

# **Statistical experiment design for animal research**

**Carlos Oscar S. Sorzano**

Natl. Center of Biotechnology (CSIC), Madrid,  
Spain

June 23, 2023

Cite as C.O.S. Sorzano. Statistical experiment design for animal research. OSF Preprints:  
[osf.io/e9s25](https://osf.io/e9s25) (2023)

*To all animals that contribute to the humankind's progress  
with their suffering and lives.*

*To all researchers who want to honor them  
by using the best scientific tools at their hands.*

# Contents

<b>Foreword</b>	<b>7</b>
<b>Part 1</b>	<b>11</b>
<b>1 Why do we need a statistical experiment design?</b>	<b>13</b>
1.1 Pilot, exploratory and confirmatory experiments	19
1.2 Independence between individuals: experimental units	25
1.3 Avoiding bias: blocking, randomization and blinding	28
1.4 Reducing variance	39
1.4.1 Variable selection	41
1.4.2 Population selection	41
1.4.3 Experimental conditions	43
1.4.4 Population scope, outliers and lack of independence	43
1.4.5 Averaging and pooling	47
1.4.6 Blocking	52
1.4.7 Paired samples	54
1.4.8 Blocking and randomization	55
1.5 Automating decision making: hypothesis testing	56
1.5.1 An intuitive introduction to hypothesis testing	60
1.5.2 Statistical power and confidence	63
1.5.3 Multiple testing	67
1.5.4 A worked example	71
1.6 A primer in sample size calculations	75
<b>2 Real experimental examples</b>	<b>83</b>
2.1 Some fully developed examples	83
2.1.1 Difference between two group means	83
2.1.2 Showing differences among different groups, ANOVA	87
2.1.3 Comparing proportions	90
2.1.4 Survival curves	94
2.1.5 A factorial design	95
2.1.6 Dose optimization	102
2.1.7 Diet optimization	103
2.2 Practical issues related to sample size calculation	104

2.2.1	An omnibus sample size . . . . .	104
2.2.2	Experiments have to be repeated three times to be statistically significant . . . . .	108
2.2.3	Treating counts as continuous variables . . . . .	112
2.2.4	Both-sexes vs One-sex or One-sex-at-a-time experiments . . . . .	114
2.2.5	Which variance to use for the calculation of the sample size? . . . . .	117
2.2.6	Foreseeing dropouts . . . . .	120
2.2.7	Pilot studies . . . . .	121
2.2.8	Decisions taken after low-powered tests . . . . .	123
2.2.9	We should not calculate the sample size for every animal experiment . . . . .	125
2.3	Practical issues related to experiment design . . . . .	126
2.3.1	Undefined goals . . . . .	126
2.3.2	Longitudinal studies of the same animal . . . . .	128
2.3.3	Unnecessarily discretizing continuous variables . . . . .	128
2.3.4	Incomplete/imbalanced factorial designs . . . . .	132
2.3.5	Animal housing constraints . . . . .	134
<b>3</b>	<b>Statistical pitfalls</b> . . . . .	<b>137</b>
3.1	Probability pitfalls . . . . .	138
3.2	Data analysis pitfalls . . . . .	141
3.3	Test selection guide . . . . .	166
<b>Part 2</b>		<b>181</b>
<b>4</b>	<b>Sample size calculations</b> . . . . .	<b>183</b>
4.1	Sample size for the mean . . . . .	184
4.1.1	Hypothesis test on the mean of one sample when the variance is known . . . . .	184
4.1.2	Hypothesis test on the mean of one sample when the variance is unknown . . . . .	185
4.1.3	Confidence interval for the mean . . . . .	187
4.1.4	Hypothesis test on the mean for paired samples . . . . .	188
4.1.5	Hypothesis test on the difference of the mean of two samples . . . . .	190
4.1.6	Hypothesis test on the mean of several groups (ANOVA) . . . . .	191
4.1.7	Unequal group sizes . . . . .	194
4.1.8	Hypothesis test on the equivalence of two means . . . . .	195
4.2	Sample size for proportions . . . . .	198
4.2.1	Hypothesis test on one small proportion . . . . .	198
4.2.2	Confidence interval for one proportion . . . . .	202
4.2.3	Hypothesis test on one proportion . . . . .	204
4.2.4	Confidence interval for the difference of two proportions . . . . .	206
4.2.5	Hypothesis test on the difference of two proportions . . . . .	208
4.2.6	Hypothesis test on the difference of two paired proportions . . . . .	209
4.2.7	Hypothesis test on the difference of multiple proportions . . . . .	211
4.2.8	Hypothesis test on the equivalence of one proportion . . . . .	213

4.2.9	Hypothesis test on the equivalence of two proportions . . . . .	214
4.3	Sample size for regression . . . . .	215
4.3.1	Linear regression: Confidence interval on the slope of a regression coefficient . . . . .	219
4.3.2	Linear regression: Hypothesis test on the regression coefficients	221
4.3.3	Logistic regression: Hypothesis test on the regression coefficients	223
4.3.4	Cox regression: Hypothesis test on the regression coefficients	227
4.3.5	Poisson regression: Hypothesis test on the regression coefficients	231
4.4	Sample size for Poisson counts . . . . .	236
4.4.1	Hypothesis test for a single population . . . . .	237
4.4.2	Hypothesis test for two populations . . . . .	239
4.5	Sample size for the variance . . . . .	240
4.5.1	Confidence interval for the standard deviation . . . . .	240
4.5.2	Hypothesis test for one variance . . . . .	241
4.5.3	Hypothesis test for two variances . . . . .	243
4.6	Sample size for correlations . . . . .	244
4.6.1	Confidence interval for correlation . . . . .	244
4.6.2	Hypothesis test on one sample correlation . . . . .	245
4.6.3	Hypothesis test for the correlations in two samples . . . . .	246
4.6.4	Hypothesis test for multiple correlation in one sample . . . . .	246
4.6.5	Confidence interval for the Intraclass Correlation (ICC) . . . . .	247
4.6.6	Hypothesis test for Cohen's $\kappa$ . . . . .	251
4.7	Sample size for survival analysis . . . . .	253
4.7.1	Confidence interval for the mean survival time . . . . .	253
4.7.2	Confidence interval for survival time percentile . . . . .	257
4.7.3	Confidence interval for survival rate . . . . .	261
4.7.4	Hypothesis test for one sample mean survival time . . . . .	263
4.7.5	Hypothesis test for one sample survival rate . . . . .	266
4.7.6	Hypothesis test for two samples with exponential survival . . . . .	268
4.7.7	Hypothesis test for two samples with log-rank test . . . . .	269
4.8	Sample size for pilot experiments . . . . .	270
4.8.1	Pilot experiments for the variance and mean . . . . .	271
4.8.2	Pilot experiments for proportions . . . . .	272
4.9	Adaptive sample size . . . . .	273
4.9.1	Hypothesis test for the superiority of a proportion . . . . .	274
4.9.2	Hypothesis test on the difference of the mean of two samples . . . . .	277
4.9.3	Hypothesis test on the difference of two proportions . . . . .	280
4.9.4	Sample size reestimation in blind experiments . . . . .	282
4.9.5	Hypothesis test on the difference of two survival curves . . . . .	282
<b>5</b>	<b>Design of experiments</b>	<b>285</b>
5.1	Basic designs . . . . .	285
5.1.1	Completely randomized design, CRD . . . . .	286
5.1.2	Regression design . . . . .	292
5.1.3	Randomized block design, RBD . . . . .	295
5.1.4	Use of covariates . . . . .	302

5.1.5	Linear models, sample size and replications . . . . .	306
5.1.6	Factorial designs, FD . . . . .	309
5.1.7	Non-orthogonal, incomplete and imbalanced designs . . . . .	320
5.2	Advanced designs . . . . .	336
5.2.1	Latin squares . . . . .	336
5.2.2	Graeco-Latin squares . . . . .	339
5.2.3	Cross-over designs . . . . .	339
5.2.4	$2^k$ Factorial designs . . . . .	344
5.2.5	$2^k$ Fractional factorial designs . . . . .	349
5.2.6	Split-unit designs . . . . .	356
5.2.7	Hierarchical or nested designs . . . . .	358
5.2.8	Fixed vs. random and mixed effects . . . . .	361
5.2.9	Multilevel models . . . . .	367
5.2.10	Loop designs . . . . .	368
5.2.11	Response surface designs . . . . .	370
5.2.12	Mixture designs . . . . .	376
5.3	Design selection guide . . . . .	380
5.4	Sample size for designed experiments . . . . .	382
5.4.1	Sample size for completely randomized and randomized block designs . . . . .	382
5.4.2	Sample size for factorial designs . . . . .	386
	Mathematical introduction to experiment design . . . . .	388

# Foreword

When we do a scientific experiment, from the statistical point of view, we can distinguish two important periods: before doing the experiment and after having done the experiment. After the experiment, we will be faced to a collection of numbers stemming from the measurements. Our goal, in a biomedical scientific context, will be to show that a drug is effective against a given disease, that a cell type is involved in some physiological process, that a gene is overexpressed under some condition, that a new vaccine is effective to protect the population, ... Statistical tools, most prominently statistical inference, will be used to discriminate the underlying signal we are interested in from the noise coming from measurement errors and biological variability.

Whether we find or not the sought effect depends on three things: 1) there is really some biological effect (e.g., the drug is really effective); 2) how much noise there is in our measurements; and 3) how much evidence we have collected to show that there is really some effect, that is, how many times we have observed this difference. From the statistical point of view, we cannot act on Point 1. But we can act on Points 2 and 3 before doing the experiment, we do not need to wait the experiment to be done to perform a “post-mortem” analysis.

Point 2 is addressed by **statistical experimental design**. This technique tries to arrange the experiment in such a way that we can identify the different sources of variation and determine which part of the variability observed in the measurements comes from our treatment (drug, vaccine, gene or cell type of interest), the signal, and which part comes from other sources such as sex, age, health condition, the experimenter doing the experiment, etc. The part of variation that we cannot explain will be the noise. We will declare that there is a biological difference if the signal is well (significantly) above the level of noise. By far, the most known experimental design in biomedical sciences is the comparison of the results from a control and a treated group. However, this is not the only one and many other designs can be conceived so that the amount of noise is minimized.

Point 3 is addressed by **the sample size calculation**. That is, how many times we repeat the experiment to determine if there is a biological effect or not. The more we measure, the surer we are about our decision (we see some effect or we do not see any effect). Hypothesis testing is a way to automate this decision in a quantitative way. Of course, we can take wrong decisions: deciding that there is an effect when there is none (false positive) or deciding that there is no effect when there is (false negative). The probability of these two types of mistakes can be controlled at will by simply choosing an appropriate sample size.

Researchers are normally much more aware of sample size calculation than statistical experimental design. Probably the reason is that it is the sample size what gives them “statistical power” to detect the differences they are interested in. However, they could have used fewer samples if they had designed the experiment following some statistical principles.

Points 2 and 3 assume that the experiment is well conducted and that the observed differences are only caused by the variable of interest (treatment, cell type, gene, ...). If there are other uncontrolled variables affecting our results (e.g., males respond to the treatment, but females do not) and these variables are not explicitly taken into account in the experimental design, then we will have biased results. That is, the observed differences are not caused by the treatment, but by something else that we do not know, making us to believe that there is a true biological effect. The presence of bias ruins our experiment as we are fooled by data whose true differences are caused by uncontrolled variables. The main three tools to fight bias are **blocking** (we identify possible variables that might affect or not the results), **blinding** (to prevent a possible biases from the researcher), and **randomization** (we randomize the samples in such a way that any possible affecting variable that has not been blocked equally affects all samples so that its effects are randomly distributed among the experimental groups).

This book addresses the statistical tools needed to tackle Points 2 and 3 as well as avoiding bias. That is, all the steps **before doing the experiment**. The book does not address data analysis (**after having done the experiment**). However, we will see that we cannot design our experiment if we do not know how the data going out from it will be analyzed. In this regard, the past and the future of the experiment are tightly linked.

Much to the despair of many biomedical researchers, Statistics is a branch of Mathematics. The simplest statistical concepts are very intuitive. However, there can be some, very profound concepts (such as the degree of freedom). Calculations can become very sophisticated requiring in practice the compulsory help of a computer. On the other hand, those researchers passionate about understanding the information coming out from a large collection of numbers will find in Statistics and Data Analysis an endless source of joy.

The book has been divided in two parts. In Part 1, we will give an intuitive overview about how to design an experiment. Very few formulas will appear, and those appearing will be as simple as possible (but not simpler than needed). Many other calculations are sufficiently sophisticated and they will appear as “black boxes” Software in this area is often used in this black-box mode. In a way, we do not need to know the specific details of the calculations. Although, we need to understand the questions asked by the software and some of them inevitably refer to deep statistical concepts. In Part 2, we present a closer overview and rationale behind some of the calculations involved in sample size calculation and statistical experimental design. We cannot cover absolutely all cases as they are many and some of them very specific. However, we cover a sufficiently wide spectrum so that the interested reader can get a glimpse of the underlying details.

Part 1 has three chapters and it is written for any researcher having to use statistical tools. In Chapter 1, we give an overview of all concepts involved in the statistical design of an experiment. In Chapter 2, we give examples of some of the most common experimental situations and address some of the most pressing questions related to the

experiment design. Most of the calculations of the sample size or related to the experiment design are performed in a “black-box” mode, those wanting to see the details are referred to the corresponding section in Part 2. In Chapter 3, we present pitfalls when understanding concepts related to probability, data analysis, and experiment design. We also give a guide for the selection of an appropriate statistical test.

Part 2 has two chapters and it is written for those researchers wanting to understand the internal calculations performed and not wanting to use the programs available as black boxes.

The book is full of examples taken from real biomedical experiments. Our goal is that researchers can identify their own work in the given examples by simply changing a few words and the context. Additionally, we have collected the main theoretical ideas into short paragraphs (marked as “Important remarks”) intuitively explaining their relevance and their consequences for the practical work in the laboratory.

Calculations have been mostly performed with NCSS PASS, one of the programs to calculate the sample size, or using homemade Matlab scripts.

Our recommendation to experimenters would be that they should read Part 1 and embrace the concepts exposed there so that they can spot possible deficiencies in their own experiment designs. However, not every researcher should become an expert in statistical experimental design and sample size calculation. They will certainly not become an expert by simply reading Part 1 of this book. Current Science is multidisciplinary, and the wide adoption of Statistics as an analysis tool should follow the same rule. We strongly recommend that researchers, in general, either they have a strong statistical background to be able to design and analyze their experiments, or they team with someone capable of doing so. The synergy coming out a positive collaboration will result in better and more ethical science, especially if we think that many experiments involve living and sentient beings.



# Part 1



# Chapter 1

## Why do we need a statistical experiment design?

Animal research is crucial for biomedical advances because animal models often show higher discrimination than many other experimental alternatives and have the necessary fidelity which may be required (Russell WMS, Burch RL. 1959. The principles of humane experimental technique. Wheathampstead (UK): Universities Federation for Animal Welfare.). Although, results with animals cannot be directly extrapolated to humans (Leist and Hartung, 2013), they provide key insight and clues about the possible behaviour of drugs and treatments in other species like ours. The European directive 2010/63/EU proposes the 3Rs (Replacement, Reduction, and Refinement) as an ethical approach to animal research, being conscious of benefits of animal experiments and the harm infringed to them.

Replacement addresses the substitution of animals by other non-sentient experimental entities (cell cultures, invertebrates, organoids, organs-on-chips, or mathematical models). For instance, it has been shown that lethal doses are better extrapolated from human cell cultures to human subjects than from animals to human subjects (Ekwall et al, 1998). Refinement refers to the way in which the experiment is conducted so that it is performed in as humane a manner as possible (distress, pain and harm are reduced; social and intelligent animals enjoy an enriched environment). The second R, reduction, is the topic of these chapters. We should aim at performing experiments with as few animals as possible consistent with achieving the aims of the experiment, which is typically to produce reproducible, publishable, statistically significant results. This reduction goal is best achieved by a thorough and detailed statistical design of the experiment, including consideration of spatial/time/treatment organization, and a calculation of the appropriate number of animals required for the experiment based on power calculation and credibility for publication. Before carrying out the experiment we should carefully design the analysis methodology so that we plan in advance the way the data will be analyzed thus preventing avoidable surprises. Another way to reduce the number of animals is by minimizing the variance of the observed values (see Sec. 1.4), or reusing them in multiple experiments. For instance, pigs and dogs

typically participate in toxicology experiments (paying special attention to the cumulated severity). Graeco-Latin squares (Sec. 5.2.2) would prevent the results of the first experiment interfering with the second one. Another example would be sharing tissues of the control animals (if a group is interested in the brain and another group in the liver, the same animal could serve for both purposes).

The refinement refers to the modification of husbandry or experimental procedures to minimize pain and distress, and enhance its welfare. It may not be easy to adhere both to reduction and refinement as these two concepts may go into opposite directions. For instance, 1) procedures can be performed such that they either inflict less harm on more animals or inflict more harm on fewer animals; 2) genetically-altered animals may be a refinement due to the increased accuracy of the animal model, but a significant number of animals are required to generate a new animal line; 3) the maintenance of a colony of immune-compromised or suffering mice while the research is temporarily suspended, because animals will be needed again in the near future.

Overall, our goal should be to perform experiments of the highest quality, reproducible and that reveal true biological effects. Much has been said about the reproducibility crisis in science. Part of it is caused by an incorrect experimental design, part of it by an incorrect use of the statistical analysis tools, part by publication bias, etc. The statistical design tools that we will learn in this book will help us in the endeavour of making science whose results can be reproduced elsewhere and whose validity can be much more trusted.

### **How to read these chapters**

The four statistical chapters of this book cover: 1) an overview of the problem and the main statistical concepts involved; 2) a calculation of the sample size; 3) a statistical plan to analyze the data that has a direct impact on the layout of the experiment; 4) some of the most common statistical pitfalls. Chapters 1 and 4 provide a very useful insight into the problem of animal experimentation and its careful design from a statistical point of view. We recommend their full reading. Chapters 2 and 3 are more reference chapters. They cover many different experimental situations and they do not need to be sequentially read. In a first pass, the reader may directly go to the examples and the important remarks. In this way, he/she will have a wide overview of the kind of problems she can encounter and solve with the tools provided by this book. All examples (more than 100 of them) have been laid in an animal research setting trying to reflect the everyday life of many researchers. The book is focused on the design of the experiment, and this aspect has been covered in depth. There would be a whole branch of statistical data analysis that has been left out of the scope of the book. Chapter 4 partially includes part of the data analysis, but only those aspects more connected to the design of the experiment and those concepts so important that failing to consider them would spoil the experiment from the very beginning.

### **Statistical experiment design**

A careful statistical experiment design involves three steps:

1. Objective design: we should clearly set from the very beginning the objective of our experiment (e.g., measure the effect on sugar concentration in blood of a new drug treatment for type II diabetes animals). With this objective in mind, we should choose:
  - the species, stocks and strain(s) of animals that will better allow extrapolation to other species, like humans.
  - the variable to measure, in this example, the concentration of glucose in blood plasma measured 4h after food intake when the treatment, at different doses, has been given for 2 weeks every 8 hours.
  - the experimental groups we will compare, e.g., a control diabetic group with an experimental diabetic, treated group. Sometimes, positive and negative control groups are included along with the treatment group. That is, a group in which there should not be any response (negative control) and a group in which all animals should respond (positive control). This can be easily accomplished in experiments involving a dose: zero dose would be the negative control and a large dose would be the positive control. Both, negative and positive, controls help to identify situations in which the treatment has been contaminated (negative control) or inactivated (positive control). We may also include groups that receive surgery without the therapeutic step (these are called sham controls), groups that receive a competitive treatment (like the reference drug in an experiment in which we are developing a new drug), or naïve controls that do not receive any treatment or operation (they give information about the effect of time, weather or other experimental conditions apart from the treatment). Multiple comparisons can be performed within a single experiment as long as all of them were planned in advance.
  - the test we will use to verify whether the treatment has an effect, e.g., a t-test for the difference in the mean assuming unequal variance in both groups.
  - a target difference so that we can determine when the treatment is successful or not, e.g., if the expected glucose level in diabetic mice is about 300 mg/dL with a standard deviation around 40 mg/dL, we want to be able to detect reductions of at least 100 mg/dL in the glucose blood concentration (we will assume that the standard deviation in the treated group is similar, although not equal, to the standard deviation in the untreated group).
2. Sample size design: To be able to detect a difference of 100 mg/dL when the standard deviation is 40, with a statistical power of 90% and a confidence level of 95%, we need 5 mice per group (Mathews, 2010)[Chap. 2]. The confidence level of 95% implies that if we repeat this experiment many times with 5 mice in each group, just by chance, we will erroneously find in 5% of them that our treatment is useful to cause such a reduction in the blood glucose level, when actually it does not have any effect. The statistical power of 90% means that in these many repetitions of our experiment we will erroneously find useless 10%

Table 1.1: Experiment design: number of animals in each one of the groups.

	Female treated	Female control	Male treated	Male control
Morning	2	1	1	2
Afternoon	1	2	2	1

of the treatments that actually have such a large effect, but simply because we had “bad luck” with our samples, the observed difference is not significant. The number 5 mice per group is a statistical constraint derived from the way we will analyze our data once the experiment is performed (t-test). However, we may add other experimental constraints, e.g., we may add an extra mouse per group to account for the fact that sometimes, for whichever experimental reason (incorrect blood extraction, environmental contamination, etc.), our measurements are not valid. We do not expect these accidents to occur very frequently and we do not foresee that they may happen more than once per experiment. Since we do not know in which group it will happen, we may add 1 mouse to each one of the groups as a safeguard that, in case such accidents occur, we still have the 5 mice per group we need for the statistical comparisons. In this way, we will perform our experiment with 6 individuals per group.

Too many animals in an experiment is a waste of economical, laboratory and human resources. Too few animals will spoil the experiment because even if the effect we seek for is present, we will not have enough samples to show that it is statistically significant. Both cases (too many and too few) call for our ethical responsibility because the treatments and conditions applied to the research animals are harsh.

3. Experimental layout design: We know that the mean glucose level in blood depends on the sex of the animal and the time of the day. If we put all male animals in the control group and all the female animals in the treated group, we cannot know if the difference observed between the two groups is caused by the treatment or by the different sex of the subjects. The same would happen if we measure all the treated animals in the morning, and all the control animals in the afternoon. The observed difference might be caused by the time of the day (morning or afternoon) that we take our samples, and not by our treatment. Statistically, this uncertainty is called confounding (we are confounding sex or daytime with the treatment), and the way to avoid it is by designing balanced, blocking experiments (we block the two variables that are not of our interest at the moment, we are interested in the effect of the treatment and not in the effect of sex or daytime). The trick is to assign the same number of treated and control animals to each one of the levels of the variables to block (3 control males and 3 treated males, 3 control females and 3 treated females, 3 control measurements in the morning and 3 treatment measurements in the morning, ...). We may organize our experiment as shown in Table 1.1. There might be other variables to block as the technician carrying out the extraction or taking care of the animals,

Table 1.2: Experimental measurement plan

Morning	<ul style="list-style-type: none"> <li>(a) Male control</li> <li>(b) Female treatment</li> <li>(c) Female treatment</li> <li>(d) Female control</li> <li>(e) Male treatment</li> <li>(f) Male control</li> </ul>
Afternoon	<ul style="list-style-type: none"> <li>(a) Male treatment</li> <li>(b) Male treatment</li> <li>(c) Female control</li> <li>(d) Female control</li> <li>(e) Male control</li> <li>(f) Female treatment</li> </ul>

work shifts of the animal facility staff, the cage from which the animals come from, the preparation of the drug which is used along the treatment, ... We can block as many variables as we suspect that may cause a difference in the measurements. Obviously, the more variables we want to block, the more animals we will need to keep the design balanced. However, there might be unsuspected variables making a difference but we did not foresee (position of the cage in the shelf, time the blood sample is waiting for analysis, ...). For this reason, it is not advised to perform the experiment following a fixed pattern. For instance, all females first, then all males, treatments always before controls (as shown in Table 1.1). A statistical mantra we should keep in mind is *control what you can, block what you cannot, and randomize the rest*. Consequently, inside each block, we should randomize the order in which measurements are performed and establish a measurement plan (see, for instance, Table 1.2). It is important that the randomization is performed by a computer and not by a person because humans tend to create regular patterns when we randomize (Schulz et al, 2012). An advantage of animal research as opposed to clinical trials is that the researcher can plan and control many more variables in the experiment than when humans are involved. This is also a responsibility because the success or failure of our experiment depend more on our ability to carefully design the experiment.

Nowadays, there are multiple software that allow us to calculate the sample size and the experiment design. However, they should be used with care. Software can be enormously helpful in taking the tedium out of power calculations, but is only effective if we understand the principles behind what we are doing. Table 1.3 lists a range of

power calculation software, applets and online resources. This is not a comprehensive list, but lists a useful range of power calculators. The licensed power calculators tend to be comprehensive, relatively easy to use, have good help files, and help available but cost money. The free calculators tend to be less comprehensive with less help available. There are also power calculators aimed at specific purposes.

Software	Cost	Platform	Uses
GWA Power	Free	R	Genome Wide Analysis
powerSurvEpi	Free	R	Survival analysis
Optimal Design	Free	Windows	CRT
Power V3.0	Free	Windows	Logistic-like regression
Russ Lenth's Power Calculator	Free	Windows	Range
G*Power	Free	Windows	Range
PS Sample size and Power	Free	Windows	Range
Powerandsamplesize.com	Free	Online	Range
Sorzano Pilot sample size	Free	Online	Pilot studies
NCSS PASS	License	Windows	Comprehensive
Nquery	License	Windows	Comprehensive
StatMate	License	Windows	Comprehensive
Design Expert	License	Windows	Comprehensive
Power and Precision	License	Windows	Range

Table 1.3: List of software tools that can be used to design an experiment. They cover sample size calculation, experimental design or both.

Compare the difference between this careful experimental design before carrying out the experiment, and the experiments performed “to see what happens” or without taking the necessary precautions (blocking and randomization). In the long term, careful statistical designs save animal lives, reduce the harm infringed on animals, reduce research time and costs, increase research quality and reproducibility, allow publication and promote ethics in science. [Gore and Stanley \(2005\)](#) shows the problems that incorrectly designed experiments may face. The ARRIVE ([Percie du Sert et al, 2020](#); [Kilkenny et al, 2010](#)) and PREPARE ([Smith et al, 2018](#)) guidelines aim at improving the design and reporting of biomedical experiments once they are performed, including the communication of their statistical aspects. From the first appearance of the ARRIVE guidelines (2010) to its current publication (2020), there has been a huge shift towards promoting statistical aspects as the most important ones in preparing the experiment. Tables 1.4 and 1.5 show what in the current guidelines are called the “Essential 10” showing that the experiment design, before the experiment is carried out, must be given primary importance ([Festing and Altman, 2002](#); [Festing, 2003](#)). The Experimental Design Assistant has been designed to enforce application of the ARRIVE guidelines and contains useful tools to help with the design and implementation of experiments, and, if followed, should lead to more reproducible and better designed experiments. Unfortunately, good statistical experimental design and reporting is not always the rule:

- [McCance \(1995\)](#) surveyed 133 papers commissioned by the editors of the Aus-

tralian Veterinary Journal. In the opinion of the statistician: 61% would have required statistical revision before publication, 5% had such serious errors that the conclusions were not supported by the data, 30% had deficiencies in design of the studies including failure to randomize, inappropriate group size, heterogeneity of subjects and possible bias, 45% had deficiencies in the statistical analysis including the use of sub-optimal methods and errors in calculation, 33% had deficiencies in presentation of the results including unexplained omission of data and inappropriate statistical methods.

- [Kilkenny et al \(2009\)](#) surveyed a random sample of 271 papers involving live mice, rats or non-human primates. They found that of the papers studied: 87% did not report random allocation of subjects to treatments, 86% did not report “blinding” where it seemed to be appropriate, 100% failed to justify the sample sizes used, 5% did not clearly state the purpose of the study, 6% did not indicate how many separate experiments were done, 13% did not identify the experimental unit, 26% failed to state the sex of the animals, 24% reported neither age nor weight of animals, 4% did not mention the number of animals used, 35% reported the numbers used, but these differed in the materials and methods and the results sections.

In this chapter we review the principles of statistical experiment design. It is aimed at biomedical researchers undertaking experiments with animals at any level, but especially those having to design the experiment; this is normally PIs, senior researchers and postdocs. Statistics is a branch of Mathematics and, as such, it is difficult to get away without any mathematical formula. However, our aim is to keep these to a minimum, showing only those that are key to understand the basic statistical concepts. We try to present the intuition behind them and its practical consequences. Most of us use our mobile phones without understanding how they work, and that does not prevent us from finding them very useful. Unfortunately, using Statistics is not like using a mobile phone: 1) it is considerably more complex; 2) we must plan costly experiments that have an important component of statistical planning; and 3) we draw conclusions from the statistical analysis with very important scientific, economical and ethical consequences. For these reasons, we cannot blindly use Statistics without a minimal understanding of its mechanisms. Normally nobody dies if we do not use our mobile phone correctly, but poor experimental design can lead to unnecessary suffering and death of many animals. We understand that not every researcher needs to be a deep expert in Statistics, and our presentation provides, we think, the minimum requirements for understanding the standard use of Statistics in a research laboratory. We strongly encourage experimental researchers to team with other researchers with a sufficient understanding of the statistical concepts they need for their work, so that we use research grants and animal lives in the most efficient and ethical manner.

## 1.1 Pilot, exploratory and confirmatory experiments

A possible classification of experiments may distinguish between **pilot**, **exploratory**, and **confirmatory** experiments.

20 CHAPTER 1. WHY DO WE NEED A STATISTICAL EXPERIMENT DESIGN?

Table 1.4: Items of the Essential 10 of the ARRIVE 2.0 guidelines. These aspects need to be determined before executing the experiment. For each one of them, we give a pointer to a section of this book that is helpful in that regard.

<p>1. Study design</p>	<ul style="list-style-type: none"> <li>• Identification of the experimental unit. Sec. 1.2.</li> <li>• Selection of the experimental groups (negative and/or positive controls, treatment groups). Sec. 1.4.</li> <li>• Selection of factors, blocks, the combination of these that will be tested and the mathematical model to be used during the experiment analysis. Chap. 5.</li> </ul>
<p>2. Sample size</p>	<ul style="list-style-type: none"> <li>• Specification of the statistical objective (hypothesis testing or construction of confidence intervals and the variable on which the objective will be calculated). Related to Items 6 and 7. Sec. 3.3</li> <li>• Calculation of the number experimental units needed to achieve a given statistical power and confidence level to be able to detect a given effect size (hypothesis testing) or to construct a confidence interval of a given precision (confidence interval). If the experimental unit is not the animal, the number of animals should also be reported to help the reader to understand the design. Chap. 4.</li> </ul>
<p>3. Inclusion and exclusion criteria</p>	<ul style="list-style-type: none"> <li>• Inclusion criteria define the eligibility of the animals to participate in the study once this has started (e.g., they must be able to perform a given task).</li> <li>• Exclusion criteria define reasons for disqualifying the animals during the study (e.g., complications during surgery, developing a motor impairment that interferes with performance measures or not meeting a quality control standard, such as insufficient sample volumes, unacceptable levels of contaminants, poor histological quality). These criteria should be defined before starting the experiment</li> </ul> <p>Both inclusion and exclusions will result in dropouts which should have been foreseen in the calculation of the sample size. Sec. 2.2.6.</p>
<p>4. Randomisation</p>	<p>Randomisation of animals into the treatment groups guarantees that each animal has the same probability of receiving any of the treatments. Statistical inference on non-random groups are not valid. Sec. 1.3.</p>
<p>5. Blinding</p>	<p>Researchers normally have a preferred hypothesis in the experiment and may unintentionally interfere with its results. Researchers should be blind to the group allocation, during the conduct of the experiment, the assessment of the outcome, and the data analysis. Sec. 1.3.</p>

Table 1.5: Items of the Essential 10 of the ARRIVE 2.0 guidelines. These aspects need to be determined before executing the experiment. For each one of them, we give a pointer to a section of this book that is helpful in that regard.

6. Outcome measures	Variable(s) of interest in the experiment. Its variability and our desired precision in statements about it (hypothesis testing or confidence interval) determine the sample size (Item 2). Sec. <a href="#">1.4</a> .
7. Statistical methods	The statistical methods employed will determine the objective of the experiment (comparing several treatments, determining an effect with a given precision, etc.) Once these methods are fixed, we will be able to calculate the sample size (Item 2). The analysis pipeline should be outlined and how missing values will be handled, too. Sec. <a href="#">3.3</a> .
8. Experimental animals	Specification (before the experiment) and reporting (after the experiment) the animal characteristics (e.g., species, strain, substrain, sex, weight, and age). The choice of these features may affect the variance of the observations. Sec. <a href="#">1.4</a> .
9. Experimental procedures	Specification of the pharmacological, surgical, pathogen infection, measurement, ... procedures. Sometimes, there are multiple ways of performing the same task and some of them may induce larger variance than others. Sec. <a href="#">1.4</a> .
10. Results	Description of the observations once the experiment is carried out. They include summary statistics, histograms, quantiles, outliers, plots, ... and any other representation that help the reader to understand the outcome of the experiment.

**Pilot experiments** are small studies (with 1-20 experimental subjects) aiming to determine the scale of a variable (e.g., if extraterrestrial aliens just arrive to Earth and they want to measure the order of magnitude of human height they just need one sample to know that humans typically measure between 1 and 3 meters, we are not in the order of the millimeter nor in the order of the kilometer). Pilot experiments are designed to allow researchers: to gain familiarity with the experimental material, make sure that the instructions are understandable and can be followed, ensure that all steps in the procedure can be performed, check that the staff is sufficiently trained in the necessary procedures, check the correct operation of the equipment, detect a floor or ceiling effect (e.g., a task is too difficult or too easy resulting in skewed results), assessing that the level of intervention is appropriate (e.g., the dose of a drug), identify adverse effects (pain, suffering, distress, or lasting harm) and the effectiveness of the actions to reduce them (e.g., analgesia dose and schedule), verify that the procedures are not too mild or severe, define early humane end-points, and gain some information on the variability (although with so few individuals this information is not sufficient to allow a robust calculation of the sample size and can only give us a “ball park” estimate, [Sorzano et al \(2017\)](#)). Given that the control in planned experiment most likely has also been the control in other experiments, it is possible and desirable to base estimates of variability on related prior data from the researcher’s own laboratory since this is typically more robust than that from a small sample. Alternatively data from similar experiments in the scientific literature may be used if existing data is not available within the researcher’s own laboratory. Typically, there will be a range of values available and one may choose an appropriate value. We may assume similar variability for the treatment group or even increase it a bit (10%-20%) as a safeguard for possible larger variance.

**Exploratory experiments** can be used to generate hypothesis for further testing. They may not have a clear objective from the beginning and they respond to “let’s see what happens if ...”. Often they measure many characteristics of the individuals and identify interesting differences between groups that may even be statistically significant. This is, for example, the case of many gene expression experiments measured with microarrays that determine the expression level of thousands of genes. However, the identified differences should be further tested in a confirmatory experiment, in which the research hypothesis is set from the very beginning. The problem with exploratory experiments is that they determine the hypothesis after seeing the data, this is called *data snooping*, *data fishing*, *data dredging* or *p-hacking* and it may lead to severe bias simply because the observed differences are not due to any underlying scientific reason, but just by chance and the specific random response of the animals at hand. That is why the new hypotheses need to be confirmed in a confirmatory experiment. Exploratory experiments are similar to an experiment in which we give an exam to a number of students, we then score the exams, and sort students according to their performance. We measure many characteristics of the students and realize that the first five students they are all Aquarius (or girls, or wear blue jeans). We cannot immediately conclude that being Aquarius (or girl, or wearing blue jeans) gives an additional advantage in this kind of exam. These observations may be produced just by chance and the fact that we have measured many characteristics, and a few of them, randomly resulted in statistically significant differences between the top of the list and the rest of it. Remember that the 95% level of confidence of hypothesis tests implies that in 5% of

the tests, the test will result in a statistically significant difference when actually there is none. This problem is called multiple testing and there are ways to minimize its impact that we will cover in subsequent sections. But if we measure the expression level of 20,000 genes and we do not take special precautions, we should expect that we find, on average, 1,000 genes reporting erroneous significant differences between groups. There might be a true reason behind these results (the gene A is really differently expressed in the two groups, or Aquarius people are born at the beginning of the year and this small difference gives them an advantage) or they may be simply analysis artifacts. In any case, the results from an exploratory experiment should be confirmed by a confirmatory experiment. Finding significant differences in exploratory experiments is not the end of the analysis, but should rather be seen as the start of the analysis procedure where differences seen should inform a coherent hypothesis which can be examined in the data set using a variety of different measures. This “triangulation” of results should reduce blind acceptance of false positives.

**Confirmatory experiments** involve comparisons between two or more groups. They are normally set to test a null hypothesis (normally the absence of difference among the groups). If the null hypothesis is rejected with a level of confidence, say 95%, it means that if the null hypothesis is true, observing differences as large as the ones observed in our experiment would only occur in 5% of the cases, meaning that very likely the difference is caused by the treatment. Still, there is a 5% of probability that the result is an artifact (called Type I error or false positives) coming from the sampling variability (this statement can be refined, and we will do it later, looking at the observed  $p$ -value). If we fail to reject the null hypothesis, it does not mean that the treatment has no effect, it means that this experiment cannot show that it has an effect (we will discuss more about this issue in Sec. 3). The experimental subjects in a confirmatory experiment must be independent from each other (technically, they must be experimental units, which will be defined below). To minimize bias, the experimental subjects should be assigned to the experiment groups at random and the experimenter should be blind with respect to the treatment being applied.

One of the assumptions of confirmatory experiments is that there are no systematic differences between the groups being compared apart from the treatment applied to each one of them. Results of these experiments are biased if the effect of some external, uncontrolled variable is confounded with the treatment. For instance, there is a significant negative correlation between the purchase of warm clothes and the purchase of ice creams. The reason is not that as people spend less money in warm clothes, they have more spare money that they can spend in purchasing ice creams. There is a common reason, weather, such that cold or hot weather is causing the purchase of warm clothes and ice creams. In general, if we find a relationship  $A \rightarrow B$ , there might be a common cause  $C$  that is causing both,  $A \leftarrow C \rightarrow B$ , and once we account for  $C$ ,  $A$  and  $B$  are unrelated. We may carry out an experiment to measure the relationship between drinking coffee and cardiovascular disease in humans. In our experiment it seems that heavier coffee drinkers have higher risk of cardiovascular disease. However, in this result smoking is an external variable we are not controlling, and it may be that heavy smokers are also heavy coffee drinkers, and the higher cardiovascular risk is caused by smoking and not by coffee. In animal experiments, we may encounter the same situation, but in a much less obvious way. We may find relationships between the expression

levels of genes *A* and *B*, but they may be causally unrelated, existing a gene *C* that is related to both *A* and *B*. We therefore need to be careful in assuming cause and effect. Other typical confounders are circadian rhythms, atmospheric pressure, the location of the animal in the animal house (as they may have different temperature, humidity and light levels), or a growing skill of the researcher doing the surgery.

Confirmatory experiments should be designed to be powerful: if there is a relevant difference, we should be able to detect it. There are three ways of making an experiment more powerful:

- *Increasing the number of animals.* Using enough individuals so that, if there is a difference, the p-value can be shown to be below the significance limit. Statistical power should be one of the parameters in the calculation of the sample size. Typical statistical powers are 90% or 80%, meaning that if the treatment makes a difference of a specified size (the 100 mg/dL in the example of Sec. 1) we will be able to detect it in 90% of the experiments (and we will miss it with probability 10%, these are called Type II errors or false negatives). More statistical power will require more independent samples in the experiment.
- *Decreasing the variance of measurements.* The statistical power depends on the effect size we look for (the larger the effect size, the smaller the number of animals) and the variability of the measurements. In this way, another way to reduce the number of animals and/or increase the statistical power is by decreasing the variability of the observations by using more precise laboratory analytic tools, measuring variables with less variance that are also related to our objective, decreasing the genetic variability of the individuals used in the experiment, etc.
- *Increasing treatment effect.* Where drug treatments are used, pilot experiments with varying doses may yield an optimal dose that maximizes treatment effect.

Confirmatory experiments should be designed with wide applicability in mind: the results should hold true independently of relevant variables like sex, strain, different diets and environments (a potent antihypertensive drug that is only useful for male, C57BL/6 mice under a very restrictive diet is not very useful for the general population). This applicability condition implies that the experiment should consider variations at least in a few relevant variables (factorial or randomized block designs help in this regard with a very little extra cost, see Secs. 5.1.3 and 5.1.6). When the results of an experiment can be extrapolated to a wider population, it is said to have external validity. Internal validity refers to the possibility of repeating the experiment and getting the same result. Unbiased and statistically powerful experiments have internal validity, meaning that there is a low probability of obtaining false positive or false negative results. Many biomedical experiments are performed in very controlled environments and with a limited number of animal strains (sometimes only one). This provides internal validity, but it does not provide external validity. There is nothing wrong with this as long as the scope of the experiment results are clearly stated, and no overstatements are done. However, there is an easy way of gaining external validity. If we identify the factors that affect our results (age, sex, initial health condition, intensity and duration of the treatment, etc.) we may vary them as much as possible within our experiment

using a factorial design (see Sec 5.1.6). In these designs, the effect of the treatment we are interested in has been tested in many other situations defined by the factors.

Before starting the analysis we should have set from the very beginning a plan for its statistical analysis (this is an absolute requirement for experiments performed under Good Laboratory Practices, [Macleod et al \(2009\)](#); [Kilkenny et al \(2010\)](#)). We will see that the calculation of the sample size directly depends on the way we will analyze the resulting data. This analysis plan should not be so complex that it is relatively easy to make a mistake along the process. The most powerful statistical techniques applicable to our problem should be employed, and the next experiment should be performed once we know the results from the previous one, so that we can refine the next experiment with the newly acquired knowledge.

## 1.2 Independence between individuals: experimental units

Experimental units are the smallest division of the experimental material such that any two experimental units can receive different treatments. If the sample size in each group required to detect a given difference is  $N = 6$ , it means that we need 6 experimental units to perform our experiment. The concept of experimental unit is better exemplified by specific cases:

- Example 1: We are studying the effect of additional supplements of a growth hormone on the body weight of mice. After regularly giving the hormone for two weeks, we will measure the weight of the treated animals and compare it to the weight of a control group.
  - Example 1.a: We have 6 animals in a cage and we give the hormone to each animal through an injection. Since each animal can receive the hormone or not independently of the others, each animal is an experimental unit and its weight provides an independent measurement for the statistical analysis. However, all 6 animals are in the same cage and there might be cage effects (subclinical infection, animal fighting, ...) that would affect all the animals in the same cage. The cage acts as a block, and its effects can be identified as shown in Sec. 1.3.
  - Example 1.b: We have 6 animals in a cage and we give the hormone through the food. Since all animals eat from the same feeder, each animal cannot receive the treatment or not independently of the others in the same cage. In this case, the weight of each animal does not provide an independent measurement. We have a single experimental unit, the cage (the reason is that we do not know how much each animal has received, for instance, it might be that the most dominant animal has received a higher dose while other animals may have received significantly less). If we need  $N = 6$  experimental units per group, we need  $N = 6$  cages. The independent measurement provided by the experimental unit, the cage, is the average of the weights of the animals inside that cage. On one side, we need more animals with respect to the case of independent treatments (Example 1.a). On the other side, this

increase is compensated by the fact that the variability of the mean of the animals in the cage is smaller than the variability of each animal (because the average divides the variance by the number of averaged elements).

- Example 2: We are studying the effect of additional supplements of a growth hormone on the body weight of mice. After regularly giving the hormone for two weeks to pregnant female mice, we will measure the birth weight of the offspring of the treated animals and compare it to the birth weight of a control group. This example is similar to Example 1.b because each of the newborns cannot be independently given the treatment. The average of all newborns from the same mother is giving a single independent measurement. The experimental unit is the mother, not each of the little mice. [Lazic \(2010\)](#) and [Lazic et al \(2018\)](#) extensively discuss this kind of designs in which the different measures are called pseudoreplications.
- Example 3: We are interested in the effect of some ophthalmic drops on the recovery of conjunctivitis. We will compare the difference between our new drops and some reference drops in the market (control). The same animal can be given the new drop in the left eye and the control drop in the right eye (or vice versa). In this way, each animal serves as its own control and the intersubject variability is strongly reduced. The experimental unit is each animal and its corresponding measurement is the difference between the treated and control measurements for each eye. Consequently, we have only one (independent) measurement per animal, not two. This kind of data is called paired data.
- Example 4: We are interested in the effect of four different analgesics (A,B,C,D). After a sufficiently long washout period, it is assumed that the effect of each analgesic is completely cleared from the animal body. We will measure the effect of the analgesic through a standard pain test. Each animal can be sequentially given the analgesics, with the corresponding washout periods, and measured its sensitivity to pain under each one of them. These designs are called cross-over designs, and again each animal serves as its own control, thus reducing the intersubject variability. Because of the randomization principle referred above, the sequences normally vary from animal to animal (ABCD, DCBA, DACB, ...). The experimental unit in this case is the combination of animal and time period because for each animal and time period a different treatment can be given, independently of the rest.
- Example 5: We are interested in the relationship between depression and pain sensitivity. For testing this association we will study the pain sensitivity through a standard pain test of two rat outbred stocks: WKY rats that are a model of depressive rats and Wistar rats that are not depressive and will serve as control. In this case, the experimental unit is the stock, because for any animal within a stock we cannot change its treatment (the treatment is actually the stock it belongs to). Consequently, in this experiment we only have  $N = 2$  experimental units.

- **Example 6:** Many electronic devices regularly record physiological parameters (for instance, blood glucose level every 5 minutes). Each one of the measurements is not an experimental unit. This kind of data is better analyzed with repeated measures ANOVA or time series techniques. In this case, the experimental unit is each one of the animals carrying the measurement device. Technically, each one of the measurements is called an observational unit.
- **Example 7:** Giving twice a new drug to the same animal does not bring two independent experimental units, because the individual is the same and the measurements are not completely independent (for instance, within the expected variability between animals of a class, this particular animal may have a particularly high response, making us think that the drug is very effective).
- **Example 8:** There are experiments in which the experimental unit may change among treatments. Let us think of the route of administration of a drug. We compare the effect of injecting the drug (the experimental unit is the animal) to the effect of giving it with the food (the experimental unit is the cage). In this case, we should compare groups with the same number of animals. A possible design would gather all the observations from a cage in which all animals were injected the treatment into a single measurement. In this way, this observation is comparable to the one in which the treatment was given in the food.

#### Important remarks

1. An experimental unit is the smallest division of the experimental material such that any two experimental units can receive different treatments.
2. When an animal is given the treatment once, and measured multiple times, each one of the measurements is called an *observational unit*. This kind of designs are called nested designs or repeated measures and should be analyzed as described in Sec. 5.2.6.
3. If the independence between samples is compromised, data appears to be less variable than it is in reality. This artificial reduction of variance can be compensated if we measure the Intra-Class Correlation (ICC). The interested reader is referred to Secs. 4.6.5 and 5.2.9 for details on how to use it.

Being extremely important for the statistical analysis, unfortunately the concept of independence is relative to our research objective. To explain this assessment let us briefly introduce DNA microarray experiments. Animals are given a treatment. We assume that different treatments will have different effects on the mRNA expression in different tissues. Then, we extract samples from the tissues of interest, isolate the mRNA, reverse transcribe it to cDNA, dye the cDNA and hybridize the cDNA with DNA probes. We may analyze several animals (biological replicates), we may repeat the process of reverse transcription and dyeing (technical replicates of the first experimental stage), and we may repeat the hybridization with the DNA probes (technical

replicates of the second experimental stage). If our goal is to characterize the effect of the animal treatments, our experimental units are the animals. However, if our goal is to characterize a particular sample, then the technical replicates of the two experimental stages (mRNA reverse transcription, dyeing and probe hybridization) can be viewed as independent samples (Churchill, 2002).

### 1.3 Avoiding bias: blocking, randomization and blinding

Technically, a statistic is biased if it is estimated in such a way that the expected value of our calculation is different from its true value. The calculations of statistics is at the core of all hypothesis testing and we may find significant differences due to other reasons other than our treatment. In this sense, the presence of statistical bias totally invalidate the conclusions from our study. We have already seen the bias induced by the confounding of other variables (see Sec. 1). However, bias can be caused by many other factors. Some of them are less important in animal experiments, but all of them are important in general biomedical research:

- **Omitted-variable bias** is caused by not including a variable in a regression when it has a significant influence on the measurements. For instance, not including the animal age in the level of some hormone in blood. The confounding bias we saw in Sec. 1 is a bias of this type since we are not accounting for the systematic differences induced by the different levels of an important variable (like performing the experiment in the morning or in the afternoon in our example).

Actually, the bias appears when there is a relationship between the predictor variable used in the regression and the variable left out in the regression. The following example gives the intuition behind this problem. Suppose we are interested in some observations  $y$  that depends on two predictor variables,  $x_1$  and  $x_2$ . The true model would be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Let us also assume that there is some linear relationship between  $x_2$  and  $x_1$  such that

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

Let us pretend that we do not know the dependence on  $x_2$  and simply explore the dependence on  $x_1$ . Then, the true relationship between  $y$  and  $x_1$  would be

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 (\gamma_0 + \gamma_1 x_1 + \delta) + \varepsilon \\ &= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) x_1 + (\varepsilon + \beta_2 \delta) \\ &= \beta'_0 + \beta'_1 x_1 + \varepsilon' \end{aligned}$$

If there is no relationship between  $x_2$  and  $x_1$ , then  $\gamma_0 = \gamma_1 = 0$ , and our estimates of the relationship between  $y$  and  $x_1$  are the same, that is,  $\beta_0 = \beta'_0$  and  $\beta_1 = \beta'_1$  (there is no bias, but there is an increase in the variance of the residuals  $\sigma_{\varepsilon'}^2 < \sigma_{\varepsilon}^2$ ).

If there is a relationship between  $x_2$  and  $x_1$ , then our estimates  $\beta'_0$  and  $\beta'_1$  will be different from the underlying true values  $\beta_0$  and  $\beta_1$ . Our reasoning has been illustrated with linear models, but the same kind of bias applies to non-linear relationships.

This kind of bias can be avoided by including in the analysis as many variables as can reasonably be related to our observations,  $y$ . That is, trying to identify as many sensible predictors as possible. This does not come for free as we need degrees of freedom (i.e., a larger sample size) to be able to estimate all the extra parameters.

- **Selection bias** is caused by some individuals being more likely to be selected than others. For instance, if the experimenter has to take an animal at random from a cage, those animals more quiet, docile or less aggressive may be chosen more often biasing the sample. We also have a selection bias if we take individuals out of the population of interest (for instance, we aim at young animals, but we also include in the study animals that cannot be considered young anymore; this situation is called overcoverage) or we systematically miss part of the population of interest (we do not have any young animal with a particular phenotype that might be relevant for the results of our study; undercoverage). Another example of selection bias occurs if we try to avoid assigning less healthy animals to the high dose group.

In metaanalyses in which we review a question, such as “What is the effect of treatment X on condition Y?”, by examining previously published research papers on the topic, we may also have an important selection bias: which previous studies are considered and which are not, what is exactly X or Y (for instance, X can be an anticoagulant or an oral anticoagulant, and Y can be heart failure, or heart failure with a given severity). Metaanalyses are also affected by publication bias, as negative results tend to be unpublished or, at most, published in the form of reports or as notes in trial registries. For this reason, it is also recommended to include some of these grey literature results. For a complete review on how to avoid selection bias in metaanalyses, the reader is referred to [McDonagh et al \(2013\)](#).

- **Performance bias** is involuntarily caused by the vested interest of researchers. If they are developing a new drug and comparing it to the vehicle alone, they inadvertently may take more care in administering and measuring those animals receiving the new drug than those animals receiving the vehicle. Another example would be if sick animals in the control group are given the benefit of the doubt and kept longer than animals in the high dose group. A final example is if one of the treatments is more difficult to apply and only one researcher is capable of applying it, while all the other animals are treated by the rest of the research group.

This kind of bias can be avoided if the researcher is blind to the treatment used in each animal.

- Observer bias is caused unconsciously by the prejudices of the experimenter, especially when the experiment requires some kind of subjective grading of animal behaviour or scoring histological section, for instance. Objective measures, like the glucose level in blood, are less prone to this kind of bias.

This kind of bias can be avoided if multiple, blind observers evaluate each one of the samples.

- Exclusion or subpopulation bias is caused by a systematic exclusion of a certain type of individuals from the study, for instance outlying measurements not coming from measurement errors but from the underlying biological variability. Actually, we may have multiple biological subpopulations in the feature we are analyzing. In Fig. 1.1 we show the statistical distribution of the observations we would have from a biological feature of interest when we have three different subgroups. The majoritarian subgroup (80%) has a mean of 14 and a standard deviation of 2. The second largest, 15%, subgroup has a slightly larger mean response, 15, and smaller standard deviation, 1. Finally, a small subpopulation, 5%, has a smaller mean response, 11, and an intermediate standard deviation, 1.5. The mean of these three subpopulations is still 14, although the most observed value is around 14.5. If we do not know these decomposition in subgroups, we may be tempted to say that the mean is 14.5 and that the left extreme from the distribution are caused by outliers and remove them from the analysis.

Additionally, most statistical tools are designed to deal with a single homogeneous population, and this is definitely true for parametric tools such as Student's t-test, ANOVA, Ordinary Least Squares regression, etc. meaning that the p-values and confidence intervals estimated by these tools are necessarily wrong when dealing with subpopulations.

If the presence of subpopulations can be determined, for instance, through clustering or classification according to some biomarkers, then it is recommended to study each one of the subgroups independently using the standard statistical tools. If this decomposition is not possible, then we should always bear in mind that biologically, it is very likely that we have a mixture of subpopulations each one responding in a different way and that, consequently, we should not take the p-values as written in stone.

- Attrition bias is caused by a systematic loss of individuals from the study for a reason related to the treatment. For instance, the treatment may induce in some animals some form of severe harm that forces their sacrifice. The measurements at the end of the study will not include the measurements from the sacrificed animals. This may be highly problematical given that any study should include humane end-points and animals may be lost from many studies. In a time course this will tend to produce a 'ceiling effect' which will tend to reduce variability and treatment effect. A derived measure such as specific growth rate or survival analysis on time to endpoint may be useful here.

In general, dropouts will affect the statistical power of the study and the balance of confounders within the group.

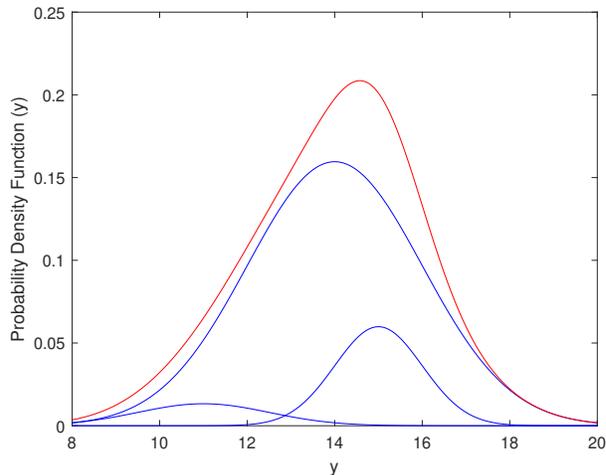


Figure 1.1: Example of the probability density function of a measurement  $y$  (red) when it is composed of multiple subpopulations (blue).

- Follow-up bias is caused by taking the decision to further pursue a research line after observing some interesting difference in a pilot experiment (Albers and Lakens, 2018). Pilot experiments usually have a small number of samples and their results have a large variance. If we do not see an interesting difference in the pilot, we will not further explore the topic (maybe missing true differences). And the opposite, if we see an interesting difference in the pilot, we will further explore it despite the fact that the apparently large difference was simply caused by the large variance induced by the low sample size.

There are other types of less technical biases like publication bias (negative results are seldom reported (ter Riet et al, 2012; Macleod et al, 2015)), but they would actually be very useful to reduce the number of false positives (Simonsohn et al (2014), if our experiment is similar to many other experiments that failed to show a given effect, we would rethink more often whether our results show a true effect or it is just the result of chance; the problem that the failing experiments are never published and we think that we are the first ones that have ever performed an experiment of the sort and, luckily or smartly enough, found a positive effect, this is called the file drawer problem as many experiments with negative results end buried in the file drawer despite the fact that all this negative evidence is very useful). Experimental bias is one of the main sources of incorrect conclusions and it has been extensively studied in Random Clinical Trials (Higgins and Green, 2011), case-control studies (Sackett et al, 1979) and experiments with animals (Sullivan et al, 2016). Hooijmans et al (2014) and Zeng et al (2015) provide useful guidelines to try to avoid, or at least identify, bias in biomedical research with laboratory animals.

The main tools to fight bias are blocking, randomization, blinding and good reporting:

1. **Blocking:** We saw in Sec. 1 how to block a couple of variables (sex and time of the day at which the experiment is performed). The systematic differences induced by discrete variables is by measuring the same number of treatment and control samples at each of level of the discrete variable, in this way the design is balanced and the omitted-variable bias is zero. Some other variables that we may want to block are: the researcher doing the experiment or measuring its results, chemical batches, animal's diets, fish tank water supplies, laboratory equipment such as an incubator, ... Blocking is also a very simple way of addressing the problem of biological subpopulations. We may think of each one of the blocks as a mini-experiment, in which all treatments (control and treatment, for instance) are applied. The data analysis tools will be able to effectively determine the contribution of the block to the variability of the observations. In this way, this variability is explained and removed from the variability of the unexplained part (residuals). If the variable to block is continuous, instead of discrete (e.g., the time of the day of the experiment recorded as any real number between 0.0000h and 23.9999h; or the room temperature at the time of the experiment), then it is called a covariate and it should participate in the regression explaining the observed measurements. Blocking variables or measuring covariates has the additional advantage that it reduces the variance of the residuals (the part of the measurements that we cannot explain with our model), normally at a cheap cost in terms of degrees of freedom (we will see this effect in Sec. 5). In this way, proving that treatments have a statistically significant effect is easier since the associated statistic outstands more from the lower level of noise. Generalized Linear Models (Dobson and Barnett, 2008) are very rich models that can handle simultaneously discrete and continuous variables in the same model. Depending on the experiment performed, the researcher may have to resort to this more advanced method.

In animal experiments the cage is an important variable to block depending on the specific experiment and the nature of the species. Some species are social and single housing is stressful and considered detrimental to the animal welfare. However, males may fight depending on the strain and husbandry conditions. Cages should be balanced with respect to the treatment. In this regard putting all the treatments or controls in the same cage is not a good idea, because we would be confounding the cage effect with the treatment. Sometimes it is our only choice because we are studying an infectious disease that will be transmitted to the rest of the animals in the same cage (especially if the experiment is too long) or we are performing a pharmacokinetic analysis and a coprophageous behaviour would distort the measurements. If we give the same treatment to all animals in a cage, then it is better to consider the cage as a single experimental unit, with smaller variance due to the several animals in the cage). On the other side, we should be aware of possible coprophageous behaviour of some animals so that untreated animals may consume metabolites of the treated ones. For this reason, sometimes very valuable animals (like those wearing a telemetry apparatus or with a very specific genetic variant) are housed with a companion animal that is not part of the study. The position of the rack (animals in racks close to the door

are more disturbed) and of the cage within the rack (top or bottom) may have an influence in the response of the animals (Gore and Stanley, 2005).

- Example 9: We are measuring the immune response of mice to 4 different conditions. Researchers suspect that the litter the animals come from may explain some of the differences observed in the animals (some of the litters may have systematically higher or lower responses). For performing this experiment, 5 litters and 4 animals per litter were used. Each one of the animals from the litter was randomly assigned to one of the conditions. This design allows the identification of the contribution of the litter (if this exists). The analysis of this design is explained in detail in Sec. 5.1.3.

When dealing with animals we may also want to block their sex, age, the care taker, the person applying the treatments, the person evaluating the results, the order in which evaluations or treatments are performed (for instance, the first animal evaluated or treated may behave differently as it does not know what comes next; but the rest of animals has seen that there is some activity and may be more active than simply induced by the treatment).

In experiments with chemical reactants, an important source of confounding can be the batch from which we prepare our chemicals or the support in which perform our reactions. If we are using different suppliers, stocks, or bottles during our experiment, the small differences in the concentration of the different batches may cause a difference in the observations that can be confounded with the treatments. The same would happen if the experiment takes a long time and the reactant may differ from the beginning of the experiment to the end (for instance, it may have been partially oxidized or its humidity, pH, ... may have changed over time). Blocking the batch and performing experiments balanced in the batches may be important in certain settings. For instance, experiments performed with microarrays are particularly sensitive to these effects (Johnson et al, 2007).

Another important source of confounding may come from the instrumentation, if the experiment involves several measuring devices, or the experiment lasts for a sufficiently long time. All instruments must be calibrated in a way that we can know the exact relationship between our observed measurements,  $y$ , and the real values of the variable being studied,  $x$  (for instance, we may relate fluorescence,  $y$ , to fluorophore concentration,  $x$ ; or measured concentration,  $y$ , to real concentration,  $x$ ). In any case, the relationship between  $y$  and  $x$  is given by a calibration function,  $f(x)$ , such that

$$y = f(x) \tag{1.1}$$

Ideally,  $f(x)$  would be a linear function or identity ( $y = x$ ). But, in practice, many devices have non-linear responses (see Fig. 1.2). We may assume our device is working in its linear response area if the difference between the actual response and the ideal response is smaller than a given threshold (in Fig. 1.2, the relative error  $|x - y|/x$ , is smaller than 10% for  $x < 0.666$ , it is smaller than 5% for  $x < 0.435$ ). How strict this threshold should be depends on the experimenter. Let us assume that we have two experimental groups, control and treatment. Let

us assume that the mean of the control group is  $\bar{x}_C = 0.1$  and the mean of the treatment group is  $\bar{x}_T = 0.9$ . The control group is clearly working on the linear zone of the instrument, but the treatment group is working on its non-linear zone. The difference of the measurements of the two groups is

$$\bar{y}_T - \bar{y}_C = 0.750 - 0.099 = 0.671$$

smaller than the true difference

$$\bar{x}_T - \bar{x}_C = 0.900 - 0.100 = 0.800$$

Additionally, you may notice that the slope in the non-linear zone is smaller than the slope in the linear zone. If the true variance in both groups is the same, this difference in slope causes a decrease in the variance of the measurements in the treatment group with respect to the variance of the control group. As a corollary, we should always, if possible, try to work on the linear response area of our instrumentation. If we are using more than one measurement device, each specific device has its own calibration curve making the problem of calibration even worse. Blocking, or at least randomizing (see Point 2 in this list), the measurement device may be important depending on the experiment.

Finally, the calibration curve may drift over time, meaning that, unless regularly recalibrated, measurements from the beginning of the experiment may differ from measurements at its end, simply by a change in the measuring instrument. These differences may confound with the differences caused by our treatments, and blocking the time at which the instrument has been used may be necessary. Laboratory instruments for precise measurements like pipets, burets, and analytical balances fall under the same category of measurement instruments, and in many laboratories they are randomized. The same can be said of laboratory technicians, animal carers, etc. Differences in their performance can be confounded with treatments and blocking or randomization is highly advisable.

Computer programs should also be calibrated. For instance, we may measure the area occupied in a fluorescence microscopy image by fluorophore marking a specific protein. This is normally done through some software implementing algorithms that have parameters. These parameters must be adapted to each illumination conditions, constitutive fluorescence due to unspecific binding of the fluorophore to other molecules in the cell, and, consequently, adjusted to the experiment at hand so that the software measurements agree with some known results (e.g., absence of the protein of interest or known concentration of this protein). In a way, this would be the equivalent operation to the calibration used in balances or pipets. However, this is seldom done. Very likely, the parameters chosen will over- or under-estimate the true area covered by the protein, making comparisons across experiments very difficult. Still, comparisons within the same experiment are still valid if all images within an experiment are analyzed in the same way, with the same parameters. Then, the relative differences or ratios (depending on whether our process is additive or multiplicative) can still be attributed to the factors defining the different experimental conditions under study.

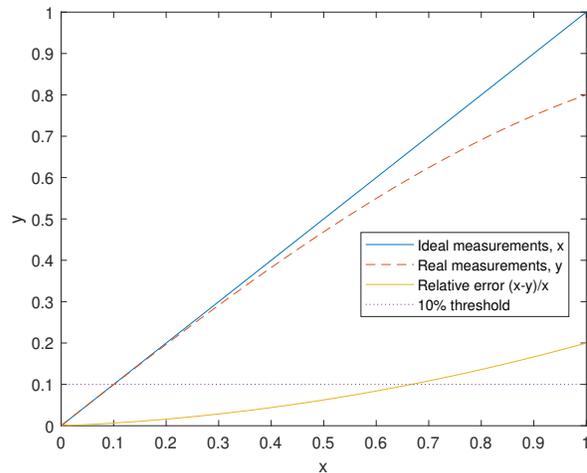


Figure 1.2: Calibration curve example.

On the other side, being usually biased, it is true that algorithms tend to produce much more reproducible results, that is, less variable, than humans. This is a great advantage when designing or analyzing an experiment. As a summary, we should be skeptical of sentences like “*the results are precise because they have been measured with a program*”.

2. **Randomization** is the process of randomly allocating the experimental units to the treatment(s) or control. For instance, let us imagine that we perform an experiment along a single day. We did not account for a variation during the day, but we realize that there is a systematic increase of the results as the day progresses. If we put all the control animals in the morning and all the treated animals in the afternoon, then there will be significant differences between the two groups because of the time of the day the experiment was carried out, and not because of the treatment. This kind of systematic errors can be avoided if the treatments (control or treatment) are randomly distributed along the day. The same would happen if there are differences due to the environment in the animal facility, the researcher handling the animals, or any other reason that we may have not thought in advance. Randomization requires that animals are uniquely identified (for instance, with a number) and the randomization should be performed by a computer (a simple random number generator will serve this purpose) since humans tend to create regular patterns when trying to randomize (Schulz et al, 2012). Once an experimental unit has been assigned to a specific treatment, we should stick to it (we cannot apply the treatment to a different animal because “it was easier to take”). Randomization addresses the selection bias, and the omitted-variable bias for those variables unknown to the researchers that may cause a difference in the outcome (e.g., subclinical infections, shipment

at different temperatures, different food suppliers, etc.) by randomly distributing confounders in the treatment and control groups.

We should randomize as much as we can. For instance, 1) the labelling of the groups (we should assign labels like A, B, C, ... to the different groups so that the treatment cannot be inferred from the label, and we should not use the same labels across experiments, e.g., always using the labels A or C for the control group); 2) the assignment of animals to treatments (this is what most researchers understand by randomization); 3) the housing of animals into cages (consider the bias caused when assigning newly arrived animals to cages, if we pick the less active animals first because they are easier to grab, and we put them together in the same cage); 4) cage location (it has been reported, see Sec. 5.2.1, that the position of the cage within the rack may induce systematic biases); 5) order of feeding, testing, etc. (performing these tasks in a systematic order may induce a systematic pattern in the measurements).

The main difference between blocking and randomization is that by blocking we will be able to measure the variability due to the block, by randomization we will not be able. In both cases, we will be “protected” against biases induced by the blocking/randomized variable, but by blocking we will determine its effect, while in randomization we will not. For instance, consider an experiment in which animals are presented a sequence of visual stimuli. We may randomize the order in which these are presented, or we may block the order in which they are presented and study the learning effects.

If we have variables to block, these must be measured before randomizing. This is called stratified randomization. For instance, if the animal sex is important, the randomization is performed within the male and the female group. In this way, we guarantee that for each of the groups, the treatment is balanced. This is particularly important in experiments in which some of the groups are rare. For instance, we do not have many old animals for our experiment. If we do the randomization before measuring the age group, we might end with most of the old animals in the control or the treatment group. The baseline response before applying the treatment can also be used for stratification before randomization.

If all units are known from the beginning, the randomization is simply performed by a random permutation performed by a computer. For instance, if we are going to study 10 animals, 5 in each group (control and treatment), we may simply sort them as

Treatment	C	C	C	C	C	T	T	T	T	T
Animal ID	6	3	7	8	5	1	2	4	9	10

However, if units arrive sequentially and we want to study  $N$  animals in total ( $N/2$  in each group), we may randomly assign them to one of the groups with a probability that depends on the number of animals already in each group. Assume that there are already  $n_C$  animals in the control group and  $n_T$  in the treat-

ment group. Then, we assign it to the treatment group with probability

$$p_T = \frac{N/2 - n_T}{N - n_T - n_C}$$

To do the random assignment we generate a random number with a uniform distribution between 0 and 1, then we assign it to the treatment group if this randomly generated number is smaller or equal  $p_T$ .

Sometimes, units arrive sequentially and we want to block some variables (e.g., sex and age). Depending on the number of animals of each kind in each one of the subgroups we assign them to one treatment or the other with different probabilities. For instance, assume that at the present moment we have already received 30 animals, and that the current number of animals assigned to each treatment per group is

Type	Control	Treatment
Male	7	10
Female	8	5
Age ( $\leq 9$ months)	5	6
Age ( $\leq 18$ months)	6	4
Age ( $\leq 27$ months)	2	4
Age ( $\geq 30$ months)	2	1

If a new male with 20 months arrives, the probability of assigning it to the treatment will be 0.5 modified by some factors that depends on the number of treatments and controls in its groups. It is a male, and there are already 10 male treatments and 7 male controls, so we will multiply 0.5 by 7/10 so that it is more likely to assign it to the control group. Similarly, in the group of animals between 19 and 27 months, there are 2 controls and 4 treatments, so that we will multiply by 2/4 so that it is more likely to assign it to the control group. Summarizing, the probability of assigning it to the treatment will be

$$p_T = 0.5 \frac{7}{10} \frac{2}{4} = 0.1750$$

Following this procedure, groups will tend to be compensated in number while the assignment is still random.

Stratified randomization is used if we foresee that a particular variable may introduce some variability. For instance, let us assume that we are comparing two treatments (control and treatment) and that we foresee that larger animals may have a larger response. If we simply randomly allocate animals to the two treatments, we might have, just by chance, larger animals in one of the groups. Instead we can create strata in the weight, for instance, we sort all the animals by weight and for every two animals, we randomly assign the two animals to the control or treatment group. This strategy may increase the variability of each one of the two treatment groups (will increase for sure if the weight really makes

a difference in the result) because the two groups now have a wide range of weights. We may follow this randomization strategy, but at the same time include the weight in the data analysis either as a block (see Sec. 5.1.3), or, for continuous variables like weight, as a covariate (see Sec. 5.1.4).

Randomization not only applies to the assignment of experimental units to treatment groups, it can also be applied to all levels of the experiment. For instance, let us imagine that we are interested in the ability of animals to recognize a specific odour, and how it is affected by a particular treatment (we will have a control group of animals and a group of treated animals). We embed our odour of interest in a sequence of 10 different odours. If we perform the test always with the same sequence of odours, we would be analyzing the difference between “the two treatments (control and treatment) under this particular sequence of odours”. It is much more interesting randomizing the sequence of odours in each test, so that we can analyze the difference between “the two treatments”.

3. **Blinding** hides the treatment information to the patient (single blinding), the patient and the experimenter (double blinding), or the patient, the experimenter and the data analyst (triple blinding). With laboratory animals, single blinding is normally unnecessary. However, if possible, blinding the experimenter from the treatment he or she is applying or evaluating drastically improves the fairness of the experiment (for instance, cages or animals should not be labelled with the applied treatment, in this way they will not be handled differently). [Bebarta et al \(2003\)](#) evaluated the outcome of 290 research studies with animals. Those studies lacking randomization, blinding or both were significantly more likely to report positive outcomes. Blinding directly addresses performance and observer bias.
4. **Good reporting.** Unfortunately, except for a few cases like survival analysis, there is no ideal technical solution for exclusion or attrition bias. At least, good reporting the experiment and its data filtering and processing may help the reader to evaluate the quality of the reported results. In this regard, [Kilkenny et al \(2010\)](#); [Hooijmans et al \(2014\)](#); [Zeng et al \(2015\)](#) provide a guideline to experiment reporting that should minimize this kind of bias.
5. **Data imputation.** A statistical approach that can be adopted, although acknowledging all its limitations, is data imputation. For instance, in attrition bias we may have the measurements from the beginning of the study, but at some point the animal ceases to give measurements because it had to be sacrificed. We may try to extend the first measurements into the future, predicting what the remaining measurements could have been. This follows the philosophy of analyzing the data according to the Intention-to-Treat (ITT), that is, all animals are analyzed irrespectively of their dropouts or, in the case of patients, noncompliance with the admission criteria). The way to extend the known measures into the future can range from: 1) copying the last observation to the missing observations, 2) imputing the missing data through some form of regression, 3) imputing the missing data from a worse case scenario (for instance, if the variable is dichotomous, like death or survival, we may assume survival for all the dropouts in

the control group and death for all the dropouts in the treatment group; if the variable is continuous, we may assume a normal measure for all the control dropouts and the worse endpoint for all the treatment dropouts). The advantage of the worse case scenario imputation is that it underestimates the treatment effect. If we can still show that the treatment is effective, then we are more confident in its efficacy.

Finally, we would like to give a word of caution against overconfidence. Some research groups think that they do not need to use these tools against bias (blocking, randomization, blinding, etc.) because they are *professional* researchers, they think that those recommendations are only valid for researchers who are prone to make mistakes. However, as we have extensively discussed along this section, these are the best tools we can use to avoid bias and the lack of their use may result in seriously incorrect conclusions.

## 1.4 Reducing variance: variable and population selection, experimental conditions, averaging, and blocking

Variability is inherent to biological populations and experimental measurements. Individuals differ among them in almost all imaginable variables: morphological, biochemical, metabolic, genomic, proteomic, physiological, microbiome, behavioral, etc. For any variable we may even find circadian variations, variations due to specific environmental factors, variations due to response to stimuli, etc. The mantra of many biologists is that “Biology is highly variable”. However, this cannot be an excuse for not doing any experiment, or accepting the results of “any” experiment. Actually, these are the variations most research studies aim to understand. Different individuals in our study, called *biological replicates*, will help us to model this biological variability, and new treatments should be tested with a sufficiently wide biological scope so that our study results can be generalized to the whole population. Some sources of the variations in the observations can be due to:

- True biological variation:
  - Genotype: outbred animals are genetically undefined, genetically defined animals include isogenic, mutant or genetically modified animals; even within this group there might be some genetic drift at the company selling the animals
  - Phenotype differences before or during the experiment: some animals start to differ in our facility due to different handling, location, bedding, diet composition and availability, feeding, cleaning frequency, type of caging, number of animals in the same cage, litter and litter size, social hierarchy within the cage, aggression, infections propagating within the animal house, subclinical situations, accidents, particular situations, etc.

- Physiological reasons: different breed, sex, age, time of day, lunar cycle, stress (homeostasis alteration to cope with an external stimulus), distress (homeostasis alteration without the capability to cope with the external stimulus), environmental factors (temperature, humidity, atmospheric pressure, light levels and cycles, noise, smells, room characteristics), suboptimal health conditions, or pain.
- Experimental variations:
  - Caused by the researchers: intra-researcher (we do not work exactly the same every day), inter-researcher (there are differences among different researchers, especially in skill and experience, gender, use of cosmetics, investigator personality), inaccuracy of the procedures, lack of validated or authenticated tools, the presence of the investigator in the animal facility may induce changes in the animals.
  - Measurement noise: due to the imprecision of measurement device, human errors, electronic noise, uncertainty in the readout of analog scales, incorrect calibration, contamination of the samples, poor documentation or data capture, etc. These measurement errors can be reduced by measuring the same experimental unit multiple times, these are called *technical replicates* or *pseudoreplications*.

Since the overall error is the sum of biological error and experimental errors, reducing technical error will reduce the overall error. Other measures to reduce variability are to have a tight control on the genetic standardisation (inbred strains, hybrid breeding (F1), coisogenic and congenic mutants, transgenic and knockout mutants, etc.), microbiological standardization (reduction of latent infections, maintenance of barrier systems, periodic health status assessments, use of sentinel animals, swabs, sampling procedures), phenotypic uniformity, acclimatisation of the animals, housing, husbandry control (husbandry-related cycles, seasonal cycle, reproductive cycle, weekend-working days cycle, cage change/room sanitation cycle, diurnal cycle, in-house transport, caging, number of animals per cage, cage material, bedding, enrichment, ventilation, temperature, humidity, air quality, odours, lighting, noise, alarm systems), nutrition control (feeding scheme, form of the diet, pellets, energy content and components of the diet, batch, sterilization (heat, irradiation, ...), storerooms and conservation of perishable feed, watering system, functioning of automatic watering systems, water quality, microbiological and salts content, etc.), training of the researchers, and standardization of the experimental procedures (acclimatisation, restraint, substance administration, biological fluids collection, anaesthesia and analgesia, surgery, euthanasia, tissues and organs sampling, analytical procedures, etc.).

Ultimately, the variability of our observations (once all variability sources have been considered) will determine the number of samples that we need to involve in our experiments. As a general rule, higher variability will require a larger sample size for detecting the same treatment effect. Alternatively, if we fix the sample size, higher variability will hinder our ability (statistical power) to detect a given treatment effect. These ideas are further discussed in Sec. 1.5.

### 1.4.1 Variable selection

As we will see in Chap. 4, the calculation of the sample size depends on the information brought in by each one of the experimental units, and the noise of our measurements. In this way, different types of observations are more informative and, generally speaking, the information order of variables would be: categorical, ordinal, discrete, and continuous. For instance, if we are studying the presence of macrophages in a given microscopy field the following measurements would bring an increasing amount of information: 1) absence or presence of macrophages (categorical); 2) qualitative number of macrophages (ordinal: none; one or two; three, four or five; more than five); 3) quantitative number of macrophages: 0, 1, 2, 3, ... (discrete); 4) area occupied by the macrophages in the field (continuous). If possible, we should work with as informative variables as possible.

Some discrete variables may be treated as (almost) “continuous” for the purposes of statistical analysis. For instance, we may measure the severity of arthritis of a single paw in a scale from 0 to 4. Each animal receives a score that is the sum of the scores of the four paws.

We must be careful with the variables we chose for the analysis, they must be as descriptive and related to our interest as possible. For instance, in behavioural studies we want to analyze how a particular treatment affects the exploration time of the animals. We compare the time spent exploring novel objects to the time spent exploring familiar objects. The discrimination index is defined as

$$DI = 100 \frac{t_{novel} - t_{familiar}}{t_{novel} + t_{familiar}}.$$

The problem with this variable is that an animal that spends 2 minutes exploring new objects vs 1 minute exploring familiar objects gets a discrimination index of 33.3%, the same as an animal that is lethargic for most of the experiment and explores the new objects for 2 seconds, and the familiar object for 1 second.

We should also work with variables related to our experimental objective with as little variance as possible. For instance, if we are interested in the appetite effect of some treatment given with food, we should prefer directly measuring the weight increase of the animals, instead of the weight of the food consumed (because animals may throw food through the cage and we would skip the variability induced by variable excrements). By avoiding the variability of unrelated events, we would reduce the biological variability of the variable of interest.

### 1.4.2 Population selection

Currently experiments can be performed on mixed stocks, outbred stocks, and inbred strains (Chia et al, 2005). Mixed stocks of animals would be the equivalent of the genetic variability encountered in large human populations (like a whole country). Outbred stocks would be the equivalent of the variability of small communities with little interaction with other communities (like Laponia). Finally, all animals of an inbred or hybrid F1 strains would be genetically identical, as human identical twins. In this

Table 1.6: Sleeping time of different animal strains after a dose of hexobarbital.

Strain	Type	Mean (min)	Std.Dev. (min)	N	Power
A/N	Inbred	48	4	23	86
BALB/c	Inbred	41	2	7	> 99
C57BL/HeN	Inbred	33	3	13	98
C3H/He	Inbred	22	3	13	98
SWR/HeN	Inbred	18	4	23	86
CFW	Outbred	48	12	191	17
Swiss	Outbred	43	15	297	13

way, the variability of our observations is mainly due to epigenetic, treatment or environmental differences between the animals. In this way, the observations variability would be strongly reduced. Except for research related to quantitative trait loci, the experimental use of outbred stocks is discouraged (Chia et al, 2005). And, currently, no research experiments are performed on mixed stocks of animals due to the large number of animals required to prove any statistically significant difference in these populations.

Actually, it is currently preferred to show the effect of our treatment in several independent inbred strains than showing it in outbred stocks. For instance, Jay Jr (1955) analyzed the sleeping time of different stocks of animals after a dose of 125 mg. per kilo body weight of hexobarbital. Table 1.6 shows the mean and standard deviations observed for each kind of animals. We note that the inbred strains cover a wide range of sleeping time (from a mean of 18 to a mean of 48), while the outbreds are centered around 43-48 (although with a large standard deviation). With this variability, we may calculate the number of animals of each kind,  $N$ , needed to detect of change of 4 minutes in sleeping time with a confidence level of 95% and a power of 90%. Similarly, if we fix the sample size to  $N = 20$  animals, we may calculate the power to detect a change of 4 minutes in the mean. We note that the sum of all animals in the inbred strain is 79 ( $=23+7+13+13+23$ ) and the statistical power is between 86 and 99% if  $N = 20$ . However, the sample size for performing a similar experiment with an outbred stock is between 200-300 animals (between 2 and 3 times more). If we fix the sample size to  $N = 20$ , then the power drops from about 90% to about 15%. For this reason, the current recommendation (Chia et al, 2005) is to show the effectiveness of our treatments on a variety of inbred strains sufficiently covering the spectrum of the physiological variability of the whole population. If maintaining several inbred strains is too costly for our experiment, at least, we should make sure to report the applicability of our results, in which conditions and with which strains is our treatment effective. The interested reader is referred to Sec. 4.6.5 for further details on how the compromise of the independence between individuals can affect the sample size.

### 1.4.3 Experimental conditions

The specific setup of the experiment may also affect the variability of our observations. For instance, [Chvedoff et al \(1980\)](#) reported an increase in the variance of the weight of mice depending on whether they were housed 1, 2, 4, or 8 animals per cage. Another example is given by [Crabbe et al \(1999\)](#). They repeated the same experiment with eight mouse strains in three different locations: Portland, Edmonton and Albany. They controlled all the experimental conditions (same research team, same inbred strains, equally calibrated apparatus, equated husbandry, same testing protocols, same age, same starting time, same protocol order) so that the experiment was as homogeneously performed as possible. They found significant differences in body weight and behavioral tests in the three experimental sites, meaning that there had been some differences escaping their control despite their careful effort to equate everything. These studies call for an homogenization of all the experimental variables we can control (same number of animals per cage, same calibration procedures, same protocols, ...) knowing that, although reducing the variability, there might still be uncontrolled variables we are not aware of and that affect our results.

### 1.4.4 Population scope, outliers and lack of independence

One of the key assumptions of all statistical tools is that the observations are a random sample of the whole population being studied. Intuitively, it means that our observations are representative of the whole population being studied in all its statistics (mean, variance, distribution, ...). Put differently, this hypothesis assumes that any individual in the general population has the same chance of being observed, and that no individual or subpopulation has a larger chance of being overrepresented. As we have seen in the previous sections, we may perform experiments in too narrow populations, with very low genetic variability, or under very strict laboratory conditions (the health, hygiene, diet, exercise, and environmental conditions of laboratory animals differ significantly from those from a general population of animals or humans). This certainly help to reduce the variance, but at the cost of reducing the scope of the whole population being studied. Our random sample of a given inbred strain is a good representative of that kind of mice, not all mice in the world. That is why it is recommended using several inbred strains to validate our research hypothesis (e.g., a drug is useful in decreasing a given disease condition). The same occurs with the homogenization of the experimental conditions. They are aimed at reducing the variance of our observations. However, they compromise the applicability of our results to a wider population. Especially, if we are testing new treatments in laboratory animals with the aim of an ultimate commercialization in humans, pet or farm animals. The variability encountered in the general human or animal populations is much larger due to the larger genetic variability, environmental conditions and different lifestyles.

Some experiments analyze cells coming from one animal or a pool of animals, for instance, the proportion of cells of a given type. These experiments should be handled with care. Seemingly, the number of cells,  $N$ , is huge and the estimated proportions seem to be very accurate (small confidence interval) due to the large number of allegedly independent events. The same problem is faced by experiments with low num-

ber of animals being analyzed resulting in large numbers of events (gene copy number, number of RNA transcripts, ...) However, the cell type is not that independent (they are coming from one or a few animals), and we may encounter a generalization problem: is the proportion of this type of cell in the whole population of animals the same as the one I have measured in my single individual or group of animals? If we have measured very few animals, we cannot guarantee that this is the case. Ideally, in these experiments we should determine the proportion of cells per animal (or pool) for several animals (or pools), and treat these proportions as a continuous variable for which we construct a confidence interval. Constructing the confidence interval using the standard proportion tools is not the best approach, because events are not independent. However, this ideal approach is not always feasible for experimental or economical reasons.

Researchers are often worried about the presence of outliers in their observation and how they should treat them. Should they be eliminated from the sample, left in the sample, or be treated separately? We should ask ourselves about the nature of those outliers.

- Have they been caused by an obvious measurement error (mistyping of the numbers, malfunctioning devices, measurement blackouts, non-sensical numbers)? If so, we should remove them as they do not really exist in the general population we want to generalize our results to.
- Have they been caused by an obvious error in the application of the treatment (e.g., not applying the correct dose, applying the treatment in a different area than the intended one, not strictly adhering to the dosage plan, artificially lengthening the surgery and having a longer post-operation recovery as a result, ...)? If so, we must be aware that these application errors could be representative of the errors that can be committed in the future in the general population of humans or animals. As such we may want to keep the outlier observations as representative of the situations that can be encountered in the application of our treatment. Or alternatively, we may discard those outliers knowing that our results then only apply to the general population of individuals for which the therapeutic plan is perfectly applied. Keeping the observations does not mean that both datasets (those that correctly received the treatment and those that received it in an incorrect way) must be analyzed together. We may create two subpopulations and draw separate conclusions for each one of them.
- Have they been caused by the different response of the individuals? If so, it means that the general population can be subdivided into smaller populations, each having a different response. As in the case above, we may divide the responses in subgroups and draw different conclusions for each one of them (e.g., 80% of the population has a strong response to our treatment, while 20% of it has a null or small response to the treatment). Stratified sampling is a statistical technique especially aimed at characterizing the overall response in the general population when several subpopulations with different responses are known to coexist. The interested reader is referred to [Thompson \(2012\)](#)[Chap. 11].

By removing outliers that really belong to the studied population we are biasing all our estimates (mean, variance, ...). On the other side, leaving outliers that do not belong to

the population also biases our estimates. Unfortunately, there is no statistical technique that can clarify the nature of an outlier. Statistical tools may indicate the presence of samples whose observations do not follow the general population trend observed among the samples. But, they cannot assess whether these anomalous observations are caused by measurement errors, errors in the application of the treatment, or different biological responses of those individuals. It is the responsibility of researchers taking a decision regarding those samples and about how they should be analyzed. The answer “I will assume that that observation is an error” is not valid in general, since it precludes a careful reasoning about the nature of that particular outlier. We admit that, unless obviously nonsensical measurements, it is difficult to distinguish *a posteriori* between a measurement error and a misapplication of the treatment, but at least we should be able to recognize differently responding subpopulations. Obviously, we can only do this with a sufficiently large population, and with small experimental sample sizes we are bound to believe that most observations correspond to the naturally observed variability.

The presence of subpopulations, if not well treated, can easily lead to incorrect conclusions similar to those obtained under the presence of outliers. Fig. 1.3 represents a possible result of a study. We are interested in the level of a given compound in blood after giving a drug to the animals. Before giving the drug, we measure the baseline level of the compound. Then, we randomize the animals into a control and treatment group, and measure the level of the compound again after treatment. The plot at the top of Fig. 1.3 shows the mean level of the compound of interest in the two groups along with their standard error of the mean. From this plot alone, one would conclude that the administration of the drug results in a higher level of the compound of interest. However, a more detailed analysis (see Fig. 1.3, bottom) reveals that before administering the drug, we could already identify two different populations (one with a high and another one with a lower baseline level), and that, just by chance, randomization assigned more than half of the high responding samples to the treatment. The random assignment along with the existence of two subpopulations falsely created the impression of a higher response of the treatment group. In Sec. 1.4.8 we will see that a correct randomization should try to produce control and treatment groups of the same size (in this particular case, this equal size groups would have saved the experiment because in this simulated data there was no difference between the two groups, and the equal size groups would have allowed us to identify this situation). In general, the existence of subgroups in the data can create many different kinds of misleading results, and we should always try, to the best of our ability, to identify this situation.

Another source of artificially observing a low variance is by violating the assumption of independence of the samples. The assumption of independence is two-fold: 1) independence between groups and 2) independence within group. The first assumption, independence between groups, would be violated by the same individual participating in several groups (control and treatment, for instance). This is obviously avoided in laboratory research.

There are also obvious ways of violating the second assumption, independence within group, for instance, collecting multiple samples from the same individual. These are technical replicates and they can be averaged to produce a single measurement, as we will see in the next section, or we may use repeated-measures ANOVA, which ba-

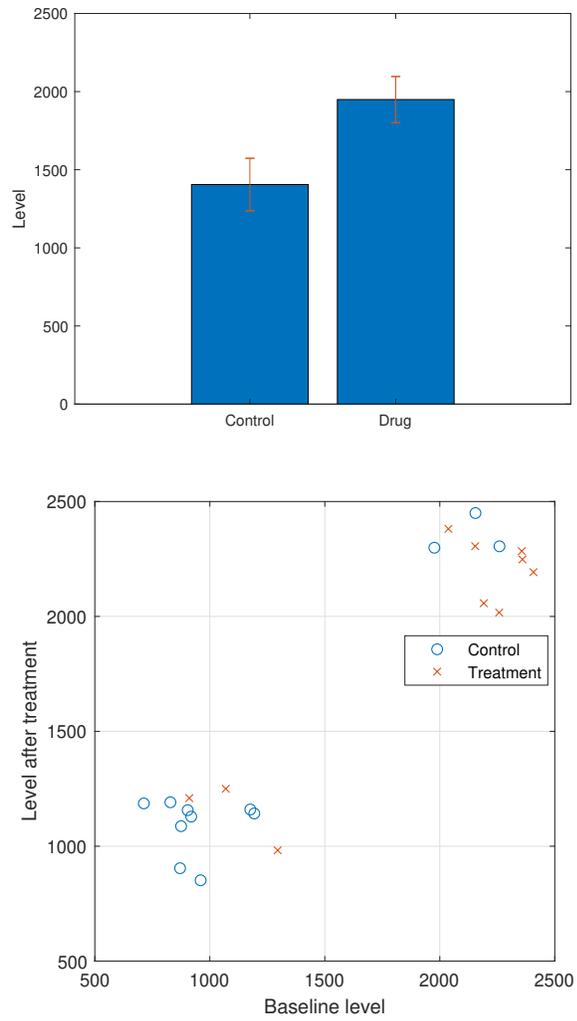


Figure 1.3: Effect of the presence of subpopulations (see text).

sically amounts to using the individual as a block (this technique is seen in Sec. 5.2.6). However, there also are more subtle ways of violating the second assumption. For instance, applying the treatment to different cultures of the same cell line. This cell line is coming from a single individual, and may even be immortalized causing biological artificial results that are not observed *in vivo*. Statistically, we see that the independence of these samples is compromised, and they only generalize to the population of cells of this cell line. The same occurs if we use different animals from the same inbred strain. They are all twins and share the same genetic background. In a way, we are measuring multiple times “almost the same” individual. Measuring animals that have been housed in the same cage, or are siblings from the same mother also compromise the independence of the observations. A human analogy would be measuring a variable in persons from the same family and living in the same house. In all cases, our statistics are biased. In terms of the mean, we are assuming that the rest of the population is responding in the same way as our observations. If they do not, then our sample mean will be a biased estimate of the population mean. In terms of variance, the variability observed within a single individual (cell line, inbred strain, ...), within a few families (all animals coming from a few mothers), or within a few cages (think of them as households) is usually smaller than the variability observed in a wider population. In this regard, as the observed variability is smaller than it should be, we tend to be overconfident on the statistical significance of the observed differences (the p-value is artificially low). If the p-value is close to the threshold of significance (typically, 0.05), we may declare as significant a result that is not truly so. It is simply an artifact caused by underestimating the population variance.

Disregarding all technical (statistical) considerations, in which many researchers get lost in, we should always apply our common sense and think to which population does this random sample generalize to: the population of all cells equal to ours (as in the case of a cell culture), the population of all mice of this strain (an inbred strain), the population of all mice under very strict environmental conditions (the laboratory conditions), ... and consequently be humble about the generalization of our results to larger populations and be prepared for failures when our treatments are tested in more general experimental conditions (Phase II and Phase III in drug developments, for instance).

### 1.4.5 Averaging and pooling

The simplest and most wide spread measurement model is the *additive noise model*:

$$y = x + n \quad (1.2)$$

where  $y$  is the observed value,  $x$  the ideal (inaccessible) value, and  $n$  a random noise variable.  $n$  is assumed to have zero mean (otherwise, it would be biasing our measurements) and to be independent of the ideal values. Under these circumstances, the variance of the observed measurements would be given by

$$\sigma_y^2 = \sigma_x^2 + \sigma_n^2 \quad (1.3)$$

that has a very natural interpretation: the variability we observe in our measurements is partly caused by the biological variability,  $\sigma_x^2$ , and by our measurement errors,  $\sigma_n^2$ .

Although it is not strictly necessary, in many experiments it is assumed that noise follows a Gaussian distribution with zero mean and variance  $\sigma_n^2$ .

If we measure the same subject  $M$  times (technical replicate), we may reduce the variance of our measurements. Each measurement would be of the form:

$$y_i = x + n_i \quad i = 1, 2, \dots, M$$

Our measurement for this individual would finally be

$$y = \frac{1}{M} \sum_{i=1}^M y_i = x + \frac{1}{M} \sum_{i=1}^M n_i \quad (1.4)$$

Consequently, the variance of our observations is

$$\sigma_y^2 = \sigma_x^2 + \frac{\sigma_n^2}{M} \quad (1.5)$$

That is, technical replicates help to reduce the variance of our observations by reducing the variance associated to the measurement errors.

- **Example 10:** We are performing an experiment in which we allocate animals into treated or control groups. Then, we sacrifice each animal, extract several tissue slices from each one of them, and score the slices. The analysis is performed with a hierarchical analysis of variance (see Sec. 5.2.7). From a previous experience we have seen that 50% of the variability comes from the scores of the different slices, 25% comes from the treatment or control, 10% from the inter-animal variability, and 15% is unexplained (residuals). Which would be the best strategy to increase the precision and statistical power of our experiment?

**Solution:** As reported above the larger proportion of variability comes from the tissue slices rather than the inter-animal variability, then it is much better to increase the number of slices extracted from each animal than increasing the number of animals. In a hierarchical analysis the calculation performed is not the same as the averaging of the technical replicates that we have described in this section. However, they have a similar spirit. As we have discussed in this section, averaging the most variable part is the action that has a higher impact on the reproducibility of the experiment.

Following the same reasoning if the largest source of variability is between cages, then we should use more cages in our experiment. However, if the largest source of variability is the variability within the cage, then we should have more animals per cage.

If we have  $N$  biological replicates and we average them into a single population mean ( $\bar{y}$ ), then the variance of the population mean is the variance of each of the observations divided by  $N$

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{N} = \frac{\sigma_x^2}{N} + \frac{\sigma_n^2}{NM} \quad (1.6)$$

If we associate costs to each of the kind of replicates ( $C_B$  for the biological replicates and  $C_T$  for the technical replicates), then we may calculate the optimum number of technical replicates per biological replicate by minimizing the variance of the average subject to a constraint in the cost

$$\begin{aligned} \min_M \quad & \frac{\sigma_x^2}{N} + \frac{\sigma_n^2}{NM} \\ \text{subject to} \quad & C_B N + C_T N M = C \end{aligned}$$

whose optimum is

$$M = \sqrt{\frac{\sigma_n^2 C_B}{\sigma_x^2 C_T}}$$

For instance, in microarray experiments we may assume that technical replicates are 4 times less variable than biological replicates ( $\frac{\sigma_n^2}{\sigma_x^2} = 1/4$ ; obviously, this estimate depends on the genes we are studying). The price of a technical replicate can be around  $C_T = 500\$$ , and the price of a biological replicate (mouse) can be  $C_B = 15 + 2d\$$  (being  $d$  the length in days of the experiment, and  $2\$/\text{day}$  the average price for animal housing). The formula above would recommend 1 technical replicate for short experiments, and 2 technical replicates for long experiments.

Microarray experiments have a complicated setup and they have represented a great technological breakthrough. In an extremely simplified description, different treatments are given to animals (in Fig. 1.4 shown as A and B). The different animals would be biological replicates (sometimes several animals are combined into a single pool as described below). The objective of the experiment is to identify differences between treatments at the level of mRNA. A sample from the tissue of interest is extracted and the mRNA isolated. The mRNA of the samples are reverse transcribed into cDNA and combined with a dye. This process is repeated twice with different dyes (red and green). This repetition is a technical replicate at the level of transcription and dye. Every microarray sees two of these combinations with different dyes (one red and one green). The microarray has several wells or spots. In each of the wells, there is a DNA sequence probe that hybridizes with the sample cDNA. The ratio of fluorescence between both dyes is measured from each well resulting in the known colored spot images (see Fig. 1.4). In the same microarray we can put the same DNA probe in two wells, this would be a technical replicate at the level of spot.

Spot technical replicates have a correlation coefficient of 95%, indicating the high reproducibility (low noise) of the probe hybridization and fluorescence measures. Let us refer to this noise as  $\sigma_{n2}^2$ . Technical replicates at the level of mRNA reverse transcription and dyeing has a correlation between 60-80% (Churchill, 2002), indicating a higher level of noise at this point, that we will refer to as  $\sigma_{n1}^2$ . Although these numbers have surely changed since 2002, the setup is still valid and illustrates a more general problem. If we have  $N$  animals,  $M_1$  transcription and dyeing replicates, and from each dyed sample we take  $M_2$  spot replicates, then the variance of the mean of the duplicated spots will be

$$\sigma_y^2 = \frac{\sigma_x^2}{N} + \frac{\sigma_{n1}^2}{NM_1} + \frac{\sigma_{n2}^2}{NM_1 M_2}$$

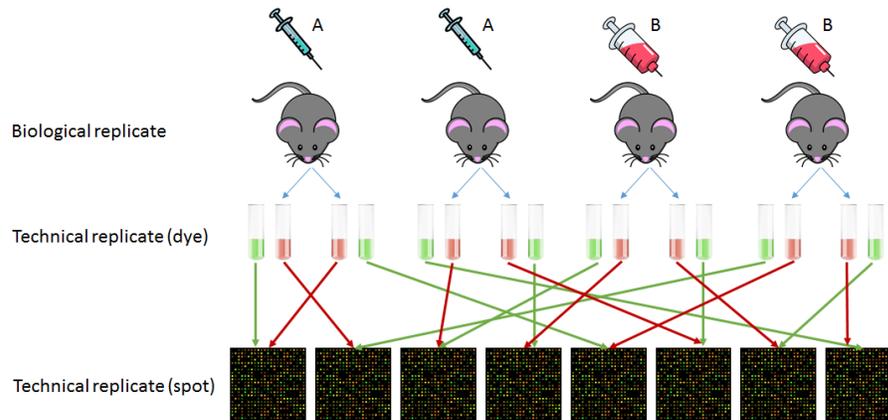


Figure 1.4: Setup of a microarray experiment. See text for a description of the experiment.

Depending on the technology, the noise of the different stages may be higher or lower. It is specially important the size of the noise with respect to the biological variability,  $\sigma_x^2$ . Sometimes, it is very tempting to avoid biological replicates because the experiment seems to be more reproducible. The apparent increase in statistical significance and power is illusory, and the statistically significant results may simply reflect the random fluctuation due to the specific animals used in our experiment.

In Eq. 1.3,  $\sigma_x^2$  could be understood as the addition of two sources of variability: within-animal and between-animal

$$\sigma_x^2 = \sigma_{x,within}^2 + \sigma_{x,between}^2 \quad (1.7)$$

The within-animal variability may be due to circadian rhythms, random fluctuations along time, different physiological conditions between measurements, etc. While the between-animal variability is expected to be caused by genetic and environmental differences between the different individuals. If we want to determine the contribution of each one of these two components, we need to measure the same animal multiple times. The differences between measurements will be due to within-animal variability as well as to the noise variance. We will not be able to disentangle these two components as they go inherently together in all our measurements. For this reason, all the argument above about reducing the variability by making repeated measures also apply to within-animal variability. Some statistical analysis techniques, like all those based on repeated measures (see Secs. 5.2.6 and 5.2.7), specifically exploit the fact that some measurements are coming from the same animal to produce better estimates of the effects of the treatments.

A useful tool to reduce the biological variability is to pool tissue or cells from several animals, and then applying the treatments to the pool or performing measures from that pool. By pooling, we are “creating” an artificial animal, let us call it  $\tilde{x}$ , whose variability is, in principle, smaller than the raw biological variability. Ideally,

the variance of the pool should be reduced by the number of pooled animals,  $K$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{K}$$

In practice, the response of the animals is correlated because they are housed in the same laboratory, treated by the same persons, fed with the same food, ... and, depending on the case, they may even be tied with familiar bonds. If the correlation between the measurements of the different animals is  $\rho$ , then the variance is not reduced by  $K$ , but by a smaller factor

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{K/(1+(K-1)\rho)}$$

Note that for  $\rho = 1$  (all animals are perfectly correlated, this would be the case of clones), the variance of the pool is the same as the variance of the original animals, while we would have the false impression of having reduced the variance of the experiment by pooling from different animals.

It has also been proposed to model the effect of pooling as

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{K^a}$$

with  $a$  being a number between 0 and 1. If  $a = 1$ , then the pooling has been maximally effective; while if  $a = 0$ , then the pooling has not helped in reducing the biological variability. In either way, through the correlation  $\rho$  or the exponent  $a$ , we see that pooling aims at reducing the biological variability of our measurements, but it may not always be maximally effective.

#### Important remarks

4. Gathering all animals into a single pool from which we will make several technical replicates is not a good statistical design, because we have no possibility to estimate the biological variability. It is much better to divide the different animals in several pools that do not share animals (otherwise, we compromise their independence).

The assumption of the additive noise model is that there are many sources of error that are added to our measurements. By the central limit theorem, the addition of many random independent variables result in a Gaussian distribution (that is why additive noise usually assumes a normal distribution). However, not all measurement errors or sources of variability are additive. Some systems have a multiplicative or exponential behavior (e.g., the decay of drug concentration in blood is normally exponential, cell divides by two after a given period of time, PCR experiments double the amount of DNA with every replication cycle). In all these experiments, changes are relative (an increase or decrease of 1%, 5%, or 50% with respect to the current level). Noise is additive in the logarithmic space

$$y = x \cdot n \rightarrow \log(y) = \log(x) + \log(n) \quad (1.8)$$

This is the *multiplicative noise model*. If  $\log(y)$  follows a Gaussian distribution, then  $y$  is said to follow a log-normal distribution. Limpert et al (2001) revised the applications of the log-normal in sciences and they include the abundance of bacteria, the latent period of a disease, survival time in cancer, sensitivity to chemicals ( $EC_{50}$ ), gene expression (Beal, 2017), and noise in imaging modalities based on the counting of photons (e.g., fluorescence microscopy, PET and SPECT, Rodrigues et al (2008); Waters (2009)). For this reason, in many fields, like microarray analysis (Quackenbush, 2002), technical replicates are averaged using a geometric mean. The geometric mean is equivalent to averaging in the logarithmic space as shown below

$$y = \sqrt[M]{y_1 y_2 \dots y_M} \Leftrightarrow \log(y) = \frac{1}{M} \sum_{i=1}^M \log(y_i) = \log(x) + \frac{1}{M} \sum_{i=1}^M \log(n_i) \quad (1.9)$$

In this way, we see that choosing between a standard average or a geometrical average depends on the nature of the noise and the way data is generated, rather than our own preference.

### 1.4.6 Blocking

Our measurements may be affected by variables that we are not interested in. For instance, we are interested in the effect of a drug, and we perform an experiment with two groups (control and treatment) with the aim of comparing the mean level of some variable  $y$  in the two groups. In each of the groups we assume that the  $i$ -th observation ( $i$  is supposed to refer to each independent biological replicate of the experiment, that is, the measurement of the experimental unit) respond to the model

$$y_i = \mu + \alpha_{x(i)} + n_i \quad (1.10)$$

where  $\mu$  is an overall mean,  $\alpha_{x(i)}$  is the effect caused by being in the control or treatment group and takes the values  $\alpha_{control}$  or  $\alpha_{treatment}$  depending on the group the  $i$ -th animal has been assigned to. The mean of the observations in the control group is

$$\bar{y}_{control} = \mu + \alpha_{control}$$

and the one of the treatment group is

$$\bar{y}_{treatment} = \mu + \alpha_{treatment}$$

To make the decomposition in Eq. (1.10) unique, we impose the constrain

$$\alpha_{control} + \alpha_{treatment} = 0$$

This constrain causes that  $\mu$  can be estimated from the overall mean of all our observations. Under this model, the variance of  $y$  can be decomposed as

$$\sigma_y^2 = \sigma_x^2 + \sigma_n^2$$

where  $\sigma_x^2$  is the variance induced by the fact of taking or not the drug, and  $\sigma_n^2$  is the variance induced by all other experimental variables (measurement errors, chemical

batches, biological variability, sex, genotype, month of the year, day or time of the day in which the experiment was performed, etc.) For example, it has been seen that the results on Mondays (after two days of quietness) are different from the rest of the days. In general, we refer to  $\sigma_n^2$  as the noise variance or the unexplained variance. Note that in the expression above we have assumed that there is no relationship between  $x$  and  $n$  (if there is, we should have included a term with the correlation between the two). This independence is not fulfilled if we presume an additive noise model, but the true model is multiplicative, for instance.

Proving that our drug is making a difference ultimately amounts to comparing  $\sigma_x^2$  to  $\sigma_n^2$  and checking whether the observed signal,  $\sigma_x^2$ , is significantly different from the noise,  $\sigma_n^2$ . We may think of this as a Signal-to-Noise Ratio, and more evolved versions of this comparison is at the core of the Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), and Generalized Linear Models (GLMs).

We may reduce the unexplained variance,  $\sigma_n^2$ , by blocking the effect of some variables that we cannot control, but that we can measure, for instance, sex. We cannot control the sex of the animals, but we can annotate it and remove the variability induced by it from the unexplained variance. For doing so, we now explain our measurements as

$$y_i = \mu + \alpha_{x(i)} + \alpha_{sex(i)} + n'_i \quad (1.11)$$

where  $\alpha_{sex(i)}$  takes the values  $\alpha_{male}$  or  $\alpha_{female}$  depending on whether the animal is male or female. Then, the variance of our measurements can be decomposed as

$$\sigma_y^2 = \sigma_x^2 + (\sigma_{sex}^2 + \sigma_{n'}^2)$$

The variance of the measurements,  $\sigma_y^2$ , does not change due to the blocking. The variance explained by the treatment or control does not change either. However, we have decomposed the previously unexplained variance,  $\sigma_n^2$ , by something that is explained by sex and something else that we cannot explain yet:

$$\sigma_n^2 = \sigma_{sex}^2 + \sigma_{n'}^2$$

The statistical comparison to assess whether our treatment is successful is performed between the variance explained by the treatment-control variable,  $\sigma_x^2$ , and the unexplained variance,  $\sigma_{n'}^2$ . After blocking, the unexplained variance is smaller than the unexplained variance before blocking

$$\sigma_{n'}^2 < \sigma_n^2$$

Consequently, if our treatment is making a difference, it will be easier to show that the difference is statistically significant.

We may think of blocking as a “research insurance”. Sex may or may not make a difference in our measurements, but if it does, by blocking, we will be able to measure this difference and subtract its effect from the unexplained variance. If it does not make a difference, it does not cause any harm (other than we consume 1 degree of freedom for estimating its effect; to have a reference value, when we perform an experiment with  $N$  experimental units, we have  $N - 1$  degrees of freedom available for our calculations). With this idea in mind we may block many different variables: sex, the device with

which we analyze the data (if we use several devices), the day in which we perform the experiment (if we expect differences between the different days), the time of the day at which we perform the experiment (for instance, weight measures early in the morning are different from weight measures just after feeding), the cage of the animal we are observing (depending on the experiment, cages may cause significant differences due to the interactions with the other animals in the cage or the position of the cage in the room), the litter the animal is coming from, the surgeon that operates the animal, ... Blocking variables are relatively “cheap” in terms of extra number of animals for our experiment, and may bring significant benefits in terms of explained variance.

Measuring covariates brings the same kind of benefits: removing unexplained variance. For instance, let us assume that we are measuring the blood pressure of animals. Blood pressure may depend on room temperature and sex. So, at the same time that we measure the blood pressure, we annotate the animal sex and the room temperature. Then, we model our measurements as

$$y_i = \mu + \alpha_{x(i)} + \alpha_{sex(i)} + \beta T_i + n_i''$$

The variance decomposition is now

$$\sigma_y^2 = \sigma_x^2 + (\sigma_{sex}^2 + \sigma_{Temp}^2 + \sigma_{n''}^2)$$

Our statistical analysis will be even more sensitive to differences caused by our treatment. With the same number of animals we will increase our statistical power. Or alternatively, for the same statistical power, we may reduce the number of animals in the experiment.

### 1.4.7 Paired samples

Paired samples can be seen as a special case of blocking in which individuals act as blocks, they serve as their own controls. This is the case of experiments in which we can measure before and after applying the treatment, or we can measure the response of the left and right eyes to different treatments. Experiments with twins, siblings or matched pairs (looking for another individual with similar characteristics) also fall under this category. Cross-over designs in which an individual is given a treatment for a period, and then another treatment in another period are also analyzed as paired samples. However, there are many detractors of cross-over designs, the main concerns are related to the washout period (does it really revert the individual to its initial condition?), and to the order in which the treatments are given (and in this regard, randomization is an important countermeasure as usual).

Repeated measurements can be seen as an extension of paired samples (they are also called pseudoreplications). An animal is given a treatment and, then, measured multiple times, at different parts of its body, or at different tasks. The different time points can be referred to the initial measurement at  $t = 0$ . An alternative analysis is through the standard block design in which the individual acts as a block. Typically repeated measures is treated as a split-plot design in which the subject is the factor “hard to change” (see Sec. 5.2.6).

By computing the difference between the two measurements we remove the inter-subject variability inherent to the analysis of two independent measurements (e.g. in two groups of animals, control and treatment, controls and treated animals are different, while in paired samples, they are the same individual).

This data is typically analyzed with a Student's t-test on the difference between the two measurements. However, this test assumes Gaussianity of the difference. Non-parametric alternatives exist. The most popular are: 1) sign tests (the test checks if the number of positive or negative signs in the difference is significantly different from what is expected at random); 2) Wilcoxon's signed rank tests, note that the sign test does not consider the magnitude of the difference, only its sign, Wilcoxon's signed rank test includes the magnitude of the difference and is statistically more powerful than the sign test; 3) McNemar's test if the responses are binary (yes/no, absent/present, ...); 4) permutation tests, in which the labels before and after, left and right, etc. are permuted, the distribution of the difference between the two situations is studied with these permutations, and finally the truly observed difference is compared to this distribution. In general, these more widely used statistical techniques should be used instead of less accepted tools as the use of ratios (Karp et al, 2012).

### 1.4.8 Blocking and randomization

We may combine the benefits of blocking and randomization by first blocking and then randomizing. Let us assume that we are performing an experiment with 80 animals, of which 35 are males and 45 are females. We want to block sex, then we first split the animals in two groups according to our blocking variable, and then we randomly assign to the control or treatment groups as shown in the following table.

All animals (80)	Male (35)	Control (17)
		Treatment (18)
	Female (45)	Control (23)
		Treatment (22)

We may block two variables simultaneously. For instance, we may block sex and the time of the day we perform the experiment (morning or afternoon) as shown in the following table.

All animals (80)	Male (35)	Morning (17)	Control (8)
			Treatment (9)
	Afternoon (18)	Control (9)	
		Treatment (9)	
	Female (45)	Morning (23)	Control (12)
			Treatment (11)
Afternoon (22)		Control (11)	
		Treatment (11)	

Depending on the characteristics of the block it may be balanced or not. For instance, the time-of-day block in the previous example is balanced (40 animals are tested in the morning and 40 in the afternoon); but the sex block is not balanced (because

we only have 35 males available for our experiment versus 45 females). If carefully randomized, we may balance our treatment (40 animals in the control group and 40 animals in the treatment group), as shown above.

In the following example, we block sex, the day at which the experiment is performed, and the time of the day. It is shown as an experiment planning in which at each cell we show the kind of animal (male or female, M or F) to be tested, and the kind of treatment (control or treatment, C or T). We assume that we have as many males as females available for the experiment. We see that every day there is the same number of males and females, and treatments and controls. The same happens for every time of the day. With this design we block three variables simultaneously and we do not confound the effect of the day, time of day, or sex with the treatment and control.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
9:00	FT	MC	FT	FT	MC	MT	MC	FC	MT	MT
12:00	MC	FT	MC	MC	MT	FC	FC	FT	MC	MT
15:00	MT	MC	FT	FT	FT	FC	FT	MC	FC	FC
18:00	FC	FT	MC	MC	FC	MT	MT	MT	FT	FC

We could finish this section on blocking and randomization with a *statistical mantra* for experiment design: “Control what you can, block what you cannot, and randomize the rest”. We can control our treatments, we can block those variables that we think may have an impact on the variability of the observations, and the rest should be randomized (e.g., position of the cages in the animal house racks, the order of feeding and treating, the person applying the treatment, the person performing the measurements, the order of measuring, etc.).

#### Important remarks

- Control what you can, block what you cannot, and randomize the rest.

## 1.5 Automating decision making: hypothesis testing

*In God we trust, all others must bring data. (Anonymous)*

In our daily research life we must continuously take decisions based on our observations: Is it worthy the new compound for the treatment of this disease? Is there a relationship between this gene and a given phenotype? Does this drug cause an adverse effect in the liver? Is the temperature in our animal house within specifications? Each one of these questions can be answered with yes or no, and our subsequent actions depend on the answer. Hypothesis testing is a statistical tool normally adopted to automate our decision making. For every research question and collection of independent observations (experimental units), the methodology will produce a number (the famous p-value) that we will compare to a prespecified threshold (typically, 0.05). If the p-value is above this threshold we will assume a state of affairs (e.g., our drug does not have any effect on the disease), and if the p-value is below the threshold, then we

will assume a different state of affairs (*e.g.*, the drug improves the disease state). The p-value is calculated assuming a particular state of the world (the null hypothesis, *e.g.* the drug does not have any effect) and it is the probability of observing results at least as extreme as the ones we have observed if the null hypothesis is true (the alternative hypothesis, *e.g.* the drug does help). Particularly important is the assumption under the null hypothesis of the statistical distribution of the observations. The p-value is correct if, and only if, under the null hypothesis the observations really behave as assumed. If they do not, then the p-value is only a good approximation of the true probability of observing some results at least as extreme as the ones we have observed if the distribution of the observations under the null hypothesis does not deviate too much from the assumed distribution. For strong deviations, the p-value is simply useless.

In hypothesis testing we must specify a null hypothesis ( $H_0$ ) and an alternative hypothesis ( $H_a$ ). The goal of the technique is to disprove that the null hypothesis really represents the state of the nature. With a controlled risk of committing an error (bounded by the level of confidence), we may reject the null hypothesis or fail to reject it. For this reason, we must always place our research hypothesis in the alternative hypothesis.

- Example 11: We are developing a new vaccine for a disease whose incidence is 10% (understood as the probability of acquiring the disease within a year). We expect that our vaccine lowers this incidence. The hypotheses to be used would be:

$$\begin{aligned} H_0 &: \pi_{\text{vaccine}} \geq 0.1 \\ H_a &: \pi_{\text{vaccine}} < 0.1 \end{aligned}$$

If we succeed in our vaccine, we will disprove the null hypothesis and accept the alternative one (the probability of acquiring the disease in one year is smaller than 10%). Except in likelihood ratio tests, by construction, the null hypothesis has to be the complement of the alternative. That is why we have  $H_0 : \pi_{\text{vaccine}} \geq 0.1$ .

- Example 12: We are developing a new vaccine for a disease. In our experiment we will have two groups of animals and both will be challenged in the same way with the pathogen. We expect that the proportion of vaccinated animals that acquire the disease is smaller than the proportion of control animals that acquire the disease. The hypotheses to test are:

$$\begin{aligned} H_0 &: \pi_{\text{vaccine}} \geq \pi_{\text{control}} \\ H_a &: \pi_{\text{vaccine}} < \pi_{\text{control}} \end{aligned}$$

- Example 13: For the disease of the previous two examples, we want to explore the effect of an environmental variable to the incidence of the disease. But, actually, we do not know whether the environmental variable will promote, inhibit or be neutral for the transmission of the disease. For our experiment, we will have two groups, control and treatment. The treated group will be exposed to the environmental variable, while the control will not. The hypotheses in this case

are:

$$H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$$

$$H_a : \pi_{\text{treatment}} \neq \pi_{\text{control}}$$

We may have noticed that Examples 11 and 12 use inequalities ( $\geq$ ) while Example 13 uses an equality ( $=$ ). This makes Examples 11 and 12 to be one-tail tests, and Example 13 a two-tails test. This technical difference makes an important experimental difference: the number of animals for one-tail tests is smaller than for two-tail tests. The reason is that in two-tails tests we want to disprove the null hypothesis in more cases (if the incidence in the treatment group increases or decreases), while in the one-tail tests we want to disprove the null hypothesis only if the proportion decreases. The extra requirements for the two-tail tests call for a larger number of animals. This is another reason to establish correctly the way we will analyze the data before carrying out the experiment, because its statistical power depends on the number of animals we have chosen, and it is not the same in a one-tail than a two-tail test.

Examples 11, and 12 are examples of superiority tests (our treatment is superior to a reference, Example 11, or a control group, Example 12). Example 13 is an example of significance tests (our treatment is significantly different from the control). Superiority and significance tests are the most common ones used in animal research. However, there are other classes of tests.

- Example 14: We are developing a generic vaccine that is supposed to work in the same way as the reference commercial vaccine in the market. If we succeed in such an endeavour we will reject the null hypothesis and accept the alternative one. Consequently, we must use the equality in the alternative hypothesis instead of the null hypothesis as we did in Example 13

$$H_0 : \pi_{\text{generic}} \neq \pi_{\text{reference}}$$

$$H_a : \pi_{\text{generic}} = \pi_{\text{reference}}$$

These kinds of tests are called equivalence tests and the way to calculate the p-value is more involved than in the case of significance tests. Similarly, the number of animals of equivalence tests is higher than for significance tests.

- Example 15: We are developing a new vaccine that is supposed to work at least as well as the reference commercial vaccine in the market. In this case the hypotheses are

$$H_0 : \pi_{\text{new}} > \pi_{\text{reference}}$$

$$H_a : \pi_{\text{new}} \leq \pi_{\text{reference}}$$

Note that this time the equality sign falls under the alternative hypothesis as opposed to Examples 11 and 12 in which the equality sign fell under the null hypothesis. These tests are called non-inferiority tests (our new drug is at least as good as the reference), and the way to calculate the p-value and the number of animals is different from the significance tests.

It is important to correctly set from the beginning the kind of test because it affects the number of animals required for our experiment, and because if incorrectly

set, we will never be able to prove our research hypothesis (remember that the research hypothesis goes to the alternative hypothesis). We can reject the null hypothesis and, consequently, accept the alternative hypothesis. We have succeeded in proving that the null hypothesis is false. But we can never accept the null hypothesis, simply we have failed to prove that the null hypothesis is false. It is like in legal trials, we accumulate evidences to disprove the innocence of the defendant, but we can never prove his innocence (many trials absolve the defendant because there not enough evidences of his guilt). In hypothesis testing, each new observation brings new evidence about the falseness of the null hypothesis until there is so much evidence (the p-value is so low) that we reject the possibility that the null hypothesis describes the state of the nature. The following example shows why failing to show the falseness of the null hypothesis does not automatically imply that it is true.

- Example 16: Michael Jordan (of the Chicago Bulls) and I go to play some basketball together. We try 7 free throws and he scores 7 (of 7) and I score 3 (of 7). Do Michael Jordan and I have the same skill in scoring free throws? Do we have the same success probability?

$$\begin{aligned} H_0 : \pi_{Jordan} &= \pi_{me} \\ H_a : \pi_{Jordan} &\neq \pi_{me} \end{aligned}$$

The p-value of this experiment is 0.062, as it is above 0.05, with a confidence level of 95% we cannot reject the hypothesis that Michael Jordan and I have the same scoring probability in free throws. But it does not mean that Michael Jordan and I *do* have the same scoring probability. It means that, with the acquired evidence, we cannot reject the hypothesis that we are equally good (if we make a longer experiment with more free throws, it would become clear that we do not have the same skills).

In animal research, the example above shows that we can never accept the null hypothesis, simply we did not accumulate enough evidence to show it is false. That is why calculating the sample size in advance is so important. We will determine the smallest difference we want to detect, then we can calculate the number of experimental units needed to detect that difference. For instance, if we want to detect a difference of at least 50% when the percentage of Michael's free throws is about 85%, then we will need at least 22 free throws each. If we want to detect a difference of at least 5% between Michael Jordan and Larry Bird (of Boston Celtics), then the number of free throws rises to 1,252. (For the curious, the historical percentage of free throws of Larry Bird was 88.6% and the one of Michael Jordan 83.5%.) The smallest size we want to detect, the 50% or 5% in the example of Michael Jordan, is called the effect size and we need to specify it in advance in order to calculate the number of animals required for our experiment. By specifying the effect size, we are specifying the sensitivity of our experiment and will adjust accordingly the number of animals to our sensitivity requirements.

**Important remarks**

6. The smaller the difference we want to detect (the effect size), the larger the number of experimental units required for the experiment.
7. We can reject or not the null hypothesis.
8. Failing to reject the null hypothesis does not make it true.

**1.5.1 An intuitive introduction to hypothesis testing**

The goal of this section is to give a non-technical insight into the hypothesis testing procedure. The reader is referred to [Ellenberg \(2014\)](#) for an excellent general public book on statistical, and mathematical in general, thinking. Ellenberg manages to smoothly introduce the reader into many complex statistical concepts.

As we have already stated, the goal of hypothesis testing is to disprove the null hypothesis. The p-value is a measure of our “surprise” to see the observed results if the null hypothesis is true. Suppose we are studying the effect of a new drug. For doing so, we follow 100 animals with a particular disease that has a mortality of 10%. Half of the animals receives the drug, while the other half does not. On average we should expect to have about the same number of deaths in both groups (about 5, that is, 10% of 50), if the drug does not help the animal to overcome the disease. Actually, seeing exactly 5 deaths in one of the groups, although it is the most likely event, it has only a probability of 18.5%, and there are other frequent events as seeing only 3, 4, 6 or 7 deaths. Also, having exactly the same number of deaths in both groups is a relatively infrequent result even if the drug does not help. Assuming that the drug does not help, the null hypothesis is true, only in 13.3% of the experiments we will see this result, while in 43.3% of the experiments we will see more deaths in the control group, and in 43.3% of the experiments we will see more deaths in the treated group. In other words, simply seeing fewer deaths among the animals receiving the drug is not a guarantee that the drug is working.

Assume we do not see any death in the treated group. Each of the animals has a survival probability of 0.9. So, if the null hypothesis is true (the drug does not cure this disease), observing 50 survivals occurs with probability  $0.9^{50} = 0.005$ , i.e., only in 1 of 200 similar experiments in which the drug does not help. Consequently, after seeing this results, we would be rather surprised that the drug does not help. The p-value quantifies this surprise (the calculation of the p-value for the comparison of two groups is different from the  $0.9^{50}$  that we have shown above, but this number illustrates the idea in a very simple manner). Once we have calculated the p-value, we need some mechanism to take the decision of whether it is worthy to continue studying this drug, or we should devote our efforts to some other candidate. This is done by comparing the p-value to a pre-established threshold (typically, 0.05, that is 1 in 20). If the p-value is below this threshold, we declare the drug effects as significantly different from no effect. Note that the goal of hypothesis testing is helping us to take a decision, not revealing the truth. The truth will always remain unknown because we might have been unlucky with our sample (Type I and II errors in the following section). However, if

our experimental design is sufficiently powered ( $1 - \beta$  in the following section, that is, if we have tested enough individuals) and the drug effect is declared non-significant, it does not mean that the drug exerts no effect at all (it is hard to think that a chemical compound goes totally unnoticed in an organism), but its effect is sufficiently small as to not be distinguished among the biological and measurement variability normally observed in animals. Consequently, very likely this compound does not deserve further efforts.

Note that the expression “statistically significant” does not mean “practically important”, it simply means that its effect is clearly different from the effect expected under the null hypothesis. The difference could be still small enough to be of practical importance. For instance, if a drug significantly increases the risk of blood clotting (could result in a severe or fatal event) with respect to another treatment by a factor two, this is not a sufficient reason to abandon the treatment. If the probability of blood clotting of the first treatment is very small, twice this probability is still very small, and the benefits of the drug in many other aspects may largely compensate the risk increase.

We have to be careful with the 0.05 threshold of the classical statistical testing. This threshold implies that, on average, in 1 out of 20 experiments in which the drug does not make a difference, we will declare its effects as significantly different from no effect. If we are screening thousands of compounds (technically this problem is known as multiple testing), this is a very large number of false positives and some correction is needed (see Sec. 1.5.3). The same happens in other contexts in which many statistical tests are performed. In functional Magnetic Resonance Imaging, fMRI, of the brain, each voxel is statistically compared to some background distribution to determine if it is activated or not, if there is brain activity at that location or not. This is very useful to map the brain regions in charge of the different cognitive or physiological tasks. However, [Bennett et al \(2009\)](#) warns against uncorrected tests as they might result in significantly activated brain areas in dead salmon! When the multiple testing corrections are performed, these significantly activated brain areas disappear, as expected from a dead body.

The problem with the threshold of 0.05 (on average, 1 in 20 experiments in which the null hypothesis is true is declared to have statistically significant results) is that it is too high, leading to many false positives. The following example is taken from [Elenberg \(2014\)](#). Imagine that we are haruspices trying to predict the outcome of given events by reading the entrails of sacrificed animals. We try to predict the price of the NYSE and we fail, to predict the next U.S. president and we fail, to predict the consumption of natural gas next winter and we fail, ... We fail in most of our predictions, but thanks to the gods, we successfully predict the occurrence of an earthquake next month. Our predictions had a p-value below the well accepted threshold of 0.05 and we are allowed to publish our results in the International Journal of Haruspicy. Reading the entrails of animals are not related at all with any of our predicted outcomes (the null hypothesis is true in absolutely all of our experiments). But as we, and the thousands of other haruspices around the world, are making “random” predictions, just by chance, on average, 1 in 20 of those predictions will fall below the 0.05 decision threshold. This is nothing but the verification of the principle that improbable events are not impossible, and actually they occur (the probability of winning the lottery is extremely small, but among the many lottery players, one of them is winning). Fisher, one of

Table 1.7: Average number of genes in each of the situations (P represents protein and D disease).

	P related to D	P not related to D	Sum
Test statistically significant	9	5,000	5,009
Test not statistically significant	1	94,990	94,991
Sum	10	99,990	100,000

the fathers of statistical inference, wrote that “A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this significance level”. The important part is “a properly designed experiment rarely fails to give”, and not “succeeds once in giving”. This calls for either extremely low p-values (say,  $10^{-6}$ , see Sec. 1.5.3, although this number is not meant in any way as a suggested threshold of anything, but just as an illustration of what a extremely low p-value means) or confirmation of the results by further experimentation. Biomedical research is not haruspicy, but in recent years there have been important alarms about the reproducibility of research studies (Ioannidis, 2005; Begley and Ioannidis, 2015; Baker, 2016). There are systemic reasons for this like systematically small experimental groups, the pressure to publish significant results, the fact that negative results cannot normally be published, the fact that the same or similar problems are studied by many groups worldwide and, just by chance, one of them gets a significant result, the fact that results tend to be published only once (if it is a positive result, a second, third, ... group cannot normally publish the confirmation of the result; and if it is a negative result it is more difficult to publish because it goes against the “established, peer-reviewed” previous result), the researcher freedom to choose the data to analyze and the analysis technique (Ioannidis, 2005; Simmons et al, 2011), etc. In the case of haruspicy we may quickly realize that our statistically significant prediction belongs to the Type I error class (the p-value was significant just by chance, because the null hypothesis -entrails cannot predict future events- is always true). But in the case of the relationship of a particular gene to a particular disease, our intuition is much less protected (because we know that genes and diseases are connected).

One of the problems of the standard hypothesis testing is that it does not take into account prior probabilities. To illustrate the problem, let us consider the following scenario. The human proteome is made of approximately 100,000 proteins. Out of which, let us say that only 10 are actually related to any particular disease (the prior probability of any randomly taken protein being related to that particular disease is  $10/100,000=10^{-4}$ ). If we make a standard statistical test with all proteins with the standard parameters (confidence level of 95% and statistical power of 90%, see next section for the formal definition of these parameters) we would end, on average, with the distribution shown in Table 1.7. The statistical test is not the problem: the proteins that are not related to the disease rarely pass the test, and most of the proteins that are related to the disease pass the test. The problem is that the number of genes not related to the disease are massively preponderant with respect to the proteins that are related to the disease. If we incorporate our knowledge about the prior probability of a randomly chosen protein being related to that particular disease (already reflected

Table 1.8: Possible situations encountered when performing an hypothesis test

	$H_0$ is false	$H_0$ is true
$H_0$ is rejected	True Positive (Correct)	False Positive (Type I error)
$H_0$ is not rejected	False Negative (Type II error)	True Negative (Correct)

in Table 1.7), what is called a Bayesian approach, then the probability that the protein is actually related to the disease knowing that the test is statistically significant is only  $9/5,009=0.0018$ . In other words, only in 0.18% of the experiments in which the statistical test states that the protein is related to that particular disease, this statement is correct. Under this point of view, we better understand that a statistically significant result opens an interesting research line that requires more investigation and confirmation by similar or related experiments. Genomics and Proteomics data analysis are well aware of this problem, and they duly take countermeasures (see Sec. 1.5.3). However, in many biomedical domains we do not know the *a priori* probabilities, and we do not simultaneously perform thousands of tests as to take multiple testing protections. But, when we consider the number of experiments performed in the lifespan of a single researcher, we wonder if we should have not increased the confidence level of our statistical tests. As researchers, we are saved by the fact that we do not perform a “randomly chosen experiment” among a “universe of 100,000 possible experiments”, but we carefully choose the experiment we will perform “among the universe of experiments that our current knowledge predicts that they have higher chance of giving a positive result”. But in any case, it is important being aware of this problem of reproducibility before we launch ourselves at reporting statistically significant findings.

The use of the p-value as the most important outcome or the only determinant of the importance of a result in any biomedical research has been heavily criticized. In Sec. 3.2 we further discuss about these criticisms and suggest alternatives.

## 1.5.2 Statistical power and confidence

Table 1.8 shows the different situations we may encounter when performing an hypothesis test. In reality,  $H_0$  can be true or false, and our hypothesis test may reject it or not. If  $H_0$  is false and we rejected it, we made a correct decision. The same if  $H_0$  is true and we cannot reject it. However, there are two situations in which we can make wrong decisions: 1) if  $H_0$  is true and we reject it (*false positive*), and 2) if  $H_0$  is false and we cannot reject it (*false negative*). In the statistical literature, the first kind of errors are called Type I errors, while the second Type II errors. A test is said to be positive if  $H_0$  is rejected, and negative if  $H_0$  cannot be rejected. In this way, Table 1.8 also labels each one of the situations as true or false positive or negative.

- **Example 17 (Type I error):** Let us assume that  $H_0$  is “the new vaccine does not reduce the probability of infection” ( $H_0 : \pi_{vaccine} \geq \pi_{control}$ , see Example 12 in the previous section). Let us also assume that in reality, this statement is true. If we commit a Type I error, after analyzing our observations we would incorrectly believe that the new vaccine reduces the probability of infection and we would keep on working on its development, even if in reality the new vaccine is useless.

- **Example 18 (Type II error):** Let us assume that  $H_0$  is “the new vaccine does not reduce the probability of infection” ( $H_0 : \pi_{\text{vaccine}} \geq \pi_{\text{control}}$ , see Example 12 in the previous section). Let us also assume that in reality, this statement is false and our new vaccine is really effective. If we commit a Type II error, after analyzing our observations we would incorrectly believe that the new vaccine is useless, and we will stop researching into it, abandoning a research line that could have led to a successful vaccine.

The statistical theory for hypothesis testing explicitly controls the probability of committing Type I and II errors assuming that we correctly identified the distribution of the observations if the  $H_0$  is true. These probabilities are called  $\alpha$  and  $\beta$  respectively. Traditionally,  $\alpha$  is set to 0.05, that is in 5% of our experiments in which the new vaccine is not useful, we will incorrectly believe it is helpful. There is nothing special about the number 0.05 except tradition. We could have lowered it to 0.5%, and we would be even more conservative stating that a new treatment is useful only if there is much evidence supporting it (actually, this suggestion has been recently proposed as a way to increase the reproducibility of biomedical experiments, [Benjamin et al \(2018\)](#); on a related topic [Simmons et al \(2011\)](#) has shown that the freedom of the researcher to choose the variables to study from a set of collected data could effectively raise the Type I error up to 60%). The complement of  $\alpha$ , that is  $1 - \alpha$ , is called the statistical confidence, and it is traditionally set to 95%. There is less consensus about  $\beta$ , but typical values are 10% or 20%, meaning that in 10% or 20% in which the new treatment is useful we will miss this effect, and incorrectly believe that it is not. Larger values of  $\beta$  are not so sensible because it would mean that in our experimentation we would miss many useful treatments and it compromises the ultimate goal of experimental research. As an extreme example, if  $\beta = 0.5$  we might as well have tossed a coin. The complement of  $\beta$ , that is  $1 - \beta$ , is called the statistical power.

These probabilities are calculated assuming that if  $H_0$  is true, we correctly know the distribution of the observations and that errors are strictly caused by sampling errors. In some experiments we may have been “unlucky” with the animals in our experiment which are “extremes” of the distribution if  $H_0$  is true, leading us to incorrect conclusions. The way of controlling the Type I and II errors is by calculating the sample size needed to maintain these probabilities under desired upper bounds (typically  $\alpha = 0.05$  and  $\beta = 0.1$  or  $0.2$ ). However, sample size calculations assume that sampling errors are the only ones in place. Systematic errors (see Section 1.3) completely invalidate the calculations and will result into much higher error probabilities.

For experiments with live animals it is essential to use 3Rs principles; carrying out a power calculation is an excellent way to produce a robust justification of animal numbers for funding bodies and regulatory authorities.

#### **Important remarks**

9. For a fixed confidence level and effect size, increasing the number of animals increases our statistical power: if our treatment makes a difference, we will detect it with more probability.

10. For a fixed confidence level and statistical power, increasing the number of animals increases our experiment sensitivity (the detectable effect size is smaller): if our treatment makes a small difference, we will detect it (with the probability specified by the statistical power).
11. By calculating the sample size before performing the experiment we can control the Type I and II errors at will, assuming that there are not systematic errors causing bias.

It is important to realize at this point that the p-value itself is also a random variable (Boos and Stefanski, 2011). If we repeat the same experiment from the same population (but different realization of the sample), we would get different results.

- **Example 19:** We perform an experiment in order to determine if there is any difference in the systolic blood pressure between two mouse strains. We want to have a statistical power of 80% if the difference between the two strains is larger than 20 mmHg. We assume that the standard deviation of the measurements in each one of the groups is also 20 mmHg. For a confidence level of 95%, we need a sample size of  $N_1 = N_2 = 17$  animals per group.

In Fig. 1.5 we show the p-values and the confidence intervals for 1,000 simulated experiments in which the true underlying difference is 20 mmHg. We can see that the p-values range from highly significant results (p-value  $< 10^{-8}$ ) to non-significant p-values (the maximum p-value is 0.935).

#### Important remarks

From the previous example we draw several conclusions:

12. The p-value is itself a random variable with a large variability. Due to random sampling we may have experiments with the same underlying truth, but some of them are not significant and others have a significance of  $10^{-8}$ . Even, among the significant results, there are several orders of magnitude of difference between experiments (the effective range of significant results span from  $10^{-1.3}$  to about  $10^{-5.5}$ ).
13. The example shows that if we get a p-value of  $10^{-5}$  in an experiment it does not mean that if we repeat the experiment we will most likely get a highly significant result (in the example, experiments with  $10^{-5}$  had the same underlying truth as experiments with  $10^{-1}$  or  $10^{-0.5}$ ). Actually, if the p-value is at the significance threshold (typically, 0.05), then there is a probability of 50% of repeating the experiment and having a result in either side (significant or non-significant, Greenwald et al (1996)).
14. The freedom of many researchers to choose the data and the variables that participate in the analysis may inflate the effective false positive rate (Type I errors,  $\alpha$ ) well above the 0.05 level (up to 0.6 as reported by Simmons et al (2011)). The solution suggested by these authors is reporting the specific

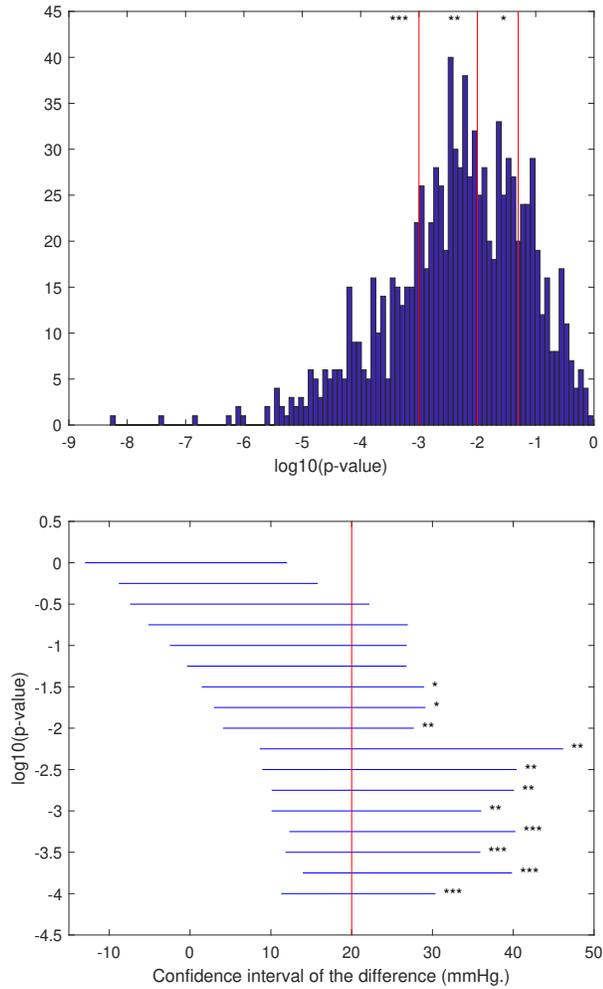


Figure 1.5: Top: Histogram of p-values (represented in logarithmic scale) of 1,000 simulated experiments as the one described in Example 19 in which the true underlying difference between the two groups is 20 mmHg. Bottom: Confidence interval of some of the experiments. On the left scale, we see the logarithm of the p-value of that experiment. We have labeled the experiments and p-values with \*\*\* if the p-value is smaller than 0.001, \*\* if it is smaller than 0.01, and \* if it is smaller than 0.05.

choices performed, all the data measured, and the analysis with and without the removed data.

15. Confidence intervals are much more stable and they do not exhibit these wild variations (Cumming, 2008). Even, some of the non-significant results show confidence intervals that point to “almost significant” results. In these cases, repeating the experiment with a higher statistical power would help elucidating whether there is really a difference between both groups or not. There have been recent alarms on the reproducibility of experiments in Science (Begley and Ioannidis, 2015; Baker, 2016) and its economical impact which has been estimated to be about 28B\$/year in the US (Freedman et al, 2015). Among many other reasons, experiments with low statistical power and poor (but significant p-values) are behind this recent concern. Several actions have been proposed:
  - Lowering the significance threshold from 0.05 to 0.005 and relabelling the experiments with p-values in the range 0.05 to 0.005 as suggestive results (Benjamin et al, 2018).
  - Increase the statistical training of practitioners (Freedman et al, 2015; Munafò et al, 2017) and this is exactly the topic of the Modules 10 and 11 of the European Directive 2010/63/EU on the protection of animals used for scientific purposes (see Table 1.9) and associations as the Education and Training Platform for Laboratory Animal Science (ETPLAS, <https://etplas.eu/>).
  - Blinding (Sec. 1.3) and pre-registration of the experiments (Munafò et al, 2017) as a way to prevent unconscious biases.
  - Encouraging collaboration and team science (Munafò et al, 2017), especially including in the team researchers without any emotional attachment to the project that can help to supervise the experiment design, monitoring, and bring different research cultures. Multisite studies help to avoid biases and to confirm the results among different groups.
  - Improving the quality and width of reports (Munafò et al, 2017) in two ways: by publishing preprints and other ways to avoid the publication bias, and by increasing the amount of raw data, analysis scripts, etc. towards a more open science.

### 1.5.3 Multiple testing

As presented above, the probability of Type I and II errors refer to a single hypothesis test. However, in current research, the technology allows us to perform many simultaneous tests. For instance, in drug screening we can test the effect of thousands of compounds on a cell culture, or microarray experiments give the expression level of thousands of genes. Let us illustrate the problem of multiple testing with microarray

---

<b>Module 10</b>	
10.1	Describe the concepts of fidelity and discrimination.
10.2	Explain the concept of variability, its causes and methods of reducing it (uses and limitations of isogenic strains, outbred stocks, genetically modified strains, sourcing, stress and the value of habituation, clinical or sub-clinical infections, and basic biology).
10.3	Describe possible causes of bias and ways of alleviating it (e.g. formal randomisation, blind trials and possible actions when randomisation and blinding are not possible).
10.4	Identify the experimental unit and recognise issues of non-independence (pseudoreplication).
10.5	Describe the variables affecting significance, including the meaning of statistical power and p-values.
10.6	Identify formal ways of determining of sample size (power analysis or the resource equation method).
10.7	List the different types of formal experimental designs (e.g. completely randomised, randomised block, repeated measures [within subject], Latin square and factorial experimental designs).
10.8	Explain how to access expert help in the design of an experiment and the interpretation of experimental result
<hr/>	
<b>Module 11. Good scientific practice</b>	
11.3	Describe the principles of a good scientific strategy that are necessary to achieve robust results, including the need for definition of clear and unambiguous hypotheses, good experimental design, experimental measures and analysis of results. Provide examples of the consequences of failing to implement sound scientific strategy.
11.4	Demonstrate an understanding of the need to take expert advice and use appropriate statistical methods, recognise causes of biological variability, and ensure consistency between experiments.
11.6	Describe situations when pilot experiments may be necessary.
11.8	Explain the importance of rigorous scientific technique and the requirements of assured quality standards such as GLP.
11.9	Explain the importance of dissemination of the study results irrespective of the outcome and describe the key issues to be reported when using live animals in research e.g. ARRIVE guidelines.

---

Table 1.9: Learning outcomes of the Modules 10 and 11 (only those directly related to the topics of this book) of the European Directive 2010/63/EU. Modules 10 and 11 is a pre-requisite for people who will be designing projects (Function B). Module 10 is also beneficial, although not a prerequisite, for scientists carrying them out (Function A).

experiments. The current methodology to analyze gene expression (Allison et al, 2006) is much more involved than the extremely simplified version exposed in this chapter. But it illustrates the need for developing more advanced statistical tools.

Let us assume that we measure the gene expression level of 20,000 genes in a group of healthy animals and a group of diseased animals. We want to identify those genes that are related to our disease (the relationship can be causal, a change on the expression of this gene is partially causing the disease, or consequential, the expression of this gene is changed because there are other genes that have also changed their expression level). Let us assume that 1,000 of the genes are truly affected by the disease. However, this number is unknown to us and that is why we are performing the experiment. For each gene, we perform an hypothesis test

$$\begin{aligned} H_0 &: \mu_{healthy} = \mu_{disease} \\ H_a &: \mu_{healthy} \neq \mu_{disease} \end{aligned}$$

Let us assume that we design our experiments with 90% of statistical power and 95% statistical confidence. Due to Type II errors, of the 1,000 related genes, we will correctly identify 900 and miss 100. Of the 19,000 unrelated genes and due to Type I errors, we will incorrectly think that 950 of them are related to the disease. All these information is shown in Table 1.10.

Table 1.10: Average number of genes in each of the situations.

	$H_0$ is false	$H_0$ is true	Sum
$H_0$ is rejected	900	950	1,850
$H_0$ is not rejected	100	18,050	18,150
Sum	1000	19,000	20,000

From the experiment we will obtain 1,850 (=900+950) positives. Once the hypothesis test rejects the null hypothesis, the probability that the gene is actually related to the disease (it is a True Positive given it is a Positive) is

$$p_{TP|P} = \frac{900}{1,850} = 48.6\%$$

That is, more than half of the related genes are False Positives, instead of the 5% used in the design. This ratio is called the False Discovery Rate . We have encountered two problems in this example: 1) Typical experiment designs do not consider the *a priori* probability of  $H_0$  being true; 2) The confidence level considers a single test and not a family of tests. The first problem can be addressed through Bayesian sample size determination (Adcock, 1997). The second problem through multiple testing correction. Most of them change the  $\alpha$  value to be used in sample size determination and hypothesis testing considering the total number of tests to be performed,  $K$ . The  $K$  tests still have a specified family Type I error (typically  $\alpha_{family} = 0.05$ ), but each individual test has a much smaller  $\alpha$ . A well-known correction is the Bonferroni correction:

$$\alpha = \frac{\alpha_{family}}{K}$$

This is too conservative and other corrections have been suggested like Sidak

$$\alpha = 1 - (1 - \alpha_{family})^{\frac{1}{K}}$$

A very popular approach to control the family Type I error is the Benjamini-Hochberg procedure. First, we sort the  $K$  p-values of the  $K$  tests in ascending order ( $p_1, p_2, \dots, p_K$ ). Second, we reject the null hypothesis for the  $k$ -th test if

$$p_k \leq k \frac{\alpha_{family}}{K}$$

Once we cannot reject the null hypothesis for the test  $k_0$ , we cannot reject it for  $k > k_0$ .

#### Important remarks

16. Significance answers the questions:

- If  $H_0$  is true, what is the probability of incorrectly rejecting it?
- Of all the experiments you could run in which  $H_0$  is true, what is the fraction in which you will reach the conclusion that the results are statistically significant?

Power answers the questions:

- If  $H_0$  is false, what is the probability of correctly rejecting it?
- Of all the experiments you could run in which  $H_0$  is false, what is the fraction in which you will reach the conclusion that the results are statistically significant?

False Discovery Rate answers the questions:

- If a result is statistically significant, what is the probability that  $H_0$  is true?
- Of all the experiments that reach a statistically significant conclusion, what is the fraction in which  $H_0$  is true?

17. Significance level, statistical power and FDR depend on the sample size, the effect size and the population variance. The following analog explains these ideas. You send your child into the basement to find a tool. He comes back and says “It isn’t there”. What do you conclude? Is the tool there ( $H_0$ ) or not ( $H_a$ )? Your conclusion depends on:

- How long the kid has been looking for. (sample size)
- How large the tool is (it is easier to find a snow shovel than a small screw-driver to fix glasses). (effect size)
- How messy the basement is. (population variance)

### 1.5.4 A worked example

Let us now illustrate all these ideas with a particular example. In the following we provide an extremely simplified model of the functioning of a thermostat that keeps constant the animal house temperature. It will serve our illustration purposes, but a real operation of a thermostat would require at least two hypothesis tests because the temperature is specified to be within a range, and not a single value as in our example, and because there are variations of the temperature along the day that are not considered by our simplified model.

Let us assume that we are in charge of the thermostat of the animal house and that our aim is to keep constant the animal house temperature at a fixed value of 21°C. Under normal operation, the temperature mean is  $\mu = 21^\circ\text{C}$ , temperature measurements are Gaussianly distributed, and they have a standard deviation of  $\sigma = 0.5^\circ\text{C}$ . We measure the temperature once every hour, and we compute an average using the last 24 measurements. In a particular day, our average is 20.76°C, that is not exactly 21°C, but it is not too far either. Should we assume that the thermostat is malfunctioning, and take the necessary compensatory actions? Doing it when it is not necessary incurs some operational costs, conversely not doing it when it is necessary biases all the experiments in the animal house.

Hypothesis testing provides a simple mechanism of taking these decisions. It computes the probability of the observing a value at least as extreme as the one we have observed, 20.76°C, if the thermostat is working correctly. This probability is known as the p-value, which in this case is 0.0188 as we will justify below. This value is smaller than 0.05, consequently, we would reject the hypothesis that the thermostat is working correctly and go for maintenance. In the following we show how we have arrived to this probability.

Let us assume that we take a single measurement of the temperature. This observation is 21.17°C. At this moment, our best estimate of the mean is

$$\hat{\mu} = 21.17$$

and our uncertainty about the mean (measured as the standard deviation of our estimate) is the same as the variability of the underlying measurements

$$\sigma_{\hat{\mu}} = \sigma = 0.5$$

Fig. 1.6 shows the presumed distribution of the temperature measurements if the thermostat is working correctly. It shows our, for the moment, single observation, and with this observation, the observed mean and the uncertainty about the location of the actual mean.

After two hours we have collected two more samples of the temperature (20.52 and 21.55). At this moment, our best estimate of the underlying mean is

$$\hat{\mu} = \frac{21.17 + 20.52 + 21.55}{3} = 21.08$$

and we have reduced our uncertainty thanks to the acquisition of more information (see Fig. 1.7)

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{3}} = \frac{0.5}{\sqrt{3}} = 0.29$$

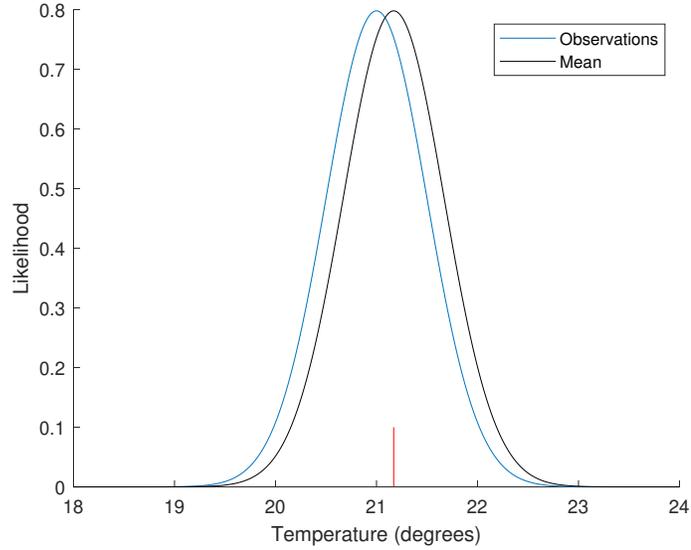


Figure 1.6: Blue: Presumed distribution of the measurements if the thermostat is working correctly. Red: One observation of the temperature. Black: *A posteriori* distribution of the mean after one observation.

As we acquire more and more samples, the uncertainty about the mean is further reduced. After 24 samples, our estimate of the mean is the average of the last 24 samples, that is, 20.76 and the uncertainty has been reduced to (see Fig. 1.8)

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{24}} = \frac{0.5}{\sqrt{24}} = 0.10$$

At the sight of this figure we see that, although 20.76°C is rather close to 21°C, with the acquired evidence, it would be rather unlikely that the true underlying mean is 21°C. We need now some mechanism to determine whether we should reject the hypothesis that the thermostat is correctly working or not.

This tool is hypothesis testing. Our null hypothesis is that the thermostat is correctly working:

$$\begin{aligned} H_0: & \mu = 21 \\ H_a: & \mu \neq 21 \end{aligned}$$

We need to know the distribution of a random variable, also called a statistic, if the null hypothesis is true. In the case that the measurements are normally distributed and their standard deviation is known, such a statistic is

$$z = \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1) \quad (1.12)$$

that is the difference between the observed mean,  $\hat{\mu}$ , and the reference mean,  $\mu_0$  (in our example  $\mu_0 = 21$ , over the standard deviation of our mean estimate is Gaussianly

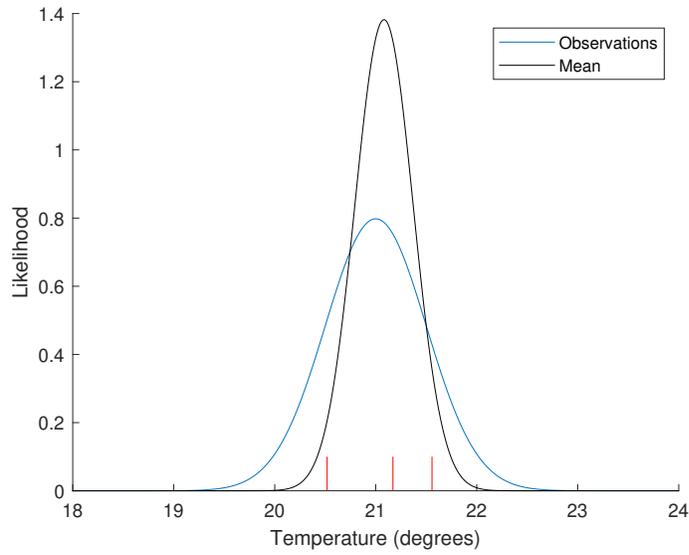


Figure 1.7: Blue: Presumed distribution of the measurements if the thermostat is working correctly. Red: Three observations of the temperature. Black: A *posteriori* distribution of the mean after three observations.

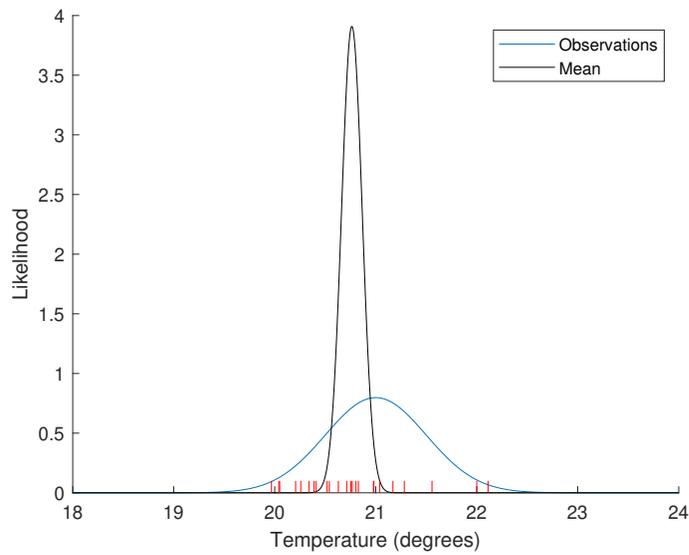


Figure 1.8: Blue: Presumed distribution of the measurements if the thermostat is working correctly. Red: 24 observations of the temperature. Black: A *posteriori* distribution of the mean after 24 observations.

distributed with 0 mean and standard deviation 1. If we plug in our data, we obtain

$$z = \frac{20.76 - 21}{0.1} = -2.35$$

The probability of observing a value as extreme as -2.35 (or lower) or 2.35 (or upper) is 0.0188, that is, if the null hypothesis is true (and the thermostat is correctly working) we would only observe a z statistic as large as 2.35 or larger in only 1.88% of the experiments taking 24 independent samples (see Fig. 1.9). This 0.0188 is the p-value. We reject the null hypothesis if this p-value is below a given threshold, typically 0.05 (=  $\alpha$ ). Consequently, in this example we would reject the hypothesis that the thermostat is correctly working and go for maintenance. The two vertical dashed lines are located at the z values for which the area in the central region is 95% (=  $1 - \alpha$ ) and they are represented as  $z_{\frac{\alpha}{2}}$  and  $z_{1-\frac{\alpha}{2}}$ , meaning that the area from  $-\infty$  to these two points are  $\frac{\alpha}{2}$  and  $1 - \frac{\alpha}{2}$ , respectively.

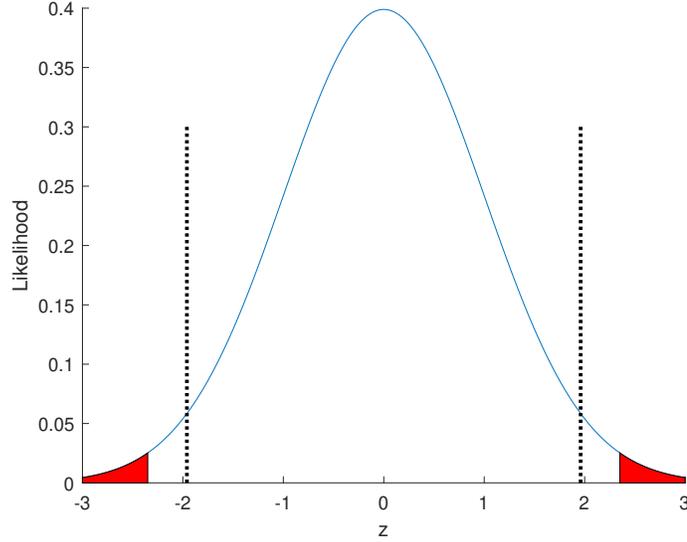


Figure 1.9: The red shaded area is the probability of observing a z statistic as large as -2.35 or larger (in absolute value) if the null hypothesis is true. The two vertical dashed lines indicate the z statistic for which that area is 0.05.

Fig. 1.9 shows the rejection area in terms of the z statistic, but we could map it back to the temperature space (see Fig. 1.10), by exploiting

$$z = \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} \Rightarrow \hat{\mu} = \mu_0 + \frac{\sigma}{\sqrt{N}}z \quad (1.13)$$

that is,

$$\hat{\mu} \sim N\left(\mu_0, \frac{\sigma^2}{N}\right)$$

In plain words, our estimate of the mean if the null hypothesis is true has mean  $\mu_0$  and variance  $\frac{\sigma^2}{N}$  (remember that the variance is the square of the standard deviation).

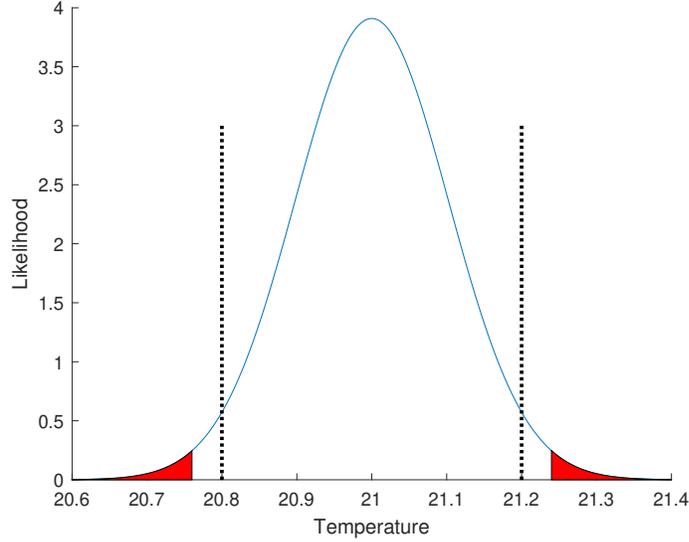


Figure 1.10: The red shaded area is the probability of observing a sample mean as far from  $21^\circ$  as  $20.76$  or further if the null hypothesis is true. The two vertical dashed lines indicate the temperatures for which that area is  $0.05$ .

## 1.6 A primer in sample size calculations

We can at this point partly understand the logic behind sample size calculation. When we do the experiment we will reject the null hypothesis if our sample mean is further than a given distance from the reference temperature,  $21^\circ\text{C}$ , see Fig. 1.10 and Eq. 1.13:

$$\frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} < z_{\frac{\alpha}{2}} \quad \text{or} \quad \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} > z_{1-\frac{\alpha}{2}} \quad (1.14)$$

Because of the symmetry of the Gaussian function this is equivalent to

$$\left| \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} \right| > z_{1-\frac{\alpha}{2}}$$

For  $\alpha = 0.05$ ,  $z_{1-\frac{\alpha}{2}}$  takes the value  $1.96$ . The distance  $|\hat{\mu} - \mu_0|$  is called the effect size, and it is the minimum difference from the reference mean that we will be able to detect with a Type I error of  $\alpha$ . Let us rewrite the effect size as  $\Delta$ . We may rearrange the Eq. 1.14 and solve for the sample size

$$N > \left( \frac{z_{1-\frac{\alpha}{2}} \sigma}{\Delta} \right)^2 = \left( \frac{z_{1-\frac{\alpha}{2}}}{\Delta/\sigma} \right)^2 \quad (1.15)$$

If we want to detect with a confidence of 95% a change of  $0.25^{\circ}\text{C}$  in our thermostat example, whose standard deviation is  $0.5^{\circ}\text{C}$ , then we simply need to plugin our specifications into Eq. 1.15

$$N > \left( \frac{1.96}{0.25/0.5} \right)^2 = 15.36$$

That is, we need at least 16 samples to detect such changes. With temperature samples, we may use more if desired, but with animal samples, we run into ethical and economical considerations (why use more animals in an experiment, whose goal has a strong likelihood of being achieved with fewer animals?).

An interesting consequence of Eq. 1.14 is that the effect size and the sample size are linked. If we fix the effect size, then we can calculate the sample size required for detecting it, as we have done in the previous paragraph. If we fix the number of samples, then the effect size adapts consequently. If we keep running our experiment with 24 samples, then we will be able to detect an effect of (see Eq. 1.14)

$$\Delta > 1.96 \frac{0.5}{\sqrt{24}} = 0.2^{\circ}\text{C}$$

As expected, with more than 16 samples, we will be more sensitive ( $0.2^{\circ}\text{C} < 0.25^{\circ}\text{C}$ ). However, this relationship is not linear, twice the number of samples does not imply a reduction the effect size to a half (the corresponding effect size for  $N = 32$  measurements is  $\Delta > 0.17^{\circ}\text{C}$ ). This nonlinear relationship comes from the square root that participates in the formula.

The example above has given us some intuition on how we may calculate the sample size for our experiment:

1. We need to know how the data will be analyzed: we will perform an hypothesis test in which the null hypothesis is of the form  $H_0 : \mu = \mu_0$ . We will assume that the samples are normally distributed, and we will reject the null hypothesis with a Type I error rate of  $\alpha$ .
2. We need to determine the effect size that we want to detect, that is, the minimum departure from the null hypothesis we want detect with the specified confidence  $(1 - \alpha)$ .

However, in this design we have not considered Type II errors (the thermostat is not working correctly, but with a small sample size, I fail to prove it). Let us assume that we want to have a statistical power of 80% in detecting an effect size of  $\Delta = 0.25^{\circ}\text{C}$ . That is in 80% of the experiments in which the departure from the reference mean,  $21^{\circ}\text{C}$ , is  $\Delta$ , we will correctly reject the null hypothesis in 80% of the cases ( $\beta = 0.2$ ). Fig. 1.11 shows this situation. The distribution of the sample mean under the null hypothesis is still represented in blue and it is centered around  $21^{\circ}\text{C}$ . Before performing the experiment we cannot know whether the thermostat is malfunctioning due to an excessively low or high temperature, and we will have to do the sample size calculation for both cases.

- Excessively low temperature. Let us assume that the thermostat is actually making the temperature to be lower than the reference. As our effect size is  $0.25^{\circ}$ ,

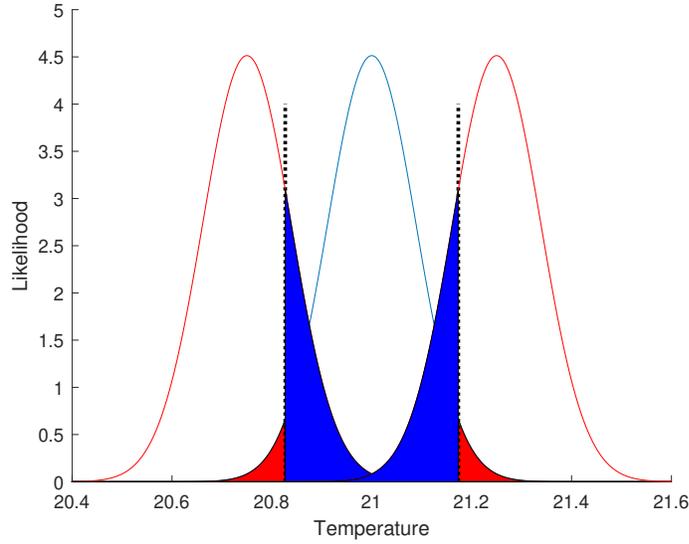


Figure 1.11: The red shaded area is the probability of observing a sample mean as far from 21°C as 0.25°C or further if the null hypothesis is true. Departures can be above or below 21° resulting in two possible distributions (see text).

the distribution of the sample mean under the alternative hypothesis is centered around  $\mu_1 = 20.75^\circ\text{C}$ . The variance is still given by  $\sigma^2/N$  because it only depends on the variance of the samples and the number of samples. When the experiment is carried out, we will reject the null hypothesis if the observed sample mean is outside a given region. Then, we must set the number of samples, such that the probability of not rejecting the null hypothesis when the thermostat is causing a lower temperature is  $\beta$ . That is, the blue area in Fig. 1.11 coming from the left Gaussian must be  $\beta$  ( $=0.2$  in our example). Summarizing, we must find a number of samples such that the red area on the left is  $\alpha/2$  ( $=0.025$ ) and the blue area is  $\beta$  ( $=0.2$ ). At the left rejection border, if the null hypothesis is true, we must have:

$$\Pr \left\{ \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} < z_{\frac{\alpha}{2}} \right\} = \frac{\alpha}{2}$$

For  $\alpha = 0.05$ ,  $z_{\frac{\alpha}{2}} = -1.96$  (note that it is a negative value). The critical value at which we will reject the null hypothesis is

$$\hat{\mu}_{crit} = \mu_0 + \frac{\sigma}{\sqrt{N}} z_{\frac{\alpha}{2}}$$

If the alternative hypothesis is true, this critical value has a normalized position given by

$$z_{crit}^a = \frac{\mu_0 + \frac{\sigma}{\sqrt{N}} z_{\frac{\alpha}{2}} - \mu_1}{\frac{\sigma}{\sqrt{N}}} = \frac{\Delta}{\frac{\sigma}{\sqrt{N}}} + z_{\frac{\alpha}{2}}$$

The Type II error probability is given by the probability under the alternative hypothesis of  $z$  being larger than  $z_{crit}^a$ . If we want this probability being  $\beta$ , we must have

$$\Pr\{z > z_{crit}^a\} = \beta = \Pr\{z^a > z_{1-\beta}\}$$

or what is the same

$$\begin{aligned} z_{crit}^a &= z_{1-\beta} \\ \frac{\Delta}{\sigma} + z_{\frac{\alpha}{2}} &= z_{1-\beta} \end{aligned}$$

From the latter equation, we deduce that

$$N = \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta/\sigma} \right)^2 \quad (1.16)$$

where we have made use of the fact  $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$ .

- **Excessively high temperature.** Let us now assume that the thermostat is making the temperature to be higher than the reference. Now the distribution of the sample mean under the alternative hypothesis is centered around  $\mu_1 = 21.25^\circ\text{C}$ . We may now make the reasoning as we did in the previous case, we must find a number of samples such that the blue area of the right Gaussian is  $\beta$  and the red area on the right is  $\alpha/2$ . At the right rejection border, if the null hypothesis is true, we must have:

$$\Pr\left\{ \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}} > z_{1-\frac{\alpha}{2}} \right\} = \frac{\alpha}{2}$$

For  $\alpha = 0.05$ ,  $z_{1-\frac{\alpha}{2}} = 1.96$ . The critical value at which we will reject the null hypothesis is

$$\hat{\mu}_{crit} = \mu_0 + \frac{\sigma}{\sqrt{N}} z_{1-\frac{\alpha}{2}}$$

If the alternative hypothesis is true, this critical value has a normalized position given by

$$z_{crit}^a = \frac{\mu_0 + \frac{\sigma}{\sqrt{N}} z_{1-\frac{\alpha}{2}} - \mu_1}{\frac{\sigma}{\sqrt{N}}} = \frac{\Delta}{\sigma} + z_{1-\frac{\alpha}{2}}$$

The Type II error probability is given by the probability under the alternative hypothesis of  $z$  being smaller than  $z_{crit}^a$ . If we want this probability being  $\beta$ , we must have

$$\Pr\{z < z_{crit}^a\} = \beta = \Pr\{z^a < z_\beta\}$$

or what is the same

$$\begin{aligned} z_{crit}^a &= z_\beta \\ \frac{\Delta}{\sigma} + z_{1-\frac{\alpha}{2}} &= z_\beta \end{aligned}$$

From the latter equation, we deduce that

$$N = \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta/\sigma} \right)^2 \quad (1.17)$$

where we have made use of the fact  $z_{1-\beta} = -z_{\beta}$ . Because of the symmetry of the distributions involved, solving for  $N$  in this case also results in the same sample size calculated in Eq. 1.16.

For the specifications of the thermostat ( $\Delta = 0.25$ ,  $\alpha = 0.05$  and  $\beta = 0.2$ ) we have

$$N = \left( \frac{1.96 + 0.84}{0.25/0.5} \right)^2 = 31.40$$

That is, we need at least 32 samples to detect a departure of  $0.25^{\circ}\text{C}$  from the reference temperature with a statistical confidence of 95% and a statistical power of 80%. In this sense, we realize now that with 24 samples, we were having a much smaller statistical power (69%) to detect deviations of  $0.25^{\circ}\text{C}$ .

### Sample size lessons

The main formula for the sample size calculation in the example above was

$$N = \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta/\sigma} \right)^2$$

This formula already shows the ideas exposed in Sec. 1.5:

#### Important remarks

18. The sample size ( $N$ ), effect size ( $\Delta$ ), statistical confidence ( $1 - \alpha$ ) and statistical power ( $1 - \beta$ ) are linked by a single formula. Fixing three of them automatically fixes the fourth one.
19. More important than the effect size in itself,  $\Delta$ , is the relationship ( $\Delta/\sigma$ ) between the effect size and the variance of the observations,  $\sigma$ . We may regard this ratio as a target Signal-to-Noise Ratio (SNR), and it is called the *normalized effect size*.
20. Increasing the statistical confidence or power results in a larger number of samples, since  $z_{1-\frac{\alpha}{2}}$  and  $z_{1-\beta}$  increase.
21. Smaller normalized effect sizes result in larger number of samples, since we want to be more sensitive.
22. Let us summarize the procedure followed to find the sample size:
  - (a) We have constructed a statistic,  $z$ , whose distribution is known under the null hypothesis.
  - (b) We have found a critical value of this statistic beyond which we will reject the null hypothesis. This critical value fulfills

$$\Pr\{z > z_{1-\frac{\alpha}{2}}\} = \frac{\alpha}{2}$$

- (c) We have translated this critical value of the statistic into a critical value of our observation,  $\hat{\mu}_{crit}$ .
- (d) Then, we have calculated the Type II errors associated to this value if the alternative hypothesis were true.

$$\Pr\{z^a < z_{crit}^a\} = \beta$$

The design equations can be summarized as finding the minimum  $N$  for which

$$\Pr\left\{z < \frac{\Delta}{\frac{\sigma}{\sqrt{N}}} + z_{1-\frac{\alpha}{2}}\right\} < \beta \quad (1.18)$$

In this case, this procedure has resulted in a closed form formula for the  $N$ . However, this is not the situation, in general. Instead, we can progressively increase the sample size until the criteria of statistical confidence ( $1 - \alpha$ ) and power ( $1 - \beta$ ) are satisfied.

The sample size is tightly connected to the data analysis procedure, in particular the hypothesis test, that we will perform once the experiment is finished. The specific hypothesis test implies a statistic, in our example the  $z$  statistic, whose distribution under the null and alternative hypotheses must be known. This knowledge is the one that allows relating the sample size to the statistical confidence and power, resulting in a useful equation that can be used to calculate the sample size.

#### Important remarks

23. Each hypothesis test implies its own sample size formula. There is no “universal” sample size formula valid for all experiments and situations. Also, we must pay careful attention to the assumptions of the hypothesis test (distribution of the observations, known parameters, the specific null and alternative hypotheses, ...).

Non-parametric tests are often used if the experimental data does not fulfill the distributional assumptions of parametric tests. Unfortunately, except for some few cases, there is no easy relationship between non-parametric hypothesis tests and the sample size. A perfect solution would be simulating the experiment many times and adjusting the number of samples to the required confidence level and statistical power. However, these simulations are normally out of the reach of many researchers. The common alternative is to design the sample size as if we were going to perform a parametric test, and then correct by some “safety” factor that increases the sample size accounting for the fact that our uncertainty is larger since we do not know the statistical distribution of the observations. In this way, the sample size is calculated as

$$N_{non-parametric} = \frac{N_{parametric}}{ARE} \quad (1.19)$$

where ARE is the Asymptotic Relative Efficiency. The following table shows the most common non-parametric tests along with their parametric counterparts and ARE:

Non-parametric	Purpose	Parametric	ARE
Mann-Whitney U test	Compare 2 independent samples	Student's t test	$3/\pi = 0.955$
Wilcoxon signed-rank test	Compare 2 dependent samples	Paired Student's t test	$3/\pi = 0.955$
Spearman correlation test	Correlation between 2 variables	Pearson's correlation test	0.91
Kruskal-Wallis ANOVA	Compare 3 or more groups	1-way ANOVA	0.864
If not in this table			0.85

There are a number of situations in which the sample size calculation fails, in particular:

**Important remarks**

24. If we assume an incorrect variance of the observations. This is a very common error and we tend to be optimistic about the variability of our experiments.
25. If we violate the assumptions of the hypothesis test, especially the distribution of the observations.
26. If we misunderstand the questions performed by the sample size calculation software. It is advisable, if possible, to use two different software or verify with some easy-to-calculate approximate formula.

The sample size calculation is performed at a stage of research in which we have not yet performed the experiment. Consequently, there is a large amount of uncertainty at this point, and the sample size calculation only gives approximate suggestions of sensible sample sizes (if the sample size calculation suggests 32 samples, we know that we cannot accomplish our goals with 10 and that we do not need as many as 100; however, we do not have precision at this point, because we only have a guess of the variability of the experiment, to determine if we need 30 or 35 samples).

Sometimes, researchers are pushed to achieve too much with limited resources. For instance, a researcher is interested in the effect of a new treatment compared to a control group. He/she will study the effect at five time points. There are a total of 20 animals. That leaves two animals per time point and treatment. However, two is typically a very low number (as we will see in the next chapter, being low or high depends on the variability of the measurements) for any useful comparison (although a full factorial experiment design may help a bit in this regard). It might be better to concentrate on fewer time points, so that the number of animals per time point and treatment is increased.

Next chapter shows the calculation methods, assumptions and consequences for the most common experimental situations encountered in animal research. It is meant to

be a reference chapter, so that we only look up the case in which we are interested at a particular moment. In a first pass over the book, the reader may go over the examples and important remarks to get an idea of the kind of problems he/she may encounter and for which there is already a good statistical solution.

## Chapter 2

# Real experimental examples

### 2.1 Some fully developed examples

#### 2.1.1 Difference between two group means

One of the most common situations is that in which we want to compare a the mean of a continuous variable in a control group and a treatment group. We will start with its most simple version, and we will progressively complicate it.

##### Basic case: showing a difference in independent two-group comparisons

- Example 20: We are interested in checking if a given eye drop we are developing has any effect on the intraocular pressure of rats. The intraocular pressure of these animals ranges is about 25 mmHg and, from previous experiments, we know that the standard deviation of the measures of a tonometer (a device to measure the intraocular pressure) is about 1.6 mmHg while the standard deviation of the intraocular pressure among animals would be around 2.1 mmHg (Pease et al, 2006). We want to determine if our eye drop compared to the vehicle alone has a systematic difference larger than 5%, that is differences larger than 1.25 mmHg ( $=0.05 \cdot 25$ ). How many animals do we need to carry out this experiment if we plan to measure two groups of similar animals: one group with the vehicle and the other group with the our eye drop? For analyzing the results we will use a two-sample Student's t-test assuming that two groups of measurements are independent. We want to have a statistical power of 90% and a confidence level of 95%.

Solution: As we explained in Sec. 1.4.5, the variance of our observations is the variance due to the biological variation between animals (whose standard deviation is  $\sigma_x = 2$  mmHg) and the variance or noise of our measurements (whose standard deviation is  $\sigma_n = 1.6$  mmHg). In this way, the standard deviation of the observations,  $\sigma_y$ , is predicted to be around

$$\sigma_y = \sqrt{2.1^2 + 1.6^2} \approx 2.6 \text{ mmHg}$$

At this point, we may use the formulas in Sec. 4.1.5 to obtain that we need  $N = 95$  animals per group.

### Paired samples case: showing differences in a one-group comparison

- **Example 21:** In the previous experiment, we realize that can avoid the inter-individual variability by measuring each animal with the two drops (one in the left eye and another one in the right eye). In this way, there is less variability between measurement. Performing the experiment in this way turns the measurements dependent by pairs (the pair of measurements coming from the same animal are dependent on each other). We will analyze the data using a paired sample Student's t-test, and we want to have the same statistical power and confidence level as above.

**Solution:** As discussed in Sec. 4.1.4, our true, independent observations are the difference between the measurements of the two kinds of drops. The predicted standard deviation of this difference is now

$$\sigma_{\Delta y} = 1.6\sqrt{2} \approx 2.3 \text{ mmHg}$$

(where we have assumed that the intraocular pressure of the left eye is the same as in the right eye; if this is not the case we could still calculate the standard deviation of the difference, but that would unnecessarily complicate this example). In this case, the sample size required for the comparison is only  $N = 37$  animals. Note that each animal will be measured twice (either the left or right eye) and the eye drops should be randomized (the left eye should randomly get the eye drop with the drug or the vehicle, and the other eye, the other treatment).

### Showing equivalence

- **Example 22:** As a researcher we are interested in showing that our drug significantly differs from the control (their differences are larger than 5%). However, as manufacturers of the tonometers we want to show that our devices are consistent among different factories: that is, the measurements from a device manufactured in Factory 1 are equal to the measurements from another device manufactured in Factory 2 (they differ in less than 5%). As discussed in Sec. 1.5, it is not the same performing a significance test (the equality between means is in the null hypothesis) than an equivalence test (the equality between means is in the alternative hypothesis). In the case of equivalence, we need to specify the limits within which we still consider the two measurements to have the same mean. Let us assume that are willing to tolerate a difference of at most 5% of the underlying true intraocular pressure, that is, 1.25 mmHg. Once we perform the measurements we will analyze the data using an equivalence test for the difference of paired samples (Sec. 4.1.8, the referred section is for the comparison of two independent means; however, the ideas there are similar to the ones of a single independent mean, although the specific calculations are a bit different).

**Solution:** For an equivalence test, we need to specify the lower and upper limit that we will still consider to be equal to each other ( $\epsilon_L$  and  $\epsilon_U$  in the notation

of Sec. 4.1.8). In this case, we will set them to be  $\pm 0.625$ , that is,  $1.25/2$ . Then, we will need  $N = 148$  animals for this experiment and on each animal two measurements will be performed: one with each tonometer. In the case of a manufacturer we may not need animals, but we can use a calibrated object mimicking the eye of an animal.

#### Important remarks

27. The effect size and the standard deviation used for the calculations must have the same units as the observations that will participate in the comparison between groups. If we are measuring the level of vasopresin in blood, the effect size must be the difference in vasopresin levels that we want to detect, if it exists, between the two groups. The standard deviation must be the variability of vasopresin in blood that we expect in our observations. It does not make sense to have effect sizes in other units (number of animals, proportion of animals showing a response, etc.)

#### Accounting for differences between researchers

- Example 23: Using the tonometer is not trivial and we suspect that there might be differences among researchers. To make sure that the differences are not biased we will take three protections:
  - Blocking: 3 different researchers will perform the same experiment.
  - Blinding: Each researcher does not know if he is measuring the values from the control or the eye drop with the drug. The two solutions will be labelled as A and B, and the researcher does not know if A is the eye drop with the drug or not.
  - Randomization: The sequence of measurements should be randomized between controls and treatments (BA, BA, AB, AB, AB, BA, AB, ...). Note that the randomization should be performed by a computer.

Should we use  $N = 39$  animals per researcher?

Solution: As discussed in Sec. 5.1.3, each researcher acts as a block. As there are 3 researchers, we will need 2 degrees of freedom to estimate their contribution. That is, we simply need to add 2 extra samples to estimate if there are biases caused by the researchers and compensate for them. To keep the number of samples a multiple of 3 (because there are 3 researchers) we will measure 42 animals. Each researcher will measure 14 of them. For each animal, we will still measure the difference between treatment and control, producing a single measurement,  $\Delta y_{ij}$  where  $i = 1, 2, 3$  represents the researcher, and  $j = 1, 2, \dots, 14$  the animal measured by that researcher. Our analysis formula should be (see Sec. 5)

$$\Delta y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

The test we want to perform to show if our treatment is changing the intraocular pressure is

$$H_0 : \mu = 0$$

As we calculated above, we need 39 animals (actually, degrees of freedom) to make this test and have a statistical power of 90% and a confidence level of 95% to detect a change of 1.25 mmHg when the standard deviations between animals is about 1.6 mmHg.

#### Important remarks

28. We need to know how the data will be analyzed (two independent samples, two dependent samples, ...; significance, equivalence or one-tail test; which test will be applied) for calculating the sample size. At the same time, we can take protections against bias by blocking, blinding, and randomization.

The structure of a design may look sometimes complicated. However, a careful look on the analysis formulas may reveal a much simpler internal structure. This is the case of the following example.

- Example 24: A researcher is interested on the effect of two different kinds of livestock (intensive or extensive) on the quality of meat. We will focus on the amount of retained water of the meat and relate it to the stress levels experienced by the animals before being transferred to the slaughterhouse and just after death. We will measure stress by the level of cortisol in blood. We will relate the amount of retained water ( $y$ ) to the levels of cortisol before and after death ( $c^B$  and  $c^A$ ) through a linear regression. In this way, the amount of retained water of the  $i$ -animal could be modelled as

$$y_i = \beta_0 + \beta_B c_i^B + \beta_A c_i^A + \varepsilon_i^Y$$

On their turn, the levels of cortisol in blood are supposed to depend on the livestock style through a 1-way ANOVA:

$$\begin{aligned} c_i^B &= \mu_B + \alpha_i^B + \varepsilon_i^B \\ c_i^A &= \mu_A + \alpha_i^A + \varepsilon_i^A \end{aligned}$$

We can combine all this information into a single observation equation

$$\begin{aligned} y_i &= \beta_0 + \beta_B(\mu_B + \alpha_i^B + \varepsilon_i^B) + \beta_A(\mu_A + \alpha_i^A + \varepsilon_i^A) + \varepsilon_i^Y \\ &= (\beta_0 + \beta_B\mu_B + \beta_A\mu_A) + && \text{Constant} \\ &\quad (\beta_B\alpha_i^B + \beta_A\alpha_i^A) + && \text{Treatment} \\ &\quad (\beta_B\varepsilon_i^B + \beta_A\varepsilon_i^A + \varepsilon_i^Y) && \text{Noise} \end{aligned}$$

This is the structure of a 1-way ANOVA with two levels, or even simpler, the difference between two groups. Then, we may use the standard Student's t-test to check whether intensive livestock retain more or less water than extensive livestock. The sample size calculation could follow a very simple approach.

### 2.1.2 Showing differences among different groups, ANOVA

- **Example 25:** Some researchers are interested in the effect of a treatment that is injected to each animal over time in a specific cardiac tissue and over time. It is suspected that there is a gene that is particularly sensitive to that treatment. For that reason, mice with three different genotypes of that specific gene will be studied (one of them is the wild-type, the second is a knock-out in the gene of interest, and the third is a knock-out in a promoter of the gene of interest). To check that is a specific compound of the treatment and not the vehicle, we will have two treatments: one with the active compound and another one with just the vehicle. Animals will be sacrificed 1h, 6h, 12h, 24h, 48h, and 72h after applying the treatment. The mean fluorescence intensity of cardiac tissue will be measured at those time points. The fluorescence of the protein of interest will be normalized with respect to the fluorescence of a reference gene. From previous experiments, we know that the normalized fluorescence has a standard deviation about 0.16. We want to detect differences in the normalized fluorescence larger than 1 with a statistical power of 90% and a confidence level of 95%. How many animals do we need for this experiment?

The experiment will be repeated 3 times to confirm its findings.

**Solution:** In a naive design, we would calculate the sample size based on a two-independent groups Student's t-test with the rationale that at some moment we will compare the fluorescence of the wild-type and the fluorescence of the other two genotypes at some particular time point. With the data above and the formulas in 4.1.5 for a two-tails test, we would come out with  $N = 3$ . We have 2 treatment levels (vehicle and compounds), 3 genotypes, and 6 time points. Then, the total number of animals needed for the experiment would seem to be  $N = 2 \cdot 3 \cdot 6 \cdot 3 = 108$ . Even more,  $N = 3 \cdot 108 = 324$  if we count the three repetitions of the experiment.

However, it is not expected that the two knock-outs express more protein than the wild-type. That is, the gene expression can only decrease or stay at the same level as the wild-type. Then, we should use a one-tail test, which decreases the sample size of a single comparison to  $N = 2$ , and for the total experiment to  $N = 2 \cdot 3 \cdot 6 \cdot 2 = 72$ . With this naive design, we see that depending on sensible choices (one-tail vs. two-tails) we can cut down the sample size by 33%.

However, a 3-way ANOVA design could have been used. Aside from the main effects of treatment, genotype, and time, we foresee that there can be pairwise interactions (the treatment or vehicle may induce different time responses; each genotype may have a different time response; and each genotype may respond differently to the treatment). However, we are not interested in triple interactions. In this way, our analysis model will be (see Chap. 5)

$$y_{ijkl} = \mu + \alpha_i^T + \alpha_j^G + \alpha_k^H + \alpha_{ij}^{TG} + \alpha_{ik}^{TH} + \alpha_{jk}^{GH} + \epsilon_{ijkl}$$

where  $\alpha_i^T$  corresponds to the main effect of control or treatment,  $\alpha_j^G$  is the main effect of the genotype,  $\alpha_k^H$  is the main effect over time, then pairwise interactions

follow. The subscript  $l$  refers to the  $l$ -th animal within the  $ijk$  combination. To estimate all these parameters we need the following number of degrees of freedom: 1 ( $\alpha_i^T$ ), 2 ( $\alpha_j^G$ ), 5 ( $\alpha_k^H$ ), 2 ( $\alpha_{ij}^{TG}$ ), 5 ( $\alpha_{ik}^{TH}$ ), and 10 ( $\alpha_{jk}^{GH}$ ). A total of 20 degrees of freedom, and there are 36 combinations. That means that if we use  $N = 1$  animal per combination, we still have 15 degrees of freedom ( $15 = 35 - 20$ ) for the residuals, and we would have a high statistical power to detect the effect size of 1 in a standard deviation around 0.16. That is, we could cut down the sample size by 67%, from 108 to 36 animals.

Many researchers feel unease by using a single animal per combination. They are afraid that:

1. The animal of a particular combination may become unusable for whatever reason losing the possibility to analyze the whole data.
2. There can be some unnoticed problem with a specific animal that corrupts its measurement.
3. Also, they like the idea of repeating the experiment three times to make sure that there is no unnoticed bias in the first execution of the experiment.

Point 1 is handled by the analysis of incomplete designs, as shown in Sec. 5.1.7. Point 2 will slightly increase the variability of the observations, although the analysis of linear models is relatively robust to small violations of the assumption that all combinations have the same variance. With respect to Point 3, as discussed in Sec. 2.2.2 it is better to divide the experiment in 3 mini-experiments, and include the repetition as part of the analysis. Now the analysis model is

$$y_{bijkl} = \mu + \alpha_b^B + \alpha_i^T + \alpha_j^G + \alpha_k^H + \alpha_{ij}^{TG} + \alpha_{ik}^{TH} + \alpha_{jk}^{GH} + \varepsilon_{ijkl}$$

where  $b$  is the repetition (block) of the experiment. Getting the effect of the block only costs 2 degrees of freedom (there are 3 repetitions), and we have 85 degrees of freedom for the residuals ( $=107 - 20 - 2$ ). This gives us an extreme sensibility to detect fluorescence changes much smaller than 1, and the security that what we are reporting is not the result seen once in our life.

A recurrent question is if it is better to perform a 3-way ANOVA or a whole series of Student's t-tests between all combinations of interest (e.g., at  $t=1h$ , wild-type treated vs knock-out treated; at  $t=24h$ , ...). The ANOVA analysis allows all these pairwise comparisons, the so-called, post-hoc analysis with an additional protection against the Type-I error inflation. Additionally, the ANOVA comparisons estimate the noise variance from the 108 observations, while the pairwise Student's t-test of each mini-experiment only had, in the original design, 3 observations in each of the groups. This makes ANOVA to have much higher statistical power.

The statistical mantra: As we saw in Chap. 1, the statistical mantra would be *control what you can, block what you cannot, and randomize the rest*. We also saw that blinding was important to get unbiased results. How to apply these principles in this experiment:

- Blinding: The researchers injecting the treatment, sacrificing the animals, and measuring the fluorescence should not know the treatment that animal received and its genotype.
- Control what you can: we have controlled the treatment, genotype, and time.
- Block what you cannot: we use the mini-experiments as blocks.
- Randomize the rest: We can randomize at two levels:
  1. Between mini-experiment: If possible, we should perform the mini-experiments in different centers, different seasons, different researchers, different equipment, ... If multiple centers are not possible, the rest are much easier to achieve.
  2. Within each mini-experiment: we should randomize the order in which we inject (for instance: 1) WT+vehicle, 2) KO+treatment, 3) WT+treatment, 4) PromoterKO+vehicle, ...) and at each time point of analysis we should also randomize the order in which we extract and analyze the cardiac tissue. We should also randomize the person applying the treatments, performing the surgery, and analyzing the data, the laboratory material used in the experiment, and even the time of application of the treatment. This latter may be difficult depending on the sampling times. For instance, in our sampling plan there is a sample after 12h, this sampling point may be difficult to combine with the normal working hours. Still, as discussed in Sec. 2.3.3, time could be treated as a continuous variable instead of a discrete one, and as such an arbitrary sampling point reveals more information about the time behaviour of the fluorescence.
- Example 26: In the previous experiment, the researchers are also interested the proportion of cells differentiated into a particular cell-type at each time point. From previous experiments, the expected proportion of differentiated cells is expected to range from 0 (in the KO) to 5% (in the wild-type), with a standard deviation in the wild-type about 2%. Can we apply the same 3-way ANOVA methodology as we did in the previous example?

Solution: Unfortunately, we cannot. There are two reasons for that. The first reason is that proportions do not follow a Gaussian distribution in general. Still, if the proportion is away from the 0 or 100% extremes, then the distribution of the observed proportions can be approximated by a Gaussian and we would be allowed to use the 3-way ANOVA approach. The second reason is that in our particular experiment, the expected proportion is very close to the 0% extreme. If we use the 3-way ANOVA, it will assume that the 2% standard deviation applies as a Gaussian to all combinations of factors. As such, the observations are expected to be between  $-3\sigma$  and  $3\sigma$ , this would imply negative values both for the wild-type and KO. When dealing with proportions, we should use the statistical tools specifically designed to deal with proportions. These are exemplified in the next example.

### 2.1.3 Comparing proportions

We have a proportion problem when we count how many individuals or events with a given property appear in a total of individuals. For instance: how many animals get infected if  $N$  animals are exposed to a pathogen dose; how many cells are differentiated into a given cell type of a total of  $N$  cells observed; how many viable cells there are in a given microscope field; ... All these experiments can be described with a binomial distribution whose parameters are the total number of observations (animals, cells, ...) and the proportion of those observations having a given feature ( $p$ , that is between 0 and 1). In a given experiment, we will observe  $x$  individuals with the feature of interest out of the  $N$  we have observed. The raw data for our observations will be the pair  $(N, x)$ . From this raw observation we may estimate the proportion of individuals in the population having the feature

$$\hat{p} = \frac{x}{N}$$

As we discussed in the last paragraph of the previous section, sometimes the binomial distribution with parameters  $N$  and  $p$  can be approximated by a Gaussian with a mean  $Np$  and variance  $Np(1-p)$  (this is allowed if  $Np > 5$  and  $N(1-p) > 5$ ). Then, we have all the set of statistical tools for the Gaussian at our disposal (Student's t-test, ANOVA analyses, Snedecor's F, ...). We could proceed as if we were comparing means. For instance, let's say that we are interested in the proportion of viable cells of a given type in cardiac tissue under conditions A and B. We would measure these proportions for  $N$  animals in each group having an estimate  $\hat{p}_{ij}$  for the  $j$ -th animal which was under the condition  $i$ . Then, we could use the standard ANOVA to analyze these estimates of the proportions (we would could calculate the mean of the estimates of the proportions, their standard deviation, etc.). If we cannot make this approximation by a Gaussian, then we cannot substitute our raw observations (pairs of the form  $(N_{ij}, x_{ij})$ ) for each animal) by the estimate of the proportion for that animal.

At the beginning of the previous example we compared the use of multiple two-independent groups Student's t-tests (which makes pairwise comparison between combinations of treatments, genotypes, and time points) with the use of a 3-way ANOVA that integrates all the information into a single model (which is much more powerful as it sees all the information at the same time). The equivalent to these two tools are multiple two-independent groups proportion comparisons (see Sec. 4.2.5) and the logistic regression (see Sec. 4.3.3). Undoubtedly, logistic regression is less known and more difficult to use and interpret and that is why many researchers prefer using the less powerful pairwise comparisons.

- Example 27: Following with the example of the previous section, let us say that we are interested in comparing the number of differentiated cells between the wild-type (WT) and the knockout (KO) genotypes, after 24h of applying the treatment. Remind that we have 1 animal for each of these combinations although we have repeated the mini-experiment 3 times. Therefore, we have the following raw observations:

$$\begin{array}{l|l} (N_{1,T,WT,24h,1}, x_{1,T,WT,24h,1}) & (N_{1,T,KO,24h,1}, x_{1,T,KO,24h,1}) \\ (N_{2,T,WT,24h,1}, x_{2,T,WT,24h,1}) & (N_{2,T,KO,24h,1}, x_{2,T,KO,24h,1}) \\ (N_{3,T,WT,24h,1}, x_{3,T,WT,24h,1}) & (N_{3,T,KO,24h,1}, x_{3,T,KO,24h,1}) \end{array}$$

We wonder which is the data with which we should run the pairwise comparisons and the logistic regression.

Solution:

The pairwise comparison between WT and KO for a fixed treatment and time should combine the information of all the mini-experiments, that is,

$$\left( \sum_{b=1}^3 N_{b,T,WT,24h,1}, \sum_{b=1}^3 x_{b,T,WT,24h,1} \right) \left| \left( \sum_{b=1}^3 N_{b,T,KO,24h,1}, \sum_{b=1}^3 x_{b,T,KO,24h,1} \right) \right.$$

The raw data for the logistic regression is a bit more involved, we must specify for each observed cell whether it was differentiated (1) or not (0) and from which treatment, genotype and time it was coming from. We should also use the data from all the mini-experiments. The raw data would look something like a very long table of which we only reproduce a few of its rows

Mini-experiment	Treatment	Genotype	Time	Differentiated
1	C	WT	1	1
1	C	WT	1	1
...				
1	C	WT	1	0
1	C	WT	1	0
...				
1	C	WT	2	1
...				
3	T	KO	72	0

The total number of rows in the table should be the total number of cells observed (differentiated or not) in the 3 mini-experiments for all combinations of treatment/vehicle, genotypes, and time points.

Finally, we wonder whether we could use these pairwise comparisons or logistic regression for the design of the number of animals needed for our experiment. It can certainly be done technically. However, experiment designs based on the number of cells needed to have a given statistical power and confidence level to detect differences in proportions of a given amount is not a good idea. Taken to the extreme, we can take as many cells from the animal as wanted. This would largely increase our statistical power. However, our conclusions only apply to the cells of that animal. We would like to generalize to a larger population of animals. Seen differently, our conclusions would be derived from a single animal (no matter how many cells we have observed from that animal and how small the p-value is in the comparison). That is why the number of animals should be designed based on some other property rather than the properties of the cells of those animals.

Still, there are experiments in which the main variable of concern is a genuine proportion and we would like to design the experiment according to that variable of interest.

- Example 28: A researcher is trying to humanize mice by making them to express a human receptor that is the key protein for a viral infection that affects humans, but it does not affect mice. If she succeeds, the genetically modified mice would be susceptible of being infected with the human virus. She will test the animals by injecting 3 times a sufficiently high viral load as to cause an infection if the animals are susceptible. She will inject the same viral load to a control group. If the genetic modification is successful, she would observe that no control animal is infected and that all genetically modified animals are. How many animals does she need to include in each group if we want to compare the proportion of infected animals in both groups and we want to have a statistical power of 90% and a confidence level of 95%?

Solution: This is an all-or-nothing response. The probabilities of infection in each one of the groups should be 1 or 0. The sample size of an experiment in which we are comparing proportions of two independent groups is described in Sec. 4.2.5. The formulas do not allow strict  $p_1 = 0$  and  $p_2 = 1$  values, because some of the calculations go to infinite. But we may use  $p_1 = 0.001$  and  $p_2 = 0.999$ . We obtain that we need a minimum of  $N = 4$  animals per group.

As always we should include blocking, randomization and blinding wherever possible in our experiment. In this example, the order in which animals are injected and measured should be randomized, the researcher should be blind to the genotype of the animal she is injecting, ideally, several researchers should inject the viral loads, not always using the same laboratory material (such as pipettes), etc.

- Example 29: We are interested on the effect of a gene on the incidence of lymphomas. These are induced with chemicals and it is known that 20% of the males develop a lymphoma after the treatment with the chemicals, while 45% of the females develop a lymphoma with the same chemical dose. We are interested on the protective effect of a gene on the development of these lymphomas and we will compare wild-type vs knock-out animals. How many animals do we need if we want to detect a change in the incidence of lymphomas of at least 15% (males would have an incidence of at least 35% and females of at least 60%).

Solution: The data from this problem can be analyzed through a logistic regression (Sec. 4.3.3). The logit of the probability of developing a lymphoma for the  $i$ -th animal is given by

$$\text{logit}(p_i) = \mu + \alpha_i^S + \alpha_i^G \quad (2.1)$$

where  $\alpha_i^S$  accounts for the effect of the sex of that animal and  $\alpha_i^G$  accounts for the effect of the genetic background (WT or KO). Although we can use the formulas in Sec. 4.3.3, these are thought for the case in which the regression has some continuous predictor variable. This is not the case here. Then, we can separate the problem into two much simpler problems. If we consider males and females

separately, then the two problems are detecting a change of 15% of incidence in two groups (Sec. 4.9.3). Using the formulas for this comparison between groups, we calculate that we need 162 KOs and 162 WT's to detect a difference of 15% in the males with a confidence level of 95% and a statistical power of 90%, and 202 female KOs and 202 female WT's.

Additionally, by separating the problem in two subproblems we will make a more efficient use of the animals, as we do not need the same number of animals in each one of the groups (female WT's, female KOs, male WT's, male KOs).

- **Example 30:** About 36% patients undergoing transcatheter aortic valve replacement may experience moderate or severe prosthesis–patient mismatch after surgery (there is a mismatch if the ratio between the effective valve area, measured by echocardiography, and the total body surface is below a given threshold). Some researchers are studying a new type of replacement hoping that they can reduce this rate below 10% with their new device (they want to detect this difference with a statistical power of 90% and a confidence level of 95%). Experiments are performed on pigs by various surgeons from two collaborating hospitals. Every week, one surgeon can operate 2 pigs on a given day. Half of the pigs will receive the current device while the other half will receive the new device. A secondary objective of the study is to check whether the replacement procedure time increases or decreases. The current procedure mean time is about 2h, with a standard deviation of 10 minutes.

Solution: We may design the experiment with an objective of detecting a difference in the mismatch rate, the procedure mean time, or both. If we want to do it with objectives in mind, we should calculate the sample size with respect to a difference in means (Sec. 4.1) and the sample size with respect to a difference in proportions (Sec. 4.2), and take the largest value. However, in this study, the procedure time is secondary. Consequently, we will design it only with the proportion difference as objective.

Using the formulas in Sec. 4.2.5 we get a sample size of  $N = 49$  per group, that we will increase to  $N = 50$  to be able to block the hospital (each hospital will perform half of the operations).

To protect ourselves against possible biases, we will put in place the following measures

- **Blocking:** The hospital will be blocked by making them operate half of the current and the new devices. Every day of the experiment at each hospital will be a block: with a surgeon and an expert measuring the valve size. Every day of experiment we will replace 1 standard and 1 new device. In this way, the skillfulness of the specific surgeon or the person measuring the effective valve area will not affect the comparison.
- **Blinding:** The surgeon will not know to the last moment which of the two implants he/she will be placing. In this way, the initial steps of the surgery will not be biased. The person measuring the effective valve size after surgery should not know the kind of implant he/she is measuring.

- Randomization: Everyday, the order of the replacements will be random (AB, BA, AB, AB, BA, AB, ...). The pairs of surgeon and measuring expert should also be randomized along the 25 weeks of the experiment.

### 2.1.4 Survival curves

- Example 31: One of the major risks of the aortic valve replacement is the occurrence of a heart failure within the next year after surgery. We think that the diet may have an effect on the reduction of heart failures within the period of study. It has been observed that after 1 year, 25% of the patients have suffered at least 1 heart failure event within that period. Pigs will be subject to an aortic valve replacement. Half of the pigs will be given Diet 1 while the other half will be given Diet 2. How many animals do we need per group if we want to detect a decrease by a factor 2 (e.g., from 25% to 12.5%) with a confidence level of 95% and a statistical power of 90%.

Solution: We could simply compare the proportion of animals having had a heart failure or more within 1 year after surgery in both groups. This would be a comparison of two independent proportions and using the same formulas as in the previous example we would come down to  $N = 198$  animals per group.

However, in this approach we are losing the time information. It is not the same 1) all animals staying healthy all over the year except for the last day in which 25% of them have a heart failure, as 2) about 2% of the animals having a heart failure every month amounting to about 25% at the end of the year. Time information is collected by the so-called, hazard rate, that can be intuitively understood as the instant probability of the event occurring at time  $t$  given that the event has not occurred up to that time. In the two situations described above: in the first situation the hazard is very low all the way to the end of the year when it sharply grows, in the second situation the hazard is constant along the year.

The log-rank test compares two samples whose hazard rates are proportional to each other. They do not need to be constant over time, but they must vary at the same speed so that their ratio is constant. In our case, we are interested in whether the hazard rate of one of the diets is at most a half of the hazard rate of the standard diet. Following the formulas in Sec. 4.7.7, we can reduce the sample size per group from  $N = 198$  to  $N = 134$ . The reason for this reduction is that we have much more information as we track the number of events over time and can compare them, instead of just counting the number of events at the end of the period.

The standard protections against bias should still be in-place. For instance, if all animals on Diet 1 are in Hospital 1 and all animals on Diet 2 are in Hospital 2, the difference could be caused by the surgeons that performed the valve replacement or the animal care of the different hospitals rather than by the diet. The hospital, surgeons, animal carers, etc. should be blocked. The person evaluating whether a heart failure has actually happened or not should be blind to the animal's diet. Otherwise, he/she may be more prone to declare a heart failure event in one case or another.

### 2.1.5 A factorial design

- **Example 32:** A researcher is interested in a novel cell therapy to treat tumours of different kinds. For this experiment, he will try with 2 different kinds of tumours. He will try 3 different kinds of cell treatments, and each one of these can be given with an adjuvant or not. He expects that the different kinds of tumours may respond differently to the treatments and to the adjuvants. He is not interested in sex or age differences although there might be. For each animal he will measure the volume of the tumour over time and will follow the animals up to 2 months after the tumour implantation. Another outcome of interest for him is the survival time during this period. Finally, we plan to repeat the experiment twice to confirm its findings.

How should we design the experiment and how many animals do we need per combination of treatments, tumour, and adjuvant?

**Solution:** In this experiment, we have 3 factors: type of tumour ( $T$  that takes 2 levels), cellular treatment ( $C$  that takes 3 levels), and the presence or absence of adjuvant ( $A$ , 2 levels). In total, we have  $12 = 2 \times 3 \times 2$  combinations of factors. We will consider sex (male/female) and age (young/old) as blocks ( $S$  and  $Y$  respectively, with two levels each). We have a total of 8 blocks. As discussed in Sec. 2.2.2, we should consider repetitions of the experiment as mini-experiments whose results will be analyzed as a whole. In this way, we have another block that is the mini-experiment,  $E$ .

Because the tumours may respond differently to the different treatments and adjuvants, we will include the interactions tumour-cell treatment ( $TC$ ) and tumour-adjuvant ( $TA$ ). Finally, we also foresee that the adjuvant is more useful in some of the therapies, we will include the interaction cell treatment-adjuvant ( $CA$ ). We will not consider triple interactions. The tumours will grow over time, but because there are two kinds of tumours they may grow differently. For that reason, we will include time in our analysis formula. Knowing that tumours grow exponentially, the correct way of modelling them is in the logarithmic space (Sec. 2.2.5). That is, our analysis equation would be (see Chap. 5)

$$\begin{aligned} \log(y_{syetcak}) &= \mu + \beta_{T_1} T_1 t + \beta_{T_2} T_2 t && \text{tumour growth over time} \\ &+ \alpha_s^S + \alpha_y^Y + \alpha_e^E && \text{block contributions} \\ &+ \alpha_t^T + \alpha_c^C + \alpha_a^A && \text{main effects of the factors} \\ &+ \alpha_{tc}^{TC} + \alpha_{ta}^{TA} + \alpha_{ca}^{CA} && \text{pairwise interactions} \\ &+ \varepsilon_{syetcak} && \text{residuals} \end{aligned}$$

We have introduced the auxiliary variables  $T_1$  and  $T_2$  that take the values 1 or 0 depending on whether that animal receives the Tumour 1 or 2.

To detect changes in any of the main effects of the factors whose size is at least 1.5 times the standard deviation of the residuals with a statistical power above 95%, we only need 50 animals in total.

We may compare this number to the standard design proposed by many researchers. Remind that we have 8 blocks (including the repetition of the experiment) and 12 treatment combinations. Using 10 animals per group, we would

obtain 960 animals. That is, we have an extremely high statistical power, which may be seen as an overkilling and unethical due to the waste of animals and economical resources. The price to pay is that our design is now incomplete and imbalanced (see Sec. 5.1.7), but computers can solve for the different contributions of the different effects (our formula above).

From previous experience, we know that about 20% of the animals will not develop a tumour after injection of the tumour cells. To account for any other unforeseen event, we will also increase the number of animals by another 10% (see Sec. 2.2.6). That is we will use 72 animals. With this increase in the number of animals, if no animal drops out, we will be able to detect changes as small as 1.25 times the standard deviation of the residual with a statistical power above 95%. To choose the distribution of animals in groups we have chosen to optimize the  $D$ -criterion of the system matrix (see Sec. 5.1.7).

Our 72 animals are distributed as shown in Tables 2.1 and 2.2. The fact that the design is incomplete and imbalanced translates into the observation that not all blocks in the table have the same number of treatments. This complicates the analysis when done by hand, but not when done with a computer, and as we have seen above we have an excellent statistical power.

We have based our design on the measurement of the tumour volume. We could have done it based on the survival analysis. We can also relate the hazard of dying from the tumour at any time,  $\lambda(t)$  to the factors of our model. Analogously to the analysis formula we gave for the tumour volume, we would have now

$$\begin{aligned} \log(\lambda(t)) = & \mu + \beta_{t_1} T_1 t + \beta_{t_2} T_2 t && \text{tumour growth over time} \\ & + \alpha_s^S + \alpha_y^Y + \alpha_e^E && \text{block contributions} \\ & + \alpha_t^T + \alpha_c^C + \alpha_a^A && \text{main effects of the factors} \\ & + \alpha_{tc}^{TC} + \alpha_{ta}^{TA} + \alpha_{ca}^{CA} && \text{pairwise interactions} \end{aligned}$$

This is called the Cox regression model (Sec. 4.3.4) However, it is much more difficult to design the experiment based on this latter equation and we have preferred to address the problem through a simpler approach based on a standard linear model. Still, once the experiment has been performed we may fit the Cox model to identify the contribution of each one of the factors on the probability of dying from the tumour.

- **Example 33:** We are interested in the effect of cage enrichment on the activity of rats (cage size, bed material, toys, etc.). We consider two different kinds of environments and we think that the enrichment may have a different impact on males or females. We will measure the voluntary locomotor time within the cage after one week Klein et al (2022). We cannot mix males and females in the same cage and all animals within the same cage will receive the same treatment. For these reasons, the experimental unit will be the cage. The activity of all animals within the same cage will be measured and averaged, resulting in the measurement of the cage. We will employ 8 female and 8 male cages. Within each cage there will be 2 rats. 8 experimental units per group allow us to identify normalized effects of 1 with a confidence of 95% and a statistical power above

Female	Old	Tumour1	TreatmentA	NoAdjuvant
Female	Old	Tumour1	TreatmentA	Adjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	Adjuvant
Female	Old	Tumour1	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentA	Adjuvant
Female	Old	Tumour2	TreatmentB	NoAdjuvant
Female	Old	Tumour2	TreatmentB	Adjuvant
Female	Old	Tumour2	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentC	Adjuvant
Female	Young	Tumour1	TreatmentA	Adjuvant
Female	Young	Tumour1	TreatmentB	NoAdjuvant
Female	Young	Tumour1	TreatmentB	Adjuvant
Female	Young	Tumour1	TreatmentC	NoAdjuvant
Female	Young	Tumour2	TreatmentA	Adjuvant
Female	Young	Tumour2	TreatmentB	NoAdjuvant
Female	Young	Tumour2	TreatmentB	Adjuvant
Female	Young	Tumour2	TreatmentC	NoAdjuvant
Female	Young	Tumour2	TreatmentC	Adjuvant
Male	Old	Tumour1	TreatmentA	NoAdjuvant
Male	Old	Tumour1	TreatmentA	Adjuvant
Male	Old	Tumour1	TreatmentB	NoAdjuvant
Male	Old	Tumour1	TreatmentB	Adjuvant
Male	Old	Tumour1	TreatmentC	NoAdjuvant
Male	Old	Tumour1	TreatmentC	Adjuvant
Male	Old	Tumour2	TreatmentC	NoAdjuvant
Male	Old	Tumour2	TreatmentC	Adjuvant
Male	Young	Tumour1	TreatmentA	Adjuvant
Male	Young	Tumour1	TreatmentB	NoAdjuvant
Male	Young	Tumour1	TreatmentB	Adjuvant
Male	Young	Tumour2	TreatmentA	Adjuvant
Male	Young	Tumour2	TreatmentB	Adjuvant

Table 2.1: Experiment 1 of Example 32. The animals have been sorted by blocks to facilitate reading.

Female	Old	Tumour1	TreatmentA	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentA	NoAdjuvant
Female	Old	Tumour2	TreatmentB	NoAdjuvant
Female	Old	Tumour2	TreatmentC	NoAdjuvant
Female	Old	Tumour1	TreatmentA	Adjuvant
Female	Old	Tumour1	TreatmentB	Adjuvant
Female	Old	Tumour1	TreatmentC	Adjuvant
Female	Old	Tumour2	TreatmentA	Adjuvant
Female	Old	Tumour2	TreatmentB	Adjuvant
Female	Old	Tumour2	TreatmentC	Adjuvant
Female	Young	Tumour1	TreatmentB	NoAdjuvant
Female	Young	Tumour1	TreatmentA	Adjuvant
Female	Young	Tumour1	TreatmentB	Adjuvant
Female	Young	Tumour1	TreatmentC	Adjuvant
Female	Young	Tumour2	TreatmentA	NoAdjuvant
Female	Young	Tumour2	TreatmentB	NoAdjuvant
Female	Young	Tumour2	TreatmentB	Adjuvant
Female	Young	Tumour2	TreatmentC	Adjuvant
Male	Old	Tumour1	TreatmentA	NoAdjuvant
Male	Old	Tumour1	TreatmentA	Adjuvant
Male	Old	Tumour1	TreatmentB	NoAdjuvant
Male	Old	Tumour1	TreatmentB	Adjuvant
Male	Old	Tumour1	TreatmentC	NoAdjuvant
Male	Old	Tumour2	TreatmentA	NoAdjuvant
Male	Old	Tumour2	TreatmentB	Adjuvant
Male	Old	Tumour2	TreatmentC	Adjuvant
Male	Young	Tumour1	TreatmentA	Adjuvant
Male	Young	Tumour1	TreatmentB	NoAdjuvant
Male	Young	Tumour1	TreatmentB	Adjuvant
Male	Young	Tumour1	TreatmentC	NoAdjuvant
Male	Young	Tumour1	TreatmentC	Adjuvant
Male	Young	Tumour2	TreatmentA	NoAdjuvant
Male	Young	Tumour2	TreatmentA	Adjuvant
Male	Young	Tumour2	TreatmentB	NoAdjuvant
Male	Young	Tumour2	TreatmentB	Adjuvant
Male	Young	Tumour2	TreatmentC	Adjuvant

Table 2.2: Experiment 2 of Example 32. The animals have been sorted by blocks to facilitate reading.

90%, both for the sex and the treatment. That is, we will be able to detect changes in the main effects and the interaction between treatment and sex whose size is 1 ( $\Delta/\sigma_\varepsilon = 1$ , see Sec. 5.4). Note that  $\sigma_\varepsilon$  here is the standard deviation of the observations of the cage, not the animals. In this design we have deliberately not stated which this value will be. However, we know that 8 experimental units per treatment will allow us to detect changes whose size are twice the level of noise, whichever this is. We cannot expect to have a high sensitivity to detect changes whose size is half the level of noise ( $\Delta/\sigma_\varepsilon = 0.5$ ). The power for detecting these changes is only 45%.

The sample size above has been calculated assuming a 2-way ANOVA model:

$$y = \mu + \alpha_s^S + \alpha_t^T + \alpha_{st}^{ST}$$

We have 2 factors, sex and treatment, with 2 levels each. That means that we will need 3 degrees of freedom to determine their parameters and we will have 12 degrees of freedom available for the residuals (see Sec. 5.1.5).

- **Example 34:** We are interested in the effect of two different environment enrichment strategies on the stress suffered by animals in the cages. The two strategies are two different frequencies of cage bed cleaning and the presence or absence of a tube in which animals can enter and find food. We will measure the cortisol level in blood before and after applying the enrichment strategies. We will have four groups (frequency 1-no tube, frequency 1-tube, frequency 2-no tube, frequency 2-tube). The key idea of this example is that each one of the animals serve as its own control. Although we have two numbers for each animal, we only have a single piece of information, that is the difference between the two cortisol levels. Then, we are left only with two factors, 2-way ANOVA, and we may use the same design and sample size calculation as in the previous example.
- **Example 35:** Wild voles eat the roots of fruit trees causing important losses to agriculture. We want to know which are the voles' preferences. For doing so, we will capture  $N$  voles, and evaluate how much they eat from 10 different kinds of roots. Female and male voles may have different preferences. This is a two-way ANOVA with two factors: sex (2 levels) and roots (10 levels). If we make a standard sample size calculation to detect normalized effect sizes of 1 (that is, changes that are of the same size of the standard deviation of the residuals), then we get  $N = 40$ . There are 20 level combinations (2 times 10), meaning that we need 2 animals per combination. That is, 2 females and 2 males with root 1, 2 females and 2 males with root 2, ... Note that each animal is different, we cannot reuse the same 2 females and 2 males. At this point, the sample size calculation meets the experiment design. If we design our experiment as a two-way ANOVA, as done above, then each animal is offered a single root (the allocation of animal to root must be at random), and we measure how much the animal eats.

However, we could have organized our experiment differently. We could capture  $N$  wild voles, and offer each one of them the 10 roots. This would be a repeated measures design (see Sec. 5.2.6, the root acts as a within-factor and sex as a

between-factor). We can offer the roots one after the other, but very likely this will bias the results towards the roots that are offered first (afterwards the vole is full and it does not want to any anymore). In this design, the sequence of roots should be randomized. Alternatively, we could offer the 10 roots simultaneously, and let the vole choose from which to eat. This design would probably cause a lot of zeroes in the results, as the vole very likely will eat only from one or two roots. Then, our observations would violate the assumptions of ANOVA. Finally, we can offer the voles the roots one at a time, during 10 consecutive days. Always at the same time, to avoid differences along the day. Again, the sequence of roots should be randomized for each vole. In this repeated-measures design, we only need 7 females and 7 males to achieve the same statistical power and confidence level to detect changes of the same size as the standard deviation. We would have reduced the number of animals from 40 to 14.

### Approximate calculations of the sample size

Although there are formulas to calculate the sample size of a multiway ANOVA (see Sec. 5.1.5), these are not easy to apply and they can be further complicated if some of the combinations of factor levels do not exist (as is the case in repeated-measures or nested designs). In these cases, it is convenient to have approximate methods at hand. We extend our previous example to include the response over time of the enrichment.

- **Example 36:** We are interested in evaluating the time dependence of the enrichment treatment during the first week. For doing so, we will measure the voluntary activity time of all animals for seven consecutive days (we will need 6 degrees of freedom to determine the main effect of the day). Our research questions include: do males and females behave differently over time (interaction between sex and time, 6 degrees of freedom)? and do the treatments have a different time profile (interaction between treatment and time, 6 degrees of freedom)? We may expand our analysis above to a 3-way ANOVA:

$$y = \mu + \alpha_s^S + \alpha_t^T + \alpha_d^D + \alpha_{st}^{ST} + \alpha_{sd}^{SD} + \alpha_{td}^{TD} \quad (2.2)$$

This is a repeated measures design (Sec. 5.2.7), in which the time factor has 7 levels (we will measure all cages every day). The calculation of the sample size using a power analysis becomes much more complicated. In this case, we need  $N = 16$  cages, 8 female and 8 male cages. Within each sex, we will assign 4 cages to each of the two treatments.

A simplified approach would count the extra number of degrees of freedom following the spirit of the resource equation, Eq. 4.13. As discussed in Sec. 2.2.1, this equation assumes a normalized effect size around 1 for the main effects.

We will need 18 degrees of freedom for the main effects of time and its interactions with sex and treatment. We have to add the 3 degrees of freedom for sex, treatments and their interaction. Finally, we should keep between 10 to 20 degrees of freedom for the noise. Additionally, we want the number of cages,  $N$ ,

to be a multiple of 4 (2 sexes times 2 treatments), so that the design is balanced. Overall, the total number of degrees of freedom becomes:

$$N - 1 = 3 + 18 + E$$

Choosing  $N = 36$  we would have  $E = 14$  degrees of freedom for the noise. With a total of  $N = 36$  cages, we would use 18 female and 18 male cages. Within each sex group we would have 9 cages treated with Treatment 1 and 9 cages treated with Treatment 2. Once we start the experiment we will measure the voluntary activity time within each cage every day, average the times for the two rats within each cage and those averages will be the observation of the cage that will participate in the analysis Eq. 2.2.

Another approximate calculation for this kind of designs is provided by the following consideration on the main effects. For instance, we may compare the difference between treatments 1 and 2. In our analysis, this comparison will be part of the post-hoc ANOVA analysis. However, at the time of design, we need to choose a number of cages and the calculation formulas for the multiway ANOVA are complicated. We may, instead, calculate the sample size assuming that this comparison is performed with a Student's t test (although it is not the right tool, many groups do actually use this technique to perform this kind of comparisons). If we want to detect a normalized effect size of 1, then we will need 23 cages in each treatment (see Sec. 4.1.5). Let us round it up to the next multiple of 4 (2 sexes times 2 treatments), that is, 24 cages in Treatment 1 and 24 in Treatment 2. These 24 cages will be split into the 2 sexes, that is, 12 cages of each sex. Finally, our design would have a total of  $N = 48$  cages.

We see that the approximate calculation methods cannot exploit the extra information due to the different combinations of the factor levels. Although, they still give sample sizes within the correct order of magnitude:  $N = 16$  (correct),  $N = 36$  (resource equation),  $N = 48$  (Student's t-test), the exact calculations allow a significant reduction of the number of animals needed.

- Example 36: A researcher is developing a new treatment against a particular kind of tumors. She will compare her new treatment to the current one in male and female animals. The average weight of males is about 30g, while the average weight of females is 25g. The standard deviation of the weight is about 10% of their weight, that is, 3g and 2.5g for males and females. For simplicity, we will calculate the sample size taking the worse case,  $\sigma_\epsilon = 3$ . She is interested in variations of the animal weight of about 20% of their weight, that is, 6 and 5 g., respectively. Again, for simplicity we will use the most restrictive change, that is, 5 g. If we calculate the sample size for one-tail, two independent groups using a Student's t-test ( $\alpha = 0.05, \beta = 0.1$ ), then we would need 7 animals per group.

A naive design would use 7 females and 7 males with the current treatment, 7 females and 7 males with the new treatment. That makes a total of 28 animals. However, as discussed in Sec. 2.2.4, estimating the effect of sex only requires

an extra degree of freedom. So, we may increase the sample size for the comparison of the treatment by 1, actually 2 to make it a multiple of 2 (because of the 2 sexes). That is, we will use 4 females and 4 males with the current treatment, and 4 females and 4 males with the new treatment. When we compare the effect of the new treatment with respect to the old treatment, we will still have 8 animals per group (4 males and 4 females). That is, even a bit more power to identify changes of 5 g. when the standard deviation of the observations is 6 g. Increasing the number of animals to 5 females and 5 males, we may even estimate the separate effect of the treatment on males and females. This is one of the advantages of factorial designs: with relatively few animals we may recognize the main effects of each one of the factors, and their interactions. Although, this latter with a lower power for the same effect size.

### 2.1.6 Dose optimization

- **Example 37:** A researcher is interested in optimizing the dose of a combined therapy of antibiotic and phages to fight an infection of a multi-drug resistant bacteria. From previous experiments, she knows that the variability of the bacterial load in the lungs of mice is around 0.333 logarithmic units, and a severely infected animal has a bacterial load about  $10^8$  (Colony Forming Units, CFUs), that is, 8 logarithmic units. The doses of interest of the antibiotic and the phage are 1-1.5 [g/(kg.day)] for the antibiotic and 1-5· $10^9$  (Plaque Forming Units, PFUs) for the phage. Which are the combinations of antibiotic and phage doses that she should try to find the optimal combination?

**Solution:** This problem addresses estimating the response of a variable of interest (bacterial load) as a function of some continuous variables (the antibiotic and phage doses). What the researcher is looking for is a Response Surface Design (Sec. 5.2.11). As explained in Sec. 5.1.2 it is numerically more stable to work with the centered doses. That is, the center points of the dose ranges of interest are 1.25 [g/(kg.day)] and 3 [ $10^9$  PFU], so we will define the centered antibiotic and phage doses as

$$\begin{aligned}\tilde{D}_a &= D_a - 1.25 \\ \tilde{D}_p &= D_p - 3\end{aligned}$$

and construct a model of the form

$$\begin{aligned}y &= \mu && \text{0-th order approximation} \\ &+ \beta_a \tilde{D}_a + \beta_p \tilde{D}_p && \text{1st order approximation} \\ &+ \beta_{aa} \tilde{D}_a^2 + \beta_{ap} \tilde{D}_a \tilde{D}_p + \beta_{pp} \tilde{D}_p^2 && \text{2nd order approximation}\end{aligned} \quad (2.3)$$

There are several ways to design optimal sampling patterns (Sec. 5.2.11). The one used here is the Central composite face centered (CCF), and the optimal sampling points are the ones depicted in Fig. 2.1. Once the model is fit, we can find the maximum of the surface and that would be the optimal dose.

As always we should apply the principles of blocking, randomization and blinding. For instance, we should randomize the order of the sampling points and do

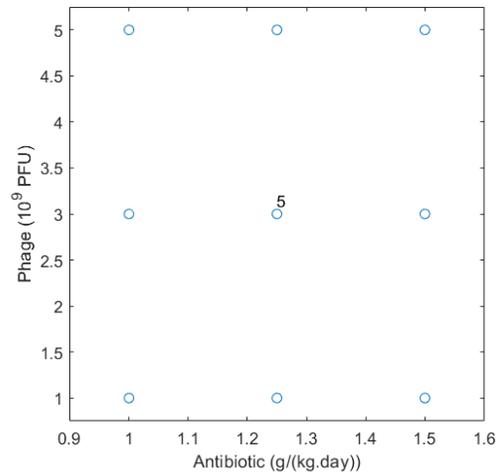


Figure 2.1: Optimal sampling points for fitting the model in Eq. 2.3. The middle point has to be sampled 5 times. The sampling in the middle allows a better estimation of the curvature of the response surface.

not follow a specific pattern (for instance, from low to high doses). We should also randomize the researcher applying the treatment, animals assigned to the treatments, the laboratory material. The researcher should be blind with respect to the treatment she is applying (for instance, a researcher could prepare the doses and another one could apply them).

It must be noted that in this problem in which we are interested in the optimum of a function, there is no comparison involved. That is why in the specification of the problem we have not given any statistical power or confidence level.

### 2.1.7 Diet optimization

- Example 38: A researcher is interested in finding the cow diet that causes a maximum production of eicosapentaenoic acid (EPA), docosahexaenoic acid (DHA), and conjugated linoleic acid (CLA) in milk. He can control the milk content of these acids by controlling the cow diet. Concisely, there are three components of the diet ( $A$ ,  $B$ , and  $C$ ), that makes the 100% of the diet content. There are some constraints,  $A < 5\%$ ,  $B < 10\%$ , and remaining will be filled with  $C$ . What are the optimal sampling points to find the optimal diet?

Solution: As in the previous example, we want to find an optimum of a response surface. We only have two independent variables, as once  $A$  and  $B$  are given, then  $C$  is automatically determined to fill the diet up to 100%. These problems where the addition of the three variables is fixed are called a mixture problem

(we want to determine the optimal composition of a mixture), and their designs are described in Sec. 5.2.12.

We will already use the centered  $A$  and  $B$  concentrations. We will have three models to fit:

$$\begin{aligned} y_{EPA} &= \mu_{EPA} + \beta_{EPA,a}\tilde{A} + \beta_{EPA,b}\tilde{B} + \beta_{EPA,aa}\tilde{A}^2 + \beta_{EPA,ab}\tilde{A}\tilde{B} + \beta_{EPA,bb}\tilde{B}^2 \\ y_{DHA} &= \mu_{DHA} + \beta_{DHA,a}\tilde{A} + \beta_{DHA,b}\tilde{B} + \beta_{DHA,aa}\tilde{A}^2 + \beta_{DHA,ab}\tilde{A}\tilde{B} + \beta_{DHA,bb}\tilde{B}^2 \\ y_{CLA} &= \mu_{CLA} + \beta_{CLA,a}\tilde{A} + \beta_{CLA,b}\tilde{B} + \beta_{CLA,aa}\tilde{A}^2 + \beta_{CLA,ab}\tilde{A}\tilde{B} + \beta_{CLA,bb}\tilde{B}^2 \end{aligned} \quad (2.4)$$

The sampling points to fit this model are shown in Fig. 2.2. Once the different models are fitted, we will find the combination of  $A$  and  $B$  that maximizes

$$y_{EPA} + y_{DHA} + y_{CLA}$$

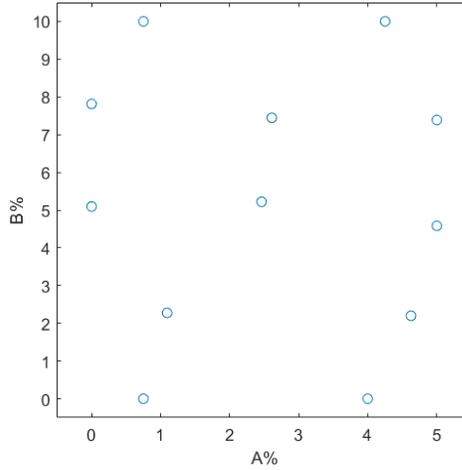


Figure 2.2: Optimal sampling points for fitting the model in Eq. 2.4.

## 2.2 Practical issues related to sample size calculation

In the following we comment on common mistakes or hot topics in the design of experiments using animals. We also provide some examples about how to design experiments in the most common situations.

### 2.2.1 An omnibus sample size

Many researchers carry out all their experiments with a fixed amount of animals per group. This is a typical situation when the experiment is carried out without any specific goal in mind and the researchers will measure: “*macroscopic, biochemical, genetic, immunohistochemical, cytometric, and permeability changes*”. On one side, it

is understandable that they are interested in all these changes. On the other side, it is remarkable that they are not particularly interested in any one of them so as to make an explicit design for that feature.

Another, more prosaic, reason for this fixed sample size is “we have always performed experiments of this size” or the more scientific, “based on previous experience, we will use this sample size”.

In the most common type of analysis, an independent two-samples, two-tails Student’s t-test, fixing the sample size amounts to fixing the relationship between statistical power and effect size (see Sec. 4.1.5). For instance, if we fix  $N = 8$  animals per group, this is equivalent to having the following statistical powers to detect the following normalized effect sizes ( $\Delta/s$ )

Normalized effect size ( $\Delta/s$ )	Statistical power ( $1 - \beta$ )
1.9	0.95
1.7	0.90
1.5	0.80
1.4	0.75
1.1	0.50
0.7	0.25

That is, we have a statistical power of 95% to detect changes whose size is 1.9 times the standard deviation of the observations in each one of the groups (assumed to be the same between both groups). We may still be able to detect smaller effects. For instance, as low as 0.7 times the observed standard deviation, but only in 25% of our experiments.

In many Omics experiments, it is typical to work with 3 animals per group or time point. Then, we will be able to detect the following normalized effect sizes with the following statistical power

Normalized effect size ( $\Delta/s$ )	Statistical power ( $1 - \beta$ )
4.0	0.95
3.6	0.90
3.1	0.80
2.9	0.75
2.1	0.50
1.4	0.25

Due to the high cost of these techniques, there is nothing wrong with working with only 3 animals (at least in terms of variance; for a discussion about the sample size in terms of bias and variance, see next section), only that we should be aware of our detection capacity (only big changes larger than 4 times the standard deviation). It is often argued that the high sensitivity of the omics techniques make working with 3 samples perfectly appropriate. High sensitivity means that the measurement noise is sufficiently low as to be able to detect small effect sizes. This being true, we must bear in mind that: 1) omics experiments may normally suffer from important batch effects (measurements from samples within a single batch may be shifted up or down with respect to measurements in another batch), this causes that measurements from

different centers (institutes, hospitals, etc.) normally differ quite importantly and must be normalized before being analyzed; 2) working with few animals makes us very dependent on the physiological state of the animals being analyzed (sex, age, diet, biorhythms, injuries, ...) turning extrapolation to a wider population of animals more difficult.

**Important remarks**

29. We lose control on the effect size that can be detected and the probability of detecting it by choosing always the same number of animals. Fixing the number of animals is equivalent to choosing an (unknown) statistical power to detect a given effect size.

A refined version of this fixed sample size for all experiments is Mead's resource equation, that applies to all experiments with linear models (Sec. 4.1.6). This time the number of degrees of freedom for the residuals is fixed between 10 and 20. This fixed number also implies an implicit choice of the effect size with respect to the level of noise. Mead's resource equation states that the number of degrees of freedom acquired with the experiment ( $N - 1$ ) is spent in estimating parameters for the factors and their interactions, the blocks, and the remaining degrees of freedom stay for the residuals. It is this last number which is fixed to be between 10 and 20.

Let us assume a simple experiment in which we have a single factor for which we need to spend  $T$  degrees of freedom estimating its parameters (there are  $T + 1$  levels in this factor), a blocking variable for which we need to spend  $B$  degrees of freedom, and  $E$  degrees of freedom left for the residuals. Mead's resource equation in this case would be

$$N - 1 = T + B + E$$

Let us refer as  $\sigma_T^2$  to the effect size of the treatment, which can be calculated as

$$\sigma_T^2 = \frac{1}{T + 1} \sum_t (\alpha_t^{(T)})^2$$

That is the variability accounted by the treatment is related to the sum of the squares of its main effects. In this situation, the normalized effect size that we can detect with different  $E$  degrees of freedom depends on the number of treatments. The following table shows the sensitivity we achieve for various combinations of  $E$  and  $T$ , in all cases  $\alpha = 0.05$

$T$	$E$	Normalized effect size ( $\sigma_T/\sigma_\epsilon$ )	Statistical power ( $1 - \beta$ )
1	10	1.16	0.95
		1.04	0.90
		0.90	0.80
		0.84	0.75
		0.63	0.50
		0.41	0.25
1	20	0.81	0.95
		0.73	0.90
		0.63	0.80
		0.59	0.75
		0.44	0.50
		0.29	0.25
3	12	1.23	0.95
		1.11	0.90
		0.97	0.80
		0.92	0.75
		0.70	0.50
		0.48	0.25
3	20	0.94	0.95
		0.85	0.90
		0.74	0.80
		0.70	0.75
		0.54	0.50
		0.37	0.25

With these two tables we recognize the following rules:

1. The effect size is implicitly chosen to be of about the same size of the standard deviation of the residuals ( $\sigma_T/\sigma_\epsilon \approx 1$ ) with a statistical power between 90 and 95%.
2. The detectable effect size increases (our detection capacity is worse) as the number of treatments grow.
3. These experiments are normally designed with the same number of animals per treatment group. That is why we cannot always achieve a desired number of degrees of freedom for a particular design. For instance, with 4 treatment groups ( $T = 3$ ), if we put 3 animals per group, we have 8 degrees of freedom left for the residuals, while if we put 4 animals per group, we have 12 degrees of freedom left for the residuals. That is, we cannot achieve exactly  $E = 10$ .

#### Important remarks

30. The resource equation leaves us in a situation in which we will be able to recognize an effect size that is in the order of the level of noise ( $\sigma_T/\sigma_E \approx 1$ ). However, the specific power and effect size depends on the number of

treatments and the number of extra degrees of freedom left for the residuals. Again, the specific choice is implicit and unknown, as it has not been based on a proper power analysis.

In analysis involving cytometry, the number of cells being analyzed is certainly huge, in the order of  $10^4$  to  $10^5$  cells. However, we should be aware that these cells are not independent as they are all coming from the same animal and they are affected by a common treatment, same severity of the injury, same diet, etc. Analyzing at the level of the number of cells fulfilling some condition (for instance, being  $CD8^+$  T cells with interferon  $\gamma$  receptors) is useful to determine the extent of the response in a single animal, there are statistically sophisticated tools for this task (Nájera et al, 2010). However, to infer the strength of the response in a population of mice, we must use the measurements of each animal as a single observation and follow the standard data analysis procedures used when other quantities such as blood glucose concentration are used in research. That is, an animal gives a single observation, not 80,000 because we have extracted 80,000 cells from it.

#### Important remarks

31. We should not be fooled by having a large number of measurements, as these measurements may not be independent of each other.

### 2.2.2 Experiments have to be repeated three times to be statistically significant

Many researchers think that claiming an effect of a treatment after having observed a positive result once is not statistically sound. They are in a way correct although for different reasons than they think. As we have seen in Secs. 1.3 and 1.4, there are two important concepts when we estimate any statistical parameter as, for instance, the difference between the means of two groups; namely, bias and variance. Sample size calculation, hypothesis tests, and p-values address variance very well, but they are totally disarmed against bias. For instance, there is no statistical method that is able to determine if the difference between the two group means is caused by a malfunctioning of the analytical device during the week of the experiment or that the differences between the control and treatment groups are due to a bottle of chemicals that someone left open inadvertently. In this sense, repeating the experiment more times randomizing everything except the variables of study will definitely help to make sure that the positive result of an experiment is not due to some unforeseen bias (see Sec. 1.3). In this direction, we may randomize the analytical device used for the measurements, the pipettes, the person doing the experiment, the position of the cages within the shelves, ... or even better doing the experiment in a different laboratory. On the other side, in the absence of bias, which unfortunately is always unknown, there is no need to repeat an experiment whose p-value is  $10^{-14}$ , it will always show that there is a difference between the two groups (see Sec. 1.5.2).

However, instead of thinking of the experiment as an experiment that has to be repeated three times, we may think of it as a larger experiment that will be split in three

times, three laboratories, or three parallel branches. The mini-experiments will have their own parameters as they are considered to be a blocking variable (Sec. 1.3), although the number of degrees of freedom required for the detection of a possible bias is very small ( $B - 1$  in the case of  $B$  mini-experiments). The sample size also has to be divided by the number of blocks. For instance, let us assume that we have determined that we need 18 individuals per group to be able to identify a given difference between the control and treatment groups. If we divide the experiment in three mini-experiments, then in each mini-experiment the number of control and treated animals will be 6 ( $=18/3$ ). We may increase a bit this number to 7 to compensate for the fact that we will need to estimate 2 extra parameters for the block contributions.

Following with the example in the previous section, let us assume that we want to detect a difference of size 3 in a standard deviation of 2 in the observations within a subgroup. We will use both sexes treated as a block, and we plan to repeat the experiment 3 times in different periods of time separated by two months and by different researchers. The different repetitions of the experiment are considered also as block variable whose levels have to be determined using 2 degrees of freedom. We wonder how many animals per group we need to perform this experiment to be able to capture the same differences between control and treatment as in the previous example. In particular we will design our experiment with:

$$\begin{aligned}\sigma_T &= 1.5 \\ \sigma_S &= 0.5 \\ \sigma_R &= 1 \\ \sigma_e &= 2\end{aligned}$$

where  $\sigma_T$  represents the variability due to control or treatment,  $\sigma_S$  the variability caused by sex,  $\sigma_R$  the variability due to the repetition of the experiment, and  $\sigma_e$  the variability within each of the subgroups ([Repetition 1, Control, Females], [Repetition 1, Control, Males], ... up to 12 subgroups). We only need  $n = 2$  animals per group making again a total of 24 animals. That is, in the first repetition of the experiment we will use 2 control females, 2 control males, 2 treated females, and 2 treated males.

We may compare this number of animals with the number of animals of repeating the one-sex-at-a-time three times ( $=72 \cdot 3 = 216$ ). Surprising as it may, the power of factorial designs comes from fact that we still have 12 animals receiving the treatment and 12 receiving the control, as in our previous design. But, we have distributed these 24 animals into 3 mini-experiments so that we are protected against possible, unknown biases in each of the repetitions. This design has a statistical power of 93.6% for the treatment/control variable, but only 21.4% for the sex contribution.

If we want to have a 90% statistical power to detect the small differences between the sexes, then we need  $n = 15$  animals per group, yielding a total of 180. The statistical power for the treatment/control group raises to 100%. 180 is still under 216 and this design the possibility to make very fine detections in sex, which are totally out of reach of the one-sex-at-a-time design.

This idea of performing mini experiments can be extended to studies in which the main result is a proportion. For instance, we are interested in the analysis of the proportion of animals surviving after 15 days of being challenged with a severe infection.

Without treatment less than 1% of the animals survive, and we want to explore the survival increase induced by different treatments. For every treatment we will compare the treatment vs a control situation with no treatment. For experiment complexity reasons, we have decided to work with 10 animals per group, and we will perform the experiment twice, to be sure that the first result is not biased by any uncontrolled variable. With 10 animals per group, we have the following statistical power to detect the following changes in the survival proportion between the two groups:

Survival proportion change ( $\Delta p$ )	Statistical power ( $1 - \beta$ )
0.64	0.95
0.58	0.90
0.50	0.80
0.47	0.75
0.36	0.50
0.26	0.25

Instead of repeating the experiment twice and having relatively small power in both replications, we could treat both replications as parts of a larger experiment with a larger sample size (and power). We may compare the proportions observed in both replications to determine if there is a significant deviation. That is, we could compare the survival proportion in the control group of Subexperiment 1 to the survival proportion of the control group of Subexperiment 2. In this comparison we have the statistical power listed in the table above. If the null hypothesis of the equality of proportion is rejected, then we can presume that there is a significant mismatch between the results observed between the first and second subexperiments. We could proceed similarly for the treated groups, only that now the detectable effect sizes change because for proportions, the statistical power and effect size are tightly coupled to the observed proportion. Let us assume that the survival proportion in the treatment group is about 50%. Then, our capacity to detect proportion changes with two groups of 10 animals is

Survival proportion change ( $\Delta p$ )	Statistical power ( $1 - \beta$ )
0.50	0.62
0.47	0.50
0.38	0.25

We agree that we do not have much power in the comparisons between the two subexperiments, but it has been our choice to work only with 10 animals per group.

If we do not reject the null hypothesis that the proportions in the two subexperiments are equal for both the control and the treatment groups, then we can merge them into a single experiment with 20 animals per group and our detection capacity increases to

Survival proportion change ( $\Delta p$ )	Statistical power ( $1 - \beta$ )
0.41	0.95
0.37	0.90
0.32	0.80
0.30	0.75
0.23	0.50
0.17	0.25

We have gained a detection power increase of about 20% (from 64% to 41%) by moving from 10 animals per group to 20 animals per group.

Actually, we are not interested in a single treatment, but in 3 treatments compared to the control (no treatment) and on the responses of 6 mouse strains to the challenge and treatments. In total we will do 18 tests ( $T_1, T_2, T_3$  vs Control for the 6 strains). This is a fair amount of tests. To avoid Type I error inflation, that is, inflation of false positives, we should decrease the confidence level to  $0.05/18=0.00277$  (Bonferroni correction). Then, our sensitivity again decreases to

Survival proportion change ( $\Delta p$ )	Statistical power ( $1 - \beta$ )
0.58	0.95
0.54	0.90
0.48	0.80
0.46	0.75
0.38	0.50
0.31	0.25

$\Delta p = 0.58$  that we have now with a statistical power of 95% is more sensitive than the  $\Delta p = 0.64$  with a statistical power of 95% that we had in our first design with two replicated experiments of size 10. The difference is that now we are considering both subexperiments as part of a larger one, with a low-powered test between the two subexperiments to detect possible biases in any one of them. We have also lowered  $\alpha$  to 0.00277 to keep the family-wise false positive error below 0.05 ( $1 - (1 - 0.00277)^{18} = 0.049$ ). Our original design had a family-wise false positive error of 0.84 ( $1 - (1 - 0.05)^{36} = 0.84$ ), meaning that we had a 84% probability of having a false positive result in our 36 tests.

If we want to keep a detection capability of  $\Delta p = 0.5$  with a statistical power of 95% and reducing  $\alpha$  to 0.00277 to account for the multiple comparisons, we would need to increase the group size to  $N = 25$  individuals per group. That is, two subexperiments of 13 individuals per group.

#### Important remarks

- To protect ourselves against bias, we should consider a large experiment that is split into smaller mini-experiments rather than repeating the experiment 3 or any other number of times. Blocking, blinding, and randomizing are our best tools to fight bias. The analysis of all the data from the

large experiment will allow us to determine whether any one of the mini-experiments has suffered any bias.

### 2.2.3 Treating counts as continuous variables

Sometimes the main variable of interest is a count, *e.g.*, the number of tumor lesions, the number of cells of a given type in a microscope field, or the number of events of a given type (for instance, fights) in a period of time. We may be interested in comparing these counts between two groups. We may be tempted to use the formulas in Sec. 4.1, which are the most standard formulas used in sample size calculations (these are related to Student's t-tests, ANOVA, etc.). However, these formulas are designed for continuous, normally distributed variables. A count variable is not continuous (we may have 0, 1, 2, ... tumor lesions, but we cannot have 1.37 tumor lesions). In this case, a more appropriate calculation would use Poisson count rates for the design as the ones shown in Sec. 4.4. There are other discrete distributions to model counts like the binomial or negative binomial. The problems related to the binomial are those in which we are interested in the proportion of individuals with a given property (for instance, the proportion of animals with allergy). Individuals should be independent of each other. That is, the presence of the property in one animal should not influence the presence or absence of that feature in another animal. This is not the case of contagious diseases (the presence of a diseased animal in a cage may cause other infections within the same cage). If our problem is one in which the proportion of animals is the main target, we should use the techniques shown in Sec. 4.2. Negative binomials are widely used because they have an advantage over Poisson or binomial distributions. Namely, they are specified by two parameters so that we have two degrees of freedom to specify the mean and standard deviation of the distribution. Poisson and binomial distributions are mostly specified by a single parameter, and in this way, the mean and standard deviation are not free to vary, they are tightly linked and given one, the other is automatically fixed.

A researcher is interested in comparing the mean number of tumoral lesions (pancreatic intraepithelial neoplasias) between two mouse genotypes. The research hypothesis is that one of the genes, missing in one of the mouse strains, helps reducing the number of lesions. For the sample size calculation, they have seen that [Guerra et al \(2011\)](#) studies the relationship between pancreas inflammation and pancreas cancer. They report a mean number of lesions around 7 with a standard deviation around 2.75. A characteristic feature of the Poisson counts is that the standard deviation is the square root of the mean, which is exactly the case here ( $\sqrt{7} = 2.66$ ). Consequently, we may use the sample size calculations in Sec. 4.4.2. They want to detect differences of at least 1 standard deviation between the two groups (the reference group would have a mean around 7, and the second group around 9.75). If we want to have a confidence level of 95% and a statistical power of 90%, then we need 19 animals per group.

If we had used the formulas for continuous variables (Sec. 4.1.5), we would have had a sample size of 22 animals per group if we account for the variance difference due to the increase of number of lesions in one of the groups or 18 animals per group, if we do not realize of this increase. In the first case, we would have an overpowered design,

while in the second we would have a slightly underpowered design. In any case, the approximation by a continuous variable was not so bad.

A count can be approximated by a continuous variable if the count is large enough, so that the steps between one observation and the next can be seen as small from the point of view of the mean count. The Poisson is well approximated by the Gaussian if the mean count is over 10, and very badly approximated if it is smaller than 5. In our case, the mean was 7, and we have already seen that the Gaussian approximation gives, at least, a ballpark estimate of the number of needed animals per group.

If the count cannot be approximated by a continuous variable, we may still calculate the sample size using the formulas for the Student's t-test or any other parametric test. But, we know that we cannot use the parametric test during the analysis because the variable does not follow the presumed distribution. Instead, we will use the corresponding non-parametric test and adjust the sample size calculated before by the Asymptotic Relative Efficiency (see Sec. 1.6).

- Example 39: The number of measurement attempts to get a successful measurement of the intraocular pressure with a tonometer can be around 1.1 with a standard deviation around 0.16 (Pease et al, 2006). As designers of a new tonometer we want to show that our new model is better than the reference (whose mean is 1.1). We will consider our design a success if the mean drops down to 1.05. We will ask a number of researchers to get 5 successful readings from our tonometer and the tonometer of reference (some of them may need 5 measurements to get the 5 successful readings, some of them may need, 6, 7, ...). The sequence of tests will be randomized. For instance, if we label the two tonometers as A and B, then a particular researcher may follow the measurement sequence AB-BAABBBAA. Then, we will make a permutation test comparing the number of attempts with A and with B. How many researchers do we need to involve in our experiment to identify this difference?

Solution: If we compute the sample size with a one-tail, two-independent samples Student's t-test to have a statistical power of 90% when the difference is at least 0.05 and the standard deviation of the observations is about 0.16, we would need 168. However, we will perform a permutation test, not a Student's t-test. Then, according to the table of Asymptotic Relative Efficiencies in Sec. 1.6, we should use

$$N = \frac{168}{0.85} = 198 \quad \text{researchers.}$$

In this example we have used a permutation test because the differences between the number of attempts to get 5 successful measures using tonometer A or using tonometer B is a discrete variable taking values ..., -2, -1, 0, 1, 2, ... We expect most of the values to be between -2 and 2, although the difference is not bounded on either side. This is clearly away from the Gaussian distribution assumption of the Student's t-test. Some researchers may be tempted to have used Wilcoxon signed-rank test, as it is non-parametric. However, this test is not free of assumptions, one of them being that the observed difference is a continuous variable, which is not the case in our example.

Very often, experiment results are graded into a scale. For instance, skin lesions or histopathological samples are graded from 0 (no lesion), 1 (mild lesion), ... to 3 (moderately affected) as in [Barton et al \(2000\)](#). None of the tests assuming continuous differences can be applied (Student's t-test, Mann-Whitney U test, Wilcoxon signed-rank test, or Kruskal-Wallis ANOVA). This leaves us with very few choices. Permutation tests could be applied. Their assumption is that in the absence of difference (null hypothesis), the distribution is symmetric around 0.

When grades are added, for instance, we evaluate multiple criteria using grades 0, 1, 2, or 3, and add the different grades into a single score (see, for instance, [Lee et al \(2006\)](#)), then the resulting variable is not continuous yet, but can be approximated by a continuous variable if the number of individual grades is sufficiently high (e.g., for 10 individual grades between 0 and 3, the total score ranges from 0 to 30). In this case, the assumptions made by non-parametric (or even parametric) tests, although not completely fulfilled, are much better preserved.

#### Important remarks

33. Statistical tools based on the Gaussianity of the observations such as Student's t-test, or ANOVAs should not be used with counts or discrete data. Non-parametric tests such as Mann-Whitney U test, Wilcoxon signed-rank test, or Kruskal-Wallis ANOVA, also assume that the underlying variables are continuous (although not Gaussian). There are specific tools for ordinal variables (cumulative odds and odds ratios, Kendall's  $\tau$  and Spearman's  $\rho$  correlations, Mantel-Haenszel test for linear trend). These tools are not easy to use. Permutation tests are extremely easy to use and they are very general. The price to pay as we move away from parametric tests is statistical power. Conversely, we will need larger sample sizes to be able to detect the same effect size with the same statistical power as a parametric test. The design of the experiment is performed assuming a Gaussian distribution of the observations, and corrected due to the non-continuous or unknown distribution of the observations.

### 2.2.4 Both-sexes vs One-sex or One-sex-at-a-time experiments

The UK Medical Research Council<sup>1</sup> and US NIH<sup>2</sup> have decided to only fund animal research that includes both male and female animals. Sex causes major differences in many animal and human studies and experiments with both sexes should be promoted as a general biomedical research principle ([Clayton, 2016](#)). Unless it is justified, for instance an study on the effects of hysterectomy (the surgical removal of uterus), there can be constitutively different responses between males and females (main effect of the factor), and treatments may have a different effect on males and females (interactions between treatment and sex).

Beside the lack of scientific knowledge by disregarding half of the animal population, there is another drawback of working only with one sex: animals of the other sex

<sup>1</sup><https://www.ukri.org/news/use-of-both-sexes-to-be-default-in-laboratory-experimental-design>

<sup>2</sup><https://grants.nih.gov/grants/guide/notice-files/not-od-15-102.html>

have to be culled. Consequently, the number of animals required to make an experiment is twice the number used in the statistical comparisons. These single-sex experiments may be standard in some domains and the standard sex in that domain varies from field to field. Some researchers have understood this mandate to use both sexes as one-sex-at-a-time. However, this cull of animals also occurs as often animals of a given age are required.

In this book we have provided statistical tools to better handle this situation. Specifically, in Sec. 5.1.6 we explain one of the most powerful techniques: factorial design. Sex and treatment are two factors, each one with 2 levels (control vs treatment, male vs female). We may identify the contribution of each one of the factors, main effects, and their interaction (e.g., the treatment has a different effect on each sex). The assumption of factorial design is that the variance of the observations in each one of the level combinations is the same. That is, the variance of control males, control females, treated males, and treated females is the same. This may easily be the case, and if not, for the calculation of the sample size we may take the worse variance. The fact that males and females may be quite different and give a large variability in the whole set of observations is irrelevant, because this variability will be accounted for by the main effects of sex. Additionally, the number of animals required for this design is not twice the number of animals used for one sex. Let us give some numerical example.

Let us assume that an experiment is traditionally performed on female animals. We are measuring a variable whose mean in the control group is 10 (in arbitrary units), with a standard deviation of 2, and we want to detect a decrease by 3, that is the mean of the treated group should be 7 or less to declare the experiment successful. We will assume that the standard deviation of treated females is also 2. Using the formulas in Sec. 4.1.5 with a confidence level of 95% and a statistical power of 90%, we would need 9 animals per group. However, to produce these 18 females, we will need to cull other 18 males, making a total of 36 animals involved in the first stage of the experiment. If the experiment with females is successful, we will repeat the experiment with males, adding up to a total of 72 animals.

Let us now plan the experiment to include both sexes assuming that the average response of males is 12, but we also want to detect an effect size of 3, and we will assume that standard deviation of all subgroups (control males, control females, treated males, and treated females) are all equal, and equal to 2. If we put  $n$  animals per subgroup, we will use a total of  $4n$  animals in the experiment. For simplicity, let assume that we do not foresee a difference of the treatment between males and females. That means that sex is a block, not a factor. We will need 2 degrees of freedom to determine the parameters of the experiment: 1 for the main effect of the control/treatment, and 1 for the main effect of sex. This time the design should use the formulas in Sec. 4.1.6. Mead's resource formula would say that we need

$$N - 1 = T + B + E = 1 + 1 + 15$$

where we have left 15 degrees of freedom for the residuals. That is, we will need 18 animals, as it is not a multiple of 4, we will use 5 animals per group, making a total of 20 animals (both males and females). It must be noted that we have not used in this calculation our aim of detecting changes of size 3, that is in Mead's formula the

effect size is implicitly defined. If we use the more complete formula in which the effect size is included, then we would get  $n = 6$  animals per group, making a total of 24, and raising the statistical power to 93.8%. We may compare the total of 24 animals including both groups to the 72 animals making one-sex-at-a-time experiments.

We may also include extra animals to be able to estimate the interaction between sex and treatment. Let us assume that this interaction would be around +1 in females and -1 in males. Then, the expected means in each one of the groups would be:

	Females	Males
Control	10	12
Treated	7+1	9-1

Our factor model is

$$y_{ijk} = \mu + \alpha_i^T + \alpha_j^S + \alpha_{ij}^{TS} + \varepsilon_{ijk}$$

That is, our observations will be an overall mean,  $\mu$ , plus a contribution of the treatment,  $\alpha^T$ , plus a contribution of sex,  $\alpha^S$ , and a contribution of the interaction between treatment and sex,  $\alpha^{TS}$ .

It must be noted that the table above is the minimum difference we want to detect with a given statistical power and confidence level. We have not performed the experiment yet, and we do not know whether this difference will occur or not, but if it happens, we want to be able to prove that it is statistically significant with this statistical power and confidence level.

To calculate the sample size, we must determine the size of each one of the contributions, for which we must estimate the different terms of the model. We will do so by computing the marginal means as shown in the table below

	Females	Males	
Control	10	12	$y_{C..} = 11 \Rightarrow \alpha_C^T = 1.5$
Treated	8=7+1	8=9-1	$y_{T..} = 8 \Rightarrow \alpha_T^T = -1.5$
	$y_{.F.} = 9 \Rightarrow \alpha_F^S = -0.5$	$y_{.M.} = 10 \Rightarrow \alpha_M^S = 0.5$	$\mu = 9.5$

The interactions can be estimated from each one of the cell means

$$\begin{aligned} y_{CF.} &= 10 = \mu + \alpha_C^T + \alpha_F^S + \alpha_{CF}^{ST} = 9.5 + 1.5 - 0.5 + \alpha_{CF}^{ST} \Rightarrow \alpha_{CF}^{ST} = -0.5 \\ y_{CM.} &= 12 = \mu + \alpha_C^T + \alpha_M^S + \alpha_{CM}^{ST} = 9.5 + 1.5 + 0.5 + \alpha_{CM}^{ST} \Rightarrow \alpha_{CM}^{ST} = 0.5 \\ y_{TF.} &= 8 = \mu + \alpha_T^T + \alpha_F^S + \alpha_{TF}^{ST} = 9.5 - 1.5 - 0.5 + \alpha_{TF}^{ST} \Rightarrow \alpha_{TF}^{ST} = 0.5 \\ y_{TM.} &= 8 = \mu + \alpha_T^T + \alpha_M^S + \alpha_{TM}^{ST} = 9.5 - 1.5 + 0.5 + \alpha_{TM}^{ST} \Rightarrow \alpha_{TM}^{ST} = -0.5 \end{aligned}$$

We may now associate a variance to each one of the contributions:

$$\begin{aligned} \sigma_T &= \sqrt{\frac{(\alpha_C^T)^2 + (\alpha_T^T)^2}{2}} = 1.5 \\ \sigma_S &= \sqrt{\frac{(\alpha_F^S)^2 + (\alpha_M^S)^2}{2}} = 0.5 \\ \sigma_{ST} &= \sqrt{\frac{(\alpha_{CF}^{ST})^2 + (\alpha_{CM}^{ST})^2 + (\alpha_{TF}^{ST})^2 + (\alpha_{TM}^{ST})^2}{4}} = 0.5 \end{aligned}$$

If we want to be able to detect the interactions between the treatment and sex, then the sample size must be increased to  $n = 43$ , or  $N = 4n = 172$ . This is much larger

than the previous experiment. This is expected because in our previous experiment we wanted to detect a difference between groups of 3 ( $\alpha_C^T - \alpha_T^T = 3$ ) when the standard deviation of the observations within a subgroup was 2. However, we now want to detect a difference 1 ( $\alpha_{TM}^{TS} - \alpha_{TF}^{TS} = 1$ ) when the standard deviation of the subgroups is still 2. This is a more complex detection scenario and we need a larger sample size.

#### Important remarks

34. Factorial designs allow us to consider both sexes without having to repeat the experiment twice. Actually, for a very little cost in terms of animals we can evaluate the influence of the sex and the interaction between treatment and sex.

Another important lesson from this experiment is that the ANOVA decomposition into different effects is not the same as the decomposition humans do. The ANOVA decomposition has the restriction that the summation of all the individual contributions have to be 0 (see Sec. 5.4.1). This was certainly our case above:

$$\begin{aligned}\alpha_C^T + \alpha_T^T &= 0 \\ \alpha_F^S + \alpha_M^S &= 0 \\ \alpha_{CF}^{TS} + \alpha_{CM}^{TS} &= 0 \\ \alpha_{TF}^{TS} + \alpha_{TM}^{TS} &= 0 \\ \alpha_{CF}^{TS} + \alpha_{TF}^{TS} &= 0 \\ \alpha_{TM}^{TS} + \alpha_{TF}^{TS} &= 0\end{aligned}$$

However, this is not the decomposition we originally did in our mind by adding +1 to the treated females and -1 to the treated males. In particular, our decomposition of effects does not fulfill the last two equations.

### 2.2.5 Which variance to use for the calculation of the sample size?

The variance of the observations is one of the key parameters in all sample size calculations (see Chap. 4). After all, hypothesis testing can be seen as a signal detection problem in the presence of noise, and the amount of noise is of primary importance to determine the number of samples we need to be able to detect a given amount of signal (the smaller the signal we want to detect with respect to the amount of noise, the larger the sample size). However, we now wonder what is the variance we should use in our calculations. This is particularly relevant in all those experiments whose primary goal is to make an statement about the mean (Sec. 4.1).

For example, [Yoshida et al \(2016\)](#) shows the mean and variability of a tumor volume as a function of time since tumor implantation. In Fig. 2.3, we show a typical tumor growth curve along with the standard deviations of the measurements at different time points after implantation. We observe that there are wild differences between the standard deviations at different time points. More importantly, the standard deviation after 28 days of implantation would imply negative tumor volumes, which are clearly nonsensical. The reason is that tumor growth is a process in which cell division plays a central role. In other words, the underlying process is multiplicative, rather than additive, and the distribution of tumor sizes is not symmetrical. For this reason, we need

to work in a logarithmic scale rather than in a natural scale. If we take the logarithm of our measurements we see that the standard deviation of these logarithms are much more stable than the standard deviations of the measurements in natural units. That means that we may apply a technique like ANOVA in the logarithmic measurements, but not in the raw measurements.

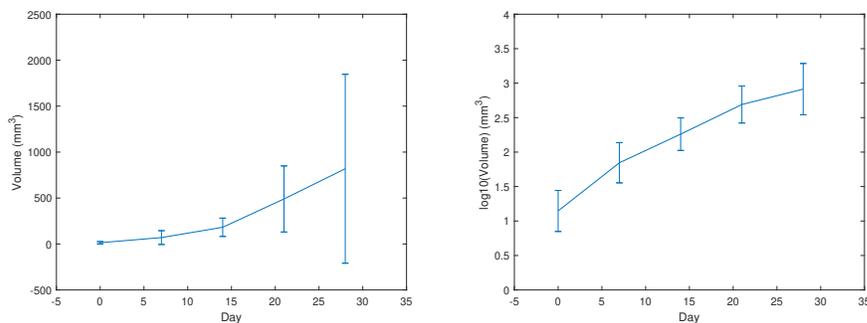


Figure 2.3: Left: Tumor volume as a function of time since implantation. Right: Same plot in logarithmic units.

If we want to detect a reduction of the tumor volume by a factor 2 after 28 days of tumor implantation, we would need  $N = 103$  individuals per group (control and treatment) if we work in natural units (an effect size of  $410 \text{ mm}^3$  with a standard deviation of 1000, confidence level=0.05 and statistical power=90%, see Sec. 4.1.5), while we would need  $N = 27$  animals per group to detect the same change in logarithmic units (an effect size of  $-0.3 = \log_{10}(0.5)$  and a standard deviation of 0.37).

The concentration of many hormones or metabolites are also log-normal. For instance, [Kalliokoski et al \(2012\)](#) showed that circulating corticosterone and fecal corticosterone metabolites are log-normally distributed. The reason is that the production of many substances are regulated by multi-stage signaling cascade giving rise to a multiplicative process. The difference between a normal or a log-normal distribution cannot be assessed with few samples (most normality tests will result in high p-values). However, the long tail of the log-normal distribution will distort the standard analyses (Student's t-tests, ANOVA tests, etc.) that assume normally distributed data.

#### Important remarks

35. We must be careful in the choice of the standard deviation in variables whose underlying process is multiplicative/divisive (titrations, cell division, dilutions, gene expression, signaling cascades, etc.). These variables tend to be log-normal, we should analyze their logarithm, and consequently we must use the standard deviation in the logarithmic space.

Another common situation is when we have the influence of a block variable. For instance, let us assume that we are measuring a variable whose response depends on the sex of the animal and the treatment applied (see Sec. 5.4.1 for a detail explanation

of linear models):

$$y_{ijk} = \mu + \alpha_i^S + \alpha_j^T + \varepsilon_{ijk}$$

That is, the observation for a particular animal  $k$  depends on its sex,  $i$ , the treatment received,  $j$ , and some noise specific to that animal. The variance we need to use for the sample size calculation is the one of  $\varepsilon$ , not the one of  $y$ . That is, the variance from the homogeneous groups (equal sex and equal treatment), and not the one of the raw observations. Let us graphically illustrate these ideas. Let us assume that we have  $\mu = 10$ ,  $\alpha_{male}^S = 2 = -\alpha_{female}^S$ ,  $\alpha_{control}^T = 0.5 = -\alpha_{treatment}^T$ , and  $\sigma_\varepsilon = 1$ . Fig. 2.4 shows the acquired data. The estimate of the noise standard deviation is 0.91 if we estimate sex differences and 2.38 if we do not. To detect an effect size of 1 ( $= \alpha_{control}^T - \alpha_{treatment}^T$ ) with a 95% confidence level and 90% statistical power if we also estimate sex differences we need 15 individuals per group, while we need 98 per group if we do not explicitly estimate sex differences.

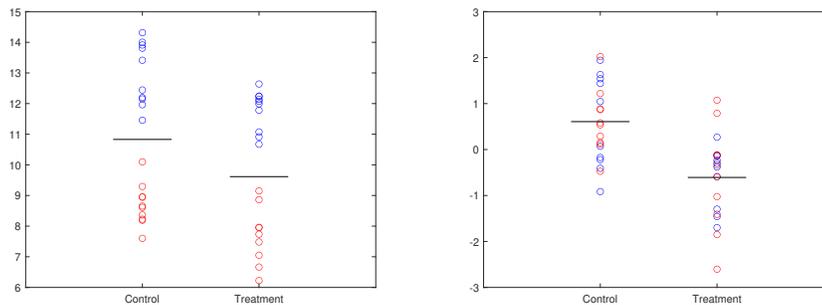


Figure 2.4: Left: Raw measurements of a control and treatment groups. Blue samples are males while red samples are females. Right: Same plot after eliminating the estimated overall mean and sex differences.

### Important remarks

36. When designing experiments with several factors or blocking variables, the standard deviation needed for the sample size calculation is the one of the residuals. This information must be gathered from previous experiments. If it is unknown, then we should use the variance of the observations, knowing that the sample size will be larger than needed. How much larger depends on the differences between blocks, whose contributions to variability we have not been able to disentangle before carrying out the experiment.

### Looking for a variance in previously published literature

As stated above, the variance is the variance of our observations. If we are measuring the concentration in blood of a given compound, it is the variance of the observed concentrations. Some practitioners use previous literature to learn about this variance,

which is a very recommended practice as it saves animals as long as the information gathered from the literature is sufficiently close to the variability we will have in our experiments. Our strain of animals may be different from the one of a specific paper. However, after looking at several papers with different strains, we can have an educated idea of what kind of variability should be expected in our experiment. When looking at previous papers, we should pay attention to the variability of their observations, and not to the sample size used in those papers. The reason is that our goals of precision and blocks will very likely be different from theirs. Still, the sample size of previous works already give an idea of the kind of effect size we should expect from different sample sizes (see Sec. 2.2.1). A mistake that we have seen in some researchers is calculating the variance of the sample size in previous works, rather than the variance of the observations in previous works. The various sample sizes among works and their variance are irrelevant for the design of an experiment about making inference on the effect of a treatment on the animals.

### 2.2.6 Foreseeing dropouts

Let us imagine that we have calculated the sample size to be  $N = 10$  animals per group for whatever comparison purpose we have designed it for. However, from previous experiences we know that some animals become untestable or unusable for our research because they get an infection of a different kind interfering with our experiment, there is an accident and we lose a cage, or any other reason. We foresee that  $p = 10 - 20\%$  of the animals could become unusable for any reason. Then, we can increase the number of animals per group such that when we lose  $p\%$  of them, still we have  $N$  per group. In this way, we would need:

$$N' = \frac{N}{1 - p} \quad (2.5)$$

In our case,  $N' = 13$  animals so that when we lose 20% of them, we will have 10 per group.

It is important two observations:

- The dropouts have to be unrelated to our treatment. If our treatment is an immunosuppressor and animals get infected because of our treatment, then by analyzing only the treated, non-infected animals we would be biasing our results.
- The dropouts cannot belong to a different subpopulation of animals. If only a fraction of the animals respond to our treatment, then by analyzing only the respondents we are biasing our results. The proper way to handle this situation would be to report a proportion of respondents (and its confidence interval), and then the effect size of the treatment in whatever variable we are measuring (and its confidence interval).

We may also consider dropouts at multiple stages. For instance, in an oncological experiment, we know that 20% of the animals do not develop a tumour even if we inject tumoral cells. In our experiment we will start a given treatment when the tumour has reached a given volume. However, we know that 15% of the tumours do not reach this volume after 4 weeks after injection. Let us refer to the dropouts of the first stage

(not developing tumour),  $p_1$ , and to the drops of the second stage (not reaching a given volume),  $p_2$ . Then, the number of animals that we need for our experiment is

$$N' = \frac{N}{(1-p_1)(1-p_2)} \quad (2.6)$$

Following a similar reasoning we can include as many consecutive stages as needed.

#### Important remarks

37. If we expect dropouts from the experiment, we can increase the sample size to compensate beforehand for these losses.

### 2.2.7 Pilot studies

A researcher is interested in evaluating the immune response to the metastasis of a given kind of tumor. For doing so, she will measure the number of immune cells of a given type (for instance, large peritoneal macrophages) in a particular anatomical location (the peritoneal cavity). Being so specific, there are no previous publications about the number of expected cells of that type in that location. So the researcher poses the analysis as pilot study in which 7 mice strains will be tested, with 11 time points to measure the progression of the number of cells. For each combination of strain and time point, 7 animals will be employed. The experiment will be repeated twice.

If we count the total number of animals, it is  $7 \times 11 \times 7 \times 2 = 1,078$ , which is far beyond the expected size for a pilot study. Another problem with this experiment is that it is not designed for any specific measurement purpose. From the data analysis point of view experiments are performed in order to detect differences between various experimental/treatment conditions, or to measure any parameter with a given precision. None of the two are explicitly mentioned in this experiment. However, the way it is formulated points to an experiment whose goal is to measure the mean number of cells of a given type. With  $N = 7$  animals, the precision achieved is about 0.92 times the standard deviation (for this purpose you may use the calculator at <http://i2pc.es/coss/Programs/SampleSizeCalculator/index.html>). That is, the 95% confidence interval about the mean will have a length whose size is  $2 \times 0.92s$ , where  $s$  is the observed standard deviation. As we discussed in Sec. 2.2.2, it is better to consider repetitions of the experiment as a mini-experiment within a larger experiment. In this case, the measurement model would be

$$y_{ij} = \mu + \alpha_i^{repetition} + \varepsilon_{ij}$$

That is, we need 1 degree of freedom to estimate the possible differences caused by the repetition. then there remain 12 degrees of freedom to estimate the confidence interval of the mean. The precision increases to 0.64, that is, the confidence interval of the mean is narrower.

If the goal of the experiment is to determine the standard deviation of the observations, then with  $N = 7$  the 95% confidence interval for the standard deviation is  $\sigma \in [0.6s, 2.2s]$ . That is, with  $N = 7$  the length of the confidence interval is  $2.8s$  for

the standard deviation and 1.8s for the mean. This may or may not be acceptable uncertainties for each one of these parameters (standard deviation or mean). But now, we are aware of the compromises we have chosen. Due to the excessive lengths of these confidence intervals, very often the same amount of information is obtained from previous publications if any similar experiment is reported. If it is truly the first time to measure a particular quantity, then the researcher must be aware of the implications in terms of uncertainty (confidence intervals) implied by a low number of animals.

In the experiment, as proposed by the researcher, there were 11 time points to perform these measurements (0, 1, 3, 5, 7, 9, 12, 15, 20, 30, and 60 days after tumor implantation). The number of animals may be reduced by considering the number of cells at each time point as a regression problem, for instance,

$$y_{ij} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_{ij}$$

That is, the number of cells observed for the animal  $j$  at time point  $t_i$ ,  $y_{ij}$ , is a function (in the example, polynomial) of time. In this way, all observations, no matter of the time, contribute to reduce the overall uncertainty. We may see that in the regression problem we need to estimate 3 parameters (the  $\beta$ 's), while in the original formulation of the problem we need to estimate 11 parameters (the mean number of cells of each of the time points). For the sample size calculation for a regression problem, see Sec. 4.3.

#### Important remarks

38. Pilot studies should be performed to: 1) check the feasibility of the study; 2) identify possible deficiencies in its design; 3) evaluate the severity of the procedures (pain, suffering, distress or lasting harm) and the actions to reduce them; 4) define humane end-points; 5) verify that the instructions are clear; 6) verify that the researchers have the sufficient training and skill to carry it out; 7) check that the equipment works as expected and that the experimenters know how to operate it; 8) check that the planned observations can be performed; or 9) if animals are required to perform any task, check that they can do it and that the task is not too difficult or too easy.
39. Pilot studies to determine statistical information (like the mean or standard deviation) are recommended only if there is absolutely no information about these parameters. If some previous information can be collected from some other source (like a similar experiment in another publication), the uncertainty from the pilot study is typically larger than the one from the publication (Sorzano et al, 2017).
40. Pilot experiments to keep “the number of animals low”, but then using 4 groups (positive control, negative control, and two treatments) do not make much sense. The “pilot” adjective cannot be the excuse to avoid the calculation of a proper sample size or to perform experiments “just to see what happens”. We should always think of the ethical implications of our research and the respect due to the animals and funding sources.

### 2.2.8 Decisions taken after low-powered tests

Pilot studies, as shown in the previous section, are ones of these studies in which we have very little statistical power. Strictly speaking, pilot studies with a statistical purpose are meant to determine sensible ranges of some parameters (mean, variance, proportion, correlation, ...) and, as such, there is not such a concept of statistical power (which is associated to a hypothesis test). However, the low sample sizes normally employed in pilot studies result in very wide confidence intervals turning the experiment not too informative.

The same situation can be encountered with hypothesis tests. Sometimes, with the best faith, we try to take data analysis decisions based on some preliminary, small experiments hoping that the results from these experiments will help us to better guide a larger, subsequent experiment. For instance, we are analyzing the immune response of mice to a viral challenge. We have multiple vias to apply our treatment. A researcher may want to perform a small experiment with only 3 animals per group using all vias to determine which combination gives the smallest variance of the results. Then, the best via will be used for the rest of the experiments, now with a larger sample size and comparing different compounds, doses, etc.

In principle, this approach sounds sensible and leading to an optimal experiment. The problem is that a sample size of 3 only allows us to detect very large differences between the standard deviation of the best via and the standard deviation of any other via. The following table shows the Effect size and its corresponding statistical power when the confidence level for the hypothesis test  $H_0 : \sigma_2^2 \geq \sigma_1^2$

Effect size ( $\sigma_2/\sigma_1$ )	Statistical power ( $1 - \beta$ )
27.2	0.95
18.7	0.90
12.5	0.80
10.8	0.75
6.2	0.50
3.6	0.25

That is, we can only reject the null hypothesis that  $\sigma_2 > \sigma_1$  if  $\sigma_2$  is 27.2 times  $\sigma_1$ , and in the best case, with very little power, if  $\sigma_2$  is 3.6 times  $\sigma_1$ . Very likely, we will not observe such large differences in standard deviation between the different vias. One of the vias will be the one with the minimum variance, but we will not be able to say which one it is because we cannot reject the hypothesis that the via with minimum variance is statistically significantly smaller than the other vias. Another way to put it is that the confidence interval on our estimates of the variance for the different vias are all too similar, they overlap too much, and we cannot safely choose one of them as the one with minimum variance. This is a consequence of the low sample size. If we want to detect at least when the standard deviation of any other group at least doubles the one of the minimum group (assuming that we have 4 possible vias and we apply a Bonferroni correction for  $\alpha$ ), we should use between 28 and 34 animals per group to have a statistical power between 90 and 95%. And still, twice the standard deviation seems to be a difference in variance that is rather unlikely to happen between different vias.

The tests for homocedasticity (all groups have the same variance) suffer from exactly the same problem, with the sample sizes typically used in groups (10, 15, 20), the standard deviations between the two groups have to differ by a factor 3.1, 2.5, or 2.1, to be rejected with a confidence level of 5% and a statistical power of 90%. If we check whether our groups are homocedastic or not, most of the time we will not reject the hypothesis that the groups have different variance. This failure to reject does not make them to have the same variance, simply that we do not have enough evidence to say that the variances are different.

The same happens with a researcher deciding on whether to use a parametric or a non-parametric test for analyzing his/her data. A seemingly sensible approach would be to acquire the data for the different groups and then checking with a normality test whether the observations follow a Gaussian distribution or not. The problem is the same of low sample size. With the typical sample sizes normally used in experiments with animals, between 10 and 20, normality tests have very little power, and either there is a huge difference between the observed data and the expected data from a Gaussian distribution, or there will not be enough evidence to reject the hypothesis that the data follows a Gaussian distribution. But this failure to reject does not make the data Gaussian.

Overall, using low-powered tests give us the false confidence that our data is normally distributed or homocedastic, but this may not be true and we cannot reject these hypothesis simply because our sample size is too small.

#### Important remarks

39. Tests with very few samples have very little power and most of the times the null hypothesis is not rejected. This may give us the false impression that our data is Gaussian, homocedastic or have any other property that we are testing. However, failing to show that our data is not Gaussian does not make it Gaussian. Simply, we do not have enough evidence to show that it is not Gaussian. In general, not being able to reject the null hypothesis should not make us believe that the true state of nature is the one described by the null hypothesis.

The following is another example of decisions taken on low-powered experiments.

- Example 40: A researcher is interested in delivering a bitter compound to mice. In order to find the optimal acceptance by the mice, the compound will be dissolved in water with 10 different concentrations of sugar (from 0.0% to 7.5% of sugar in mass). 3 mice will be tested in every group (groups are defined by the solution, that is, water+compound+X% of sugar). The group that has no rejection will be selected as the optimal sugar concentration for the rest of the experiment.

The problem with this approach is that even in the most extreme case (there is a group in which the solution was rejected by the 3 animals and another group in which the solution was accepted by the 3 animals), the p-value of this comparison (two-tails Fisher's exact test) is 0.1. That is, we are not sure of which solution is more accepted by the animals.

### 2.2.9 We should not calculate the sample size for every animal experiment

That is true. There are some kind of animal work whose size does not need to be calculated using power analysis. For instance:

- Colony maintenance and management: mouse strains must be maintained in the facilities so that researchers has enough animals to perform their work (Weichbrod et al, 2017). They must calculate how many mice must be mated, kept, etc. As there is no measurements, comparisons, etc. there is no need to use a statistical tool to calculate this number of animals.
- Producing genotypes: in the era of genetic manipulations, the generation of transgenic, knockout, knockin, conditional animals or specific genotypes, we also need to calculate the number of matings, probability of the offspring having a particular gene or mutation, etc. (Pinkert, 2014). However, no measurements and comparisons are involved and there is no need for a statistical tool.

However, some other experiments seem to fall in this category of experiments in which the sample size does not need to be calculated. However, these experiments have an ambiguous objective. Let's see an example:

- Example 41: A researcher is interested in the neural connection between the retina and specific parts of the brain that are known to be connected in normal animals. The question is whether these connections are modified in any way for animals having a genetic disease that causes visual impairment. To answer this question a fluorescent neural marker will be injected in the retina of the animals, then they will be sacrificed and slices of the optic nerve and the corresponding brain part will be analyzed at the microscope. Histological analysis of slices of normal and diseased animals will be analyzed to compare their differences. The researcher thinks that the sample size of this experiment does not need to be calculated because the experiment is purely qualitative: looking at the histological sections and reporting the observed differences. Concisely, the researcher will check if the presence or absence of visual fiber tracts on both sides of the brain.

However, the qualitative description that the researcher is suggesting lends itself to a quantitative measure. We can compare the proportion of diseased animals in which we observe no visual fiber tracts, only one tract on the left side of the brain, only one tract on the right side of the brain, or both tracts and compare these proportions to the same proportions in normal animals. Actually, if there is no quantitative comparison between both groups, it is hard to conclude anything from the experiment. Readers would not know what to think if we report that on 4 normal animals we have always seen the tract on both sides, and that we have not seen it in 4 diseased animals. We should accompany this statement with a p-value (in this case, a two-tails Fisher's exact test comparing the proportion in both groups gives a p-value of 0.029; the p-value is too close to the 0.05 threshold to be sure that the result is biologically true, especially because we know that the p-value can be very unstable, see Fig. 1.5).

Reality can be even worse imagine that we have seen 4 normal animals with visual tracts on both sides of the brain, and in the diseased animals we have seen only 1 tract (on the left or the right side of the brain). Then, of the 8 possible tracts, we have seen 8 in the normal animals and 4 in the diseased animals. Now the p-value of a two-tails Fisher's exact test comparing the proportion in both groups gives a p-value of 0.077, which would say that we cannot declare the difference significant.

All these hassle could have been avoided if we had designed the experiment from the beginning choosing a number of animals that would allow us to detect differences in proportions of a given size.

## 2.3 Practical issues related to experiment design

### 2.3.1 Undefined goals

Compare these two research objectives:

- Our objective is to study the effect of corticosteroids in osteoporosis.
- Our objective is to determine the effect of different plasma levels of cortisone in the bone density of the tibia of BALB/C mice.

The second objective sounds much more concise and manageable. The same may happen in many research experiments although their appearance sounds profound. Let us consider the following experiment.

- Example 42: Some researchers are interested in the inflammatory response of the intestine when a colitis is induced in the animals. For this purpose, they will measure: 1) macroscopic changes, 2) biochemical and genetic changes, 3) immunohistochemical changes, 4) cytometric changes, and 5) permeability changes after 0, 5, 10 and 15 days after the induction of the colitis. They will use 8 animals per group and they will repeat this analysis in 6 different mouse strains. They plan to analyze the results with a Student's t test. In total, they will need  $6 \times 2 \times 4 \times 8 = 384$  animals.

In the previous description of the experiment there are a number of ambiguities or aspects that could be better handled:

- It is implicit that for each time point there will be a control group so that they will compare for each strain and each time point, the inflammation in the control vs the treatment (colitis) group. This is, in principle, a good design, but it should be explicitly stated.
- The use of a Student's t test at each time point prevents them from attributing different contributions to different factors. In particular, they could have used a 3-way ANOVA analysis with interactions. The generation model would be

$$y_{ijkl} = \mu + \alpha_i^S + \alpha_j^T + \alpha_k^D + \alpha_{ij}^{ST} + \alpha_{ik}^{SD} + \alpha_{jk}^{TD} + \alpha_{ijk}^{STD} + \epsilon_{ijkl}$$

That is, the observation for a given animal is decomposed by the main effects of its strain,  $\alpha_i^S$ , its treatment,  $\alpha_j^T$  (control or colitis), the day of the measurement,  $\alpha_k^D$ , and interactions between strain and treatment,  $\alpha_{ij}^{ST}$ , between strain and day  $\alpha_{ik}^{SD}$ , and between treatment and day,  $\alpha_{jk}^{TD}$ . In this way, we may recognize that some strains have constitutively larger or smaller values of any of the measured variables ( $\alpha_i^S$ ), that the treatment has an overall inflammatory effect ( $\alpha_j^T$ ), although some strains are more or less sensitive to the colitis ( $\alpha_{ij}^{ST}$ ). We may also identify the inflammatory time profile in general ( $\alpha_{jk}^{TD}$ ) and if some strains recover more or less quickly ( $\alpha_{ijk}^{STD}$ ).

In terms of parameters the 3-way ANOVA requires to estimate 48 parameters, the same as the pairwise Student's t-test. However, the full factorial 3-way ANOVA allows interpreting the contribution of each one of the factors as well as their interactions. An important difference between the 3-way ANOVA and the pairwise Student's t-test, is that the former assumes that the variance in all combinations of factors is similar. As we saw in the previous section, this assumption is very difficult to verify with a small number of samples, unless there are very large differences in standard deviation. This situation favours the use of the 3-way ANOVA.

- Although not explicitly stated, using 8 animals per group allows to identify effect sizes between the two groups whose size is 1.8 times the observed standard deviation (see Sec. 4.1.5), disregarding which are the two groups that we are comparing. Moreover, this difference has not been chosen by the researcher, who may not be aware of it, but by the selection of the number of samples.

#### Important remarks

40. When analyzing the response of a variable on multiple factors (strain, control/treatment, and time), a factorial design can reveal much more information (main effects of each of the variables and interactions between pairs or triples of variables) than pairwise group comparisons for the same cost in number of animals.

We may also consider the following experiment.

- Example 43: A researcher is interested in studying the connectivity of neurons in the telencephalon of normal mice and of mice with a gene defect that mimic a rare human disease. In particular they will measure the distribution of a given cell type marked with a fluorophore in both, healthy and diseased, telencephalons. To calculate the sample size, they have obtained data from the literature arriving to the conclusion that the data has a standard deviation of 5%. They will look for differences larger than 10%. A sample size calculator using a two-tails Student's t-test with 90% of statistical power and 95% confidence level yields  $N = 7$  animals per group.

The problem with the description above is that it is unclear how “the distribution of the cells in the telencephalon” will be translated into a continuous variable whose

differences in mean are the ones tested by the Student's t-test. It is also unclear how its standard deviation is expressed as a percentage (Is it a percentage of the mean value? Is it a percentage because the observed number is a percentage itself (e.g., the percentage of the area of the microscope covered by the cells)? Can this number take values outside the 0-100% range? Is it normally distributed?).

**Important remarks**

41. Under the appearance of well-defined objectives and calculations, there can be very ambiguous experiment designs.

### 2.3.2 Longitudinal studies of the same animal

In some experiments we may follow the same animal over a time (for instance, at  $t = 0, 7, 14, 28, 60$  days of treatment). Because different individuals may respond differently to the same treatment and its disease progression may be different, we should account for the contribution of the animal itself to the variability of our observations. In this way, the animal itself could be blocked. The data acquisition model would be

$$y_{ijk} = \mu + \alpha_i^A + \alpha_j^T + \alpha_k^D + \alpha_{jk}^{TD} + \varepsilon_{ijk}$$

that is, the observation is a contribution of the animal itself,  $\alpha_i^A$ , the treatment,  $\alpha_j^T$ , the day of measurement,  $\alpha_k^D$ , and a possible different behaviour of the treatments over time,  $\alpha_{jk}^{TD}$ . Ideally, the contribution of the animals,  $\alpha_i^A$ , should be accounted by a single parameter,  $\sigma_A^2$ , estimated from a random-effects linear model (see Sec. 5.2.8). Alternatively, a fixed-effects model could be used, although at the cost of extra degrees of freedom that have to be borrowed from the residuals, losing statistical power.

The same model would apply if we have to measure the animal at multiple locations, for instance, the visual and the somatosensory cortex of the brain. The location would play the role of measurement day in our previous example.

**Important remarks**

42. When an animal is measured multiple times, the differences between animals can be explicitly accounted. Ideally through a random effect linear model (that only consumes 1 degree of freedom). As these models are not so much known, at least, a fixed-effect linear model can be fitted (needing  $N - 1$  degrees of freedom, where  $N$  is the number of animals).

### 2.3.3 Unnecessarily discretizing continuous variables

Very often the factors affecting a given response are continuous. For instance, we are interested in the immunological response of animals to a viral challenge. We will quantitatively assess the response through the gene expression level of a given protein involved in an immunological signalling pathway. We think that the sex and age of the animals (young=3 months old, or old=20 months old) may influence the response, as

well as the viral dose. We want to measure this response at multiple time points (0, 12, and 48h after challenging).

We need to establish the data generation model that will be used both for the experiment design and its posterior analysis. We may use a factorial design such as (see Sec. 5.1.6 on factorial designs)

$$y_{ijklm} = \mu + \alpha_i^{(sex)} + \alpha_j^{(age)} + \alpha_k^{(dose)} + \alpha_l^{(time)} + \alpha_{im}^{(sex,time)} + \alpha_{jm}^{(age,time)} + \varepsilon_{ijklm}$$

We have included an interaction between sex and time as it is possible that males and females may respond over time differently. The same applies to the age of the animals, maybe younger animals respond differently over time than older ones. The dose will take two values:  $10^4$  or  $10^5$  viral particles per mL. A possible experiment design with this factorial model and 20 animals would look like (note that at this moment we are not making any sample size consideration, as it would distract us from the main argument of this example):

Sex	Age	$\log_{10}(\text{dose})$	Measurement time
Male	Young	4	0
Male	Young	4	0
Male	Old	5	0
Female	Young	5	0
Male	Old	5	0
Female	Old	4	0
Female	Old	4	0
Male	Old	4	12
Male	Old	5	12
Female	Young	4	12
Female	Young	4	12
Female	Young	5	12
Male	Old	4	12
Female	Young	4	48
Male	Young	5	48
Male	Young	4	48
Female	Old	5	48
Female	Old	5	48
Male	Young	5	48
Male	Old	4	48

This table gives us an easy execution plan for our experiment and the analysis is also rather easy through the standard data analysis tools available. However, this design does not allow us to determine what happens in between 3 and 20 months old mice and with other doses different from  $10^4$  or  $10^5$ . Similarly, we do not know how the response changes in between the sampling times. All these variables are intrinsically continuous. However we have discretized them in order to have a simple experimental design. A much more informative model would have used these variables as continuous, that is

(see Sec. 5.1.2 on regression designs)

$$y_m = \mu + \alpha_m^{(sex)} + \beta^{(age)} age_m + \beta^{(dose)} dose_m + \beta^{(time)} time_m + \beta^{(sex,time)} sex_m time_m + \beta^{(age,time)} age_m time_m + \varepsilon_m$$

The term  $\alpha_m^{(sex)}$  is exactly the same as the one of the previous design and accounts for the main effect of sex. However, age, dose and time are now considered as continuous variables, and in our model we have assumed a first-degree polynomial relationship between the response and these variables. However, this is not a hard constraint and any other functional relationship might have been used (see Sec. 5.1.2).

There are two interaction terms of different nature: the interaction sex-time is between a discrete and a continuous variable, while the interaction age-time is between two continuous variables. Both are treated in the same way by adding an extra term that combines both variables in a multiplication. The interaction with a discrete variable is simply modelled by adding an indicator variable that takes the value 0 if the animal is male and 1 if it is female,  $sex_m$ .

In the equation above it is interesting to see that the dependence of the response with time is of the form:

$$y_m = \dots + (\beta^{(time)} + \beta^{(sex,time)} sex_m + \beta^{(age,time)} age_m) time_m$$

The first term accounts for the main effect of the time evolution (which is independent of sex and age) while the other two terms explicitly account for the sex and age differences in time response, respectively.

For the experimental design we should randomly sample the interval of interest (for instance, any time between 0 and 48h, or any dose between  $10^4$  and  $10^5$  viral particles/mL). For the age we may apply a stratified sampling. If we define young mice between 2 and 8 months, and old mice between 8 and 20 months, then a random sampling would place twice as many old animals than young animals (because  $20-8=12$  is twice  $8-2=6$ ). To avoid this unbalance we could take 10 random samples between 2 and 8 months, and 10 random samples between 8 and 20 months.

With all these considerations the experimental design would look like

Sex	Age	$\log_{10}(\text{dose})$	Measurement time
Male	5.6	4.29	6.6
Male	3.6	4.76	7.2
Male	17.9	4.75	9.4
Female	5.9	4.38	11.7
Male	14.5	4.57	12.1
Female	20.0	4.08	12.2
Female	8.9	4.05	12.4
Male	13.3	4.53	16.8
Male	9.3	4.78	16.9
Female	6.1	4.93	22.7
Female	6.5	4.13	26.3
Female	4.7	4.57	26.4
Male	19.5	4.47	28.1
Female	2.5	4.01	29.6
Male	3.4	4.34	39.1
Male	7.5	4.16	39.9
Female	8.1	4.79	40.4
Female	17.3	4.31	44.0
Male	2.9	4.53	44.6
Male	17.8	4.17	46.0

This table implies that for instance, for the first row, by design, we will wait for a male animal to reach 5.6 months-old to inject a dose of  $19500 \approx 10^{4.29}$  viral particles/mL. Then we will measure the gene expression of the gene of interest 6.6h after injection. We may be happy with this design or we may think that this experiment may take too long (waiting for 20 months so that we can measure the last animal). We may want to perform the experiment at once. Then, we may treat age as a covariate, that is, we no longer design it, but we measure it. That is, we take a young (between 2 and 8 months-old), male animal, inject 19500 viral particles/mL and measure gene expression 6.6h after injection. Note that we have not selected a young animal of a specific age, but we annotate its age in our laboratory notebook so that later in the analysis we can estimate what is the contribution of age to the immunological response.

The design above allows us to visualize what is the behavior of our response in between the extremes (2-8 months, 8-20 months,  $10^4$ - $10^5$  dose, 0-48h). However, if we will fit a simple first-degree polynomial on the age, dose and measurement time variables, then the most informative sampling points, the ones that minimize the variance of the polynomial coefficients are the ones described in Sec. 5.2.11. The random sampling is an easy substitute of more sophisticated designs that allow us to explore the dependence of the response as a function of the continuous variables, rather than presuming an *a priori* functional relationship and then deciding the most informative sampling points (as done in Sec. 5.2.11).

Finally, let us extend here the technique to include interactions between discrete and continuous variables when there are more than two levels for the discrete variables. For instance, let us assume that we are interested in the immunological response as a function of the genotype of the animal (we have 3 different genotypes A, B, and C) and,

for simplicity, let us drop the dependence on sex and age. We think that there might be differences in the time response of these three genotypes, then we could use the model

$$y_m = \mu + \alpha_m^{(genotype)} + \beta^{(time)}time_m + \beta^{(A,time)}A_mtime_m + \beta^{(B,time)}B_mtime_m + \varepsilon_m$$

The indicator variables  $A_m$  and  $B_m$  take the values 1 and 0 if the animal is of genotype A or B, respectively. In this way, we can see that the response of the different genotypes would be

$$y_m = \begin{cases} \mu + \alpha_A^{(genotype)} + (\beta^{(time)} + \beta^{(A,time)})time_m + \varepsilon_m & \text{if genotype=A} \\ \mu + \alpha_B^{(genotype)} + (\beta^{(time)} + \beta^{(B,time)})time_m + \varepsilon_m & \text{if genotype=B} \\ \mu + \alpha_C^{(genotype)} + \beta^{(time)}time_m + \varepsilon_m & \text{if genotype=C} \end{cases}$$

#### Important remarks

43. Continuous variables such as dose, time, age, etc. can be modelled through a regression. This has the advantage that allow us to understand the animal response in between the sampling points. The optimal sampling points should be calculated as a function of the range of interest (see Sec. 5.1.7).

Moreover, we can turn an intrinsically discrete analysis into a continuous one. For instance, some ELISA experiments proceed by a series of dillutions and measuring the absorbance of the solution. This gives an absorbance-dillution factor curve. By construction the dillution factor is discrete and many analyses compare the dillution factor at which the absorbance drops below a given threshold. However, we can compute the area under the absorbance curve, which is naturally continuous and make the comparison at the level of this new variable. This is the approach proposed by the absorbance summation technique (Hartman et al, 2018). In this way, we see that by looking for a clever trick we can turn discrete observations into more informative continuous ones.

### 2.3.4 Incomplete/imbalanced factorial designs

In Sec. 5.1.7 we discuss about the necessity of blocks sharing treatments in order to be able to compare the different results. For instance, let us assume that we are comparing 3 different treatments to a control situation. We think that the litter may make a difference. We plan to have 5 individuals per treatment. Then a complete, balanced experimental design would be

	Treatments
Litter 1	C, T <sub>1</sub> , T <sub>2</sub> , T <sub>3</sub>
Litter 2	C, T <sub>1</sub> , T <sub>2</sub> , T <sub>3</sub>
Litter 3	C, T <sub>1</sub> , T <sub>2</sub> , T <sub>3</sub>
Litter 4	C, T <sub>1</sub> , T <sub>2</sub> , T <sub>3</sub>
Litter 5	C, T <sub>1</sub> , T <sub>2</sub> , T <sub>3</sub>

We have 19 degrees of freedom, and we need 3 of them to estimate the treatment effects, and 4 for the litters, meaning that we still have 12 for the residuals (we could have modelled the litter as a random effect, consuming in this way a single degree of freedom). These are not many but they give a statistical power to detect changes about twice the standard deviation.

The litters act as blocks within which all treatments are applied. For some reason, only 3 treatments can be applied per litter. Then, we must use an imbalanced design. In our design above, we used 20 animals, but 20 is not a multiple of 3, we will use now 24 animals and construct an incomplete, but balanced design

	Treatments
Litter 1	$C, T_1, T_2$
Litter 2	$C, T_1, T_3$
Litter 3	$C, T_2, T_3$
Litter 4	$T_1, T_2, T_3$
Litter 5	$C, T_1, T_2$
Litter 6	$C, T_1, T_3$
Litter 7	$C, T_2, T_3$
Litter 8	$T_1, T_2, T_3$

Each treatment appears 6 times. We have 23 degrees of freedom from the experiment, and we need 3 and 7 for the treatments and litter blocks, respectively. So we are in a similar situation to the previous one, with 13 degrees of freedom for the residuals.

Let us consider now a different experimental situation. We are given two compounds,  $A$  and  $B$ , that we use in a standard two-groups comparison. 6 months later we receive other two compounds,  $C$  and  $D$ , that we also test and want to compare to the results of  $A$  and  $B$ . Our experimental situation is

	Treatments
Period 1	$A, B$
Period 2	$C, D$

We discussed in Sec. 5.1.7 that these treatments cannot be compared ( $A$  vs  $C$  or  $D$ ;  $B$  vs  $C$  or  $D$ ). This is true if there is block effect, that is something that depends on the specific period in which the treatments were tested. For instance, the humidity in the first period might be different from the one in the second period. If humidity affects the results from the experiment, then the results of the first period might be systematically shifted up (or down) with respect to the results of the second period making the comparison useless. Still, we may be interested in comparing the 4 treatments. Then, the best solution would be to report the mean and its confidence interval of the four treatments, making the disclaimer that  $A$  and  $B$  were tested on a different period than  $B$  and  $C$  and that there might be block effects. We should not make any hypothesis test between  $A$  and  $C$  or  $D$ , but still the report may be of interest. For instance,  $C$  may cause a much stronger response than  $A$  with such a difference that cannot be explained by the standard blocking variables (humidity, operator, pipettes, or any other experimental tools), which should induce relatively small differences if properly controlled. In any case, if we are very interested in a comparison with  $A$ , we may consider including  $A$  in the second period.

Incomplete designs make the assumption that the mechanism of action in all blocks is present and it is the same. For instance, let us assume that we need to apply 3 treatments (control+2 treatments,  $C, T_1, T_2$ ) during 2 weeks. The experiment is such that every week we can only apply two treatments. We are afraid that there could be important block effects between weeks, so we make the following design

	Treatments
Week 1	$C, T_1$
Week 2	$C, T_2$

According to Sec. 5.1.7, by having the control in common between the two weeks, we can compare the effects of  $T_1$  and  $T_2$ . We are assuming that the mechanism through which the block (week) interacts with the treatment is present in both weeks (different humidity, temperature, operator, chemical batch, etc.).

However, there are situations in which this assumption does not hold. This is particularly true in biological responses. For instance, let us say that the block is sex. Our design would look like

	Treatments
Males	$C, T_1$
Females	$C, T_2$

However,  $T_2$  may affect females through a hormonal signalling pathway that is only present in females. Then, by having the control,  $C$ , in common between males and females, we cannot infer what will be the effect of  $T_2$  on males (which presumably will be none). The same would apply if  $C, T_1$ , and  $T_2$  interact with different proteins in a biochemical pathway, or if they are ligands that bind to different pockets of the same protein. In all these cases, we are violating the assumption that “the mechanism by which treatments and blocks interact is conserved” between blocks.

#### Important remarks

44. Factorial designs assume that the mechanism of action in all levels of a given factor can be extrapolated to the other levels. This assumption may hold in many occasions (week, period, litter, center, researcher, etc.), but it may also not hold (if the mechanism of action is different in the different blocks, e.g., sex or age). This does not mean that sex or age cannot be used in incomplete designs. It means that we should be aware of possible differences that cannot be identified by the factor analysis.

### 2.3.5 Animal housing constraints

When dealing with mice several constraints may appear: 1) males cannot be housed with females; 2) young animals cannot be housed with older animals because they fight each other; 3) animals that are not related by family relationships also fight; ... These housing constraints have to be taken into account when designing our experiment. They can be done in two ways:

- Cages with all treatments: This would be the most straightforward approach. Let us imagine we are interested in comparing the pharmacokinetic properties of 4 formulations of the same compound. We will use 8 individuals per formulation. We may treat sex and age as blocks, gather 4 homogeneous (same sex, age, and family) animals into the same cage and apply one of the formulations to each one of them. We may treat the cage as a block (paying the price of 7 degrees of freedom) or not.

Cage	Sex & Age	Formulations
Cage 1	Male, Young	$F_1, F_2, F_3, F_4$
Cage 2	Male, Young	$F_1, F_2, F_3, F_4$
Cage 3	Male, Old	$F_1, F_2, F_3, F_4$
Cage 4	Male, Old	$F_1, F_2, F_3, F_4$
Cage 5	Female, Young	$F_1, F_2, F_3, F_4$
Cage 6	Female, Young	$F_1, F_2, F_3, F_4$
Cage 7	Female, Old	$F_1, F_2, F_3, F_4$
Cage 8	Female, Old	$F_1, F_2, F_3, F_4$

- Cages with a single treatment: The previous design cannot be applied if there is a cross-talk between treatments within the same cage due to transmission through feces, urine, or any other reason such as dominance of the control animals on the diseased ones. Then, the same cage must have the same treatment.

Cage	Sex & Age	Formulations
Cage 1	Male, Young	$F_1, F_1$
Cage 2	Male, Young	$F_2, F_2$
Cage 3	Male, Young	$F_3, F_3$
Cage 4	Male, Young	$F_4, F_4$
Cage 5	Male, Old	$F_1, F_1$
Cage 6	Male, Old	$F_2, F_2$
Cage 7	Male, Old	$F_3, F_3$
Cage 8	Male, Old	$F_4, F_4$
Cage 9	Female, Young	$F_1, F_1$
Cage 10	Female, Young	$F_2, F_2$
Cage 11	Female, Young	$F_3, F_3$
Cage 12	Female, Young	$F_4, F_4$
Cage 13	Female, Old	$F_1, F_1$
Cage 14	Female, Old	$F_2, F_2$
Cage 15	Female, Old	$F_3, F_3$
Cage 16	Female, Old	$F_4, F_4$

These designs are also useful in virology experiments in which animals within the same cage will infect each other. In this case, we may find useful treating the cage as the experimental unit, with two measurements coming from it that will be averaged to produce a single number for the experimental unit. For instance, we may be interested in three viral concentrations ( $C = 0$ ,  $D_1 = 10^4$ , and  $D_2 =$

$10^5$  viral particles/mL). Then, both animals in the same cage receive the same treatment so that we avoid contagion between animals in the same cage treated differently.

Cage	Sex & Age	Formulations
Cage 1	Male, Young	$C, C$
Cage 2	Male, Young	$C, C$
Cage 3	Male, Young	$D_1, D_1$
Cage 4	Male, Young	$D_1, D_1$
Cage 5	Male, Young	$D_2, D_2$
Cage 6	Male, Young	$D_2, D_2$
Cage 7	Male, Old	$C, C$
Cage 8	Male, Old	$C, C$
Cage 9	Male, Old	$D_1, D_1$
Cage 10	Male, Old	$D_1, D_1$
Cage 11	Male, Old	$D_2, D_2$
Cage 12	Male, Old	$D_2, D_2$
Cage 13	Female, Young	$C, C$
Cage 14	Female, Young	$C, C$
Cage 15	Female, Young	$D_1, D_1$
Cage 16	Female, Young	$D_1, D_1$
Cage 17	Female, Young	$D_2, D_2$
Cage 18	Female, Young	$D_2, D_2$
Cage 19	Female, Old	$C, C$
Cage 20	Female, Old	$C, C$
Cage 21	Female, Old	$D_1, D_1$
Cage 22	Female, Old	$D_1, D_1$
Cage 23	Female, Old	$D_2, D_2$
Cage 24	Female, Old	$D_2, D_2$

In this design we have 23 degrees of freedom in total, that are spent in estimating the effect of the treatments (2), sex (1), and age (1), leaving 19 degrees of freedom available for the residuals.

#### Important remarks

45. The cage may be used as a block (all treatments are applied in the same cage) or not (all animals in the same cage receive the same treatment). Ideally, in the second design the whole cage would be treated as a single experimental unit, instead of each animal of the cage being treated as the experimental unit. The decision to use one or the other design depends on whether we expect interactions of the treatment within the block or not.

## Chapter 3

# Statistical pitfalls

Evolution did not select *Homo sapiens* because we were good at solving statistical problems, the evolutionary pressure was on the solution of other kind of problems. Although in an economical context, the work of Daniel Kahneman and Richard Thaler, Nobel Memorial Prizes in Economy in 2002 and 2017 respectively, showed us how difficult is probability and statistics for us. Both have outreach books in which they summarize part of their work and the interested reader will find in them many surprising and funny situations in which humans are found when dealing with these issues (*Thinking, fast and slow*, D. Kahneman; *Misbehaving: the making of behavioral economics*, R. Thaler). As a consequence, Statistics manages to regularly pervade our intuition. We tend to quickly jump into conclusions from characteristics of a small population, we tend to be overconfident in our own experiments, we tend to see patterns in random data, we do not realize that coincidences are common, and we tend to ignore alternative explanations. Actually, one of the main points in Statistics is to decide when we can generalize to a large population, the observations we have made in a small laboratory sample. Along these chapters we have seen how to fight bias, calculate the sample size that allows this generalization and how to organize these samples so that we avoid influences from nuisance factors.

Once the experiment is carried out, we will analyze its results using also statistical tools. The scope of this book is on the design of the experiment, and not on data analysis. However, there is a non-negligible intersection of ideas related to both data analysis and experimental design. If these ideas are overlooked, then the experiment runs the risk of being spoiled. In this section we collect a number of issues to keep in mind during our statistical analysis, and planning of the research experiment.

The reader interested in general statistical topics normally used in the analysis of biomedical data are referred to the Points of Significance collection of Nature Methods (<https://www.nature.com/collections/qghhqm/pointsofsignificance>), and the Statistics at Square One of BMJ (<https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one>). On the same topic of this book, experiment design, the reader may find interesting the Special Issue of the Institute for Animal Laboratory Research Institute Journal (ILAR J) published in 2014 (<https://academic.oup.com/ilarjournal/issue/55/>

3).

### 3.1 Probability pitfalls

**We are not good at recognizing ambiguously defined probabilities.** When we say that a test for a given disease is 98% accurate, we normally failed to recognize that this statement alone is ambiguous. In a frequentist approach, the probability is defined as the ratio between positive cases and all possible cases. For instance, the probability of being born a male is the ratio between the number of all male newborns and the number of all newborns. When we say that the disease test is 98% accurate, we do not know which the numerator and denominator are. There are four possible interpretations and they all make sense. Simply, the statement “98 % accurate” does not indicate which one of them we are referring to:

- Interpretation 1: Sensitivity. Numerator: Correctly identified disease cases in a group of animals with the disease. Denominator: Number of tested animals (all of them had the disease).
- Interpretation 2: Specificity. Numerator: Correctly identified non-disease cases in a group of animals not having the disease. Denominator: Number of tested animals (none of them had the disease).
- Interpretation 3: Predictive value of positive test. Numerator: Correctly identified disease cases. Denominator: Number of animals whose result with this test was positive.
- Interpretation 4: Predictive value of negative test. Numerator: Correctly identified non-disease cases. Denominator: Number of animals whose result with this test was negative.

It is even more ambiguous when we assign to probabilities the notion of “belief”. The surgeon told the first patient ever having a heart transplant that his probabilities of survival were of 70%. In a frequentist approach (number of successes divided by number of possible cases), this statement has no sense, since there has not been any previous experience. In many research situations, our probabilities resemble more a degree of belief (consequently not supported by accurate measurements of previous data), rather than a true frequentist probability. There is nothing wrong with this approach (very likely, 70% was the expected value of the surgeon considering all his experience in similar patients and what he could infer from other, necessarily different, operations), as long as we recognize its limitations: it is simply a useful way of handling uncertainty, and that probability is much more solidly invoked when it is based on past observations or in well defined models as illustrated in the next paragraph.

**We normally fail to consider the assumptions of probabilities.** We can compute the probability of an event based on a model of how the world is or based on counting positive results and dividing it by the total number of possible outcomes. For instance, for the probability of being born male, we may make the following assumptions: 1)

Each ovum has an X chromosome and none has a Y chromosome; 2) Half the sperm have an X chromosome and the other half have a Y chromosome; 3) Only one sperm will fertilize the ovum; 4) Each sperm has an equal chance of fertilizing the ovum; 5) If the winning sperm has a Y chromosome, then the embryo will be XY (male); 6) If the winning sperm has a X chromosome, then the embryo will be XX (female); 7) Any miscarriage or abortion is equally likely to happen to male or female fetuses. Our prediction with this model is that there is 50% chances of being a male or a female. We have come to this probability reasoning on a model of the world.

However, reality is that in 2012 worldwide, 51.7% of the newborns were male, and 48.3% female. There is something wrong in our model that does not faithfully represent the real world. If we now set the probability of humans being male to 51.7%, because it has been carefully measured experimentally, we are further adopting more assumptions. We assume that this probability does not change: 1) over the years, 2) along the year, 3) across races, 4) and across world regions.

Realizing how we have come to a given probability (model or data based) is important in order to understand the limitations of the predictions based on this probability.

**We tend to confuse conditional probabilities.** Conditional probabilities are normally expressed as  $\text{Prob}\{A|B\}$  and it is understood as the probability of  $A$  occurring when we know that  $B$  has occurred. It is also read as the probability of  $A$  given  $B$ . Our problem is that we find it difficult to distinguish  $\text{Prob}\{A|B\}$  from  $\text{Prob}\{B|A\}$ . In some contexts it is very easy: it is not the same the probability of a boring book (given,  $B$ ) being about Statistics ( $A$ ) and the probability of a Statistics book being boring (I leave to the reader the decision about which one is higher). In some other contexts it is more difficult: the probability that a heroin addict (given,  $B$ ) first used marijuana ( $A$ ) is not the same as the probability of a marijuana user (given) will later become addicted to heroin. And in technical contexts it becomes absolutely incomprehensible: the probability of a study for which the null hypothesis is true (given,  $B$ ) having a p-value smaller than 0.05 ( $A$ ) is not the same as the probability of the null hypothesis being true for a study in which the p-value is smaller than 0.05 (given).

**We do not naturally calculate with conditional probabilities.** We regularly monitor for the presence of a rare disease in our animal house. We have a test that correctly identifies 99% of the infected animals, and incorrectly gives a true positive in 0.2% of the non-diseased animals. There must be something wrong with these numbers, 99% and 0.2%, because they do not add up to 100%.

This intuition is incorrect because they are not complementary probabilities. 99% is the probability of identifying the disease with the test (positive result of the test) knowing that the animal has the disease, while 0.2% is the probability of incorrectly identifying the disease knowing that the animal does not have the disease. These probabilities are formally written as

$$\begin{aligned} &\text{Prob}\{positive|disease\} \\ &\text{Prob}\{positive|healthy\} \end{aligned}$$

It is not true that

$$\text{Prob}\{positive|disease\} + \text{Prob}\{positive|healthy\} = 1$$

which cannot be read as “Given a positive result of the test, for sure, either the animal has the disease or not”. The complements of these probabilities are  $\text{Prob}\{negative|disease\}$  and  $\text{Prob}\{negative|healthy\}$ , respectively, for which

$$\begin{aligned}\text{Prob}\{positive|disease\} + \text{Prob}\{negative|disease\} &= 1 \\ \text{Prob}\{positive|healthy\} + \text{Prob}\{negative|healthy\} &= 1\end{aligned}$$

which can be read as “Given an animal with the disease, for sure, either the test is positive or negative” (the same for a healthy animal).

**We do not naturally do Bayesian calculations.** If the test of the previous example is positive for one of the animals in our laboratory, what is the probability that it actually has the disease? We know that this disease affects only 0.1% of the animals.

The correct answer to this question is given by Bayesian theorem:

$$\text{Prob}\{disease|positive\} = \frac{\text{Prob}\{positive|disease\}\text{Prob}\{disease\}}{\text{Prob}\{positive\}}$$

The probability of having a positive result

$$\begin{aligned}\text{Prob}\{positive\} &= \text{Prob}\{positive|disease\}\text{Prob}\{disease\} \\ &\quad + \text{Prob}\{positive|healthy\}\text{Prob}\{healthy\} \\ &= 0.99 \cdot 0.001 + 0.002 \cdot 0.999 = 0.003\end{aligned}$$

Substituting back into the Bayesian formula, we have

$$\text{Prob}\{disease|positive\} = \frac{0.99 \cdot 0.001}{0.003} = 0.331$$

That is, we have a very accurate test (only failing to detect 1% of the diseased animals and with very few false positives), but if the test is positive, the probability of actually having the disease is less than 1/3. The problem is that the disease is rare and because of that there are more false positives ( $\text{Prob}\{False\ positive\} = \text{Prob}\{positive|healthy\}\text{Prob}\{healthy\} = 0.001 \cdot 0.999 = 0.001998$ ) than true positives ( $\text{Prob}\{True\ positive\} = \text{Prob}\{positive|disease\}\text{Prob}\{disease\} = 0.99 \cdot 0.001 = 0.00099$ ).

**We are not good at recognizing Bayesian setups.** One third of the laboratory accidents happen to 1st year Ph.D. students. Consequently, it seems that 1st year Ph.D. students are more careful in their laboratory handling than their more experienced colleagues that are responsible for two thirds of the laboratory accidents. But we tend to forget that 1st year Ph.D. students are only 5% of the researchers in the laboratory. 1/3 is the probability of being in 1st year knowing that there has been an accident ( $\text{Prob}\{1st|Acc\}$ ). How reliable is a 1st year Ph.D. student in the laboratory is not given by this probability, but by  $\text{Prob}\{Acc|1st\}$ . With the data available we cannot calculate

this latter probability, but we can calculate the odds ratio of a laboratory accident between 1st year Ph.D. students and the rest of researchers in the laboratory. For doing so, we exploit Bayes rule

$$\text{Prob}\{1st|Acc\} = \frac{\text{Prob}\{Acc|1st\}\text{Prob}\{1st\}}{\text{Prob}\{Acc|1st\}\text{Prob}\{1st\} + \text{Prob}\{Acc|rest\}\text{Prob}\{rest\}}$$

which can be transformed into

$$\frac{\text{Prob}\{Acc|1st\}}{\text{Prob}\{Acc|rest\}} = \frac{\text{Prob}\{1st|Acc\}}{1 - \text{Prob}\{1st|Acc\}} \left( \frac{1}{\text{Prob}\{1st\}} - 1 \right)$$

Substituting the data known, we get

$$\frac{\text{Prob}\{Acc|1st\}}{\text{Prob}\{Acc|rest\}} = \frac{1/3}{1 - 1/3} \left( \frac{1}{0.05} - 1 \right) = 9.5$$

That is, 1st year Ph.D. students have a risk of laboratory accident that is 9.5 times larger than that for the rest of researchers.

## 3.2 Data analysis pitfalls

**We get confused by variance and subpopulations.** We tend to be overwhelmed by the biological variability observed in some populations. This variability is also increased if there exist several subpopulations within the whole populations with very different characteristics. We tend to constrain statistical analysis to very limited and homogeneous experimental conditions, we think that we cannot “merge” results from different replications of the same experiment because they are “too different”. But along this reasoning we have forgotten a fundamental point of experimental research: the validity of our results. If we are developing a new vaccine, and its protection effects can only be shown in very narrow experimental conditions, then our vaccine cannot be used in a general population. Actually, that is the whole point of statistical analysis: can we generalize the results observed in a small sample of individuals to the whole population, or at least, to a subgroup of it with more homogeneous characteristics? The key assumption of the statistical analysis is that the individuals studied in our experiment are representative (random samples) from the whole population (or a part of it). If this is true, then population variance can be compensated by a larger sample size. If this is not true, then we say that our results are biased. As we saw in Secs. 1 and 1.3, there can be several sources of bias. Bias invalidates our generalization capability. In Sec. 1.4.4 we gave an example of an invalid analysis due to the presence of subpopulations and an incorrect randomization. The use of all the information present in the experiment, very often in a graphical way, will allow us to understand the relationships between the different variables and, possibly, reduce the number of animals in future experiments by a better understanding of their behaviour.

In very few cases, we need to analyze data with no variance. This could be the case for instance if we measure the time that an animal takes to perform a given task. We have an upper limit beyond which we stop the experiment, and in this particular case,

all the animals reached that limit. The appropriate tool to analyze this data is through a survival analysis with censored data. The censoring will handle correctly the lack of variability in the dataset. In any case, the example just described should be analyzed with survival analysis.

**We misunderstand the meaning of a confidence interval.** Instead of a point estimate, it is much better reporting a confidence interval (CI). For instance, instead of saying that the survival probability after 6 months is 79%, it is much more informative to say that the 95% confidence interval of the survival probability after 6 months is [64, 89]%. The true survival proportion lies or lies not in the 95% CI, but there is no way to know if it does or not. If we repeat the experiment (calculating the CI) many times, in 95% of the occasions, our CI contain the true survival proportion (although we do not know which ones). Actually, the confidence is on our procedure to construct intervals, not about this particular interval. 95% is, consequently, the probability that our CI contains the true proportion. There is nothing special about 95% (except tradition). If the true parameter is outside our CI, it is due to bad luck with our samples (sampling error). This occurs in 5% of the cases.

95% is not the probability that the true proportion is in our CI (note the difference between this latter statement and that our CI contains the true proportion with probability 95%). Once we have performed the experiment, the true proportion is or is not inside the CI, it is no longer a matter of probability. The 95% probability relates to the construction procedure of CIs, not to a specific CI. A 95% CI does not mean that 95% of the sample data falls within this interval. A 95% CI does not mean that with probability 95% if we repeat the experiment, the estimated proportion falls within this interval. It does not mean, either, that 95% of the population has this survival proportion.

**We misunderstand the meaning of a p-value.** If we compare two groups (treatment and control) and we get a p-value of 0.03. This means that ...

- If the two population means were identical (null hypothesis), there is a 3% chance of observing a difference as large as you observed (or larger).
- Random sampling from identical populations would lead to a difference smaller than what you observed in 97% of the experiments, and larger than you observed in 3% of the experiments.

and it does not mean that ...

- There is a 97% chance that there is a real difference between the two populations and 3% chance that the difference is a random coincidence.
- The p-value is the probability that the result is due to sampling error.
- The p-value is the probability that the null hypothesis is true.
- The probability that the alternative hypothesis is true is not 1-p-value.
- The probability that the experiment will hold up when repeated is not 1-p-value.

- A high p-value does not prove that the null hypothesis is true.

There are also some common mistakes related to the use of p-values:

- **Stargazing:** Considering results in a paper only important if they have 1, 2, 3, ... stars. p-values are not as reproducible as confidence intervals (see Sec. 1.5.2), and they only mean that the result is not generated under the null hypothesis, not that the result is relevant.
- **Significance is not relevance:** Being statistically significant does not mean that the result is relevant, because the difference between the treatment and control groups may be too small to be useful, for instance. The following two examples illustrate this idea.
  1. We compare the responding proportion in a control and treatment group. We report the sample size, the proportion of responding animals in the control and treatment groups, the p-value, and the 95% confidence interval

Sample size per group	Control	Treatment	pval	CI 95%
10	10%	80.0%	0.006	[44.39,97.48]%
100	10%	26.0%	0.006	[17.74,35.73]%
1000	10%	14.1%	0.006	[12.00,16.41]%
10000	10%	11.2%	0.006	[10.59,11.83]%

They all have the same p-value, but their relevance are rather different (e.g., the last one is seldom interesting, the effect is too small, while the first one is rather interesting despite its small sample size that translated into a large uncertainty about the true proportion of respondents). Conversely, if the result is not statistically significant, it is a warning on its biological relevance (since we cannot discard that we have observed these differences just by chance, the null hypothesis cannot be rejected).

2. In Fig. 3.1 top we show a possible result for an experiment. All statistical measures (parametric or non-parametric difference between means or distributions) would indicate that there is a highly significant difference between the two groups. However, we see in Fig. 3.1 bottom that they overlap too much to be of practical relevance. The treatment implies a slight shift to the right of the animal response, which is correctly identified by the statistical tools. But it is the responsibility of the researcher to decide whether this difference is important in real life, or even, if it exists, it is too small to be relevant. Unfortunately, Statistics cannot give any number that capture the importance of a statistically significant finding. This importance depends on the physiological meaning of the underlying variables and their impact in the quality of life of the animals.
- The rejection of the null hypothesis does not automatically make the alternative hypothesis true. For instance, when the null hypothesis states that the mean of the control group is equal to the mean of the treatment group and we reject the null

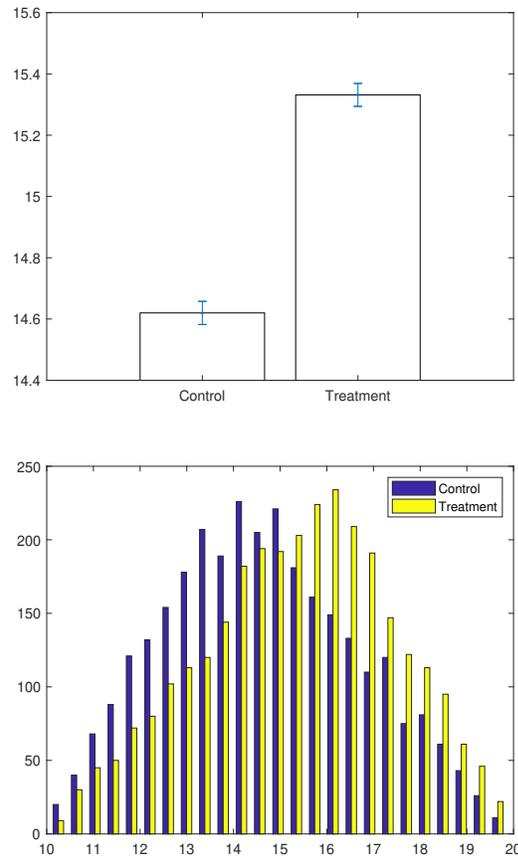


Figure 3.1: Example of a comparison between two experimental situations in which all statistical measures indicate a highly significant difference, but whose practical relevance may not be too important. The top figure represents a typical bar plot with the mean and its SEM (Standard error of the mean) of the results from the two experiments and the bottom figure represents their histogram.

hypothesis using a Student's t-test, we are also hypothesizing that the observations deviate normally from the means of their respective groups. If the p-value is very small it may mean that the two means are certainly different or that the data does not follow a normal distribution around their respective means. There are at least two conditions for not rejecting the null hypothesis: having the same mean and following a Gaussian distribution. We reject the null hypothesis if the observations do not meet any of the two conditions. But we cannot automatically conclude that the two means are different because the null hypothesis has been rejected.

- p-hacking to obtain significance: Trying different hypothesis tests to see if one of them proves to be significant, dynamic sample size (adding more and more data until the result is significant), taking subsets of the data on which the difference is significant, playing with the definition of outliers, changing from a two-sided hypothesis to a one-sided, or preprocess the data in multiple ways. In general, we cannot change our data analysis plan after seeing the results of the experiment. This is called *data snooping*. [Simmons et al \(2011\)](#) argue that part of the problem is that we currently measure many variables and are free to choose which ones to present in the article. [Simonsohn et al \(2014\)](#) and [Head et al \(2015\)](#) tried to quantify the extent of p-hacking in publications reaching the conclusion that it increases as the reported p-value approaches the threshold 0.05, very likely meaning that it might have originally been above 0.05 and that, by probably well intended purposes, the researchers made an effort to find the analysis/data-selection combination that turned the data just significant.

[Nieuwenhuis et al \(2011\)](#) shows a number of pitfalls in the reporting of statistical tests, especially in the use of the p-value and the selection of the appropriate statistical test.

One of the measures that have been proposed to prevent p-hacking and the selective reporting of results is the preregistration of all animal studies ([Bert et al, 2019](#); [van der Naald et al, 2021](#)). However, as of today, preregistration of animal studies is hardly known and most studies are not registered. This contrasts with the situation in clinical studies, where all of the studies have to be preregistered.

As a final comment, we cannot be fundamentalist of not testing any comparison that was not plan from the beginning. We are free to do so in order to discover interesting relationships that we might not have thought when we designed the experiment. However, we cannot stop after the first statistically significant result. This exploration has given us a suggestive new research line. We should now design a new experiment in which we will confirm our finding.

- The p-value has a high variance. The p-value is calculated from the result of an experiment whose observations are random by nature and, consequently, the p-value is also random. Additionally, there is a highly non-linear relationship between the effect observed and the p-value. [Fig. 1.5](#) shows how unstable the p-value can be. Its dynamic range goes over 8 orders of magnitude.

- Post-hoc power analysis is the estimation of the statistical power once the experiment has been performed. We have observed some effect size, and now we calculate what would be the statistical power if the true underlying effect size was the one observed. Unfortunately, post-hoc power is simply another way of reporting the p-value. There is a close relationship between the observed power and the observed p-value (Hoenig and Heisey, 2001). If we want to look at our experiment retrospectively, we should better look at the confidence interval. On the other side, post-hoc power analysis is useful as a prospective tool to design new experiments (Hoenig and Heisey, 2001).
- The p-value cannot substitute biological understanding, we still need to use our biological knowledge to make sense of the p-value. For instance, we are exploring whether a given antibody that works on tumor cell cultures helps to reduce the size of the same kind of tumor in vivo. If we get a p-value of 0.03, or even 0.07, we would understand that the antibody is also working in vivo, as it does in vitro. The fact that the p-value is close to the significance threshold (0.05) is probably caused by a small sample size that resulted in a low statistical power considering the effect size and the size of the fluctuations of the observations (biological and measurement variability). As a negative control of the previous experiment, we also treat with an analgesic instead of the tumor antibody. When compared to the vehicle, the analgesic also reaches a p-value of 0.03. Given our understanding of Biology, this low p-value is simply a false positive caused by the possibility of rejecting the null hypothesis when this is true (what happens with probability 0.05). However, if the analgesic causes a difference with respect to the vehicle whose p-value is  $10^{-14}$ , then we should revise our biological knowledge and investigate why the analgesic is causing such a significant difference.
- Very often we study the effect of a treatment in multiple ways: survival time of the animal, tumor size over time, overall clinical score, gene expressions of biomarkers, immunohistochemical analyses, etc. We still only have a control and treatment group, but we measure their difference in multiple ways. For each one of these ways we will perform a hypothesis test and associate a p-value to each comparison. However, there is no way by which we can combine the different p-values into a single number to take an “automated” decision. It must be us who must consider all the results at hand and finally decide whether the treatment is effective and worthy to pursue further. In a way, we may make the following analogy. We can evaluate the economic health of a country by measuring the unemployment rate, the GDP, the GDP per capita, the trading balance between imports and exports, ... Each one of these perspectives give evidence about the economical phenomenon we are interested in. But, in the end, it must be us who, at the sight of the different evidences, must decide if the country is performing economically well or not. There is no way that we can combine all this information into a single number that summarizes the economical health of the country in a “statistically correct” way. Although we may combine different information sources in many ways, none of them is necessarily the right one, and definitely,

no numerical combination is a substitute of our understanding of the underlying processes.

**We abuse of the p-value as a proxy for relevant research.** In the recent years there has been a strong debate about “the reign of the p-value” (Lakens, 2015; Karpen, 2017; Halsey, 2019; Price et al, 2020; Smith, 2020). Note that the debate is not about biologically relevant results that are discarded by statistical tests rejecting them for being non-significant (p-value above 0.05). Actually, if a result is statistically non-significant, it may be pointing in an interesting biological direction, but we still do not have enough evidence to claim that there is a true biological effect and further investigation is required. Very often these biological interesting, but statistically non-significant results are caused by an insufficient statistical power (that is, a low sample size for the detecting the effect size of the phenomenon we are interested in) or an incorrect statistical analysis that leaves out some useful information.

The p-value is criticized for the following reasons:

- It is difficult to understand: in a previous paragraph we discussed about the correct interpretation of the p-value as the probability of observing this experimental evidence when the null hypothesis is true,  $\text{Prob}\{Data|H_0\}$ . However, what the researcher would like to calculate is the probability of the null hypothesis being true at the sight of the experimental evidence  $\text{Prob}\{H_0|Data\}$  (Cohen (1994); see Sec. 3.1). Additionally, many biomedical textbooks with which biomedical researchers are trained define the p-value in an ambiguous or incorrect way (Price et al, 2020). The problem with the rejection of the null hypothesis comes from the following line of reasoning (Cohen, 1994):

1. If  $H_0$  is true, then this result would probably not occur.
2. This result has occurred.
3. Then,  $H_0$  is probably not true.

However, this reasoning is identical to

1. If a person is an American, then he is probably not a member of Congress.
2. This person is a member of Congress.
3. Then, this person is probably not an American.

The problem of wanting to compute  $\text{Prob}\{H_0|Data\}$  is that, according to Bayes’ theorem, we need the prior probability of  $\text{Prob}\{H_0\}$ , which we do not have.

- If we use a large enough sample, the p-value will always be significant (e.g., the test will reject the hypothesis that the means of the control and treatment groups are equal). Although, being true, we may be looking at the 5th decimal of the mean. It is true that the 5th decimal of the mean is not equal in both groups, but that does not make the two results, control and treatment, relevantly different.

- Its extreme instability as shown in Fig. 1.5 makes it a bad arbiter with respect to whether a finding is relevant and whether it deserves to be published. The effect size and its confidence interval is proposed as a much more stable report of the outcome of an experiment (Halsey, 2019; Smith, 2020). And there are methods based on bootstrap resampling that help us to construct these confidence intervals (Ho et al, 2019), even online (<https://www.estimationstats.com>).
- The threshold of 0.05 shows too weak evidence against the null hypothesis. This threshold implies that 1 in 20 experiments in which the differences are not significant, they will be declared significant. More stringent thresholds like 0.01, 0.005, or 0.001 are advocated. The use of the False Discovery Rate (FDR) is also a very useful way of fighting the inflation of the Type I error rate (False Positives). The FDR is the probability of a positive test (i.e., the null hypothesis is rejected) truly corresponding to a false positive. The FDR is routinely used in experiments in which many hypothesis tests are performed simultaneously (like in proteomics, genomics, or drug screening). When a single test is performed, we may calculate it as a reference value for giving more or less credibility to the result of our experiment. For doing so, we need to presume an *a priori* probability of the null hypothesis being true (see Sec. 1.5.3 for an example). Let us assume that  $p_0$  is the probability of the null hypothesis being true, while  $p_1 = 1 - p_0$  is the probability of the alternative hypothesis being true. Then, given the p-value,  $p$ , the False Positive rate ( $\alpha$ , also called confidence level) and the False Negative rate ( $\beta$ , the complementary of the statistical power), then the FDR can be calculated as (Vidgen and Yasseri, 2016):

$$FDR = \frac{p_0\alpha}{(1-p_0)(1-\beta) + p_0\alpha} \quad (3.1)$$

Typical values for  $\alpha$  and  $\beta$  are 0.05 and 0.1, respectively. Let us assume that  $p_0 = 0.9$ , that is only half of our tested treatments truly makes a difference ( $p_0 = 0.5$ ). Then, the FDR is

$$FDR = \frac{0.5 \times 0.05}{0.5 \times 0.1 + 0.9 \times 0.05} = 1/3$$

The probability of a positive result being a false positive is 33%. Probably, this is still too optimistic. If only 1 out of 10 tested treatments makes a difference, the probability of a positive result being a false positive goes up to 82%. If we lower the confidence level to  $\alpha = 0.001$ , then the FDR even in the adverse situation of  $p_0 = 0.9$  goes down to less than 9%.

We may use our p-value instead of  $\alpha$  in Eq. 3.1 (Halsey, 2019) to get an idea of how reliable our positive result is. For instance, if the p-value is 0.03 when half of our treatments result in a positive difference ( $p_0 = 0.5$ ), the probability of being a false positive is 23%. However, if the p-value is 0.0052 or smaller, then the same probability is at most 5%.

- It hides bad scientific practices such as attempting a study multiple times and only reporting the one in which the p-value is significant; analyzing many different variables and reporting only the ones whose p-value is significant; dropping

outliers or changing the analysis plan once the data has been observed; splitting, merging, or transforming the data until a significant result is obtained; conducting hypothesis tests during the data collection and stopping the experiment as soon as a significant result is achieved. All these are bad statistical practices and, very likely, they are carried out by ignorance and not by bad faith with the aim of cheating. [Lakens \(2015\)](#) argues that the observed increase in the recent years of papers in which the p-value is between 0.041 and 0.049 is caused by publication bias and the file drawer problem (experiments whose results are not significant, at the confidence level of 0.05, are never published).

- The sample size is typically kept as low as possible, and often too low for the effect size we want to detect. Let us consider an experiment for which our sample size is calculated to give a statistical power of 80% ( $\beta = 0.2$ ). Let us assume that the alternative hypothesis is true, then the following table shows the probability of rejecting the null hypothesis 0, 1, or 2 times:

No. Rejections	Probability
0	$\beta^2 = 0.04$
1	$2\beta(1 - \beta) = 0.32$
2	$(1 - \beta)^2 = 0.64$

[Button et al \(2013\)](#) argues that the median power in neuroscience articles is between 8-31% with its obvious implications of lack of reproducibility and ethical consequences of an inefficient experimentation.

Despite all problems of hypothesis testing and null hypothesis rejection, there is not any objective mechanistic ritual that can replace it ([Cohen, 1994](#)). The best we can do is to be more strict on the confidence level (e.g.  $\alpha = 0.001$ ), report effect sizes, their confidence intervals, and use our brains to determine whether the data supports our scientific claims.

**We do not know how to interpret not significant results.** Two groups of pregnant women:

- One of the groups received routine ultrasound twice during pregnancy. In 4.98% (=383/7685) of the cases, an adverse outcome was detected.
- The other group received ultrasound only when indicated by clinical reasons. In 4.91% (=373/7596) of the cases, an adverse outcome was detected.

The null hypothesis is that the risk of adverse outcome is the same in both groups. The relative risk is 1.01 (=4.98/4.91) and has a 95% confidence interval (CI) [0.88,1.17] and the p-value is 0.86. There are three possible interpretations and there is no way to decide which is correct:

1. The CI contains 1. Routine ultrasounds are not helpful nor harmful. They could be skipped.

2. The CI is compatible with a relative risk of 0.88, that is there is a 12% reduction in the risk of adverse outcome by routine use of ultrasounds.
3. The CI is compatible with a relative risk of 1.17, that is there is an increase of 17% in the risk of adverse outcome. May ultrasounds be harmful to the fetus?

**We do not realize the assumptions made by statistical analysis.** Most statistical analyses commonly used in research assume that:

- The samples in our experiment are representative of a larger population. There is no bias as those discussed in Secs. 1 and 1.3. The conclusions we draw from our data can only be extrapolated to a population for which our sample was representative. For instance, the survival proportion after 6 months for a given disease depends on the state-of-art treatments at that moment, the clinical cares given to the patient, ... changes in these variables will induce changes in the survival proportion.
- Samples are independent, and as we saw in Sec. 1.2 there are obvious and not so obvious ways of breaking this independence.
- Data is accurate. Beside obvious errors as mistypes, there are more subtle ways of breaking this assumption.
  - Counting a specific kind of cell in a microscopy field is sometimes undefined because we have doubts of whether a cell is really of the type of interest.
  - If we are interested in studying the survival proportion after 6 months for a given treatment, we may make more efforts to keep a treated individual alive from 5 months to 6 months. If it dies after 6 months and 1 day, it counts as a survival (and, consequently, a success) for our treatment.
  - For survival analysis, the entry criteria does not change over time. For instance, it may be the detection of the first metastasis. But the acquisition of new equipment in the middle of the study allows us to detect earlier these metastases.
  - The same happens with the end point, if we are interested in cancer deaths, but our diseased animals die from an infection totally unrelated to cancer, do we count them for survival? Counting them or not counting, both make sense, but the decision has to be taken before performing the experiment. In these cases, it is also recommended making the survival analysis twice (counting and not counting them) and checking whether the two results significantly differ.
- There are no outliers. A sample is an outlier if it comes from
  - Invalid data (transposed digits, shifted decimal point, sensor blackout, ...).
  - Experimental mistake (bad pipetting, a voltage spike, a hole in a filter).

It is not an outlier if it comes from

- Random chance (just by chance some values are larger/smaller than rest).
- Biological diversity (the population is really variable).
- Invalid assumption (we assume it is normal, but it is log-normal).

Removing data because it does not fit our “expectations” is cheating. But, leaving outliers may lead to invalid results, it is another way of “cheating” (see Fig. 3.2). We do not cheat when the decision to remove an outlier is based on rules and methods established before the data was collected. In this regard, the systematic use of robust statistics (statistical procedures specifically designed to be robust to extreme values) may help.

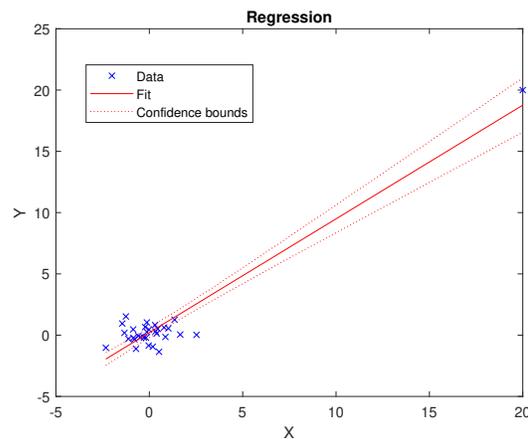


Figure 3.2: Example of a dataset for which leaving an outlier incorrectly results into an artificially inflated coefficient of determination,  $R^2 = 0.91$ , giving the impression of a clear relationship between  $X$  and  $Y$ . If we remove this outlier, there is no relationship between  $X$  and  $Y$ .

Linear models and ANOVA-like models assume that the variance is homocedastic (that is, the variance of the residuals of the different combinations of factors and blocks is the same independently of the specific combination). This assumption is sometimes violated, especially by the negative or the positive control groups. There are ways to explicitly account for the differences in variance (e.g., Welch’s ANOVA). If the classical ANOVA is used, which assumes equal variance, then these control groups may be removed from the analysis.

**We fail to realize that non-parametric tests are not assumption free.** Nonparametric methods have several advantages or benefits over parametric methods: they may be used on all types of data including nominal, ordinal, interval and ratio scaled; they make fewer and less stringent assumptions than their parametric counterparts; they may be

almost as powerful as the corresponding parametric procedure when the assumptions of the latter are met and when this is not the case, they are generally more powerful. This has led to their being used as a first resort when there are any problems with data distribution, such as non-normality. Note, however, that there is a restricted range of non-parametric equivalents of parametric tests, and while there are very efficient and effective equivalents for simple comparisons, there are no such simple equivalents for more complicated designs commonly encountered in ANOVA. Note also that, while the non-parametric tests may be distribution-free, they are not assumption-free.

Consider using a parametric method when:

- The assumptions for the population probability distribution hold true.
- The sample size is large enough for the central limit theorem to lead to normality of averages.
- The data is non-normal but can be transformed.

Consider using a non-parametric method if the data is

- Distinctly non-normal and cannot be transformed.
- From a sample that is too small for the central limit theorem to lead to normality of averages.
- From a distribution not covered by parametric methods.
- From an unknown distribution.
- Nominal or ordinal.

It is generally believed that non-parametric tests are immune to data assumption violations and the presence of outliers. While non-parametric methods require no assumptions about the population probability distribution functions, they are based on some of the same assumptions as parametric methods, such as randomness and independence of the samples.

Equally important is that many non-parametric tests are sensitive to the shape of the populations from which the samples are drawn. For example, the 1-sample Wilcoxon test can be used when the team is unsure of the population's distribution, but the distribution is assumed to be symmetrical. For the Kruskal-Wallis test, samples must be from populations with similar shapes and equal variances. Problems with data that lead to non-normality, for example low averages caused by treatment leading to data "piling up" against the lower limit, will typically lead to differences in both shape and variance of the distributions, which may invalidate the assumptions of non-parametric tests.

Table 3.1 contains the most commonly used parametric tests, their non-parametric equivalents and the assumptions that must be met before the non-parametric test can be used.

**We fail to use the correct statistical distribution.** Not all data follow the Gaussian distribution, which is one of the underlying assumption behind t-tests, ANOVA tests,  $\chi^2$  tests, etc. As we saw in Sec. 1.4.5, there are many situations in which log-normal is the appropriate distribution. For counting data we have other distributions like Poisson,

Parametric test	Non-parametric equivalent	Non-parametric data assumptions
1-Sample z-test or t-test	1-Sample sign test	Bivariate random variables are mutually independent. The measurement scale is at least ordinal.
	1-Sample Wilcoxon test	Random, independent sample is from a population with a symmetric distribution.
2-Sample t-test	Mann-Whitney test	Mutually independent random samples from two populations that have the same shape, whose variances are equal and a scale that is ordinal.
Paired t-test	Paired Wilcoxon test	Random, independent samples are from populations with symmetric distributions.
1-Way Analysis of Variance (ANOVA)	Kruskal-Wallis test	Random, mutually independent samples are from populations whose distribution functions have the same shape, equal variances. Each sample consists of five or more measures. Kruskal-Wallis is more powerful than Mood's for data from many distributions, but less robust against outliers.
	Mood's median test	Independent random samples from population distributions that have the same shape. Mood's median test is robust against outliers.
2-Way ANOVA	Friedman test	Responses for each of the block-treatments are from populations whose distribution functions have the same shape and equal variances. Treatments must be assigned within the blocks.

Table 3.1: Table of some the most used parametric tests, their non-parametric counterparts, and some of the assumptions of the non-parametric tests.

binomial or negative binomial. If we do not know the distribution of our data, we may resort to non-parametric tests. They are less statistically powerful (because they use less *a priori* knowledge), but they do not assume any particular distribution (although they still do the standard assumptions of a representative and independent samples, and accurate data). The decision on parametric or non-parametric is most important with small sample sizes, but with small sample sizes most normality tests cannot show that the data is not Gaussian (high p-values), they simply do not have enough evidence, due to the small sample size, to show that the data is not Gaussian. This gives a false confidence on the use of parametric analysis, because failing to reject Gaussianity does not make the data Gaussian.

Particularly important to this issue is the use of proportions as continuous data. There are situations in which our raw measurements are proportions. For instance, we treat animals in various ways and measure the proportion of cells expressing a given protein. For each animal, we may analyze hundreds of thousands of cells, and the proportion of cells expressing this protein is almost a continuous variable. It does not make sense to use the standard tools for discrete variables (binomial, Poisson, ...) which are aimed at dealing with a few observations (not hundreds of thousands). Instead, we should use the standard tools aimed for continuous variables (ANOVA, linear models, Student's t-tests, ...). But this tools assume Gaussianity of the underlying values, which very likely are not the case if the proportions are close to the 0 or 1 limits. It is traditional to transform these proportions with an arcsine or logit transformation, which makes these proportions to be more Gaussianly distributed. In recent years, there has been a movement favouring logit (see [Warton and Hui \(2011\)](#) and references therein).

**We incorrectly report variability.** Our data is variable, we never observe exactly the same value in all animals, and it is common to report not only the mean of our observations, but also some notion of variability. We have to distinguish between the variability of our observations and the variability of our estimate of the mean. The variability of the observations is the variability inherent to the animals we are studying. Assume we study  $N = 10$  animals, each one with an observation  $x_i$  ( $i = 1, 2, \dots, 10$ ). Our estimate of the mean would be

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

The estimate itself is another random variable (if we take a different group of 10 animals, our estimate of the mean will be different due to the random sampling). If the standard deviation of the observations is  $\sigma_x$ , the standard deviation of the estimate of the mean is  $\sigma_x/\sqrt{N}$ , independently of the distribution of the observations. This standard deviation is called the standard error of the mean (SEM).

If we want to report the variability of the observations we can give an interval based on percentiles of their distribution (for instance, the interval defined by the 2.5% and 97.5% percentiles). This interval is not a confidence interval. Some researchers report the mean and standard deviation (SD) of our observations ( $\hat{\sigma}_x$ ). Reporting the standard deviation has the disadvantage that makes the reader assume that the distribution of observations is symmetric around the mean, and that we only know an approximate range

of the observations for the Gaussian (which we know that in practice is approximately limited to  $\mu \pm 3\sigma$ ).

If we report the CI associated to the mean estimate or the SEM, we are only reporting our uncertainty about our estimate of the mean. That is, the variability of the estimate of the mean as a random variable. As the number of samples in our experiment grows we reduce the uncertainty about our estimate of the mean. If we simply report this uncertainty, we may give the false impression of low variability samples, when what we have is low variability estimates of the mean of the samples.

**We misuse correlation.** Pearson's correlation is a very useful tool to identify the association of two variables. Intuitively, we want to measure how much information do we gain on the variable  $Y$  given the value of another variable  $X$ . Correlation is between -1 and 1. The closest its absolute value is to zero, the less information we have. However, we do not realize that these statements are true only for linear dependencies. Correlation can only account for linear relationships between two continuous variables. It is a tool designed to capture relationships like the ones in the top row of Fig. 3.3, in which both  $X$  and  $Y$  are continuous, random variables. If there is a perfect linear relationship between both variables (second row of Fig. 3.3), then the correlation is either -1, 0 or 1, depending on the sign of the slope of  $Y$  over  $X$ . However, as illustrated in the third and fourth rows of Fig. 3.3, Pearson's correlation has the same value for datasets with very different characteristics. For instance, it cannot capture non-linear relationships as the ones in the third row. A better suited tool for this is the coefficient of determination ( $R^2$ ), which is defined for any kind of regression (linear or non-linear), or the mutual information (which is well-defined for any pair of random variables, continuous or not). The fourth row shows that a high correlation does not necessarily imply a strong linear relationship between two variables. Correlation is easily fooled by the presence of outliers and non-linear relationships.

Additionally, we tend to interpret correlation as causation, for instance if the expression of two genes, A and B, are highly correlated we tend to interpret one as a cause of the other. But, this relationship may not be necessarily so: both genes may be caused by a common gene C that we have not measured. For instance, the budget spent on ice-creams is highly correlated (negatively) with the budget spent on warm clothes. The reason is not that if we stop buying ice creams, then we have more spare money that we can use to buy warm clothes. There is a common cause, summer, that makes both variables to be highly and negatively correlated. This same effect occurs if we introduce common information in the variables being correlated. For instance, for a number of villages we may measure the number of babies, storks and women. Then we construct the variables  $X = \text{Babies/Storks}$  and  $Y = \text{Women/Storks}$ , that is the number of babies and women per stork. Variables  $X$  and  $Y$  are highly correlated, but not because storks bring babies to women, but because the same information (the number of storks) is seen by both variables  $X$  and  $Y$ .

We also misuse the correlation coefficient when we use it to measure the association between two discrete variables or one discrete and one continuous variables. In these cases it is better to use other tools:

- For one continuous and one discrete variables: we may use ANOVA using the

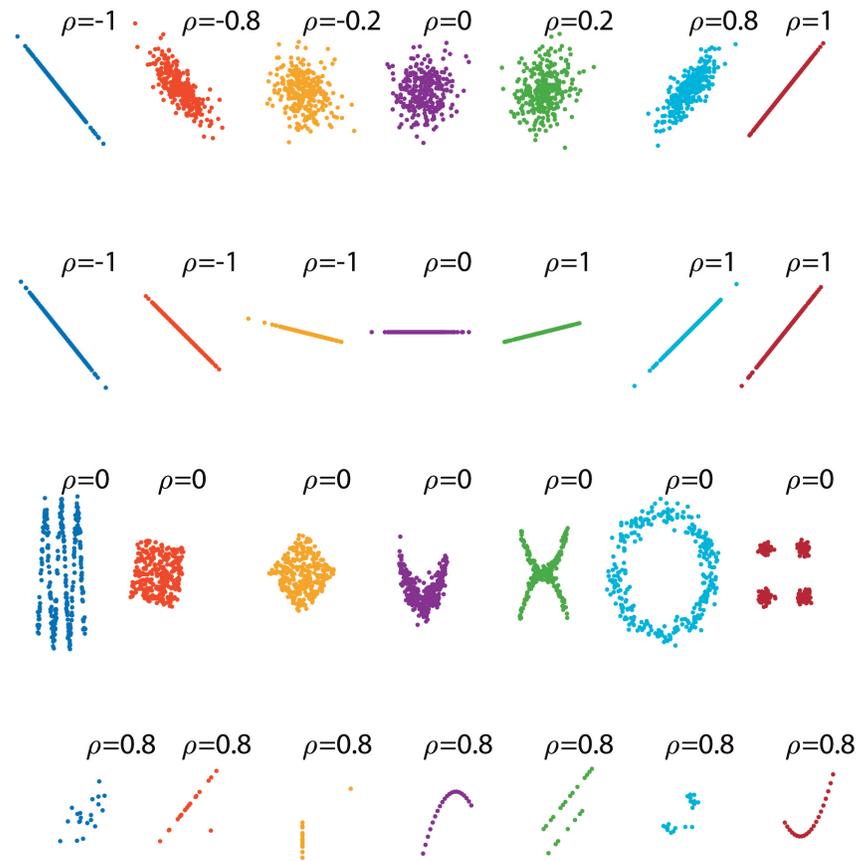


Figure 3.3: Example of several datasets and their corresponding correlation values.

discrete variable as factor and measure the  $R^2$  of the model.

- For two ordinal variables: we may use Kendall's or Spearman's rank correlation coefficient.
- For two categorical variables: we may use a  $\chi^2$  test for association.

The correlation coefficient answers the question "How much information do we gain on  $Y$  if we know the value of the variable  $X$ ?". There are more advanced versions of the correlation coefficient:

- Multiple correlation coefficient: "How much information do we gain on  $Y$  if we know the value of the variables  $X_1$  and  $X_2$ ?"
- Partial correlation coefficient: "How much information do we gain on  $Y$  if we know the value of the variable  $X_1$  once I have removed from  $Y$  the variability due to  $X_2$ ?"
- Part correlation coefficient: "How much information do we gain on  $Y$  if we know the value of the variable  $X_1$  once I have removed from  $X_1$  the variability due to  $X_2$ ?"

**We do not check the assumptions of regression.** Regression analysis, as all statistical techniques, makes assumptions about the observed data and the data generation model. Some of the assumptions are hard to know whether they are really fulfilled or not, but some other are very easy by simply inspecting the residuals of the regression. Generally speaking, any regression can be seen as the prediction of a variable  $Y$  as a function of some predictors (continuous or discrete)  $X_1, X_2, \dots, X_p$ . The difference between our prediction  $f(X_1, X_2, \dots, X_p)$  and the observations are the residuals,  $\varepsilon$

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$

Linear models like ANOVA and other related models presented along the book share many of the assumptions explained below and the same caution should be taken with them. In the following paragraphs we discuss these assumptions:

- Representative data. As all statistical techniques, regression assumes that the observed data is representative of a population (for instance, mice with hypertension or mice whose cholesterol level in blood was between 1.3 mg/mL and 1.7 mg/mL). If there are animals that do not belong to this population, then the regression results are biased. Analogously, if our sample does not fully represent the whole population it aimed to, our regression is biased. For the same reason, the regression results are only valid within the population for which the sample was representative. Applying the regression formula,  $f$ , to a different population (non-hypertensive mice or mice with a cholesterol level different from the observed range) is considered as an extrapolation. As such, extrapolation is not necessarily bad, but we should always be cautious about the validity of the "internal causes" driving the relationship between the predicted and predictor variables outside of the population for which the regression was performed.

- Predictors are not noisy. The standard regression tools are based on Least Squares (LS) optimization. The goal is to find the function  $f$  that minimizes the distance between the predicted and the observed values. Let us define the vector of predictors  $\mathbf{X} = (X_1, X_2, \dots, X_p)$ , and let us assume that we have  $N$  observations of predictor-observation pairs  $(\mathbf{X}_i, Y_i)$ , with  $i = 1, 2, \dots, N$ . Then, Least Squares can be formulated as

$$f_{LS}^* = \arg \min_f \sum_{i=1}^N (Y_i - f(\mathbf{X}_i))^2 = \arg \min_f \sum_{i=1}^N \varepsilon_i^2$$

This is minimizing the vertical distance between the observed and predicted values (see Fig. 3.4). Although other regression techniques exist, all those based on the minimization of something only related to  $\varepsilon$  implicitly assume that the measurement of the predictors are perfectly performed (without any noise), while the measurements of the observations are noisy.

If this assumption is not true (the measurements of the predictors are also noisy), then we should use Total Least Squares (or any of its variants). This technique minimizes the distance (not the vertical distance) between the predicted and observed values (see Fig. 3.4).

$$f_{TLS}^* = \arg \min_f \sum_{i=1}^N \|\mathbf{X}_i, Y_i - (\mathbf{X}_i, f(\mathbf{X}_i))\|^2$$



Figure 3.4: Illustration of the minimization implied by Least Squares (left) and Total Least Squares (right).

- Predictors are linearly independent. Another assumption is that none of the predictors can be expressed as a linear combination of the rest of predictors. If they are not linearly independent, then one of the predictors is redundant and can be removed from the regression because it does not bring any new information (for instance, in the regression of a mouse length as a function of the mouse weight in grams and ounces, the weight in ounces does not bring any new information that was not brought in by the weight in grams). The set of linearly dependent variables is said to be multicollinear, and the corresponding regression coefficients are very poorly determined (because the information can be arbitrarily shifted from one variable to the rest of the variables in the linearly dependent set). We

may detect multicollinearity through the condition number of the system matrix or the Variance Inflation Factor (see the Appendix of Chapter 5). Collinear variables (or almost collinear variables) should be condensed in a smaller set of linearly independent variables (if you have more predictors than samples, you are guaranteed to have a multicollinearity problem). This is done by a previous step of dimensionality reduction (Principal Component Analysis, Non-negative Matrix Factorization, Independent Component Analysis, Autoencoders, ...) or by Partial Least Squares (that has an embedded linear dimensionality reduction step).

An important source of multicollinearity is the use of non centered predictors. Assume that we have a model of the form

$$y = \beta_0 + \beta_1 x$$

and the mean of  $x$  is different from zero. Then, we could write the model as

$$y = \beta_0 + \beta_1(\mu_x + \tilde{x}) = (\beta_0 + \beta_1\mu_x) + \beta_1\tilde{x}$$

where  $\tilde{x}$  has now a zero mean. That is,  $x$  is collinear with the model constant, and the distribution of this shared information between  $\beta_0$  and  $\beta_1$  is arbitrary. Additionally, the  $R^2$  may get confused by this shared information and be artificially inflated to high values. We recommend to perform regressions with centered variables, both  $y$  and  $x$ . That is, we define the centered variables as  $\tilde{y} = y - \bar{y}$  and  $\tilde{x} = x - \bar{x}$ , and then fit the model

$$\tilde{y} = \beta_1 \tilde{x}$$

These models tend to have much fewer problems of multicollinearity, as one of the main sources of it (non-zero means) have been removed.

- Residuals are homocedastic. That is, they have the same variance across all values of the predictors (see Fig. 3.5 top for an example of heterocedastic residuals). If this is not the case, it is normally because the data generation model is not correct. Sometimes, this is corrected by some data transformation (of the predictor and the predicted variables). If not, you may try to use Weighted Least Squares, in which the residuals are multiplied by a factor that depends on the predictor value such that the corrected residual has the same variance across the predictor range.

$$f_{WLS}^* = \arg \min_f \sum_{i=1}^N w_i (Y_i - f(\mathbf{X}_i))^2$$

- Residuals are uncorrelated. Uncorrelated to the predictors and uncorrelated to the residuals themselves (see Fig. 3.5 bottom for an example of autocorrelated residuals). Plots of the autocorrelation function of the residuals or the cross-correlation between the residuals and predictors should reveal the violation of this assumption. This normally indicates that the family of explored functions  $f$  does not truly explain the data generation model and that we should resort to

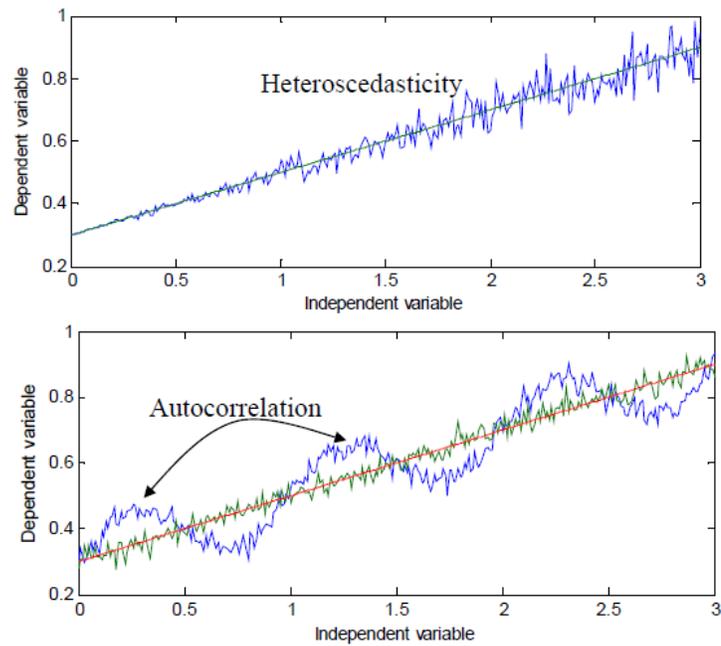


Figure 3.5: Illustration of a regression with heteroscedastic residuals (top, note that the variance of the residuals changes across the predictor value), and autocorrelated residuals (bottom, residuals are not independent of other residuals).

some other family. If this cannot be done, then we may use Generalized Least Squares in which a matrix  $W$  compensates the correlation among residuals. Let us refer to all the  $N$  observations as the vector  $\mathbf{Y}$  and to the  $N$  predictions as the vector  $\mathbf{F}$  such that the  $i$ -th component of this vector is  $F_i = f(\mathbf{X}_i)$ . Then, the Generalized Least Squares minimizes

$$f_{GLS}^* = \arg \min_f (\mathbf{Y} - \mathbf{F})^T W^{-1} (\mathbf{Y} - \mathbf{F})$$

A common mistake is performing a regression on smooth data (see Fig. 3.6). The smoothing can artificially create trends. Additionally, the smoothing introduces a local correlation among residuals increasing the coefficient of determination,  $R^2$ , and decreasing the p-value.

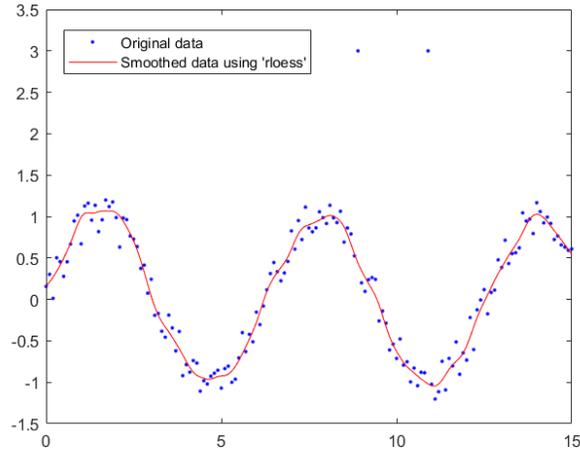


Figure 3.6: Example of smoothed data using rloess (a robust version of local regression using weighted linear least squares and a 2nd degree polynomial model).

- Residuals are normally distributed. Least Squares is tightly linked to the assumption of normality (Gaussian) distribution of the residuals. We should check that there are not outlier among the residuals. Useful tools for checking for the presence of outliers are the leverage, the studentized residual, and Cook's  $D$ . Remind that a data point is an outlier if it can be explained by reasons unrelated to the underlying population (like measurement errors, data transcription typos, ...) If after removing outliers residuals are still not Gaussian, for linear regression you may use a Generalized Linear Model (GLM) that are valid for any distribution of the residuals from the exponential family (the univariate members of this family are the Gaussian,  $\chi^2$ , Bernoulli, exponential,  $\beta$ ,  $\Gamma$ , and Poisson distributions). For non-linear regression, we need to formulate the problem in a Maximum Likelihood framework using the specific distribution of the residuals.

**We misuse regression.** We misuse regression when we overinterpret its results. For instance, we should use at least the p-value and the coefficient of determination,  $R^2$ , to fully understand a regression (remind that the  $R^2$  is the fraction of the data variance explained by the regression model). For instance, Fig. 3.7 shows a dataset for which a linear regression has very low p-value (i.e., it is highly significant) but it also has a very low coefficient of determination (i.e., it cannot explain the variations observed in the data). If we only consider the p-value, we overestimate the explanatory power of the model.

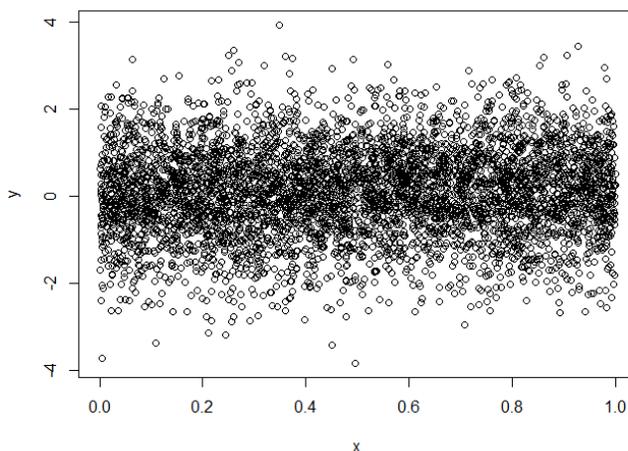


Figure 3.7: Example of a dataset for which a linear regression ( $Y = b_0 + b_1X$ ) has very low p-value (0.000105) and very low coefficient of determination,  $R^2$ , (0.003005). This low coefficient of determination implies that the model cannot explain even 0.5% of the observed variance.

However, choosing a model by maximizing the  $R^2$  is not a good practice because models with more parameters tend to have higher  $R^2$ . At some moment, with too many parameters we may perform overfitting as shown in Fig. 3.8. As a rule of thumb it is recommended to have between 10 and 20 observations per model parameter. For linear models, a more detailed sample size calculation has been given in Sec. 4.3. There is no formal definition of overfitting although the Vapnik-Chervonenkis (VC) dimension or sample complexity theory are formal frameworks related to overfitting. Informally, we may say that a model is overfitted if its complexity is not justified by the data. In this way, we may penalize a model for having too many parameters. An objective way of implementing this idea is by some formula that takes into account the explained variance (sum of squares) and the number of parameters. As we saw in Eq. 5.3 we may decompose the total sum of squares into a part that depends on the model and a part that depends on the residual. Actually,  $R^2$  is the fraction of the total sum of the squares

explained by the model.

$$R^2 = \frac{SS_{model}}{SS_{total}} = \frac{SS_{total} - SS_{residuals}}{SS_{total}} = 1 - \frac{SS_{residuals}}{SS_{total}}$$

If we have  $N$  observations and a model with  $p$  parameters, the adjusted  $R^2$  is defined as

$$R_{adj}^2 = 1 - \frac{SS_{residuals}/(N-p)}{SS_{total}/(N-1)}$$

In this way, we only augment the number of parameters if the decrease in sum of squares of the residuals sufficiently justifies the “cost” of an extra parameter. Other model selection tools exist like Akaike’s Information Criterion (AIC), Schwarz’s Bayesian Information Criterion (BIC), Minimum Description Length (MDL), or Mallow’s  $C_p$ , each one with different properties and assumptions.

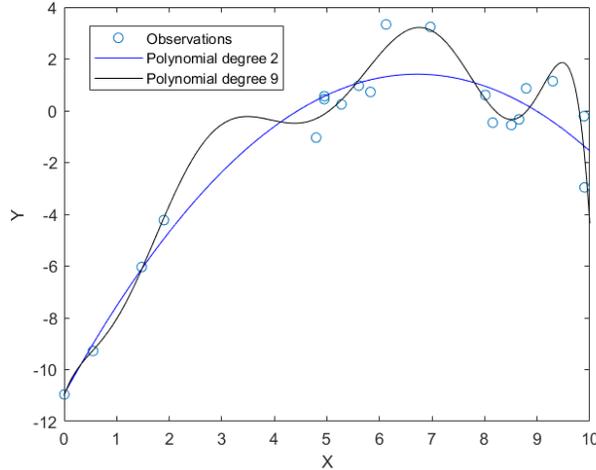


Figure 3.8: Example of overfitting. The same observations have been fitted with polynomials of degree 2 and 9. The more complex model goes closer to the observations thanks to the larger number of parameters, but it does not generalize well the value of the function between samples or outside the measurement region.

Two models are nested if one of them is a particular case of the other, for instance,  $Y = b_0 + b_1X$  is nested in  $Y = b_0 + b_1X + b_2X^2$ . For nested models we may also check if the extra parameter is justified through an hypothesis test called *Partial F test*. If we refer to the simple model as “reduced” and the more complex model as “full”, then under the null hypothesis ( $H_0$ : there is no difference between the explanatory power of both models) the statistic

$$F = \frac{\frac{SS_{residuals}^{reduced} - SS_{residuals}^{full}}{p_{full} - p_{reduced}}}{\frac{SS_{residuals}^{full}}{N - p_{full}}}$$

is distributed as a Snedecor's  $F$  with  $p_{full} - p_{reduced}$  and  $N - p_{full}$  degrees of freedom. Likelihood ratio, Wald or Score (Lagrange multiplier) tests can also be employed for this task. BIC and MDL were also designed for the selection of nested models. For the comparison of non-nested models we may use AIC or the Relative Likelihood test. Bear in mind that we should always compare models fitted to the same dataset.

Coming from the machine learning field, some other techniques like  $K$ -fold cross validation, leave-one-out, or bootstrapping can also be used to assess the validity of our regression model. These techniques follow a strategy in which the regression is performed on a subset of the data (training phase) and tested on the remaining part of the data (test phase). This process is repeated several times by randomly changing the subsets used for fitting and testing. By analyzing the performance of the regressor on the test data across the multiple runs, we can determine the capacity of the regressor to generalize to unseen data and assess the degree of overfitting.

We also misuse regression if we fit scientifically non-sensible models to data simply because the function fits well the data. For instance, the natural regression for the reaction rate of a chemical reaction is of the form

$$V = V_{max} \frac{[S]}{K_m + [S]}$$

where  $[S]$  is the concentration of the substrate,  $V$  is the reaction speed, and  $V_{max}$  and  $K_m$  are the model parameters. This family of functions have a chemical reasoning behind (with its own assumptions) that, if the assumptions are justified in our case, should explain the observed reaction rate values. We cannot use the regression model

$$V = b_0 + b_1 \log([S])$$

simply because it has a smaller  $R^2$ . In the absence of a family of regressors based on physical reasoning, we should not fit many regression functions and see which one fits the best. This is another way of data snooping.

**We misuse linear models.** Linear models assume that the observations can be explained by the sum of a collection of underlying factors (see Sec. 5.1). The most widely known is ANOVA. ANOVA can be 1-way, 2-ways, multi-way. The number of factors refers to the number of factors. For instance, we treat males and females at  $t = 0$  with a given treatment, and then measure the concentration of a given hormone in the blood of the animals at times  $t = 6h$ ,  $t = 12h$  and  $t = 24h$ , then we can say that our observation at a time  $t$  and sex  $s$  can be modelled as

$$y_{tsk} = \mu + \alpha_s^S + \alpha_t^T + \alpha_{ts}^{ST} + \epsilon_{tsk}$$

In the expression above,  $\alpha_s^S$  will take two values,  $\alpha_{male}^S$  and  $\alpha_{female}^S$ , explaining the different response of males and females (independently of time). Similarly,  $\alpha_t^T$  will take three values, corresponding to the different measuring times. These three values explain the effect of the treatment over time on the animals, disregarding their sex. Finally, the interactions will take 6 values explaining the different effect over time on females and males. This experiment truly calls for a two-way ANOVA design and

analysis as we will have two factors, sex and measuring time, that could explain our measurements.

If instead of the concentration of the hormone, we measure gene expression, we may be tempted to add an extra factor to the model

$$y_{gtsk} = \mu + \alpha_g^G + \alpha_s^S + \alpha_t^T + \alpha_{ts}^{ST} + \epsilon_{tsk}$$

For simplicity of exposition, we have only added the main effect of the factor. However, there is no difference between measuring an hormone or measuring a gene expression. The gene is an output, not an input to the model. It plays the same role as the hormone concentration. In the case of measuring the hormone, the parameter  $\mu$  is an overall mean of the hormone level irrespectively of sex and time. And this overall mean makes sense. In the case of the gene expression,  $\mu$  would be an overall mean of the gene expression of all genes, irrespectively of the gene, sex and time. This second parameter does not make sense. We should analyze the results of each gene independently, and we should use a two-way ANOVA, in the same way as we did in the case of measuring an hormone concentration.

Another example of the misuse of ANOVA is the following. Let us assume that we have four groups of measurements: Control-Treatment and Before-After applying the treatment. We may be tempted to analyze this data with a two-way ANOVA with the two factors treatment and time. However, the measurements before and after the treatment are paired. The two-way ANOVA disregards this matching. The correct way of analyzing this data is by subtracting (or dividing, depending on the nature of the data) the measurements before treatment from the measurements after treatment. We would obtain a single observation per animal and only a factor left, control or treatment. Then, we would go for a comparison of the mean of two groups with a Student's t-test or 1-way ANOVA (parametric tests that assume that the observations are Gaussian), or a Mann-Whitney test (non-parametric test without the Gaussian assumption).

**What to do with missing values?** There two types of missing values: 1) observations whose value cannot be measured, and 2) observations whose value have not been measured. For instance, titer of viral load that is below our detection threshold is a missing value of the first kind, while the concentration of a protein that has been mis-read in a proteomic study is of the second kind. There are several approaches to the treatment of missing values of the second kind: removing the whole individual (deleting a row of the data matrix), removing the whole protein (deleting a column of the data matrix), or imputing the missing value based on the value of similar individuals in the database. For this latter operation there are powerful algorithms based on regression trees or forests, k-Nearest Neighbours, etc. (Donders et al, 2006). Although imputing a value is not ideal, if there are enough data, this is not a bad option, especially if values are lost at random (the loss pattern is not caused by anything related with our treatments) and are relatively few. However, for missing values of the first kind, there is little we can do. The reason is that we cannot "invent" the values we cannot measure. If we substitute, for instance, values below the detection limit by the detection limit or a fraction of it, we are artificially lowering the variability of the observations. All parametric tests (ANOVA, Student's t-test, etc.) will be fooled by the low variability of the

imputed values and will underestimate the true observation variance. In this case, it is better to separate the analysis in two parts. In a first analysis, we report the proportion of samples for which the variable could be measured and compare this among groups. In the second analysis, we compare the observations that could be measured among the different groups.

### 3.3 Test selection guide

Although data analysis is not within the scope of this book, the question of which should be the appropriate statistical test for this data is so common, that we here include some guide to select some suitable statistical tool for the data analysis. This is not the ultimate guide and the researcher wise selection is always encouraged. The interested reader is referred to [Sheskin \(2004\)](#) and [Kanji \(2006\)](#) for a more comprehensive review of the tests available and their applicability. In this guide we simply give the test name and not the reference, the reader must look for it in his/her software tool as well as the theory behind it.

In many cases, there are parametric and non-parametric tests available. Parametric tests assume some statistical distribution for the observed data (usually normality). If this assumption is correct, these tests are more powerful than their non-parametric equivalents since they have more information about the data being analyzed. Having more statistical power is useful if the effect size (for instance, the difference between the different groups) is relatively small. If the effect size is large, then, both kind of tests (parametric and non-parametric) should be able to identify it as significant. On the other side, if the Gaussian assumption is largely violated, p-values calculated by parametric tests are incorrect.

#### S.1 What kind of analysis do you want to perform?

- Test about continuous variables (e.g., height, temperature, blood pressure), go to [S.2](#).
- Multiple dependent and independent variables (e.g., regression problem with multiple predictors and predicted variables), go to [S.11](#).
- Test about ordinal variables (e.g. mild, medium, severe), go to [S.12](#).
- Test about count data (e.g. number of visits to a maze room in 10 minutes), go to [S.18](#).
- Test about discrete/categorical variables (e.g., male/female, yes/no, red/green/blue), go to [S.19](#).
- Test about correlation/association (e.g., relationship between height and weight; animal sex and hormone level), go to [S.27](#).
- Test about survival (e.g., time before a tumor grows to a given size), go to [S.33](#)
- Sequential tests (e.g., we have to test 80 animals, but we will do interim tests to see whether we can take a decision earlier), go to [S.34](#).

**S.2** Test about continuous variables. Which is the number of variables you are analyzing?

- One variable (e.g., weight), go to [S.3](#).
- One variable, but it is an angle, go to [S.4](#).
- Two or more variables (e.g., weight and height), go to [S.10](#).

**S.3** Test about one continuous variable. A sample is a set of independent measurements. For instance if you are measuring the level of an hormone in a group of animals, you have 1 sample. If you are comparing the level of the hormone in two groups (control and treatment), then you have 2 samples. If you are comparing the level of the hormone for multiple doses and for each dose you test a number of animals, then you have 3 samples.

Two samples are independent if the measurements are on different animals. Two samples are dependent if an animal is measured before and after treatment, or we measure the left and right eye of the same animal. Two samples are also dependent if for each animal, we look for a matched animal with almost identical characteristics (for instance, twins or siblings).

How many samples do you have?

- 1 Sample, go to [S.5](#).
- 2 Independent samples, go to [S.6](#).
- 2 Dependent samples, go to [S.7](#).
- 3 or more independent samples, go to [S.8](#).
- 3 or more dependent samples, go to [S.9](#).

**S.4** Test about one continuous, angular variable. Due to the special nature of angles (0 and 360 degrees denote the same orientation), special tests are particularly suited for this special case.

- Randomness: to check if the set of angles tend to cluster or they are uniformly distributed you may use V-test or Watson's  $U_n^2$  test.
- Two samples:
  - Watson-Williams test or Mardia-Watson-Wheeler test to check if the mean angles of two independent groups significantly differ.
  - Watson's  $U^2$  test to test if two groups of angles significantly differ with respect to their mean direction or angular variance.
- Three or more samples: Harrison-Kanji-Gadsden is an ANOVA-like technique for angular data.

**S.5** Test about one continuous variable measured in one sample. For instance, measuring the response time of a group of animals or the concentration of a drug in serum in a single group of animals after having received a fixed dose.

- Parametric tests.

- Tests about the mean:
  - \* z-test: if you want to compare the mean of your sample to a predefined value and you know the variance before doing the experiment.
  - \* Student's t-test: if you want to compare the mean of your sample to a predefined value and you have to estimate the variance from the data itself.
- Test about the variance:  $\chi^2$  test if you want to compare the variance of your variable with some predefined value.
- Test about the skewness: skewness test if you want to know if the distribution your data is coming from is symmetric or not about its centre. A symmetric distribution has skewness=0.
- Test about the kurtosis: kurtosis test if you want to compare the kurtosis of your sample to a predefined value (e.g., the standardized Gaussian distribution has a kurtosis of 3).
- Normality test if you want to check if the distribution the data comes from is compatible with the Gaussian distribution. There are a number of them: 1) D'Agostino-Pearson test, 2) Jarque-Bera test, 3) Anderson-Darling test, 4) Cramér-von Mises criterion, 5) Lilliefors test, 6) Shapiro-Wilk test, 7) Pearson's  $\chi^2$  test, 8) Fisher's cumulant test, 9) the w/s test.
- Distribution test: if you want to test that your data is compatible with a particular continuous distribution, you may use Kolmogorov-Smirnov test.
- Test about outliers: Dixon's Q test, Grubbs' test, Peirce and Chauvenet's criteria for identifying the presence of outliers.
- Non-parametric tests.
  - Tests about the mean: The following tests are focused on showing if the mean or median of the variable being analyzed,  $X$ , is zero or not. If we are interested in showing that the is larger than a threshold  $\mu_0$ , we analyze the sign of the variable  $X - \mu_0$ . The following are different possibilities for this test: 1) A permutation test for the mean, 2) Wilcoxon signed-rank test, 3) the sign test for a median.
  - Test about randomness: for example, are the residuals of a regression really random, or they follow some pattern with the predictor variable? Residuals must be first sorted with respect to the predictor (for instance, time). Possibilities for this test are: 1) mean-square successive difference test, 2) the adjacency test for randomness of fluctuations, 3) the serial correlation test for randomness of fluctuations, 4) the turning point test for randomness of fluctuations, 5) the difference sign test for randomness, 6) the run test on successive differences, 7) the run test, 8) the Wilcoxon-Mann-Whitney rank sum test for the randomness of signs, 9) the rank correlation test for randomness.

**S.6** Test about one continuous variable measured in two independent samples. Some examples comparing an hormone level between two different mouse strains or comparing some physiological variable in a control and treatment groups.

- Parametric tests.
  - Test about the mean: you want to compare the mean in both samples. In case of multiple tests you should adopt some protection against family error inflation like 1) Bonferroni correction, Holm-Bonferroni, 3) Sidak, 4) Benjamini-Hochberg (False Discovery Rate).
    - \* z-test: if you know the variance of each group before doing the experiment.
    - \* Student's t-test: if you have to estimate the variance of each group from the data itself.
    - \* The single-factor, between-subjects ANCOVA (Analysis of Covariance), if you have measured another variable that may help in the comparison between the two groups (e.g., the weight of each animal).
    - \* If the ANOVA rejects the equality of the means of all groups, then post-hoc comparisons will be performed between pairs of groups. In this case, you may use: 1) Least Significant Difference, 2) Tukey's Honestly Significant Difference to compare all vs all groups (also known as Tukey's range, or Tukey-Kramer), 3) Link-Wallace test to compare all vs all groups, 4) Dunnett's test to compare all vs control, 5) Hsu's test to compare all vs best, 6) Scheffe's test to perform unplanned comparisons, 7) Brown-Forsythe if the variance of the two groups is different, 8) Duncan's Multiple Range test, 9) Newman-Keuls adapts the test to the size of the difference between the two groups.
  - Test about the variance: you want to compare the variance in both samples.
    - \* Snedecor's F test for two population variances. There is a variant of this test that includes the correlation between measurements in both groups.
    - \* Hartley's  $F_{max}$  test for the homogeneity of variance.
    - \* Bartlett's test.
    - \* Levene's test.
- Non-parametric tests.
  - Test about the mean: Rather than the mean, the following tests normally address the equality of the median. 1) Wilcoxon-Mann-Whitney or Mann-Whitney U test or rank-sum test, 2) Tukey-Duckworth test, 3) Mood's median test, 4) Rank-sum test for the difference between the largest mean and the rest.
  - Test about the variance: 1) Siegel-Tukey test for equal variability, 2) Moses test for equal variability

- Test about the distribution: Do both samples come from the same distribution? You may use: 1) Kolmogorov-Smirnov test, 2) the median test of two populations, 3) Wilcoxon inversion test (U-test), 4) van der Waerden normal-scores test.

**S.7** Test about one continuous variable measured in two dependent samples. For example, comparing the effect of a drug before and after treatment, or comparing the recovery of two dermal lesions in an animal (one treated and the other untreated). Typically both measurements (e.g., before and after) are combined into a single variable that represents the difference between the two stages. However, this step depends on the specific tool used.

- Parametric tests.
  - Test about the mean: 1) z test for two dependent samples if you assume you know beforehand the variance of the difference, 2) Student's t test for two dependent samples if you will estimate the variance from the data itself, 3) Sandler's A test, 4) the single-factor, between subjects Analysis of Variance (1-way ANOVA), 5) the single-factor, between subjects Analysis of Covariance (1-way ANCOVA) if you have also measured some covariate (e.g., the animal's age).
- Non-parametric tests.
  - Test about the mean: 1) Wilcoxon matched-pairs signed-rank test, 2) binomial sign test for two dependent samples, 3) Kruskal-Wallis one-way Analysis of Variance (ANOVA), 4) Jonckheere-Terpstra test for ordered alternatives.

**S.8** Test about one continuous variable measured in three or more independent samples. For example, comparing the effect on weight of five different diets.

- Parametric tests.
  - Test about the mean: you want to compare the mean in all groups. The rejection of the null hypothesis normally implies that not all means are equal, meaning that there are at least two that are different to each other. A post-hoc analysis then follows trying to identify the pair(s) that is(are) different (see S.6).
    - \* The single-factor, between subjects Analysis of Variance (1-way ANOVA).
    - \* The single-factor, between-subjects Analysis of Covariance (1-way ANCOVA), if you have measured another variable that may help in the comparison (e.g., the weight of each animal before starting the diet).
    - \* If groups are defined by several independent variables, you may use 2-way ANOVA, 3-way ANOVA. For example, groups are defined by diet and sex.

- \* If groups have a nested nature (e.g., we take several individuals from each group, and take several measurements from each individual; the measurements form a subgroup within larger groups), the you may use nested ANOVA.
  - Test about the variance: you want to check if the variance in all groups is the same. Some possibilities are: 1) Hartley's  $F_{max}$ , 2) Bartlett's test, 3) Cochran's C test, 4) Levene's test.
- Non-parametric tests.
  - Test about the mean: 1) Kruskal-Wallis 1-way ANOVA, 2) ordered logistic regression, 3) Steel test for comparing K treatments with a control, 4) median test of K populations, 5) Jonckheere-Terpstra test for ordered alternatives.
  - Test about the variance: Brown-Forsythe test.
  - Test about the distribution: van der Waerden normal-scores test.

**S.9** Test about one continuous variable measured in three or more dependent samples. For example, measuring the blood pressure of animals before treatment and 1, 2, 4 and 8 hours after a drug bolus.

- Parametric tests.
  - Test about the mean: The single-factor, between subjects Analysis of Variance (1-way repeated measures ANOVA).
- Non-parametric tests.
  - Test about the mean: 1) Friedman two-way Analysis of Variance (ANOVA) by ranks, 2) Page test for ordered alternatives.

**S.10** Test about two or more continuous variables. For the hypothesis about the difference between two or more population means you may use:

- The between-subjects factorial analysis of variance (1-way MANOVA). The associated test is the Hotelling's  $T^2$  test. Example: Comparing the height and weight of two or more mouse strains.
- The within-subjects factorial analysis of variance (Repeated MANOVA, the same animal is measured multiple times along time, and the time changes are sought). Example: Comparing the height and weight of a group of mice as they grow at 1, 5, 10 weeks old.
- The factorial analysis of variance for a mixed design (One-way and Repeated MANOVA). Example: Comparing the height and weight of two or more mouse strains as they grow at 1, 5, 10 weeks old.

**S.11** Regression. In the following, we understand regression in a very wide sense. We will include many different problems that they all share a common characteristic: they all can be understood as trying to find some functional relationship between sets of variables. We will distinguish between the analysis of dependence (that tries to find an explicit dependence  $(Y_1, Y_2, \dots) = f(X_1, X_2, \dots)$ )

and the analysis of interdependence (that tries to find an implicit dependence  $(Z_1, Z_2, \dots) = f(X_1, X_2, \dots)$ ). In the following the variables  $Y_1, Y_2, \dots$  refer to the experimentally observed, dependent variables, while  $X_1, X_2, \dots$  refer to experimentally observed, independent variables. The variables  $Z_1, Z_2, \dots$  refer to unobserved (latent) independent variables. Continuous variables are represented by capital letters ( $X_1, X_2, \dots$ ), while non-continuous variables are represented by small letters ( $x_1, x_2, \dots$ )

- Analysis of dependence.
  - Regression:  $Y_1 = f(X_1)$ . A continuous variable is predicted from another continuous variable (e.g. the length of an animal is predicted from his weight). The function  $f$  may be linear or non-linear. If the function is linear and the residuals are Gaussian, there are closed-form tests for the regression coefficients. If not, bootstrapping may be used to test that the regression coefficients are significantly different from 0. The significance of the different parameters of a linear regression can be performed with 1) F-test for non-additivity, 2) F-test for main effects and interaction effects, 3) F-test for nested or hierarchical classification, 4) F-test for the linearity of regression, 5) t-tests and Z-tests may also be used if the residuals are normal. To see if the residuals of a time regression are autocorrelated you may use the Durbin-Watson test.
  - Multiple regression  $Y_1 = f(X_1, X_2, \dots)$ . A continuous variable is predicted from other continuous variables (e.g., the length of an animal is predicted from his weight and his age). See comments for regression above.
  - Structural Equation Modelling  $(Y_1, Y_2, \dots) = f(X_1, X_2, \dots, Z_1, Z_2, \dots)$  Dependent variables are predicted from observed and unobserved variables (e.g., (cholesterol LDL, cholesterol HDL)=f(weight, enzyme A activity [unobserved])). See comments for regression above.
  - MANOVA  $(Y_1, Y_2, \dots) = f(x_1, x_2, \dots)$ . Several continuous variables are predicted by several categorical variables (e.g., (Concentration hormone A, concentration compound B)=f(strain, sex, age group)). Associated tests to MANOVA are 1) Wilks'  $\Lambda$ , 2) Pillai-Bartlett trace, 3) Lawley-Hotelling trace, 4) Roy's greatest root, 5) Hotelling's  $T^2$ .
  - MANCOVA  $(Y_1, Y_2, \dots) = f(x_1, x_2, \dots, X_1, X_2, \dots)$ . Several continuous variables are predicted by several categorical variables and some continuous covariates (e.g., (Concentration hormone A, concentration compound B)=f(strain, sex, age group, concentration compound C)). See comments for MANOVA above.
  - Discriminant analysis (and in general any classification algorithm)  $y_1 = f(X_1, X_2, \dots)$ . A discrete, binary variable is predicted from several continuous variables (e.g. disease (or healthy)=f(gene A expression, gene B expression)). Testing if the classifier performs significantly better than random can be done though McNemar's or Fisher's exact tests.

Cross-validation and bootstrapping also help to validate a classifier. There is also a Discriminant test for the origin of a sample (e.g., is one sample generated by model A or by model B).

- Logistic/Logit regression  $Y_1 = f(X_1, X_2, \dots)$ . A discrete variable  $y_1$ , approximated by a continuous variable  $Y_1$  (sometimes interpreted as the probability of belonging to one class) is predicted by several continuous variables (e.g., probability of disease=f(gene A expression, gene B expression)). See comments for regression.
- Canonical correlation, Partial Least Squares  $(Y_1, Y_2, \dots) = f(X_1, X_2, \dots)$ . Several continuous variables are predicted by several continuous variables (e.g. (gene A expression, gene B expression)=f(gene C expression, gene D expression)). See comments on regression above.
- Conjoint analysis  $y_1 = f(X_1, X_2, \dots, x_1, x_2, \dots)$ . An ordinal variable is predicted by several categorical/ordinal/metric variables (e.g., severity=f(animal movement, water intake, food intake)). See comments on discriminant analysis above.
- Analysis of interdependence. Although there are no tests associated, the following techniques are useful data analysis techniques that are used in many contexts.
  - Dimensionality reduction (Principal Component Analysis, Factor Analysis, Independent Component Analysis, Non-negative Matrix Factorization, ...)  $Z_1, Z_2, \dots = f(X_1, X_2, \dots)$ . Several continuous latent factors are sought from the continuous input data (e.g. (curiosity,intelligence)=f(exploration time, exercise time, sleeping time)).
  - Correspondence analysis  $Z_1, Z_2, \dots = f(x_1, x_2, \dots)$ . Several continuous latent factors are sought from discrete input data (e.g. (gene expression A, gene expression B)=f(hair colour, eye colour, skin colour)).
  - Clustering  $z_1 = f(X_1, X_2, \dots)$ . A categorical variable, the cluster label, is predicted from several numerical variables (e.g. animals having similar characteristics are put into the same cluster).

**S.12** Tests about ordinal variables. An ordinal variable is one in which the values are ordered (mild, medium, severe, irreversible), but the distance from one value to the next does not have any meaning. For this reason, ordinal variables convey much less information than continuous variables. They are also more related to subjective evaluations (for instance an animal procedure may seem medium to a veterinarian and severe to another one), and in this way subjected to a higher level of noise. Many non-parametric tests of continuous variables (like temperature or height) treat the variables like ordinal. In a way, most tests for ordinal variables can be considered non-parametric, and generally speaking, there are many fewer options than for continuous variables.

In the following all cases refer to experiments with a single ordinal variable that will be tested. A sample is an independent group that has received a treatment. For instance, if you are comparing the severity of a procedure in three groups that are receiving different treatments, you have three samples.

How many samples do you have?

- (a) 1 sample, go to [S.13](#).
- (b) 2 independent samples, go to [S.14](#).
- (c) 2 dependent samples, go to [S.15](#).
- (d) 3 or more independent samples, go to [S.16](#).
- (e) 3 or more dependent samples, go to [S.17](#).

**S.13** Tests about one ordinal variable in one sample. For example, which is the severity of a procedure evaluated for a single treatment.

- Test about the median: if you want to compare the median of the observations to a predefined median, then you may use Wilcoxon signed-rank test.
- Test about the distribution: if you want to compare the frequency of the observations to some predefined discrete distribution (e.g., mild 50%, medium 30%, severe 15%, irreversible 5%), you may use  $\chi^2$  goodness-of-fit test.

**S.14** Tests about one ordinal variable in two independent samples. For example, which is the severity of a procedure evaluated for two treatments or for two research centers.

- Test about the median: you want to compare the median of the observations in the two groups. Possibilities are: 1) Mann-Whitney U-test, 2) Permutation test, 3) Ordered logistic regression.
- Test about the variance: you want to test if both groups have the same variability. Possibilities are: 1) Bootstrap, 2) Jackknife, 3) Siegel-Tukey test for equal variability, 4) Moses test for equal variability.
- Test about the distribution: you want to test if both groups have the same discrete distribution. Possible tests are: 1) Kolmogorov-Smirnov test for discrete distributions, 2) Bootstrap Kolmogorov-Smirnov test, 2)  $\chi^2$  test.

**S.15** Tests about one ordinal variable in two dependent samples. For example, which is the severity of a procedure evaluated by two veterinarians, they independently evaluate the same animals.

- Test about median: you want to test if the median of the difference between both evaluations is zero. You may use: 1) Wilcoxon matched-pairs signed-rank test, 2) binomial sign test for two dependent samples, 3) permutation test.

**S.16** Tests about one ordinal variable in three or more independent samples. For example, which is the stress level of animals evaluated for three treatments.

- Test about median: you want to test if all groups have the same median. Possibilities are: 1) Kruskal-Wallis one-way Analysis of Variance (ANOVA) by ranks, 2) Jonckheere-Terpstra test for ordered alternatives, 3) van der Waerden normal-scores test, 4) factorial logistic regression.

**S.17** Tests about one ordinal variable in three or more dependent samples. For example, which is the stress level of animals evaluated by three veterinarians.

- Test about median: you want to test if all groups have the same median. Possibilities are: 1) Friedman two-way Analysis of Variance (ANOVA) by ranks (it is the non-parametric equivalent to the Repeated measures ANOVA, one of the variables is the animal being evaluated, the other is the evaluator), 2) Page test for ordered alternatives.

**S.18** Test on count data. Count data is of the form 0, 1, 2, ... For example, number of photons arriving a detector, number of cells of a given type in a microscope field, number of visits to a maze room in 10 minutes, etc. The usual way of dealing with count data is by assuming it follows a given discrete distribution and by fitting the parameters of that distribution. The fitting can be by a constant (and then two or more groups can be compared) or by regression. Distributions normally considered are:

- Poisson: number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.
- Negative binomial: number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures occurs.
- Zero-Inflated Count Models: on top of Poisson and Negative Binomial count models we may add the concept of zero-inflation. For instance, we may count the number of fishes caught by visitors to a national park. The number of fishes can be modelled with a Poisson of a given parameter calculated as a function of the number of children in the group, the number of nights in the park, and the number of persons in the group. If a group takes 0 fishes, it may be because they tried to fish, but did not catch any fish (true zero), or because they went with the children to see the park and they did not try even fishing (inflated zero).
- Zero-truncated Count Models: There are situations in which the value 0 is not amenable for the Poisson or Negative Binomial count models. For instance, the number of nights at hospital can be 1, 2, 3, ... but cannot be 0. The probability distribution has to be adjusted to exclude the value 0.
- Hurdle models: there are two models: one generating the zeros, and another one generating the positive values.
- Random effects count models: the parameter of the Poisson or negative binomial is assumed not to be constant, but the realization of an underlying random variable.

The following tests can be performed:

- Single sample:

- Test on the value of the model parameter. For example, has the count rate parameter departed from a reference situation?
- Test on the significance of a regression parameter. For example, does the ingestion of magnesium increase the number of immune cells circulating in blood?
- Two samples: for example, is the event rate in one group larger than in the other group? If you are comparing two samples, and assume that both have Poisson counts with different rates, then you may compare the two rates with: 1) z-test if the the Poisson can be approximated by a Gaussian, 2) F-test.

**S.19** Test about discrete/categorical variables. A discrete variable is one that takes a finite number of values (yes/no, control/disease, red/blue/green, male/female, ...). There is no logical order among the different possibilities.

How many discrete variables are you considering?

- 1, go to [S.20](#).
- 2, go to [S.26](#).

**S.20** Test about one discrete/categorical variable. A sample is a set of independent measurements that have received the same treatment. For instance, if you are studying the number of responding animals for three different treatments, you have three samples.

How many samples do you have?

- 1 sample, go to [S.21](#).
- 2 independent samples, go to [S.22](#).
- 2 dependent samples, go to [S.23](#).
- 3 or more independent samples, go to [S.24](#).
- 3 or more dependent samples, go to [S.25](#).

**S.21** Test about one discrete/categorical variable in one sample.

- Test about the distribution: For example, is the distribution of phenotypes concordant with a Mendelian inheritance? or is the proportion of males with a given behavior equal to 50%?
  - For binary variables: 1) Binomial sign test, 2) z-test (if the sample size is large enough so that the number of observations can be approximated by a Gaussian).
  - For multivalued (including binary) variables: 1) exact test of goodness-of-fit, 2)  $\chi^2$  goodness-of-fit, 3) G-test of goodness-of-fit (if the sample size is large).

- Test about randomness: if you are testing if the sequence of observations is random or, on the contrary, it has a time pattern. Possible tests are: 1) Single sample runs test, 2) The difference sign test for randomness, 3) the Wilcoxon-Mann-Whitney rank sum test for the randomness of signs, 4) the rank correlation test for randomness, 5) frequency test, 6) gap test, 7) Poker test, 8) maximum test, 9) Coupon's collector test.

**S.22** Test about one discrete/categorical variable in two independent samples. For example, you are testing if the number of respondents to a drug in two groups are similar, or if the choice preferences in two groups are similar.

- Test about distribution:
  - For binary variables: 1) Fisher's exact test, 2) z-test (if the sample size is large enough so that the number of observations can be approximated by a Gaussian), 3) z-test for correlated proportions.
  - For multivalued (including binary) variables: 1)  $\chi^2$  for homogeneity of groups, 2)  $\chi^2$  for the independence of groups, 3) G-test of independence (if the sample size is large).

**S.23** Test about one discrete/categorical variable in two dependent samples. For example, scoliosis is a disease in which the spine has an excessive curvature. One of the problems is the development of fibrosis. Are the probabilities of developing fibrosis in the concave and convex sides of the spine equal? We have samples from the convex and concave sides of the same animal.

- Test about distribution: 1) McNemar test, 2) Gart test for order effects, if the dependence is introduced by the order in which treatments are applied, 3) Bowker test of internal symmetry, 4) Stuart-Maxwell test of marginal homogeneity.

**S.24** Test about one discrete/categorical variable in three or more independent samples. For example, you are testing if the number of respondents to three different drugs are similar.

- Test about distribution: 1)  $\chi^2$  for the compatibility of  $K$  counts, 2) Cochran's test for the consistency of an  $K \times 2$  contingency table, 3)  $\chi^2$  test for the independence of a  $p \times q$  contingency table, 4) Cochran-Mantel-Haenszel test (it adds an extra variable, for instance, you are testing if the number of respondents to three different drugs are similar, and you will repeat this experiment at different research centers).

**S.25** Test about one discrete/categorical variable in three or more dependent samples. For example, you are testing if the number of respondents to two drugs along time (repeated measures) is the same.

- Test about distribution: 1) Cochran's Q test, 2) Repeated measures logistic regression.

**S.26** Test about two discrete/categorical variables in two samples. For example, count the number of animals with a given phenotype among three possibilities in a genetic cross (expected to follow a 1:2:1 ratio), do multiple crosses. One of the variables is the phenotype, the other the cross number.

- Test about distribution: Repeated G-tests of goodness-of-fit.

**S.27** Test about correlation/association. What kind of variables do you want to test?

- 1 variable among multiple groups, go to [S.28](#).
- 2 or more continuous variables, go to [S.29](#).
- 2 ordinal/rank variables, go to [S.30](#).
- 2 categorical variables, go to [S.31](#).
- 1 continuous and 1 categorical variables, go to [S.32](#).

**S.28** Test about correlation/association of one variable among multiple groups. We ask several researchers to evaluate the pain level of a given procedure. What is the correlation (consistency) between the different answers? We have a collection of measurements for each rater and the measurements from the same rater make a group.

This problem of correlation between groups can be addressed by:

- For continuous variables: Test on the intraclass correlation coefficient.
- For ordinal variables: 1) Kendall's coefficient of concordance test, 2) the rank correlation test for agreement, 3) Friedman's test.

**S.29** Test about correlation/association of two or more continuous variables. The correlation tests normally are based on linear associations between the variables being studied (see Sec. [3.2](#), about the misuse of the correlation).

- Parametric tests: they normally assume gaussianity of the two variables being compared.
  - Correlation between two variables:
    - \* Test of the Pearson correlation coefficient. For example, are the length and weight of an animal correlated?
    - \* Test of the Partial correlation coefficient. For example, are the length and weight of an animal correlated when we remove the effect of the age from both variables?
    - \* Test of the Semipartial correlation coefficient. For example, are the length and weight of an animal correlated when we remove the effect of the age from the length?
  - Correlation between a variable and a set of variables.
    - \* Test of the Multiple correlation coefficient. For example, what is the correlation between weight and (length, waist size, and neck size)?

- Non-parametric tests: go to [S.30](#).

**S.30** Test about correlation/association of two ordinal/rank variables. For instance, what is the correlation between the assessment of the severity of a procedure and the level pain?

This problem is addressed by: 1) test on the Spearman's rank-order correlation coefficient, 2) test on Kendall's  $\tau$ , 3) test on Goodman and Kruskal's  $\gamma$ .

**S.31** Test about correlation/association of two categorical variables.

- For binary variables: for example, is having a gene active or not related to a disease state? This is addressed by: 1)  $\chi^2$  test of independence, 2) Fisher's exact test, 3) test on the contingency coefficient, 4) test on Cramer's  $\phi$  coefficient, 5) test on Yule's Q, 6) test on the Odds Ratio, 7) test on Cohen's  $\kappa$ .
- For multivalued (including binary) variables: is the phenotype of an animal related to its social behavior classified into 4 different categories? This is addressed by: 1) test on the contingency coefficient, 2) test on Cramer's  $\phi$  coefficient, 3) test on the Odds Ratio, 4) test on Cohen's  $\kappa$ .

**S.32** Test about correlation/association of one continuous and one categorical variables. Is the animal length related to sex? Or to a specific phenotype?

- For binary variables: 1) test on Cohen's d index, 2) test on Cohen's g index.
- For multivalued (including binary): 1) Test on the coefficient of determination,  $R^2$ , of a 1-way ANOVA model, 2) test on  $\Omega^2$ , 3) test on  $\eta^2$ , 4) test on Cohen's f index.

**S.33** Test about survival. Survival data is analyzed by fitting a survival model to the observed data and then making inference on the fitted parameters. The parameters may be assumed to be constant or to be a function of other variables. In this latter case, tests on the significance of the regression parameters may also be performed.

The following models are normally used in survival analysis: 1) Exponential survival, 2) Weibull survival, 3) Normal survival, 4) Log-logistic survival, 5)  $\Gamma$  survival, 6) Exponential-logarithmic survival.

The regression of the model parameters can be done through: 1) Cox proportional hazards regression, 2) Parametric survival models, 3) Survival trees.

The following tests can be performed:

- Single sample:
  - Test on the value of the model parameter. For example, has the survival parameter departed from a reference situation?
  - Test on the significance of a regression parameter. For example, does the ingestion of iron relate to the survival after stroke?

- Two samples: For example, is the survival in one group larger than the survival in another group? This can be done by: 1) Test on the comparison of two exponential models (one for each group), 2) Log-rank test, there is no assumption about the specific survival model but it is assumed that the ratio between the hazards in both groups is constant.

**S.34** Sequential tests. Sequential tests are performed as a way to early stopping the experiment if there are not good chances of having found a useful treatment or if the evidence of having found it is so overwhelming that we do not need to go to the end of the planned experimental size.

There are sequential tests to verify:

- That the mean of the treatment is different from a reference mean.
- That the variance of the treatment is different from a reference variance.
- That the proportion of individuals with a given label is different from a reference proportion (a Bernouilli variable).
- That the coefficient of variation (standard deviation divided mean) is within pre-specified limits.

Some of these tests and the corresponding sample size calculations are presented in Sec. 4.9.

## **Part 2**



## Chapter 4

# Sample size calculations

In this chapter we will review the most common cases encountered in animal experiments. The sample size depends on the objective of our study. We will distinguish between two different kinds of studies:

1. Hypothesis test: these studies aim at rejecting a null hypothesis and, consequently, accepting the alternative hypothesis. It is the most common situation in experimental research: we want to prove that our vaccine or drug is effective, that a given gene is related to a disease, that a given diet has some particular effect on individuals, or that some environment causes some specific phenotype. The result of an hypothesis test is binary: the null hypothesis is rejected or it cannot be rejected.
2. Confidence intervals: these studies aim at identifying a range of values that characterize a parameter of interest: which is the average temperature of the laboratory, which is the average number of leukocytes per mL. of blood, which is its variance, which is the proportion of animals that get infected with a virus at a given virus dose, or which is the correlation between the expression level of a given gene and a phenotype of interest.

Actually, both kinds of studies are related and both are based on the same statistical inference theory. In fact, the hypothesis test can be calculated by computing a confidence interval on a statistic and checking if this statistic includes the value specified by the null hypothesis.

We will review the sample size when the test or confidence interval is on the mean of a given variable, a proportion, a regression coefficient, a variance, a Poisson count, or a survival rate. We will also see how to design the sample size for a pilot study in which there is no prior knowledge about the experiment results, and how to design experiments with early stopping criteria if we see that the treatment is not effective enough or we have already collected enough evidence that the treatment is effective.

The chapter is written as a reference and there is no need to read it all together. However, in a first reading, we recommend to see the examples to get an idea of the kind of problems that can be successfully solved and that cover a wide spectrum of experimental situations.

## 4.1 Sample size for the mean

### 4.1.1 Hypothesis test on the mean of one sample when the variance is known

This is exactly the case of Sec. 1.6. For completeness, let us reproduce it here. The hypothesis test is of the form

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_a : \mu &\neq \mu_0 \end{aligned} \quad (4.1)$$

The sample size formula was

$$N = \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\tilde{\Delta}} \right)^2 \quad (4.2)$$

where  $\alpha$  is the Type I error probability (rejecting the null hypothesis when it is wrong) and  $\beta$  is the Type II error probability (not rejecting the null hypothesis when it is false).  $z_x$  is the point of the standardized Gaussian curve (zero mean and standard deviation one) whose area under the curve from  $-\infty$  to that point is  $x$ .  $\tilde{\Delta} = \Delta/\sigma$  is the effect size normalized by the standard deviation of the observations,  $\sigma$  is the standard deviation of our observations and  $\Delta$  is the effect size we want to detect with power  $1 - \beta$  and statistical confidence  $1 - \alpha$ . That is if  $\mu$  departs from  $\mu_0$  in at least  $\Delta$ , then we will detect it with the specified power and statistical confidence. The use of a normalized effect size,  $\tilde{\Delta}$  highlights the fact that how large or small an effect size is, is relative to the amount of noise present in our measurements. Large effect sizes buried in a lot of noise are as difficult to detect as smaller effect sizes with less noise.

The sample size design formula above is valid only for two-tail tests (those in which the null hypothesis uses an equal sign). For one-tail tests,

$$\begin{aligned} H_0 : \mu &\geq \mu_0 \\ H_a : \mu &< \mu_0 \end{aligned} \quad (4.3)$$

or

$$\begin{aligned} H_0 : \mu &\leq \mu_0 \\ H_a : \mu &> \mu_0 \end{aligned} \quad (4.4)$$

the formula must be slightly modified to

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\tilde{\Delta}} \right)^2 \quad (4.5)$$

As we have discussed,  $z_{1-\alpha}$  is smaller than  $z_{1-\frac{\alpha}{2}}$  and, consequently, the number of samples required for one-tail tests is smaller than that for two-tail tests.

- **Example 44:** In the example of the laboratory temperature, the standard deviation of the thermostat is  $\sigma = 0.5^\circ$ , we wanted to detect a deviation of  $\Delta = 0.25^\circ$ ,

with a statistical power of 80% and a statistical confidence of 95%. As we showed above, this requires 32 samples

$$N = \left( \frac{z_{0.975} + z_{0.8}}{\bar{\Delta}} \right)^2 = \left( \frac{1.96 + 0.84}{0.25/0.5} \right)^2 = 31.40$$

However, our thermometer also has a measurement error whose standard deviation is  $0.2^\circ$  that adds to the standard deviation of the thermostat. Independent additive variables (true temperature+thermometer error) add their variances, so that the variance of our observations will be

$$\sigma^2 = \sigma_{\text{thermostat}}^2 + \sigma_{\text{thermometer}}^2 = 0.5^2 + 0.2^2 = 0.29$$

The standard deviation of our measurements become now

$$\sigma = \sqrt{0.29} = 0.54$$

And the sample size

$$N = \left( \frac{1.96 + 0.84}{0.25/0.54} \right)^2 = 36.57$$

That is, we would need 37 samples to take the decision of stopping the thermostat or not. This is 5h later (remember that we take a sample every hour) than in the case of a perfect thermometer. This is due to the extra uncertainty introduced by the measurement process. However, we may reduce the reaction time to the same 32h as in the case of a perfect thermometer. For doing so, we simply need to take 8 samples every hour of the current temperature, and average them. The averaging will reduce the uncertainty due to the thermometer, but it cannot reduce the uncertainty due to the thermostat

$$\sigma^2 = \sigma_{\text{thermostat}}^2 + \sigma_{\text{thermometer}}^2/8 = 0.5^2 + 0.2^2/8 = 0.255$$

and now the required sample size is

$$N = \left( \frac{1.96 + 0.84}{0.25/\sqrt{0.255}} \right)^2 = 31.99$$

That is  $N = 32$ .

#### 4.1.2 Hypothesis test on the mean of one sample when the variance is unknown

This case is much more common than the previous one. Although, in order to design the experiment we must have a guess of the standard deviation of the observations, when the experiment is performed, this standard deviation will normally be estimated from the samples themselves. This acknowledges our uncertainty on the prior we have used. For instance, in the previous example we assumed that the standard deviation of

the thermostat was  $0.5^{\circ}\text{C}$  and that of the thermometer  $0.2^{\circ}\text{C}$ . However, in reality, we may be uncertain about the absolute correctness of these numbers, and we may prefer estimating them from the data once the experiment is performed. Let us refer to the observations as  $y_i$  ( $i = 1, 2, \dots, N$ ). Let the population mean and standard deviation be  $\mu$  and  $\sigma$ . We do not have access to these parameters, but we may estimate them as

$$\begin{aligned}\hat{\mu} &= \frac{1}{N} \sum_{i=1}^N y_i \\ \hat{\sigma} &= \sqrt{\frac{1}{N-1} \sum_{i=1}^N (y_i - \hat{\mu})^2}\end{aligned}\quad (4.6)$$

These statistics are called the sample mean and standard deviation, respectively. As opposed to the sample parameters, that are fixed numbers, our estimates of the parameters are random variables, with their own distributions around the true population parameters. Our statistical test is still on the hypotheses in Eq. 4.1. However, along the analysis we will substitute the population parameters by the sample parameters. Instead of the  $z$  statistic that uses the population standard deviation

$$z = \frac{\hat{\mu} - \mu_0}{\sigma}$$

we use the  $t$  statistic that uses the sample standard deviation

$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}}$$

The larger the sample, the more similar our estimates are to the population parameters. Additionally, the fact that our estimates are random variables means that the sample size design formulas employ different distributions with respect to their ideal sample size design formula. But the design principles are still valid (see Eq. 1.18 and the reasoning around): 1) given a statistical confidence  $(1 - \alpha)$  we find the threshold of the statistic such that the probability of observing a statistic at least as large as the threshold if the null hypothesis is true is  $\alpha$  (or  $\alpha/2$  if we have a two-sided test); let us refer to this threshold as the “critical” value; 2) with the critical value above, the probability of not rejecting the alternative hypothesis at a given effect size is at most  $\beta$ . This latter probability depends on the sample size,  $N$ , and we must find the sample size such that this latter constraint is fulfilled. In the case of the Student’s  $t$ , the sample size design in the ideal case, Eq. 1.18, now becomes

$$\Pr \left\{ t_{\lambda, N-1} < t_{1-\frac{\alpha}{2}, 0, N-1} \right\} < \beta \quad (4.7)$$

The probability is measured with a Student’s  $t$  with  $N - 1$  degrees of freedom and a non-centrality parameter of  $\lambda = \frac{\Delta}{\sigma} \sqrt{N}$ , while the threshold is coming from a centered Student’s  $t$  with  $N - 1$  degrees of freedom. This sample size design formula is not so similar to its ideal counterpart. In particular, it has some important differences:

1. It calls for the sample standard deviation at a stage in which we have not yet performed the experiment, and we need an educated guess of it. In practice, we

tend to be overoptimistic about our experiments, and the guess of the standard deviation used for the sample size design is normally smaller than the true one.

2. It uses the percentiles of the Student's t distribution, which are more difficult to work by hand than the standardized normal distribution. The latter distribution is parameter free (it has zero mean and standard deviation one) and its percentiles are well known (for instance its 95% percentile is  $z_{0.95} = 1.64$ , meaning that a random number drawn from the standardized Gaussian distribution has a probability of 95% of being smaller than 1.64). However, the Student's t distribution has two free parameters, the centrality parameter,  $\lambda$ , and the number of degrees of freedom,  $\nu$ . In this way, the 95% percentile of the Student's t distribution is now  $t_{0.95,\lambda,\nu}$ . To know this number we must specify these two free parameters. In the design formula, the number of degrees of freedom is  $\nu = N - 1$  for both, the confidence and power terms; while the centrality parameter is  $\lambda = 0$  for the confidence term (the Student's t is centered and symmetric) and  $\lambda = \frac{\Delta}{\sigma} \sqrt{N}$  for the power term (the distribution is no longer symmetric).
  3. More importantly, the number of samples,  $N$ , is also a parameter of the distribution (the Student's t has  $N - 1$  degrees of freedom), and there is no analytical solution of this equation, meaning that its exact solution can only be found through numerical algorithms (normally implemented by a computer program).
- Example 45: For the example of the previous section in which the standard deviation of the observations was supposed to be close to 0.54 and we wanted to detect deviations of at least 0.25°C, we would require

$$\Pr \left\{ t_{\frac{0.25}{0.54}\sqrt{N}, N-1} < t_{0.975, 0, N-1} \right\} < 0.1 \Rightarrow N = 51$$

As expected, this sample size is larger than the one in Example 44,  $N = 37$ , because we have to estimate the standard deviation from the data, instead of assuming it is known. Less prior information results in larger sample sizes. We are now using t rather than z and t is always larger than the corresponding z value.

The sample size design formula above is valid for two-tail tests. For one-tail tests, the percentile must be changed from  $t_{1-\frac{\alpha}{2}, 0, N-1}$  to  $t_{1-\alpha, 0, N-1}$ .

### 4.1.3 Confidence interval for the mean

The sample size design for hypothesis test on a single mean with unknown variance (Eq. 4.7) can be used to calculate the sample size needed to estimate a confidence interval of the mean with a given precision. Let us assume that the goal of our research is to determine the mean of a given normal variable, *e.g.* the room temperature at which the thermostat is regulated reducing the uncertainty associated to this determination to a value smaller than a certain limit. This is achieved by a confidence interval. Once we perform the experiment with  $N$  samples, we will have an estimate of the mean and the

standard deviation of the population (see Eq. 4.6). Then, we can construct a confidence interval with confidence  $1 - \alpha$  as

$$\left( \hat{\mu} - t_{1-\frac{\alpha}{2},0,N-1} \frac{\hat{\sigma}}{\sqrt{N}}, \hat{\mu} + t_{1-\frac{\alpha}{2},0,N-1} \frac{\hat{\sigma}}{\sqrt{N}} \right)$$

Let us call  $\Delta$  to the maximum deviation from the mean of the confidence interval

$$\Delta = t_{1-\frac{\alpha}{2},0,N-1} \frac{\hat{\sigma}}{\sqrt{N}}$$

Then, we can easily calculate the number of samples required to achieve this maximum width as

$$N = \left( \frac{t_{1-\frac{\alpha}{2},0,N-1}}{\tilde{\Delta}} \right)^2 \quad (4.8)$$

with  $\tilde{\Delta} = \Delta/\hat{\sigma}$ .

- **Example 46:** Consider the thermostat Example 45, in which we want to construct a 95% confidence interval whose maximum half-width is  $0.25^\circ\text{C}$ . We presume that the standard deviation of the observations will be close to 0.54. The number of samples required for this experiment is

$$N = \left( \frac{t_{0.95,0,N-1}}{0.25/0.54} \right)^2 \Rightarrow N = 21$$

Note that it is much smaller than in the Example 45, the reason being that we only want to construct a confidence interval, rather than testing if the thermostat is malfunctioning with an hypothesis test.

#### Important remarks

46. Constructing confidence intervals is much cheaper in terms of sample size than testing an hypothesis. The reason being that we require the test to have a given power if the difference is at least  $\Delta$ , while the construction of the confidence interval does not care about alternative hypotheses.

#### 4.1.4 Hypothesis test on the mean for paired samples

- **Example 47:** Let us assume we are studying the effect of a new compound A on the intraocular pressure of a mouse strain that serves as a model of glaucoma. The compound is administered in eye drops so that one eye of the mouse can be given the new compound while the other one may serve as its control with only the vehicle being administered. Since an animal serves as its own control, we reduce the variability between subjects and the effect of the treatment is easier to detect due to the lower variance. Let us assume that the intraocular pressure without treatment is about 14.8 mmHg with a standard deviation about 2.2. We want to detect pressure reductions of 0.5 mmHg, and let us assume that the standard deviation with and without treatment is the same.

For each animal we will get two observations (one from the treated eye and another one from the control eye; a good experimental design would randomize for each animal whether the treated eye is the left or right one). Let us call these two observations as  $y_{1i}$  and  $y_{2i}$  where  $i$  refers to the  $i$ -th animal. For each animal we will calculate the difference

$$\Delta y_i = y_{1i} - y_{2i}$$

and the compound A is interesting if the mean of the  $\Delta_i$ 's is negative (there is a decrease of the intraocular pressure after applying the treatment)

$$\hat{\mu}_{\Delta y} = \frac{1}{N} \sum_{i=1}^N \Delta y_i$$

Consequently, our hypothesis test will be of the form

$$\begin{aligned} H_0 : \mu_{\Delta y} &\geq 0 \\ H_a : \mu_{\Delta y} &< 0 \end{aligned} \quad (4.9)$$

Our independent observations are the  $\Delta y$ 's and not the individual  $y$ 's. When analyzing the data we will transform the  $y$  measurements into  $\Delta y$ 's and continue the analysis with them. We no longer have two population of independent measurements (treatment and control), but a single population of measurements, the difference between the two groups, and we want to detect a deviation of this difference from a reference situation. Consequently, we are in the same case as in the previous section and the sample size design formula is the one in Eq. 4.7. However, the important standard deviation for the sample size design is not the one of  $y$ , but the one of  $\Delta y$ , because these are the measurements upon which we will perform the statistical test.  $\Delta y_i$  is constructed as  $\Delta y_i = y_{1i} - y_{2i}$  and its variance is

$$\sigma_{\Delta y}^2 = \sigma_1^2 + \sigma_2^2 = 2\sigma^2$$

This comes from the facts that we have assumed that the variance before and after treatment are equal ( $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ) and that for any random variable constructed in the form

$$Y = aY_1 + bY_2$$

its variance is

$$\sigma_Y^2 = a^2 \sigma_{Y_1}^2 + b^2 \sigma_{Y_2}^2$$

- Example 47 (continued): In this way, we can calculate the sample size as

$$\Pr \left\{ t_{\frac{0.5}{\sqrt{2.2.2}}} \sqrt{N, N-1} < t_{0.95, 0, N-1} \right\} < 0.1 \Rightarrow N = 333$$

333 is the number of pairs we need to study. In this case, the number of pairs coincides with the number of animals, since each animal is providing the two eyes. The large number of pairs, 333, comes from the fact that the normalized effect size is relatively small  $\hat{\Delta} = 0.5/(\sqrt{2} \cdot 2.2) = 0.16$ . The larger the normalized effect size, the smaller the sample size.

### 4.1.5 Hypothesis test on the difference of the mean of two samples

This is, probably, the most common kind of test in biomedical and animal research. We study the difference between the mean of two groups, typically a treatment and a control group. The difference with the case of the previous section is that each subject is not its own control anymore, and the animals in both groups are different. This is, for example, the case of the development of most new drugs. The drug is tested on a treatment group and its effect is compared to a control group.

- **Example 48:** For instance, we may study the systolic blood pressure of mice. In the standard population, it should be around 120 mm Hg with a standard deviation of about 6 mm Hg (although this standard deviation depends on the strain). NZO/HILtJ is a mouse strain with a systolic blood pressure around 130 mm Hg. We are studying the effect of a new drug against hypertension and we want to determine the dose at which the blood pressure drops 5 mm Hg. How many mice of this strain do we need in each group to find these differences with a statistical confidence of 95% and a statistical power of 90%?

As we did in the previous section, let us call  $y_{1i}$  and  $y_{2i}$  the  $i$ -th measurement in the Groups 1 and 2, respectively. As opposed to the previous section, this time the  $i$ -th animal in Group 1 is not the  $i$ -th animal of Group 2. Actually, each animal belongs to only one group. We can estimate the mean and standard deviation of each group as

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{N_1} \sum_{i=1}^{N_1} y_{1i} \\ \hat{\mu}_2 &= \frac{1}{N_2} \sum_{i=1}^{N_2} y_{2i} \\ \hat{\sigma}_1 &= \sqrt{\frac{1}{N_1-1} \sum_{i=1}^{N_1} (y_{1i} - \hat{\mu}_1)^2} \\ \hat{\sigma}_2 &= \sqrt{\frac{1}{N_2-1} \sum_{i=1}^{N_2} (y_{2i} - \hat{\mu}_2)^2}\end{aligned}$$

In the most general case, we will assume that the variance of each group is different. Then, we will estimate the difference between the two groups as

$$\hat{\mu}_{\Delta y} = \hat{\mu}_1 - \hat{\mu}_2$$

and our hypothesis test is

$$\begin{aligned}H_0 &: \mu_{\Delta y} \geq 0 \\ H_a &: \mu_{\Delta y} < 0\end{aligned}\tag{4.10}$$

The variance of the difference between the two groups is

$$\hat{\sigma}_{\hat{\mu}_{\Delta y}}^2 = \frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}$$

The calculation of the sample size boils down to a Student's  $t$  with  $\nu$  degrees of freedom

$$\nu = \frac{\left(\hat{\sigma}_{\hat{\mu}_{\Delta y}}^2\right)^2}{\frac{1}{N_1-1} \left(\frac{\hat{\sigma}_1^2}{N_1}\right)^2 + \frac{1}{N_2-1} \left(\frac{\hat{\sigma}_2^2}{N_2}\right)^2}$$

and a non-centrality parameter

$$\lambda = \frac{\Delta}{\hat{\sigma}_{\mu_{\Delta y}}}$$

The sample size design formulas for this case is

$$\Pr \left\{ t_{\lambda, v-1} < t_{1-\frac{\alpha}{2}, 0, v-1} \right\} < \beta \quad (4.11)$$

Apart from the specific details of the formulas, which are irrelevant from a user perspective because these formulas are implemented in software programs that help the researcher to design the experiment, there are some important lessons to learn from the sample size design formulas seen so far:

#### Important remarks

47. The sample size formula depends on how the data will be analyzed. Specifically, on the test that will be performed (a test on the mean of a sample, on the mean of the difference, on the difference between two means, ...; one-tail or two-tails).
48. It is necessary to know the distribution of the statistic upon which the decision will be taken. For means, the important distributions are the Gaussian (if the variance of the measurements is known) or the Student's t (if the variance is to be estimated from the observations).
49. The normalized effect size plays a crucial role in all designs and each specific case has its own normalization rules.

- Example 48 (continued): In our example  $\Delta = 5$  mmHg. We will assume that the standard deviation in both groups are the same  $\hat{\sigma}_1 = \hat{\sigma}_2 = 6$  mmHg. With this information, we can calculate the sample size that turns out to be  $N = 24$  in each group.

#### 4.1.6 Hypothesis test on the mean of several groups (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique that allows to test whether the mean of a collection of groups, normally called treatments, are all equal. This is a rather common situation in science, and technically it is called 1-way ANOVA because we have only one variable defining the groups (the different treatment applied to each group).

- Example 49: Continuing with the example of the previous section on blood pressure (Example 48), we are simultaneously studying multiple drugs. Each group receives one of the drugs. If at least one of them reduces the blood pressure 5 mm Hg (from 130 of hypertensive mice to 125), then we want to detect this change with a statistical confidence of 95% and a statistical power of 80%?

If there are  $T$  treatments, the ANOVA hypotheses are

$$\begin{aligned} H_0: & \mu_1 = \mu_2 = \dots = \mu_T \\ H_a: & \mu_i \neq \mu_j \text{ for at least two of the treatments} \end{aligned} \quad (4.12)$$

For just two groups, ANOVA is equivalent to the hypothesis test on the difference of the mean of two independent samples (see the previous section). For more than two groups, if the ANOVA test rejects the null hypothesis, then at least one of the groups is different from the rest, but we do not know which one. Then we will perform the so-called *post-hoc* tests to identify which are the two groups that are different. These *post-hoc* tests explicitly account for the multiple comparisons inflation of the Type I error (see Sec. 1.5.3). Amongst the post-hoc procedures Tukey's honestly significant difference test is one of the most popular, but many other exist.

In this section we give the ANOVA sample size formula when all groups have the same size, more general designs with variable group sizes are available. As we will see at length in Sec. 5.1, the ANOVA test ultimately finishes in a Snedecor's F statistic,  $f$ . Under the null hypothesis, this statistic is distributed as a central Snedecor's F with  $T - 1$  and  $N - T$  degrees of freedom. Under the alternative hypothesis, we need to hypothesize some result for which we want to have a specific statistical power. For the example above, we wanted to detect a change of 5 mmHg., in one of the groups. Let  $\alpha_i$  denote the difference between the  $i$ -th treatment and the overall mean. In our example, we had  $\alpha_1 = -5$  and  $\alpha_i = 0$  for all the rest. Then,

$$\sigma_\alpha^2 = \frac{1}{T} \sum_i \alpha_i^2$$

is a sort of "effect size" (what is the average variance with respect to the overall mean by any individual in any of the groups). Under the alternative hypothesis,  $f$  is distributed as a non-central Snedecor's F with  $T - 1$  and  $N - T$  degrees of freedom and with non-central parameter

$$\phi = N \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}$$

where  $\sigma_\varepsilon^2$  is the variance within each one of the treatment groups (for instance, in Example 48, we assumed that the variance within each group was 6 mmHg.) The sample size,  $N$  must be such that

$$F_{T-1, N-T, 0, 1-\alpha} = F_{T-1, N-T, N \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}, \beta}$$

If we apply this formula to the example above we have the following results depending on the number of groups

- Example 49 (continued):

$T$	2	3	4	5	6	7	8	9
$N$	32	29	29	29	30	31	31	32

In this table we recognize three important features:

1. For  $T = 2$ , the ANOVA design requires  $N = 32$  while the two sample design of previous section (Example 48) required only  $N = 26$ . The reason is that the ANOVA is a two-tail test, while in Example 48 we planned the experiment as a one-tail test, and consequently it required less samples.
2. If we have  $T = 3$  groups, we are really interested in the two-tail test, and ANOVA rejects the null hypothesis, then we will perform pairwise comparisons to identify the pair that is making a difference. But, we have only  $N = 29$  samples per group, and this number of animals per group lacks power to identify a difference of  $\Delta = \pm 5\text{mmHg}$  (we require at least  $N = 32$  individuals per group for this test). In this way, we may face the situation in which ANOVA rejects the hypothesis that all groups are the same, but the p-value of all pairwise comparisons are not statistically significant.
3. There is a non-linear relationship between the number of groups and the number of samples per group.

#### Important remarks

50. Sample size designs based on ANOVA are specifically aimed at rejecting the ANOVA null hypothesis (all means are the same), but may not be useful for the post-hoc tests.
51. If post-hoc tests are important in our research, we should design the experiment using the two sample designs of previous section taking into account that we may incur in a Type I error inflation due to multiple testing.

As a simplified design, Mead's resource equation has been proposed. This equation states that the number of samples,  $N$ , must fulfill

$$N - 1 = T + B + E \quad (4.13)$$

where  $T$  is the number of treatments,  $B$  the number of blocks and  $E$  the number of degrees of freedom available for the residuals, which should be between 10 and 20. This equation is based on the number of degrees of freedom consumed by each one of the different components of the variance (see Sec. 5 for a detailed explanation of this decomposition). As can be easily seen, this design does not make any consideration of effect size and power. Although we cannot give an exact number for the effect size addressed by this formula, this can be estimated to be (depending on the number of treatments and blocks) between 1.5 and 2 with a statistical power of 90%. That is, this design is capable of identifying changes in the mean of one of the groups if this change is at least 1.5 times the standard deviation of the observations for each one of the treatments.

A different perspective of similar problems on the sample size calculation for designed experiments is given in Sec. 5.4.

### 4.1.7 Unequal group sizes

In the previous section, the variance of our estimate of the difference is

$$\sigma_{\hat{\mu}_{\Delta y}}^2 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

Actually, we may try to minimize this variance while keeping fixed the total number of samples

$$\min_{N_1, N_2} \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} \quad \text{subject to } N_1 + N_2 = \text{constant}$$

The solution is

$$N_2 = N_1 \frac{\sigma_2}{\sigma_1} \quad (4.14)$$

#### Important remarks

52. That is, we should put more samples in the more variable groups, and if the two groups are equally variable, then the number of samples in both groups will be the same  $N_1 = N_2$ .

Another situation in which we may want to have different group sizes is when the cost of getting samples from Group 1 is different from the cost of getting samples in Group 2. This cost may represent a real economical cost, or the difficulty to find animals with a given condition (animals from Group 1 are 10 times more rare than animals from Group 2, then the cost of Group 1 is 10 times higher than the cost of Group 2). Let us represent the cost of both groups as  $C_1$  and  $C_2$ . Then, we may minimize the variance of the estimate of the difference keeping the cost constant:

$$\min_{N_1, N_2} \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} \quad \text{subject to } C_1 N_1 + C_2 N_2 = \text{constant}$$

The solution is

$$N_2 = N_1 \sqrt{\frac{C_1}{C_2}} \quad (4.15)$$

Note that if  $C_1 > C_2$ , then  $N_2 > N_1$ .

53. That is, we should put more samples in the less costly group.

Finally, if different treatments are to be compared to a control group. Let us assume that we have  $T$  groups receiving  $T$  different treatments and a control group. Let the number of samples in each of the treatment groups be  $N_T$ , while  $N_0$  represents the number of animals in the control group. Similarly to the previous paragraph, the variance of each of the comparisons is

$$\sigma_{\hat{\mu}_{\Delta y}}^2 = \frac{\sigma^2}{N_0} + \frac{\sigma^2}{N_T}$$

We may minimize this variance while keeping fixed the total number of samples

$$\min_{N_0, N_T} \frac{\sigma^2}{N_0} + \frac{\sigma^2}{N_T} \quad \text{subject to } N_0 + TN_T = \text{constant}$$

The solution is (Bate and Karp, 2014)

$$\boxed{N_0 = N_T \sqrt{T}} \quad (4.16)$$

#### Important remarks

54. That is, we should put more samples in the control group, since it will participate in many more comparisons, and diminishing its variance will result into more powerful comparisons.

### 4.1.8 Hypothesis test on the equivalence of two means

Many research experiments respond to the significance test paradigm

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_a &: \mu_1 \neq \mu_2 \end{aligned}$$

If we reject the null hypothesis, then we presume that the true state of affairs is the alternative hypothesis, and the mean in the group of the new treatment is different from the one in the control.

However, some studies respond to the equivalence test paradigm

$$\begin{aligned} H_0 &: \mu_1 \neq \mu_2 \\ H_a &: \mu_1 = \mu_2 \end{aligned}$$

Note that the equal sign has moved from the null hypothesis to the alternative hypothesis. If we reject the null hypothesis, then we presume that the true state of affairs is the alternative hypothesis, and the mean of the treatment and control groups are not different. This is the case, for example, of bioequivalence: we need to show that the effect of our new drug is not different, within limits, from the effect of the reference drug.

Technically, equivalence tests are more difficult than their significance test counterparts. The reason is that the null hypothesis of equivalence tests imply two different tests. To see how this arises, let us first define when two means are considered to be “the same”. Normally, it is assumed that two means are the same if their difference,  $\Delta\mu = \mu_1 - \mu_2$  is small

$$\begin{aligned} H_0 &: \Delta\mu \leq \varepsilon_L \text{ or } \Delta\mu \geq \varepsilon_U \\ H_a &: \varepsilon_L < \Delta\mu < \varepsilon_U \end{aligned}$$

According to the European Medicines Agency Guideline CPMP/EWP/QWP/1401/98, a drug (normally, a new generic coming into the market) is bioequivalent to another (the reference drug) if the effect of the new drug is within a limit from 80% (=0.8) to 125% (=1/0.8) of the effect of the reference (see Fig. 4.1).

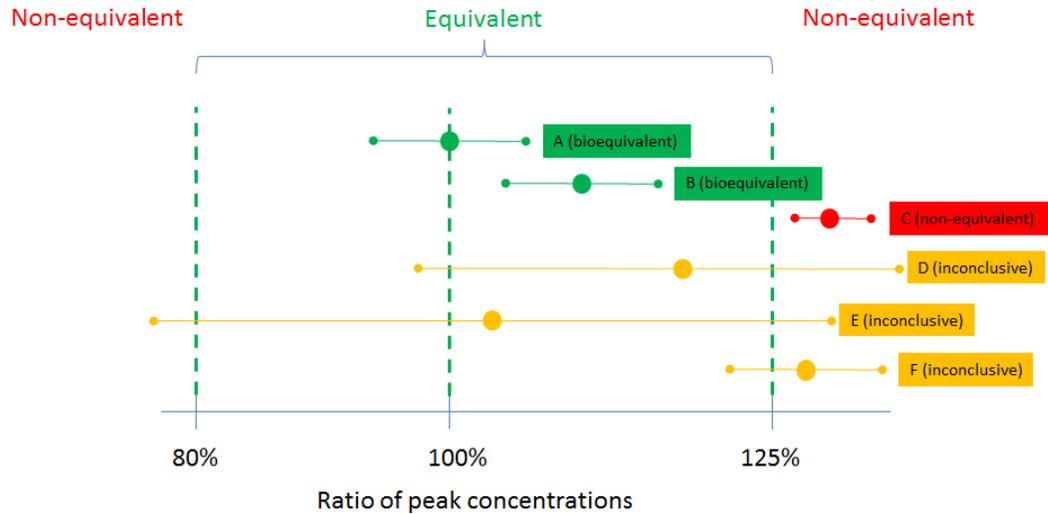


Figure 4.1: Two drugs are said to be bioequivalent if the 95% confidence interval of the ratio of variables of relevance (peak concentration, effect, etc.) is inside the bioequivalent area defined between 80% and 125%. The figure shows six different possible confidence intervals and the interpretation of each one of the results.

- Example 50: We are developing a generic of a drug against hypertension. The reference drug is capable of lowering the mean systolic blood pressure of a mouse model of hypertension from 130 mmHg to 120 mmHg (see Example 48). The effect size of the reference drug is  $\Delta = -10$  mmHg. The new drug is a bioequivalent of the reference if its effect size is between 8 and 12.5 mmHg. From this data, we can compute the lower and upper limits for the equivalence tests.  $\Delta\mu = \mu_{reference} - \mu_{generic}$  and it must be

$$\begin{aligned} 120 - 122 &< \Delta\mu < 120 - 117.5 \\ -2 &< \Delta\mu < 2.5 \end{aligned}$$

Equivalence tests are usually translated into two one-sided t-tests (TOST) by checking two other hypothesis tests

$$\begin{aligned} H_{01} &: \Delta\mu \leq \varepsilon_L \\ H_{a1} &: \Delta\mu > \varepsilon_L \end{aligned}$$

and

$$\begin{aligned} H_{02} &: \Delta\mu \geq \varepsilon_U \\ H_{a2} &: \Delta\mu < \varepsilon_U \end{aligned}$$

Our new drug is bioequivalent to the reference drug if we can reject the two null hypotheses  $H_{01}$  and  $H_{02}$ , as a consequence it must be  $\varepsilon_L < \Delta\mu < \varepsilon_U$ . Each one of these tests is a one-sided t-test of two samples as the one we designed in Eq. 4.11. We will

not give at this moment explicit design formulas as the sample size design software implement them and we have already settled the main ideas of sample size calculations.

- **Example 50 (continued):** For our drug bioequivalence problem we will need  $N_{reference} = N_{generic} = 166$  observations (statistical power of 90% and statistical confidence of 95%). We may compare this sample size with the one of Example 48,  $N = 26$ .

Fig. 4.2 shows the statistical distributions of the two null hypothesis and the alternative hypothesis when  $\Delta\mu = 0$ . Compared to significance tests (Fig. 1.11) we see that in significance tests, the null hypothesis results in a centered distribution of the statistic and the alternative hypotheses are on each side. However, for equivalence tests, it is just the opposite.

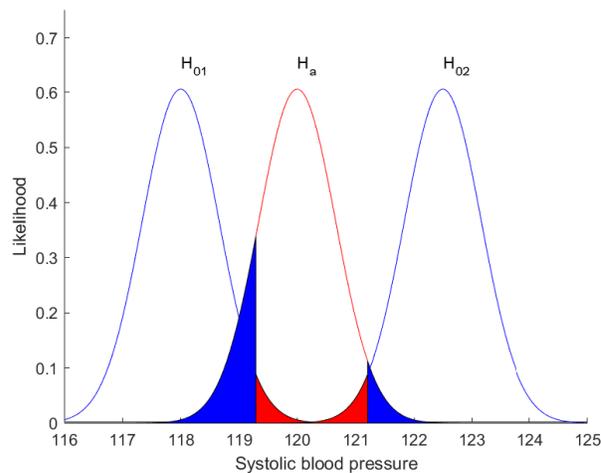


Figure 4.2: The red shaded area is the probability of rejecting any of the null hypotheses if they are true (this area is the complement of the statistical confidence). The blue shaded area is the probability of not rejecting the null hypotheses when the alternative hypothesis is true (only represented for  $\Delta\mu = 0$ ). The symmetry is broken by the 80% and 125% requirement of the guideline.

#### Important remarks

55. Although equivalence tests use the same “ingredients” as significance tests (statistical confidence and power, one-tail statistical tests), they are used in a different manner. Most importantly, significance tests have a single null hypothesis, while equivalence tests have two.
56. It is much more difficult to show equivalence than significance: the number of samples in equivalence tests is normally much higher.

## 4.2 Sample size for proportions

Many research studies aim at identifying the proportion of a population that responds to a given treatment, that has a certain phenotype or that have a given characteristic. As we did with means, experiments with proportions can be performed with one group (we analyze the proportion within a single group) or two groups (we analyze the difference in proportions between two groups). The mathematics associated to proportions are more difficult than those associated to means. We will only show their complexity in its simplest form as a way to grasp the main ideas related to proportions. Many computer programs implement the formulas required for sample size calculations with proportions, and knowing the exact design formulas is not needed in general. We will also give some approximated formulas so that researchers can have an order of magnitude of the sample size required.

It is important to distinguish proportions from other quantities that can also be expressed as percentages. Proportions represent probabilities of events. An animal can be infected with a probability of 50%, or equivalently, in a large population of animals, the proportion of them that can get an infection is 50%. If the area of a skin lesion in an animal increases by 50%, this latter 50% does not have the same nature as the proportion of 50%. The first one refers to a probability, while the second one refers to a variation expressed as a percentage. Proportions are bounded between 0 and 100%, while variations are not.

### 4.2.1 Hypothesis test on one small proportion

- Example 51: We are developing a vaccine against a pathogen. We are only interested in vaccines for which the probability of infection when directly exposed to the pathogen is below 1%. How many individuals do we need to show that a given vaccine is useful?

In this example, the hypothesis test we need is

$$\begin{aligned} H_0 : & p \geq 0.01 \\ H_a : & p < 0.01 \end{aligned}$$

where  $p$  is the probability of infection when directly exposed to the pathogen.

This test is of the form

$$\begin{aligned} H_0 : & p \geq p_U \\ H_a : & p < p_U \end{aligned}$$

where  $p_U$  is an upper bound of the probability of infection. The infection of each animal is modelled by a Bernoulli distribution. It is infected with probability  $p$  and not infected with probability  $1 - p$  (Bernoulli is the distribution of a fair coin flip, we obtain heads with probability  $p = 50\%$  and tail with probability  $1 - p = 50\%$ ). If we have a collection of  $N$  independent Bernoulli events, the number of infections follow a Binomial distribution of parameters  $N$  and  $p$ .

The Bernoulli and Binomial distributions are said to be discrete distributions, as opposed to continuous distributions. Discrete distributions describe the probability of

variables that take discrete values (*e.g.* infected/not infected, number of infections equal to 0, 1, 2, ...); while continuous distributions describe the probability of variables that take continuous variables (*e.g.*, the systolic blood pressure of mice can take any value between 110 and 130 mmHg). Discrete probability distribution assign a probability to each of the possible outcomes of the experiment. In the example of the vaccine, the probability of observing  $x$  infections among the  $N$  mice is

$$\Pr\{X_{\text{infections}} = x\} = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}$$

where  $x!$  is the factorial of the number  $x$  ( $x! = x \cdot (x-1) \cdot (x-2) \cdot \dots \cdot 2 \cdot 1$ , for instance,  $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ ). If we use  $N = 300$  mice and the true probability of infection is  $p = 0.01$ , then Fig. 4.3 shows the probability of observing 0, 1, 2, ... infections. The expected number of infections is

$$\mathbb{E}\{X_{\text{infections}}\} = Np$$

That is, in our example we expect to see  $300 \cdot 0.01 = 3$  infected animals if the true probability of infection after being vaccinated is  $p = 0.01$ . In Fig. 4.3 we can see that  $X = 3$  is the most probable result, with a probability around 22.5%. Interestingly, observing just 2 infections has a probability of 22.4%. We understand this due to the random nature of our observations, and observing 2 or 3 infections in 300 if the probability of infection is 1% seems to be very logical. Observing  $X = 0$  infections would happen with probability 4.9%, and observing  $X = 15$  or more would only happen in 1% of the cases (if we observe 15 or more infections, it would be very unlikely that the true probability of infection is  $p = 1\%$  and we would expect that it is a number closer to  $p = 5\%$ , because  $300 \cdot 0.05 = 15$ ).

Fig. 4.3 shows the distribution under the null hypothesis for  $p = 1\%$ . But the null hypothesis include any value with  $p \geq 1\%$ . For instance, Fig. 4.4 shows a possible alternative distribution ( $p = 0.1\%$ ) and another possible null hypothesis ( $p = 4\%$ )

As we did with the hypotheses for the mean, we need to understand how the data will be analyzed. In this problem, we will use  $N$  animals, and we will reject the null hypothesis ( $H_0 : p \geq p_U$ ), if the probability of observing a number of infections as extreme as  $x_0$  (our actual observations when we do the experiment) is lower than a given threshold,  $\alpha$  (typically,  $\alpha = 0.05$ ). Extreme values of the null hypothesis in this case are small values. For instance, if the true probability of infection is  $p = 4\%$ , then observing only 5 infections or less is only 1.86% (see Fig. 4.4). Note that  $p_U$ , in our example  $p_U = 1\%$ , is the value of the null hypothesis that produces the lowest value, and consequently is the worse case. When we perform the experiment, we will reject the null hypothesis if

$$\sum_{x=0}^{x_0} \Pr\{X_{\text{infections}} = x\} = \sum_{x=0}^{x_0} \frac{N!}{x!(N-x)!} p_U^x (1-p_U)^{N-x} < \alpha \quad (4.17)$$

Note that this equation has two unknowns:  $N$ , the sample size we are trying to calculate, and  $x_0$ , the number of infections we observe when the experiment is performed. But we have not performed the experiment yet! This design equation already highlights several interesting ideas about the design of the sample size for proportions:

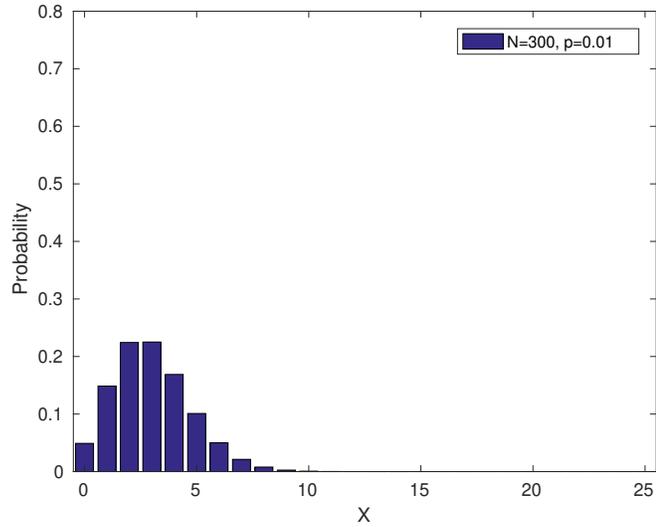


Figure 4.3: Probability of observing  $x = 0, 1, 2, \dots$  infections in  $N = 300$  animals when the probability of being infected is  $p = 0.01$ .

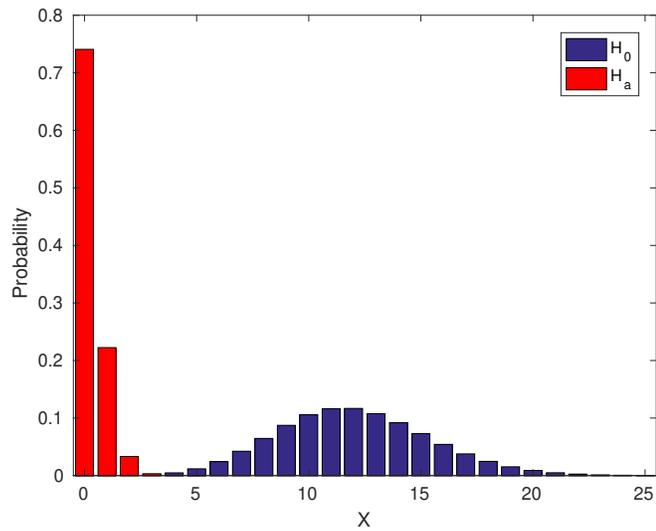


Figure 4.4: Probability of observing  $x = 0, 1, 2, \dots$  infections in  $N = 300$  animals when the probability of being infected is  $p = 0.1\%$  (red) and  $p = 4\%$  (blue).

**Important remarks**

57.  $N$  depends on a free parameter,  $x_0$ , that is chosen by the user at the moment of design. Different choices of the free parameter results in different sample sizes. In this example, the smallest sample size is attained for  $x_0 = 0$ .
58. Making the choice  $x_0 = 0$  does not imply that we are foreseeing that the number of infections will be 0 before performing the experiment. It means that if we perform the experiment and observe  $x_0 = 0$  infections, then we will reject the null hypothesis, with a Type I error smaller than  $\alpha$ .

For  $x_0 = 0$ , the equation above simplifies to

$$(1 - p_U)^N < \alpha$$

and solving for  $N$

$$N > \frac{\log(\alpha)}{\log(1 - p_U)} \quad (4.18)$$

for  $\alpha = 0.05$  and small  $p_U$  (such that  $\log(1 - p_U) \approx -p_U$ ), this equation can be approximated by

$$N > \frac{3}{p_U}$$

that is the famous *rule of 3* used in Epidemiology.

- Example 51 (continued): In our example we would need

$$N > \frac{\log(0.05)}{\log(0.99)} = 298.07$$

that is, we need  $N = 299$  mice. If we perform the experiment of exposing each one of them to the pathogen and none is infected ( $x_0 = 0$ ) we will reject the null hypothesis and assume that the true state of affairs is that the probability of infection is smaller than  $p_U = 1\%$ . If we observe one or more infections, then we cannot reject the null hypothesis, and our vaccine will not be interesting.

**Important remarks**

59. Proving that an event is very rare,  $p_U$  is very low, requires a lot of samples, and the number of samples grows with the inverse of the probability of the event, which can easily grow very quickly as  $p_U$  approaches 0.

We can easily turn a problem with a large proportion into a problem of a low proportion by simply changing the event we look for.

- Example 52: We want to show that more than 99% of the animals in our animal facility are correctly labelled in their cages. Our test would be of the form

$$\begin{aligned} H_0 : & p < 0.99 \\ H_a : & p \geq 0.99 \end{aligned}$$

Instead of having an upper bound of the probability (as in the case of infections), we have a lower bound. In principle, we have not developed the theory for handling these situations, but we can easily do by changing the event we look for. Instead of looking for correctly labeled mice, we may look for mislabeled mice. Then, the test would turn into

$$\begin{aligned} H_0 &: p \geq 0.01 \\ H_a &: p < 0.01 \end{aligned}$$

#### Important remarks

60. We can turn superiority tests into inferiority tests or viceversa simply by looking at a different event.

### 4.2.2 Confidence interval for one proportion

Sometimes we are interested in determining a proportion with a given precision.

- **Example 53:** We are interested in determining the proportion of animals that will develop cancer when they are directly exposed to a given carcinogen. We want to report a confidence interval rather than a point estimate, and we want that our confidence interval is at most 5% wide (for instance, if this proportion is 15%, we want the 95% confidence interval to be between 12.5 and 17.5%). How many animals do we need to expose to achieve this precision?

An important observation when addressing this problem is that solving for the sample size in this problem requires an *a priori* estimate of the proportion we are looking for. It may sound counter-intuitive that we need an estimate of the proportion before performing an experiment whose goal is to estimate it. However, the uncertainty about a proportion increases as the proportion approaches 50%, and in this way, we may achieve the same precision with fewer animals if the proportion we look for is closer to the extremes (0% or 100%). In the example of the carcinogen, let us assume that we expect the proportion we want to estimate to be around 15%.

Let us assume that when we perform the experiment we observe  $x_0$  cancers in  $N$  animals. The estimate of the proportion will be

$$\hat{p} = \frac{x_0}{N}$$

and the 95% confidence limit will be of the form  $[p_L, p_U]$  (lower and upper bound, respectively) such that the probability of observing values as extreme as  $x_0$  is  $\alpha/2$  for  $p_L$  and for  $p_U$ :

$$\begin{aligned} \sum_{x=x_0}^N \frac{N!}{x!(N-x)!} p_L^x (1-p_L)^{N-x} &= \frac{\alpha}{2} \\ \sum_{x=0}^{x_0} \frac{N!}{x!(N-x)!} p_U^x (1-p_U)^{N-x} &= \frac{\alpha}{2} \end{aligned} \tag{4.19}$$

We have now two unknowns ( $x_0$  and  $N$ ) with two equations and we must find values such that  $p_U - p_L < \Delta_p$ , being  $\Delta_p$  the desired width of the confidence interval (in our example,  $\Delta_p = 5\%$ ). Obviously, this is not an easy task, and we may try to find some alternative procedure that results in a more easy approach. Such a procedure is offered by approximations. The Binomial distribution of parameters  $N$  and  $p$  can be safely approximated by a Gaussian of mean  $\mu = Np$  and variance  $\sigma^2 = Np(1-p)$  if  $Np > 5$  and  $N(1-p) > 5$  (see the binomial distribution for  $H_0$  in Fig. 4.4). In this case, we may design the sample size using the standard sample size design for the Gaussian means. Without entering into the mathematical details, the solution of this problem is

$$N > \left( \frac{z_{1-\frac{\alpha}{2}}}{\frac{\Delta_p/2}{\sqrt{p(1-p)}}} \right)^2 \quad (4.20)$$

This formula resembles Eq. 1.15: a numerator that depends on the confidence level and a denominator that is a normalized effect size.

#### Important remarks

61. Designing the sample size for discrete variables can be rather cumbersome mathematically, but in some situations we may find alternative, approximated, procedures that provide a useful answer for the problem at hand.
62. However, we should not forget that these approximations are just approximations. They provide an order of magnitude and not a precise answer.
63. Additionally, the sample size calculation requires an initial guess of the proportion we are looking for.

- Example 53 (continued): Now it is very easy to calculate the sample size with the approximate formula:

$$N > \left( \frac{z_{1-\frac{0.05}{2}}}{\frac{0.05/2}{\sqrt{0.15 \cdot 0.85}}} \right)^2 = \left( \frac{1.96}{\frac{0.025}{0.357}} \right)^2 = 783.7$$

That is, we need to expose 784 animals to the carcinogen to have such a precise confidence interval (the exact solution is 822, we see that the approximated solution is in the same order of magnitude, but this time fell a bit short). The reason for this large number is that the effect size,  $0.05/2 = 0.025$  is relative small compared to the standard deviation of the estimate of the proportion, 0.357. If we cannot afford such a large number of animals, we will have to sacrifice precision. If the maximum number of animals we can afford is 100, then the precision will drop down to  $\Delta_p = 14\%$

$$N > \left( \frac{z_{1-\frac{0.05}{2}}}{\frac{0.14/2}{\sqrt{0.15 \cdot 0.85}}} \right)^2 = \left( \frac{1.96}{\frac{0.025}{0.357}} \right)^2 = 99.9$$

That is, if the proportion of animals with cancer is, as expected,  $p = 15\%$ , then the 95% confidence interval will extend from  $p_L = 8\%$  to  $p_U = 22\%$ .

We should now verify that the conditions for the approximation hold

$$\begin{aligned} Np &= 100 \cdot 0.15 = 15 > 5 \\ N(1-p) &= 100 \cdot 0.85 = 85 > 5 \end{aligned}$$

If they do not hold, then the sample size calculated by the approximated procedure will not be close to the true sample size.

If we do not want to make any assumption about the expected proportion, we may make the design in the worse case, for which the uncertainty is maximum, that is  $p = 0.5$ , but the sample size quickly grows as shown by the following calculation:

$$N > \left( \frac{z_{1-\frac{0.05}{2}}}{\frac{0.14/2}{\sqrt{0.5 \cdot 0.5}}} \right)^2 = \left( \frac{1.96}{\frac{0.025}{0.5}} \right)^2 = 195.9$$

#### Important remarks

64. There is a trade-off between sample size and precision of the confidence interval. More precise confidence intervals, smaller  $\Delta_p$ , require more samples; conversely, experiments with a low number of samples result in less precise confidence intervals for the proportion.
65. It is easier to be precise in the confidence interval of proportions as they go away from the region of maximum uncertainty,  $p = 50\%$ . The number of samples for these proportions will be smaller than for proportions close to 50%.

### 4.2.3 Hypothesis test on one proportion

- Example 54: The infection rate of a given pathogen is 5% when adult animals are directly exposed to it. We suspect that the infection rate of newborns is higher. How many newborns do we need to study to test this hypothesis with a 95% confidence level and if we want to have a statistical power of 90% if the infection rate goes above 10%?

Our test is of the form

$$\begin{aligned} H_0 &: p \leq 0.05 \\ H_a &: p > 0.05 \end{aligned}$$

As in all tests, there will be a distribution associated to the null hypothesis and another one to the alternative hypothesis. There will be a threshold,  $x_0$ , beyond which we will reject the null hypothesis for being very unlikely (with a Type I error rate, that is, the null hypothesis is actually true, but with the evidence collected we reject it, which in the example is 5%). Fig. 4.5 shows this situation. The area of the  $H_0$

distribution to the right of  $x_0$  is  $\alpha = 0.05$ , while the area of  $H_a$  to the left of  $x_0$  is  $\beta = 0.1$  (the Type II error). Our task is to find  $N$  and  $x_0$  that fulfill these constraints

$$\begin{cases} \sum_{x=x_0}^N \frac{N!}{x!(N-x)!} p_0^x (1-p_0)^{N-x} < \alpha \\ \sum_{x=0}^{x_0} \frac{N!}{x!(N-x)!} p_a^x (1-p_a)^{N-x} < \beta \end{cases} \quad (4.21)$$

where  $p_0$  is the upper limit of the null hypothesis ( $p_0 = 0.05$  in our example) and  $p_a$  is the probability at which we want to have a given statistical power ( $p_a = 0.1$  in our example).

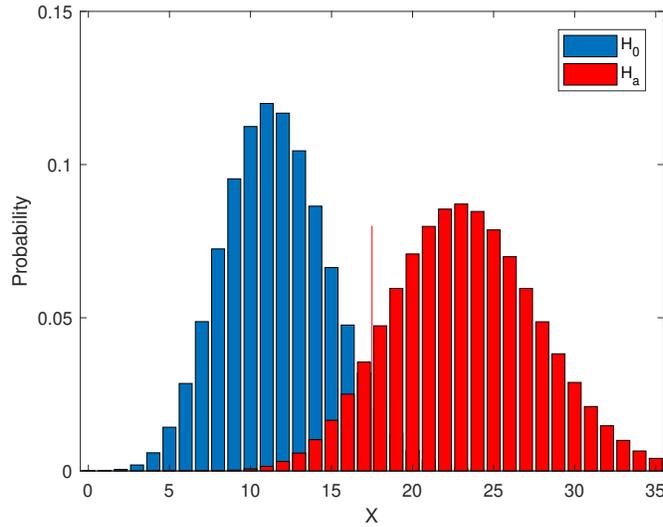


Figure 4.5: Probability of observing  $x = 0, 1, 2, \dots$  infections in  $N = 233$  animals when the probability of being infected is  $p = 5\%$  (blue) and  $p = 10\%$  (red).

However, this sample size design does not lend itself to easy calculations. As we did in the previous section, we may use approximate methods. If the binomials can be approximated by Gaussians (this can be done if  $Np > 5$  and  $N(1-p) > 5$ ), then we may use a Gaussian sample size calculation formula

$$N \geq \left( \frac{z_{1-\alpha} \sqrt{p_0(1-p_0)} + z_{1-\beta} \sqrt{p_a(1-p_a)}}{p_a - p_0} \right)^2 \quad (4.22)$$

that is much better suited for hand calculations.

- Example 54 (continued): The approximated method gives for this case

$$N \geq \left( \frac{z_{0.95} \sqrt{0.05 \cdot 0.95} + z_{0.90} \sqrt{0.1 \cdot 0.9}}{0.1 - 0.05} \right)^2 = 221$$

The exact method gives  $N = 233$  and  $x_0 = 18$ , meaning that when we perform the experiment, if we observe less than 18 infections, we cannot reject the hypothesis that the probability of infection is smaller or equal 5%. If we observe 18 infections or more, then we reject that hypothesis and, as suspected, the infection rate in newborns is larger than in adults. The Gaussian approach has the disadvantage that it does not give the threshold  $x_0$  in advance. However, this is a minor drawback since all the statistical software for analyzing the results of the experiment will calculate it, probably internally, and report the p-value of our observations.

#### Important remarks

66. Hypothesis tests with proportions operate in the same way as with the mean: there is a distribution associated to the null hypothesis, another one with the alternative hypothesis, and there is a threshold that is used to take the decision to reject the null hypothesis or not. The sample size is calculated such that the area of the Type I and II errors are the ones specified at the design,  $\alpha$  and  $\beta$ . Unlike the designs for the mean, the threshold itself is part of the design problem and, if the sample size is calculated in an exact way, it must be also calculated along with the  $N$ .

#### 4.2.4 Confidence interval for the difference of two proportions

- Example 55: For the example above, let us say that we want to construct a 95% confidence interval on the difference between the two proportions: the proportion of infected adults and the proportion of infected newborns:

$$p_L < p_{\text{newborn}} - p_{\text{adults}} < p_U$$

For doing so, we will study two groups (adults and newborns) and estimate the proportion of infections in each of the groups. We foresee that  $p_{\text{newborn}}$  is around 10% and  $p_{\text{adults}}$  around 5%. Since both proportions are rather close, we want the confidence interval to be very precise such that  $p_U - p_L < 5\%$ .

In this problem, the statistical variable of interest is

$$\Delta p = p_1 - p_2$$

Let us call the interval width as  $\Delta$

$$\Delta = p_U - p_L$$

$\Delta$  is our main design parameter and it represents how precise we want to be around the observed difference. If the Gaussian approximation of the binomial can be applied ( $Np > 5$  and  $N(1-p) > 5$ ), then the sample size design formulas are

$$\begin{aligned} N_1 &= \left( \frac{\frac{z_{1-\frac{\alpha}{2}}}{\Delta/2}}{\sqrt{p_1(1-p_1)+p_2(1-p_2)R}} \right)^2 \\ N_2 &= \left( \frac{\frac{z_{1-\frac{\alpha}{2}}}{\Delta/2}}{\sqrt{p_1(1-p_1)\frac{1}{R}+p_2(1-p_2)}} \right)^2 \end{aligned} \quad (4.23)$$

where  $R = N_1/N_2$  is a ratio between the size of the two groups. This ratio is very convenient if it is easier to obtain animals in one of the groups than in the other (*e.g.* it is easier to have access to adults than newborns). From this sample size design formula we can reinforce concepts we already have

#### Important remarks

67. Counterintuitively, for determining the sample size we need to assume the proportion of infections in both groups,  $p_1$  and  $p_2$ , before doing the experiment! This is a problem of most sample size calculations working with proportions, and there is no workaround. The reason is that the mean and the variance of the underlying distribution, the Binomial, depends on the proportion (for instance, the Gaussian distribution does not have this property: the variance of the Gaussian does not depend on its mean). In practice, it means that before doing the experiment we need to foresee based on previous experiments or the literature, which are reasonable estimates for these proportions and use them in the sample design. When we actually perform the experiment, we will see how correct or incorrect we were about our initial guess, and how correct or incorrect our sample size actually is.
68. The sample size grows with the inverse of the precision,  $\Delta$ . As  $\Delta$  approaches 0, the number of samples for the experiment rockets.
69. Very high or very low proportions have lower uncertainty than proportions around  $p = 50\%$  (the term  $\sqrt{p_1(1-p_1)+p_2(1-p_2)}$  is smaller for  $p_1$  and  $p_2$  close to 0 or 1 than for  $p_1$  and  $p_2$  close to 0.5).

- **Example 55 (continued):** Continuing with the example and assuming that we will study the same number of animals on both groups we require

$$N_1 = N_2 = \left( \frac{\frac{z_{0.975}}{0.05/2}}{\sqrt{0.05 \cdot 0.95 + 0.1 \cdot 0.9}} \right)^2 = 846$$

That is, we require 846 animals per group. Although,  $p_1$  and  $p_2$  are close to 0, the required precision,  $\Delta = 0.05$ , is relatively high resulting in a very large

number of animals. On the other side, since  $p_1$  and  $p_2$  are expected to be so close to each other, having a smaller precision (for instance,  $\Delta = 0.15$  results in only 94 animals per group) would probably make the results of the experiment useless, because the confidence interval would probably include 0, leaving us with the uncertainty if there is really a difference between the infection rate in adults and newborns.

#### Important remarks

70. Calculating the sample size before performing the experiment allows us to take the decision of embarking into the experiment (and enrolling at least 846 animals per group) or not (because we cannot afford so many animals and the scientific knowledge gain from fewer animals does not justify the experiment).

### 4.2.5 Hypothesis test on the difference of two proportions

- Example 56: Following the last two examples, we are interested in testing if there is a difference in the infection rate of a pathogen in adults and newborns. We expect the infection rate in newborns to be higher than the one in adults (which is expected to be around 5%). If the difference is larger than 5% (that is, the infection rate in newborns raises above 10%), we want to be able to see it with a statistical power of 90%. The confidence level is set to the standard 95%.

Our statistical test is of the form

$$\begin{aligned} H_0 &: p_{\text{newborn}} \leq p_{\text{adults}} \\ H_a &: p_{\text{newborn}} > p_{\text{adults}} \end{aligned}$$

- Example 56: We are interested in checking if transportation of the animals affect the proportion of pregnancy in females afterwards. For doing so we will have two groups of females: one will be transported in a trip of 24h and another one will stay in the animal facility. We expect that the number of pregnancies is about 80%. How many animals do we need in each group if we want to detect differences of at least 10% with a confidence level of 95% and a statistical power of 90%?

We may extend the sample design formula for confidence intervals in Eq. 4.23 to hypothesis tests

$$N_1 = N_2 = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\Delta}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}} \right)^2 \quad (4.24)$$

where we have assumed the same number of samples in both groups ( $R = 1$ ). As a technical note, there are many possible designs for this research problem, this one is called Z-test unpooled variance. The other designs differ in the assumptions, the calculations and the results, although all of them should result in a similar sample size.

Note an important difference between the design for the confidence interval and the design for the superiority test:  $\Delta/2$  in the confidence interval design (Eq. 4.23) has turned into  $\Delta$  for the superiority test (Eq. 4.24). This results in a large reduction of the sample size.

- Example 56 (continued): The required sample size for this example would be

$$N_1 = N_2 = \left( \frac{z_{0.95} + z_{0.9}}{\frac{0.05}{\sqrt{0.05 \cdot 0.95 + 0.1 \cdot 0.9}}} \right)^2 = 472$$

#### Important remarks

70. In Examples 54, 56, and 56, we have seen three different flavours of the same problem: 1) comparing the proportion of a group to a reference (Example 54), 2) computing a confidence interval for the difference of two groups (Example 56), and 3) showing that the proportion of a group is larger than the proportion in another group (Example 56). The sample size varies wildly ( $N = 221, 846,$  and  $472,$  respectively). This highlights, once again, the need to plan the experiment in advance and decide exactly which is the goal of our experiment.

#### 4.2.6 Hypothesis test on the difference of two paired proportions

- Example 57: We are interested in knowing if a drug has an effect in a particular symptom of a disease. For doing so, we will check a number of diseased animals and check whether they had the symptom or not before treatment with the drug. Then, we will administer the drug and check whether the symptom is present or not. Before administering the treatment, we expect that 50% of the animals have the symptom. We want to have a statistical power of 90% if the presence of the symptom drops to 20%. The statistical confidence of the test is set to 95%. How many animals do we need for this test?

After performing the experiment we can organize the observations in a table depending on whether the animals have the symptom or not before and after treatment:

		After	
		Absent=0	Present=1
Before	Absent=0	$n_{00}$	$n_{01}$
	Present=1	$n_{10}$	$n_{11}$

where the  $n_{ij}$  are counts of individuals. The total count of individuals is

$$N = n_{00} + n_{01} + n_{10} + n_{11}$$

There is a fundamental difference between this contingency table and a standard contingency table: the animals before and after treatment are the same. This is the equivalent

for proportions of paired measurements. This situation also happens if we measure the absence/presence of a feature in two different locations of the same animal, or the absence/presence of a feature in an animal and its sibling. The same individual serves as its own control or, at least, there is a matched control (as in the case of siblings). The standard tools for contingency tables (like the  $\chi^2$ -test) do not apply because those tools are designed for independent samples, and not multiple measures on the same individual. The appropriate tool is McNemar's test that determines if the row and column marginal distributions are equal. Technically, the test is

$$\begin{aligned} H_0 &: p_{01} = p_{10} \\ H_a &: p_{01} \neq p_{10} \end{aligned}$$

where  $p_{01} = n_{01}/N$  and  $p_{10} = n_{10}/N$ . Note that  $p_{01}$  and  $p_{10}$  are the discordant frequencies (they had symptoms before but not after, or viceversa). So McNemar's test checks whether the proportion of discordant events is the same in both directions.

The number of animals for the experiment can approximately be calculated with the help of two proxy variables: the total proportion of discordant events and the odds ratio between both kinds of discordant events

$$\begin{aligned} p_D &= p_{10} + p_{01} \\ OR &= p_{10}/p_{01} \end{aligned}$$

Then,

$$N = \left( \frac{z_{1-\frac{\alpha}{2}}(OR+1) + z_{1-\beta} \sqrt{(OR+1)^2 - (OR-1)^2 p_D}}{(OR-1)\sqrt{p_D}} \right)^2 \quad (4.25)$$

- Example 57 (continued): We must translate our previous expectations into proportions in each one of the cells. The following table shows this decomposition

		After		
		Absent=0	Present=1	
Before	Absent=0	$p_{00}=40\%$	$p_{01}=10\%$	50%
	Present=1	$p_{10}=40\%$	$p_{11}=10\%$	50%
		80%	20%	

For the sample size calculation we have:

$$\begin{aligned} p_D &= 0.4 + 0.1 = 0.5 \\ OR &= 0.4/0.1 = 4 \\ N &= \left( \frac{z_{0.975}(4+1) + z_{0.9} \sqrt{(4+1)^2 - (4-1)^2 0.5}}{(4-1)\sqrt{0.5}} \right)^2 = 55 \end{aligned}$$

The exact design formula (not shown here) gives  $N = 59$ .

**Important remarks**

71. As usual in the sample size design with proportions, we need to make use of the expected proportions in each of the cases before doing the experiment. This is a bit unnatural to researchers, but there is no other way of making the design, and it is better having a rough guess of the number of samples to show a given effect than taking a fixed number, *e.g.*  $N = 30$ , for discovering later (*post-mortem* analysis) that we have fallen too short or large the number of samples required for our purposes used too few or too many animals.
72. The term  $(OR - 1)\sqrt{p_D}$  plays the role of the effect size. As this term approaches to 0 (because the two discordant proportions are very similar and  $OR \approx 1$ , or because there are very few discordant events,  $p_D \approx 0$ ), the number of samples required for our experiment grows very quickly. This is logical because in these two cases, it will be very difficult to show that  $p_{01} \neq p_{10}$ .

**4.2.7 Hypothesis test on the difference of multiple proportions**

- **Example 58:** We want to verify if there is a relationship between the incidence of a given pathology and genotype and sex. We will study four genotypes ( $G_1$ ,  $G_2$ ,  $G_3$ , and  $G_4$ ) that we will assume equiprobable. If there is no relationship, then we should observe 50% of male and female diseased animals at all genotypes. If there is, then in some of the genotypes we may observe a deviation from this 50%. We want to have a statistical power of 90% if the deviation is larger than 10%. We want to have a statistical confidence of 95%. How many diseased animals do we need to observe to test this hypothesis?

This kind of studies are addressed through a *contingency table*, in the example above of diseased animals. When we perform the experiment we record in this table how many animals we have observed of each kind

		Genotype			
		$G_1$	$G_2$	$G_3$	$G_4$
Sex	Male=1	$n_{11}$	$n_{12}$	$n_{13}$	$n_{14}$
	Female=2	$n_{20}$	$n_{21}$	$n_{23}$	$n_{24}$

At the moment of experiment design we cannot input the number of animals observed because the experiment has not started yet. Instead, we will input the expected probabilities at each of the cells. If there is no higher or lower incidence of the disease with sex and/or genotype, all cells should have the same probability as shown below.

		Genotype			
		$G_1$	$G_2$	$G_3$	$G_4$
Sex	Male=1	$p_{11}^0 = 0.125$	$p_{12}^0 = 0.125$	$p_{13}^0 = 0.125$	$p_{14}^0 = 0.125$
	Female=2	$p_{21}^0 = 0.125$	$p_{22}^0 = 0.125$	$p_{23}^0 = 0.125$	$p_{24}^0 = 0.125$

We will refer to these probabilities as  $p_{ij}^0$ , the  $i$  and  $j$  refer to the cell and  $^0$  to the fact that this is our *a priori* assumption, and it will be the distribution under the null hypothesis. Note that the probability in each of the cells is  $0.125 = 0.5 \cdot 0.25$ , 0.5 because this is the *a priori* probability of male and female, 0.25 because each of the 4 genotypes is equiprobable, and the product because we assume that sex and genotype are independent (they would not be independent if for a given genotype, more males are born than females or viceversa, and this is different to what happens in the species in general).

If in any of the groups the probability of diseased males and females is unbalanced (*e.g.* males suffer more frequently the disease than females), then we would observe a different distribution of probabilities. Just for illustrating the idea, let us assume that this happens in Genotype 1, and that the deviation is 10%. Then the expected table of probabilities would be

		Genotype			
		$G_1$	$G_2$	$G_3$	$G_4$
Sex	Male=1	$p_{11}^a = 0.6 \cdot 0.25 = 0.15$	$p_{12}^a = 0.125$	$p_{13}^a = 0.125$	$p_{14}^a = 0.125$
	Female=2	$p_{21}^a = 0.4 \cdot 0.25 = 0.1$	$p_{22}^a = 0.125$	$p_{23}^a = 0.125$	$p_{24}^a = 0.125$

We have labelled this probability distribution as  $^a$  because it is associated to the alternative hypothesis. Remember that at this specific distribution we wanted to have a statistical power of 90%. We may compute the difference between the two distributions (null and alternative) as

$$w = \sqrt{\sum_{ij} \frac{(p_{ij}^0 - p_{ij}^a)^2}{p_{ij}^0}}$$

We have named the difference as  $w$  and it is the effect size. In the example above,  $w = 0.1$ . Then, we need some equation that allows us to find the sample size. This equation is given, as usual, by a threshold such that from this threshold to the side of the alternative hypothesis, the null hypothesis has a probability  $\alpha$ ; and from this threshold to the side of the null hypothesis, the alternative hypothesis has a probability  $\beta$ . The test we will do to analyze this data is a  $\chi^2$ , and the sample size design equation is

$$\chi_{1-\frac{\alpha}{2}, 0, df}^2 = \chi_{\beta, Nw^2, df}^2 \quad (4.26)$$

where  $df = (R - 1)(C - 1)$  is the number of degrees of freedom of the  $\chi^2$  and it is calculated as a function of the number of rows,  $R$ , and columns,  $C$ , of the contingency table (in our example,  $df = (2 - 1)(4 - 1) = 3$ ), and the parameter  $Nw^2$  is the non-centrality parameter of the  $\chi^2$ .

- **Example 58 (continued):** In our example,  $w = 0.1$ ,  $\alpha = 0.05$  and  $\beta = 0.1$ , and the sample design equation is

$$\chi_{0.975, 0, 3}^2 = \chi_{0.1, N \cdot 0.01, 3}^2$$

The solution is  $N = 1,418$ . The reason for such a high number is that the effect size is very small. Once again, we realize of the importance of designing

the experiment in advance and adjusting our expectations to the sample size, or adjusting the sample size to our requirements.

### 4.2.8 Hypothesis test on the equivalence of one proportion

- Example 59: We are exploring a new administration route for a drug. Normally,  $p = 60\%$  of the animals respond to the drug. How many animals do we need to study to show that the new route is equivalent to the previous one? We want to have a power of 90% when the number of responders is  $p_1 = 50\%$  or  $p_2 = 70\%$ .

As was shown in Sec. 4.1.8, equivalence tests are translated into two tests and they normally require more samples than standard significance tests. In the case of proportions, this is also the case. The equivalence test can be written as

$$\begin{aligned} H_0 : p &\leq p_{0L} \text{ or } p \geq p_{0U} \\ H_a : p_{0L} &< p < p_{0U} \end{aligned}$$

$p_{0L}$  and  $p_{0U}$  are the lower and upper bounds such that the proportion is still considered to be the same as the nominal value. This test is translated into two one-sided tests (TOST, see Fig. 4.6).

$$\begin{aligned} H_{0L} : p &\leq p_{0L} \\ H_{aL} : p_{0L} &< p \end{aligned}$$

and

$$\begin{aligned} H_{0U} : p &\geq p_{0U} \\ H_{aU} : p &< p_{0U} \end{aligned}$$

If both null hypotheses are rejected, then we would accept the alternative hypothesis and, in this way, we would have shown that the proportion is within the specified limits ( $p_{0L}$  and  $p_{0U}$ ).

$H_0$  is rejected if the number of observed samples is in the region  $x_1 < x < x_2$  with

$$\begin{aligned} \sum_{x=x_1}^N b(x; N, p_{0L}) &< \alpha \\ \sum_{x=0}^{x_2-1} b(x; N, p_{0U}) &< \alpha \end{aligned}$$

and the power of the test

$$\sum_{x=x_1}^{x_2} b(x; N, p) < 1 - \beta$$

where  $b(x; N, p)$  is the probability of observing  $x$  successes when  $N$  samples are drawn from a binomial distribution, and the probability of success is  $p$ . We need to find numbers  $x_1$ ,  $x_2$  and  $N$  satisfying the equations above.

These equations can be solved numerically with a computer. An approximate solution of the number of samples is obtained if the three binomials can be approximated by Gaussians and  $p = \frac{p_{0L} + p_{0U}}{2}$ , then we can solve for  $N$  as

$$N \geq \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{p_{0U} - p_{0L}}{\sqrt{(p_{0L} + p_{0U})(2 - p_{0L} - p_{0U})}}} \right)^2 \quad (4.27)$$

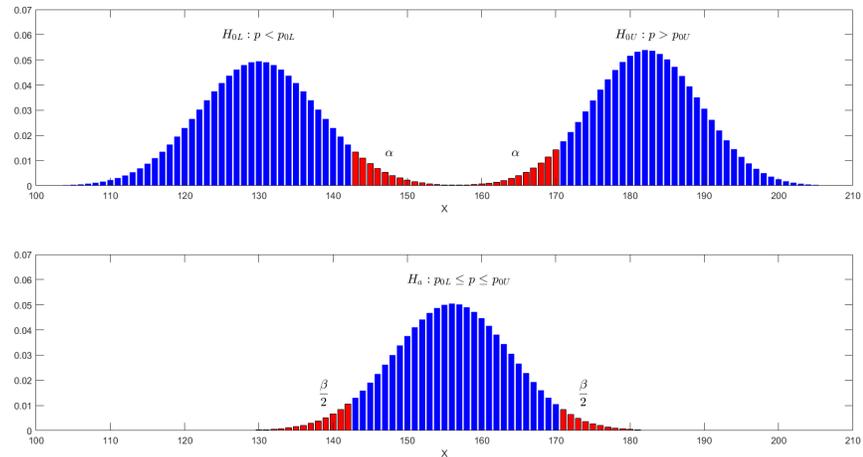


Figure 4.6: Top: Representation of the two null hypotheses for the equivalence of two proportions. Bottom: Representation of the alternative hypothesis.

- Example 59 (continued): The exact solution requires  $N = 255$  samples. We would reject the null hypothesis if the number of respondents is between  $x_1 = 142$  and  $x_2 = 165$ . The approximate solution gives

$$N \geq \left( \frac{z_{0.95} + z_{0.95}}{\frac{0.7-0.5}{\sqrt{(0.5+0.7)(2-0.5-0.7)}}} \right)^2 = 260$$

#### Important remarks

73. From the design, we get the number of samples  $N$ , and two limits,  $x_1$  and  $x_2$ , such that we can decide when the experiment is done whether we should reject the null hypothesis or not.

### 4.2.9 Hypothesis test on the equivalence of two proportions

- Example 60: We can solve the same problem as in Example 59, but estimating the proportion from two populations: one with the standard administration route and another one with the alternative route.

An appropriate hypothesis test is

$$\begin{aligned} H_0 &: |p_1 - p_2| \geq \Delta \\ H_a &: |p_1 - p_2| < \Delta \end{aligned}$$

where  $\Delta$  is the maximum difference between the proportions in the two groups,  $p_1$  and  $p_2$ , such that both groups are still considered to have the same proportion. As in the previous case, we can solve the problem with two one-sided tests (TOST), for which an exact solution exist based on binomial counting (as in the previous case). In this section we will not give the formulas, which are more complicated than in the previous section, but they have the same flavour.

If the binomials associated to the different options can be approximated by Gaussians, then  $\widehat{\Delta p} = p_1 - p_2$  is approximately normal with variance

$$\sigma_{\widehat{\Delta p}}^2 = \frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}$$

In the equivalence case,  $p_1$  is supposed to be equal to  $p_2$  (and referred to as  $p$ ), and the optimal allocation size gives  $N_1 = N_2 = N$  so that

$$\sigma_{\widehat{\Delta p}}^2 = \frac{2p(1-p)}{N}$$

The sample size for a given power  $1 - \beta$  is

$$N \geq \left( \frac{z_{1-\alpha} + z_{1-\frac{\beta}{2}}}{\frac{\Delta}{\sqrt{2p(1-p)}}} \right)^2 \quad (4.28)$$

where  $\Delta$  is the maximum deviation for which the two proportions are still considered to be equivalent (that is,  $|p_1 - p_2| < \Delta$ )

- Example 60 (continued): The exact solution requires  $N = 520$  samples per group. The approximate solution gives

$$N \geq \left( \frac{z_{0.95} + z_{0.95}}{\frac{0.1}{\sqrt{2 \cdot 0.6(1-0.6)}}} \right)^2 = 520$$

#### Important remarks

74. If we compare the sample size for an experiment with just one proportion (Example 59 or two proportions (the example above), we see that the size for two proportions is much larger. The reason is that with two proportions there is much more “uncertainty” involved since we need to estimate the difference in proportion for two groups, instead of just one.

### 4.3 Sample size for regression

Regression is a statistical procedure to estimate the relationship between random variables. In its simplest form, let us consider a random variable  $Y$ , called the dependent

or predicted variable, and another random variable  $X$ , called the independent variable or the predictor. Then, we can find a linear relationship between the two:

$$Y = b_0 + b_1X \quad (4.29)$$

An important observation is that  $X$  is supposed to be a random variable and cannot be controlled by the experimenter. For instance, the weight and the length of the mice can both be measured by the experimenter, but he cannot influence any of the two. Regression in this case determines if there exists a linear relationship between the weight ( $Y$ ) and length ( $X$ ). We can extend this model to  $p$  predictors

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

This is different to an experiment in which the researcher checks on the effect of different growth hormone doses on the weight of the mice. He tries different hormone doses ( $X$ ) and measures the observed weights ( $Y$ ) for each of the dose levels. The assumption is that for each hormone level ( $X$ ), there corresponds a “unique and unobservable” weight value ( $b_0 + b_1X$ ) that is “corrupted” by measurement errors and the intrinsic variability associated to each individual animal ( $E$ )

$$Y = b_0 + b_1X + E \quad (4.30)$$

In this model,  $X$  is not a random variable, but deterministic and totally specified by the experimenter. In this model, the randomness in the observations ( $Y$ ) is only caused by the observation errors ( $E$ ). Fig. 4.7 illustrates the case of random and deterministic predictors. Both problems are statistically called a regression problem, and the sample design formula follows the same principles in both cases.

Consider a pair of observations  $(X_i, Y_i)$  for a given animal. The predicted value,  $\hat{Y}_i$  (line in Fig. 4.7), for this predictor,  $X_i$ , is

$$\hat{Y}_i = b_0 + b_1X_i$$

and the vertical distance between the predicted value and the true observation (circles in Fig. 4.7) is called the *residual*,  $\varepsilon_i$

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Note that this residual is well defined for the case of random predictors as well as for deterministic predictors. Least squares is one of the most widespread regression techniques, although more sophisticated goal functions exist addressing different properties of the data. The goal in least squares is to minimize the square of the vertical distance between the observations and the predicted model.

Pearson correlation coefficient is associated to a linear regression model as the one in Eq. 4.29. In this way, it would be incorrect to calculate the correlation between the animal weight and the hormone dose (whose model is Eq. 4.30). Regression analysis can handle both cases under a number of assumptions:

1. The observed samples are representative of the whole population. For the case of weight and length (random predictor), the meaning of this assumption is clear.

For the case of weight and dose (deterministic predictor), it means that the doses administered during the experiment are representative of the doses of interest (*e.g.* we can artificially manipulate the slope  $b_1$  by adding too large doses).

2. For the model with deterministic predictors, it is assumed that error is zero mean.
3. The predictor is measured without noise. If our experiment violates this assumption we should use *errors-in-variables* or *total least squares* models, instead of least squares. The interested reader is referred to [Fox \(2015\)](#) for a more detailed explanation of these models.
4. If there are several predictors (*e.g.*,  $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3\dots$ ), they are linearly independent (that is, none of them can be expressed as a linear combination of the rest; technically, we cannot find  $a_i$  coefficients such that  $a_1X_1 + a_2X_2 + a_3X_3 + \dots = 0$ ). Linear dependence indicates redundancy in the information brought by the predictors. We do not need one (or some) of them, since the information brought by the redundant variables is already contained in the rest. Partial least-squares is a technique that removes the linear independence of the predictors.
5. Regression residuals are uncorrelated to the predictors. If this is not the case, it probably indicates that our regression model does not truly represent the underlying relationship between the predictor and the predicted variable (for instance, we have assumed a linear relationship  $Y = b_0 + b_1X$  when the true relationship is of the form  $Y = b_1\sqrt{X}$ ).
6. The variance of the residuals is constant across the predictor (see [Fig. 4.7](#)). This assumption is technically known as homoscedasticity, and when it is not fulfilled the data is said to be heterocedastic (see [Fig. 4.8](#)). Weighted Least Squares is a technique that has been specifically designed for heterocedastic data.

As we have presented it now, this kind of regression assumes that the predicted variable,  $Y$ , is continuous. This is, probably, the most common kind of regression. In particular, we have only presented linear regression. Non-linear regression also assumes continuous  $Y$ . However, there exist other regression variants: like the logistic regression ( $Y$  is a probability, see [Sec. 4.3.3](#)), the Cox regression ( $Y$  is a survival rate, see [Sec. 4.3.4](#)), and Poisson regression ( $Y$  is a count rate, see [Sec. 4.3.5](#)). These regressions are much more specialized and have more restricted applications. However, there are problems for which these are the right tools, and any other kind of analysis is much less powerful or, simply, incorrect.

Interestingly, we can extend the regression tools to handle categorical predictors, or even a mixture of continuous and categorical predictors. This is called generalized regression (Generalized Linear Models are a particular case of the generalized regression). The sample size design for these more advanced problems is out of the scope of this chapter. The interested reader is referred to [Dobson and Barnett \(2008\)](#).

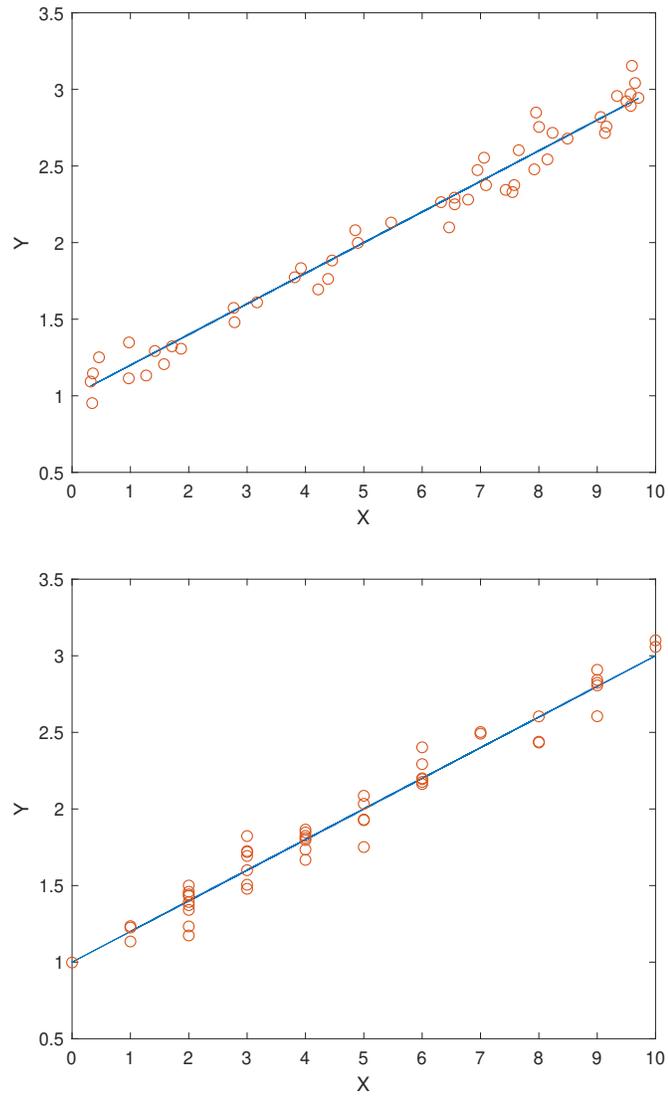


Figure 4.7: Top: Regression example in which both  $Y$  and  $X$  are random variables. Bottom: Regression example in which  $Y$  is random, but  $X$  is deterministic

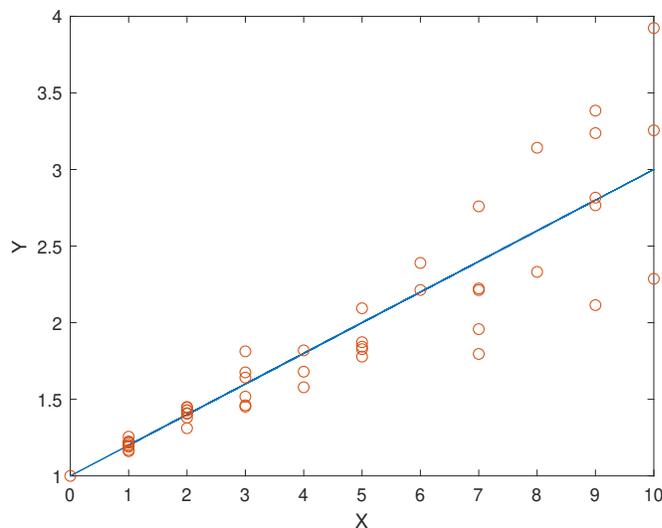


Figure 4.8: Example of heteroscedasticity, the variance of the residuals is different for different values of the predictor  $X$ .

#### 4.3.1 Linear regression: Confidence interval on the slope of a regression coefficient

- **Example 61:** We are interested in the effect of a given drug on the Forced Expiratory Volume (FEV) of a lung disease mouse model. FEV is defined as the expired air volume during 0.1 seconds, and it is expected to be around 1 mL. We expect that a new drug we are developing helps to increase the lung capacity of the animals with a maximum change of 0.5 mL if we give a maximum dose of 20 mg/kg. We want to determine a 95% confidence interval on the slope, whose maximum half width is of size 0.005. We expect the standard deviation of the regression residuals to be around 0.066 mL. How many animals do we need for this experiment?

As the  $Y$  observations are realizations of a random variable, and consequently this makes the estimated coefficients of the regression ( $\hat{b}_0, \hat{b}_1, \dots$ ) to be also random variables having their own statistical distribution. Assuming that the residuals are normally distributed with zero mean, then the regression coefficients are also normally distributed. The confidence interval of level  $1 - \alpha$  for a coefficient associated to a predictor is given by

$$\left[ \hat{b}_1 - t_{1-\frac{\alpha}{2}, N-p-1} s_{\hat{b}_1}, \hat{b}_1 + t_{1-\frac{\alpha}{2}, N-p-1} s_{\hat{b}_1} \right] \quad (4.31)$$

where  $p$  is the number of predictors of the regression (for  $b_0 + b_1X$  the number of predictors is  $p = 1$ ). That is, it is a symmetric confidence interval centered on our estimate and whose width is given by a percentile of the central Student's  $t$  distribution

(as expected for means calculated from Gaussian data) with  $N - p - 1$  degrees of freedom, and the sample standard deviation of the corresponding coefficient. This sample standard deviation is given by

$$s_{\hat{b}_1} = \frac{s_\varepsilon / \sqrt{N - 1}}{s_X}$$

where

$$s_\varepsilon = \sqrt{\frac{1}{N - p - 1} \sum_i \varepsilon_i^2}$$

$$s_X = \sqrt{\frac{1}{N - 1} \sum_i (X_i - \bar{X})^2}$$

being  $\bar{X}$  the mean of the  $X_i$  observations. If  $X$  is a Gaussian random variable, then a good estimate of  $s_X$  is

$$s_X = \sigma_X$$

that is, the standard deviation of the underlying Gaussian. If  $X$  is uniformly distributed between  $X_m$  and  $X_M$ , then an estimate of  $s_X$  is

$$s_X = \frac{X_M - X_m}{\sqrt{12}}$$

Finally, if  $X$  is deterministic and linearly distributed between  $X_m$  and  $X_M$  with  $N_X$  steps, then the exact value of  $s_X$  is

$$s_X = \frac{X_M - X_m}{\sqrt{12}} \frac{\sqrt{N_X(N_X + 1)}}{N_X - 1}$$

From these equations we draw some interesting conclusions:

#### Important remarks

75. As expected, as the number of samples grows, the uncertainty around  $\hat{b}_1$  decreases, because the term  $s_\varepsilon / \sqrt{N - 1}$  decreases.
76. Before doing the experiment we need a lot of previous knowledge about the result: 1) we need a guess of the sample standard deviation of the residuals,  $s_\varepsilon$ ; and 2) we need an estimate of the sample standard deviation of the predictor,  $s_X$ . If the predictor is deterministic, we can have an exact value of  $s_X$  because, as researchers, we are fixing these values; but if the predictor is random, then we need some guess of this statistical parameter before doing the experiment.
77. In any case, the sample size is calculated in a situation requiring many assumptions about the results of an experiment that has not been performed yet, and the specific number has to be taken as an order of magnitude rather than a precise calculation.

The sample size,  $N$ , is found by assuming that the semi-length of the confidence interval is smaller than the specified half-width,  $\Delta$

$$\boxed{t_{1-\frac{\alpha}{2}, N-p-1} \frac{s_{\varepsilon}/\sqrt{N-1}}{s_X} < \Delta} \quad (4.32)$$

- **Example 61 (continued):** For the example above, we guess that  $s_{\varepsilon} = 0.066$  (we may have found an estimate of this value in previous experiments from our laboratory, or from figures in papers of similar experiments). We will test doses from 2 to 20 mg/kg in steps of 2 (2, 4, 6, ..., 20; 10 doses in total). The sample standard deviation of these doses is

$$s_X = \frac{20-2}{\sqrt{12}} \frac{\sqrt{10 \cdot 11}}{9} = 6.06$$

With these estimates, we calculate the sample size to be

$$t_{0.975, N-2} \frac{0.066/\sqrt{N-1}}{6.06} < 0.005 \Rightarrow N = 22$$

We have 10 different doses, so we may perform 2 measurements at each dose level, and 2 more extra samples at any of the doses (preferably the first and last one, because they increase the precision of the estimate of the  $b_1$  coefficient).

A rough estimate of the  $b_1$  coefficient before doing the experiment can be calculated because we expect an increase of the FEV in 0.5 mL when the dose is 20 mg/kg, this yields an *a priori* estimate of

$$\hat{b}_1 = 0.5/20 = 0.025$$

This means that the half-width of the confidence interval will be about one fifth of the estimated slope, and this is achieved with about 22 samples.

### 4.3.2 Linear regression: Hypothesis test on the regression coefficients

- **Example 62:** Following with the Example 61, we plan an hypothesis test to check if the regression coefficient, which is expected to be relatively small (around  $\hat{b}_1 = 0.025$ ), could actually be 0. If this is the case, then our drug would not be having any effect on the FEV.

$$\begin{aligned} H_0 : & b_1 = 0 \\ H_a : & b_1 \neq 0 \end{aligned}$$

If the coefficient is larger than 0.005, we want to have a statistical power of 90%. How many animals do we need to test this hypothesis?

For a single predictor, this design is based on a Snedecor's F distribution. Under the null hypothesis, the regression should not explain more variability than the mean of the

samples. Then, we must find a size such that the Type I error probability is  $\alpha$  and the Type II error probability is  $\beta$ .

$$\boxed{F_{1-\alpha,1,N-2} = F_{\beta,\phi,1,N-2}} \quad (4.33)$$

where  $F_{1-\alpha,1,N-2}$  is the  $1 - \alpha$  percentile of a central Snedecor's F with 1 and  $N - 2$  degrees of freedom, and  $F_{\beta,\phi,1,N-2}$  is a non-central Snedecor's F with 1 and  $N - 2$  degrees of freedom and non-centrality parameter

$$\phi = N \left( \frac{s_X b_1^a}{s_\varepsilon} \right)^2$$

being  $b_1^a$  the coefficient of the alternative hypothesis for which we already want to have a specific power,  $s_X$  the standard deviation of the predictor  $X$ , and  $s_\varepsilon$  the standard deviation of the residuals (obviously, at the time of design an educated guess of these two parameters must be used).

The same idea can be extended to the linear multiple regression, only that the specific formulas are more complicated. We are primarily concerned here, not so much with the formulas, but with the consequences that derive from them.

#### Important remarks

78. The sample size for testing the significance of a regression coefficient decreases if the standard deviation of the residuals,  $s_\varepsilon$ , decreases. This result is logical: if the observed data is better fitted by the regression line, then we have less uncertainty about the slope of this line.
79. The sample size also decreases if the standard deviation of the predictor,  $s_X$ , increases. This is also logical: if we study the relationship between  $Y$  and  $X$  for a wider range of  $X$ , it will be easier to detect the line that relates the two variables.
80. The sample size also increases if the number of predictors,  $p$ , grows, because we need to estimate more parameters and this decreases the degrees of freedom available from a fixed sample size.

- Example 62 (continued): For the example above, the non-centrality parameter is

$$\phi = N \left( \frac{6.06 \cdot 0.005}{0.066} \right)^2 = 0.21N$$

We simply need to find the  $N$  that satisfies

$$F_{0.95,1,N-2} = F_{0.1,0.21N,1,N-2} \Rightarrow N = 42$$

As expected, the sample size for this experiment is larger than for the previous one, because we are putting a constraint on the statistical power required at a distance  $b_1^a = 0.005$ .

### 4.3.3 Logistic regression: Hypothesis test on the regression coefficients

Logistic regression addresses the problem of predicting the probability of an event. For instance, let  $Y$  be the event of having a cardiovascular disease in mice ( $Y = 1$  if the animal has the disease, and  $Y = 0$  if it does not). We want to study the relationship between the probability of suffering the disease and the animal weight,  $X$ . We expect this probability to be low for animals with normal weight (20-30 g) and to increase as the body weight increases (see Fig. 4.9). Logistic regression is a way of representing this dependence. This technique expresses the probability of  $Y = 1$  as a function of the predictor using the so-called logistic function

$$\Pr\{Y = 1\} = \frac{e^{b_0+b_1X}}{1 + e^{b_0+b_1X}}$$

For instance, the function represented in Fig. 4.9 is

$$\Pr\{Y = 1\} = \frac{e^{-10.26+0.27X}}{1 + e^{-10.26+0.27X}}$$

Note that this function is simulated and it does not represent a real probability of cardiovascular disease, but it serves to illustrate our argument. For simplicity of notation, let us call  $p = \Pr\{Y = 1\}$ . The *logit* of  $p$  is defined as

$$\text{logit}(p) = \frac{p}{1-p}$$

that is, the logarithm of the odds of disease vs. non-disease. If  $p$  is defined as a logistic function, then the *logit* of  $p$  becomes

$$\text{logit}(p) = b_0 + b_1X \quad (4.34)$$

- Example 63: We are interested in checking if there is a relationship between the body weight of a mouse and the probability of suffering a cardiovascular disease. For doing so, we will perform a logistic regression as the one in Eq. 4.34. We will perform a test of the form

$$\begin{aligned} H_0 : & b_1 = 0 \\ H_a : & b_1 \neq 0 \end{aligned}$$

If we cannot reject the null hypothesis, then we cannot exclude the possibility that body weight does not have any effect on the probability of suffering a cardiovascular disease. If the odds ratio increases above 3 per standard deviation, we want to have a statistical power of 90%. How many animals do we need to test this hypothesis?

Under the null hypothesis, the *logit* of  $p_0$  is

$$\text{logit}(p_0) = b_0$$

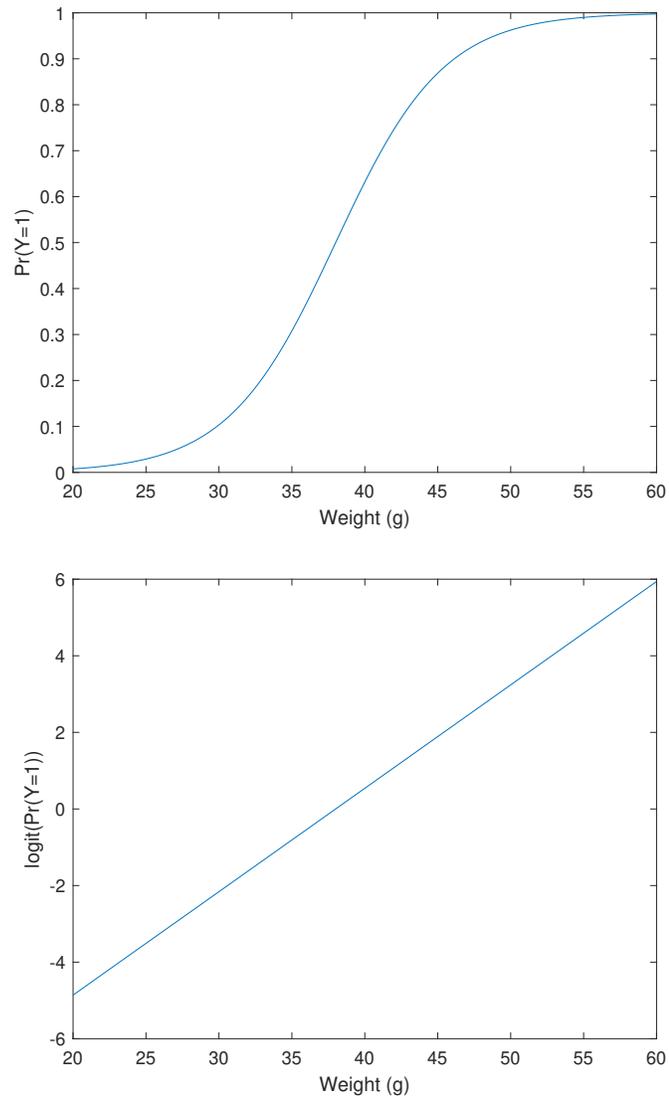


Figure 4.9: Top: Probability of having a cardiovascular disease ( $Y = 1$ ) as a function of the mouse body weight in grams. Bottom: Same probability represented as *logit*.

because under the null hypothesis there is no relation between body weight and probability of cardiovascular disease. Under the alternative hypothesis, the *logit* of  $p_a$  is

$$\text{logit}(p_a) = b_0 + b_1X$$

We may compute the difference between the two

$$\text{logit}(p_a) - \text{logit}(p_0) = b_1X$$

This is also the expression of the logarithm of the odds ratio

$$\text{logit}(p_a) - \text{logit}(p_0) = \log\left(\frac{p_a}{1-p_a}\right) - \log\left(\frac{p_0}{1-p_0}\right) = \log\left(\frac{p_a/(1-p_a)}{p_0/(1-p_0)}\right) = \log(OR)$$

Consequently,

$$OR = e^{b_1X}$$

So testing if  $b_1 = 0$  is the same as testing if the odds ratio is equal to 1. In order to compute a sample size, we need to further assume that  $X$  is normalized to have 0 mean and standard deviation 1 (note that we can always normalize our variables, and in practice, this normalization is not a constraint). Then,  $e^{b_1}$  is the increase in odds ratio of the disease for every increase of  $X$  in one standard deviation. Under the null hypothesis and if  $X$  is normalized, it can be shown that the estimate of  $b_1$ ,  $\hat{b}_1$  is approximately distributed as a Gaussian with zero mean and standard deviation  $\frac{1}{\sqrt{p_{\mu_X}(1-p_{\mu_X})}}$ . Consequently, we can design the number of samples as

$$N > \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{b_1^a}{\frac{1}{\sqrt{p_{\mu_X}(1-p_{\mu_X})}}}} \right)^2 \quad (4.35)$$

where  $b_1^a$  is the effect size we want to detect with a given statistical power (see the example below), and  $p_{\mu_X}$  is the expected probability of disease at the mean of the predictor,  $X$ .

- Example 63 (continued): We want to have a power of 90% if the odds ratio goes above 3 when the weight is one standard deviation above the mean. This means

$$e^{b_1^a} = 3 \Rightarrow b_1^a = 1.1$$

Our animals will have a mean weight of 25 g. with a standard deviation of 2 g. At 25 g., we expect that only 3% of the animals suffer from a cardiovascular disease. The sample size for this experiment would be

$$N > \left( \frac{z_{0.975} + z_{0.9}}{\frac{1.1}{\frac{1}{\sqrt{0.03 \cdot 0.97}}}} \right)^2 \Rightarrow N = 299$$

That is with  $N = 299$  we will be able to check if there is an increase in odds ratio as small as 3 when the weight is increased by one standard deviation. Remember that the odds ratio is defined as

$$OR = \frac{p_a/(1-p_a)}{p_0/(1-p_0)}$$

Then

$$p_a = \frac{p_0 OR}{p_0(OR-1)+1} = \frac{0.03 \cdot 3}{0.03(3-1)+1} = 0.085$$

That is, the probability of suffering the cardiovascular disease has to grow from 3% at 25 g. to above 8.5% at 27 g. (one standard deviation of weight) if we want to be able to detect it with a statistical power of 90% and a statistical confidence of 95%. For this detection we need  $N = 299$  mice whose weight is normally distributed with mean 25 g. and standard deviation 2 g.

#### Important remarks

81. This example highlights an important ethical concern of animal research: before doing an experiment we should evaluate the cost, harmful procedures and scientific knowledge gained from the experiment, and decide whether the experiment is worthy to be carried out. Many results of animal disease models are supposed to be extrapolated to humans, although they do not always correlate so well with human results. In this example, we need many animals to gain relatively little knowledge, which we are unsure of being able to extrapolate to humans. The same experiment with humans would have been much cheaper (we still need to observe  $N = 299$  humans, but the maintenance cost of the experiment is zero as opposed to the maintenance of the research animals) and the results are directly obtained for humans (they do not have to be extrapolated). In this case, there are no harmful procedures on the animals, although there are many experiments in which there are.

The logistic regression can be performed with several predictors and the sample size formula would be identical simply taking into account the inflation of Type I errors due to the multiple testing on the different coefficients (see Sec. 1.5). We can also perform a logistic regression with non Gaussian predictors or deterministic predictors (for instance, a drug dose). If  $b_1$  is positive, then the probability of the event will increase with larger values of the predictor. If  $b_1$  is negative, then the probability of the event will decrease with larger values of the predictor. Unfortunately, the sample size formula only exists for Gaussian predictors. For the rest of cases, we may resort to a simulation. We would fix the curve (or slope) of the alternative hypothesis, then we would simulate the appearance or not of the event in  $N$  animals according to this probability distribution. We would then perform the hypothesis test, as if we were analyzing real data, and compute the statistical power and confidence of our simulations by repeating this simulated experiment many times. The procedure is computationally costly, but it would allow us computing the sample size for any possible predictor.

**Important remarks**

82. Simulation of the experiment is an alternative to the calculation of the sample size using formulas. Simulation is valid for any kind of predictor (deterministic or random) and any kind of statistical distribution, while the sample size formulas we are giving in this chapter are only valid under the conditions for which they were derived. However, simulation requires some programming skills, unless some software performing these simulations is available.

**4.3.4 Cox regression: Hypothesis test on the regression coefficients**

Survival analysis is a statistical technique that tries to explain the expected duration of time until an event of interest happens. The technique takes its name from the expected duration of an individual until death. However, the event does not need to be death, but any other event of interest can be defined: time to cure, time to first visit of a room in a maze, time to first relapse, ... For convenience, we will keep the standard nomenclature related to death and survival. Cox regression (after its creator, the statistician David Cox) is a technique that tries to explain the survival time as a function of some predictors like the dose of a drug, the concentration of some hormone in blood, or the kind of received treatment. It is also called proportional hazards regression.

Before getting into the details of sample size for Cox regression, let us briefly introduce the main concepts of survival analysis. Let us define the random variable  $T$  as the time of death. The function  $S(t)$  is defined as the probability of dying after a time  $t$

$$S(t) = \Pr\{T > t\}$$

Once the survival function is defined, we may define the hazard function,  $\lambda(t)$ . Suppose that an individual has survived until time  $t$ , the hazard function indicates the probability that it will not survive after an infinitesimally small time,  $dt$ . It can be calculated as

$$\lambda(t) = -\frac{S'(t)}{S(t)}$$

where  $S'(t)$  is the derivative  $S(t)$  with respect to  $t$ . The hazard function can be understood as a kind of instantaneous probability of death. This probability may vary over time or be fixed. If it varies over time, it may be higher early in time and then decrease (this is, for example, the case of diseases affecting youngsters more than adults); or it may be lower early in time and then increase (this is the case of diseases affecting more the elderly). Constant hazards are related to external causes (like accidents, pathogens, ...) whose probability is not affected by the survival time. See Fig. 4.10 for an example of the three kinds of hazards, and their corresponding survival functions. Note that the hazard of external causes is constant, while the hazard of infant diseases is larger for younger animals, and the hazard of elder diseases is larger for elder animals. Note that, as expected, if a population is affected by infant mortality, the survival function is lower in the early days. Once the reason of infant mortality decreases because animals are older, then this survival function passes to occupy the upper position. The

opposite happens when the cause of mortality affect more the elderly. Interestingly, a constant hazard results in an exponentially decaying survival function. Note also that the cause of mortality may not affect the mean survival time, which in the three examples is  $\mu = 500$  days. Note also that at  $t = 500$ , the survival probability is not 50%, that is the mean survival time is not the time at which 50% of the animals still survive. Interestingly, for a constant hazard the mean survival time is

$$\mu = \frac{1}{\lambda}$$

Presume that we will perform an experiment in which we will observe the animals for 1,000 days and annotate their time of death. After 1,000 days some of the animals may still be alive (in the examples of Fig. 4.10 between 5 and 25% of the animals are still alive depending on the nature of the cause of death). But we planned the experiment for 1,000 days, and we stop it. The remaining animals for which we did not observe their death time, are said to be *censored* from the experiment. The same would happen if another researcher accidentally takes one of our animals on the 300<sup>th</sup> day, so that it disappears from our experiment. This animal, for which we did not observe its death time, is said to be censored (although for a different reason than the end of the experiment).

Cox regression relates a number of predictors to the hazard function. In particular, the hazard function is supposed to be of the form

$$\lambda(t) = \lambda_0(t)e^{b_1X_1+b_2X_2+\dots+b_pX_p}$$

That is, the hazard is a baseline hazard,  $\lambda_0(t)$ , times an exponential that depends on a linear combination of the predictors. The predictors can be continuous (*e.g.*, drug dose) or discrete (*e.g.*, receiving treatment or not). Note that the baseline hazard may change over time, it depends on  $t$ . The assumption of this model is that the predictors affect the baseline hazard in a multiplicative way, that is why it is called proportional hazards. We can manipulate the hazard expression to

$$\log\left(\frac{\lambda(t)}{\lambda_0(t)}\right) = b_1X_1 + b_2X_2 + \dots + b_pX_p$$

which has the more familiar expression of a linear regression, only that the left hand side is the logarithm of the ratio between the true hazard and the baseline hazard.

- **Example 64:** We are interested in knowing if the time mice spend “training” in a wheel helps them to better solve a maze. For doing so, we will measure the training time of each animal, and measure the time they take to escape from a simple maze. The event of interest is the maze escape (not death), and the survival time is the time they take to solve it. Our predictor,  $X_1$ , is the training time, and the Cox regression model

$$\log\left(\frac{\lambda(t)}{\lambda_0(t)}\right) = b_1X_1$$

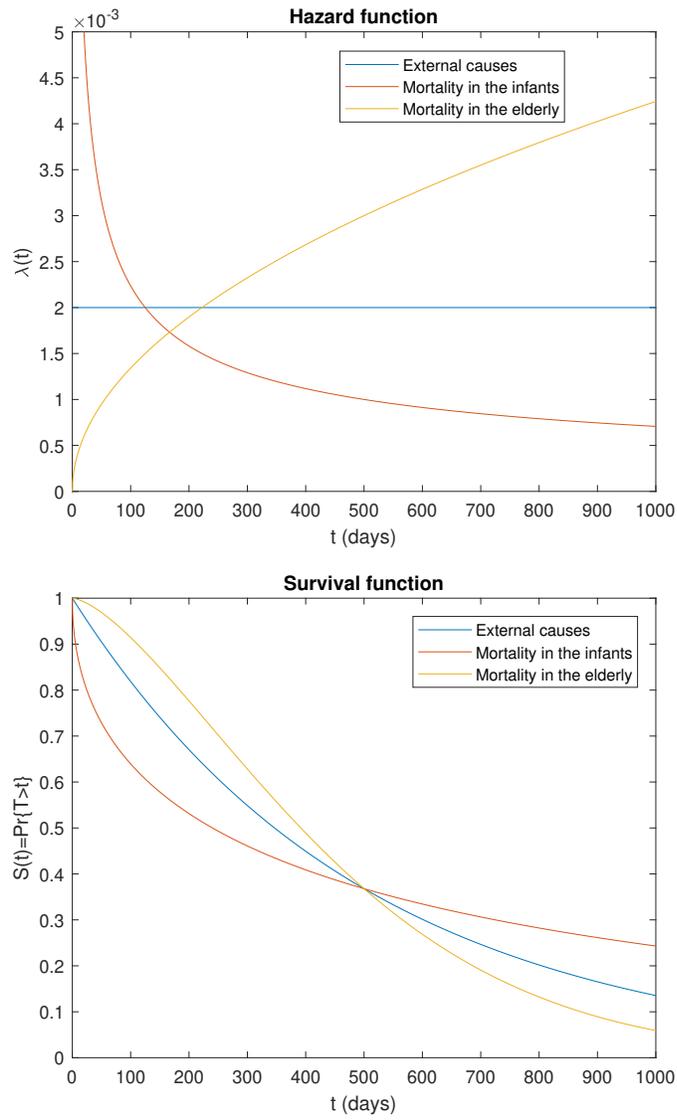


Figure 4.10: Top: hazard function for external causes (constant), for higher infant mortality, and higher mortality in the elderly. Bottom: Corresponding survival functions.

The average training time is 5 minutes with a standard deviation of 1.5 minutes. To know if there is a relationship between training and maze solving we can test the hypothesis:

$$\begin{aligned} H_0: & b_1 = 0 \\ H_a: & b_1 \neq 0 \end{aligned}$$

We want to have a statistical power of 90% if the hazard increases by a factor 1.5 per extra trained minute. The statistical confidence of the test will be 95%. We plan to stop the experiment after 150 seconds. If the animal has not found the exit of the maze within this time, the animal will be taken out. We expect that 10% of the animals will not be able to solve the maze in this time.

We can calculate the sample size with a design formula very similar to the one in Eq. 4.35

$$D > \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{b_1^a}{\frac{1}{\sigma_{X_1}}}} \right)^2$$

$b_1^a$  is the coefficient at which we already want to have a statistical power of 90% (see the example below), and  $\sigma_{X_1}$  is the standard deviation of the predictor. This sample design calculates the required number observed events,  $D$ . But as we saw in Fig. 4.10, at the end of the experiment time limit (150 seconds in the example), some of the animals may still have not found the exit. Let us call, this probability  $S(t_{limit})$ . Then, the number of animals required is increased to account for this failure probability.

$$N = \frac{D}{1 - S(t_{limit})} > \frac{1}{1 - S(t_{limit})} \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{b_1^a}{\frac{1}{\sigma_{X_1}}}} \right)^2 \quad (4.36)$$

- **Example 64 (continued):** In the example of the mice in the wheel, we expected 10% of the animals not to be able to solve the maze in less than 150 s. (these are the censored animals). So at  $t_{limit} = 150$  s. we have  $S(150) = 0.1$ . We wanted to have a statistical power of 90% if the regression coefficient is such that the hazard ratio increases by a factor 1.5 for every minute of extra training. The increase in hazard ratio is given by

$$e^{b_1^a X_1} = \left( e^{b_1^a} \right)^{X_1}$$

that is, for every minute of training, the hazard ratio is multiplied by  $e^{b_1^a}$ , meaning that

$$e^{b_1^a} = 1.5 \Rightarrow b_1^a = 0.41$$

We can now calculate the number of samples

$$N > \frac{1}{1 - 0.1} \left( \frac{z_{0.975} + z_{0.90}}{\frac{0.41}{\frac{1}{1.5}}} \right)^2 \Rightarrow N = 32$$

Note that in our example we have used for simplicity the two-sided sample design formula (we have used  $z_{1-\frac{\alpha}{2}}$  instead of  $z_{1-\alpha}$ , so that with  $N = 32$  we will be able to detect increases of the hazard ratio by a factor 1.5, but also decreases by a factor 1.5. If we are only interested in the increase side, then our test becomes one-sided and the sample size decreases to  $N = 26$ .

#### Important remarks

83. As happened with the linear regression, the sample size also decreases if the standard deviation of the predictor,  $s_{X_1}$ , increases. This is logical: if we study the survival time for a wider range of  $X_1$ , it will be easier to detect the relationship between both variables.
84. If many individuals still survive at the end of our study (in the example, many mice cannot solve the maze in 150 s.), we will need more individuals to have a number of observations large enough so that we can estimate the relationship between survival time and the predictors with the required statistical confidence and power.

### 4.3.5 Poisson regression: Hypothesis test on the regression coefficients

The result of many experiments is expressed as a count: how many young are born to a mother in a litter? how many prizes an animal can find in a maze during a fixed period of time? how many times an animal visits a given room in a maze during a fixed period of time? Poisson regression addresses the problem of verifying if these counts depend on some controllable predictors like the treatment given to the mothers having litters, the training time of mice or the presence or absence of an abuse drug in the visited room.

As we did in the previous section, let us briefly introduce the main ideas around count data. Let us call  $Y$  the events we are counting (*e.g.*, number of young in the litter, number of prizes or visits).  $Y$  will take values 0, 1, 2, ... When we perform an experiment of this kind with  $N$  animals, our result will be a count. The  $i$ -th animal will give a single count  $y_i$  (*e.g.*, a specific mother had  $y_i = 7$  young in the litter, a specific mouse found  $y_i = 3$  prizes in 60 s.). To analyze the data, let us count the number of animals with the same count which we will refer to as  $n_y$  (for instance,  $n_5 = 7$  mothers out of  $N = 40$  had  $Y = 5$  young in their litter). Then, we can estimate the frequency of observing any number of young as

$$\hat{p}_y = \frac{n_y}{N}$$

In our example, the observed frequency of having 5 young in a litter is  $7/40=17.5\%$ . This frequency must be an approximation of the true underlying probability of having 5 young in a litter (see Fig. 4.11).

$$\hat{p}_y \approx \Pr\{Y = y\}$$

The sequence  $\{\hat{p}_0, \hat{p}_1, \hat{p}_2, \dots\}$  gives an empirical estimate of the underlying distribution. As  $N$  grows, the empirical approaches more the true underlying distribution.

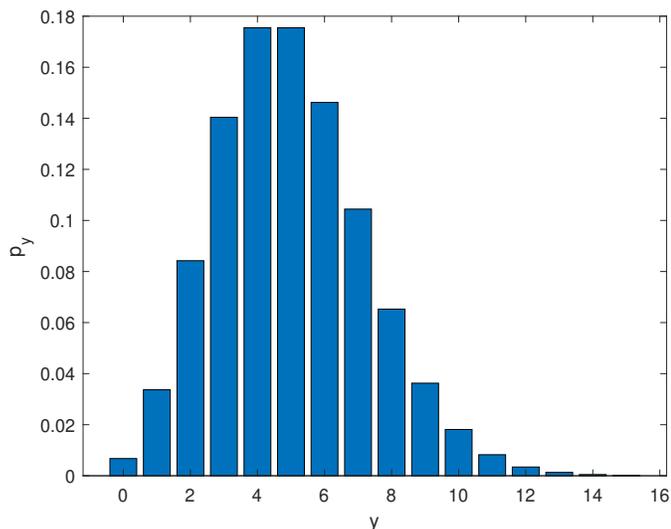


Figure 4.11: Empirical probability of observing  $y = 0, 1, 2, \dots$  young in a litter.

In order to make the analysis and experiment design tractable, we need to model the observed frequencies by some discrete distribution. The most common are:

- **Binomial distribution:** This distribution was already introduced in Sec. 4.2 when we calculated the sample size for proportion experiments. This distribution describes the probability of observing  $y$  events in  $N$  trials when the probability of each event is independent of the occurrence of other events and the probability of each event occurring is  $p$ . For instance, consider a cell culture with  $N$  bacteria observed for 1 minute. Let us call  $p$  to the probability of any of the cells acquiring a mutation during this time. These events are supposed to be independent, that is, the fact of a cell having a mutation does not cause or preclude another cell of acquiring another mutation. The probability of observing  $y$  mutated cells would be given by

$$\Pr\{Y = y\} = \frac{N!}{y!(N-y)!} p^y (1-p)^{N-y}$$

When dealing with proportions, we have seen that the binomial distribution gives raise to rather impractical expressions involving summations. This is particularly problematic for large  $N$  (e.g., there can be between  $N = 10^9 - 10^{10}$  *E. coli* cells per mL.). Additionally, sometimes we are interested in the cell culture as a whole (we cannot count the exact number of cells in the sample, like  $N = 3,895,206,312$  bacteria in my sample). It is more convenient to define a

mutation rate per billion of cells, for instance

$$\lambda = Np$$

being  $N = 10^9$ . This would give us a mutation rate per unit of amount (1 billion cells) and time (1 minute). Like this, we have already paved the way for introducing the Poisson distribution.

- Poisson distribution: This distribution expresses the probability of observing  $y$  events in fixed interval of time, space or amount, if these events occur at a constant rate and they are independent of any other event. In the example above of a cell culture and the count of cells acquiring a mutation, we would need to assume that the mutation rate is constant over time (there are not periods with larger or smaller mutation rates). Then, the number of mutations observed over a fixed period of time and a number of cells,  $N$  can be described by a Poisson distribution. If  $\lambda$  is the event rate per unit time and unit amount of cells (in the example above 1 minute and 1 billion cells), then the probability of observing  $y$  events during a period  $T$  and  $N$  billion cells is

$$\Pr\{Y = y\} = e^{-\lambda NT} \frac{(\lambda NT)^y}{y!}$$

$\lambda$  can also represent the event rate per unit area, then in a fixed area  $A$  the probability of observing  $y$  events would be the same as in the previous expression simply substituting  $NT$  by  $A$ . In general, we talk of  $T$ ,  $N$  and  $A$  as unit of exposure. However, knowing  $N$ ,  $T$  and  $A$  are not strictly necessary for the use of the Poisson distribution. For instance, we may analyze the litter size data in Fig. 4.11, and find a suitable rate,  $\lambda$ , that describes this data. In this example,  $\lambda = 5$  and it is the average rate of young per litter. The probability of observing  $y$  newborns in a litter would be

$$\Pr\{Y = y\} = e^{-\lambda} \frac{\lambda^y}{y!}$$

Note that we cannot use the binomial distribution to analyze the newborns per litter data because  $N$  is unclear. This data is not generated by  $N$  embryos, each one with a probability  $p$  of being born and, then, observing  $y$  newborns (successes) out of the  $N$  trials. Although we have introduced the Poisson as a natural continuation of the binomial, it can also be used to describe the probability of events of a different nature than the binomial distribution.

The formula above is the general expression of the Poisson distribution of parameter  $\lambda$ . We see that the expression above with  $N$  and  $T$  is equivalent to this latter one by defining  $\lambda' = \lambda NT$ , that is  $\lambda'$  is the event rate in a particular sample with  $N$  billion cells and observed for  $T$  minutes, while  $\lambda$  is the event rate per unit of time and amount. The mean and variance of a Poisson of parameter  $\lambda$  would be both  $\lambda$ .

- **Negative binomial distribution:** This distribution calculates the probability of observing  $y$  independent successes before  $r$  failures are observed. The probability of success of each trial is  $p$  (e.g., in a coin flip sequence, the probability of observing  $y$  heads before  $r = 3$  tails are observed). The probability of observing  $y$  events is

$$\Pr\{Y = y\} = \frac{(y+r-1)!}{y!(r-1)!} p^y (1-p)^r$$

The main advantage of this distribution with respect to the Poisson is that it can describe events of the same nature as the Poisson (like the number of newborns in a litter), and it has two parameters ( $p$  and  $r$ ). The mean and the variance of  $Y$  are different (as opposed to the Poisson, in which the mean and variance are equal), and we can find the  $p$  and  $r$  parameters that reproduce the empirical distribution observed in the data. In this situation, in which the mean and variance are different, it is said that the count data is *overdispersed*, and the negative binomial is used to model read counts in DNA-seq and RNA-seq experiments for this reason.

#### Important remarks

85. Apart from the specific formulas of probability of each one of the distributions, we now know that we have three tools (binomial, Poisson, and negative binomial) to model count data. Each one of them has their domain of application and they model data generated under particular assumptions.
86. Real data does not need to follow any of these models, in the same way as real measurement errors or any continuous variable of interest do not need to be Gaussian. However, making the assumption that the experimental data follows a particular, known distribution allows us to design the sample size, and to analyze the data.
87. Non-parametric discrete data analysis techniques exist, in the same way as they exist for continuous data. These techniques allow analyzing the data without making the assumption that the counts follow any known distribution. As in the case of non-parametric techniques for continuous data, non-parametric techniques are less powerful than their parametric counterparts. If the data really follows one of the known distributions, we would be losing statistical power by employing a non-parametric technique.

We can now address the problem solved by Poisson regression: does the count rate depend on some predictors? Poisson regression assumes that the count rate can be expressed as

$$\lambda = e^{b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p}$$

or what is the same

$$\log(\lambda) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

- **Example 65:** We are interested in knowing if female mice receiving a particular treatment give birth to fewer newborns per litter. To do this, we will choose a number of female mice, randomly assign them to the treatment or control groups (with probability 50%), and observing the number of newborns in each one of the groups. Then, we will fit the model

$$\log(\lambda) = b_0 + b_1 X_1$$

The baseline count rate is  $\lambda = 5$ , that is, without the treatment, mouse mothers normally have a number of youngsters as shown in Fig. 4.11. To know if there is a relationship between the treatment and the number of newborns, we can test the hypothesis:

$$\begin{aligned} H_0 : & b_1 \geq 0 \\ H_a : & b_1 < 0 \end{aligned}$$

We want to have a statistical power of 90% if the count rate decreases to one half of the baseline count rate. The statistical confidence of the test will be 95%.

In this example, the predictor  $X_1$  takes a binary value:  $X_1 = 0$  if the mouse does not receive the treatment, and  $X_1 = 1$  if it receives it. The sample size formula is

$$N > \left( \frac{z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}\{b_1|H_0\}} + z_{1-\beta} \sqrt{\text{Var}\{b_1|H_a\}}}{\frac{b_1^a}{\sqrt{T e^{b_0}}}} \right)^2 \tag{4.37}$$

where  $\text{Var}\{b_1|H_0\}$  and  $\text{Var}\{b_1|H_a\}$  is the variance of  $b_1$  under the null and alternative hypotheses, respectively;  $T$  is the total time of observation (or total area, or total amount, that is, since the  $\lambda$  is a rate per unit time, area or amount,  $T$  in this formula accounts for the fact that we may be analyzing a wider sample); and note that  $e^{b_0}$  is the baseline count rate. The variance of  $b_1$  under the two hypotheses are

	$\text{Var}\{b_1 H_0\}$	$\text{Var}\{b_1 H_a\}$
$X_1$ is Gaussian	$\frac{1}{\sigma_{X_1}^2}$	$\frac{1}{\sigma_{X_1}^2} e^{-b_1 \mu_{X_1} + \frac{b_1^2 \sigma_{X_1}^2}{2}}$
$X_1$ is Binomial	$\frac{1}{p(1-p)}$	$\frac{1}{1-p} + \frac{1}{p e^{b_1}}$

**Important remarks**

88. If the time of observation increases,  $T$  grows, the number of samples,  $N$ , decreases. The same happens if the baseline count rate increases. This effect is logical because short observation times or small count rates lead to very variable observations of the number of events.
89. If the variance of a Gaussian predictor,  $\sigma_{X_1}^2$ , increases, the number of samples decreases. We have already seen this behavior in other regressions, when we study the relationship between a predicted variable and its predic-

tors, wide range of the predictors allow a better identification of the relationship.

90. If a binary predictor is equally likely,  $p$  approaches 0.5 and the number of samples decreases. In the opposite direction, when one of the values,  $X_1 = 0$  or  $X_1 = 1$ , is very unlikely to occur, it is more difficult to determine the relationship between the predicted variable and its predictor.

- **Example 65 (continued):** Our predictor is binary, and since each animal has a probability of 50% of being in the control or treatment group, we have  $p = 0.5$ . In this example, the count rate is for the whole experiment (not per unit time or per unit amount), so that  $T = 1$ . The baseline count rate is  $\lambda_0 = 5 = e^{b_0}$ . Finally, we want to have a statistical power of 90%, if the count rate with the treatment drops by a factor 0.5. We note that for  $X_1 = 1$  we have

$$\lambda_a = e^{b_0+b_1} = \lambda_0 e^{b_1} \Rightarrow \frac{\lambda_a}{\lambda_0} = e^{b_1}$$

In our particular case,

$$\frac{\lambda_a}{\lambda_0} = 0.5 = e^{b_1} \Rightarrow b_1^a = -0.6931$$

We will use the one-sided version of the sample size design formula

$$N > \left( \frac{z_{0.95} \sqrt{\frac{1}{0.5(1-0.5)}} + z_{0.9} \sqrt{\frac{1}{1-0.5} + \frac{1}{0.5e^{-0.6931}}}}{\frac{-0.6931}{\frac{1}{\sqrt{1.5}}}} \right)^2 \Rightarrow N = 18$$

## 4.4 Sample size for Poisson counts

Poisson distribution is the natural choice for the count of discrete events when the probability of each event is independent and very low. In the following we will show how the Poisson distribution arises as the limit when the probability of the events go to zero, but the overall average of the count remains constant. In many research laboratories, radioactive substances are used as a way to visualize the location or the presence of several compounds (radiolabelling). For a given amount of radioactive material, let us assume that we observe  $\mu$  disintegrations in 1 hour. We now divide the observation time, 1 hour, into  $N$  small pieces of width  $\Delta t$ . If the pieces are small enough, then the probability of observing two or more disintegrations in the same time slot will be zero. For these small time slots, we assume that the probability of observing one disintegration,  $p$ , is very small. Additionally, we will assume that observations are independent such that the observation of one event does not depend on the time passed from the previous observation. We can calculate the probability of observing  $k$  events in the  $N$  time slots with the help of a binomial distribution

$$\Pr\{X = k\} = \binom{N}{k} p^k (1-p)^{N-k}$$

According to the binomial distribution, the expected number of events is  $Np$ , but we know that this must be  $\mu$ , so that

$$p = \frac{\mu}{N}$$

As the slot width  $\Delta t$  goes to zero, the number of slots  $N$  goes to infinity, but their product is constant and equal to 1 hour. For the same reason, the probability of observing an event at any of the time slots,  $p$ , goes to zero, but the overall mean  $Np$  remains constant and equal to  $\mu$ . The Poisson limit states that the probability of observing exactly  $k$  events in 1 hour can be calculated as

$$\Pr\{X = k\} = \lim_{N \rightarrow \infty} \binom{N}{k} \left(\frac{\mu}{N}\right)^k \left(1 - \frac{\mu}{N}\right)^{N-k} = \frac{e^{-\mu} \mu^k}{k!}$$

$\mu$  is the average number of events observed in 1 hour, that is a fixed period of time,  $T$ . It is customary to define an average number of events per unit time,  $\lambda$ , and then calculate the expected number of events in a period  $T$  as

$$\mu = \lambda T \Rightarrow \Pr\{X = k\} = \frac{e^{-\lambda T} (\lambda T)^k}{k!}$$

Although we have introduced the Poisson distribution in a time setting, it can also represent spatial events (for instance, the number of radioactive detections in a detector of surface  $A$ ).

Some observations in Biology are known to follow a Poisson distribution, for instance the number of mutations per DNA nucleotide after a given amount of radiation, the number of deaths per day in a given age population (assuming deaths are independent from each other, for example, they are not related to an infectious contagious disease), or the number of cases of adverse effects observed for a drug. There are some other experiments that are also modelled as a Poisson like the number of times an animal visits a maze, the number of macrophages in a microscopy field, the number of receptors in cell membrane, etc. The advantage of making this assumption is that it allows us to handle count data in a much better way than statistical tools not designed to handle count variables (like the standard hypothesis tests designed for Gaussian variables or regression tools). The main drawback of Poisson modelling is that it assumes independence of the events, and this assumption may be violated by our system (for instance, mice may learn in the maze example and they visit the room more frequently (or less frequently) the room as expected by a random, independent visit pattern; or several macrophages may gather called by chemiotaxis). In these situations, other statistical distributions may be used for the modelling as the negative binomial or quasi-Poisson distributions. Still, the sample sizes calculated for the Poisson distribution constitute a good starting point for more complicated distributions.

#### 4.4.1 Hypothesis test for a single population

- Example 66: We are developing a new drug for veterinarian use and we want to determine if an adverse effect is common ( $>1\%$ ) or not. How many animals do we need to observe to ascertain this question? We want to have a statistical

power of 90% if the proportion of adverse effect is larger than 2%. The statistical confidence level is set to the standard 95%.

We need an hypothesis test of the form

$$\begin{aligned} H_0 : \lambda &< \lambda_0 \\ H_a : \lambda &\geq \lambda_0 \end{aligned}$$

$\lambda$  can be understood as the probability of adverse effects per animal, and  $\lambda_0$  would be the lower limit to be considered a common effect.

As in all other discrete cases, the exact design formulas imply complicated summations. We must find  $k$  (critical threshold of the number of observed adverse effects to reject the null hypothesis) and  $N$  (the sample size) such that

$$\begin{cases} \Pr(0 \leq X \leq k; N\lambda_0) \geq 1 - \alpha \\ \Pr(0 \leq X \leq k; N\lambda_a) \leq \beta \end{cases} \quad (4.38)$$

where  $X$  is the number of observed adverse effects (which follows a Poisson distribution),  $\lambda_0$  is the rate at the null hypothesis, and  $\lambda_a$  is the alternative rate at which we already want to have a specified power. Dealing with these summations is difficult, although it can be done with a computer. If the product  $N\lambda$  is large, then we can use the square-root method that exploits the fact that the square root of  $X$  is normally distributed

$$\sqrt{X} \sim N(\sqrt{N\lambda}, 0.25)$$

From here, we may deduce

$$N \geq \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\sqrt{\lambda_a} - \sqrt{\lambda_0}}{\sqrt{0.25}}} \right)^2 \quad (4.39)$$

- Example 66 (continued): In the example above we must find  $k$  and  $N$  such that

$$\begin{aligned} \Pr(0 \leq X \leq k; 0.01N) &\geq 0.95 \\ \Pr(0 \leq X \leq k; 0.02N) &\leq 0.1 \end{aligned}$$

The exact solution is  $N = 1,296$  and  $k = 19$ , that is, we must give the drug to  $N = 1,296$  animals. If more than  $k = 19$  have adverse effects, then we must reject the null hypothesis and declare that the adverse effect is common (its incidence is larger than 1%). The approximate method gives

$$N \geq \left( \frac{z_{0.95} + z_{0.9}}{\frac{\sqrt{0.02} - \sqrt{0.01}}{\sqrt{0.25}}} \right)^2 = 1,248$$

### 4.4.2 Hypothesis test for two populations

- **Example 67:** We want to determine the best way of maintaining an animal house with a low incidence of a given infection. For doing so, we will observe two animal houses: one follows the procedure 1, while the other follows the procedure 2. The base proportion of infected animals should be around 1%, and we would like to have a statistical power of 90% if the infection rate deviates from each other more than 0.5%.

We need an hypothesis test of the form

$$\begin{aligned} H_0 : \lambda_1 &= \lambda_2 \\ H_a : \lambda_1 &\neq \lambda_2 \end{aligned}$$

The exact solution is given by a binomial method whose details fall outside the scope of this book, but it follows similar expressions than the exact method of the previous section. If we can use the square-root approximation ( $N\lambda_1$  and  $N\lambda_2$  are large), then

$$\frac{\sqrt{\lambda_2} - \sqrt{\lambda_1}}{\frac{1}{2}\sqrt{\frac{2}{N}}} \sim N(0, 1)$$

The optimal number of samples per group is given by

$$N = \left( \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\sqrt{\lambda_2} - \sqrt{\lambda_1}}{\frac{1}{2}\sqrt{2}}} \right)^2 \quad (4.40)$$

For an superiority/inferiority test

$$\begin{aligned} H_0 : \lambda_1 &\geq \lambda_2 \\ H_a : \lambda_1 &< \lambda_2 \end{aligned}$$

We exploit that

$$\frac{\sqrt{\lambda_2} - \sqrt{\lambda_1}}{\frac{1}{2}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim N(0, 1)$$

The optimal number of samples is given by

$$N_1 = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\sqrt{\lambda_2} - \sqrt{\lambda_1}}{\frac{1}{2}\sqrt{1+R}}} \right)^2 \quad N_2 = \frac{N_1}{R} \quad \text{with} \quad R = \frac{\lambda_2}{\lambda_1} \quad (4.41)$$

As the expected rate in Group 2 is larger than in Group 1, we need fewer samples in Group 2 than in Group 1 ( $N_2 < N_1$ ).

- **Example 67 (continued):** We will perform two designs one for positive deviations of Group 2 with respect to Group 1, and another one for negative deviations:

$$N = \left( \frac{\frac{z_{0.975} + z_{0.9}}{\sqrt{0.015} - \sqrt{0.01}}}{\frac{1}{2}\sqrt{2}} \right)^2 = 10,420$$

$$N = \left( \frac{\frac{z_{0.975} + z_{0.9}}{\sqrt{0.005} - \sqrt{0.01}}}{\frac{1}{2}\sqrt{2}} \right)^2 = 6,125$$

We see that the most restrictive case is when  $\lambda_2$  is larger than  $\lambda_1$ . Consequently, we need to observe  $N = 10,420$  in each animal house.

#### Important remarks

91. Poisson distribution is also called the distribution of rare events. Consequently, working with it implies very large sample size, simply because the number of observations is very low (the expected value is  $N\lambda$ , which also happens to be its variance).

## 4.5 Sample size for the variance

The following set of procedures aim at designing the sample size for situations in which very little is known about the experiment. Note that in many other sample size designs, the variance of the observations is a key parameter (this is the case of all sample size designs for the mean and for regression). However, there are experimental situations in which even this variance is unknown. The following sample size calculations will allow us to design an experiment by which we will gain some insight into the variability we should expect from our observations.

### 4.5.1 Confidence interval for the standard deviation

For instance, let us assume this is the first time, ever in history, that the expression level of a given gene is studied. How many individuals should we study to determine the standard deviation with a confidence interval whose two-sided width is smaller than a given desired precision. When we perform the experiment, we will be able to calculate the sample standard deviation,  $\hat{\sigma}$  as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Then, we will construct a two-sided  $1 - \alpha$  confidence interval (e.g., 95% confidence interval) as

$$\left( \hat{\sigma} \sqrt{\frac{N-1}{\chi_{1-\frac{\alpha}{2}, N-1}^2}}, \hat{\sigma} \sqrt{\frac{N-1}{\chi_{\frac{\alpha}{2}, N-1}^2}} \right)$$

The confidence interval for the ratio  $\frac{\sigma}{\hat{\sigma}}$  is

$$\left( \sqrt{\frac{N-1}{\chi_{1-\frac{\alpha}{2}, N-1}^2}}, \sqrt{\frac{N-1}{\chi_{\frac{\alpha}{2}, N-1}^2}} \right)$$

whose width is by design to be smaller than  $\delta$ , so the sample size design equation must be

$$\sqrt{\frac{N-1}{\chi_{\frac{\alpha}{2}, N-1}^2}} - \sqrt{\frac{N-1}{\chi_{1-\frac{\alpha}{2}, N-1}^2}} \leq \delta \quad (4.42)$$

that must be solved numerically.

- Example 68: We want to determine a 95% confidence interval for the standard deviation of the gene expression level of a given gene with a two-sided precision less than  $\delta = 1$ . Then, we need  $N = 12$  samples. With this number of samples the 95% confidence interval for the  $\frac{\sigma}{\hat{\sigma}}$  ratio is

$$(0.71, 1.70)$$

That is, the true standard deviation could be as small as  $0.71\hat{\sigma}$  or as large as  $1.70\hat{\sigma}$ . Having more precision in our confidence interval rapidly increases the sample size. For instance, to have only a 10% of two-sided width, the sample size would grow up to  $N = 774$  individuals. Then, the confidence interval would be

$$(0.95, 1.05)$$

#### Important remarks

92. A large precision for the variance or standard deviation rapidly increases the number of samples. For small sample sizes, we need to accept a relatively large uncertainty about the true underlying variability of our population.

### 4.5.2 Hypothesis test for one variance

- Example 69: We regularly monitor the precision of the optical densitometer of our laboratory. Historically, the standard deviation of the measurements has been  $\sigma = 0.05$  (arbitrary units). How many samples do we need to detect an increase of variance larger than 50% of the nominal variance with a statistical power of 90% and a confidence of 95%?

In this setting we will perform an hypothesis test (see Fig. 4.12):

$$\begin{aligned} H_0 &: \sigma^2 \leq 0.05^2 \\ H_a &: \sigma^2 > 0.05^2 \end{aligned}$$

The sample size can be calculated by analyzing the equation of the critical value beyond which we would reject the null hypothesis

$$\frac{1}{N-1} \sigma_0^2 \chi_{1-\alpha, N-1}^2 \leq \frac{1}{N-1} \sigma_a^2 \chi_{\beta, N-1}^2$$

That is, the sample size design formula is given by the smallest  $N$  satisfying

$$\boxed{\frac{\chi_{1-\alpha, N-1}^2}{\chi_{\beta, N-1}^2} \leq \frac{\sigma_a^2}{\sigma_0^2}} \quad (4.43)$$

An approximate solution when  $N$  is large is given by the Gaussian approximation

$$\boxed{N > \frac{1}{2} \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\log \frac{\sigma_a}{\sigma_0}} \right)^2} \quad (4.44)$$

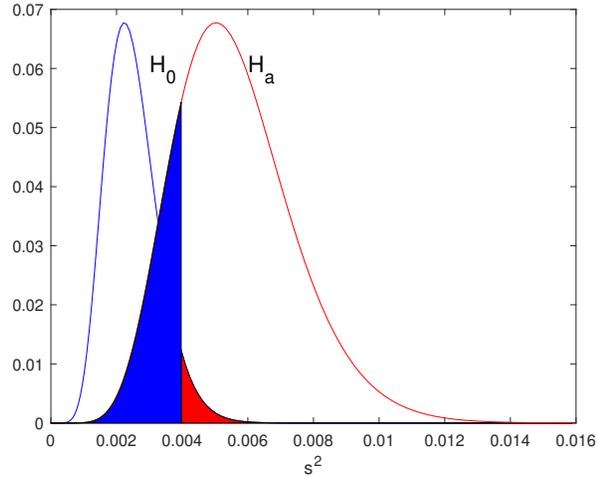


Figure 4.12: Example of hypothesis test for a one sample variance. The two distributions show the expected values of the sample variance,  $s^2 = \hat{\sigma}^2$ , if the null ( $H_0$ ) or the alternative ( $H_a$ ) hypotheses are true. The shaded areas represent the probability of Type I (red) and Type II (blue) errors.

- Example 69 (continued): With this data we have  $\sigma_a^2/\sigma_0^2 = 1.5$ , then we must find  $N$  such that

$$\frac{\chi_{0.95, N-1}^2}{\chi_{0.1, N-1}^2} \leq 1.5$$

whose solution is  $N = 105$  samples. The approximated design also gives the same result

$$N > \frac{1}{2} \left( \frac{z_{0.95} + z_{0.9}}{\log \sqrt{1.5}} \right)^2 = 105$$

The design formula for a test checking the decrease of the variance is given by

$$\boxed{\frac{\chi_{\alpha, N-1}^2}{\chi_{1-\beta, N-1}^2} \geq \frac{\sigma_a^2}{\sigma_0^2}} \quad (4.45)$$

For a test checking the change of the variance (two-sided, that is, either increase or decrease), we would have the maximum  $N$  from

$$\boxed{\frac{\chi_{1-\frac{\alpha}{2}, N-1}^2}{\chi_{\beta, N-1}^2} \leq \frac{\sigma_a^2}{\sigma_0^2} \quad \text{and} \quad \frac{\chi_{\frac{\alpha}{2}, N-1}^2}{\chi_{1-\beta, N-1}^2} \geq \frac{\sigma_a^2}{\sigma_0^2}} \quad (4.46)$$

### 4.5.3 Hypothesis test for two variances

- Example 70: We are buying a new optical densitometer that claims to be more precise than our old model. How many samples do we need to take from each densitometer to test if this claim is true? We want to have a statistical power of 90% if the new variance is 50% smaller than the old one.

Now the hypothesis test is given by comparing the variance of both samples. In the following test we refer to the variance of the old equipment as  $\sigma_1^2$  and to the variance of the new equipment as  $\sigma_2^2$

$$\begin{aligned} H_0 : & \sigma_1^2 \leq \sigma_2^2 \\ H_a : & \sigma_1^2 > \sigma_2^2 \end{aligned}$$

Under the null hypothesis the statistic

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

is distributed as a Snedecor's F with  $N_1 - 1$  and  $N_2 - 1$  degrees of freedom. If the two groups have the same size,  $N_1 = N_2 = N$ , then the sample size is the smallest  $N$  satisfying

$$\boxed{\frac{F_{1-\alpha, N-1, N-1}}{F_{\beta, N-1, N-1}} \leq \frac{\sigma_1^2}{\sigma_2^2}} \quad (4.47)$$

When  $N$  is large, this can be approximated by the Gaussian design

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\log \frac{\sigma_1}{\sigma_2}} \right)^2 \quad (4.48)$$

- Example 70 (continued): In this example, we must find  $N$  such that

$$\frac{F_{0.95, N-1, N-1}}{F_{0.1, N-1, N-1}} \leq \frac{1}{0.5}$$

whose solution is  $N = 74$ . The approximate formula gives

$$N = \left( \frac{z_{0.95} + z_{0.9}}{\log \sqrt{\frac{1}{0.5}}} \right)^2 = 72$$

## 4.6 Sample size for correlations

### 4.6.1 Confidence interval for correlation

- Example 71: We are interested in detecting a weak correlation between aldosterone (an steroid hormone produced by the adrenal gland) concentration in blood plasma and blood pressure. We expect the correlation to be around 0.25. How many individuals do we need to study to determine the correlation with a precision of 0.05 and a level of confidence of 95%.

We are looking for a confidence interval of the form  $[\rho_L, \rho_U]$  where L and U refer to the lower and upper bounds respectively. As with other sample design formulas, for the correlation we need to foresee beforehand which will be approximately the result of the experiment. So that in our case, if we expect the correlation to be around 0.25, the lower and upper bounds will be  $[0.2, 0.3]$ . With this information we can use Fisher's  $Z$  transform that is distributed approximately as a Gaussian

$$Z = \tanh^{-1}(\hat{\rho}) = \frac{1}{2} \log \frac{1+\hat{\rho}}{1-\hat{\rho}} \sim N \left( \tanh^{-1}(\rho), \frac{1}{N-3} \right)$$

In this way, we transform the confidence interval problem on  $\rho$  into a confidence interval problem for  $Z$

$$\Pr\{\rho_L < \rho < \rho_U\} = 1 - \alpha = \Pr\{Z_L < Z < Z_U\}$$

We already know its solution which is

$$\begin{aligned} Z_L &= \frac{1}{2} \log \frac{1+\rho_L}{1-\rho_L} = \frac{1}{2} \log \frac{1+\rho}{1-\rho} - z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{N-3}} \\ Z_U &= \frac{1}{2} \log \frac{1+\rho_U}{1-\rho_U} = \frac{1}{2} \log \frac{1+\rho}{1-\rho} + z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{N-3}} \end{aligned}$$

If we now subtract the first equation from the second, we have the sample size design formula

$$Z_U - Z_L = 2z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{N-3}} \Rightarrow N = \left( \frac{2z_{1-\frac{\alpha}{2}}}{Z_U - Z_L} \right)^2 + 3 \quad (4.49)$$

- Example 71 (continued): In our example

$$\begin{aligned} Z_L &= \frac{1}{2} \log \frac{1+(0.25-0.05)}{1-(0.25-0.05)} = 0.2027 \\ Z_U &= \frac{1}{2} \log \frac{1+(0.25+0.05)}{1-(0.25+0.05)} = 0.3095 \\ \Delta Z &= Z_U - Z_L = 0.1068 \\ N &= \left( \frac{2z_{0.975}}{0.1068} \right)^2 + 3 = 1,351 \end{aligned}$$

#### Important remarks

93. The sample size needed for low correlations is very large precisely because the correlation is so low that it requires many samples to be sure that the detected small correlation is not by chance. For large correlations this is not the case: with relatively few animals, the large correlation quickly becomes apparent.

### 4.6.2 Hypothesis test on one sample correlation

- Example 72: We suspect that the correlation between the length and weight of an animal is smaller than 0.9. How many individuals do we need to inspect to show so if we want to have a test power of 90% if the correlation is actually 0.8?

We are making a test of the form:

$$\begin{aligned} H_0 &: \rho \geq \rho_0 \\ H_a &: \rho < \rho_0 \end{aligned}$$

This is a test with a single sample (we are not comparing the correlation between length and weight in two groups). Then, we simply have to extend the formula of the previous section to include the statistical power

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{Z_0 - Z_a} \right)^2 + 3 \quad (4.50)$$

where  $Z_0$  is the Fisher's  $Z$  transform of  $\rho_0$  and  $Z_a$  is the Fisher's  $Z$  transform of the correlation for which we already want to have a given statistical power.

- Example 72 (continued): In our example

$$\begin{aligned} Z_0 &= \frac{1}{2} \log \frac{1+0.9}{1-0.9} = 1.4722 \\ Z_a &= \frac{1}{2} \log \frac{1+0.8}{1-0.8} = 1.0986 \\ N &= \left( \frac{z_{0.95} + z_{0.9}}{1.4722 - 1.0986} \right)^2 + 3 = 65 \end{aligned}$$

### 4.6.3 Hypothesis test for the correlations in two samples

- **Example 73:** The correlation between length and weight in the general population is about 0.8 (Group 1). We wonder if this same correlation holds among diabetes type II animal models (Group 2) because these animals tend to be fatter. How many control and diseased animals do we need to study to check if the correlation is lower in diabetes type II animals? We want a power of 90% if the correlation drops below 0.7.

We are making a test of the form:

$$\begin{aligned} H_0 : \rho_1 &\leq \rho_2 \\ H_a : \rho_1 &> \rho_2 \end{aligned}$$

We now have two populations (control and diseased animals). After transforming the observed correlations we will finally compare the difference between both:

$$\Delta Z = Z_1 - Z_2$$

and the test can be reformulated as

$$\begin{aligned} H_0 : \Delta Z &\geq 0 \\ H_a : \Delta Z &< 0 \end{aligned}$$

The variance of  $\Delta Z$  is

$$\sigma_{\Delta Z}^2 = \frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}$$

. If  $N_1 = N_2$ , then the sample design formula is given by

$$\Delta Z = (z_{1-\alpha} + z_{1-\beta})\sigma_{\Delta Z}$$

that is

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\Delta Z}{\sqrt{2}}} \right)^2 + 3 \quad (4.51)$$

- **Example 73 (continued):** In the example above

$$\begin{aligned} Z_1 &= \frac{1}{2} \log \frac{1+0.8}{1-0.8} = 1.0986 \\ Z_2 &= \frac{1}{2} \log \frac{1+0.7}{1-0.7} = 0.8673 \\ \Delta Z &= 0.2313 \\ N &= \left( \frac{z_{0.95} + z_{0.9}}{0.2313} \right)^2 + 3 = 324 \end{aligned}$$

### 4.6.4 Hypothesis test for multiple correlation in one sample

- **Example 74:** We are interested in predicting the time to recover from pneumonia in mice when two drugs are administered in combination. How many samples do we need to do so if we want to have a power of 90% if there is a multiple correlation coefficient larger than 0.7?

The multiple correlation coefficient is used to determine how well a given variable,  $Y$ , can be predicted from a linear combination of other variables,  $X_1, X_2, \dots$ . For instance, in the example above, we could predict the time to recover,  $T$ , as a linear combination of the daily dose drugs A and B,  $D_A$  and  $D_B$ , respectively, and the age of the animal in months,  $Age$

$$T = \mu + \beta_A D_A + \beta_B D_B + \beta_{age} Age$$

The multiple correlation coefficient is the square root of the coefficient of determination,  $R^2$ , that is the fraction of the total sums of squares explained by the model

$$R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{\epsilon}}{SS_{total}}$$

The sum of squares corresponding to the residuals has been noted as  $SS_{\epsilon}$ . Our hypothesis test is

$$\begin{aligned} H_0 : R^2 &= 0 \\ H_a : R^2 &> 0 \end{aligned}$$

We construct the  $F$  statistic

$$F = \frac{SS_{model}/df_{model}}{SS_{\epsilon}/df_{\epsilon}}$$

Under  $H_0$  it is distributed as a Snedecor's  $F$  with  $df_{model} = p$  and  $df_{\epsilon} = N - p - 1$  degrees of freedom, where  $N$  is the number of samples, and  $p$  the number of predictor variables in the model. Under  $H_a$  it is distributed as  $F_{\phi, p, N-p-1}$  where  $\phi = N \frac{R^2}{1-R^2}$  is the non-centrality parameter. Consequently, we need to find  $N$  such that

$$\boxed{F_{1-\alpha, p, N-p-1} = F_{\beta, N \frac{R^2}{1-R^2}, p, N-p-1}} \quad (4.52)$$

- **Example 74 (continued):** In the example above, we have  $p = 3$  predictor variables and we need to find  $N$  such that

$$F_{0.95, 3, N-4} = F_{0.1, N \frac{0.72}{1-0.72}, 3, N-4}$$

That is  $N = 20$ .

#### Important remarks

94. When applicable, regression is a relatively powerful statistical tool because it may have a high explanatory power at a very low cost in terms of degrees of freedom (the number of predictors). For a few predictors, very low sample sizes are required.

### 4.6.5 Confidence interval for the Intraclass Correlation (ICC)

- **Example 75:** In animal research, a qualified professional must evaluate the pain state of animals and, in general, their welfare. The professional must rate the severity of the procedures and the state of the animals either from their behaviour

or their facial expression. Different professionals may diverge in their evaluation. Let us assume that they rate the severity of the animal state from 0 (normal state) to 10 (extremely painful). We want to determine the coherence of the evaluation of the veterinarians from four animal facilities. This is measured by the ICC. For doing so,  $N$  animals will be evaluated by the four veterinarians. How many animals must be evaluated in order to construct a 95% confidence interval whose half width is 0.1? We expect the intraclass correlation to be around 0.8.

We must first define the intraclass correlation,  $ICC$ . It is normally defined in an ANOVA setting. Let  $y_{ij}$  denote the rate of animal  $i$  given by the veterinarian  $j$ . We will assume that these rates can be modelled as

$$y_{ij} = \mu + \alpha_i + \tau_j + \varepsilon_{ij}$$

The ICC is defined as

$$ICC = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}$$

- ICC is close to 1 when there are small differences within raters ( $\sigma_\tau^2 \gg \sigma_\varepsilon^2$ ).
- ICC is close to 0 when there are large differences within raters ( $\sigma_\tau^2 \ll \sigma_\varepsilon^2$ ).

Let us call  $r$  to the number of raters, then the transformation

$$Z = \frac{1}{2} \log \frac{1 + (r-1)ICC}{1 - ICC}$$

is approximately normally distributed with zero mean and variance

$$\sigma_Z^2 = \begin{cases} \frac{1}{N-3/2} & r = 2 \\ \frac{r}{2(r-1)(N-2)} & r > 2 \end{cases}$$

Now, we can construct our sample size design formula

$$N \geq \begin{cases} \left( \frac{2z_{1-\frac{\alpha}{2}}}{\Delta Z} \right)^2 + \frac{3}{2} & r = 2 \\ \left( \frac{2z_{1-\frac{\alpha}{2}}}{\frac{\Delta Z}{\sqrt{\frac{r}{2(r-1)}}}} \right)^2 + 2 & r > 2 \end{cases} \quad (4.53)$$

- Example 75 (continued): In our example, we have  $r = 4$  raters, assuming that the ICC will be around 0.8, then the lower and upper bounds will be around 0.7 and 0.9 respectively.

$$\begin{aligned} Z_U &= \frac{1}{2} \log \frac{1+3 \cdot 0.9}{1-0.9} = 1.8055 \\ Z_L &= \frac{1}{2} \log \frac{1+3 \cdot 0.7}{1-0.7} = 1.1677 \\ \Delta Z &= Z_U - Z_L = 0.6378 \\ N &\geq \left( \frac{2z_{0.975}}{\frac{0.6378}{\sqrt{\frac{4}{2 \cdot 3}}}} \right)^2 + 2 = 28 \end{aligned}$$

The ICC plays an important role in the sample size calculation of other parameters, for instance the mean, because it reduces the variance within group artificially giving a false impression of low variability. For example, in Sec. 4.1.4 we discussed the sample size calculated for studying the effect of a new drug on the intraocular pressure of a mouse model of glaucoma. For the design, we used the fact that the standard deviation of the intraocular pressure was around 2.2 mmHg. We wanted to detect changes in the intraocular pressure of 0.5 mmHg. The sample size calculation was based on the formula

$$\Pr \left\{ t_{\lambda, N-1} < t_{1-\frac{\alpha}{2}, 0, N-1} \right\} < \beta$$

with  $\lambda = \frac{\Delta}{\hat{\sigma}} \sqrt{N}$ . The result in that case was  $N = 333$  meaning that 333 animals would be used in the experiment. The intraocular pressure of the two eyes of each one of them would be measured. One of the eyes will receive the new drug, while the other will not.

The problem of this experiment is that if the same researcher makes all the measurements, then the independence between samples may be compromised. Actually, the independence may be compromised by many other factors. For instance, if we start at 9AM and start measuring animals, the observations may be correlated due to the time of the day at which the measurement is performed (it might be that the intraocular pressure varies along the day), the level of tiredness of the researcher, his growing skill in measuring the pressure as more animals have been measured, etc. Also, if all animals come from the same animal facility, this may also introduce some correlation between measurements. All these effects result in an ICC, that is within the group of animals measured in a single place there is a small correlation which makes the variance to be apparently smaller than it should if the measurements were truly independent. The observed variance,  $\sigma_{obs}^2$ , would be

$$\sigma_{obs}^2 = \sigma^2(1 - ICC)$$

where  $\sigma^2$  is the variance that would be observed in the absence of correlation (this is the one that we have normally used along the book). In the limit, note that if all samples are perfectly correlated,  $ICC = 1$ , then the observed variance drops to 0. Correlation between samples taken in the same laboratory is a major concern in multicentric studies.

In these circumstances, in order to have the same statistical power and confidence we should increase the sample size to account for the apparent decrease of variance. For the intraocular pressure example, we should modify the centrality parameter to

$$\lambda = \frac{\Delta}{\frac{\hat{\sigma}}{\sqrt{1-ICC}}} \sqrt{N}$$

and solve the sample size again. Approximately we should have

$$N_{correlated} \approx N \frac{1}{1 - ICC}$$

where  $N_{correlated}$  refers to the sample size needed in the presence of correlated samples, and  $N$  in the absence of sample correlations.

In the example, we have given of  $N = 333$  mice to measure a decrease of 0.5 mmHg. in intraocular pressure, if the  $ICC=0.1$ , we should increase the sample size to  $N = 370$ .

The same happens with multicentric experiments (sometimes called cluster randomized control trials). The concern is that measurements within the same center are not independent and they have some correlation ( $ICC$ ).

**Important remarks**

95. If samples are not independent, and this can be measured through the intraclass correlation, the sample size of any of the experiments must be larger in order to compensate for the apparent reduction of variability.

- **Example 76:** We are developing a new diet for laboratory animals that should keep the cholesterol levels in blood around a concentration of 250 mg/dL with a standard deviation of 30 mg/dL. We want to have a statistical power of 90% for detecting deviations of 25 mg/dL. How many animals do we need to test for verifying this hypothesis?

The hypothesis is of the form

$$\begin{aligned} H_0 : \mu &= 250 \\ H_a : \mu &\neq 250 \end{aligned}$$

This kind of problems was analyzed in Sec. 4.1.2, and we would have obtained a sample size of  $N = 18$ , that is, we need to check the cholesterol level of  $N = 18$  animals to take decision of whether our new diet is performing according to its specifications.

However, our experiment is carried out in  $K = 4$  different laboratories, and we expect an intra-class correlation of  $ICC = 0.1$  between samples from the same laboratory. How many animals per laboratory should we study to have the same statistical power as if the experiment were performed with independent animals?

We know that the variance of the estimate of the mean is reduced as the number of samples grow as

$$\sigma_{\bar{\mu}}^2 = \frac{\sigma_X^2}{N}$$

If there are  $K$  centers, with  $N_K$  samples in each center, it should be  $N = KN_K$ . Additionally, if there is intra-class correlation, the variance of the mean estimate is modified to

$$\sigma_{\bar{\mu}}^2 = \frac{\sigma_X^2}{N} (1 + (N_K - 1)ICC)$$

The multiplicative factor  $1 + (N_K - 1)ICC$  is called the Variance Inflation Factor (VIF). To have the same confidence level and power as in a completely random study we would need

$$\frac{1}{N} = \frac{1}{KN_K} (1 + (N_K - 1)ICC) \Rightarrow N_K = N \frac{1 - ICC}{K - N \cdot ICC} \quad (4.54)$$

- **Example 76 (continued):** In our example  $K = 4$ , then number of animals per center with an  $ICC = 0.1$  would be

$$N_K = 18 \frac{1 - 0.1}{4 - 18 \cdot 0.1} \Rightarrow N_K = 8$$

That is, we will need  $KN_K = 4 \cdot 8 = 32$  animals, 8 per center, rather than the  $N = 18$  required for completely independent animals.

#### 4.6.6 Hypothesis test for Cohen's $\kappa$

- **Example 77:** We are interested in the consistency of criteria among veterinarians of two different centers. For testing the coherence in their evaluations, two veterinarians from two centers are asked to assess the severity of different procedures. A procedure can be assessed as mild, moderate, severe or non-recovery (4 labels in total). We want to check if the concordance as measured by Cohen's  $\kappa$  is larger than 0.8

$$\begin{aligned} H_0 : \kappa &\leq 0.8 \\ H_a : \kappa &> 0.8 \end{aligned}$$

How many procedures should they evaluate if we want to have a statistical power of 90% if  $\kappa$  goes above 0.9.

Cohen's  $\kappa$  is a possible way of measuring the association between categorical variables. Note that Pearson's correlation would be incorrectly applied in this case, since it was developed for continuous variables. Let  $K$  be the number of categories (in our example above  $K = 4$ ). Let  $N$  be the number of procedures to evaluate by both veterinarians. Let  $p_{ij}$  denote the proportion of cases in which the first veterinarian assigned a label  $i$  and the second a label  $j$ . Let  $p_o$  denote the observed proportion of agreement

$$p_o = \sum_{i=1}^K p_{ii}$$

Let  $p_e$  denote the expected proportion of agreement by chance

$$p_e = \sum_{i=1}^K p_{i \cdot} p_{\cdot i}$$

where  $p_{i \cdot} = \sum_{j=1}^K p_{ij}$  and  $p_{\cdot j} = \sum_{i=1}^K p_{ij}$ . The empirical Cohen's  $\kappa$  is defined as

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}$$

For a test of the kind

$$\begin{aligned} H_0 : \kappa &\leq \kappa_0 \\ H_A : \kappa &> \kappa_0 \end{aligned}$$

the variance of the estimate of  $\kappa$  is approximately

$$\sigma_{\hat{\kappa}}^2 \approx \frac{p_o(1-p_o)}{N(1-p_e)^2} = \frac{[\hat{\kappa}(1-p_e) + p_e][1 - (\hat{\kappa}(1-p_e) + p_e)]}{N(1-p_e)^2} = \frac{f(\hat{\kappa})}{N}$$

where we have made use of

$$p_o = \hat{\kappa}(1-p_e) + p_e$$

and we have defined

$$f(\hat{\kappa}) = \frac{[\hat{\kappa}(1-p_e) + p_e][1 - (\hat{\kappa}(1-p_e) + p_e)]}{(1-p_e)^2}$$

The number of samples can be solved from the standard equation for Gaussian variables with known variance

$$\kappa_0 + z_{1-\alpha}\sigma_{\hat{\kappa}_0} = \kappa_1 - z_{1-\beta}\sigma_{\hat{\kappa}_a}$$

From where

$$N \geq \left( \frac{z_{1-\alpha}\sqrt{f(\kappa_0)} + z_{1-\beta}\sqrt{f(\kappa_a)}}{\kappa_a - \kappa_0} \right)^2 \quad (4.55)$$

- **Example 77 (continued):** Let us assume that the probability of both veterinarians of assessing a procedure as mild is 0.5, moderate is 0.3, severe is 0.15, and non-recovery 0.05. Then

$$\begin{aligned} p_e &= 0.5^2 + 0.3^2 + 0.15^2 + 0.05^2 = 0.365 \\ f(\kappa_0) &= \frac{[0.8(1-0.365)+0.365][1-(0.8(1-0.365)+0.365)]}{(1-0.365)^2} = 0.2750 \\ f(\kappa_1) &= \frac{[0.9(1-0.365)+0.365][1-(0.9(1-0.365)+0.365)]}{(1-0.365)^2} = 0.1475 \\ N &\geq \left( \frac{z_{0.95}\sqrt{0.2750} + z_{0.9}\sqrt{0.1475}}{0.9-0.8} \right)^2 \Rightarrow N = 184 \end{aligned}$$

That is,  $N = 184$  different procedures need to be evaluated by both veterinarians.

If we want to estimate Cohen's  $\kappa$  with a confidence interval of length  $\Delta$  (between minimum and maximum) and a confidence level  $1 - \alpha$ , then we must use

$$N \geq \left( \frac{2z_{1-\frac{\alpha}{2}}\sqrt{f(\kappa_0)}}{\Delta} \right)^2 \quad (4.56)$$

where  $\kappa_0$  is an approximate, expected value of Cohen's  $\kappa$  once we perform the experiment.

#### Important remarks

96. As happened with other sample size calculations related to proportions, the sample size formula for the Cohen's  $\kappa$  also requires that before doing the experiment we have some clue of which the results will be approximately.

In particular, which will be the expected frequency of each one of the categories.

## 4.7 Sample size for survival analysis

Survival analysis is a statistical technique that models the time elapsed until an event occurs. One of the events of interest that originally drove the development of the technique was death (and, therefore, its name “survival”). However, the event may be any other one like time to stop a habit, or time to first visit. In Sec. 4.3.4 we briefly introduced the main concepts associated to survival analysis, and the interested reader is referred to that section before going into the details of the sample size calculations.

### 4.7.1 Confidence interval for the mean survival time

- **Example 78:** We are developing a new antitumoral therapy. Animals start receiving the treatment at a given dose when the tumor has grown to a given size. Then, we measure the time to the disappearance of the tumor. Some of the tumors do not respond to the treatment. We will observe the animals for 6 months and we expect that after this time, 85% of the tumors have disappeared. How many animals do we need to study if we want to construct a 95% confidence interval for the mean time to disappearance whose half width is only 20% of the nominal value?

The event of interest is in this case the disappearance of the tumor. if we have  $N$  animals, let us refer to the time to disappearance in the animal  $i$  as  $t_i$ . We can have two stopping criteria:

1. By maximum observation time: If our experiment reaches a maximum time,  $t_{max}$ , e.g., six months.
2. By maximum number of events: If our experiment reaches a maximum number of events,  $r_{max}$ , e.g., 10 disappeared tumors.

In any case, let us assume that we have been observing a time  $t_{obs}$  and that within this time  $r$  tumors have disappeared. For the tumors that did not disappear we will have a “censored” measurement  $t_i = t_{obs}$ . We estimate the mean survival time (understood as mean time until the event) as

$$\hat{\mu} = \frac{1}{r} \sum_{i=1}^N t_i \quad (4.57)$$

### Exponential survival

For an exponential survival, the survival function is given by

$$S(t) = \Pr\{T > t\} = e^{-\frac{t}{\mu}}$$

where  $\mu$  is the mean survival time. Under this model, the instantaneous hazard is constant and equal to

$$\lambda(t) = \frac{1}{\mu}$$

The mean survival time for this model is

$$MST = \mu$$

For the exponential survival, it can be shown that the  $1 - \alpha$  confidence interval for the mean survival time is

$$\Pr \left\{ \frac{2r\hat{\mu}}{\chi_{1-\frac{\alpha}{2},v}^2} < \mu < \frac{2r\hat{\mu}}{\chi_{\frac{\alpha}{2},v}^2} \right\} = 1 - \alpha$$

where  $v = 2r$  if the test is terminated after  $r_{max}$  events, or  $v = 2(r + 1)$  if the test is terminated after a fixed time  $t_{max}$ .

This interval is asymmetric: the lower bound is closer to  $\mu$  than the upper bound. In this way, we can reorganize the upper bound,  $\mu_U$ , so that we calculate its width as a fraction of the nominal value  $\hat{\mu}$

$$\frac{\mu_U}{\hat{\mu}} = \frac{2r}{\chi_{\frac{\alpha}{2},2(r+1)}^2}$$

If we want this width to be smaller than a given value  $1 + \Delta$ , then we need a number of events  $r$  such that

$$\boxed{\frac{2r}{\chi_{\frac{\alpha}{2},2(r+1)}^2} < 1 + \Delta} \quad (4.58)$$

### Weibull survival

Exponential survival assumes a constant hazard along the animal life, or the duration of the experiment. As we saw in Sec. 4.3.4, there are situations in which the hazard is larger at the early or late parts of the experiment. Weibull survival generalizes exponential survival and can adapt to any of these situations. The survival function is in this case

$$S(t) = \Pr\{T > t\} = e^{-\left(\frac{t}{\mu}\right)^\beta}$$

The corresponding hazard is

$$\lambda(t) = \frac{\beta}{\mu} \left(\frac{t}{\mu}\right)^{\beta-1}$$

For  $\beta = 1$ , Weibull survival is exactly the same as the exponential survival. For  $\beta < 1$ , hazard decreases over time (infant mortality); while for  $\beta > 1$ , hazard increases over

time (mortality in the elderly). See Fig. 4.10 for a representation of these curves. The mean survival time corresponding to this model is

$$MST = \mu \Gamma\left(1 + \frac{1}{\beta}\right)$$

where  $\Gamma$  is the gamma function (a generalization of the factorial, for integer values  $x$ , we have  $\Gamma(1+x) = x!$ ). The Mean Survival Time is now determined by two parameters  $\mu$  (a scale parameter, the larger  $\mu$ , the larger the MST) and  $\beta$  (a shape parameter). It now makes sense to determine confidence intervals for both parameters. It can be shown that for the scale parameter we have

$$\Pr\left\{\hat{\mu}\left(1 - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{r}}\right)^{\frac{1}{\beta}} < \mu < \hat{\mu}\left(1 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{r}}\right)^{\frac{1}{\beta}}\right\} = 1 - \alpha$$

while for the shape parameter we have

$$\Pr\left\{\hat{\beta}\left(1 - \frac{\sqrt{6} z_{1-\frac{\alpha}{2}}}{\pi \sqrt{r}}\right) < \beta < \hat{\beta}\left(1 + \frac{\sqrt{6} z_{1-\frac{\alpha}{2}}}{\pi \sqrt{r}}\right)\right\} = 1 - \alpha$$

We can calculate the number of required number of events  $r$  by fixing the desired half-width  $\Delta$  either in the determination of  $\mu$  or  $\beta$

$$\begin{aligned} \left(1 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{r}}\right)^{\frac{1}{\beta}} = 1 + \Delta &\Rightarrow r = \left(\frac{z_{1-\frac{\alpha}{2}}}{(1+\Delta)^{\beta} - 1}\right)^2 \\ 1 + \frac{\sqrt{6} z_{1-\frac{\alpha}{2}}}{\pi \sqrt{r}} = 1 + \Delta &\Rightarrow r = 6 \left(\frac{z_{1-\frac{\alpha}{2}}}{\pi \Delta}\right)^2 \end{aligned} \quad (4.59)$$

### Gaussian survival

Some variables have a “delayed” hazard that causes the event to be concentrated around a particular time (e.g., expiration dates of food, the probability that a product is spoiled is negligible at the beginning of its life). We expect the event to occur around a time  $\mu$  with a standard deviation  $\sigma$ . This can be modelled with a Gaussian survival curve. The survival and hazard functions are given by

$$\begin{aligned} S(t) &= 1 - F\left(\frac{t-\mu}{\sigma}\right) \\ \lambda(t) &= \frac{f\left(\frac{t-\mu}{\sigma}\right)}{1 - F\left(\frac{t-\mu}{\sigma}\right)} \end{aligned}$$

where  $F(x)$  and  $f(x)$  are the cumulative and probability density functions of the standardized Gaussian, respectively.

An approximate confidence interval for the location parameter is the standard one for a Gaussian

$$\Pr\left\{\hat{\mu} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{N}} < \mu < \hat{\mu} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{N}}\right\} = 1 - \alpha$$

$\hat{\sigma}$  is the standard sample estimate of the time to the event of interest. If we want the half-width to be  $\Delta$ , then the number of animals required is

$$N = \left( \frac{z_{1-\frac{\alpha}{2}}}{\Delta/\sigma} \right)^2 \quad (4.60)$$

The experiment must be carried out until all animals have experienced the event of interest.

### Number of animals

Except for the Gaussian case, these formulas tell us the number of events required, but not the number of animals required. Not all tumors respond to the treatment and, consequently, we will need more animals than  $r$ . In particular, if  $S_{max}$  tumors have not disappeared after a time  $t_{max}$ , we need

$$N = \frac{r}{1 - S_{max}} \quad (4.61)$$

- **Example 78 (continued):** In our example we wanted a precision of  $\Delta = 0.2$  **Exponential survival**  
In our case, the desired width is 20%, which implies

$$\frac{2r}{\chi_{0.025, 2(r+1)}^2} = 1.2 \Rightarrow r = 115$$

$$N = \frac{115}{1-0.15} = 136$$

That is we will use  $N = 136$  animals with tumors. After 6 months of treatment, we expect that 85% of them have disappeared. With the number of disappearances we will be able to construct a confidence interval whose maximum deviation from the nominal mean survival time,  $\hat{\mu}$  is 20%.

### Weibull survival

Let us assume now that our drug accumulates in the tumor so that it is more effective as time passes. This implies that the shape parameter  $\beta > 1$ . From previous studies, it is not unreasonable to assume  $\beta \geq 1.5$  (the most conservative design, larger number of samples, is obtained for the lower bound, that is  $\beta = 1.5$ ). If we design for the confidence interval of the scale parameter, then the required sample size would be

$$r = \left( \frac{z_{0.975}}{(1+0.2)^{1.5} - 1} \right)^2 = 39$$

$$N = \frac{39}{1-0.15} = 46$$

We see that the sample size drastically changes if we assume that the drug is more effective as it accumulates in the tumor.

- **Example 79:** We are interested in the mean time at which young male mice surpass a weight of 12 grams (an adult mouse can weigh over 25 grams). This

should happen about day 21, with a standard deviation of about 2 days. We want to determine a 95% confidence interval with a maximum half-width of 1 day. How many animals do we need for this?

#### Normal survival

In this example we have  $\Delta = 1$  and  $\sigma = 2$ . The sample size, would be

$$N = \left( \frac{z_{0.975}}{1/2} \right)^2 = 16$$

All animals must be followed until they all surpass the 12 grams of weight.

#### Important remarks

97. As is the case in many of the sample size calculations involving proportions, the sample size for the mean survival time requires a prior guess of the proportion of animals for which the event has not been observed at the end of the experiment. For Weibull survival we must also assume a prior value for the survival shape parameter  $\beta$ .

### 4.7.2 Confidence interval for survival time percentile

- **Example 80:** Following Example 78, we are interested in constructing a confidence interval for the time at which 25% of the tumors have disappeared. We want that the maximum half-width of the interval is 20% of its central value. How many animals do we need for this? Let us assume that we expect the tumors to disappear after 3 months of treatment on average. We plan to conduct our experiment for about 2 months, but the experiment will be terminated by a given number of tumors disappearing, not by maximum time.

For fully understanding the following paragraphs, we recommend the reader to be familiar with the different survival models described in the previous section.

#### Exponential survival

We would normally perform our survival experiment and estimate the mean survival time,  $\hat{\mu}$  as indicated in Eq. 4.57. Let us call  $S_p$  the “surviving tumors” at the percentile time of interest,  $t_p$ . In our example,  $S_p = 75\%$ . For an exponential survival model (see previous section), our estimate of the corresponding time  $t_p$  to reach this situation is given by

$$S_p = e^{-\frac{t_p}{\mu}} \Rightarrow t_p = \mu(-\log(S_p)) \Rightarrow \hat{t}_p = \hat{\mu}(-\log(S_p))$$

For the exponential survival it can be shown that

$$\Pr \left\{ \hat{t}_p \frac{2r}{\chi_{1-\frac{\alpha}{2}, v}^2} < t_p < \hat{t}_p \frac{2r}{\chi_{\frac{\alpha}{2}, v}^2} \right\} = 1 - \alpha$$

where  $v = 2r$  if the test is terminated after  $r$  events and  $v = 2(r + 1)$  if it is terminated by maximum time.

This problem is formally identical to the one of the previous section interchanging  $\mu$  by  $t_p$ . As we reasoned in the previous section, the number of required events is

$$\boxed{\frac{2r}{\chi_{\frac{\alpha}{2}, v}^2} < 1 + \Delta} \quad (4.62)$$

where  $\Delta$  is the half-width, in our example  $\Delta = 0.2$ .

### Weibull survival

The calculations needed if the survival follows a Weibull model are formally identical to those followed above for the exponential survival, only that the expressions slightly change

$$S_p = e^{-\left(\frac{t_p}{\mu}\right)^\beta} \Rightarrow t_p = \mu(-\log(S_p))^{\frac{1}{\beta}} \Rightarrow \hat{t}_p = \hat{\mu}(-\log(S_p))^{\frac{1}{\beta}}$$

and the confidence interval

$$\Pr \left\{ \hat{t}_p \left( \frac{2r}{\chi_{1-\frac{\alpha}{2}, v}^2} \right)^{\frac{1}{\beta}} < t_p < \hat{t}_p \left( \frac{2r}{\chi_{\frac{\alpha}{2}, v}^2} \right)^{\frac{1}{\beta}} \right\} = 1 - \alpha$$

where  $v = 2r$  if the test is terminated after  $r$  events and  $v = 2(r + 1)$  if it is terminated by maximum time. The required number of observations must fulfill

$$\boxed{\left( \frac{2r}{\chi_{\frac{\alpha}{2}, v}^2} \right)^{\frac{1}{\beta}} < 1 + \Delta} \quad (4.63)$$

### Gaussian survival

As for the previous distributions, let us refer to the time as  $t_p$  the time percentile for which the survival rate is  $S_p$ . Let us define  $z_{1-S_p}$  as the  $1 - S_p$  percentile of the Gaussian. An approximate confidence interval for the time percentile is

$$\Pr \left\{ \hat{t}_p - z_{1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{N} \left( 1 + \frac{1}{2} z_{1-S_p}^2 \right)} < t_p < \hat{t}_p + z_{1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{N} \left( 1 + \frac{1}{2} z_{1-S_p}^2 \right)} \right\} = 1 - \alpha$$

If we want the half-width to be  $\Delta$ , then the number of animals required is

$$\boxed{N = \left( \frac{z_{1-\frac{\alpha}{2}}}{\Delta/\hat{\sigma}} \right)^2 \left( 1 + \frac{1}{2} z_{1-S_p}^2 \right)} \quad (4.64)$$

For the Gaussian survival, the experiment must be carried out until all animals have experienced the event of interest. Once the experiment is performed, we will estimate the sample mean and standard deviation of the time to the event. Then, the estimate of  $t_p$  will be

$$\hat{t}_p = \hat{\mu} + z_{1-S_p} \hat{\sigma}$$

**Number of animals**

To calculate the number of animals for the exponential and Weibull survivals we will use a different reasoning, that could also have been followed in the previous section. Let us assume that we have a prior guess of the survival distribution ( $\mu_0$  for the exponential survival, and  $\mu_0$  and  $\beta_0$  for Weibull survival). Then, if we start with  $N$  animals, we expect this number of failures to occur at a time  $t_r$  given by

$$S(t_r) = 1 - \frac{r}{N} = e^{-\frac{t_r}{\mu_0}}$$

or

$$S(t_r) = 1 - \frac{r}{N} = e^{-\left(\frac{t_r}{\mu_0}\right)^{\beta_0}}$$

Assume that we want to wait about a fixed  $t_r$ , then the number of required animals will be

$$\begin{aligned} N &= \frac{r}{1 - e^{-\frac{t_r}{\mu_0}}} \\ N &= \frac{r}{1 - e^{-\left(\frac{t_r}{\mu_0}\right)^{\beta_0}}} \end{aligned} \quad (4.65)$$

- Example 80 (continued): In our example we desire a precision of  $\Delta = 0.2$ .

**Exponential survival**

If the Mean Survival Time is 3 months, then  $\mu_0 = 3$ . We plan to conduct our experiment for about  $t_r = 2$  months. Then, the required number of disappearing tumors is

$$\begin{aligned} \frac{2r}{\chi_{0.025, 2r}^2} &= 1.2 \Rightarrow r = 127 \\ N &= \frac{127}{1 - e^{-\frac{2}{3}}} = 262 \end{aligned}$$

That is we will use  $N = 262$  animals with tumors. After 2 months of treatment, we expect that  $e^{-\frac{2}{3}} = 51.3\%$  of them have disappeared. In any case, we will wait until  $r = 127$  tumors have disappeared. Then, we will estimate the time at which 25% of the tumors have disappeared as

$$\hat{t}_p = -\hat{\mu} \log(0.75)$$

If we were not too wrong about our initial guess of the mean survival time, it should be about

$$\hat{t}_p \approx -\mu_0 \log(0.75) = -3 \log(0.75) = 0.86 \text{ months}$$

The confidence interval will be in any case

$$\left( \hat{t}_p \frac{2r}{\chi_{1-\frac{\alpha}{2}, 2r}^2}, \hat{t}_p \frac{2r}{\chi_{\frac{\alpha}{2}, 2r}^2} \right) = (0.8466\hat{t}_p, 1.1995\hat{t}_p)$$

**Weibull survival**

Let us assume that our drug is more effective as it accumulates in the tumor. Previous studies have shown that we may assume  $\beta \geq 1.5$ . We will do the design in the worse case,  $\beta = 1.5$ . If the Mean Survival Time is 3 months, then we can calculate a first estimate of the scale parameter  $\mu_0$  as (see the introduction of the Weibull survival in the previous section)

$$3 = \mu_0 \Gamma\left(1 + \frac{1}{1.5}\right) \Rightarrow \mu_0 = 3.32$$

Then, the sample size can be calculated by

$$\left(\frac{2r}{\chi_{0.025, 2r}^2}\right)^{\frac{1}{1.5}} < 1.2 \Rightarrow r = 59$$

$$N = \frac{59}{1 - e^{-\left(\frac{3}{3.32}\right)^{1.5}}} = 158$$

Again, we observe that the sample sizes for an exponential survival (constant hazard) and Weibull survival (increasing hazard, in this example) are rather different. Additionally, if we are right about our initial estimates of the survival parameters, our expected value for the time at which 25% of the tumors have disappeared would be about

$$\hat{t}_p = 3.32(-\log(0.75))^{\frac{1}{1.5}} = 1.44 \text{ months}$$

As expected, since the drug is more effective by accumulation, we would expect the early percentile times (we are studying the 25% percentile) to be larger than for the exponential case. The drug appears to be initially slower.

- **Example 81:** We are interested in the time at which 95% of young male mice surpass a weight of 12 grams (an adult mouse can weigh over 25 grams). The average time at which half of the population surpasses this weight is about day 21, with a standard deviation of about 2 days. We want to determine a 95% confidence interval with a half-width of 1 day. How many animals do we need for this?

**Normal survival**

In this example we have  $\Delta = 1$  and  $\sigma = 2$ . The sample size, would be

$$N = \left(\frac{z_{0.975}}{1/2}\right)^2 \left(1 + \frac{1}{2} z_{0.95}^2\right) = 37$$

All animals must be followed until they all surpass the 12 grams of weight. If we were right about our prior,  $\mu = 21$  and  $\sigma = 2$ , then our estimate should be close to

$$\hat{t}_p = 21 + z_{0.95}2 = 21 + 1.64 \cdot 2 = 24 \text{ days}$$

### 4.7.3 Confidence interval for survival rate

- Example 82: Following the example in the previous two sections, we are interested in constructing a confidence interval for the proportion of tumors that have not disappeared after 6 months of treatment. We expect the mean time of disappearance to be about 3 months. We want to construct a confidence interval whose full width is 10% (for instance, if the final survival is about 15%, a possible confidence interval could be from 10 to 20%).

The sample size for the survival rate at a given time is based on the confidence interval for the scale parameter  $\mu$  of the survival model as was presented in Sec. 4.7.1. Assume that the scale parameter confidence interval is  $(\mu_L, \mu_U)$ . Then, the confidence interval for the survival rate is defined by the survival rates corresponding to the lower and upper limits of the scale parameter.

#### Exponential survival

The confidence interval for the scale parameter is

$$(\mu_L, \mu_U) = \left( \frac{2r\hat{\mu}}{\chi_{1-\frac{\alpha}{2}, \nu}^2}, \frac{2r\hat{\mu}}{\chi_{\frac{\alpha}{2}, \nu}^2} \right)$$

where  $\nu = 2r$  if the test is terminated after  $r_{max}$  events, or  $\nu = 2(r+1)$  if the test is terminated after a fixed time  $t_{max}$ . The corresponding confidence interval for the survival rate would be

$$(S_L, S_U) = \left( e^{-\frac{t}{\mu_L}}, e^{-\frac{t}{\mu_U}} \right)$$

The number of required events,  $r$ , determines the confidence interval of the scale parameter, and in their turn, the width of the confidence interval of the survival rate. For the calculation of the sample size, we can fix the maximum width of this confidence interval

$$\boxed{e^{-\frac{t}{2r\hat{\mu}}} - e^{-\frac{t}{2r\hat{\mu}}} < \Delta} \quad (4.66)$$

#### Weibull survival

In this case, we need to assume that the shape parameter is known,  $\beta_0$ . The confidence interval for the scale parameter is

$$\left( \hat{\mu} \left( 1 - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{r}} \right)^{\frac{1}{\beta_0}}, \hat{\mu} \left( 1 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{r}} \right)^{\frac{1}{\beta_0}} \right)$$

The corresponding confidence interval for the survival rate would be

$$(S_L, S_U) = \left( e^{-\left(\frac{t}{\mu_L}\right)^{\beta_0}}, e^{-\left(\frac{t}{\mu_U}\right)^{\beta_0}} \right)$$

and the sample size design formula

$$\boxed{e^{-\left(\frac{t}{\hat{\mu}}\right)^{\beta_0} \frac{1}{1-\frac{\alpha}{2}}} - e^{-\left(\frac{t}{\hat{\mu}}\right)^{\beta_0} \frac{1}{1+\frac{\alpha}{2}}} < \Delta} \quad (4.67)$$

### Gaussian survival

An approximate confidence interval for the survival rate is

$$\Pr \left\{ \hat{S}(t) - z_{1-\frac{\alpha}{2}} f(\hat{z}) \sqrt{\frac{1}{N} \left(1 + \frac{1}{2} \hat{z}^2\right)} < S(t) < \hat{S}(t) + z_{1-\frac{\alpha}{2}} f(\hat{z}) \sqrt{\frac{1}{N} \left(1 + \frac{1}{2} \hat{z}^2\right)} \right\} = 1 - \alpha$$

where  $\hat{z} = \frac{t - \hat{\mu}}{\hat{\sigma}}$  and  $f(x)$  is the probability density function of the standardized Gaussian function. If we want the half-width to be  $\Delta$ , then the number of animals required is

$$\boxed{N = \left( \frac{z_{1-\frac{\alpha}{2}}}{\Delta / f(\hat{z})} \right)^2 \left(1 + \frac{1}{2} \hat{z}^2\right)} \quad (4.68)$$

The experiment must be carried out until all animals have experienced the event of interest.

### Number of animals

As usual, the number of animals for the exponential and Weibull survivals is given by

$$\boxed{N = \frac{r}{1 - S}} \quad (4.69)$$

where  $S$  is the expected survival at the end of the experiment.

- **Example 82 (continued):** In our example we desire a precision of  $\Delta = 0.1$ . The Mean Survival Time is about 3 months, and the experimental time  $t = 6$  months.

### Exponential survival

From the MST we estimate  $\mu_0 = 3$ . We will stop the experiment exactly after 6 months, so the number of degrees of freedom for the  $\chi^2$  is  $\nu = 2(r+1)$ . Then, we must find  $r$  such that

$$e^{-\frac{6\chi_{0.025,2(r+1)}^2}{2r \cdot 3}} - e^{-\frac{6\chi_{0.975,2(r+1)}^2}{2r \cdot 3}} < 0.1 \Rightarrow r = 111$$

Then, the number of animals is

$$N = \frac{111}{1 - e^{-\frac{6}{3}}} = 129$$

### Weibull survival

Following the examples above, let us assume that  $\beta = 1.5$ , that is, the drug is

more effective as it accumulates. In Example 80 we showed that the estimate for  $\mu_0$  was  $\mu_0 = 3.32$ . Then we must find  $r$  such that

$$e^{-\left(\frac{6}{3.32}\right)^{1.5} \frac{1}{1 - \frac{z_{0.975}}{\sqrt{r}}}} - e^{-\left(\frac{6}{3.32}\right)^{1.5} \frac{1}{1 + \frac{z_{0.975}}{\sqrt{r}}}} < 0.1 \Rightarrow r = 31$$

The number of animals needed is

$$N = \frac{31}{1 - e^{-\left(\frac{6}{3.32}\right)^{1.5}}} = 34$$

- **Example 83:** We are interested in the proportion of young males that have not surpassed a weight of 12 grams after 24 days (this would be delayed animals). The average time at which half of the population surpasses this weight is about day 21, with a standard deviation of about 2 days. We want to determine a 95% confidence interval with a half-width of 5%. How many animals do we need for this?

#### Normal survival

In this example we have  $\Delta = 0.05$  and  $\sigma = 2$ . We are interested at the survival rate at day  $t = 24$ . If we are right about our initial estimates of the mean and standard deviation, then the corresponding  $\hat{z}$  would be close to

$$\hat{z} = \frac{24 - 21}{2} = 1.5$$

The sample size, would be

$$N = \left( \frac{z_{0.975}}{0.05/f(1.5)} \right)^2 \left( 1 + \frac{1}{2} 1.5^2 \right) = 55$$

All animals must be followed until they all surpass the 12 grams of weight. If we were right about our prior,  $\mu = 21$  and  $\sigma = 2$ , then our estimate should be close to

$$\hat{S}(24) = 1 - F(1.5) = 6.7\%$$

### 4.7.4 Hypothesis test for one sample mean survival time

For a definition of the survival models presented in this chapter, the reader is referred to Sec. 4.7.1

- **Example 84:** We are interested in the duration of some intradermal electronic implants in animals. These implants deteriorate over time due to biodegradation of the implant material. We want to show that the mean duration is larger than 3 years. How many animals do we need to test for 6 months and how should we carry the experiment? After 6 months of experiment, the implants will be recovered and their state (deteriorated or not) will be determined.

Reliability demonstration tests are performed by testing  $N$  (to be calculated) units for a prespecified time  $t_{max}$ , or by testing  $N$  prespecified units for a time  $t_{max}$  (to be calculated). The test responds to the general scheme

$$\begin{aligned} H_0 &: MST \leq MST_0 \\ H_A &: MST > MST_0 \end{aligned}$$

where  $MST$  is the mean survival time.

If  $H_0$  is true,  $H_0$  is rejected if

$$\sum_{r=0}^R b(r; N, t_{max}) < \alpha$$

where  $N$  is the number of tested individuals,  $t_{max}$  is the time of observation,  $R$  is the maximum allowed number of failing individuals, and  $b(r; N, t_{max})$  is the binomial probability

$$b(r; N, t_{max}) = \binom{N}{r} (1 - S(t_{max}))^r (S(t_{max}))^{N-r} \quad (4.70)$$

$S(t_{max})$  is the probability of surviving at  $t_{max}$  if  $\mu = \mu_0$ .

If we fix two of the three parameters ( $N, t_{max}, R$ ), the other one can be solved through the rejection equation

$$\boxed{\sum_{r=0}^R b(r; N, t_{max}) < \alpha} \quad (4.71)$$

It is customary to fix  $R = 0$  to obtain a minimum sample size,  $N$ . Then

$$\boxed{S^N(t_{max}) = \alpha \Rightarrow N > \frac{\log(\alpha)}{\log(S(t_{max}))}} \quad (4.72)$$

The following table is useful to calculate  $t_{max}$  as a function of  $S(t_{max})$  or viceversa.

Exponential	$t_{max} = -\mu_0 \log(S(t_{max}))$	$S(t_{max}) = e^{-t_{max}/\mu_0}$
Weibull	$t_{max} = \mu_0 (-\log(S(t_{max})))^{1/\beta_0}$	$S(t_{max}) = e^{-(t_{max}/\mu_0)^{\beta_0}}$
Normal	$t_{max} = \mu_0 + z_{max} \sigma$	$S(t_{max}) = 1 - F(z_{max})$ $z_{max} = \frac{t_{max} - \mu_0}{\sigma}$

- Example 84 (continued): In our example,  $\mu_0 = 3$  years, and  $t_{max} = 0.5$  years.

#### Exponential survival

For exponential survival, if the MST is 3 years, then  $\mu_0 = 3$  years. We can calculate the expected survival at 6 months, if the mean survival time is 3 years

$$S(0.5) = e^{-\frac{0.5}{3}} = 85\%$$

The number of required animals is calculated from Eq. 4.72

$$N = \frac{\log(0.05)}{\log(0.85)} = 18$$

That is, we will put an implant to 18 animals. After 6 months, we will recover the implants, if none of them ( $R = 0$ ) are deteriorated, we reject the hypothesis that the mean survival time of the implants is less than 3 years.

#### Weibull survival

Let us assume that the deterioration is progressive so that the hazard of failure of the implant grows over time. From previous studies, we have determined that  $\beta_0 = 1.5$  represents the increase of hazard of failure as time progresses. If the MST is 3 years, then the scale parameter of the Weibull distribution can be calculated by

$$MST = \mu_0 \Gamma(1 + \beta_0) \Rightarrow \mu_0 = \frac{3}{\Gamma(2.5)} = 2.26 \text{ years}$$

Then, the expected survival after 6 months will be

$$S(0.5) = e^{-\left(\frac{0.5}{2.26}\right)^{1.5}} = 90\%$$

The number of animals is now

$$N = \frac{\log(0.05)}{\log(0.90)} = 29$$

We would carry out the experiment in exactly the same way as for the exponential survival.

- **Example 85:** In the previous example, we want to reduce the number of animals to just  $N = 10$ . After how much time should we recover the implants to see if they are deteriorated?

#### Exponential survival

From Eq. 4.72 at the end of the study period, the survival rate must be

$$S^{10}(t_{max}) = 0.05 \Rightarrow S(t_{max}) = 74\%$$

From the table relating  $t_{max}$  and  $S(t_{max})$  we obtain that we need to study these animals for

$$t_{max} = -3 \log(0.74) = 0.90 \text{ years} = 329 \text{ days}$$

That is, we will put an implant to 10 animals. After 329 days, we will recover the implants, if none of them ( $R = 0$ ) are deteriorated, we reject the hypothesis that the mean survival time of the implants is less than 3 years.

#### Weibull survival

The survival rate at  $t_{max}$  is the same as in the exponential case. And we only need to calculate  $t_{max}$  from the table equations:

$$t_{max} = 2.26(-\log(0.74))^{1/1.5} = 1.01 \text{ years} = 370 \text{ days}$$

We would carry out the experiment in exactly the same way as for the exponential survival.

### 4.7.5 Hypothesis test for one sample survival rate

- **Example 86:** Following the Example 84, we want to give a guarantee time for the implants. We want to show that 90% of the implants work after 6 months. How many implants do we need to perform if we are going to monitor them for 3 months?

Reliability demonstration tests are performed by testing  $N$  (to be calculated) units for a prespecified time  $t_{max}$  or by testing  $N$  prespecified units for a time  $t_{max}$  (to be calculated). The test responds to the general scheme

$$\begin{aligned} H_0 &: S(t_0) \leq S_0 \\ H_a &: S(t_0) > S_0 \end{aligned}$$

We need to distinguish between the time of experimentation  $t_{max}$  (in our example 3 months) and the time we are interested in for the test  $t_0$  (in our example 6 months). The following table helps us to translate survivals at  $t_0$  ( $S(t_0)$ ) into survivals at  $t_{max}$  ( $S(t_{max})$ ). It also gives us expressions to calculate  $t_{max}$  if we fix  $t_0$ ,  $S(t_0)$  and  $S(t_{max})$ .

Exponential	$t_{max} = t_0 \frac{\log(S(t_{max}))}{\log(S(t_0))}$	$S(t_{max}) = (S(t_0))^{t_{max}/t_0}$
Weibull	$t_{max} = t_0 \left( \frac{\log(S(t_{max}))}{\log(S(t_0))} \right)^{1/\beta}$	$S(t_{max}) = (S(t_0))^{(t_{max}/t_0)^\beta}$
Normal	$t_{max} = t_0 + (z_0 + z_{max})\sigma$	$S(t_{max}) = 1 - F(z_{max})$ $z_0 = z_{1-S(t_0)}; z_{max} = z_0 + \frac{t_{max}-t_0}{\sigma}$

Then, the goal is to construct a confidence interval of the form

$$\Pr\{S_{max} < S(t_{max}) < 1\} = 1 - \alpha$$

If  $H_0$  is true,  $H_0$  is rejected if

$$\sum_{r=0}^R b(r; N, t_{max}) < \alpha$$

A in the previous section,  $R$  is the number failures observed during the experiment, and  $b(r; N, t_{max})$  is the binomial probability of observing exactly  $r$  failures in  $N$  individuals at time  $t_{max}$  when the probability of survival is  $S_{max}$  (see Eq. 4.70). Given two of the variables ( $R, N, t_{max}$ ) we can solve for the other. The minimum  $N$  is obtained for  $R = 0$ , so that the sample size design formula is given by

$$S_{max}^N < \alpha \Rightarrow N > \frac{\log(\alpha)}{\log(S_{max})} \quad (4.73)$$

- **Example 86 (continued):** In our example  $t_0 = 0.5$  years,  $S_0 = 90\%$ , and  $t_{max} = 0.25$  years.

#### Exponential survival

We now translate the threshold survival for  $t_0$  into a threshold survival for  $t_{max}$

$$S_{max} = (0.9)^{0.25/0.5} = 0.9487$$

Then, we require

$$N = \frac{\log(0.05)}{\log(0.9487)} = 57$$

That is, we perform  $N = 57$  implants. If in 3 months we do not observe any deterioration, then we reject the hypothesis that the survival rate at 6 months is smaller than 0.9.

#### **Weibull survival**

For a Weibull survival with an expected  $\beta = 1.5$ , we would have

$$S_{max} = (0.9)^{(0.25/0.5)^{1.5}} = 0.9634$$

Then, we require

$$N = \frac{\log(0.05)}{\log(0.9634)} = 81$$

- Example 87: If we think that  $N = 57$  or  $N = 81$  are too many mice, we can extend the period of the experiment. Let us say that we are willing to use  $N = 15$  animals. Then, we can adapt the experimentation time  $t_{max}$  as follows. The threshold at which we would reject the null hypothesis would be given by

$$S_{max}^{15} = 0.05 \Rightarrow S_{max} = 0.8190$$

Using the formulas in the table, we now look for the time at which this threshold is attained

#### **Exponential survival**

$$t_{max} = 0.5 \frac{\log(0.8190)}{\log(0.9)} = 0.95 \text{ years} = 346 \text{ days}$$

That is, we perform  $N = 15$  implants. If in 346 days we do not observe any deterioration, then we reject the hypothesis that the survival rate at 6 months is smaller than 0.9.

#### **Weibull survival**

For a Weibull survival with an expected  $\beta = 1.5$ , we would have

$$t_{max} = 0.5 \left( \frac{\log(0.8190)}{\log(0.9)} \right)^{1/1.5} = 0.77 \text{ years} = 280 \text{ days}$$

For a short period of experimentation,  $t_{max} = 3$  months, we need more animals for the Weibull survival. However, for longer periods, as the Weibull (with  $\beta > 1$ ) accumulates failures more quickly than the exponential survival, we need fewer days or animals compared to the exponential case.

### 4.7.6 Hypothesis test for two samples with exponential survival

- **Example 88:** We are developing a new analgesic. To test it, we have two groups of animals: one will receive the drug (treatment group), while the other one will not (control group). We will place animals in a hot plate, and they normally jump out of the plate, so that the mean survival time (MST) in the plate is about 4 seconds. Our hypothesis is that the animals with the drug will stay longer (to avoid burns, we take the animals out of the plate if they stay more than 10 seconds). We want to have a statistical power of 90% if the MST increases to 6 seconds or more. How many animals do we need in each of the groups?

Our test is now of the form

$$\begin{aligned} H_0 : \mu_1 &\leq \mu_2 \\ H_a : \mu_1 &> \mu_2 \end{aligned}$$

The test uses the statistic

$$F = \frac{\hat{\mu}_1}{\hat{\mu}_2}$$

where

$$\hat{\mu}_i = \frac{1}{r_i} \sum_{j=1}^{N_i} t_{ij}$$

$r_i$  and  $N_i$  are the number of events (animals jumping out) and samples of the  $i$ -th treatment, and  $t_{ij}$  is the time-to-event (or the observation time, 10 seconds in our example, if a given individual does not jump out) of the  $j$ -th sample of the  $i$ -th treatment.  $F$  follows a Snedecor's F distribution with  $2r_1$  and  $2r_2$  degrees of freedom. If the event (jumping out) occurred in all animals, then the number of samples can be calculated from the following equation

$$\mu_1 F_{\beta, 2N, 2N} = \mu_2 F_{1-\alpha, 2N, 2N} \quad (4.74)$$

where  $F_{\beta, 2N, 2N}$  is the  $\beta$  percentile of the Snedecor's F with  $2N$  and  $2N$  degrees of freedom, analogously for  $F_{1-\alpha, 2N, 2N}$ .

- **Example 88 (continued):** We can calculate the sample size by solving the equation

$$6F_{0.1, 2N, 2N} = 4F_{0.95, 2N, 2N} \Rightarrow N = 105$$

If we expect that some animals still survive, they have not jumped out and we will have to take them out, at the end of the experiment, 10 seconds in our example, then we should modify the design equation to account for the surviving rates in the two groups,  $S_1$  and  $S_2$ .

$$\mu_1 F_{\beta, 2N(1-S_1), 2N(1-S_2)} = \mu_2 F_{1-\alpha, 2N(1-S_1), 2N(1-S_2)} \quad (4.75)$$

Additionally, since the survival is exponential, we have  $S_1 = S_2 e^{-\frac{\mu_2}{\mu_1}}$ .

- Example 88 (continued): With  $t_{max} = 10$  seconds, we expect some animals that have not jumped out

$$\begin{aligned} S_1 &= e^{-\frac{10}{4}} = 8.2\% \\ S_2 &= e^{-\frac{10}{6}} = 18.9\% \end{aligned}$$

$$6F_{0.1,2N-0.92,2N-0.81} = 4F_{0.95,2N-0.92,2N-0.81} \Rightarrow N = 122$$

#### 4.7.7 Hypothesis test for two samples with log-rank test

If we do not want to assume a particular distribution for the survival time, we may perform a log-rank test (also called Mantel-Cox test)

$$\begin{aligned} H_0 &: \lambda_1(t) \geq \lambda_2(t) \\ H_a &: \lambda_1(t) < \lambda_2(t) \end{aligned}$$

For this test, the log-hazard ratio is calculated

$$r(t) = \frac{\log(S_2(t))}{\log(S_1(t))}$$

The log-rank test assumes that this ratio is constant over time, this is called the proportional hazards assumption. For the sample size, we exploit the fact that the sampling distribution of  $\log(r)$  is asymptotically normal with variance

$$\sigma_{\log(r)}^2 = \frac{1}{N_1(1 - S_1(t_{max}))} + \frac{1}{N_2(1 - S_2(t_{max}))}$$

where  $N_1$  and  $N_2$  are the number of animals in the two groups, and  $t_{max}$  is the time of experimentation. The sample size can be calculated from the equation

$$z_{1-\alpha} \sqrt{\frac{1}{N_1(1 - S_1^0(t_{max}))} + \frac{1}{N_2(1 - S_2^0(t_{max}))}} = \log(r^a) - z_{1-\beta} \sqrt{\frac{1}{N_1(1 - S_1^a(t_{max}))} + \frac{1}{N_2(1 - S_2^a(t_{max}))}} \quad (4.76)$$

where  $S_i^0$  and  $S_i^a$  refer to the survival in the group  $i$  under the null and alternative hypotheses, respectively;  $r^a$  is the ratio between the survivals under the alternative hypothesis at time  $t_{max}$  and for the value for which the statistical power is specified. If we assume  $N_1 = N_2$ , there are two possible solutions that, in their turn, make different assumptions about the mathematical problem

Schoenfeld's method: Assumes  $S_i^0(t_{max}) = S_i^a(t_{max})$ , then

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\log(r^a)}{\sqrt{\frac{1}{1-S_1(t_{max})} + \frac{1}{1-S_2(t_{max})}}}} \right)^2 \quad (4.77)$$

Lachin's method: Assumes exponential survivals

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\frac{1-r^a}{1+r^a}}{\frac{1}{\sqrt{2-S_1(t_{max})-S_2(t_{max})}}}}} \right)^2 \quad (4.78)$$

- Example 89: Following with Example 88, if we do not want to explicitly assume any particular distribution of the survival time, we may use the sample size designs presented in this section. Let us assume that we expect in the control group about 10% of the animals reaching the limit of 10 seconds, and about 20% of the animals in the group with the analgesic drug. Then

$$r^a = \frac{\log(0.1)}{\log(0.2)} = 1.4307$$

**Schoenfeld's method**

$$N = \left( \frac{z_{0.95} + z_{0.9}}{\frac{\log(1.4307)}{\sqrt{\frac{1}{0.9} + \frac{1}{0.8}}}} \right)^2 = 158$$

**Lachin's method**

$$N = \left( \frac{z_{0.95} + z_{0.9}}{\frac{\frac{1-1.4307}{1+1.4307}}{\frac{1}{\sqrt{2-0.1-0.2}}}}} \right)^2 = 161$$

**Important remarks**

98. As expected, the design for the log-rank test results in a larger number of samples due to its non-parametric nature.

## 4.8 Sample size for pilot experiments

When very little is known about the statistical behavior of the variables of interest, it is recommended to look for similar experiments in the literature to learn the basic statistical description of what should be expected (for instance, the mean and variance of the control group). However, there are occasions in which we cannot find experiments similar to ours, and we need to perform a pilot experiment to get some information from a few animals. Pilot experiments are normally recommended for other purposes (logistics, check viability, ... see Chap. 1). In this section we will see how to design and use pilot studies to gain statistical information for further design. For a full discussion on the sample size calculation for pilot studies see [Sorzano et al \(2018\)](#).

### 4.8.1 Pilot experiments for the variance and mean

- **Example 90:** We are interested in a particular gene that has never been studied before, and there is no information about its expression level in a particular tissue. We are interested in designing a drug that reduces its expression level to a half of its normal expression level. How should we design the experiment and how many animals do we need? The gene expression level will be measured by RNA-seq.

Our final goal is a comparison between the mean of two groups: control and treatment, and we want to find a reduction of at least 50% in the mean. But we do not know which is the mean of the control, nor its variance, so we cannot make at this stage any sample size calculation. Instead, we will first perform a pilot study with a small number of untreated animals (say  $N = 10$ ) to know the mean and variance of the gene expression level in the control group. With the knowledge gained in the pilot experiment we will calculate the sample size required for the main experiment.

When we perform an experiment with  $N$  animals, we will be able to estimate the mean and variance of the underlying populations,  $\hat{\mu}$  and  $\hat{\sigma}^2$ , respectively. These estimates will have an associated confidence interval that are given by

$$\begin{aligned}\sigma^2 &\in \left[ \hat{\sigma}^2 \frac{N-1}{\chi_{1-\frac{\alpha}{2}, N-1}^2}, \hat{\sigma}^2 \frac{N-1}{\chi_{\frac{\alpha}{2}, N-1}^2} \right] \\ \mu &\in \left[ \hat{\mu} - t_{1-\frac{\alpha}{2}, N-1} \frac{\hat{\sigma}}{\sqrt{N}}, \hat{\mu} + t_{1-\frac{\alpha}{2}, N-1} \frac{\hat{\sigma}}{\sqrt{N}} \right]\end{aligned}\quad (4.79)$$

- **Example 90 (continued):** In RNA-seq the observed gene expression is expressed as the logarithm in base 2 of the number of reads. For the sake of illustration, let us assume that for the  $N = 10$  animals in our experiment we have the estimates

$$\begin{aligned}\hat{\sigma}^2 &= 0.25 \\ \hat{\mu} &= 10\end{aligned}$$

Then, the corresponding 99% confidence intervals are

$$\begin{aligned}\sigma^2 &\in \left[ 0.25 \frac{9}{\chi_{0.995, 9}^2}, 0.25 \frac{9}{\chi_{0.005, 9}^2} \right] = [0.10, 1.30] \\ \mu &\in \left[ 10 - t_{0.995, 9} \frac{0.5}{\sqrt{10}}, 10 + t_{0.995, 9} \frac{0.5}{\sqrt{10}} \right] = [9.49, 10.51]\end{aligned}$$

Reducing the expression level to one half, in logarithmic base 2 units is looking for an effect size of

$$\Delta = \log_2(0.5) = -1$$

We can now make two sample size designs: one for the best case (minimum variance) and another one for the worse one (maximum variance). We now need to make a design for the difference of the means of two groups (Eq. 4.11). If the drug is successful, the mean of the treated group will decrease, and most likely the variance too. But we do not know by which factor, so we adopt the

conservative approach of keeping the same variance. We use the typical values of 90% of statistical power and 95% of statistical confidence.

$$\begin{aligned} N_{best} &= 6 \\ N_{worse} &= 46 \end{aligned}$$

In the best case, we would have needed only  $N_1 = N_2 = 3$  animals per group ( $N = 6$  animals in total). In the worse case, we will need  $N_1 = N_2 = 23$  animals per group. There is an almost 8-fold difference between the number of animals in the best and worse cases. Being conservative, we should go for the worse case design and employ  $N = 46$  animals in the main experiment). To reduce the difference to only a 3-fold difference, we would need  $N = 30$  animals in the pilot experiment, whose size is no longer considered appropriate for a pilot experiment.

#### Important remarks

99. Performing a pilot experiment in the previous example only gave an approximate number of animals for the main experiment. The difference between the best and worse case is very large and we did not gain much statistical information for the next stage. For this reason, pilot experiments, only to learn statistical knowledge of the underlying distribution, are discouraged and, almost the same uncertainty is obtained from the standard literature. In the example above, we could have taken the average variance of gene expression in a large number of genes.

### 4.8.2 Pilot experiments for proportions

- Example 91: We are developing a new therapy for a disease. The treatment is absolutely new, with no prior reference and we do not know the proportion of respondents. In a pilot study, we will treat 20 animals, and estimate the proportion of respondents. Then, we will use this knowledge to design a wider study to construct a confidence with a maximum width of 10% with a 95% of confidence level. How many animals do we need for this experiment?

We saw in Eq. 4.19 the exact method and, then, approximate methods to design the sample size for constructing confidence intervals. At the end of the main experiment, the confidence interval will be of the form

$$p \in [p_L, p_U]$$

Our specification is that the maximum width needs to be smaller than 10%, that is,

$$p_U - p_L < 0.1$$

- Example 91 (continued): In the pilot experiment with  $N = 20$  animals,  $x_0 = 3$  of them responded. This gives us an estimate of the proportion of respondents

$$\hat{p} = \frac{3}{20} = 0.15$$

But its 99% confidence interval (see the section on confidence intervals for proportions around Eq. 4.19) is

$$p \in [0.018, 0.450]$$

Note that the 95% confidence level is for the confidence interval constructed in the main experiment. For the pilot experiment, we may be more restrictive. With this information we may now design the main experiment. The best case is provided by the smallest proportion, and the worse case is given by the largest proportion. The sample size is calculated according to the formula in Eq. 4.19).

$$\begin{aligned} N_{best} &= 52 \\ N_{worse} &= 398 \end{aligned}$$

Again we find an almost 8-fold difference in the sample size. If we want to reduce the difference to a 3-fold difference, the number of animals in the pilot study would need to increase numbers up to  $N = 80$ , which is no longer considered a pilot study.

## 4.9 Adaptive sample size

In all the sections of this chapter we have assumed that we design the experiment, calculate the sample size (which is fixed), and then the experiment is carried out. Once we collect all the measurements, we analyze the resulting data, and take the consequent decisions. This is certainly one of the most common procedures in science and research with animals. However, modern research has shown that we may intermix data acquisition and decision taking. This is particularly important for long or expensive experiments, like clinical trials. We may stop the experiment if we clearly see that there is no difference between the control and treatment groups. Beside the economical benefits, there is the benefit of not giving a useless treatment to a patient or to an animal.

The data analysis of this kind of experiments is more complicated than the standard hypothesis testing or the construction of standard confidence intervals. In an extremely simplified manner, we could say that there are planned hypothesis tests along the experiment. These tests are like the normal ones, but with a confidence level that is different from the one we would use at the end of the experiment (typically 95%). In the following sections we present some of the simplest examples of adaptive sample size designs. For further details on this kind of designs, the interested reader is referred to [Wassmer and Brannath \(2018\)](#). In general, these experimental designs are called *sequential designs* or methods.

As such, these methods belong to the data analysis domain. However, they have a huge impact in the effective sample size of the experiment, and that is why they have been included in this section.

### 4.9.1 Hypothesis test for the superiority of a proportion

- Example 92: We are testing a new drug and we are interested in testing if it produces a sufficiently high response. We plan to do a two-stage experiment. In the first stage we will test the new drug in a few animals, if there is enough evidence that the drug is working, then we will go on to the second stage with more animals. How many animals should we have in the first stage and how many responding if we want to go on to the second stage if  $p > p_a = 0.25$ , and we will stop the experiment if  $p < p_0 = 0.05$ ?

The theory corresponding to this two-stage design was developed in [Simon \(1989\)](#). The inference test is of the form

$$\begin{aligned} H_0 &: p \leq p_0 \\ H_a &: p > p_0 \end{aligned} \quad (4.80)$$

where  $p$  is the proportion of animals responding to the treatment, and  $p_0$  is some threshold below which it is not worthy to go on analyzing this drug. We want to have a statistical power of  $1 - \beta$  if  $p$  is larger than a value  $p_a$

First stage analysis:

Let  $N_1$  be the number of animals in the first stage. Let us denote as  $R_1$  the maximum number of animals responding in the first stage, under which we will not proceed to the second stage (if  $R_1$  or fewer animals respond, then we will not continue). Let us assume that the true proportion of responding animals is  $p_1$ . We will terminate the experiment in the first stage with a Probability of Early Termination after stage 1 (*PET*)

$$PET = \Pr\{r_1 \leq R_1\} = \sum_{r_1=0}^{R_1} b(r_1; p_a, N_1)$$

where  $b(r_1; p_a, N_1)$  is the probability of observing  $r_1$  responses with a Binomial distribution with parameters  $p_a$  and  $N_1$ . On the contrary, we will continue with the experiment if  $p = p_0$  with probability

$$PC_1 = \sum_{r_1=R_1+1}^{N_1} b(r_1; p_0, N_1)$$

We want to design  $R_1$  and  $N_1$  such that  $PET < \beta$  and  $PC_1 < \alpha$ .

Final analysis:

If we decide to go to the second stage, then we will study  $N_2$  animals more. Let us denote as  $R$  the total number of respondents including stages 1 and 2 under which we will reject the null hypothesis. The total probability (including stages 1 and 2) of not being able to reject the null hypothesis when the alternative is true with  $p = p_a$  is

$$PT = PET + \sum_{r_1=R_1+1}^{\min(N_1, R)} b(r_1; p_1, N_1) \left( \sum_{r_2=0}^{R-r_1} b(r_2; p_1, N_2) \right)$$

The total probability (including stages 1 and 2) of rejecting the null hypothesis when this is true is (for  $p = p_0$ )

$$PC = PC_1 \left( \sum_{R'=R+1}^{N_1+N_2} \sum_{r_1=0}^{\min(N_1, R')} b(r_1; p_0, N_1) b(R' - r_1; p_0, N_2) \right)$$

We want to design  $R$  and  $N_2$  such that  $PT < \beta$  and  $PC < \alpha$ .

We may minimize the average number of animals

$$\boxed{\min N_1 + (1 - PET)N_2 \quad \text{subject to} \quad PET, PT < \beta; PC_1, PC < \alpha} \quad (4.81)$$

or minimize the maximum number of animals (minimax)

$$\boxed{\min N_1 + N_2 \quad \text{subject to} \quad PET, PT < \beta; PC_1, PC < \alpha} \quad (4.82)$$

The solution of these problems is given by enumeration (we explore all possible  $R_1, N_1, R$ , and  $N_2$  satisfying the equations above for specific  $p_0, p_a, \alpha$  and  $\beta$ ). The following table shows the resulting thresholds  $R_1$  and  $R$  and sample sizes  $N_1$  and  $N$  for different values of  $p_0$  and  $p_a$ . The table has been calculated for  $\alpha = 0.05$  and  $\beta = 0.1$ .

		Min. Average				Minimax			
$p_0$	$p_a$	$R_1$	$N_1$	$R$	$N = N_1 + N_2$	$R_1$	$N_1$	$R$	$N = N_1 + N_2$
0.05	0.25	0	9	3	30	0	15	3	25
0.10	0.30	2	18	6	35	2	22	6	33
0.20	0.40	4	19	15	54	5	24	13	45
0.30	0.50	8	24	24	63	7	24	21	53
0.40	0.60	11	25	32	66	12	29	27	54
0.50	0.70	13	24	36	61	14	27	32	53
0.60	0.80	12	19	37	53	15	26	32	45
0.70	0.90	11	15	29	36	13	18	26	32

The minimax design may be preferred if the animal accrual rate is low. If animal accrual is not a problem, then the minimum average design should be favored.

- **Example 92 (continued):** Looking at the table above in the column of minimum average, in the first stage we will use  $N_1 = 9$  animals. If none of them,  $R_1 = 0$ , respond to the drug, we will stop the experiment because the drug seems to be useless. If one or more respond, then we will go to the second stage. In the second stage we will use  $N_2 = 21$  animals ( $N_1 + N_2 = 30$ ). If the total number of respondents including the two stages is less or equal  $R = 3$ , then we cannot reject the null hypothesis ( $H_0 : p < 0.05$ ).

If the null hypothesis is true, for  $p = p_0 = 0.05$ , the probability of early termination in the first stage is 63% (that is, 63% of the experiments for which  $p = 0.05$  are terminated after only 9 animals), and the average number of animals is 16.8. The same quantities for the minimax design are 46% and 20.4. In this way, we see a clear advantage of the minimum average design over the minimax design.

We can extend this idea to multiple stages instead of just two. At each stage, we decide whether to go on with the experiment, or whether to stop it (Fleming, 1982). In Eq. 4.22 we presented the sample size required for a test like the one introduced at the beginning of this section (Eq. 4.80). This sample size assumes that the test is performed only once after collecting the information from the  $N$  animals. Dividing the experiment into multiple stages allows stopping the experiment not only because of low response (evidence that the true  $p$  is smaller than  $p_0$ ), but also because of high response (evidence that the true  $p$  is larger than  $p_a$ ).

If we divide the experiment into  $K$  stages with  $N_1, N_2, \dots, N_K$  animals in each stage, then let us denote the number of respondents up to stage  $k$  as  $R_k$ . We will also stop the experiment because of low response if  $R_k \leq R_k^0$ . In this case, we cannot reject  $H_0$ . We will stop the experiment because of high response if  $R_k \geq R_k^a$ . These limits can be calculated by

$$\begin{aligned} R_k^0 &= \left\lceil \sum_{i=1}^k N_i p_0 + z_{1-\alpha} \left( \sum_{i=1}^k N_i p_0 (1-p_0) \right)^{1/2} \right\rceil + 1 \\ \tilde{p}_0 &= \frac{((N p_0)^{1/2} + (1-p_0)^{1/2} z_{1-\alpha})^2}{N + z_{1-\alpha}^2} \\ R_k^a &= \left\lceil \sum_{i=1}^k N_i \tilde{p}_0 - z_{1-\alpha} \left( \sum_{i=1}^k N_i \tilde{p}_0 (1-\tilde{p}_0) \right)^{1/2} \right\rceil + 1 \end{aligned}$$

where  $\lceil x \rceil$  is the round up of  $x$ . The following table shows some designs with three stages for  $\alpha = 0.05$  and  $\beta = 0.1$ . Some more designs are shown in Fleming (1982).

$p_0$	$p_a$	$R_1^0$	$R_1^a$	$N_1$	$R_2^0$	$R_2^a$	$N_1 + N_2$	$R_3^0$	$R_3^a$	$N_1 + N_2 + N_3$
0.05	0.20	-1	4	15	2	5	30	4	5	40
0.10	0.30	0	5	15	3	6	25	6	7	35
0.20	0.40	2	10	20	9	13	35	15	16	50
0.30	0.50	5	12	20	12	17	35	20	21	50

- **Example 92 (continued):** Following with the same example, with  $p_0 = 0.05$  and  $p_a = 0.20$ , in the first stage we would test the drug on  $N_1 = 15$  animals. If there are more than  $R_1^a = 4$  responses, we would reject the null hypothesis and stop the experiment because there is already enough evidence of the effectivity of the drug. If not, we continue to Stage 2, with  $N_2 = 15$  more animals. If between Stages 1 and 2 there has been  $R_2^0 = 2$  responses or less, we would stop the experiment and we cannot reject the null hypothesis (the drug is useless). If there has been  $R_2^a = 5$  responses or more, we would stop the experiment and reject the null hypothesis (declare the drug effective). If the number of responses is 3 or 4 (between the two stages), we would continue to Stage 3, with  $N_3 = 10$  more animals. If there are in total  $R_3^0 = 4$  responses or less, we cannot reject the null hypothesis. If there are  $R_3^a = 5$  responses or more, we would reject the null hypothesis.

**Important remarks**

100. Inner wedge tests are extremely efficient to reduce the sample size. From the beginning of the experiment we can stop it if there is enough evidence of the effectivity of the treatment (this feature is shared with all sequential designs). From the middle of the experiment, approximately, we can also stop it if we consider that we will not be able to reject the null hypothesis.

**4.9.2 Hypothesis test on the difference of the mean of two samples**

- Example 93: We are developing a new drug for lowering LDL cholesterol concentration in blood. We would like to detect a decrease of 10 mg/dL in the cholesterol level between the new drug and the reference one. The standard deviation of the population is 20 mg/dL. We would like to divide the experiment in 5 sequential experiments (each one uses all the data available) and perform 4 interim tests. If at any of the tests there is much evidence in favour of the new drug, we stop.

We start doing the design as if there were a single final test at the end of the experiment. For that we need,  $N = 70$  samples per group (see Section 4.1.5). That is, 70 individuals per group. We would now divide the whole experimental lot into  $K = 5$  subgroups. After testing each subgroup we perform an interim hypothesis test. If there is enough evidence for the difference between the two drugs we stop the experiment.

O'Brien and Fleming (1979) developed the theory for carrying out these interim hypothesis tests without compromising the statistical confidence level and power. The interim tests ( $k = 1, 2, \dots, K$ ) are based on the statistic

$$Z_k = \frac{\sum_{j=1}^{N_k} x_{1j} - \sum_{j=1}^{N_k} x_{2j}}{\sqrt{N_k(\sigma_1^2 + \sigma_2^2)}}$$

where  $x_{ij}$  is the value of the  $j$ -th observation in treatment  $i$ , and  $N_k$  is the accumulated number of observations. Under the null hypothesis  $Z_k$  is distributed as

$$Z_k \sim N\left((\mu_1 - \mu_2)\sqrt{\frac{N_k}{\sigma_1^2 + \sigma_2^2}}, 1\right)$$

The procedure is as follows:

- At the interim tests ( $k = 1, 2, \dots, K - 1$ )
  - If  $|Z_k| > C_B(k, \alpha)\sqrt{K/k}$ , then stop and reject  $H_0$
  - Otherwise, continue to subgroup  $k + 1$
- At the final test ( $k = K$ )

- If  $|Z_K| > C_B(K, \alpha)$ , then reject  $H_0$
- Otherwise, we cannot reject  $H_0$

$C_B(K, \alpha)$  is a constant that depends on  $K$  and  $\alpha$  and that takes into account the Type I error inflation that occurs due to the multiple testing and the use of accumulated data. The following table gives the value of  $C_B(k, \alpha)$  for two values of  $\alpha$ .

$k$	$C_B(k, \alpha = 0.01)$	$C_B(k, \alpha = 0.05)$
1	2.576	1.960
2	2.580	1.977
3	2.595	2.004
4	2.609	2.024
5	2.621	2.040
6	2.631	2.053
7	2.640	2.063
8	2.648	2.072
9	2.654	2.080
10	2.660	2.087

The number of samples per group must also be modified to account for the multiple tests.

$$N_{\text{sequential}} = NR_B(K, \alpha, \beta) \quad (4.83)$$

where  $R_B(K, \alpha, \beta)$  is a number given in the following table (for other values, consult [Chow et al \(2008\)](#)[Chap. 8])

$k$	$R_B(k, \alpha = 0.01, \beta = 0.1)$	$R_B(k, \alpha = 0.05, \beta = 0.1)$
1	1.000	1.000
2	1.001	1.007
3	1.006	1.016
4	1.010	1.022
5	1.014	1.026
6	1.016	1.030
7	1.018	1.032
8	1.020	1.034
9	1.021	1.036
10	1.022	1.037

- Example 93 (continued): We will perform  $K = 5$  sequential tests. For this reason, we need to increase a bit the sample size to account for the multiple testing

$$N_{\text{sequential}} = 1.026 \cdot 70 = 72$$

We will have  $72/5 = 15$  animals per stage. In the first stages, we will be more stringent, and reject the null hypothesis only if there is a large evidence against it (see Fig. 4.13). At the end,  $K = 5$ , our rejection threshold is also larger, 2.04, than it would normally be for a single test,  $z_{0.975} = 1.96$ .

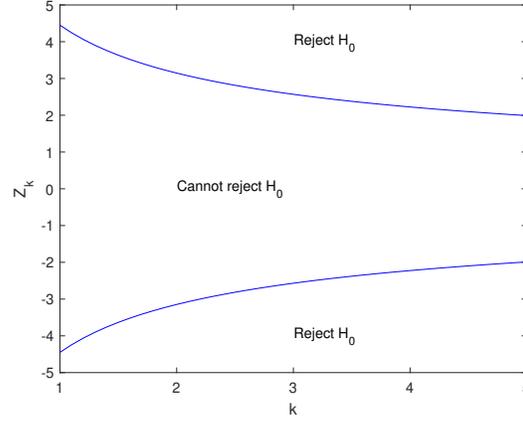


Figure 4.13: Rejection regions for the variable  $Z_k$  at the different interim tests.

- **Example 94:** While performing the experiment, it would also be helpful to stop the trial if we realize that by adding more animals we will not, probably, be able to prove that the new drug is more effective. In this way, we save costs and we avoid animals to be unnecessarily treated. How to proceed and how many animals do we need?

A possible solution to this problem is given by the inner-wedge test, which is also based on the  $Z_k$  statistic defined above. The test is parametrized by a constant  $\Delta$  that has to be chosen (normally between -0.5 and 0.5, for  $\Delta = 0$  this test is related to O'Brien and Fleming's test, and for  $\Delta = 0.5$  with Pocock's test). Let us define the constants

$$a_k = (C_{W1}(K, \alpha, \beta, \Delta) + C_{W2}(K, \alpha, \beta, \Delta))\sqrt{\frac{k}{K}} - C_{W2}(K, \alpha, \beta, \Delta) \left(\frac{k}{K}\right)^{\Delta-1/2}$$

$$b_k = C_{W1}(K, \alpha, \beta, \Delta) \left(\frac{k}{K}\right)^{\Delta-1/2}$$

The data analysis procedure is as follows:

- At the interim tests ( $k = 1, 2, \dots, K - 1$ )
  - If  $|Z_k| \geq b_k$ , then stop and reject  $H_0$
  - If  $|Z_k| < a_k$ , then stop and  $H_0$  cannot be rejected
  - Otherwise, continue to subgroup  $k + 1$
- At the final test ( $k = K$ )
  - If  $|Z_K| \geq b_k$ , then reject  $H_0$
  - Otherwise, we cannot reject  $H_0$

As in the previous case we need to increase the total sample size to

$$N_{sequential} = NR_W(K, \alpha, \beta, \Delta) \tag{4.84}$$

The following table gives the values of the constants  $C_{W1}$ ,  $C_{W2}$  and  $R_W$ , for  $\alpha = 0.05$  and  $\beta = 0.1$

$\Delta$	$k$	$C_{W1}(k, \alpha, \beta, \Delta)$	$C_{W2}(k, \alpha, \beta, \Delta)$	$R_W(k, \alpha, \beta, \Delta)$
-0.50	1	1.960	1.282	1.000
	2	1.960	1.282	1.000
	3	1.952	1.305	1.010
	4	1.952	1.316	1.016
	5	1.952	1.326	1.023
	10	1.958	1.351	1.042
-0.25	1	1.960	1.282	1.000
	2	1.957	1.294	1.006
	3	1.954	1.325	1.023
	4	1.958	1.337	1.033
	5	1.960	1.351	1.043
	10	1.975	1.379	1.071
0.00	1	1.960	1.282	1.000
	2	1.958	1.336	1.032
	3	1.971	1.353	1.051
	4	1.979	1.381	1.075
	5	1.990	1.385	1.084
	10	2.013	1.428	1.127
0.25	1	1.960	1.282	1.000
	2	2.003	1.398	1.100
	3	2.037	1.422	1.139
	4	2.058	1.443	1.167
	5	2.073	1.477	1.199
	10	2.119	1.521	1.261

- Example 94 (continued): We will perform  $K = 5$  sequential tests, and to be able to compare with the previous case, we will choose  $\Delta = 0$ . Consequently, we need to increase the sample size to

$$N_{\text{sequential}} = 1.084 \cdot 70 = 76$$

We will have  $76/5 = 16$  animals per stage. As in the O'Brien and Fleming test, in the first stages, we will be more stringent. However, now we will be able to stop the experiment earlier if  $Z_k$  does not approach the rejection limit (see Fig. 4.14).

### 4.9.3 Hypothesis test on the difference of two proportions

- Example 95: We say that an animal responds to a LDL cholesterol drug if its cholesterol level drops more than 30 mg/dL. We know that 30% of the animals respond to the reference drug. We wonder if at least 40% animals respond to a new drug. A single stage experiment would require  $N = 386$  animals (see Sec.

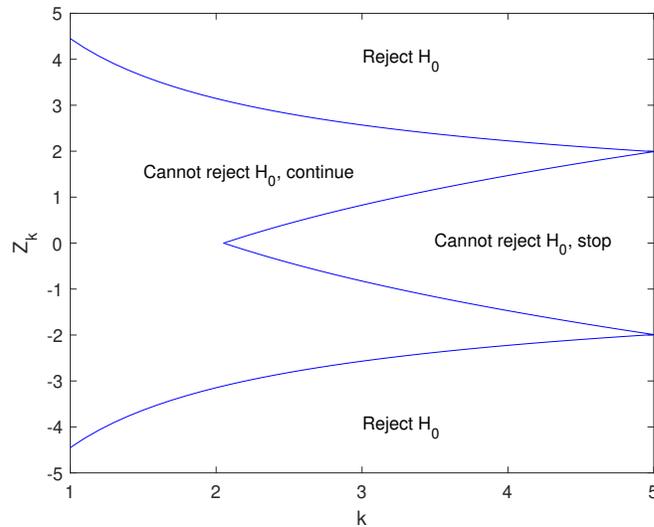


Figure 4.14: Rejection regions for the variable  $Z_k$  at the different interim tests. If  $Z_k$  is in the middle wedge, the experiment is stopped because it does not seem to be very promising.

4.2.5). How many animals do we need if we plan to perform  $K = 10$  subgroups and make interim hypothesis tests between stages?

We now define the variable

$$Z_k = \frac{\sqrt{N_k}(\hat{p}_{1,k} - \hat{p}_{2,k})}{\sqrt{\hat{p}_{1,k}(1 - \hat{p}_{1,k}) + \hat{p}_{2,k}(1 - \hat{p}_{2,k})}}$$

where  $N_k$  is the number of accumulated samples up to the  $k$ -th interim test and  $\hat{p}_{i,k}$  is the estimate of the proportion of the  $i$ -th population (1 or 2) at the  $k$ -th interim test. This variable can be directly used in O'Brien and Fleming's test or in the Inner Wedge test of the previous section.

- **Example 95 (continued):** The smallest increase in sample size is obtained in the Inner Wedge test with  $\Delta = -0.5$ . We will increase, then, the number of animals to

$$N_{\text{sequential}} = 1.042 \cdot 386 = 403$$

We will have  $403/10 = 41$  animals per stage. This experiment is particularly large,  $N = 410$  animals, but interestingly from the 5<sup>th</sup> group (typically, the stopping region of the Inner Wedge is active when half the experiment has been performed) we will be able to stop the experiment if we see that the response to the new drug is not significantly larger than to the reference once. Additionally, we may reject the null hypothesis from the first interim test if the response is sufficiently strong.

#### 4.9.4 Sample size reestimation in blind experiments

- Example 96: In Example 89 we were designing an experiment to see the effectiveness of an analgesic. Animals will be placed in a hot plate and we will record the time they take to jump out of it. We will compare a control to a treatment group using a log-rank test. We calculated the sample size to be  $N = 158$ . If we divide the experiment in 10 subgroups, can we use a sequential design with interim tests between stages?

Let  $d_k$  be the total number of failures in both groups (1 and 2) up to the  $k$ -th test. Let us index the failing subjects (in both groups) as  $i = 1, 2, \dots, d_k$ . Let  $t_i$  be the failing time of the  $i$ -th individual, and  $r_{i,1}$  and  $r_{i,2}$  the remaining number of samples in groups 1 and 2 at time  $t_i$ , respectively. Let  $\delta_{i,2}$  be 1 if the  $i$ -th individual is from Group 2 and 0 if it is from Group 1. The log-rank statistic at the  $k$ -th interim analysis can be written as

$$S_k = \sum_{i=1}^{d_k} \left( \delta_{i,2} - \frac{r_{i,2}}{r_{i,1} + r_{i,2}} \right)$$

The observed information variable is defined as

$$I_k = \sum_{i=1}^{d_k} \frac{r_{i,1}r_{i,2}}{(r_{i,1} + r_{i,2})^2}$$

The variable

$$Z_k = \frac{S_k}{I_k}$$

can be used in the O'Brien and Fleming's test or the Inner Wedge test.

- Example 96 (continued): We will use an Inner Wedge test with  $\Delta = -0.5$ . We will increase, then, the number of animals to

$$N_{\text{sequential}} = 1.042 \cdot 158 = 165$$

We will have  $165/10 = 17$  animals per stage. Now we can use the standard procedure for the Inner Wedge test.

#### 4.9.5 Hypothesis test on the difference of two survival curves

- Example 97: We are evaluating the efficacy of a new drug with respect to a reference drug. We have estimated the sample size to be  $N = 18$  per treatment and we are performing an interim test ( $K = 2$ ). We record whether an animal responds or not to the treatment or control. The study is being performed in a double blind way, and disclosing at any moment the labels would compromise the efficacy of the study. How can we, at the interim test, use the collected data to readjust the sample size with the knowledge collected?

This is a problem known as “Sample size re-estimation”. The solution depends on each kind of study being performed. Here the solution for proportions is presented (Chow et al, 2008)[Chap. 8].

Let denote  $p_1$  and  $p_2$  the actual proportion of responders in group 1 (treatment) and 2 (control). Let  $y_j$  denote the response ( $y_j = 1$ ) or not ( $y_j = 0$ ) of the  $j$ -th individual (note that we do not know if it is receiving the treatment or control drug). We assign it randomly to Stratum A with probability  $\pi$  (with  $\pi \in (0, 0.5)$ ) and to Stratum B with probability  $1 - \pi$ . We now calculate

$$p_A = \Pr\{y_j = 1 | j \in \text{Stratum A}\} = \pi p_1 + (1 - \pi)p_2 \approx \frac{1}{N_A} \sum_{j \in \text{Stratum A}} y_j = \hat{p}_A$$

$$p_B = \Pr\{y_j = 1 | j \in \text{Stratum B}\} = (1 - \pi)p_1 + \pi p_2 \approx \frac{1}{N_B} \sum_{j \in \text{Stratum B}} y_j = \hat{p}_B$$

We can solve for  $p_1$  and  $p_2$  (or its estimates) as

$$\hat{p}_1 = \frac{\pi \hat{p}_A - (1 - \pi) \hat{p}_B}{2\pi - 1}$$

$$\hat{p}_2 = \frac{\pi \hat{p}_B - (1 - \pi) \hat{p}_A}{2\pi - 1}$$

Finally, we re-estimate the sample size as explained in Sec. 4.2.5 and reproduced here for convenience

$$N = \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}}} \right)^2$$

- **Example 97 (continued):** Assume that we use  $\pi = 0.4$  and we observe  $\hat{p}_A = 0.6$  and  $\hat{p}_B = 0.5$ . Then,

$$\left. \begin{array}{l} 0.4p_1 + 0.6p_2 = 0.6 \\ 0.6p_1 + 0.4p_2 = 0.5 \end{array} \right\} \Rightarrow \hat{p}_1 = 0.3, \hat{p}_2 = 0.8$$

The re-estimated sample size is

$$N = \left( \frac{z_{0.95} + z_{0.9}}{\frac{0.3 - 0.8}{\sqrt{0.3 \cdot 0.7 + 0.8 \cdot 0.2}}} \right)^2 = 13$$

So that we can shorten the study from  $N = 18$  to  $N = 13$  samples per group.

#### Important remarks

101. Sample size reestimation allows shortening our experiment as evidence accumulates against the null hypothesis. In the same way, we can prolong the experiment if we see that our first guess of the variability of the samples fell short and that we need a larger sample size.



## Chapter 5

# Design of experiments

The previous section told us how many animals we need to achieve our research goals. If we do not have this number of animals, the experiment will simply fail to show what we wanted to show just by lack of statistical power. This section tells us how to distribute these animals in manageable groups so that we avoid bias and perform our experiment in statistically optimal conditions. Failing to design the experiment may compromise our results (for instance, through external biases). However, of the two statistical issues involved in the design (sample size and experiment design), sample size is much more important, and that is why researchers are primarily concerned with it, and are much less familiar with the experiment design. But having a good design will allow us to have a wider applicability of our results (because we have made experiments in more conditions) and better estimates of the treatment effects (because by carefully balancing, or at least designing, the experiment we will reduce the variability of the comparisons of interest).

As with the sample size calculation, experiment design needs to know how the data will be analyzed once the experiment is performed. A prominent role in this analysis is played by linear models, which were already introduced in Sec. 1.3, when we discussed about blocking. In this chapter we will introduce the concepts associated to linear models as we need them, so that the theory associated to them is not concentrated in one single section, but spread along the chapter. Notwithstanding, the Appendix of this chapter contains some of the most important mathematical results related to experimental design. The interested reader may start with this more mathematical summary, although practitioners may skip it. We will provide multiple examples so that researchers can identify their problems in the examples given.

### 5.1 Basic designs

In this section we will briefly revise the main concepts associated to the design of experiments by presenting the most basic designs.

### 5.1.1 Completely randomized design, CRD

This is the most basic design and most likely the most widely used for its simplicity. It allows the simultaneous comparison of multiple groups, and decomposing the observed data into different components that can be verified whether or not they have a statistically significant impact on the observations.

- **Example 98:** We are testing a new drug against cholesterol levels in blood. Control animals have a concentration of 250 mg/dL with a standard deviation of 30 mg/dL. We will test two doses of our drug ( $D_1$  and  $D_2$ ). We will refer to the control animals as  $D_0$ , and they only receive the vehicle of the drug (not the active compound). We will analyze 10 animals per group.

**Design summary.** In this kind of design we have multiple groups, each one receiving a different treatment. Animals are randomly assigned to each one of the treatments. Typically, the same number of animals are analyzed in each group, but this is not a strong requirement of the design.

As discussed in Section 1.3, it is important that the experiment is performed in a randomized manner (not all controls are studied first, then treatment 1, then treatment 2, ...).

If the treatment is defined by a genetic characteristic (like wild type vs. knock-out animals), animals cannot be randomly assigned to the treatment group, but the design is still considered to be completely randomized.

This design is recommended for relatively small experiments where the total number of individuals is sufficiently small so that they all “fit” in a setting of homogeneous conditions (same laboratory, same experimenter, same day, same batch of chemicals, ...) If the experiment span for multiple days, multiple centers, multiple technicians, ... the day, center, technician, etc. may cause a difference in the results and it is better to perform a design with blocks (see Sec. 5.1.3).

Once we perform the experiment we may use ANOVA (Analysis of Variance) to determine if the drug was effective. For doing so, we may distribute the data as shown in the following table. We have three columns corresponding to the three different treatments (control and two drug doses). Within each column we have 10 observations of cholesterol measurements, one from each of the animals in the group. We may refer to an individual measurement as  $y_{ij}$  meaning that it has received the  $i$ -th treatment, and it is the  $j$ -th animal within the group. ANOVA is not restricted to the same number of animals per group, but it is a common practice in research laboratories and, for simplicity, we will illustrate the technique with the same number of animals per group.

$y_{0,1}$	$y_{1,1}$	$y_{2,1}$
$y_{0,2}$	$y_{1,2}$	$y_{2,2}$
$y_{0,3}$	$y_{1,3}$	$y_{2,3}$
$y_{0,4}$	$y_{1,4}$	$y_{2,4}$
$y_{0,5}$	$y_{1,5}$	$y_{2,5}$
$y_{0,6}$	$y_{1,6}$	$y_{2,6}$
$y_{0,7}$	$y_{1,7}$	$y_{2,7}$
$y_{0,8}$	$y_{1,8}$	$y_{2,8}$
$y_{0,9}$	$y_{1,9}$	$y_{2,9}$
$y_{0,10}$	$y_{1,10}$	$y_{2,10}$

ANOVA uses a linear model that decomposes the observed data as

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (5.1)$$

This equation means that the observation of the  $j$ -th animal in the  $i$ -th group can be explained as an overall mean, plus something that depends on the treatment received, plus something that cannot be explained and it is specific to each individual (this latter term is called the residual). Fig. 5.1 shows a possible result of the experiment. Each data point is a measurement from one of the animals. We have three groups (control, and the two doses). Each of the points is a  $y_{ij}$ . The figure shows the overall mean (in black) and the average of each one of the groups. These averages are estimated from the data. For instance, the overall mean can be estimated as

$$y_{..} = \frac{1}{N} \sum_{ij} y_{ij}$$

$N$  is the total number of individuals in the three groups (in our example,  $N = 30$ ). The dot notation implies that we are averaging over the corresponding index. Since we are averaging in  $i$  and  $j$  we use two dots for the overall mean. Similarly, we can compute the average of each one of the groups (represented in red in Fig. 5.1)

$$y_{i.} = \frac{1}{N_i} \sum_j y_{ij}$$

being  $N_i$  the number of individuals in the  $i$ -th group. Note that according to the linear model in Eq. 5.1 we should have the following expected values

$$\begin{aligned} \mathbb{E}\{y_{..}\} &= \mu \\ \mathbb{E}\{y_{i.}\} &= \mu + \alpha_i \end{aligned}$$

$\alpha_i$  is the difference between the overall mean and the mean of each group, and it is the part attributed to the treatment. It can be estimated as

$$\hat{\alpha}_i = y_{i.} - y_{..} \quad (5.2)$$

In the following we will refer to estimates of the different parameters with a hat (as in  $\hat{\alpha}_i$ ).  $\alpha_i$  is called the *main effect* of the treatment (we have one main effect for every

level of the treatment, control,  $D_1$ , and  $D_2$ ). For instance, for the dataset in Fig. 5.1, we have

$$\begin{aligned} y_{..} &= 205.06 \\ y_0 &= 246.43 = 205.06 + \hat{\alpha}_0 \Rightarrow \hat{\alpha}_0 = 246.43 - 205.06 = 41.37 \\ y_1 &= 201.08 = 205.06 + \hat{\alpha}_1 \Rightarrow \hat{\alpha}_1 = 201.08 - 205.06 = -3.98 \\ y_2 &= 167.67 = 205.06 + \hat{\alpha}_2 \Rightarrow \hat{\alpha}_2 = 167.67 - 205.06 = -37.39 \end{aligned}$$

Note that, by construction, the sum of all treatments (and their estimates) add up to 0

$$\alpha_0 + \alpha_1 + \alpha_2 = 0$$

This is a constraint of the ANOVA model, because there can be infinite decompositions of the observed data compatible with the linear model ( $y = \mu + \alpha_i + \varepsilon_{ij}$ ). For instance, we could have explained each group using  $\mu = 246.43$ , and we would have obtained

$$\begin{aligned} y_0 &= 246.43 = 246.43 + \hat{\alpha}_0 \Rightarrow \hat{\alpha}_0 = 246.43 - 246.43 = 0 \\ y_1 &= 201.08 = 246.43 + \hat{\alpha}_1 \Rightarrow \hat{\alpha}_1 = 201.08 - 246.43 = -45.35 \\ y_2 &= 167.67 = 246.43 + \hat{\alpha}_2 \Rightarrow \hat{\alpha}_2 = 167.67 - 246.43 = -78.76 \end{aligned}$$

This decomposition would look more “human”: if we do not apply any treatment (control), then we get an average of 246.43; with a small dose,  $D_1$ , we get a reduction of 45.35; and with a larger dose,  $D_2$ , we get a larger reduction of 78.76. However, this is not the way that ANOVA makes the linear decomposition.  $\mu$  has to be the overall mean (and not the mean of one of the groups, as in the “human” decomposition). A consequence of this choice is that the addition of all the effects is zero.

Then, we can explain any of the observations as a function of the overall mean, the treatment received and the remaining residual. For instance, for a particular individual we have observed  $y_{01} = 189.67$ . We may explain this observation as

$$\begin{aligned} y_{01} &= \hat{\mu} + \hat{\alpha}_0 + \hat{\varepsilon}_{01} \\ 249.42 &= 205.06 + 41.37 + 2.99 \end{aligned}$$

That is, the observation 249.42, out of which 205.06 is explained by the overall mean, then 41.37 is explained by being in the control group, and the remaining 2.99 can only be explained as a specificity of that measure (including the subject, measurement errors, etc.). Residuals can be positive or negative, actually their mean is zero.

ANOVA is a technique that tries to determine if there are statistically significant differences among the treatments. For instance, we would expect that there are significant differences between the groups in Fig. 5.1 and there are not between the groups in Fig. 5.2. The reason is that, although it seems that there is a small reduction of the cholesterol with the dose, we cannot guarantee that these differences are not caused by random sampling (measurement noise, and differences among individuals and the particular groups we have sampled). Visually, we expect that the differences in Fig. 5.1 are real (they are caused by the treatment), while we are not so sure about the truth of the differences in Fig. 5.2.

ANOVA puts a number on this certainty, it is the p-value of the hypothesis test

$$\begin{aligned} H_0 &: \alpha_0 = \alpha_1 = \dots = \alpha_p \\ H_a &: \exists i, j | \alpha_i \neq \alpha_j \end{aligned}$$

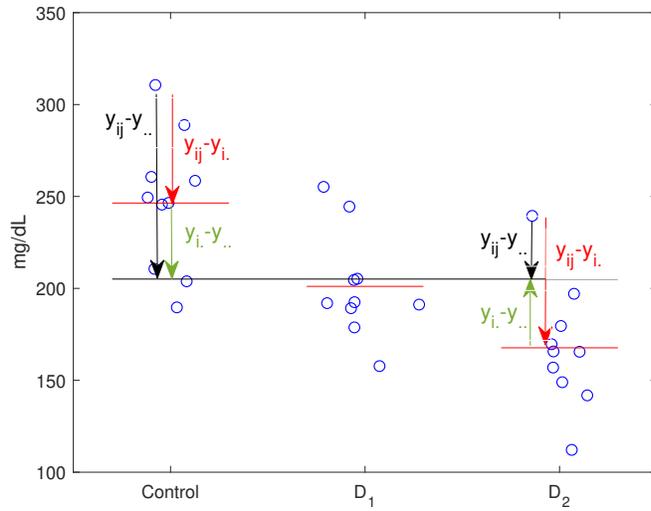


Figure 5.1: Example of data analysis by ANOVA. There are three groups (control and two doses) with 10 observations each. The horizontal black line is the overall mean. The horizontal red lines are the mean of each of the groups. Each observation is noted as  $y_{ij}$  meaning that it is the  $j$ -th observation in the  $i$ -th group. The mean of the  $i$ -th group is noted as  $y_i$ , and the overall mean as  $y_{..}$ .

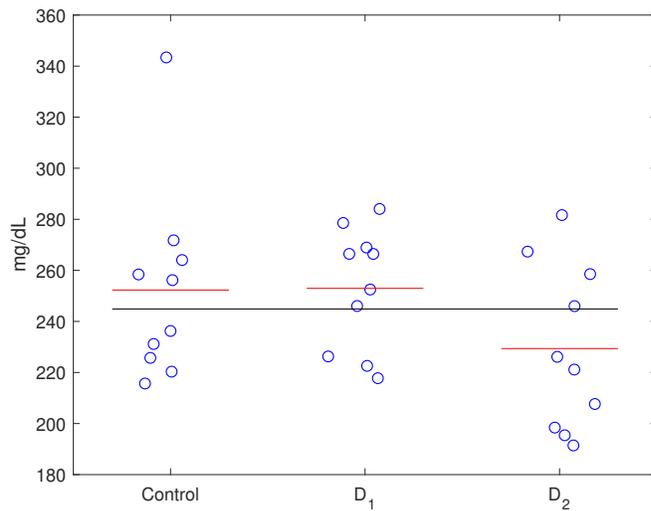


Figure 5.2: Example of data analysis by ANOVA.

That is, the null hypothesis is that there is no difference in the main effects of the different groups. The alternative hypothesis is the opposite, there are at least two treatments,  $i$  and  $j$ , whose main effects are different. If the p-value is smaller than a given threshold we would reject the null hypothesis (observing this data if there is no difference in the treatments would be so unlikely that we reject this hypothesis). If the p-value is not smaller than the threshold, we cannot reject the hypothesis that there is no difference between the treatments and that the observed differences are just caused by measurement errors and variability within the observed population.

ANOVA calculates this p-value by analyzing the distances between the observations  $y_{ij}$ , their group means  $y_{i.}$ , and the overall mean  $y_{..}$ . In particular, it exploits a property of the sum of squares

$$\sum_{ij} (y_{ij} - y_{..})^2 = \sum_{ij} (y_{i.} - y_{..})^2 + \sum_{ij} (y_{ij} - (\hat{\mu} + \hat{\alpha}_i))^2$$

This expression means that the sum of the squares of the distance from each observation to the overall mean (black arrow in Fig. 5.1) can be decomposed as the sum of the squares of the distance from its group mean to the overall mean (green arrow in the same figure) plus the sum of the squares of the distance from each observation to its group mean (red arrow). We have calculated the residual as

$$\hat{\epsilon}_{ij} = y_{ij} - (\hat{\mu} + \hat{\alpha}_i)$$

that is, the sample  $y_{ij}$  is predicted to have a value of  $\hat{\mu} + \hat{\alpha}_i$  for being in the  $i$ -th group. The difference between this prediction and the actual observation is the residual for that individual.

Note that the total sum of squares (left hand side of the equation) is related to the total variance of the observations (how different are the observations from the overall mean), and it will be noted as  $SS_T$  (total sum of squares). The equation above means that each observation contributes to the total sum of squares with a term that depends on the group it belongs to (how different is each group mean from the overall mean), plus a term that depends on the residual (how different is this observation from its group mean). We can rewrite the decomposition above more clearly as

$$\begin{aligned} \sum_{ij} (y_{ij} - y_{..})^2 &= \sum_{ij} \hat{\alpha}_i^2 + \sum_{ij} \hat{\epsilon}_{ij}^2 \\ SS_T &= SS_\alpha + SS_\epsilon \end{aligned} \quad (5.3)$$

Each experiment has an associated number of degrees of freedom. Loosely speaking, degrees of freedom are like “tokens of information”. Every new observations gives us a new token. As we accumulate observations, we are also accumulating tokens of information, evidence. If we have an experiment with  $N$  animals, then we originally have  $N$  tokens. Now, we spend these tokens in estimating different parameters. Every parameter costs a token. We need to estimate the overall mean, then this costs a token, so that the number of degrees of freedom associated to the total sum of squares,  $SS_T$ , is no longer  $N$  but  $N - 1$ . We need to estimate the three main effects as we pointed out in Eq. 5.2. Then, we would need three tokens for this. But we can play a trick; we can estimate only two of them and remember that they have to add up to zero

$$\sum_i \alpha_i = 0$$

so that given two of the main effects, we can automatically calculate the other and save a token. In general if we have  $T$  treatment groups, we only need  $T - 1$  degrees of freedom to calculate the main effects of all the groups. We had  $N - 1$  degrees of freedom to calculate parameters, we have consumed  $T - 1$ , and the remaining  $N - T$  are left for the residuals. In this way, we have a decomposition of the degrees of freedom similar to the decomposition of the sum of squares

$$\begin{aligned} N - 1 &= (T - 1) + (N - T) \\ df_T &= df_\alpha + df_\epsilon \end{aligned} \quad (5.4)$$

We may now wonder how much each degree of freedom has “bought” in terms of sum of squares. Regarding the main effects of the treatments, we have spent  $T - 1$  degrees of freedom in explaining  $SS_\alpha$ . Each degree of freedom is explaining a *mean squares* of

$$MS_\alpha = \frac{SS_\alpha}{T - 1}$$

Similarly, each of the degrees of freedom of the residuals is explaining an amount of variance given by

$$MS_\epsilon = \frac{SS_\epsilon}{N - T}$$

We can summarize all this information in the following table, which analyzes the amount of variance explained by each source of information:

Source	Sum of squares	Degrees of freedom	Mean squares ( $MS = SS/df$ )
Treatments	$SS_\alpha = \sum_{ij} \hat{\alpha}_i^2$	$T - 1$	$MS_\alpha = SS_\alpha / (T - 1)$
Residuals	$SS_\epsilon = \sum_{ij} \hat{\epsilon}_{ij}^2$	$N - T$	$MS_\epsilon = SS_\epsilon / (N - T)$
Total	$SS_T = \sum_{ij} (y_{ij} - y_{..})^2$	$N - 1$	

ANOVA checks if the mean contribution of each of the degrees of freedom of the treatments “pays” for its calculation, that is, it is significantly different from the contribution of each of the degrees of freedom of the residuals

$$f = \frac{MS_\alpha}{MS_\epsilon}$$

If the null hypothesis is true (there is no difference between the group means), then the statistic  $F$  is distributed as an Snedecor’s  $F$  with  $T - 1$  and  $N - T$  degrees of freedom. The p-value is the probability of observing under the null hypothesis an  $F$  value at least as extreme as the one we have observed

$$\text{p-value} = \Pr\{F_{T-1, N-T} > f\}$$

We reject the null hypothesis if this p-value is below a given threshold (typically, 0.05).

Additionally, the sum of squares decomposition allows the definition of the *coefficient of determination*, normally denoted as  $R^2$ , that is the proportion of the total variance that is explained by the predictions. It is defined as

$$R^2 = 1 - \frac{SS_\epsilon}{SS_T} \quad (5.5)$$

This value goes from 0 to 1.  $R^2 = 1$  implies a perfect prediction since all residuals would be zero.

- Example 98 (continued): The ANOVA table for the data shown in Fig. 5.1 would be

Source	SS	df	MS
Treatments	31252	2	15626
Residuals	30600	27	1133
Total	61852	29	

The associated  $f$  would be  $f = \frac{15626}{1133} = 13.79$  and its corresponding p-value  $p = 6.18 \cdot 10^{-5}$ . Consequently, we would reject the null hypothesis and accept that at least two treatments are different to each other. *Post-hoc* analysis would now look for the pair or pairs of treatments that are different from each other. Showing this second part of the analysis is out of the scope of this chapter since it would divert us from our main objective, design of experiments. The interested reader is referred to [Doncaster and Davey \(2007\)](#). The  $R^2$  of this ANOVA model is  $R^2 = 1 - 30600/61852 = 0.51$  meaning that it explains a little bit more than 50% of the original variability.

If we repeat this analysis for the data in Fig. 5.2, we would have obtained

Source	SS	df	MS
Treatments	3606	2	1808
Residuals	27205	27	1008
Total	30811	29	

The associated  $f$  is  $f = \frac{1808}{1008} = 1.79$  and the corresponding p-value  $p = 0.18$ , so that we cannot reject the hypothesis that there is no difference between the group means, and consequently the two dose groups are not significantly different from the control group. This ANOVA model only explains  $R^2 = 1 - 27205/30811 = 0.12$ , that is, 12% of the original variability.

### 5.1.2 Regression design

Often times we are interested in many different values of a variable. For instance, we may design an experiment to test 11 levels of a drug in Example 98, from  $D = 0$  mg. (control),  $D = 10$  mg.,  $D = 20$  mg., ...,  $D = 100$  mg. As we saw in the previous section,

we could address this design with an ANOVA design of 11 levels for the treatment variable

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

However, this would require 10 degrees of freedom for estimating the 11 main effects. We note that the dose is continuous and that the response must be a function of the dose, so that we can turn the linear model above into a regression problem with a generic function,  $f(x)$ . For instance, we may use a degree 2 polynomial

$$y_{ij} = f(D_i) = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \varepsilon_{ij}$$

where  $D_i$  is the dose given to the  $i$ -th individual. We only need to estimate 3 parameters ( $\beta_0$ ,  $\beta_1$  and  $\beta_2$ ), that is, this model is much cheaper in terms of degrees of freedom (so that we may use fewer individuals for our experiment). It has two other advantages over the ANOVA linear model: 1) the regression can predict the cholesterol level for values in between the doses used in the experiment (for instance,  $D = 15$  mg.); 2) regression analysis can check whether any of the regression coefficients is significantly different from 0 (see Eq. 4.31), meaning that we may simplify the model if we see that we have overparameterized it.

As a numerical note, it is much better numerically behaved to use centered variables than the raw variables. For instance, let us assume that the mean of all used doses is  $\mu_D$ , then we would construct the centered dose as

$$\tilde{D} = D - \mu_D$$

and then estimate the  $\beta$  parameters on the centered doses

$$y_{ij} = f(\tilde{D}_i) = \beta_0 + \beta_1 \tilde{D}_i + \beta_2 \tilde{D}_i^2 + \varepsilon_{ij}$$

From now on, to avoid obfuscating the notation and presentation of the different topics we will drop this constraint of working with centered, continuous variables. But this recommendation should be kept in mind in a real analysis.

Assuming the regression model has  $P$  parameters, and that the model is linear in these parameters (the  $\beta$ 's do not participate in a non-linear way), the data analysis table is given by

Source	Sum of squares	Degrees of freedom
Regression	$SS_\beta = \sum_{ij} (f(D_i) - y_{..})^2$	$P - 1$
Residuals	$SS_\varepsilon = \sum_{ij} (y_{ij} - f(D_i))^2$	$N - P$
Total	$SS_T = \sum_{ij} (y_{ij} - y_{..})^2$	$N - 1$

The coefficient of determination,  $R^2$ , is still well defined as in Eq. 5.5. The Snedecor's F test can tell us whether the model is significant or not.

**Design summary.** We may organize the experiment as in the case of the completely randomized design (multiple groups of randomly assigned animals, with all animals in the group receiving the same treatment). However, since the predictor is a continuous variable we are not constrained to give the same treatment to all animals. Actually, the experiment gives more information if the treatment is randomly selected from the range covered from the multiple groups (for instance, if we plan to test from 0 to 100 mg. of a dose, then selecting a random number in that range).

As in the case of the completely randomized design, this design assumes that the only source of variation is the treatment. If the experiment can be affected by blocking variables, the design can be extended with the addition of block terms (see the section below treating with covariates).

- **Example 99:** Fig. 5.3 shows the results of analyzing the effect of 11 dose levels of a new drug on the cholesterol level in blood (see Example 98). There are 5 individuals per dose level. The data analysis is performed by regression of the results with a polynomial of degree 2. The fitted polynomial is

$$y = 228.2 - 2.567D + 0.0145D^2$$

The following table shows the sum of squares decomposition for this case.

Source	SS	df	MS
Regression	77629	2	38815
Residuals	45967	47	978
Total	123596	49	

We have  $f = \frac{38815}{978} = 39.68$ , and the associated p-value  $p = 8 \cdot 10^{-11}$ , which is extremely significant. The model explains  $R^2 = 1 - 45967/123596 = 63\%$  of the original variance. Additionally, the confidence intervals for each one of the regression parameters are

$$\begin{aligned}\beta_0 &\in [207.9, 248.5] \\ \beta_1 &\in [-3.513, -1.621] \\ \beta_2 &\in [0.005391, 0.02361]\end{aligned}$$

None of these intervals include the zero value, then, all regression coefficients are statistically significant.

Note that regression models should only be used within the range of predictors for which they were constructed. For instance, this polynomial was fitted in the region  $D \in [0, 100]$  mg., and within this region, the predicted values are relatively accurate (these predicted values are said to be *interpolated*). Outside this region, the polynomial may make sensible or absurd predictions (these values are said to be *extrapolated*). Extrapolated values are not always necessarily bad, it depends on the capability of the fitted function to generalize the system behavior outside the observed region. For instance, this polynomial predicts a cholesterol value of 295 mg/dL for a dose  $D = 200$  mg., which obviously makes no biological sense.

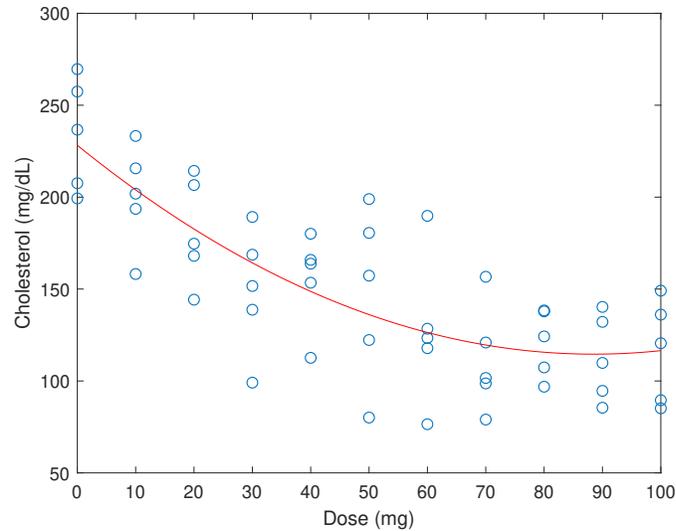


Figure 5.3: Experiment in which we are testing the effect of 11 different doses of a new drug on the cholesterol level in blood. The modelling of the response is performed by regression analysis and the resulting fitted response is shown as a solid, red line.

### 5.1.3 Randomized block design, RBD

We introduced blocking in Sec. 1.4 as a very effective way of reducing variance. The idea is to try to explain the total variability of the observations by other sources.

- Example 100: Following with the previous example, we suspect that there might be differences in the cholesterol level in blood depending on the mouse sex. We design the experiment so that we block sex as a nuisance factor. For example, we plan to carry out the experiment in the following order (randomly generated by a computer)

Female	$D_2$
Male	$D_2$
Female	Control
Female	$D_1$
Male	$D_1$
Female	$D_1$
Male	$D_1$
Female	Control
Male	Control
Male	Control
Male	$D_1$
Female	Control
Female	$D_1$
Female	Control
Female	$D_1$
Male	$D_1$
Female	Control
Female	$D_2$
Male	$D_2$
Male	$D_2$
Female	$D_2$
Male	$D_2$
Male	Control
Male	$D_2$
Female	$D_2$
Female	$D_2$
Male	$D_1$
Male	Control
Male	Control
Female	$D_1$

It can be seen that each dose level has the same number of males and females. This is a balanced design.

After performing the experiment, sex seems to have an effect on the results as seen in Fig. 5.4. We have used the same data as in Fig. 5.2 to illustrate the difference between blocking a nuisance factor, and not blocking it. Within each group, males seem to be systematically above females, and consequently part of the variability can be explained by the animal sex.

We need to extend our linear model to include the sex. Sex is now a blocking variable, also called *nuisance factor*. The following formula extends Eq. 5.1 with the effects of a blocking variable:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk} \quad (5.6)$$

Observations have now three indexes:  $i$  for the treatment (control, dose 1 or dose 2),  $j$  for the block (male or female), and  $k$  for the individual within the treatment and block.

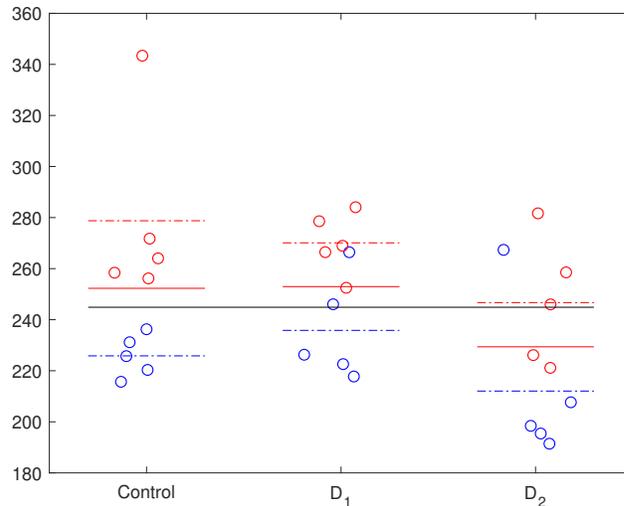


Figure 5.4: Same data as in Fig. 5.2 with sex labels per group (male in red, female in blue). For each group we have drawn the sex average with a dashed line.

It is important to note that the number of animals in each treatment and block is the same (in Fig. 5.4 we see there are 10 animals per treatment, and 15 males and 15 females).

**Design summary.** As in the randomized designs, in this kind of design we have multiple groups, each one receiving a different treatment. Animals are randomly assigned to each one of the treatments. However, in these experiments we foresee that there are nuisance factors that affect our measurements (like performing the experiment in multiple days, multiple researchers each treating only a part of the animals, multiple centers, male or female animals, ...). We design the experiment such that each level of the factors of interest (dose in our example) appears the same number of times in each of the levels of the nuisance factors (for instance, for every dose level we test the same number of males and females). A design meeting this condition is said to be *balanced*. This is not a strong requirement, but it makes the analysis easier. In this section we will assume that each level of the nuisance factors sees all treatments. If it does not, the design is said to be *incomplete*. Incomplete and unbalanced designs will be treated in Sec. 5.1.7.

As we saw in Sec. 1.3, it is important to keep the randomization within each block (treatments are not applied in the same order).

Randomized Complete Block Design (RCBD) is a special case of these designs in which each block sees each treatment exactly once.

We can estimate the main effects of the treatment in the same way as in the previous

case (Eq. 5.2)

$$\hat{\alpha}_i = y_{i..} - y_{...} \quad (5.7)$$

We can estimate the block main effects in the same way

$$\hat{\gamma}_j = y_{.j.} - y_{...} \quad (5.8)$$

If there are  $B$  levels for the blocking variable, then we need  $B - 1$  degrees of freedom to estimate them, because, in the same way as the treatments, the main effects of the blocks (and their estimates) also have to add up to 0

$$\sum_j \gamma_j = 0$$

We now have a decomposition of the sum of squares given by

$$\begin{aligned} \sum_{ijk} (y_{ijk} - y_{...})^2 &= \sum_{ijk} \hat{\alpha}_i^2 + \sum_{ijk} \hat{\gamma}_j^2 + \sum_{ijk} \hat{\epsilon}_{ijk}^2 \\ SS_T &= SS_\alpha + SS_\gamma + SS_\epsilon \end{aligned} \quad (5.9)$$

The residuals, as in the previous case, is the difference between the actual measurements and the predicted value for having received a given treatment and belonging to a particular block

$$\hat{\epsilon}_{ijk} = y_{ijk} - (\hat{\mu} + \hat{\alpha}_i + \hat{\gamma}_j)$$

Finally, we can now extend the ANOVA table to include the blocking variable

Source	Sum of squares	Degrees of freedom
Treatments	$SS_\alpha = \sum_{ijk} \hat{\alpha}_i^2$	$T - 1$
Blocks	$SS_\gamma = \sum_{ijk} \hat{\gamma}_j^2$	$B - 1$
Residuals	$SS_\epsilon = \sum_{ijk} \hat{\epsilon}_{ijk}^2$	$N - T - B + 1$
Total	$SS_T = \sum_{ijk} (y_{ijk} - y_{...})^2$	$N - 1$

The p-value is calculated in the same way as before, by comparing the mean squares of the treatment to the mean squares of the residuals. The advantage of adding blocks is that the treatments will not modify their sum of squares (they still explain the same amount of variability as if no block is considered; this is only true for orthogonal designs, see Sec. 5.1.7, but most experiments with animals are balanced and, consequently, orthogonal), but the variability explained by the block is taken from the residuals. In this way, the sum of squares of the residuals is reduced (remember the connection of blocking and the reduction of variance of Sec. 1.4), and it will be easier to show that the treatment is significant (the associated  $f$  value will be higher).

- Example 100 (continued): If we perform the ANOVA analysis on the data of Fig. 5.4 we obtain

$$\begin{aligned} \hat{\mu} &= 244.88 \\ \hat{\alpha}_0 &= y_{0..} - y_{...} = 7.42 \\ \hat{\alpha}_1 &= y_{1..} - y_{...} = 8.08 \\ \hat{\alpha}_2 &= y_{2..} - y_{...} = -15.50 \\ \hat{\gamma}_0 &= y_{.0.} - y_{...} = 10.30 \\ \hat{\gamma}_1 &= y_{.1.} - y_{...} = -10.30 \end{aligned}$$

In this way, a male mouse without treatment is predicted to have a cholesterol level of

$$y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\gamma}_j = 244.88 + 7.42 + 10.30 = 262.6$$

If we now calculate the sum of squares associated to each one of the terms we would have

Source	SS	df	MS
Treatments	3606	2	1808
Sex	3184	1	3184
Residuals	24021	26	923
Total	30811	29	

We see that the sum of squares of the treatment has not changed, and the sum of squares of the block has been taken out of the sum of squares of the residuals at the cost of a single degree of freedom. We can now calculate the  $f$  statistic associated to the treatments

$$f_{\alpha} = \frac{1808}{923} = 1.96 \Rightarrow \text{p-value} = 0.16$$

We can compare the  $f$  value in this experiment (1.96), with the  $f$  value in the analysis of the same data without the block (1.79; see Example 98). We have slightly improved the p-value, from 0.18 to 0.16. However, we cannot reject the hypothesis that there is no difference between treatments.

In this case, the blocking variable explained only a small portion of the residual variability (3184 out of the 27205, see Example 98, that is, 11.7% of it), and that is why blocking in this case did not help much. However, there are cases in which blocks explain a large portion of the total variance. In these cases, blocking is very useful. Before doing an experiment we cannot know whether blocking will be helpful or not. However, if we suspect that a variable (like sex in our example) could significantly affect the variability of the observations, blocking it and including it in the analysis does not do any harm, and it is a sort of “insurance” just in case we were right about its importance.

Additionally, ANOVA also allows us to analyze the significance of the blocking variables in the same way as we have done for the treatments

$$f_{\gamma} = \frac{3184}{923} = 3.46 \Rightarrow \text{p-value} = 0.07$$

That is, the block was almost significant (the threshold for statistical significance is usually set to 0.05), but with this data we cannot reject the hypothesis that there are no differences in the cholesterol levels of the different sexes. This lack of significance may indicate a lack of statistical power (i.e., not enough animals in each of the groups), or a real lack of difference. The distinction between these two situations is left to the judgment of the researcher, and there is no statistical tool capable of distinguishing between these two situations.

**Important remarks**

102. The analysis of the results of an experiment may give four possible situations.

- Significant p-value, and biologically sensible hypothesis: we designed an experiment to investigate if there was a difference in the different groups, and the data proved that we were right.
- Non-significant p-value, and biologically non-sensible hypothesis: this kind of experiments are more rarely performed. We foresee that there is no difference in the different groups, and normally we do not carry out any experiment to show that this is the case. However, if we still carry out the experiment, and the p-value is not significant, it would confirm our prior biological intuition.
- Non-significant p-value, and biologically sensible hypothesis: we designed an experiment to examine if there was a difference in the different groups, but the data does not support our intuition.
  - If the p-value is almost significant, as in the case of the sex block in the example, 0.07, this result probably indicates a lack of sample size (and, consequently, of statistical power). A larger sample size would most likely result in a significant difference.
  - If the p-value is far from being significant, for instance, 0.6, then we should revise our biological knowledge and try to understand the mechanism that invalidated our prior intuition (we were expecting differences between groups that were not confirmed by the data). We should try to identify some nuisance factors that might have spoiled the experiment.
- Significant p-value, and biologically non-sensible hypothesis: the data seems to support differences between the groups that we did not foresee beforehand.
  - If the p-value is almost non-significant, for instance, 0.04, this result probably indicates a Type I error (rejecting the null hypothesis when it is true). Most likely, we would have not obtained this result with a larger sample size.
  - If the p-value is extremely significant, like  $10^{-6}$ , then we should revise our biological knowledge and try to understand the reasons that make the groups so different and that were not identified before conducting the experiment. You also may suspect confounding.

We may block the animals according to some continuous variable by defining different groups of the continuous variable as in the following example.

- Example 101: We are interested in the effect of four different diets on the growth of the animals. We will use in total 48 animals and will give the same diet

to all animals within a pen. We think that the weight of the animals before performing the experiment may cause a difference in the results. For this reason, we measure the weight of the animals before the experiment and divide them into three groups: light, medium, and heavy animals with 16 animals in each of the groups (this division is easily carried out if we sort the animals by their weight and assign 16 to each one of the groups). These groups are our blocks, and within each block we apply the four different treatments by randomly assigning an animal to one of the four diets and gathering the four animals with the same diet in a pen. In this way, we will have three blocks of weights, and within each block, four pens with the four different diets, and four animals per pen.

The block design above should be preferred to one in which animals are randomly assigned to any of the diets because this latter design might, by chance, assign more heavy animals to one of the diets, making us think that it was the diet (and not the previous weight) what made the animal to grow.

An incorrect design would assign all light animals to the same diet (that is, within a block there is no variety of treatments), because we would be confounding the effect of the block and the effect of the treatment.

Blocking a variable is not always a perfect solution as illustrated by the following example.

- Example 102: We are studying the effect of an hormone in the weight of animals. We will have two groups (control,  $C$ , and treatment,  $T$ ). We have calculated that we need 4 animals per group and they will be put in two cages. We are thinking of two designs

	Cage 1	Cage 2
Design A	CCCC	TTTT
Design B	CCTT	CCTT

From the point of view of eliminating possible cage effects, we would favor Design B over Design A. However, suppose that the hormone does not have an effect on the metabolism of the animals, but on their behavior. Suppose that animals receiving the hormone are more aggressive or docile. Then, the effect on the animal weight is due to the competition between control and treated animals.

#### **Important remarks**

103. Linear models assume there is no interaction between different levels of the same variable. In the previous example of more docile or more aggressive animals, there is an interaction between the control and treated levels, which are confounded in the Design B with the cage block.

We can easily extend the linear model to include several blocking factors (for simplicity of notation, we have dropped the subindexes, but the reader must have in mind

that each observation has been given a different treatment and that it belongs to different levels in each of the blocking variables)

$$y = \mu + \alpha + \gamma^{(1)} + \gamma^{(2)} + \dots + \gamma^{(p)} + \varepsilon \quad (5.10)$$

The analysis of each of the blocking variables is similar to the sex example given above. Typical blocking variables in animal research are sex, strain, age group (although age can be treated as a covariate), the litter the animal is coming from, or its cage. In multicenter experiments, the center is also a typical blocking variable.

- Example 102 (continued): Actually, the problem of distinguishing the effect of the hormone treatment on growth with a higher aggressiveness can be solved by using all possible combinations

	Cage 1	Cage 2	Cage 3	Cage 4
Design A+B	CCCC	TTTT	CCTT	CCTT

and introducing in the analysis an extra variable that reports the effect of mixing animals with different treatments

$$y = \mu + \alpha^{(treatment)} + \gamma^{(cage)} + \alpha^{(mixed)} + \varepsilon$$

- Example 103: We are interested in the evolution over time of a constituent of chicken blood. Our study will span 25 weeks and will involve 9 animals. Every week we can only sample 6 chickens (not the 9), and we expect batch effects due to the sampling day (i.e. variations due to the blood analysis over weeks that are unrelated to the natural time evolution of the blood constituent in the animals).

Multiple blocking can help us to correctly analyze this data. Let  $\mu$  be the base level of that blood constituent. Let  $\alpha_i$  with  $i = 1, 2, \dots, 25$  be the time variation over the base level that we are interested in. We may block the different animals ( $\gamma_j^{(chicken)}$ ,  $j = 1, 2, \dots, 9$ ) and the sampling day ( $\gamma_k^{(day)}$ ,  $k = 1, 2, \dots, 25$ ). Then, we would have the observation model

$$y_{ijk} = \mu + \alpha_i + \gamma_j^{(chicken)} + \gamma_k^{(day)} + \varepsilon_{ijk}$$

And the estimation of its parameters can be done by Least Squares as shown in Sec. 5.1.7.

### 5.1.4 Use of covariates

In Sec. 1.4 we introduced blocking and using covariates as two ways of reducing the variability of the residuals. Blocking addresses discrete variables that might affect our measurements (like sex in the example of the previous section). Covariates can be thought of as blocking for continuous variables. For instance, in the example of the previous section, assume that we foresee that sex and animal weight,  $w$ , may both affect the level of cholesterol in blood. In general we should annotate along with the main results all the variables that might have an influence (sex, weight, cage position in

the rack, health status, operators, handlers, ...) for an eventual posterior analysis trying to understand the variability of the observed data.

We may extend the linear model in Eq. 5.6 to include the body weight. As with the sex, weight is not a variable we can control, but we can measure it and remove from the observations the variability introduced by it:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \beta_w(w_{ijk} - \bar{w}) + \varepsilon_{ijk} \quad (5.11)$$

where  $\bar{w}$  is the mean of the weights, and its presence guarantees that the mean of  $y_{ijk}$  is  $\mu$ . In this model, we are considering the main effects of our treatment (control or two doses, covered by  $\alpha$ ), the effect of a blocking variable (sex,  $\gamma$ ), and the effect of a covariate (body weight,  $\beta_w$ ).

**Design summary.** The use of covariates does not imply an experiment design in itself. It can be used with any design and it only requires the measurement of continuous nuisance factors that could affect our observations.

There are several ways of estimating all the parameters. The following procedure is called *hierarchical ANOVA* and it illustrates a useful way of thinking about residuals, sums of squares and progressive explanation of the variance. The hierarchical procedure iteratively explains parts of the observations. Let us denote the original observations as

$$y_{ijk}^{(0)} = y_{ijk}$$

meaning that the unexplained part of the original observations are the observations themselves.

The most basic model, called the null model, would be their mean. When we have explained the mean, there will still be some unexplained part that we will refer to as  $y_{ijk}^{(1)}$ . Then, for every observation, we have

$$y_{ijk}^{(0)} = \mu + y_{ijk}^{(1)}$$

The unexplained part is supposed to have zero mean (if it did not have, its mean would be absorbed into  $\mu$ ). The equation above is valid for all observations, and we may add all the equations obtaining

$$\sum_{ijk} y_{ijk}^{(0)} = N\mu + \sum_{ijk} y_{ijk}^{(1)}$$

from where the best estimate of  $\mu$  is

$$\hat{\mu} = \frac{1}{N} \sum_{ijk} y_{ijk}^{(0)} = y_{\dots}^{(0)}$$

The unexplained part is now

$$y_{ijk}^{(1)} = y_{ijk}^{(0)} - \hat{\mu}$$

The sum of squares unexplained by this null model is

$$RSS(\mu) = \sum_{ijk} \left( y_{ijk}^{(1)} \right)^2$$

This is called the Residual Sum of Squares because, from the point of view of the model, the unexplained parts are the residuals. This is exactly the total sum of squares,  $SS_T$ , that we have computed in the previous sections. We have consumed 1 degree of freedom to estimate the model because it has only 1 parameter.

We try now to explain the variance induced by the covariates, with the model

$$y_{ijk}^{(1)} = \beta_w(w_{ijk} - \bar{w}) + y_{ijk}^{(2)}$$

The least squares solution for  $\beta_w$  is

$$\hat{\beta}_w = \frac{\sum_{ijk} y_{ijk}^{(1)}(w_{ijk} - \bar{w})}{\sum_{ijk} (w_{ijk} - \bar{w})^2} = \frac{\sum_{ijk} (y_{ijk} - \mu)(w_{ijk} - \bar{w})}{\sum_{ijk} (w_{ijk} - \bar{w})^2}$$

The unexplained part is now

$$y_{ijk}^{(2)} = y_{ijk}^{(1)} - \hat{\beta}_w(w_{ijk} - \bar{w})$$

and the sum of squares unexplained by this first model is

$$RSS(\mu, \beta_w) = \sum_{ijk} \left( y_{ijk}^{(2)} \right)^2$$

and the part explained by  $\beta_w$  when  $\mu$  is given is

$$SS(\beta_w | \mu) = RSS(\mu) - RSS(\mu, \beta_w)$$

We have also consumed only 1 degree of freedom to estimate  $\beta_w$ .

We now explain the part corresponding to the blocking variable (in the example, male and female)

$$y_{ijk}^{(2)} = \gamma_j + y_{ijk}^{(3)}$$

Note that  $\gamma_j$  takes a different value depending if the animal is male ( $j = 0$ ) or female ( $j = 1$ ). For each of the two groups, we add all the equations corresponding to that group

$$\sum_{ik} y_{ijk}^{(2)} = N_j \gamma_j + \sum_{ik} y_{ijk}^{(3)}$$

where  $N_j$  is the number of animals with the block level  $j$ . The best estimate for the block effect would be

$$\hat{\gamma}_j = \frac{1}{N_j} \sum_{ik} y_{ijk}^{(2)} = y_{.j}^{(2)} = \frac{1}{N_j} \sum_{ik} (y_{ijk} - (\hat{\mu} + \hat{\beta}_w w_{ijk}))$$

Since all  $\gamma_j$  (and their estimates) must fulfill the condition

$$\sum_j \gamma_j = 0$$

We need to estimate only  $B - 1$   $\gamma$ 's, because the last one can be inferred from the condition above. In this way, we consume  $B - 1$  degrees of freedom. The unexplained part is now

$$y_{ijk}^{(3)} = y_{ijk}^{(2)} - \hat{\gamma}_j$$

and the sum of squares unexplained by the second model is

$$RSS(\mu, \beta_w, \gamma) = \sum_{ijk} \left( y_{ijk}^{(3)} \right)^2$$

Consequently, the part of the variance explained by the blocks  $\gamma_j$  given  $\mu$  and  $\beta_w$  would be

$$SS(\gamma|\mu, \beta_w) = RSS(\mu, \beta_w) - RSS(\mu, \beta, \gamma)$$

We could go on with this procedure solving for all variables involved in the linear model. In the last step  $n$ , we would still have some unexplained part, these would be the residuals

$$\hat{\epsilon}_{ijk} = y_{ijk}^{(n+1)}$$

and the sum of squares of these residuals would be the finally unexplained sum of squares

The ANOVA table for this model would be

Source	Sum of squares	Degrees of freedom
Covariates	$SS(\beta_w \mu) = RSS(\mu) - RSS(\mu, \beta_w)$	1
Blocks	$SS(\gamma \mu, \beta_w) = RSS(\mu, \beta_w) - RSS(\mu, \beta_w, \gamma)$	$B - 1$
Treatments	$SS(\alpha \mu, \beta_w, \gamma) = RSS(\mu, \beta_w, \gamma) - RSS(\mu, \beta_w, \gamma, \alpha)$	$T - 1$
Residuals	$RSS(\mu, \beta_w, \gamma, \alpha)$	$N - T - B$
Total	$SS_T = RSS(\mu)$	$N - 1$

- Example 104: Let us analyze the data of Fig. 5.4 as we did in Example 100, but introducing the mouse weight as a covariate that is fitted before blocks and treatments. The new parameter estimates are

$$\begin{aligned} \hat{\mu} &= 244.88 \\ \hat{\beta}_w &= 4.24 \\ \hat{\alpha}_0 &= 7.04 \\ \hat{\alpha}_1 &= 6.18 \\ \hat{\alpha}_2 &= -13.22 \\ \hat{\gamma}_0 &= 4.01 \\ \hat{\gamma}_1 &= -4.01 \end{aligned}$$

You may compare these results, with the ones of Example 100. They are different, and notably in the estimate of the effect of sex. The ANOVA table is

Source	$SS$	$df$	$MS$	$f$	p-value
Weight	9866	1	9866	13.8	0.001
Treatments	2626	2	1313	1.84	0.179
Sex	482	1	482	0.68	0.419
Residuals	17837	25	713		
Total	30811	29			

The weight covariate is highly significant, and the sex blocks, that were almost significant in Example 100, have lost most of its significance. The reason is that sex is also correlated with weight. As we have estimated the regression with the weight before sex, then most of the information between sex and cholesterol has been explained by the relationship between weight and cholesterol.

#### Important remarks

104. Linear models can be understood as an attempt to progressively explain variance by adding terms that may have an impact in the variability of the observations.
105. When we follow a sequential procedure as the one presented in this section, the parameter estimates,  $\alpha, \beta, \gamma$  depend on the order in which the parameters are fitted. They do not depend on the order only if the design is *orthogonal* (orthogonal designs are introduced later in Sec. 5.1.7).
106. A consequence is that the sum of squares of the ANOVA table must be understood as sum of squares when the variability explained by the previously fitted parameters have been removed. This is highlighted by the notation  $SS(\alpha|\mu, \beta_w, \gamma)$ , this is the sum of squares explained by the treatments  $\alpha$  when the variability explained by the mean, covariates and blocks has been removed.

### 5.1.5 Linear models, sample size and replications

#### Important remarks

107. Each of the animals in each one of the treatment groups is a replication of the experiment with a given treatment. There is sometimes a confusion between researchers that they need to replicate the experiment three times in order to have statistically significant results. As we saw in Sec. 1.5.2, if the p-value of an experiment is very low, it does not matter how many times we repeat the experiment that we will always find significantly different treatments. This is not the case if the p-value is close to the significance threshold. Instead of repeating the experiment three times, we should design the sample size (see Sec. 4.1.6) such that we have enough power to detect an effect size of our interest. This is the kind of replication we should be interested in, rather than replicating the whole experiment. Still, if the p-value

is close to the significance level, we may repeat the experiment one more time, now increasing the sample size to increase its statistical power, rather than repeating it two more times with the same sample size.

- **Example 104:** In this example we illustrate the relationship between sample size and the analysis of the ANOVA table. Assume, we are comparing a new treatment versus a control, and we only use 3 animals per group. After analyzing the data we note that treatments only explain 40% of the total sum of the squares. As illustrated in the following table

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	p-value
Treatments	2667	1	2667	2.67	0.1778
Residuals	4000	4	1000		
Total	6667	5			

We might have stopped here and reasoned that if treatments only explain 40% of the sum of squares, it is logical that they are not statistically significant: there is more noise (residuals), than signal (treatments), and consequently a low Signal-to-Noise Ratio.

If we had performed the same experiment with 5 animals per group, we would now have

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	p-value
Treatments	5334	1	5334	5.34	0.0497
Residuals	8000	8	1000		
Total	13334	9			

And our treatments would have been just significant (it is only slightly below 0.05). With 30 animals per group, our treatments are extremely significant:

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	p-value
Treatments	38667	1	38667	38.67	$6 \cdot 10^{-8}$
Residuals	58000	58	1000		
Total	96667	59			

### Important remarks

108. In the three cases treatment only explains 40% of the sum of squares, but its significance is not determined by the explained fraction of the sum of squares, but by the explained fraction per degree of freedom. As the number of degrees of freedom of the treatments stays fixed, while the one of the residuals grow, large experiments with many experimental units, can be very sensitive to small fractions of the sum of squares explained by the treatments. This is only another way of expressing that large experiments have a large statistical power (the probability of not rejecting the null hypothesis when it is false, is very low). The reasoning we did in the previous

example of a low Signal-to-Noise Ratio is correct if we think in terms of Mean Squares, not in terms of Sum of Squares.

109. As we saw in the sections dedicated to the sample size calculation, there is a relationship between sample size and statistical power. We should design the sample size in advance before doing the experiment (see Sec. 4.1.6), and not arbitrarily setting the number of animals per group, and then *a posteriori* realizing the reasons why our experiment is just not significant.
110. As a rule of thumb, it is recommended that in laboratory experiments (that are normally small experiments), there are at least between 10 and 20 degrees of freedom for the residuals. Below 10, very likely our experiment will not have sufficient statistical power to prove any useful hypothesis. Beyond 20, our experiment will be a waste of resources for the effect sizes normally sought in research. This Resource Equation Method may be used as a simple method to justify sample size in complicated ANOVAs for which it may be difficult to determine a sample size.

If we reject the ANOVA null hypothesis (none of the treatments is different to the rest), we will look for at least one pair of treatments that are different from each other (the *post-hoc* analysis). In this analysis, we will perform differences between treatment pairs

$$\hat{\Delta}_{i i'} = \hat{\alpha}_i - \hat{\alpha}_{i'}$$

whose associated variance under the null hypothesis (there is no difference between treatments  $i$  and  $i'$ ) is

$$\sigma_{\Delta_{i i'}}^2 = \sigma_{\varepsilon}^2 \left( \frac{1}{N_i} + \frac{1}{N_{i'}} \right)$$

where  $N_i$  and  $N_{i'}$  are the number of animals in the  $i$ -th and  $i'$ -th treatments.

#### Important remarks

111. We see that the number of animals per group will not only help to reject the hypothesis that the different treatments make no difference among the groups (ANOVA null hypothesis), but, as expected, it will also help to identify which treatments are significantly different from each other.

In the following example we explore the danger of dealing with replicates which are not really experimental units.

- Example 105: We are interested in the effect of an abuse drug in the cognitive abilities of mice. We have a control group and two drug doses. They are put in a maze and we measure the time in seconds they take to find the way out.

We study 3 animals per group (identified with labels from A=1 to I=9), and the following table shows the results

Control			Dose 1			Dose 2		
A	B	C	D	E	F	G	H	I
27	43	38	41	30	47	46	34	50

whose ANOVA table is

	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	p-value
Treatments	81	2	40	40/74=0.54	0.61
Errors	447	6	74		

The results are not significant. However, we think that it is a problem of number of samples, and we repeat the time measurements 4 times with the same mice.

Control			Dose 1			Dose 2		
A	B	C	D	E	F	G	H	I
27	43	38	41	30	47	46	34	50
25	43	36	43	35	42	48	37	44
30	46	37	44	31	46	46	38	52
31	44	41	45	35	48	45	35	49

whose ANOVA table is

	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	p-value
Treatments	288	2	144	144/42=3.43	0.04
Errors	1394	33	42		

Now, the results are statistically significant.

#### Important remarks

112. However, we have a serious flaw in the experiment; we are essentially measuring the same thing four times. While this will improve measurement error (and we can see that there is variation in the four measurements per mouse, which could be averaged to give a better estimate of the true value), this scheme does not give independent measurements and the theory used to calculate the p-value does not apply. What we have calculated is not a true p-value. Repeated measures ANOVA, see Sec. 5.2.6, is a more appropriate design tool for this experiments in which the animal itself is considered a block.

#### 5.1.6 Factorial designs, FD

**Design summary.** We are interested in the effect of many factors on the measurements. These factors are discrete (yes or no, several dose levels, ...). If we study the same number of animals under every possible combination of the levels of all factors, the design is said to be balanced. However, this is not a strong requirement of the technique. In this section we will assume a balanced design, and imbalanced designs will be treated later in Sec. 5.1.7. As will be shown later, all balanced designs are orthogonal.

- **Example 106:** Following Example 98, we want to know if there are differences in the cholesterol reduction of the second dose ( $D_2$ , see Fig. 5.1), if the drug is taken fasted or fed, and in combination with a special diet rich in fiber. Additionally, we want to know if any of the combinations is particularly useful/useless?

The dose will no longer be a treatment, because all animals in our experiment will receive  $D_2$  mg. Now, the two treatments are: 1) the stomach state (fasted or fed), and 2) the fiber diet (yes or no). We can extend the linear model to include two predictors (in the nomenclature of ANOVA, each predictor is called a factor). We will refer to the first variable as  $P$  and to the second as  $Q$

$$y_{ijk} = \mu + \alpha_i^{(P)} + \alpha_j^{(Q)} + \varepsilon_{ijk} \quad (5.12)$$

$y_{ijk}$  is the observation of the  $k$ -th animal in the group of animals that have received the  $i$ -th treatment in  $P$  and the  $j$ -th treatment in  $Q$ . Each of the combinations of the levels of the factors is called an experimental treatment (e.g., fasted animals with a fiber rich diet is called an experimental treatment). As usual, the decomposition is restricted such that

$$\sum_i \alpha_i^{(P)} = \sum_j \alpha_j^{(Q)} = 0$$

. This linear model is called two-way ANOVA, as opposed to Eq. 5.1 that is called one-way ANOVA. The number of ways is the number of predictor variables that we have to explain our measurements.

Suppose we have already performed the experiment, we can organize the means of the cells as in the following table (assume that each mean has been computed from  $n = 5$  animals, for example)

	$P = \text{fed}$	$P = \text{fasted}$
$Q = \text{diet}$	170	170
$Q = \text{no diet}$	170	170

Each one of the cells is defined by the indexes  $i$  and  $j$  and it would contain as many observations as animals in each one of the cells. The table above shows the mean of the observations of each of the cells. We can now explain each of the cell means as a function of the main effects of the two predictor variables:

	$P = \text{fed}$	$P = \text{fasted}$	
$Q = \text{diet}$	$170=170+0+0$	$170=170+0+0$	$\hat{\alpha}_{\text{diet}}^{(Q)} = 0$
$Q = \text{no diet}$	$170=170+0+0$	$170=170+0+0$	$\hat{\alpha}_{\text{no diet}}^{(Q)} = 0$
	$\hat{\alpha}_{\text{fed}}^{(P)} = 0$	$\hat{\alpha}_{\text{fasted}}^{(P)} = 0$	$\mu = 170$

If the results of the experiment were these, we would have to admit that the diet or stomach state did not have any effect on the cholesterol level when animals were taking a dose  $D_2$  of our new drug.

Let us now suppose that the experiment results were different as shown in the table below, which already contains the explanation in terms of the linear model.

	$P = \text{fed}$	$P = \text{fasted}$	
$Q = \text{diet}$	$160=170-10+0$	$180=170+10+0$	$\hat{\alpha}_{\text{diet}}^{(Q)} = 0$
$Q = \text{no diet}$	$160=170-10+0$	$180=170+10+0$	$\hat{\alpha}_{\text{no diet}}^{(Q)} = 0$
	$\hat{\alpha}_{\text{fed}}^{(P)} = -10$	$\hat{\alpha}_{\text{fasted}}^{(P)} = 10$	$\mu = 170$

In this case, the stomach state seems to have had an effect (whether this effect is statistically significant is a separate issue that will be address later), but the diet does not affect the results.

Let us now suppose that the experiment gave different results

	$P = \text{fed}$	$P = \text{fasted}$	
$Q = \text{diet}$	$140=170-10-20$	$160=170+10-20$	$\hat{\alpha}_{\text{diet}}^{(Q)} = -20$
$Q = \text{no diet}$	$180=170-10+20$	$200=170+10+20$	$\hat{\alpha}_{\text{no diet}}^{(Q)} = 20$
	$\hat{\alpha}_{\text{fed}}^{(P)} = -10$	$\hat{\alpha}_{\text{fasted}}^{(P)} = 10$	$\mu = 170$

Now, both stomach state and diet have had an impact on the results. We can perfectly explain the cell means only with the main effects of the two predictor variables. However, this is not the case of the following possible results

	$P = \text{fed}$	$P = \text{fasted}$	
$Q = \text{diet}$	$180 \neq 170 + 0 + 0$	$160 \neq 170 + 0 + 0$	$\hat{\alpha}_{\text{diet}}^{(Q)} = 0$
$Q = \text{no diet}$	$160 \neq 170 + 0 + 0$	$180 \neq 170 + 0 + 0$	$\hat{\alpha}_{\text{no diet}}^{(Q)} = 0$
	$\hat{\alpha}_{\text{fed}}^{(P)} = 0$	$\hat{\alpha}_{\text{fasted}}^{(P)} = 0$	$\mu = 170$

The main effect of each one of the predictor variables is 0, because the mean of the fed animals is 170, which is no different from the overall mean (the same happens for all other main effects). Our linear model in Eq. 5.12 cannot explain these observations, and we need to extend the model. In this case, it is said that there are *interactions* between the two factors. Fig. 5.5 shows the results of the two previous experiments (with and without interactions). If there are no interactions, the two represented lines are parallel to each other. For small interactions, the two lines start to be slightly non-parallel. And for strong interactions, the two lines are clearly non-parallel (in this case, they intersect, but intersection is not a necessary condition for the existence of interactions).

We can now include in the formula the interactions between predictors

$$y_{ijk} = \mu + \alpha_i^{(P)} + \alpha_j^{(Q)} + \alpha_{ij}^{(PQ)} + \varepsilon_{ijk} \quad (5.13)$$

The new term  $\alpha_{ij}^{(PQ)}$  is particular to each one of the cells and it represents positive or negative synergies of the combinations of the predictor levels, that is, this particular combination of levels is particularly good or bad. Another way to describe interactions is that the effect of the  $P$  factor depends on the level of the  $Q$  factor, and viceversa. With the interactions we can now explain the cell means (the last number in each cell is the interaction between the two corresponding levels of  $P$  and  $Q$ )

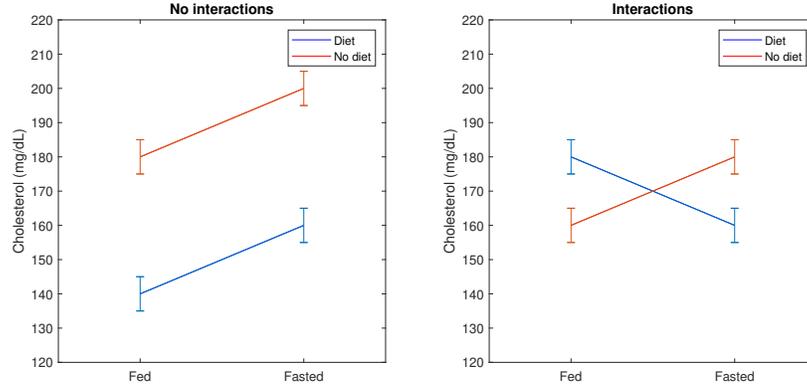


Figure 5.5: Example of results with two factors without (left) and with (right) interactions.

	$P = \text{fed}$	$P = \text{fasted}$	
$Q = \text{diet}$	$180 = 170 + 0 + 0 + 10$	$160 = 170 + 0 + 0 - 10$	$\hat{\alpha}_{\text{diet}}^{(Q)} = 0$
$Q = \text{no diet}$	$160 = 170 + 0 + 0 - 10$	$180 = 170 + 0 + 0 + 10$	$\hat{\alpha}_{\text{no diet}}^{(Q)} = 0$
	$\hat{\alpha}_{\text{fed}}^{(P)} = 0$	$\hat{\alpha}_{\text{fasted}}^{(P)} = 0$	$\mu = 170$

As usual, interactions are also constrained by the ANOVA model to meet for all  $i$

$$\sum_i \alpha_{ij}^{(PQ)} = 0$$

and for all  $j$

$$\sum_j \alpha_{ij}^{(PQ)} = 0$$

That is, the sum of the interactions of each row and column must be zero. We can estimate the different components of the linear model in Eq. 5.13 in the usual way, i.e., by removing from the mean of interest the part we have already explained:

$$\begin{aligned} \hat{\mu} &= y_{\dots} \\ \hat{\alpha}_i^{(P)} &= y_{i..} - \hat{\mu} \\ \hat{\alpha}_j^{(Q)} &= y_{.j.} - \hat{\mu} \\ \hat{\alpha}_{ij}^{(PQ)} &= y_{ij.} - (\hat{\mu} + \hat{\alpha}_i^{(P)} + \hat{\alpha}_j^{(Q)}) \\ \hat{\epsilon}_{ijk} &= y_{ijk} - (\hat{\mu} + \hat{\alpha}_i^{(P)} + \hat{\alpha}_j^{(Q)} + \hat{\alpha}_{ij}^{(PQ)}) \end{aligned}$$

As usual, the ANOVA table explains the decomposition of the sum of squares and number of degrees of freedom

Source	Sum of squares	Degrees of freedom
Treatments $P$	$SS_{\alpha^{(P)}} = \sum_{ijk} \left( \hat{\alpha}_i^{(P)} \right)^2$	$P - 1$
Treatments $Q$	$SS_{\alpha^{(Q)}} = \sum_{ijk} \left( \hat{\alpha}_j^{(Q)} \right)^2$	$Q - 1$
Interactions $PQ$	$SS_{\alpha^{(PQ)}} = \sum_{ijk} \left( \hat{\alpha}_{ij}^{(PQ)} \right)^2$	$(P-1)(Q-1)$
Residuals	$SS_{\varepsilon} = \sum_{ijk} \hat{\varepsilon}_{ijk}^2$	$N - PQ$
Total	$SS_T = \sum_{ijk} (y_{ijk} - y_{...})^2$	$N - 1$

Note that there are 4 interactions in the example above. In general, there are  $PQ$  interactions. However, estimating these many interactions is particularly cheap in terms of degrees of freedom due to their constraints in the sum of the interactions per row and columns. In this example, it only costs 1 degree of freedom.

For any of the rows of the ANOVA table we can test the hypothesis that that row has a statistically significant contribution to the explanation of variability of the observed data. This is done by comparing the corresponding  $MS$  to  $MS_{\varepsilon}$ . Under the null hypothesis (all the treatments in the row are 0), this ratio is distributed with a Snedecor's  $F$  with the number of degrees of freedom corresponding to the row and to the residuals.

- Example 106 (continued): Let us assume that the observed cell means are as shown in the table below

	$P = \text{fed}$	$P = \text{fasted}$	
$Q = \text{diet}$	$150 = 170 - 10 - 20 + 10$	$150 = 170 + 10 - 20 - 10$	$\hat{\alpha}_{\text{diet}}^{(Q)} = -20$
$Q = \text{no diet}$	$170 = 170 - 10 + 20 - 10$	$210 = 170 + 10 + 20 + 10$	$\hat{\alpha}_{\text{no diet}}^{(Q)} = 20$
	$\hat{\alpha}_{\text{fed}}^{(P)} = -10$	$\hat{\alpha}_{\text{fasted}}^{(P)} = 10$	$\mu = 170$

Let us assume that we have 5 individuals per cell, and that the ANOVA table is

Source	$SS$	$df$	$MS$	$f$	p-value
Treatments $P$	1500	1	1500	1500/900	0.215
Treatments $Q$	6000	1	6000	6000/900	0.020
Interactions $PQ$	1500	1	1500	1500/900	0.215
Residuals	14400	16	900		
Total	18900	19			

From this table we see that, with this sample size only the diet,  $Q$ , significantly explains the variability observed in the measurements.

Another typical example of a two-way ANOVA with interactions is the study of the response of males and females (Factor  $P$ ) to a control substance or treatment (factor  $Q$ ). Males and females may respond differently to the treatment and this different response is modelled through the interactions of the two factors.

We can easily extend the two-way ANOVA model to multiple factors. For instance, we may study the stomach state ( $P$ : fed or fasted), special diet ( $Q$ : yes or no), 3 dose levels ( $R$ :  $D_1$ ,  $D_2$  and  $D_3$ ), and an exercise program ( $S$ : yes or no). Additionally, we may include the linear model interactions between pairs of factors (second order interactions), triples (third-order interactions), etc. High-order interactions are normally unexpected and they may easily lead to overfitting.

- **Example 107:** Let us assume that in the study of cholesterol we analyze the 4 predictors above ( $P$ ,  $Q$ ,  $R$ , and  $S$ ), with all second order interactions. The linear model would be

$$y = \mu + \alpha^{(P)} + \alpha^{(Q)} + \alpha^{(R)} + \alpha^{(S)} + \alpha^{(PQ)} + \alpha^{(PR)} + \alpha^{(PS)} + \alpha^{(QR)} + \alpha^{(QS)} + \alpha^{(RS)} + \varepsilon$$

where we have dropped the subindexes for simplicity of notation. Let us assume that we have 2 animals per combination of levels (there are  $24 = 2 \cdot 2 \cdot 3 \cdot 2$  combinations). We organize the animals as shown in Table 5.1.

It can be verified that this design is balanced (every level of each treatment appears the same number of times in any of the combination of the other variables). Note that randomization is an important tool to fight against unknown sources of variations not controlled by the blocking or the studied factors.

Let us assume that, after performing the experiment, the ANOVA table (along with the  $f$  and p-value) is

Source	$SS$	$df$	$MS$	$f$	p-value
Treatments $P$ (stomach)	7200	1	7200	8	0.008
Treatments $Q$ (diet)	28800	1	28800	32	$3 \cdot 10^{-6}$
Treatments $R$ (dose)	50000	2	25000	27.8	$8 \cdot 10^{-8}$
Treatments $S$ (exercise)	6000	1	6000	6.7	0.015
Interactions $PQ$	4000	1	4000	4.4	0.043
Interactions $PR$	5000	2	2500	2.8	0.077
Interactions $PS$	2000	1	2000	2.2	0.146
Interactions $QR$	3000	2	1500	1.7	0.204
Interactions $QS$	3000	1	3000	3.3	0.077
Interactions $RS$	1000	2	500	0.6	0.579
Residuals	29700	33	900		
Total	137700	47			

It is customary to merge all the non-significant rows into a single one called *Lack of fit* (in the example,  $PR$ ,  $PS$ ,  $QR$ ,  $QS$ ,  $RS$ ). Note that it is different merging these rows from “returning” them to the residuals, which would affect the significance of the other rows.

Fasted	Fiber diet	$D_1$	No exercise
Fed	Fiber diet	$D_3$	Exercise
Fed	Fiber diet	$D_2$	No exercise
Fasted	Fiber diet	$D_2$	Exercise
Fed	Fiber diet	$D_1$	Exercise
Fasted	Fiber diet	$D_2$	No exercise
Fasted	Normal diet	$D_1$	Exercise
Fasted	Normal diet	$D_2$	No exercise
Fed	Normal diet	$D_2$	No exercise
Fed	Fiber diet	$D_3$	No exercise
Fed	Normal diet	$D_3$	Exercise
Fed	Normal diet	$D_1$	No exercise
Fed	Fiber diet	$D_2$	No exercise
Fasted	Fiber diet	$D_2$	Exercise
Fed	Fiber diet	$D_1$	No exercise
Fasted	Normal diet	$D_3$	Exercise
Fasted	Normal diet	$D_1$	No exercise
Fed	Fiber diet	$D_3$	Exercise
Fasted	Fiber diet	$D_1$	No exercise
Fasted	Normal diet	$D_2$	No exercise
Fasted	Normal diet	$D_3$	No exercise
Fed	Normal diet	$D_2$	No exercise
Fed	Normal diet	$D_2$	Exercise
Fed	Normal diet	$D_3$	Exercise
Fed	Normal diet	$D_2$	Exercise
Fasted	Normal diet	$D_2$	Exercise
Fasted	Normal diet	$D_1$	Exercise
Fed	Normal diet	$D_1$	Exercise
Fed	Fiber diet	$D_2$	Exercise
Fed	Fiber diet	$D_3$	No exercise
Fasted	Fiber diet	$D_3$	No exercise
Fasted	Fiber diet	$D_3$	Exercise
Fed	Normal diet	$D_1$	No exercise
Fed	Fiber diet	$D_2$	Exercise
Fasted	Fiber diet	$D_3$	Exercise
Fed	Normal diet	$D_3$	No exercise
Fasted	Normal diet	$D_2$	Exercise
Fed	Normal diet	$D_3$	No exercise
Fasted	Fiber diet	$D_1$	Exercise
Fasted	Fiber diet	$D_3$	No exercise
Fasted	Normal diet	$D_3$	No exercise
Fasted	Fiber diet	$D_1$	Exercise
Fed	Fiber diet	$D_1$	No exercise
Fed	Fiber diet	$D_1$	Exercise
Fasted	Fiber diet	$D_2$	No exercise
Fasted	Normal diet	$D_1$	No exercise
Fasted	Normal diet	$D_3$	Exercise
Fed	Normal diet	$D_1$	Exercise

Table 5.1: Example of factorial design with four factors.

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>	p-value
Treatments <i>P</i> (stomach)	7200	1	7200	8	0.008
Treatments <i>Q</i> (diet)	28800	1	28800	32	$3 \cdot 10^{-6}$
Treatments <i>R</i> (dose)	50000	2	25000	27.8	$8 \cdot 10^{-8}$
Treatments <i>S</i> (exercise)	6000	1	6000	6.7	0.015
Interactions <i>PQ</i>	4000	1	4000	4.4	0.043
Lack of fit	14000	8	1750	1.9	0.086
Residuals	29700	33	900		
Total	137700	47			

From this table we see that the main effects of all variables make a difference in the cholesterol level in blood, and that the interactions between having the dose fed/fasted and diet has an effect on the final result.

#### Important remarks

113. We can easily extend the model to include blocking variables. Actually, the difference between a factor and a blocking variable is more semantic than mathematical. Blocking variables are treated in the same way as factors, except that we assume that blocks do not interact with factors and, therefore, we are not interested in calculating the interactions of blocks with any other block or factor. For instance, if we consider the cage as a blocking variable, we are not interested in interactions between cage and dose (this particular dose in this particular cage is particularly good or bad).
114. Factorial designs are very effective in recognizing statistically significant results, even with a very low number of animals per cell (2 in the example). The reason is that we have 'hidden' replication where each level of the predictors has been tested in many different scenarios (combined with many other predictors) and the treatment effects are therefore aggregated over a much larger number of replicates than each individual treatment combination.
115. We should not forget to randomize the animals among the different combinations of factors. The randomization will help to avoid the bias induced by uncontrolled factors.

Among all possible designs, factorial designs give the smallest variance in the comparison of any of its components. Consider the following example:

- **Example 108:** We are interested in the effect of a mammalian hormone for water balance in amphibians. We will study two amphibians species (toads and frogs), and we will examine the difference between performing the experiment when animals are dry before making the experiment and when they have been 30 minutes immersed in water before administering the hormone. The control group will not receive the hormone, but only the vehicle. After receiving the treatment,

animals will be immersed in water for one hour. We will measure the change in weight of the animals after this time.

We have three factors: species ( $S$ ), moisture state ( $M$ ) and hormone treatment ( $H$ ), and we are interested only in the main effects. The linear model we will analyze will be

$$y = \mu + \alpha_S + \alpha_M + \alpha_H + \varepsilon$$

where  $y$  is the weight difference. We have resources for 24 animals, and we want to compare three different designs:

- Design 1: One variable changes at a time

4 Frogs, Dry, No hormone	vs	4 <u>Toads</u> , Dry, No hormone
4 Frogs, Dry, No hormone	vs	4 Frogs, <u>Wet</u> , No hormone
4 Frogs, Dry, No hormone	vs	4 Frogs, Dry, <u>Hormone</u>

Each one of the rows is an experiment in which we may compare two groups (the one on the left with the one on the right). Each experiment changes only one variable, and we may study the effect of that variable. However, note that each level has been tested in only one combination of the other variables. For instance, the difference between hormone and no hormone can only be assessed for dry frogs, and we do not know the effect on other species or previous moisture states. When we analyze the data, we will want to perform a comparison between two groups (see the section around Eq. 4.9). If the variance of water uptakes is  $\sigma_{\Delta w}^2$ , then the expected variance associated to this comparison will be

$$2 \frac{\sigma_{\Delta w}^2}{4}$$

The observations variance is divided by 4 because we have 4 animals in each group, and it is multiplied by 2 because we are comparing two groups.

- Design 2: We realize that in the previous design we have four groups, but one of the groups (“Frogs, Dry, No hormone”) has been repeated three times. We may have only for groups, with a larger number of animals, and for each variable perform the comparison between the two corresponding groups

6 Frogs, Dry, No hormone
6 <u>Toads</u> , Dry, No hormone
6 Frogs, <u>Wet</u> , No hormone
6 Frogs, Dry, <u>Hormone</u>

The expected variance of the comparison of the effect of the hormone is improved to

$$2 \frac{\sigma_{\Delta w}^2}{6}$$

However, this design has the same problem as the Design 1: each level has been tested in only one combination of the other variables.

- Design 3: We perform a factorial design (we should not forget about the randomization when actually performing the experiment, we only report here the number of animals per group)

3 Frogs, Dry, No hormone
3 Frogs, Dry, <u>Hormone</u>
3 Frogs, <u>Wet</u> , No hormone
3 Frogs, <u>Wet</u> , <u>Hormone</u>
3 <u>Toads</u> , Dry, No hormone
3 <u>Toads</u> , Dry, <u>Hormone</u>
3 <u>Toads</u> , <u>Wet</u> , No hormone
3 <u>Toads</u> , <u>Wet</u> , <u>Hormone</u>

The expected variance of the comparison of the effect of the hormone is improved to

$$2 \frac{\sigma_{\Delta w}^2}{12}$$

Additionally, the hormone treatment has been tested with many other levels of the other variables (frogs, toads, dry and wet). Our conclusions from this experiment will be more general than the ones from any of the other two designs.

Interestingly, all the designs use the same number of animals, but they do not have the same statistical power (shown by the variance of the variable being compared, which is the smallest for the factorial design). Also, they do not have the same experimental support (in the factorial design our statements about the effect of the hormone treatment has a wider basis than in Designs 1 and 2, because the treatment has been tested under many more conditions).

- Example 109: Factorial designs can also be used when setting up a new animal model. We may need to decide the sex of the animals, their age range, how to treat the animals before experimentation with the treatment of interest, ...

#### Important remarks

116. In research it is known the strategy of changing one variable at a time holding all the rest fixed (more similar to Design 1 above). Factorial designs seem to contradict this rule. However, they do not. They propose to hold everything fixed, except those variables of interest. These variables of interest should be combined in all possible ways. The analysis is performed at the end after collecting the data from all groups (note that in the Design 1, we could perform the analysis of each one of the rows immediately after collecting the data only for that row, only 8 animals). The collective analysis of the 24 animals is much more powerful than the analysis of individual small experiments as in Design 1.
117. Small factorial designs are used when we are interested in estimating pos-

sible interactions between factors. For designs with many parameters, we should first determine which factors effectively contribute to the final result. This is done with a fractional (or screening) factorial design (see Sec. 5.2.5).

118. The difference between a factor and a block is purely semantic. We may be interested in studying the interactions between factors (e.g., sex and treatment), but we are not interested in the interactions between blocks and factors (e.g., we are not interested in the fact that the treatment (factor) had a slightly different result on a given week (block)).

### Single replicate factorial designs

Some experiments involve a large number of factors and/or levels. If we do not expect high order interactions, we may have a single animal in each one of the combinations and fit a reduced model in which the high order interactions are not estimated (they are confounded with the residuals). It is generally advised to have at least three animals per combination. However, depending on the number of factors, it might well be that the “hidden replications” due to absence of high order interactions allow having a single animal per combination as shown in this section.

- Example 110: We are interested in maximizing the delivery of a drug so that the exposure is maximum. We have identified a few factors that might influence its absorption: salt form ( $P$ , we have identified 3 different forms of the drug that might have different absorption properties), particle size ( $Q$ , by changing the particle size after disintegration of the tablet, the surface area of the microparticles facilitate the absorption, we plan to explore 5 different particle sizes), crystallization form ( $R$ , we have identified 2 polymorphic forms and 1 amorphous form), method of granulation ( $S$ , we may use 2 different methods of granulation), compression force ( $T$ , we will explore 4 different forces).

The total number of combinations is  $3 \cdot 5 \cdot 3 \cdot 2 \cdot 4 = 360$ ). We do not foresee interactions of order higher than 2. We may, then, fit a model only with main effects and second order interactions. For every combination we will analyze a single animal. This may seem surprising, but, as we show in Table 5.2, there are more than enough degrees of freedom for the residuals.

### Important remarks

119. If we can neglect high order interactions, we may drastically reduce the number of samples to just one animal per combination, because the high order interactions act as residuals. However, due to the lack of replication we cannot construct an unbiased estimate of the noise. That is, if we do not foresee high order interactions but in reality there are, then our estimate of the noise variance is biased, confounded, by the presence of these high order interactions. Another difficulty of these designs is that we cannot

Source	<i>df</i>
Salt form ( <i>P</i> )	2
Particle size ( <i>Q</i> )	4
Crystallization form ( <i>R</i> )	2
Method of granulation ( <i>S</i> )	1
Compression force ( <i>T</i> )	3
Interactions <i>PQ</i>	8
Interactions <i>PR</i>	4
Interactions <i>PS</i>	2
Interactions <i>PT</i>	6
Interactions <i>QR</i>	8
Interactions <i>QS</i>	4
Interactions <i>QT</i>	12
Interactions <i>RS</i>	2
Interactions <i>RT</i>	6
Interactions <i>ST</i>	3
Residuals (=3rd, 4th, 5th order interactions)	292
Total	359

Table 5.2: Degrees of freedom associated to a design with multiple factors and their second order interactions.

eliminate the effects of blocks, because we need all treatments applied to all block levels and, therefore, there can only be one block (or at least, as shown in the following section, some of the treatments should be applied to several levels of the blocking variables).

120. Even if we use a single animal per combination, the number of combinations can be very high, 360 in our example above. In Sec. 5.2.5, we will see a method by which we can even further reduce the number of experiments by using *fractional factorial designs*. In this book we only present the theory for factors with two levels ( $2^k$  fractional factorial designs). But the theory can be extended to factors with an arbitrary number of levels.

### 5.1.7 Non-orthogonal, incomplete and imbalanced designs

#### Non-orthogonal designs

In Sec. 5.1.4, we have seen that one way of estimating linear models is by progressively explaining variance of the observations by adding new terms that might be related to the variability observed in the data. Least-squares simultaneously solves all the parameters at once. It is based on trying to solve a linear equation so that the error in each one of the equations is minimized. Understanding Least Squares will allow us to grasp the importance of balanced experiments in producing “easy to handle” equations for

estimating the different parts of the linear model. For explaining this technique, let us introduce an extremely simplified problem, so that the expressions that appear are relatively easy to manage.

- Example 111: Let us consider an experiment in which we are measuring the time in minutes to perform a new surgical treatment compared to a control group receiving the standard surgical treatment. For simplicity of equations, we will only study 3 animals per treatment. We want to generalize our results, so the new surgical operation will be performed by three different researchers, so that each researcher only does one operation of each kind. We expect variations on the data depending on the researcher, so that we will treat the researcher as a block. The following linear model represents our problem:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}$$

where  $i$  is the treatment received ( $i = 1$  is the reference treatment, while  $i = 2$  the new one),  $j$  is the researcher ( $j = 1, 2, 3$ ), and  $k$  is the individual number within that researcher and treatment ( $k$  is always 1, because we have only one animal per treatment and researcher).

After performing the experiment, we collect the following operation times:

	Researcher 1	Researcher 2	Researcher 3
Reference Procedure	32	35	38
New Procedure	26	25	26

We can write these results in matrix form

$$\begin{pmatrix} 32 \\ 35 \\ 38 \\ 26 \\ 25 \\ 26 \end{pmatrix} = \begin{bmatrix} \mu & \alpha_1 & \alpha_2 & \gamma_1 & \gamma_2 & \gamma_3 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{121} \\ \varepsilon_{131} \\ \varepsilon_{211} \\ \varepsilon_{221} \\ \varepsilon_{231} \end{pmatrix}$$

where we have noted on top of the columns of the *system matrix*, the variable by which it is multiplied. In this way, we can read the first row as

$$32 = \mu + \alpha_1 + \gamma_1 + \varepsilon_{111}$$

and similarly the rest of rows. At this point, we remember that we have the constraints

$$\begin{aligned} \alpha_1 + \alpha_2 = 0 &\Rightarrow \alpha_2 = -\alpha_1 \\ \gamma_1 + \gamma_2 + \gamma_3 = 0 &\Rightarrow \gamma_3 = -\gamma_1 - \gamma_2 \end{aligned}$$

so that we can eliminate some of the unknowns of the previous equation system

$$\begin{pmatrix} 32 \\ 35 \\ 38 \\ 26 \\ 25 \\ 26 \end{pmatrix} = \begin{bmatrix} \mu & \alpha_1 & \gamma_1 & \gamma_2 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{121} \\ \varepsilon_{131} \\ \varepsilon_{211} \\ \varepsilon_{221} \\ \varepsilon_{231} \end{pmatrix} \quad (5.14)$$

We need to solve this equation system for the unknowns  $(\mu, \alpha_1, \gamma_1, \gamma_2)$  as well as for the residuals  $(\varepsilon_{ijk})$ . But the solution is not unique, because we have more unknowns than equations. We will need to impose later some constraint so that we can uniquely solve the equation system.

Note that this formalism easily adapts to covariates. Let us assume, that we expect the number of years of research experience of the researcher could affect the results. We can easily extend the model to include it

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \beta(re_{ijk} - \bar{re})\varepsilon_{ijk}$$

where  $re_{ijk}$  is the research experience in years of the researcher, and  $\bar{re}$  its mean value. The following table shows the same results with the research experience in parenthesis

	Researcher 1 (9)	Researcher 2 (5)	Researcher 3 (4)
Reference Procedure	32	35	38
New Procedure	26	25	26

The average research experience is 6. Then, the equation system is

$$\begin{pmatrix} 32 \\ 35 \\ 38 \\ 26 \\ 25 \\ 26 \end{pmatrix} = \begin{bmatrix} \mu & \alpha_1 & \gamma_1 & \gamma_2 & \beta \\ 1 & 1 & 1 & 0 & 3 \\ 1 & 1 & 0 & 1 & 3 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & -2 \\ 1 & -1 & -1 & -1 & -2 \end{bmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \gamma_1 \\ \gamma_2 \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{121} \\ \varepsilon_{131} \\ \varepsilon_{211} \\ \varepsilon_{221} \\ \varepsilon_{231} \end{pmatrix} \quad (5.15)$$

Both equation systems (without and with covariate) are of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where bold letters represent vectors. The vector of residuals can be expressed as

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$$

and the sum of squares of the residuals would be

$$SS_{\varepsilon} = (\mathbf{y} - X\boldsymbol{\theta})^T (\mathbf{y} - X\boldsymbol{\theta})$$

The condition we mentioned above for having a unique solution is that this sum of squares is minimum. This problem is known as the Least Squares solution of the equation system  $\mathbf{y} = X\boldsymbol{\theta}$ , which is given by

$$\hat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \mathbf{y} \quad (5.16)$$

- **Example 111 (continued):** The estimates of the parameters for the experiment with covariates would be

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 30.33 \\ 6.08 \\ -0.49 \\ 0.09 \\ -0.85 \end{pmatrix}$$

Immediately, we can infer the two remaining parameters

$$\begin{aligned} \hat{\alpha}_2 &= -\hat{\alpha}_1 = -6.08 \\ \hat{\gamma}_3 &= -\hat{\gamma}_1 - \hat{\gamma}_2 = 0.40 \end{aligned}$$

The residuals would be

$$\begin{pmatrix} \hat{\varepsilon}_{111} \\ \hat{\varepsilon}_{121} \\ \hat{\varepsilon}_{131} \\ \hat{\varepsilon}_{211} \\ \hat{\varepsilon}_{221} \\ \hat{\varepsilon}_{231} \end{pmatrix} = \begin{pmatrix} -1.38 \\ 1.04 \\ 0.35 \\ 1.38 \\ -1.04 \\ -0.35 \end{pmatrix}$$

We would now use the ANOVA table as we have seen in the previous sections to determine if these variables (treatments, blocks, and covariates) are significantly explaining part of the variability of the observed data.

#### Important remarks

120. Least squares gives a computationally efficient method to simultaneously solve for all the model parameters without the need to decide the order in which they will be estimated as we did in Sec. 5.1.4.

Let us consider for the moment, the experiment without covariates (see Eq. 5.14). Note that the columns of the different variables are orthogonal to each other

$$\langle \boldsymbol{\mu}, \boldsymbol{\alpha}_1 \rangle = \langle \boldsymbol{\mu}, \boldsymbol{\gamma}_1 \rangle = \langle \boldsymbol{\mu}, \boldsymbol{\gamma}_2 \rangle = \langle \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1 \rangle = \langle \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_2 \rangle = 0$$

$\boldsymbol{\gamma}_1$  and  $\boldsymbol{\gamma}_2$  are not orthogonal because they are levels of the same block. Orthogonality is required only between levels of different variables. Remember that two vectors are

orthogonal if their dot product is zero, in the following example we calculate the dot product of two of those vectors

$$\langle \mu, \alpha_1 \rangle = 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot (-1) + 1 \cdot (-1) + 1 \cdot (-1) = 1 + 1 + 1 - 1 - 1 - 1 = 0$$

The reader may easily verify the rest.

Note that the system matrix,  $X$ , depends only on our experiment design (3 researchers will perform 1 operation of each kind) and not on the specific results obtained when the experiment is done. If our design fulfills this orthogonality condition, then the design is said to be orthogonal. We may now study the matrix  $(X^T X)^{-1} X^T$  for an orthogonal example such as the one of Eq. 5.14.

$$(X^T X)^{-1} X^T = \frac{1}{6} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 2 & -1 & -1 & 2 & -1 & -1 \\ -1 & 2 & -1 & -1 & 2 & -1 \end{pmatrix}$$

This matrix will be multiplied by the observed data in order to give the parameter estimates, and again, this matrix depends only on our experiment design

$$\begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 2 & -1 & -1 & 2 & -1 & -1 \\ -1 & 2 & -1 & -1 & 2 & -1 \end{pmatrix} \begin{pmatrix} y_{111} \\ y_{121} \\ y_{131} \\ y_{211} \\ y_{221} \\ y_{231} \end{pmatrix}$$

The first row implies

$$\hat{\mu} = \frac{1}{6}(y_{111} + y_{121} + y_{131} + y_{211} + y_{221} + y_{231}) = y_{..}$$

That is the Least Squares estimate of the mean,  $\mu$ , is the overall mean, as expected. Similarly, we have

$$\begin{aligned} \hat{\alpha}_1 &= \frac{1}{6}(y_{111} + y_{121} + y_{131} - y_{211} - y_{221} - y_{231}) \\ &= \frac{1}{3}(y_{111} + y_{121} + y_{131}) - \frac{1}{6}(y_{111} + y_{121} + y_{131} + y_{211} + y_{221} + y_{231}) \\ &= y_{1..} - y_{..} \end{aligned}$$

Again, as expected, the Least Squares estimate of  $\alpha_1$  is the difference between the mean of those observations that received treatment 1 and the overall mean. Finally,

$$\begin{aligned} \hat{\gamma}_1 &= \frac{1}{6}(2y_{111} - y_{121} - y_{131} + 2y_{211} - y_{221} - y_{231}) \\ &= \frac{1}{2}(y_{111} + y_{211}) - \frac{1}{6}(y_{111} + y_{121} + y_{131} + y_{211} + y_{221} + y_{231}) \\ &= y_{.1.} - y_{..} \end{aligned}$$

We verify once more that the Least Squares estimate of the block effect for Researcher 1, is the difference between the treatments operated by Researcher 1 and the overall mean. We leave the verification of  $\gamma_2$  to the reader.

**Important remarks**

121. We have verified that the Least Squares solutions for the main effects of treatments and blocks are the same as the well known rule of “mean of the group receiving that treatment (or block) minus the overall mean”. Although, we have not verified it, this result extends to interactions of any order.

- **Example 112:** The three researchers participating in the previous study are now so kind to offer themselves to perform an extra operation so that we can better estimate the time reduction in the new surgical procedure, if it exists. Since we have 3 researchers and 2 operation procedures, each one of them will randomly perform one of them. The results of the extra operation are in the following table:

	Researcher 1	Researcher 2	Researcher 3
Reference Procedure	32,30	35	38, 35
New Procedure	26	25,27	26

We now extend the equation system in Eq. 5.14 to include the extra measurements

$$\begin{pmatrix} 32 \\ 35 \\ 38 \\ 26 \\ 25 \\ 26 \\ 30 \\ 27 \\ 35 \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{121} \\ \varepsilon_{131} \\ \varepsilon_{211} \\ \varepsilon_{221} \\ \varepsilon_{231} \\ \varepsilon_{112} \\ \varepsilon_{222} \\ \varepsilon_{132} \end{pmatrix} \quad (5.17)$$

We realize that  $\mu$  is no longer orthogonal to  $\alpha_1$ , and  $\alpha_1$  is not orthogonal to  $\gamma_2$ . The new  $(X^T X)^{-1} X^T$  matrix yields the estimates

$$\begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix} = \frac{1}{108} \begin{pmatrix} 11 & 10 & 11 & 14 & 13 & 14 & 11 & 13 & 11 \\ 9 & 18 & 9 & -18 & -9 & -18 & 9 & -9 & 9 \\ 22 & -16 & -14 & 28 & -10 & -8 & 22 & -10 & -14 \\ -8 & 32 & -8 & -20 & 20 & -20 & -8 & 20 & -8 \end{pmatrix} \begin{pmatrix} y_{111} \\ y_{121} \\ y_{131} \\ y_{211} \\ y_{221} \\ y_{231} \\ y_{112} \\ y_{222} \\ y_{132} \end{pmatrix}$$

The Least Squares estimate no longer has an obvious logic, and we definitely need a computer to estimate the model parameters ( $\mu$ ,  $\alpha$ 's, and  $\gamma$ 's). Even more surprisingly, the Least Squares estimate of  $\mu$  is not the overall mean of all observations.

If we now look at the model with covariates (Eq. 5.15), we see that it is not orthogonal either due to the covariate.

**Important remarks**

122. The loss of orthogonality implies that the estimation of the model parameters is extremely complicated to perform manually and requires a computer. There is nothing wrong with this, only that we can no longer attempt to understand the logic of their estimation.
123. Another property of orthogonal designs is that if we follow a sequential procedure of estimation, as we did in Sec. 5.1.4, the estimates of the parameters do not change whichever is the sequence we follow. Non-orthogonal designs lose this property, and the model parameters vary depending on the order they are fitted.
124. Designs with covariates are almost never orthogonal because their orthogonality depend on the actual measurements observed in the individuals.

Given the equation system

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

we now know how to compute the Least Squares estimate of  $\boldsymbol{\theta}$  (see Eq. 5.16). If the residuals are Gaussian, independent from each other and from the predictors, then this estimate has an associated covariance matrix given by

$$\Sigma_{\boldsymbol{\theta}} = \sigma_{\boldsymbol{\varepsilon}}^2 (X^T X)^{-1} \quad (5.18)$$

where  $\sigma_{\boldsymbol{\varepsilon}}^2$  is the variance of the residuals.

**Important remarks**

125. The whole point of experiment design is designing the system matrix  $X$  such that the uncertainty associated to the model parameters  $\boldsymbol{\theta}$  is minimum. Unfortunately, the uncertainty is given by a matrix, and we cannot “minimize” a matrix. We may minimize its trace ( $A$ -optimality), its determinant ( $D$ -optimality), minimize its maximum eigenvalue ( $E$ -optimality), etc. These different objectives give raise to different designs, with different properties. For some computer programs, the optimality criterion is one of the choices offered to the user. These criteria ( $A, D, E$ ) are also referred to as the efficiency criteria .
126. The experiment designs seen in this chapter (completely randomized, randomized block, factorial, etc.) are simple “precooked” designs that guarantee good properties of the covariance matrix of the model parameters.

**Incomplete designs**

Incomplete designs are useful when for experimental reasons, we cannot test all treatments in all blocks. For instance, let us consider a factorial design with three factors

A, B, and C that can be present (yes) or absent (no). All the possible treatment groups are shown in the table below.

	Factor A	Factor B	Factor C
<del>Treatment 0</del>	<del>no</del>	<del>no</del>	<del>no</del>
Treatment 1	no	no	yes
Treatment 2	no	yes	no
Treatment 3	no	yes	yes
Treatment 4	yes	no	no
Treatment 5	yes	no	yes
Treatment 6	yes	yes	no
<del>Treatment 7</del>	<del>yes</del>	<del>yes</del>	<del>yes</del>

However, experimentally it may not make sense to assess the combination (no,no,no) or (yes,yes,yes). We can skip these two treatments and perform only those that make experimental sense. The Least Squares analysis presented in this section can be directly applied. Incomplete designs are a special case of non-orthogonal designs.

We may also use incomplete designs for complicated factorial designs in which not all combinations are to be tested. Additionally, the number of replicates in each one of the combinations may be different.

- Example 113: We are studying the effect on asthma of two drugs (O and E) that are inhaled, at three different doses ( $D_1, D_2$ , and  $D_3$ ). We also want to study the effect of two different sprayers ( $SP_1$  and  $SP_2$ ). Additionally, Drug O must be given with a surfactant and we want to study two surfactants ( $S_1$  and  $S_2$ ). We are interested in the main effects of each one of the factors and not the interactions between dose. We may design an experiment with the following treatments (each row is a treatment)

Drug	Surfactant	Sprayer	Dose	Number of animals
O	$S_1$	$SP_1$	$D_1$	1
O	$S_1$	$SP_1$	$D_2$	1
O	$S_1$	$SP_1$	$D_3$	1
O	$S_1$	$SP_2$	$D_1$	1
O	$S_1$	$SP_2$	$D_2$	1
O	$S_1$	$SP_2$	$D_3$	1
O	$S_2$	$SP_1$	$D_1$	1
O	$S_2$	$SP_1$	$D_2$	1
O	$S_2$	$SP_1$	$D_3$	1
O	$S_2$	$SP_2$	$D_1$	1
O	$S_2$	$SP_2$	$D_2$	1
O	$S_2$	$SP_2$	$D_3$	1
E		$SP_1$	$D_1$	2
E		$SP_1$	$D_2$	2
E		$SP_1$	$D_3$	2
E		$SP_2$	$D_1$	2
E		$SP_2$	$D_2$	2
E		$SP_2$	$D_3$	2
Control		$SP_1$		5
Control		$SP_2$		5

Before performing the experiment we may analyze the properties of the matrix  $(X^T X)^{-1} X^T$  and decide the number of animals such that we have a given statistical power if the effect size of any of the factors is some specified value (see Sec. 4.1.6).

- Example 110 (continued): Using D-optimality of the matrix  $X^T X$  (see this section above and estimating only the main effects and the 2nd order interactions, we may reduce the number of samples of Example 110 from 360 to 78, still with 1 sample per treatment and with a power higher than 99.5% for all the main effects with an effect size of 2 standard deviations. To give an impression of the kind of treatments applied we show the first 5 samples. Only one animal would receive each of the combinations listed below:

Animal 1	Salt3, Size4, Crystal1, Granulation2, Compression4
Animal 2	Salt2, Size5, Amorphous, Granulation2, Compression3
Animal 3	Salt1, Size2, Amorphous, Granulation2, Compression1
Animal 4	Salt1, Size1, Amorphous, Granulation2, Compression3
Animal 5	Salt2, Size4, Crystal2, Granulation2, Compression1
...	...

These treatments are not randomly, but carefully selected to maximize the determinant of  $X^T X$ . In Sec. 5.2.5 we will extend this idea to factors with only two levels. We will construct these incomplete designs in a more systematic way and call them *fractional factorial designs*, because we only perform a fraction of all the experiments implied by the full factorial design.

### Imbalanced designs

Imbalanced designs are useful when we cannot study all possible combinations of treatments and blocks for economical or ethical reasons or any other consideration. Imbalanced designs can also be analyzed by Least Squares. Before introducing the analysis of imbalanced designs, let us review again a balanced one. Let us suppose we are studying the difference between two treatments (A and B). We expect differences between males and females, and we want to block them. The following design shows the distribution of treatments per block

Male	A B
Female	A B

Let us assume we use the same number of individuals per treatment and block ( $N = N_{male,A} = N_{male,B} = N_{female,A} = N_{female,B}$ ). This is a balanced design because each treatment appears in each block the same number of times. A model for analyzing this data could be:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}$$

For balanced designs, we know that the best estimates of the model parameters are

$$\begin{aligned}\hat{\mu} &= y_{...} \\ \hat{\alpha}_A &= y_{A..} - y_{...} \\ \hat{\alpha}_B &= y_{B..} - y_{...} \\ \hat{\gamma}_{Male} &= y_{.Male.} - y_{...} \\ \hat{\gamma}_{Female} &= y_{.Female.} - y_{...}\end{aligned}$$

In the post-hoc analysis (see Sec. 5.1.5), when we compare the difference between treatments A and B, we will construct the difference between the two:

$$\hat{\Delta}_{AB} = \hat{\alpha}_A - \hat{\alpha}_B = y_{A..} - y_{B..}$$

whose associated variance is

$$\sigma_{\Delta_{AB}}^2 = \sigma_{y_{A..}}^2 + \sigma_{y_{B..}}^2 = \frac{\sigma_{\varepsilon}^2}{2N} + \frac{\sigma_{\varepsilon}^2}{2N} = \frac{\sigma_{\varepsilon}^2}{N} \quad (5.19)$$

where  $N$  is the number of samples per group, and  $\sigma_{\varepsilon}^2$  the variance of the residuals. The variance in each one of the treatment groups is  $\frac{\sigma_{\varepsilon}^2}{2N}$  because for each treatment there are  $N$  females and  $N$  males.

We are now interested in studying three treatments (A, B and C) and we plan to do it with the following design

Male	A C
Female	B C

Repeating the Least Squares analysis, we would reach the conclusion that the best

estimates of the model parameters are

$$\begin{aligned}\hat{\mu} &= \frac{1}{6}(4y_{...} + y_{A..} + y_{B..}) \\ \hat{\alpha}_A &= \frac{1}{3}(4y_{A..} - 6y_{Male.} - 2y_{B..} + 4y_{...}) \\ \hat{\alpha}_B &= \frac{1}{3}(-2y_{A..} + 6y_{Male.} + 4y_{B..} - 8y_{...}) \\ \hat{\alpha}_C &= \frac{1}{3}(-2y_{A..} - 2y_{B..} + 4y_{...}) \\ \hat{\gamma}_{Male} &= \frac{1}{2}(4y_{Male.} - y_{A..} + y_{B..} - 4y_{...}) \\ \hat{\gamma}_{Female} &= -\frac{1}{2}(4y_{Male.} - y_{A..} + y_{B..} - 4y_{...})\end{aligned}$$

We are not surprised by the strange coefficients for the estimation of the model parameters, as we already know that they come from matrix  $(X^T X)^{-1} X^T$ . It is noteworthy that for non-orthogonal designs, like this one, the estimate of the effect of treatment A, involves the mean of the animals receiving A (as expected), and the mean of the animals receiving B (unexpected). Similarly, the estimate of the effect of treatment C does not require the mean of the animals receiving C. The reason is that the constraints  $(\alpha_A + \alpha_B + \alpha_C = 0$  and  $\gamma_{Male} + \gamma_{Female} = 0)$  link the mean of all the groups, so that the mean of group C can be calculated from the knowledge of A and B.

In the post-hoc analysis to compare A and B we will construct their difference

$$\hat{\Delta}_{AB} = \hat{\alpha}_A - \hat{\alpha}_B = 2y_{A..} - 4y_{Male.} - 2y_{B..} + 4y_{...}$$

whose associated variance is

$$\begin{aligned}\sigma_{\Delta_{AB}}^2 &= 4\sigma_{y_{A..}}^2 + 16\sigma_{y_{Male.}}^2 + 4\sigma_{y_{B..}}^2 + 16\sigma_{y_{...}}^2 \\ &= 4\frac{\sigma_{\epsilon}^2}{N} + 16\frac{\sigma_{\epsilon}^2}{2N} + 4\frac{\sigma_{\epsilon}^2}{N} + 16\frac{\sigma_{\epsilon}^2}{4N} = 20\frac{\sigma_{\epsilon}^2}{N}\end{aligned}\quad (5.20)$$

Now the variance of the comparison between A and B is 20 times the one of a balanced design (see Eq. 5.19).

#### Important remarks

127. There is an important increase in the variance associated with the comparison between treatments A and B. This increase is partly because A and B are now tested on only  $N$  animals (while before they were tested in  $2N$  animals) and partly because with  $4N$  animals before we were testing only 2 treatments, and now 3 treatments with the same number of animals.
128. The more number of times in which A and B occur in the same block, the more efficient the design will be for assessing the difference between these two treatments.

If we now compare A and C we will have

$$\hat{\Delta}_{AC} = \hat{\alpha}_A - \hat{\alpha}_C = 2y_{A..} - 2y_{Male.}$$

whose associated variance is

$$\begin{aligned}\sigma_{\Delta_{AC}}^2 &= 4\sigma_{y_{A..}}^2 + 4\sigma_{y_{Male.}}^2 \\ &= 4\frac{\sigma_{\epsilon}^2}{N} + 4\frac{\sigma_{\epsilon}^2}{2N} = 6\frac{\sigma_{\epsilon}^2}{N}\end{aligned}\quad (5.21)$$

We obtain the same variance for the comparison BC.

**Important remarks**

129. Not all comparisons between treatments, AB or AC, have the same level of uncertainty (variance). C is the link between the two blocks and it has been tested twice (one with males, another one with females). This link makes that the comparison of treatments within the same block (AC or BC) is much less variable than the comparison of treatments in different blocks (AB).
130. The fact that there is a common treatment in the two blocks, C, makes the comparison between A and B possible. The variance of this comparison decreases as the number of common treatments increases. For instance, the comparison between AB in the previous example, where only C is in common, is larger than the variance of the same comparison for the design with four treatments shown below, where there are two treatments in common.

Male	A C D
Female	B C D

Let us analyze the design of four treatments

Male	A B
Female	C D

We note that there is no treatment in common between the two groups. From the design itself, we may infer that we are confounding the blocks with the treatments (if males give a higher response than females, we cannot know if it is because they are males or because of the treatments A and B). Furthermore, we may verify that the matrix  $X^T X$  is not invertible, meaning that there is not a Least Squares solution. Actually, the system is ill-defined and there are infinite solutions, all of them of the form:

$$\begin{aligned}
 \hat{\mu} &= y_{..} \\
 \hat{\alpha}_A &= 2y_{A..} - 2y_{.Male.} + \alpha_C \\
 \hat{\alpha}_B &= -y_{A..} + y_{B..} + 2y_{.Male.} - 2y_{..} - \alpha_C \\
 \hat{\alpha}_C &= \alpha_C \\
 \hat{\alpha}_D &= -y_{A..} - y_{B..} + 2y_{..} - \alpha_C \\
 \hat{y}_{Male} &= -y_{.A.} + 2y_{.Male.} - y_{..} - \alpha_C \\
 \hat{y}_{Female} &= -(-y_{.A.} + 2y_{.Male.} - y_{..} - \alpha_C)
 \end{aligned}$$

$\alpha_C$  is a free variable, we may give any value to it and the rest of the model parameters would adjust themselves accordingly, in this way we have infinite solutions compatible with the observed data. After performing the experiment we would realize that this experiment cannot be analyzed and that it does not give any information about the main effects of the treatments, nor the blocks: a total catastrophe for a researcher.

**Important remarks**

131. There are experimental designs that cannot be analyzed and in which the

effects of the treatments and blocks are confounded. Confusion of factors is normal in screening experiments (see Sec. 5.2.5). But they are especially designed to confound in a controlled way.

- **Example 114:** We want to determine the effect on the growth of animals with three different hormone doses ( $D_1$ ,  $D_2$ , and  $D_3$ ) and a control ( $C$ ). We will measure five animals per group. We think that the litter animals come from may cause a difference. For this reason, we will take four animals from five litters. The most efficient design (the one that allows the comparison of any pair of treatments with equal variability) would be the balanced and complete one:

	Treatments
Litter 1	$C, D_1, D_2, D_3$
Litter 2	$C, D_1, D_2, D_3$
Litter 3	$C, D_1, D_2, D_3$
Litter 4	$C, D_1, D_2, D_3$
Litter 5	$C, D_1, D_2, D_3$

The following imbalanced design would be necessarily more inefficient, note that in some of the blocks there are only two different treatments. Although inefficient, data analysis is still possible because there is a treatment linking all blocks.

	Treatments
Litter 1	$C, D_1, D_1, D_1$
Litter 2	$C, D_1, D_1, D_2$
Litter 3	$C, D_2, D_2, D_2$
Litter 4	$C, D_2, D_3, D_3$
Litter 5	$C, D_3, D_3, D_3$

Finally, the following imbalanced design is incorrect because there are blocks that receive a single treatment. In this way, the litter effect is confounded with the treatment.

	Treatments
Litter 1	$C, C, C, C$
Litter 2	$C, D_1, D_1, D_1$
Litter 3	$D_1, D_1, D_2, D_2$
Litter 4	$D_2, D_2, D_2, D_3$
Litter 5	$D_3, D_3, D_3, D_3$

### Balanced incomplete block designs

As we have seen, having a balanced design helps us to keep the estimation equations understandable. Additionally, it does not favour any comparison between treatments. If our blocks cannot hold all treatments, then we may try to find a balanced incomplete block design. A design is *balanced* if:

1. All treatments are applied the same number of times.
2. All pairs of treatments appear in the same number of blocks.

For instance, the following design is balanced because: 1) each treatment is applied 5 times; and 2) each pair (AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF, EF) appears 2 times.

	Treatments
Block 1	A B C
Block 2	A B D
Block 3	A C E
Block 4	A D F
Block 5	A E F
Block 6	B C F
Block 7	B D E
Block 8	B E F
Block 9	C D E
Block 10	C D F

- Example 115: We want to determine the effect on weight gain of two levels of protein supplement (high or low, represented as P and p, respectively) and vitamin supplement (high or low, represented as V and v). There are four different treatments in total (all possible combinations of protein and vitamin levels, that we will represent as A, B, C, and D) and we will use three animals for each of the treatments (twelve animals in total). We think that the genetics of the animal may cause a difference and to account for it, we will use six pairs of siblings. The sibling pair is our block, but we can only test two treatments on each block. The following design is a balanced incomplete design suitable for our needs:

	Treatments
Sibling pair 1	A(pv) B(pV)
Sibling pair 2	A(pv) C(Pv)
Sibling pair 3	A(pv) D(PV)
Sibling pair 4	B(pV) C(Pv)
Sibling pair 5	B(pV) D(PV)
Sibling pair 6	C(Pv) D(PV)

That is, one of the animals of the first sibling pair will receive treatment A (low protein and vitamin supplements) and the other will receive treatment B (low protein and high vitamin supplements). It can easily be seen that each treatment is applied to exactly 3 animals and that all pairs of treatments appear exactly once.

Let us define these designs in general. For doing so, let us define the following variables

$t$	No. Treatments
$b$	No. Blocks
$r_i$	No. of blocks containing treatment $i$ For a balanced design $r_i = r$ for all treatments
$k$	Size of the block
$\lambda_{ii'}$	No. of blocks containing treatments $i$ and $i'$ For a balanced design $\lambda_{ii'} = \lambda$ for all pairs

The designs are named  $(t, b, r, k, \lambda)$ -designs. A balanced design must fulfill:

$$\begin{aligned} bk &= tr \\ r(k-1) &= \lambda(t-1) \end{aligned}$$

The first equation simply states that the number of blocks times their size must be equal to the number of treatments and their repeats.  $r - \lambda$  is the *order* of the design. There does not exist a solution for all possible combinations of  $t, b, r, k$ , and  $\lambda$ . For instance, for  $t = 6$  with  $k \leq t/2$  and  $b \leq 30$ , the only solutions are

$t$	$b$	$r$	$k$	$\lambda$
6	10	5	3	2
6	20	10	3	4
6	30	15	3	6

For a comprehensive list of existing solutions see [Zwillinger \(1996\)](#)[Sec. 3.4.2].

A necessary condition to be balanced is that the row and column sums of the incidence matrix are all equal as in the following example

	Treatments
Block 1	A B C
Block 2	A B D
Block 3	A C E
Block 4	A D F
Block 5	A E F
Block 6	B C F
Block 7	B D E
Block 8	B E F
Block 9	C D E
Block 10	C D F

whose incidence matrix is

Block \ Treatment	A	B	C	D	E	F	
Block 1	1	1	1				3
Block 2	1	1		1			3
Block 3	1		1		1		3
Block 4	1			1		1	3
Block 5	1				1	1	3
Block 6		1	1			1	3
Block 7		1		1	1		3
Block 8		1			1	1	3
Block 9			1	1	1		3
Block 10			1	1		1	3
	5	5	5	5	5	5	

However, this condition is not sufficient as shown by the following example

	Treatments
Block 1	A C
Block 2	B D
Block 3	A C
Block 4	B D

whose incidence matrix is

Block \ Treatment	A	B	C	D	
Block 1	1		1		2
Block 2		1		1	2
Block 3	1		1		2
Block 4		1		1	2
	2	2	2	2	

The pair AC appears 2 times ( $\lambda_{AC} = 2$ ), while AB or AD do not appear ( $\lambda_{AB} = \lambda_{AD} = 0$ ).

An easy way to design experiments is by starting with an initial block (for instance, ABD) and adding 1 to each treatment modulo the number of treatments (that is,  $A+1=B$ ;  $B+1=C$ ;  $C+1=D$ ;  $D+1=E$ ;  $E+1=A$ ). This is called a *cyclic design*. For example, for 5 blocks of size 3 with 5 treatments we would start with the initial block ABD. Then, by adding 1 to each of the treatments we would obtain, BCE. The rest of blocks are obtained by adding 1 to the previous block as shown in the following table

	Treatments
Block 1	A B D
Block 2	B C E
Block 3	C D A
Block 4	D E B
Block 5	E A C

Note that not all initial blocks give raise to a balanced incomplete block design, and you may need to test several initial blocks before finding one.

Another easy way to generate balanced incomplete designs are based on lattices. These designs are called *lattice designs*. For example, for 7 blocks of size 3 with 7

treatments, we construct a Latin square with 7 treatments (see Sec. 5.2.1 and Fig. 5.6). Then, we take 3 columns (not any 3 are valid) and construct the different blocks. These rectangles are called *Youden squares*.

A	B	C	D	E	F	G
B	C	D	E	F	G	A
C	D	E	F	G	A	B
D	E	F	G	A	B	C
E	F	G	A	B	C	D
F	G	A	B	C	D	E
G	A	B	C	D	E	F



A	B	D
B	C	E
C	D	F
D	E	G
E	F	A
F	G	B
G	A	C

Figure 5.6: Example of Youden square. We start from a Latin square of the total number of treatments (left). Then, we select a number of columns equal to the block size (right). If we choose wisely the columns, the resulting design is balanced.

Although outside of the scope of this chapter, for a large number of treatments (a few hundreds), the interested reader may look for Cubic lattice designs and Alpha lattice designs for large-scale variety trials.

#### Important remarks

132. Balanced designs are important to keep estimation equations understandable. If we need to use blocks in which not all treatments fit, balanced incomplete block designs help us to keep these two objectives (using blocks and having a balanced design). However, these designs only exist for given combinations of the number of treatments, blocks and size of the block.

## 5.2 Advanced designs

### 5.2.1 Latin squares

**Design summary.** Latin squares is a special kind of design in which there is a single treatment factor with  $L$  levels, and two blocking variables, each one with as many levels as the treatment factor.

- **Example 116:** We want to study the time in hours to recover from a small surgical operation. We can perform it in 4 different ways (A, B, C, or D), and we will block the researcher performing the operation (4 different researchers will be employed). We expect variations depending on the time the operation is performed (9:00, 12:00, 15:00, 18:00) that we also want to block. We may use the design shown below

Researcher \ Time	9:00	12:00	15:00	18:00
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Each treatment appears only once in all rows and columns (that is the property that defines a Latin square). Note that every researcher performs all operations, and that all operations are done at a given time. If we perform only one operation per cell in the table, we would have the following table of degrees of freedom

Source	<i>df</i>
Treatments	3
Researcher	3
Time	3
Residuals	6
Total	15

Having only 6 degrees of freedom for the residuals has not much statistical power for the standard effect sizes sought in research experiments. After calculating the sample size (see Sec. 4.1.6), the total number of samples is  $N = 25$ , we decide to increase it to  $N = 32$  in order to have a balanced design and have two samples per combination of blocks and treatments. Instead of repeating twice the same Latin table, we may use a different Latin square as shown below (the upper and lower parts of the table are Latin squares).

Researcher \ Time	9:00	12:00	15:00	18:00
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C
1	B	D	A	C
2	A	B	C	D
3	D	C	B	A
4	C	A	D	B

We may increase the generalizability of the experiment by studying the treatments with a wider range of researchers and times

Researcher \ Time	9:00	12:00	15:00	18:00	8:00	11:00	14:00	17:00
1	A	B	C	D				
2	B	C	D	A				
3	C	D	A	B				
4	D	A	B	C				
5					B	D	A	C
6					A	B	C	D
7					D	C	B	A
8					C	A	D	B

The table of degrees of freedom would be

Source	$df$
Treatments	3
Researcher	7
Time	7
Residuals	14
Total	31

The number of Latin squares of a given size  $L$  is limited, and beyond a number of replications we need to repeat some of them. For  $L = 2$ , there are only 2 Latin squares; for  $L = 3$ , there are 12; for  $L = 4$ , there are 576; for  $L = 5$ , there are 161,280; for  $L = 6$ , there are 812,851,200. Latin squares have been proved to be more efficient than their complete block designs (see Sec. 5.1.3) counterparts (Giesbrecht and Gumpertz, 2004)[p. 125].

- **Example 117:** Let us consider an experiment in which we analyze the learning ability of animals under three different levels of wheel exercise: no exercise, moderate, and intense. Due to circadian rhythms, the result may depend on the time of the day the experiment is carried out. Also, it may depend on the body weight of the animal. These are two nuisance factors we would like to get rid of. For this purpose, we can define three levels for the time of the day, and three levels for the body weight (for doing so, we simply sort the animals by their body weight and divide the set of animals in three groups of equal size). Then, we use a Latin square design for our experiment. In this way, we are sure to remove the possible biases induced by these two nuisance factors.
- **Example 118:** Another example that calls for a Latin squares design is if we are using mice whose cages are placed on racks. It has been reported (Gore and Stanley, 2005) that the amount of water intake of the animals depended on the row position of the cage within the rack, and that the body temperature depended on the column of the rack. If we want to block these two effects, we may use a Latin squares design within each rack with 5 rows and 5 columns. As stated above, Latin squares can only be applied if we are studying the same number of treatments as the number of blocks in rows and columns. If this is not the case, we need to use a block design (see Sec. 5.1.3) or one of their incomplete or imbalanced versions (see Sec. 5.1.7).

### 5.2.2 Graeco-Latin squares

**Design summary.** Graeco-Latin squares result from the superposition of two Latin squares and they allow us to simultaneously perform two different experiments with just one treatment factor and two nuisance factors, or to consecutively perform experiments.

- Example 119: We are studying the effect of four different cleaning products on the stress of the animals in an animal facility. Four centers participate in the study, and each one of them has four rooms with cages. Simultaneously, we are making a different study, also on the stress of animals, with four different types of cages. Can we perform this two experiments simultaneously without any one of them interfering with the other?

We may use two mutually orthogonal Latin squares: one with the four cleaning products (A, B, C, D) and the other one with the four types of cages ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ). This kind of designs are called Graeco-Latin squares:

Center \ Room	1	2	3	4
1	A $\alpha$	D $\delta$	B $\gamma$	C $\beta$
2	C $\delta$	B $\alpha$	D $\beta$	A $\gamma$
3	D $\gamma$	A $\beta$	C $\alpha$	B $\delta$
4	B $\beta$	C $\gamma$	A $\delta$	D $\alpha$

Note that each treatment of one kind (cleaning product or cage) appears exactly once with all treatments of the other kind. The Latin letters form a Latin square, as well as the Greek letters. These two Latin squares are said to be mutually orthogonal and each combination of pairs of treatments (A $\alpha$ , A $\beta$ , ..., D $\delta$ ) appears only once. Each of the cages in the same room would be considered an experimental unit receiving the combined treatment. As we mentioned in the section above, the number of Latin squares of a given size is limited, and the pairs of orthogonal Latin squares even more limited. For  $L = 2$ , there is only 1 pair of mutually orthogonal Latin squares; for  $L = 3$ , 2; for  $L = 4$ , 3; for  $L = 5$ , 4; for  $L = 6$ , 1.

We have presented Graeco-Latin squares in a context of two simultaneous experiments. But they are also used for consecutive experiments: we first perform the experiment on the cleaning products, and when it is finished, we perform the experiment on the cage types. However, we use a Graeco-Latin design so that there is no carryover effect from the first experiment to the second.

### 5.2.3 Cross-over designs

**Design summary.** In cross-over designs we block time and individuals. In this way, we eliminate the inter-subject variability from the analysis because an individual is its own control and reduce the number of subjects if we keep fixed the

statistical power, or increase the statistical power if we keep fixed the number of subjects.

- Example 120: We are studying the pain reduction caused by an analgesic. There are two treatments: control (with only the vehicle) and treatment (with the drug). We plan to perform a cross-over design in which an animal receives first one of the treatments, and we perform the measure of pain reduction. Then, we wait for a wash-out period such that there is no interference between the first and second treatment. Finally, we give the second treatment and measure again. The execution plan is as follows

Period \ Subject	Subject											
	1	2	3	4	5	6	7	8	9	10	11	12
1	C	T	T	C	C	T	C	T	T	T	C	C
2	T	C	C	T	T	C	T	C	C	C	T	T

- Example 121: We are carrying out an experiment to determine taste preferences of mice. We have 5 possible tastes (all of them dissolved in water). In each cage, we have two bottles: one with the solution and another one with water. Bottles are refilled and swapped every day, and each solution will be tested in a cage for 1 week. Everyday we will annotate the difference between the weight of the remaining liquid in the bottle with the solution and the bottle with water. All cages will taste each one of the solutions one after the other.
- Example 122: Another example in which cross-over designs may be valuable is when animals need to be trained for performing the experiment. In this case, the animals are a very valuable asset. If we are testing several drugs, we would like to test several of these drugs on each animal, so that the training time is not wasted. We may test several drugs on the same animal as long as the animal is not “fundamentally modified” and we leave a wash-out period between tests. Examples of experiments in which the animal is not fundamentally modified are experiments in which we measure the preference of animals with respect to several types of nests (the same animal can subsequently test different kinds of nests) or experiments in which we measure the gastrointestinal transit rate under different physiological conditions or the ADME (Absorption, Distribution, Metabolism, and Excretion) properties of a drug that is fully eliminated. In any case, as shown in the following paragraphs, we may take statistical protections against order, carry-over, and learning effects.

Cross-over designs can only be used when there is no interference from the first treatment to the second. In a way, the animal seeing the first treatment is not the “same” animal that sees the second, even if it is the same individual. Interferences can be of three kinds:

- Order effects: For instance, if we are using diseased animals and the first treatment cures the disease, we cannot apply the second or if we apply, its application is useless. The order in which we apply the treatments modifies in an irreversible way the state of the animal.

- Carry-over effects: There is still some of the first treatment leftover when we apply the second (for instance, the drug has not been completely eliminated from the body). These negative effects are easily removed by sufficiently long washout periods, or by the use of statistical designs aimed at removing 1st, 2nd, ... order carryover effects, as we will see below.
- Learning effects: Another example is with mice in a maze when one of the rooms has some abuse substance. The study is on the amount of time spent in each of the rooms of the maze. In the second treatment, mice remember which was the configuration of the maze under the first treatment, and this memory modifies the time that naive animals spend in each of the rooms under the second treatment.

Our observation model this time would be

$$y_{ijk} = \mu + \alpha_i + \gamma_j^{(period)} + \gamma_k^{(individual)} + \varepsilon_{ijk}$$

and it could be analyzed using the Least Squares method presented in Sec. 5.1.7. We see that this model blocks the individual (and consequently, the intersubject individual) and time (that is, the order in which treatments are applied). To achieve this latter goal it is important that the design is balanced (for instance, it has as many CT as TC orderings in the example above).

A design is *balanced with respect to 1st order carryover effects* if each treatment precedes any other treatment the same number of times. For instance, with four treatments (A, B, C, D), a design based on the following sequences is not balanced with respect to 1st order carryover effects:

Period			
1	2	3	4
A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

The reason is that A precedes B three times, while B never precedes A. However, we can find suitable sequences for four treatments like

Period			
1	2	3	4
A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

In this design A precedes B, C, and D only once, and the same happens with all other treatment pairs. Once we have found appropriate sequences of treatments, we must assign the same number of individuals to each of the sequences.

These sequences can be found with the help of Latin squares (see Sec. 5.2.1), the two previous examples of sequences of four elements were both Latin squares. However, not all Latin squares produce designs balanced with respect to 1st order carryover

effects. For an even number of treatments we can find such sequences with the help of a single Latin square. For an odd number of treatments we require the help of two Latin squares.

Period		
1	2	3
A	B	C
B	C	A
C	A	B
A	C	B
B	A	C
C	B	A

The upper half of the table is a Latin square as well as the lower half. For three treatments, a design balanced with respect to 1st order carryover effects requires six sequences of length three. We may verify that A precedes B and C exactly twice, and the same happens with all other pairs of treatments.

We may prefer designs that are *strongly balanced with respect to 1st order carryover effects*. A design is of this type if each treatment precedes all other treatments, including itself, the same number of times. For instance, a design based on the sequences TC and CT is not strongly balanced because C precedes T once, but it does not precede itself. However, a design based on the sequences TCC and CTT is strongly balanced.

Additionally, we may require the design to be 1) *uniform within sequences*, that is, each treatment appears the same number of times in each sequence, and 2) *uniform within periods*, that is, each treatment appears the same number of times in each period. For instance, a design based on TCC and CTT is not uniform within sequences, because C appears twice in one of the sequences and once in the other. A strongly balanced, uniform within sequences and periods design is the one based on the sequences CTTC, TCCT, CCTT, and TTCC.

Period			
1	2	3	4
C	T	T	C
T	C	C	T
C	C	T	T
T	T	C	C

In the following digression let us show that if there is any carryover effect from a previous treatment, this effect is eliminated in the analysis. The observation model is

$$y = \mu + \alpha + \gamma^{(order)} + \gamma^{(period)} + \lambda^{(carryover)} + \varepsilon$$

where  $\mu$  is the overall mean,  $\alpha$  is the treatment effect in which we are interested in,  $\gamma^{(order)}$  is a possible effect due to the ordering of the treatments,  $\gamma^{(period)}$  is a possible effect due to the period,  $\lambda^{(carryover)}$  is a carryover from one period to the next, and, as usual,  $\varepsilon$  is the residual. The following table shows the decomposition of each of the cells according to this model (we omit the overall mean and residuals for simplicity

of the notation). We are modelling only 1st order carryover effects, that is, carryover effects from the immediately previous treatment in the sequence. 2nd order carryover effects would analyze the previous two treatments.

Sequence	Period 1	Period 2	Period 3	Period 4
CTTC	$\alpha_C + \gamma_{CTTC} + \gamma_1$	$\alpha_T + \gamma_{CTTC} + \gamma_2 + \lambda_C$	$\alpha_T + \gamma_{CTTC} + \gamma_3 + \lambda_T$	$\alpha_C + \gamma_{CTTC} + \gamma_4 + \lambda_T$
TCCT	$\alpha_T + \gamma_{TCCT} + \gamma_1$	$\alpha_C + \gamma_{TCCT} + \gamma_2 + \lambda_T$	$\alpha_C + \gamma_{TCCT} + \gamma_3 + \lambda_C$	$\alpha_T + \gamma_{TCCT} + \gamma_4 + \lambda_C$
CCTT	$\alpha_C + \gamma_{CCTT} + \gamma_1$	$\alpha_C + \gamma_{CCTT} + \gamma_2 + \lambda_C$	$\alpha_T + \gamma_{CCTT} + \gamma_3 + \lambda_C$	$\alpha_T + \gamma_{CCTT} + \gamma_4 + \lambda_T$
TTCC	$\alpha_T + \gamma_{TTCC} + \gamma_1$	$\alpha_T + \gamma_{TTCC} + \gamma_2 + \lambda_T$	$\alpha_C + \gamma_{TTCC} + \gamma_3 + \lambda_T$	$\alpha_C + \gamma_{TTCC} + \gamma_4 + \lambda_C$

We can estimate the main effect of the treatment as the average of all those cells having received a T, in the following equation the subscripts indicate the order and the period, and  $\bar{y}$  is the average of all individuals in that order and period

$$\begin{aligned}
 y_T &= \frac{1}{8}(\bar{y}_{CTTC,2} + \bar{y}_{CTTC,3} + \bar{y}_{TCCT,1} + \bar{y}_{TCCT,4} + \bar{y}_{CCTT,3} + \bar{y}_{CCTT,4} + \bar{y}_{TTCC,1} + \bar{y}_{TTCC,2}) \\
 &= \frac{1}{8}[(\mu + \alpha_T + \gamma_{CTTC} + \gamma_2 + \lambda_C) + (\mu + \alpha_T + \gamma_{CTTC} + \gamma_3 + \lambda_T) \\
 &\quad (\mu + \alpha_T + \gamma_{TCCT} + \gamma_1) + (\mu + \alpha_T + \gamma_{TCCT} + \gamma_4 + \lambda_C) \\
 &\quad (\mu + \alpha_T + \gamma_{CCTT} + \gamma_3 + \lambda_C) + (\mu + \alpha_T + \gamma_{CCTT} + \gamma_4 + \lambda_T) \\
 &\quad (\mu + \alpha_T + \gamma_{TTCC} + \gamma_1) + (\mu + \alpha_T + \gamma_{TTCC} + \gamma_2 + \lambda_T)] \\
 &= \mu + \alpha_T + \frac{1}{8}(2\gamma_{CTTC} + 2\gamma_{TCCT} + 2\gamma_{CCTT} + 2\gamma_{TTCC}) \\
 &\quad + \frac{1}{8}(2\gamma_1 + 2\gamma_2 + 2\gamma_3 + 2\gamma_4) + \frac{1}{8}(3\lambda_C + 3\lambda_T)
 \end{aligned}$$

We remember at this moment the ANOVA constraints on the different factors

$$\begin{aligned}
 \gamma_{CTTC} + \gamma_{TCCT} + \gamma_{CCTT} + \gamma_{TTCC} &= 0 \\
 \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 &= 0 \\
 \lambda_C + \lambda_T &= 0
 \end{aligned}$$

Then, we have

$$y_T = \mu + \alpha_T$$

**Important remarks**

133. In cross-over designs, animals are tested more than once reducing the total number of animals needed. Additionally, we can estimate the treatment effects without being affected by the between-animal variability.

134. Thanks to a strongly balanced, uniform within sequences and periods design, we can have estimates of the main effects of the treatments that are unconfounded with the order, period, and carryovers from the immediately previous treatment (1st order).

135. If we relax the constraints of the design (strongly balanced, uniform within sequences and periods), we confound the treatment effects with the order, the period, or carryover effects from the previous treatment.

If we repeat this analysis for this design with second order carryover effects, we see that it is not balanced with respect to them.

### Incomplete cross-over designs

Sometimes, we cannot test all treatments on the same animal, but just a few of them. These are called incomplete cross-over designs. The way to analyze them is exactly the same as exposed for the complete cross-over designs above. This kind of designs was discussed in [Afsarinejad \(1983\)](#) and the following example gives a glimpse of their possible use.

- Example 123: We are testing the effect of three different doses of a compound ( $D_1$ ,  $D_2$  and  $D_3$ ) plus a control dose ( $C$ ) without the compound. Each animal will receive a sequence of two doses. Since the number of treatments (four) is larger than the sequence (two), each animal will not receive the full set of treatments, and the design is incomplete. A possible design could be:

Animal	Test period 1	Test period 2
Animal 1	$C$	$D_1$
Animal 2	$D_1$	$C$
Animal 3	$C$	$D_2$
Animal 4	$D_2$	$C$
Animal 5	$C$	$D_3$
Animal 6	$D_3$	$C$
Animal 7	$D_1$	$D_2$
Animal 8	$D_2$	$D_1$
Animal 9	$D_1$	$D_3$
Animal 10	$D_3$	$D_1$
Animal 11	$D_2$	$D_3$
Animal 12	$D_3$	$D_2$

### 5.2.4 $2^k$ Factorial designs

**Design summary.** This is a standard factorial design in which the effect of multiple variables, and their possible interactions, are studied at the same time. The characteristic of this design is that each variable only has two levels. For each combination of the different treatments we will assume that  $N$  animals are studied.

In this section we will study a very common particular case of factorial design in which all factors have only two levels (yes/no, absent/present, ...). If we have  $k$  factors, the total number of treatments is  $2^k$ . This kind of designs can be analyzed in the standard way introduced in Sec. 5.1. However, we will introduce a new notation for the analysis of these designs that will help us later to perform *fractional factorial designs*, designs in which not all the combinations are tested.

- Example 124: We want to know the optimal way of reducing conflicts between animals in cages. For each combination, we will measure the average number of daily conflicts, and we will run our experiment for ten days. We are interested in

the effect of three factors related to the animals in the cage: sex ( $P$ ), age ( $Q$ ), and number ( $R$ ). For each of the factors we have two levels which we will encode as 0 or 1:

- Sex ( $P$ ): all animals are of the same sex (0) or different sex (1).
- Age ( $Q$ ): all animals are within a range of three months (0) or the age difference is larger than three months (1).
- Number ( $R$ ): two animals per cage (0) or four animals per cage (1).

For every treatment we will have two cages of that kind so that we have three observations per treatment. We can arrange the observations as

Sex ( $P$ )	Age ( $Q$ )	Number ( $R$ )	Observations		
0	0	0	$y_{0001}$	$y_{0002}$	$y_{0003}$
0	0	1	$y_{0011}$	$y_{0012}$	$y_{0013}$
0	1	0	$y_{0101}$	$y_{0102}$	$y_{0103}$
0	1	1	$y_{0111}$	$y_{0112}$	$y_{0113}$
1	0	0	$y_{1001}$	$y_{1002}$	$y_{1003}$
1	0	1	$y_{1011}$	$y_{1012}$	$y_{1013}$
1	1	0	$y_{1101}$	$y_{1102}$	$y_{1103}$
1	1	1	$y_{1111}$	$y_{1112}$	$y_{1113}$

We will consider the full factorial model with all interactions

$$y = \mu + \alpha_P + \alpha_Q + \alpha_R + \alpha_{PQ} + \alpha_{PR} + \alpha_{QR} + \alpha_{PQR} + \varepsilon \quad (5.22)$$

The table of degrees of freedom is

Source	$df$
$P$	1
$Q$	1
$R$	1
$PQ$	1
$PR$	1
$QR$	1
$PQR$	1
Residuals	16
Total	23

An interesting feature of  $2^k$  factorial designs (in the example  $k = 3$  because we have three factors of interest) is that all model parameters cost only one degree of freedom due to the constraints imposed by linear models. In the following paragraphs we will develop an alternative way of estimating the model parameters. At the end, they will produce similar estimates to the standard approach through the  $\alpha$  parameters, but the intermediate variables used for the analysis will be different.

Let us call:

- (1) the average of all observations with no treatment applied ( $P = Q = R = 0$ ).
- $p$  the average of all observations receiving the treatment  $P = 1, Q = 0, R = 0$  and  $\hat{P}$  the effect size of applying  $P = 1$ , this was  $\hat{\alpha}_p$  in our previous notation.
- $pq$  the average of all observations receiving the treatments  $P = 1, Q = 1, R = 0$  and  $\widehat{PQ}$  the effect size of applying  $P = 1, Q = 1$ , this was  $\hat{\alpha}_{pQ}$  in our previous notation.
- ...

In this way, we can write the different means as shown in the following table. It is constructed by placing + if the treatment is applied to construct that mean and - if the treatment is not applied.

Average in the new notation	$P$	$Q$	$R$	Average in the previous notation
(1)	-	-	-	$y_{000}$ .
$r$	-	-	+	$y_{001}$ .
$q$	-	+	-	$y_{010}$ .
$qr$	-	+	+	$y_{011}$ .
$p$	+	-	-	$y_{100}$ .
$pr$	+	-	+	$y_{101}$ .
$pq$	+	+	-	$y_{110}$ .
$pqr$	+	+	+	$y_{111}$ .

We may estimate the effect size of  $P = 1$  as the difference between those animals having received  $P = 1$  and those animals not having received it

$$\hat{P} = p - (1)$$

That is, the mean of the animals receiving  $P = 1, Q = 0, R = 0$  and the animals receiving  $P = 0, Q = 0, R = 0$ . However, this is not the only way to estimate  $\hat{P}$ , we could have also estimated it in the following ways

$$\begin{aligned} \hat{P} &= \frac{p+pq}{2} - \frac{1+q}{2} = \frac{1}{2}(p-1)(1+q) \\ \hat{P} &= \frac{p+pq+pr}{3} - \frac{1+q+r}{3} = \frac{1}{3}(p-1)(1+q+r) \\ \hat{P} &= \frac{pq+pr+pqr}{3} - \frac{q+r+qr}{3} = \frac{1}{3}(p-1)(q+r+qr) \end{aligned}$$

The estimator with least variance is the one that utilizes all the samples available

$$\hat{P} = \frac{p+pq+pr+pqr}{4} - \frac{1+q+r+qr}{4} = \frac{1}{4}(p-1)(1+q+r+qr)$$

We may extend the sign table above to construct a recipe to calculate any of the effects. We extend the table by adding columns corresponding to the overall mean ( $\hat{\mu}$ ), the second order interactions ( $\widehat{PQ}$ ,  $\widehat{PR}$  and  $\widehat{QR}$ ) and third order interactions ( $\widehat{PQR}$ ).  $\hat{\mu}$  is simply all +, the second and third order interactions are calculated as the product of the corresponding columns (i.e.,  $\widehat{PQ}$  is the product of column  $\hat{P}$  times column  $\hat{Q}$ )

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R}$	$\widehat{PQ}$	$\widehat{PR}$	$\widehat{QR}$	$\widehat{PQR}$
(1)	+	-	-	-	+	+	+	-
$r$	+	-	-	+	+	-	-	+
$q$	+	-	+	-	-	+	-	+
$qr$	+	-	+	+	-	-	+	-
$p$	+	+	-	-	-	-	+	+
$pr$	+	+	-	+	-	+	-	-
$pq$	+	+	+	-	+	-	-	-
$pqr$	+	+	+	+	+	+	+	+

Each column gives the signs of the average for estimating that effect. For instance, to estimate  $\widehat{PQR}$ , the least variance estimator is given by

$$\widehat{PQR} = \frac{p + q + r + pqr}{4} - \frac{1 + qr + pr + pq}{4} = \frac{1}{4}(p - 1)(q - 1)(r - 1)$$

In general, the estimation formula for any of the effects is of the form

$$\widehat{effect} = \frac{1}{2^{k-1}}(p \pm 1)(q \pm 1)(r \pm 1)$$

where  $k$  is the number of factors ( $k = 3$  with our example of three factors,  $P$ ,  $Q$ , and  $R$ ), and the sign depends on whether the corresponding variable ( $p$ ,  $q$ , or  $r$ ) is on the effect we want to estimate (-) or not (+). In this way, we have

$$\begin{aligned} \hat{P} &= \frac{1}{4}(p - 1)(q + 1)(r + 1) \\ \hat{Q} &= \frac{1}{4}(p + 1)(q - 1)(r + 1) \\ \hat{R} &= \frac{1}{4}(p + 1)(q + 1)(r - 1) \\ \widehat{PQ} &= \frac{1}{4}(p - 1)(q - 1)(r + 1) \\ \widehat{QR} &= \frac{1}{4}(p + 1)(q - 1)(r - 1) \\ \widehat{PR} &= \frac{1}{4}(p - 1)(q + 1)(r - 1) \\ \widehat{PQR} &= \frac{1}{4}(p - 1)(q - 1)(r - 1) \\ \hat{\mu} &= \frac{1}{8}(p + 1)(q + 1)(r + 1) \end{aligned}$$

The mean is different only in the sense that the denominator is  $2^k$  instead of  $2^{k-1}$ . Interestingly, the variance of all these effects are identical and they take the value

$$\text{Var}\{\widehat{effect}\} = \frac{\sigma_\epsilon^2}{2^{k-2}N} \tag{5.23}$$

where  $N$  is the number of individuals per group and  $\sigma_\epsilon^2$  the variance of the residuals. The sum of squares explained by this effect is

$$SS_{effect} = N2^{k-2}\widehat{effect}^2 \tag{5.24}$$

Once we have the effect estimates we can construct the prediction for the full factorial model. Given these effect estimates we can model the value for a sample  $y_{ijkl}$ , where  $i$ ,  $j$  and  $k$  take the values 0 or 1 depending on the treatment applied for the factors  $P$ ,  $Q$

and  $R$ , and  $l$  takes the values 1, 2, ...,  $N$  implying that there are  $N$  individuals per  $P$ ,  $Q$  and  $R$  combination. The model decomposes the observation as

$$\begin{aligned}
 y_{ijkl} = & \hat{\mu} && \text{overall mean} \\
 & + \frac{1}{2} \left( (-1)^{i-1} \hat{P} + (-1)^{j-1} \hat{Q} + (-1)^{k-1} \hat{R} \right) && \text{main effects} \\
 & + \frac{1}{2} \left( (-1)^{i+j-2} \widehat{PQ} + (-1)^{i+k-2} \widehat{PR} + (-1)^{j+k-2} \widehat{QR} \right) && \text{2nd order interactions} \\
 & + \frac{1}{2} \left( (-1)^{i+j+k-3} \widehat{PQR} \right) && \text{3rd order interactions} \\
 & + \varepsilon_{ijkl} && \text{residual}
 \end{aligned} \tag{5.25}$$

This is totally consistent with the standard ANOVA decomposition in Eq. 5.22. Additionally, we already have ways to estimate the main effects ( $\alpha_P = \frac{1}{2}(-1)^{i-1}\hat{P}$ , ...), second order interactions ( $\alpha_{PQ} = \frac{1}{2}(-1)^{i+j-2}\widehat{PQ}$ , ...) and third order interactions ( $\alpha_{PQR} = \frac{1}{2}(-1)^{i+j+k-3}\widehat{PQR}$ ). For instance, in Example 124, the expected value of the observations of the average daily conflicts of animals of the same sex ( $P = 0$ ), with age differences of more than three months ( $Q = 1$ ) and four animals per cage ( $R = 1$ ) is

$$\begin{aligned}
 \mathbb{E}\{y_{011l}\} &= \hat{\mu} \\
 &+ \frac{1}{2} \left( (-1)^{0-1} \hat{P} + (-1)^{1-1} \hat{Q} + (-1)^{1-1} \hat{R} \right) \\
 &+ \frac{1}{2} \left( (-1)^{0+1-2} \widehat{PQ} + (-1)^{0+1-2} \widehat{PR} + (-1)^{1+1-2} \widehat{QR} \right) \\
 &+ \frac{1}{2} \left( (-1)^{0+1+1-3} \widehat{PQR} \right) \\
 &= \hat{\mu} - \frac{1}{2} \hat{P} + \frac{1}{2} \hat{Q} + \frac{1}{2} \hat{R} - \frac{1}{2} \widehat{PQ} - \frac{1}{2} \widehat{PR} + \frac{1}{2} \widehat{QR} - \frac{1}{2} \widehat{PQR}
 \end{aligned}$$

Similarly, the expected value of observations of animals of the same sex ( $P = 0$ ), with age differences of more than three months ( $Q = 1$ ) and two animals per cage ( $R = 0$ ) is

$$\mathbb{E}\{y_{010l}\} = \hat{\mu} - \frac{1}{2} \hat{P} + \frac{1}{2} \hat{Q} - \frac{1}{2} \hat{R} - \frac{1}{2} \widehat{PQ} + \frac{1}{2} \widehat{PR} - \frac{1}{2} \widehat{QR} + \frac{1}{2} \widehat{PQR}$$

- **Example 125:** If we want to check if grouping animals of the same sex with age differences more than three months in groups of two or four makes a difference, we will have to construct the difference between these two expected values. This is called a *contrast*

$$\begin{aligned}
 c &= \mathbb{E}\{y_{011l}\} - \mathbb{E}\{y_{010l}\} \\
 &= \left( \hat{\mu} - \frac{1}{2} \hat{P} + \frac{1}{2} \hat{Q} + \frac{1}{2} \hat{R} - \frac{1}{2} \widehat{PQ} - \frac{1}{2} \widehat{PR} + \frac{1}{2} \widehat{QR} - \frac{1}{2} \widehat{PQR} \right) \\
 &= - \left( \hat{\mu} - \frac{1}{2} \hat{P} + \frac{1}{2} \hat{Q} - \frac{1}{2} \hat{R} - \frac{1}{2} \widehat{PQ} + \frac{1}{2} \widehat{PR} - \frac{1}{2} \widehat{QR} + \frac{1}{2} \widehat{PQR} \right) \\
 &= \hat{R} - \widehat{PR} + \widehat{QR} - \widehat{PQR}
 \end{aligned} \tag{5.26}$$

Remember that each of these terms has a variance given by Eq. 5.23. As we have four of these terms, the variance associated to  $c$  is

$$\text{Var}\{c\} = 4 \frac{\sigma_\varepsilon^2}{2N}$$

However, if our model does not consider third order interactions, then the term  $\widehat{PQR}$  disappears from the contrast (Eq. 5.26), and the associated variance is

$$\text{Var}\{c\} = 3 \frac{\sigma_{\varepsilon}^2}{2N}$$

If we do not model second order effects, then the  $\widehat{PR}$  and  $\widehat{QR}$  terms also disappear, and the variance of  $c$  reduces to

$$\text{Var}\{c\} = \frac{\sigma_{\varepsilon}^2}{2N}$$

This change of variance has a real effect on the analysis of the data. Depending on our model, the same data analyzed with the three models (full factorial, main effects and second order interactions, or only main effects) rejects the hypothesis that the contrast above is significant or not as shown in the table below

Model	$c$	$\sqrt{\text{Var}\{c\}}$	95% Confidence interval
Full model	6.7	5.9	(-6.3, 20.3)
Main+2nd order	10.0	5.1	(-1.8, 21.8)
Main	7.4	2.9	(0.6, 14.2)

In the case of the model with only main effects, there is a significant difference between grouping animals in groups of two or four. However, this difference is declared as non-significant with the other two models.

#### Important remarks

136. Choosing a model for the observations has important consequences on the statistical power of the analysis. If we foresee 2nd, 3rd, ... order analysis, factorial designs allow estimating all of these interactions. However, if we do not foresee these interactions, choosing an overcomplex model decreases our statistical power, which is our capacity to recognize significant effects.
137. Interactions whose order is larger than two are normally not expected. But, obviously, this depends on the specific system being studied.
138. We should choose the model (main effects, main effects plus second order interactions, ..., full factorial) before observing the experimental data. We cannot take the decision after seeing the data, this is called *data snooping* and it constitutes a severe flaw of the analysis.

### 5.2.5 $2^k$ Fractional factorial designs

**Design summary.** For a large number of two-level factors, we may reduce the number of experiments (we only perform a fraction of the experiments in the full

factorial design), if we sacrifice the estimation of high order interactions.

- **Example 126:** We are studying the effect of eleven different factors on the spread of spinal anesthesia. The factors of interest are: 1) baricity of the anesthetic solution, 2) drug dosage, 3) temperature of the solution, 4) viscosity of the solution, 5) animal positioning during injection, 6) animal positioning after injection, 7) site of injection, 8) needle type, 9) needle direction, 10) intrathecal catheters, 11) intraabdominal pressure. For each of the factors we have two levels.

The full factorial design implies  $2^{11} = 2048$  experiments. However, we can strongly reduce this number if we do not foresee high order interactions. As we saw in the previous section, the number of degrees of freedom consumed to estimate the main effects of a  $2^k$  factorial design is 1 per factor, and the same for the second order interactions. For  $k$  factors, the number of second order interactions is given by the number combinations of  $k$  elements taken in groups of 2

$$df_{2\text{nd order}} = C(k, 2) = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}$$

For  $k = 9$  factors, we need 36 degrees of freedom.

For simplicity of the presentation of the theoretical ideas involved, let us first present a fractional factorial design in which we will only perform half the number of experiments of the full factorial design with only  $k = 3$  factors. As we saw in the previous section, the full factorial design would estimate all the second order and third order interactions according to the following table

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R}$	$\widehat{PQ}$	$\widehat{PR}$	$\widehat{QR}$	$\widehat{PQR}$
(1) = $y_{000}$ .	+	-	-	-	+	+	+	-
$r = y_{001}$ .	+	-	-	+	+	-	-	+
$q = y_{010}$ .	+	-	+	-	-	+	-	+
$qr = y_{011}$ .	+	-	+	+	-	-	+	-
$p = y_{100}$ .	+	+	-	-	-	-	+	+
$pr = y_{101}$ .	+	+	-	+	-	+	-	-
$pq = y_{110}$ .	+	+	+	-	+	-	-	-
$pqr = y_{111}$ .	+	+	+	+	+	+	+	+

Let us assume that we only perform one half of the experiments, those indicated in the following table

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R}$	$\widehat{PQ}$	$\widehat{PR}$	$\widehat{QR}$	$\widehat{PQR}$
$r = y_{001}$ .	+	-	-	+	+	-	-	+
$q = y_{010}$ .	+	-	+	-	-	+	-	+
$p = y_{100}$ .	+	+	-	-	-	-	+	+
$pqr = y_{111}$ .	+	+	+	+	+	+	+	+

That is we will try the treatments  $P = 0, Q = 0, R = 1, P = 0, Q = 1, R = 0, P = 1, Q = 0, R = 0$ , and  $P = 1, Q = 1, R = 1$ . There are four treatment combinations that we will

not test. Remember that the signs in the columns of the different model parameters give the estimation formula for that parameters. For instance, the estimate of  $\hat{\mu}$  is

$$\hat{\mu} = \frac{1}{4}(r + q + p + pqr)$$

But, this is exactly the same estimation formula for  $\widehat{PQR}$ . This means that we are estimating the addition of these two quantities at the same time

$$\widehat{\mu + PQR} = \frac{1}{4}(r + q + p + pqr)$$

The two quantities,  $\hat{\mu}$  and  $\widehat{PQR}$ , are said to be confounded, we cannot tell from the average  $\frac{1}{4}(r + q + p + pqr)$  which part corresponds to  $\hat{\mu}$  and which part corresponds to  $\widehat{PQR}$ . We have confounded the zero order parameter with the third order interactions, but we did not expect third order interactions, so we expect this latter contribution to be zero. The same happens with the main effects and the second order interactions:  $\hat{P}$  is confounded with  $\widehat{QR}$ ,  $\hat{Q}$  is confounded with  $\widehat{PR}$ , and  $\hat{R}$  is confounded with  $\widehat{PQ}$ . Confounding is also called aliasing and, in this example, it is said that the main effects are aliased with the second order interactions, and that the overall mean is aliased with the third order interactions.

**Important remarks**

139. Confounding is not a problem if we do not expect high order interactions since they will be zero. The problem comes if we do not expect them, but they are not zero in reality. In this case, they bias the estimate of the low order terms.

Not any fraction of the full factorial design makes experimental sense. For instance, if we only perform the upper half of the table,

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R}$	$\widehat{PQ}$	$\widehat{PR}$	$\widehat{QR}$	$\widehat{PQR}$
(1) = $y_{000}$ .	+	-	-	-	+	+	+	-
$r = y_{001}$ .	+	-	-	+	+	-	-	+
$q = y_{010}$ .	+	-	+	-	-	+	-	+
$qr = y_{011}$ .	+	-	+	+	-	-	+	-

then, the main effects of  $P$  are confounded with the overall mean (note that two columns are aliased if their signs are exactly the same or exactly the opposite). But these two quantities are supposed to be important and not negligible, thus invalidating our design.

Fractional designs are not unique. The following design also reduces the number of experiments by one half and has the same confusion pattern as the previous one (main effects are aliased with the second order interactions, and the overall mean is aliased with the third order interactions).

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R}$	$\widehat{PQ}$	$\widehat{PR}$	$\widehat{QR}$	$\widehat{PQR}$
(1) = $y_{000}$ .	+	-	-	-	+	+	+	-
$qr = y_{011}$ .	+	-	+	+	-	-	+	-
$pr = y_{101}$ .	+	+	-	+	-	+	-	-
$pq = y_{110}$ .	+	+	+	-	+	-	-	-

The previous two fractional designs are symmetric in the sense that all main effects are aliased with second order interactions. But we may favour one of the factors, if this factor is more important for us. For instance, in the following design  $P$  is confounded with a third order interaction (which is supposed to be smaller than second order interactions). In this case, it is the overall mean which is slightly sacrificed by confounding it with a second order interaction ( $\widehat{QR}$ ).

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R}$	$\widehat{PQ}$	$\widehat{PR}$	$\widehat{QR}$	$\widehat{PQR}$
$r = y_{001}$ .	+	-	-	+	+	-	-	+
$q = y_{010}$ .	+	-	+	-	-	+	-	+
$pr = y_{101}$ .	+	+	-	+	-	+	-	-
$pq = y_{110}$ .	+	+	+	-	+	-	-	-

#### Important remarks

140. Not all fractions of the full factorial are valid. Some of them confound the overall mean with the main effects, or main effects among themselves.
141. Fractional factorial designs are not unique. There exist several of them with similar aliasing properties, and we may even find fractions that favour one of the factors with respect to the rest.
142. The previous example is called a  $2^{k-1}$  fractional design because we have performed only one half  $1/2 = 2^{-1}$  of a  $2^k$  full factorial design.

In the previous paragraphs we have introduced  $2^{k-1}$  fractional designs by removing experiments from the full factorial design of  $k$  factors. But we can also construct a fractional design by taking the full factorial design of  $k - 1$  factors and adding an extra factor. In the former procedure (removing experiments) we were able to analyze which effects were aliased with which other effects, but other than intelligently removing experiments, we could not decide which would be the confusion pattern. In the second procedure (directly starting from the  $k - 1$  full factorial design) we will be able to decide which effects will be confounded with which other effects. For simplicity, let us also work with the  $k = 3$  factors. We would start from the full factorial design with 2 factors

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\widehat{PQ}$
(1) = $y_{00}$ .	+	-	-	+
$q = y_{01}$ .	+	-	+	-
$p = y_{10}$ .	+	+	-	-
$pq = y_{11}$ .	+	+	+	+

Now, we decide to confound the new factor  $R$  with  $PQ$ , which is written as

$$\hat{R} \equiv \widehat{PQ}$$

Then, the sign table is

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R} \equiv \widehat{PQ}$
(1) = $y_{00}$ .	+	-	-	+
$q = y_{01}$ .	+	-	+	-
$p = y_{10}$ .	+	+	-	-
$pq = y_{11}$ .	+	+	+	+

Finally, we extend the table by constructing the new columns  $\widehat{PR}$ ,  $\widehat{QR}$  and  $\widehat{PQR}$ , in the new table we already annotate the columns with which they are confounded.

Average	$\hat{\mu}$	$\hat{P}$	$\hat{Q}$	$\hat{R} \equiv \widehat{PQ}$	$\widehat{PR} \equiv \hat{Q}$	$\widehat{QR} \equiv \hat{P}$	$\widehat{PQR} \equiv \hat{\mu}$
(1) = $y_{00}$ .	+	-	-	+	-	-	+
$q = y_{01}$ .	+	-	+	-	+	-	+
$p = y_{10}$ .	+	+	-	-	-	+	+
$pq = y_{11}$ .	+	+	+	+	+	+	+

The treatments that we will test are defined by the signs of the  $P$ ,  $Q$  and  $R$  variables, that is, 001, 010, 100, 111.

We can multiply the both sides of the design equation by the signs of any of the columns involved, obtaining the equivalent design equations (note that the multiplication of a column by itself is unity)

$$\begin{aligned}
 \hat{R} &\equiv \widehat{PQ} \\
 \widehat{QR} &\equiv \hat{P} \\
 \widehat{PR} &\equiv \hat{Q} \\
 \widehat{PQR} &\equiv \hat{\mu}
 \end{aligned}$$

The first equation is our confusion design equation, while all the rest are consequences of this design equation. All of them are equivalent because all of them induce the same aliasing pattern. The last equation above is the canonical form of the design equation. We observe that the word on the left of the canonical form has three variables. This implies that this factorial design is of resolution III. The number of variables of the canonical form defines the resolution of the experiment. The following table shows the capabilities of the different resolutions

Resolution	Example	Capabilities
I	$\hat{P} \equiv \hat{\mu}$	Not useful: an experiment of exactly one run only tests one level of a factor and hence cannot even distinguish the difference between the 0 and 1 levels of the factor.
II	$\widehat{PQ} \equiv \hat{\mu}$	Not useful: main effects are confounded with other main effects.
III	$\widehat{PQR} \equiv \hat{\mu}$	Estimates main effects, but they are confounded with 2nd order interactions.
IV	$\widehat{PQRS} \equiv \hat{\mu}$	Estimates main effects free from 2nd order interactions. 2nd order interactions are confounded with other 2nd order interactions.
V	$\widehat{PQRST} \equiv \hat{\mu}$	Estimates main effects free from 3rd order interactions. 2nd order interactions are confounded with 3rd order interactions.
VI	$\widehat{PQRSTU} \equiv \hat{\mu}$	Estimates main effects free from 4th order interactions. 2nd order interactions are confounded with 4th order interactions. 3rd order interactions are confounded with other 3rd order interactions.

To design a  $2^{k-1}$  fractional factorial design we just need 1 aliasing equation. But, we can reduce further the number of experiments, to reduce it to one fourth of the full factorial ( $2^{k-2}$  fractional factorial design) we need 2 aliasing equations. And, in general, we can reduce the number of experiments of the full factorial by a factor  $2^{-p}$  by choosing  $p$  aliasing equations. Obviously, the  $p$  aliasing equations cannot be equivalent to each other. The resolution of the experiment is given by the smallest word of the  $p$  canonical forms.

- Example 126 (continued): In the example with 11 factors, the full factorial requires  $2^{11}$  runs. However, we can reduce it by a factor 64 and have a Resolution IV factorial design with only 32 ( $= 2^{11-6}$ ) runs. With this design we are able to estimate the main effects free from 2nd order interactions (but they are confounded with 3rd order interactions). Calling the 11 factors from A to K, the 6 aliasing equations are

$$\begin{aligned}
 \widehat{ABCF} &\equiv \hat{\mu} \\
 \widehat{BCDG} &\equiv \hat{\mu} \\
 \widehat{CDEH} &\equiv \hat{\mu} \\
 \widehat{ACDI} &\equiv \hat{\mu} \\
 \widehat{ADEJ} &\equiv \hat{\mu} \\
 \widehat{BDEK} &\equiv \hat{\mu}
 \end{aligned}$$

Note that we cannot reach any of the other equations by multiplying by the factors involved in the aliasing equations. Consequently, they all are independent.

**Important remarks**

143. Using fractional factorial designs we can reduce the cost of an experiment by a factor  $2^{-p}$  if we sacrifice the estimation of high order interactions. This reduction is performed in a controlled way through the design of the aliasing equations. In the example above, the number of experiments was reduced from 2048 to 32. This small experiment is of Resolution IV meaning that we can estimate the main effects free from 2nd order interactions, although they are confounded with 3rd order interactions.
144. Using a theory similar to the Least Squares presented in Sec. 5.1.7, we can even further reduce the number of treatments in the previous example to 26. These designs are called Minimum-Run Resolution IV screening. A Minimum-Run Resolution V design for the same example requires 68 individuals.
145. If we can even neglect 2nd order interactions, we can have Resolution III designs. Plackett-Burman designs is a very well-known possibility to design these experiments. These designs can be used to rapidly identify factors that may have an effect on the outcome of the experiment, and then perform a higher resolution experiment with only those factors. These experiments with very low resolution are also called *Screening designs*. Taguchi designs are also very popular for screening variables in which only the main effects can be estimated.
146. Screening designs focus on a minimum number of treatments. However, we still need to consider how many animals per treatment are necessary taking into account the requirement of having enough degrees of freedom for the residuals so that we have a desired statistical power for a given effect size of the main effects.

- Example 127 (continued): In the example of 11 factors, Plackett-Burman Resolution III designs reduces the number of treatments from 2048 to just 12. In these example, the 12 runs are

Run	A	B	C	D	E	F	G	H	I	J	K
Run 1	0	1	1	1	0	0	0	1	0	1	1
Run 2	1	1	0	1	1	1	0	0	0	1	0
Run 3	0	0	0	0	0	0	0	0	0	0	0
Run 4	0	1	1	0	1	1	1	0	0	0	1
Run 5	1	0	1	1	1	0	0	0	1	0	1
Run 6	0	1	0	1	1	0	1	1	1	0	0
Run 7	1	1	0	0	0	1	0	1	1	0	1
Run 8	0	0	1	0	1	1	0	1	1	1	0
Run 9	0	0	0	1	0	1	1	0	1	1	1
Run 10	1	0	0	0	1	0	1	1	0	1	1
Run 11	1	0	1	1	0	1	1	1	0	0	0
Run 12	1	1	1	0	0	0	1	0	1	1	0

Note that to have statistical power we need at least two animals per treatment combination. With just one animal per treatment, we would not have any degree of freedom available for the residuals.

### 5.2.6 Split-unit designs

**Design summary.** We have an experiment with two factors. One of them requires large experimental units, while the other one small ones. Additionally, the second factor can be applied to a “small portion” of the experimental units of the first factor.

- Example 127: We are investigating the effect of light and diet on the growth of mice.
  - The experimental unit for the light factor is the whole room, all cages receive the same treatment (number of light hours).
  - The experimental unit for the diet is the cage, all mice in the same cage receive the same treatment.

Experiments with repeated measures can be analyzed with this class of designs (in which individuals are considered as fixed effects) or as mixed designs (Sec. 5.2.8, in which individuals are considered as random effects). These experiments include those in which an animal is measured multiple times, at multiple regions of its body, or at multiple tasks. If we follow a specific measurement pattern (like measuring exactly at the same location in the case of multiple regions of the body), we may expect correlations (in this case, spatial correlations) among the different measurements. We may consider randomizing the location of the measurement within a confined region so that we do not always measure at the same place. In all these cases, the animal is the factor “hard to change” and acts as a blocking factor. We may even have several repeated measures factors. For instance, we may have an experiment in which the same animal is measured at multiple times simultaneously at different regions of its body (through multiple sensors, for example).

Let us call  $P$  the factor applied to large units (called whole-units or whole-plots) and  $Q$  the factor applied to small units (called split-units or split-plots). They are also called the between-factor,  $P$ , and the within-factor,  $Q$ . We need to realize that the sample size of the different units (whole and split) are different, and consequently the precision of each one of the factors is different. For completeness of the example, let us assume that we have several blocks in which we will analyze the different treatments (see Fig. 5.7). Each measurement is indexed as  $ijk$  where  $i$  is the block number,  $j$  is the  $P$  treatment and  $k$  is the  $Q$  treatment.

$$y_{ijk} = \mu + (\gamma_i + \alpha_j^{(P)} + \eta_{ij}) + \alpha_k^{(Q)} + \alpha_{jk}^{(PQ)} + \varepsilon_{ijk}$$

This model implies that a measurement is affected by the block,  $\gamma_i$ , the treatment applied to the large unit,  $\alpha_j^{(P)}$ , and some noise affecting at the level of large units,  $\eta_{ij}$ . All these terms in the parenthesis define the contribution of the large unit. Then, we have the contribution of the split-units given by the  $Q$  treatment,  $\alpha_k^{(Q)}$ , the interaction between the factors  $P$  and  $Q$ ,  $\alpha_{jk}^{(PQ)}$ , and some noise at the level of split-unit,  $\varepsilon_{ijk}$ .

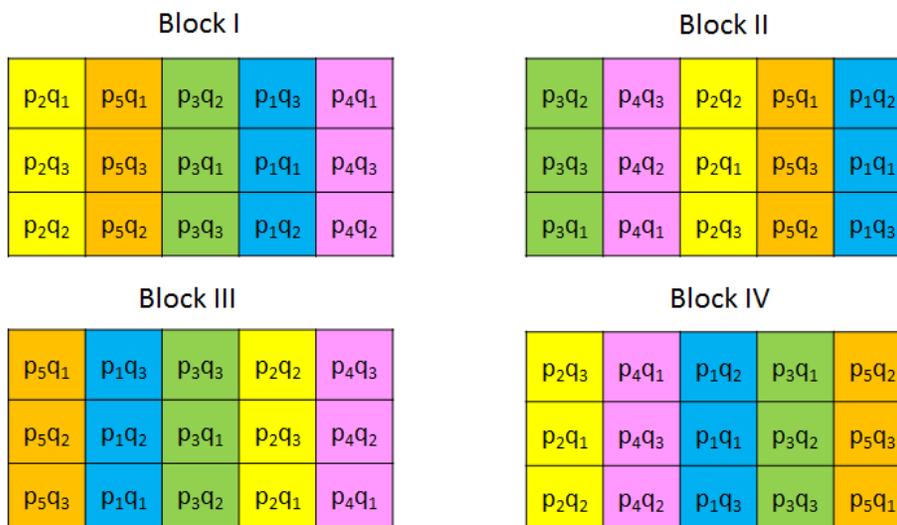


Figure 5.7: Example of split-unit design with four blocks. In each of the blocks we apply five treatments of the “hard-to-change” factor  $P$  (note that all treatments in the same column are the same) and three treatments of the factor  $Q$ . The colour of each cell is given by the  $P$  treatment.

If both treatments require relatively large experimental units, we may apply one of them on the columns, and another one on the rows as shown in Fig. 5.8. These designs are sometimes called *criss-cross designs*.

We may also nest several variables requiring increasingly small experimental units like the design in Fig. 5.9. The design is nested because for each whole-unit of the previous factor, we apply each of the possible treatments of the next factor. Obviously

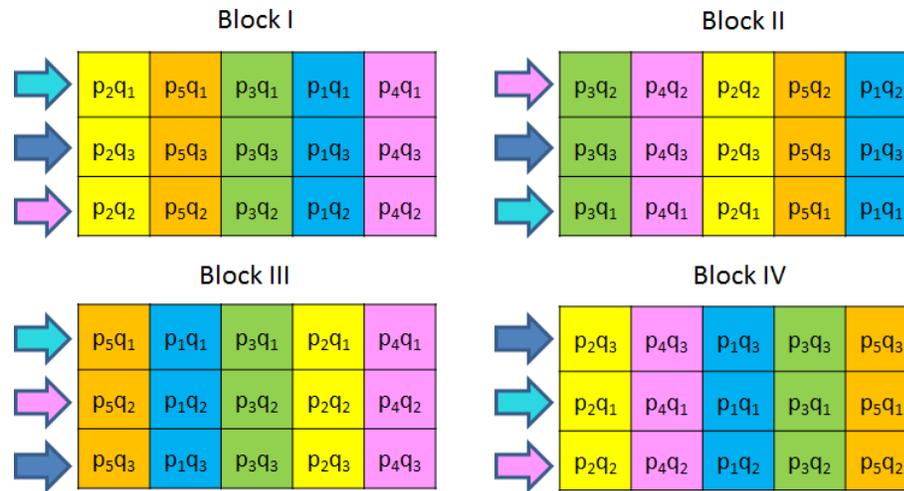


Figure 5.8: Example of split-unit design with four blocks. Both  $P$  and  $Q$  are “hard-to-change” so that they are applied in large units: each column of the design receives the same  $P$  treatment and each row of the design receives the same  $Q$  treatment.

the units of each one of the levels are increasingly small ( $P$  units are larger than  $Q$  units, and these are larger than  $R$  units).

### 5.2.7 Hierarchical or nested designs

**Design summary.** Hierarchical designs are similar to split-unit designs, only that we do not find all possible combination between factors.

- **Example 128:** We are investigating the effect of a drug on the concentration of a given protein in the liver. We have two groups (control and treatment). From every animal we will take several measurements from the liver. Which is a suitable model for this design?

Repeated measures is a design in which the same animal is measured multiple times (as in the example of multiple measures from the liver), and sometimes with a time in between measurements (e.g., 0h, 2h, 6h, 12h, 24h). In this case, the animal becomes the factor hard-to-change, and it cannot receive multiple treatments (for instance, either it is in the control or the treatment group). In these designs, it is important to identify the animal as a block variable because the animal may heavily influence the results and samples coming from the same animal are not independent.

Let us consider the treatment  $i$ , animal  $j$ , and measurement  $k$  within the same animal. Then, the observations can be modeled as

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}$$

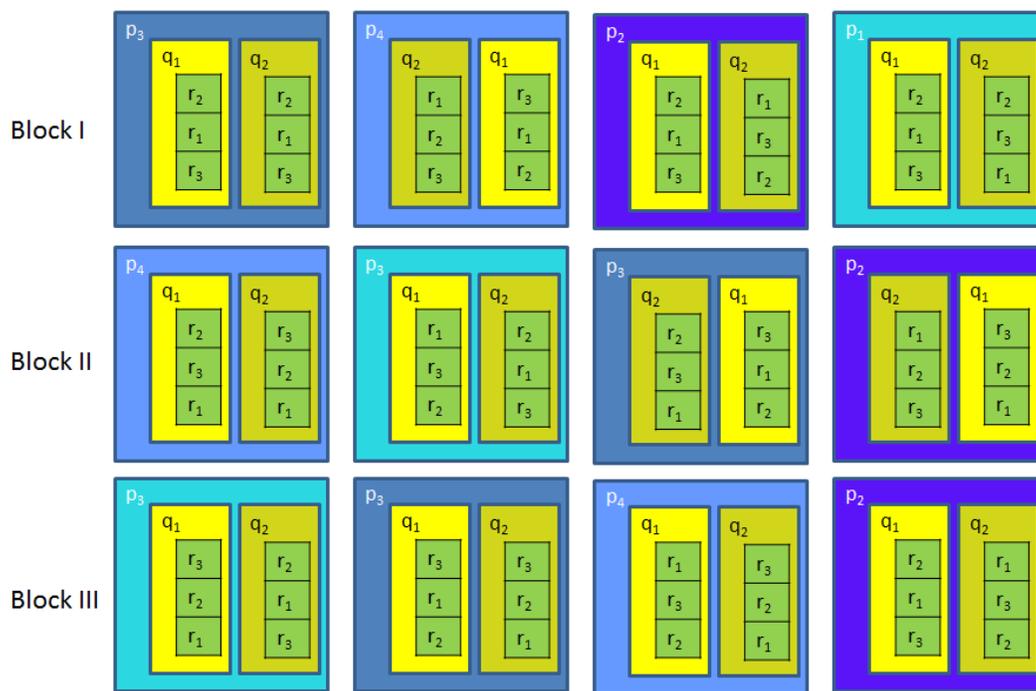


Figure 5.9: Example of nested split-unit design with three blocks, and three factors  $P$  (4 levels),  $Q$  (2 levels), and  $R$  (3 levels).

Note that each animal receives a single treatment so that inside a given  $j$  we can only find one value of  $\alpha$ . For this reason, it is said that the design is hierarchical or nested (the treatment is nested within the animal). In a way the treatment  $i$  is given once we select the animal  $j$ , that is,  $i(j)$ . This is the most distinctive feature of nested designs: the absence of all possible combinations of factors.

We can add more nested variables as shown in the example below.

- **Example 129:** Following with the previous example, the experiment will be carried out by two technicians (A and B), and each animal is analyzed by a single technician. How should we modify the model?

We add an extra factor  $\gamma_l^{(tech)}$ , the specific technician is also specified by the animal,  $l(j)$ , so that now the two variables technician and treatment are nested within the animal. The model would be modified to

$$y_{ijkl} = \mu + \alpha_i + \gamma_j^{(animal)} + \gamma_l^{(tech)} + \varepsilon_{ijkl}$$

It is sometimes useful to make a pictorial representation of the experiment as the one shown in Fig. 5.10. The figure helps to write the analysis equations and identify the different blocks and factors of the experiment. These representations are useful to recognize pathological problems of our design. Let us assume that the technician A performs all the control experiments and the technician B all the experiments with the drug. We would be confounding the effect of the technician with that of the treatment. Visually, we would see in the diagram that each of the levels of the technician corresponds to a single level of the treatment (there is a single line coming out from each of the technicians).

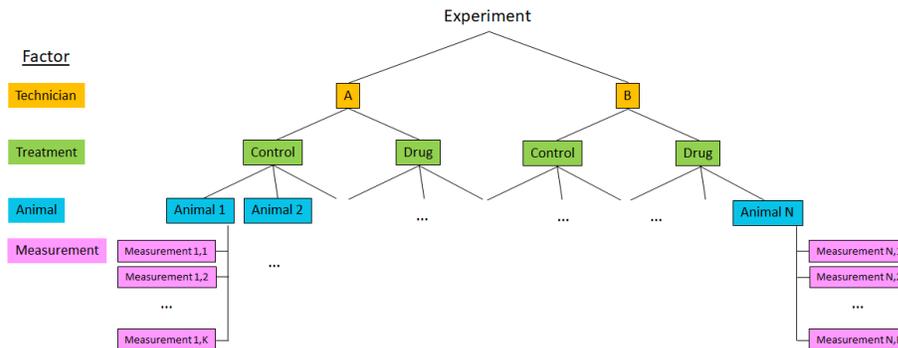


Figure 5.10: Pictorial representation of a nested design.

- **Example 130:** We are interested in the effect of a particular diet on the development of the animal development during pregnancy. For analyzing its effect, we divide the animals in two groups: control (6 cages) and treatment (6 cages). There are 5 pregnant animals per cage, and we will analyze the effect at 5 different time points. At each time point we kill one of the pregnant animals, and

study the development of the fetuses. There are two nested split-units. The first one is the cage as we may suspect that there can be some cage effect. The second one is the animal itself. The fetuses coming from the same animal may have differences due to characteristics of their common mother.

### 5.2.8 Fixed vs. random and mixed effects

**Design summary.** Fixed effects is the most common ANOVA design. In these designs the specific effect of the treatment is of research interest. Random effects refer to treatments in which the specific level of the treatment is supposed to come from a larger population. More than the specific contribution of the treatment, it is more interesting to know about its variability.

All the linear models presented so far are called of *fixed parameters*. The treatment effects were supposed to be unknown, but fixed, and all our efforts have concentrated in designing experiments capable of estimating these parameters with the least amount of uncertainty. Typical fixed effect factors are the strain of the animal (wildtype vs. transgenic), the age group of the animal (2 months vs. 6 months vs 1 year), time of the experiment, diet, supplier, experimenter performing the observations or operations, etc. However, there are situations in which these parameters are not of so much interest, as the factor levels themselves can be regarded as a sample from a larger population. For instance, in Example 128 an specific animal may be regarded as uninteresting *per se*, it is interesting as a random representative of a wider population of animals. In the same way, in Example 129, the technician itself can be regarded as random representative of a larger population of technicians. We shift the focus from the individual technicians (fixed effects), to the population of all technicians (random effects). We are interested, then, in the variance of the treatments, e.g., what is the variation from technician to technician?

- **Example 131:** We are interested in the inheritance of birth weight. For analyzing the effect of the female on it, we analyze the birth weight of the descendents of  $a = 5$  female mice, mated with different male animals. For each of the females, we measure the birth weight of  $n = 10$  of their offspring. How should we analyze the data?

#### One-way, random effects

With a single factor, the ANOVA model seen so far, called fixed effects, was of the form

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (5.27)$$

where  $y_{ij}$  was the observation for the birth weight of the  $j$ -th descendent of the  $i$ -th mother (the treatments),  $\mu$  was the overall mean,  $\alpha_i$  was the effect of the mother, and  $\varepsilon_{ij}$  the residuals. In fixed effects ANOVA,  $\alpha_i$  was the parameter of interest. This makes sense if the treatment is a drug or control, because it reveals the differences in means between the two groups. However, in this experiment with animals, rather than the difference between the specific 5 animals, it is much more interesting to assume

that these 5 animals are a random sample from a large population of females, and the parameter of interest is the variance of this large population. That is, it is assumed that  $\alpha_i$  has been randomly drawn from a  $N(0, \sigma_A^2)$  distribution, i.e., a Gaussian with zero mean and variance  $\sigma_A^2$ . In random effects ANOVA, the observation model is still the one in Eq. 5.27. However, several interesting properties stem from the change of the nature of  $\alpha_i$  (from deterministic to random) as shown in the following table:

	Fixed effects	Random effects
$\text{Var}\{y_{ij}\}$	$\sigma_\varepsilon^2$	$\sigma_A^2 + \sigma_\varepsilon^2$
$\text{Corr}\{y_{ij}, y_{i'j'}\}$	0	$\begin{cases} 0 & i \neq i' \\ \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\varepsilon^2} & i = i', j \neq j' \\ 1 & i = i', j = j' \end{cases}$

Specifically: 1) the variance of the observations is increased by the treatment variance,  $\sigma_A^2$ ; and 2) there is a correlation of the samples within the same treatment (in our example, the birth weight of all descendents from the same mother are correlated, which should not surprise us), this correlation is higher as the treatment variance grows.

For random effects, balanced designs, in which there are  $n$  observations per treatment, we would construct the ANOVA table in the standard way (note that there would be  $N = an$  animals in total):

Source	SS	df	MS = SS/df	$\mathbb{E}\{MS\}$
A	$SS_A = \sum_{ij} (y_i - y_{..})^2$	$a - 1$	$MS_A$	$n\sigma_A^2 + \sigma_\varepsilon^2$
Residuals	$SS_\varepsilon = \sum_{ij} (y_{ij} - y_i)^2$	$N - a$	$MS_\varepsilon$	$\sigma_\varepsilon^2$
Total	$SS_T = \sum_{ij} (y_{ij} - y_{..})^2$	$N - 1$		

The calculation of the sum of squares for the random effects is performed in exactly the same way as for the fixed effects case (see Sec. 5.1.1). The fourth column is new in this presentation of random effects ANOVA, although it could also have been calculated for fixed effects ANOVA. It shows which is the expected value of the Mean Squares column. This column gives us ways to estimate the model parameters. In particular,

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= MS_\varepsilon \\ \hat{\sigma}_A^2 &= \frac{MS_A - MS_\varepsilon}{n} \end{aligned}$$

This estimate is called the method of the moments, and it has the disadvantage that it may result in negative estimates of  $\sigma_A^2$ , which theoretically is impossible because it is a variance. Alternatively, we might use restricted maximum-likelihood estimates (REML), although the theory of these estimates is out of the scope of this chapter. As a general idea, we should realize that estimating variances is difficult because we need a relatively large sample size to be able to have a reasonable accuracy in the estimation. In this regard, it is better to have many animals with few samples, than a few animals with many samples.

The estimate of the treatment effects can also be performed as in the fixed effect case:

$$\hat{\alpha}_i = y_i. - y_{y.}$$

However, in the fixed effects ANOVA we needed to constrain the main effects to have a unique solution with  $\sum_i \hat{\alpha}_i = 0$ . In the random effects, this constrain is not necessary.

To test whether the treatments are statistically significant, we have to compare them to the residuals, in exactly the same way as we did in fixed effects ANOVA. The hypothesis contrast is now

$$\begin{aligned} H_0 : & \sigma_A^2 = 0 \\ H_a : & \sigma_A^2 > 0 \end{aligned}$$

For this test, we also calculate the  $F$  statistic, as in fixed effects ANOVA,

$$F = \frac{MS_A}{MS_\varepsilon}$$

Under the null hypothesis, this statistic is distributed as a Snedecor's  $F$  with  $a - 1$  and  $N - a$  degrees of freedom.

**Important remarks**

147. Random effects allows estimating the variability of a population, rather than specific means of particular individuals.

**Two-way, random effects**

We may consider in the model the effect of the mother (factor A) and father (factor B) on the birth weight of their offspring. We mate  $a = 5$  mothers with  $b = 5$  fathers, and for each cross we analyze  $n$  youngsters. We consider both factors as random samples from a larger population of mothers and fathers. The observation model of the two-way random effects ANOVA is the same as for the fixed effects

$$y_{ijk} = \mu + \alpha_i^{(A)} + \alpha_j^{(B)} + \alpha_{ij}^{(AB)} + \varepsilon_{ijk} \tag{5.28}$$

However, now the parameters are assumed to be independent from each other and come from Gaussian distributions:  $\alpha_i^{(A)} \sim N(0, \sigma_A^2)$ ,  $\alpha_j^{(B)} \sim N(0, \sigma_B^2)$ , and  $\alpha_{ij}^{(AB)} \sim N(0, \sigma_{AB}^2)$ . Now, the variance and correlation of the samples is modified to

	Fixed effects	Random effects
Var $\{y_{ijk}\}$	$\sigma_\varepsilon^2$	$\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2$
Corr $\{y_{ijk}, y_{i'j'k'}\}$	0	$\begin{cases} 0 & i \neq i', j \neq j' \\ \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2} & i = i', j \neq j' \\ \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2} & i \neq i', j = j' \\ \frac{\sigma_{AB}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2} & i = i', j = j', k \neq k' \\ 1 & i = i', j = j', k = k' \end{cases}$

That is, the correlation between two observations increases as the two samples share more common treatments. Actually, this correlation is the ratio between the sum of the variance of the common components and the total observation variance.

In the same way we did for one-way, random effects ANOVA, we can now calculate the ANOVA table for balanced designs with  $n$  animals per treatment combination as follows (note that there would be  $N = abn$  animals in total):

Source	SS	df	MS = SS/df	$\mathbb{E}\{MS\}$
A	$SS_A = \sum_{ijk} (y_{i..} - y_{...})^2$	$a - 1$	$MS_A$	$bn\sigma_A^2 + n\sigma_{AB}^2 + \sigma_\epsilon^2$
B	$SS_B = \sum_{ijk} (y_{.j.} - y_{...})^2$	$b - 1$	$MS_B$	$an\sigma_B^2 + n\sigma_{AB}^2 + \sigma_\epsilon^2$
AB	$SS_{AB} = \sum_{ijk} (y_{ijk} - y_{i..} - y_{.j.} + y_{...})^2$	$(a - 1)(b - 1)$	$MS_{AB}$	$n\sigma_{AB}^2 + \sigma_\epsilon^2$
Residuals	$SS_\epsilon = \sum_{ijk} (y_{ijk} - y_{ij.})^2$	$N - ab$	$MS_\epsilon$	$\sigma_\epsilon^2$
Total	$SS_T = \sum_{ijk} (y_{ijk} - y_{...})^2$	$N - 1$		

From this table, using the method of moments, we could easily obtain estimates of all the variances

$$\begin{aligned}\hat{\sigma}_\epsilon^2 &= MS_\epsilon \\ \hat{\sigma}_{AB}^2 &= \frac{MS_{AB} - MS_\epsilon}{n} \\ \hat{\sigma}_A^2 &= \frac{MS_A - MS_{AB}}{bn} \\ \hat{\sigma}_B^2 &= \frac{MS_B - MS_{AB}}{an}\end{aligned}$$

To test whether the treatments are statistically significant, we have to compare their  $MS$  to a suitable  $MS$ . In fixed effects and one-way, random effects ANOVA, this suitable reference  $MS$  was given by the residuals. However, in two-way, random effects ANOVA, the situation changes. The reason is that if we inspect the expected variance of the  $MS_A$ , it is

$$bn\sigma_A^2 + n\sigma_{AB}^2 + \sigma_\epsilon^2$$

If we compare it to  $MS_\epsilon$  and it is significantly larger, we do not know if the difference is due to  $\sigma_A^2$  or to  $\sigma_{AB}^2$ . We must find some comparison for which the only difference is caused by  $MS_A$ . This comparison is given by  $MS_{AB}$ , and consequently, the  $F$ -statistic for calculating the significance of  $A$  and  $B$  would be respectively

$$\begin{aligned}F_A &= \frac{MS_A}{MS_{AB}} \sim F_{a-1, (a-1)(b-1)} \\ F_B &= \frac{MS_B}{MS_{AB}} \sim F_{b-1, (a-1)(b-1)}\end{aligned}$$

Following the same strategy, the correct comparison for deciding on the significance of the interaction effects uses the residuals

$$F_{AB} = \frac{MS_{AB}}{MS_\epsilon} \sim F_{(a-1)(b-1), N-ab}$$

- Example 132: We are developing a new spectrophotometer to measure the concentration of some compounds (like triglycerides) in biological samples. We are

interested in the consistency of measurements from day to day and among different machines. We randomly select  $a = 5$  machines and  $b = 5$  days. Every day we measure  $n = 2$  samples per machine. All samples in a single day come from the same serum sample. Should we use a random or fixed effects ANOVA?

In this design we are not interested in the particular performance of a particular machine. We are not interested either on the effect of a particular day. They are simply random samples from a larger population of machines and days. If  $\sigma_{machine}^2 > 0$ , then it would indicate that our manufacturing process produces machines with different measurement properties. If we cannot remove this variability, then each machine needs to be calibrated when it is bought.  $\sigma_{day}^2$  represents the variability in triglyceride concentration among the samples used in the individual days.  $\sigma_{machine,day}^2 > 0$  would indicate that each machine needs to be calibrated on daily basis (just one calibration at the beginning of the machine life is not enough).

### Two-way, mixed effects

- **Example 133:** As in Example 131, we are interested in the inheritance of birth weight. We foresee that the mother has an important effect on the descendent birth weights, and we would also like to simultaneously analyze the effect of  $b = 4$  different nutritional complements given to the mothers. As in the previous example we will analyze  $a = 5$  mothers and  $n = 10$  descendents per mother. Our experiment will include 4 pregnancy cycles of the same mothers.

In this experiment, the mother is a random effect (they are samples from a larger population of mothers), but the nutritional complements are fixed effects (they are not samples from a larger population of nutritional complements, and we are really interested in the effect of *these particular* nutritional complements).

Repeated measures may also be analyzed with this model. The individual is considered to be a random effect, while the treatments are considered as fixed effects. Alternatively, the individual may be considered as fixed effects using the split-unit design (Sec. 5.2.6).

The observation model is

$$y_{ijk} = \mu + \alpha_i^{(A)} + \alpha_j^{(B)} + \alpha_{ij}^{(AB)} + \varepsilon_{ijk} \quad (5.29)$$

As in the two-way, random effects model, we assume  $\alpha_i^{(A)} \sim N(0, \sigma_A^2)$ . But now we assume that  $\alpha_j^{(B)}$  are fixed, deterministic values. The interactions are also supposed to be random and distributed as  $\alpha_{ij}^{(AB)} \sim N(0, \frac{b-1}{b} \sigma_{AB}^2)$ . Always that we have fixed effects, we need to impose some constraints to uniquely determine the model parameters. In the two-way, fixed effects ANOVA model, these constraints were

$$\sum_i \alpha_i^{(A)} = \sum_j \alpha_j^{(B)} = \sum_i \alpha_{ij}^{(AB)} = \sum_j \alpha_{ij}^{(AB)} = 0$$

With random effects, we do not need so many constraints, only

$$\begin{aligned}\sum_j \alpha_j^{(B)} &= \sum_j \alpha_{ij}^{(AB)} = 0 \\ \text{Corr}\{\alpha_{ij}^{(AB)}, \alpha_{ij}^{(AB)}\} &= -\frac{1}{b}\end{aligned}$$

Let us define  $\sigma_B^2 = \frac{1}{b-1} \sum_j (\alpha_j^{(B)})^2$ . As was introduced in the case of random effects, there is a correlation between two observations  $y_{ijk}$  and  $y_{i'j'k'}$  given by

$$\text{Corr}\{y_{ijk}, y_{i'j'k'}\} = \begin{cases} 0 & j \neq j' \\ \frac{\sigma_B^2 - \frac{1}{b} \sigma_{AB}^2}{\sigma_A^2 + \sigma_B^2 + \frac{b-1}{b} \sigma_{AB}^2 + \sigma_\varepsilon^2} & i \neq i', j = j' \\ \frac{\sigma_B^2 + \frac{b-1}{b} \sigma_{AB}^2}{\sigma_A^2 + \sigma_B^2 + \frac{b-1}{b} \sigma_{AB}^2 + \sigma_\varepsilon^2} & i = i', j = j', k \neq k' \\ 1 & i = i', j = j', k = k' \end{cases}$$

The mixed effects ANOVA table for a balanced design with  $n$  observations per treatment combination is exactly the same as the one for random effects except that the expected value of the  $MS_A$  is  $bn\sigma_A^2 + \sigma_\varepsilon^2$ . Consequently, the estimate of the variance of  $A$  is

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_\varepsilon}{bn}$$

and the test is performed with the  $F$  statistic

$$F_A = \frac{MS_A}{MS_\varepsilon} \sim F_{a-1, N-ab}$$

- **Example 134:** We are exploring the effect of wheel exercise in the development of neurons in the brain of mice. We have three levels of exercise (no exercise, moderate and intense). We will study the differences between males and females. For each animal, we will assess the effect by analyzing four histological sections of the brain. Sex and exercise are treated as fixed effects, while the animal and the slides are treated as random effects. If the sections are taken from specific locations (cortex, midbrain, hippocampus, ...) instead of at random, then we could treat them as a fixed effect and determine the effect of exercise in each one of the regions.

### Should we use fixed or random effects?

From a theoretical point of view it is unclear whether we should use fixed or random effects (Clark and Linzer, 2015; Gomes, 2022). On one side, fixed effects yields unbiased estimates of the effects while random effects yields biased results. On the other side, the estimates of fixed effects have larger variance than the ones of random effects. In this way, random effects would tend to be better for small sample sizes, while fixed effects would be better for larger sample sizes. Bias in random effects appears if the treatments of a mixed-effects model are correlated with the random-effects. Another problem with random effects is that practitioners are less familiar with this kind of models. However, they are usually available in many data analysis programs.

### 5.2.9 Multilevel models

In Sec. 1.2 we emphasized the need of the samples being independent. However, some experiments do not meet this requirement and the observations are interrelated through some hierarchical level structure.

- **Example 135:** We want to study the effect of the paternal diet on the weight of their descendents. For doing so, we will study a number of mothers (level 1), several of their litters (level 2), and we will weight the pups in each one of the litters (level 3).

All the statistical techniques studied in this book assume that samples are independent. In the previous example, samples are not independent at levels 2 and 3. At level 2, all the pups from a given pregnancy are affected by the nutritional and health state of the mother at that particular pregnancy. At level 3, the weight of a particular pup is affected by the weight of its siblings (a pup may grow at the womb at the cost of some other pup). In Sec. 1.2 we presented the experimental unit as the minimum amount of experimental material where we can change the treatment. In the previous example would be the mother (level 1), although our measurements come from level 3. In Sec. 1.4.5 we introduced the idea that the simplest way to deal with this data is to average all the observations coming from the same mother and treat the averaged observations (from different mothers) as independent.

Using only the observations at level 1 loses the richness of information given by the many numbers of level 3. On the other hand, using the observations at level 3 as independent violates the assumption of independence of the most common statistical techniques (Student's t-test, ANOVA,  $\chi^2$ , ...) as these observations are correlated. Classical analysis tools will underestimate the variance of the residuals, and underestimate consequently the p-value, resulting in a higher rate of False Positives. Between these two extremes (working at level 1 or level 3), there is a third option that is to model the dependence of the observations at the various levels. This is done through the so-called multilevel models (Simsek and Firat, 2011; Pearl, 2014; Aarts et al, 2015).

To simplify the presentation of the technique let us consider a two-level model.

- **Example 136:** We want to study the effect of the injection to the animals (level 1) of a growth factor on the growth of neurites (level 2). For doing so, we will isolate neurons from the animals and measure the length of its neurites.

Let us start modelling the observations of the first level. For simplicity, we will start with a 0-th order model, but later we will expand it to more coefficients. The observation model for the second level is

$$y_{ij}^2 = \mu_j^1 + \varepsilon_{ij}^2 \quad (5.30)$$

where  $j$  denotes the cluster  $j$  (the neuron, level 1),  $i$  refers to the length of the  $i$ -th neurite,  $\mu_j^1$  is the average of the  $j$ -th cluster, and  $\varepsilon_{ij}^2 \sim N(0, \sigma_\varepsilon^2)$ . The cluster mean is modelled as

$$\mu_j^1 = \mu + \alpha_j^1 \quad (5.31)$$

where  $\alpha_j^1$  is a random factor that follows a Gaussian,  $N(0, \sigma_{\alpha^1}^2)$ . We may combine both models into a single equation

$$y_{ij}^2 = (\mu + \alpha_j^1) + \varepsilon_{ij}^2 \quad (5.32)$$

A measure of the dependency of the measurements at level 2 is given by the intraclass correlation, ICC

$$ICC = \frac{\sigma_{\alpha^1}^2}{\sigma_{\alpha^1}^2 + \sigma_{\varepsilon}^2} \quad (5.33)$$

If  $\sigma_{\alpha^1}^2 \gg \sigma_{\varepsilon}^2$ , then most variation between the neurite lengths is caused by the difference between neurons. If  $\sigma_{\alpha^1}^2 \ll \sigma_{\varepsilon}^2$ , then differences between neurons are relatively small compared to the differences between the neurites within the same neuron.

The model above does not account for the injection of the growth factor yet. Let us define a new variable  $\alpha_j^G$  that accounts for the effect of the growth factor received by neuron  $j$ . As usual, the sum over all levels of  $\alpha^G$  must be zero. Then, the observation model becomes

$$y_{ij}^2 = (\mu + \alpha_j^1) + \alpha_j^G + \varepsilon_{ij}^2 \quad (5.34)$$

We can extend this idea not only to the presence or absence of the growth factor, but to the amount of the growth factor, turning the model from an ANOVA-like to a regression-like structure.

$$y_{ij}^2 = (\mu + \alpha_j^1) + (\beta_1^2 + \beta_{1j}^1)G_j + \varepsilon_{ij}^2 \quad (5.35)$$

where  $G_j$  is the dose of growth factor injected into the animal,  $\beta_1^2$  accounts for a general, fixed effect of the growth factor, and  $\beta_{1j}^1$  is the particular effect of the growth factor on the  $j$ -th neuron. Both, the 0th and 1st order effect on the  $j$ -th neuron are supposed to be random factors

$$\begin{pmatrix} \alpha_j^1 \\ \beta_{1j}^1 \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha^1}^2 & \sigma_{\alpha^1, \beta_1^1} \\ \sigma_{\alpha^1, \beta_1^1} & \sigma_{\beta_1^1}^2 \end{pmatrix} \right) \quad (5.36)$$

The reader interested in the details of how to estimate the different components of the model is referred to [Snijders and Bosker \(2011\)](#). Moreover, multilevel models are not restricted to linear regression. The same ideas can be applied to logistic (Sec. 4.3.3), Cox (Sec. 4.3.4) or Poisson regression (Sec. 4.3.5).

## 5.2.10 Loop designs

Microarray experiments face an important problem of cost and large variability of the technical replicates. However, this has been solved with a smart experiment design that has received the name of *loop designs* ([Churchill, 2002](#)). We illustrate these designs here as a way to fight very noisy measurements and possible biases caused by confounding experimental errors with the treatment effects. In Sec. 1.4.5 we presented the setup of this kind of experiments. Fig. 1.4 shows the two stages of the experimental procedure. We extract two aliquots from the same biological sample. One of the

aliquots is dyed in red and the other one in green. This dyeing (along with the whole first stage of the experiment) introduces a high level of noise. A single dyeing from a sample would confound the biological effects present in the specific sample being analyzed with the effect of the specific dyeing. For this reason, the dyeing process is repeated several times, and the different results are compared pairwise with some other dyeing (see green and red arrows in Fig. 1.4).

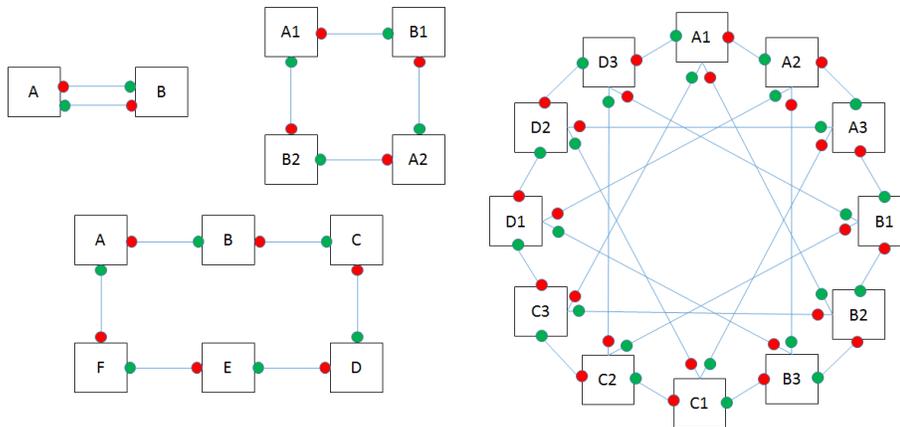


Figure 5.11: Some examples of loop designs. Capital letters (A, B, C, ...) represent different treatments (or varieties in the microarray lexicon). A1, A2 represent two subsamples from the same treatment. Each edge represents a microarray experiment, the red and green balls indicate which sample is dyed in red and green, respectively, in that experiment.

Loop designs organize the dyeing of the samples in such a way that the experimental design is balanced and we can determine the effect of the array and the dye (as nuisance/block variables) and the effect of the treatment and the gene (see Fig. 5.11). In a microarray experiment, the treatment is called the variety and it may represent different drugs, time points, physiological conditions, etc. We can also estimate second order interactions between array, dyes and genes (some genes may be particularly well or badly read/intensified in some of the arrays of dyeing processes). The overall model is of the form (Kerr and Churchill, 2001)

$$\log(y_{adv}) = \mu + \alpha_a^{(A)} + \alpha_d^{(D)} + \alpha_v^{(V)} + \alpha_g^{(G)} + \alpha_{ag}^{(AG)} + \alpha_{dg}^{(DG)} + \alpha_{vg}^{(VG)} + \varepsilon_{adv}$$

where  $\alpha_a^{(A)}$  is the main effect of the  $a$ -th array,  $\alpha_d^{(D)}$  is the main effect of the  $d$ -th dye,  $\alpha_v^{(V)}$  is the main effect of the  $v$ -th variety, and  $\alpha_g^{(G)}$  is the main effect of the  $g$ -th gene. Then, we have the corresponding second order interactions. The information of interest is in the interaction  $VG$ , that is, how the variety  $v$  affects the gene expression of the gene  $g$ . All effects related to the array and dye are “experimental artifacts”. The main effect of the variety is uninteresting because it would represent a general shift up or down of all genes. Similarly, the main effect of the gene is uninteresting because it represents genes that are shifted up or down irrespective of the variety.

### 5.2.11 Response surface designs

**Design summary.** These designs can be seen as the sampling plan for a surface regression. If we have multiple continuous factors,  $X_1, X_2, \dots, X_k$ , then these designs plan which samples to take from the different factors to optimally fit a response surface  $Y = f(X_1, X_2, \dots, X_k)$ .

- **Example 137:** We are preparing a formulation for a drug that must be delivered as an emulsion. We may dissolve the drug in two compounds simultaneously. The goal is to determine the optimal concentration of each of the two compounds such that the efficiency of the amount released is maximized. Fig. 5.12 shows a possible result of the experiment.  $X_1$  and  $X_2$  vary from 0 to 1 (from no concentration to a maximum concentration defined for each compound). For every combination of  $X_1$  and  $X_2$  the efficiency,  $Y$ , of the release is different. When the experiment is done we will find a functional relationship between  $y$  and our control variables:

$$Y = f(X_1, X_2)$$

We want to find sampling points (blue points) in the figure so that we can optimally determine the coefficients defining the response surface.

Interestingly, although we have not sampled at the optimum combination of  $X_1$  and  $X_2$  concentrations, by maximizing  $f$  we will be able to identify the optimal combination of these two solvents so that the liberation of the drug is maximum.

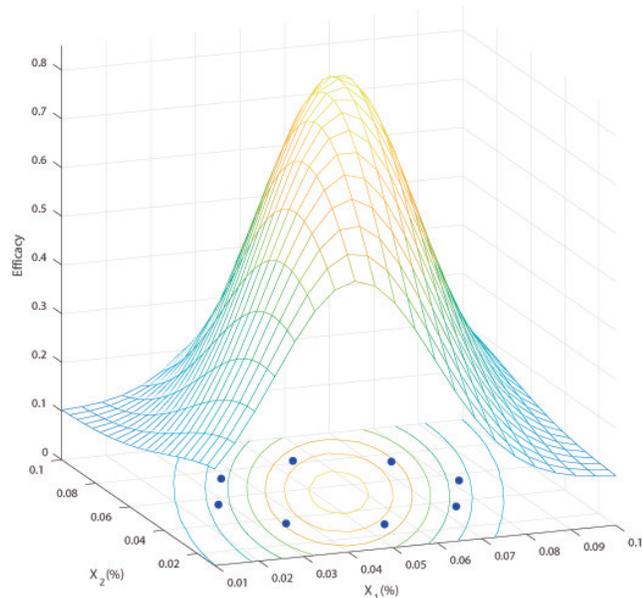


Figure 5.12: Response surface example.

Response surface designs choose a sampling pattern in the  $(X_1, X_2, \dots, X_p)$  space such that some property of the fitted surface is optimized. We must first choose the family of surfaces that we will explore. For instance, we may look for planes of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

This is a linear model in the regression parameters  $\beta$  (the dependence of  $Y$  on the  $\beta$ 's is linear). We may also allow for more complicated surfaces like

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon \quad (5.37)$$

This model allows for quadratic dependencies with  $X_1$  and  $X_2$ , but it is still linear in the regression coefficients. All linear models can be estimated by Least Squares. Once we have performed the experiment, assume we have  $N$  measurements of the form

$$(X_{1i}, X_{2i}, \dots, X_{pi}) \rightarrow Y_i$$

meaning that the  $i$ -th run of the experiment used the factor values  $(X_{1i}, X_{2i}, \dots, X_{pi})$  and for this combination we observed the response  $Y_i$ . We can set an equation system to solve for the regression coefficients. For example, for the quadratic model above, it would be

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & X_{11}^2 & X_{11}X_{21} & X_{21}^2 \\ 1 & X_{12} & X_{22} & X_{12}^2 & X_{12}X_{22} & X_{22}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & X_{1N} & X_{2N} & X_{1N}^2 & X_{1N}X_{2N} & X_{2N}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{12} \\ \beta_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

that is of the form

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Its Least Squares solution is, as we saw in Eq. 5.16,

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

Some domains require specific surface families. For instance, enzymatic reactions normally call for Michaelis-Menten functions, which can be written as

$$\frac{1}{Y} = \beta_0 + \beta_1 \frac{1}{X_1} + \varepsilon$$

or the modelling of population or tumor growth may call for a logistic regression

$$-\log\left(\frac{1}{Y} - 1\right) = \beta_0 + \beta_1 X_1 + \varepsilon$$

As we saw in Eq. 5.18, for linear models in the regression parameters and Gaussian, independent residuals, the covariance matrix of these regression parameters is

$$\Sigma_{\boldsymbol{\beta}} = \sigma_{\varepsilon}^2 (X^T X)^{-1}$$

That is, the uncertainty around the regression coefficients only depend on the variance of the residuals (which we cannot know before performing the experiment), and the system matrix  $X$  which is directly related to the sampling pattern. For linear models, the inverse of this matrix is called the Fisher's Information matrix

$$I_{\beta} = \Sigma_{\beta}^{-1}$$

The covariance of the predictions is

$$\Sigma_y = X^T \Sigma_{\beta} X$$

As we saw in Sec. 5.1.7, there are several optimization criteria

D-optimal	Maximize the determinant of $I_{\beta}$
A-optimal	Minimize the trace of $\Sigma_{\beta}$
T-optimal	Maximize the trace of $I_{\beta}$
E-optimal	Maximize the minimum eigenvalue of $I_{\beta}$
G-optimal	Minimize the maximum entry of $\Sigma_Y$
I-optimal	Minimize the trace of $\Sigma_Y$

None of them is necessarily better than the rest and our choice depends on our experimental objectives.

In the following we will assume that the  $X_i$  variables range between -1 and 1. If this is not the case, for instance, the solvents in Example 137 may go from a concentration of 0.01 to 0.1%, we can easily construct a new variable  $X'_i$  going from -1 to 1 by transforming the  $X_i$  values as

$$X'_i = -1 + 2 \frac{X_i - m_i}{M_i - m_i}$$

where  $m_i$  and  $M_i$  are the minimum and maximum values, respectively, of the variable  $X_i$ . We do the experiment design in the  $X'_i$  variables and transform them back to their natural range by undoing the transformation

$$X_i = m_i + \frac{M_i - m_i}{2} (X'_i + 1)$$

Once each factor is between -1 and 1, and if our model is linear like the one in Eq. 5.37, we might think of a  $2^k$  factorial design like the one shown in the following table and Fig. 5.13

$X'_1$	$X'_2$
-	-
-	+
+	-
+	+

However, pure  $2^k$  factorial designs do not allow estimating quadratic terms of the form  $X_i^2$ , extra samples need to be added. These extra samples are typically centered at

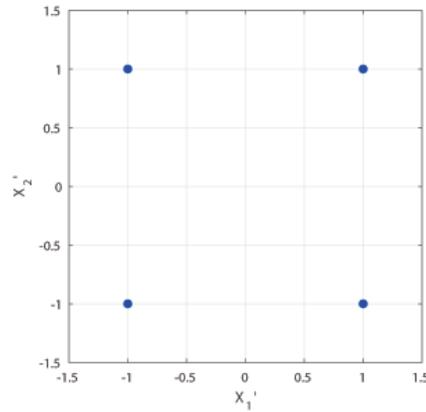


Figure 5.13: Example of a pure  $2^k$  factorial design for a response surface sampling.

0 so that the  $2^k$  factorial design (with only two levels -1 and 1) becomes a  $3^k$  factorial design (with three levels -1, 0, and 1). However, we do not need to run the full  $3^k$  factorial design (all 9 possible combinations). It is enough to perform a fractional run. Typically, only the center point is added, and several replicates of this are added as a way to measure possible drifts over time of the experiment, the inherent variability, and the curvature of the surface. Center points should be added at the beginning and the end of the experiment and evenly spread along the sequence as shown in the following table and Fig. 5.14. In general, there should be 3-5 center points along the experiment, or more if the total number of experiments is large.

$X'_1$	$X'_2$
0	0
-	-
-	+
0	0
+	-
+	+
0	0

The full factorial  $3^k$  factorial design allows the estimation of third order interactions

$$\begin{aligned}
 Y = & \beta_0 && \text{overall mean} \\
 & +\beta_1 X'_1 + \beta_2 X'_2 && \text{main effects} \\
 & +\beta_{11}(X'_1)^2 + \beta_{12} X'_1 X'_2 + \beta_{22}(X'_2)^2 && \text{2nd order interactions} \\
 & +\beta_{111}(X'_1)^3 + \beta_{112}(X'_1)^2 X'_2 + \beta_{122} X'_1 (X'_2)^2 + \beta_{222}(X'_2)^3 && \text{3rd order interactions} \\
 & +\varepsilon && \text{residual}
 \end{aligned}$$

However, the problem with these designs is that the number of runs quickly grows with the number of factors, while staying at the level of second order interactions keeps the number of runs at an acceptable level, as shown in the following table.

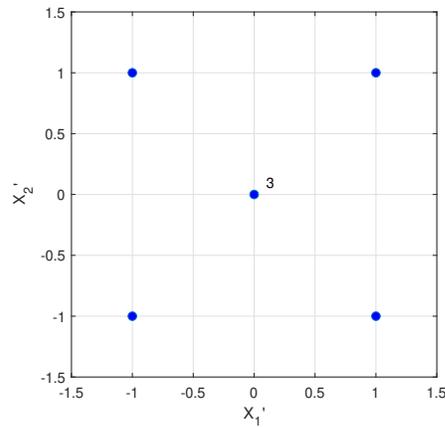


Figure 5.14: Example of a  $2^k$  factorial design with center points for a response surface sampling. The center point is run 3 times: once at the beginning of the experiment, once in the middle, and once at the end. In this way we can monitor the stability of the process.

$k$ # Factors	# Runs $3^k$ full factorial	# Runs Quadratic terms
2	9	6
3	27	10
4	81	15
5	243	21
6	729	28

Box-Wilson central composite designs can be used to perform the sampling capable of estimating up to the quadratic terms of the regression. Fig. 5.15 shows two of such designs. For  $k = 2$ , the number of runs of the Box-Wilson CCI and CCF designs coincides with the ones of the  $3^k$  full factorial. But this does not happen, in general, for higher  $k$ 's. The CCI design can be thought of as a  $2^k$  full factorial design (in Fig. 5.15 of levels  $\pm 1/\sqrt{2}$ , represented in red), plus some replicates of the center point, plus a set of axial points opposite to each other in different directions (represented in green). The CCF design can be thought of as a  $2^k$  full factorial design (of levels  $\pm 1$ , represented in red), plus a set of axial points located at the “faces” of the  $2^k$  full factorial (represented in green).

The treatments in the design can also be randomized, but not the central points, which should be evenly spread along the runs. For instance, a Box-Wilson CCI design with 5 center points could be run as

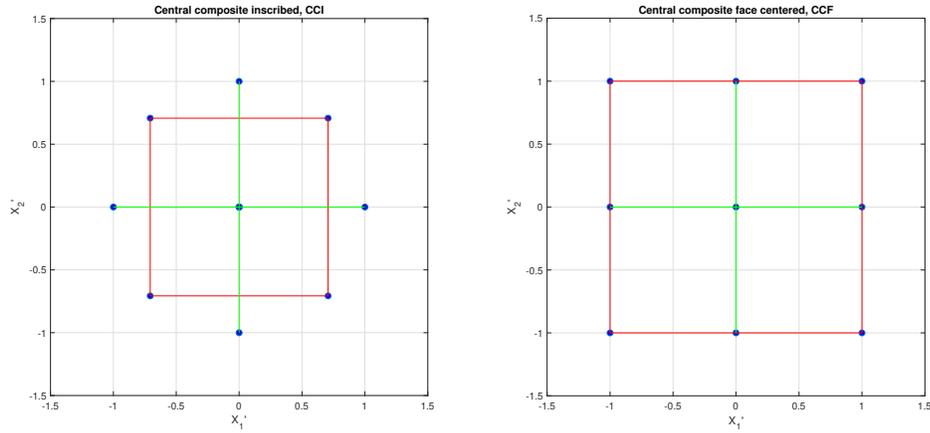


Figure 5.15: These Box-Wilson central composite designs allow estimating up to second order interactions between the factors.

$X'_1$	$X'_2$
0	0
0	0
0	$-1/\sqrt{2}$
0	$1/\sqrt{2}$
-1	-1
-1	1
1	-1
$-1/\sqrt{2}$	0
1	1
0	0
0	0
$1/\sqrt{2}$	0
0	0

In general, the further the sampling points from the origin, the more determined is the matrix  $X^T X$ . In this regard, the CCF design would be preferable to the CCI.

The Box-Behnken designs are an alternative to the CCF designs in which all samples are centered at the edges of the cube (hypercube) define by the  $2^k$  full factorial. Figure 5.16 shows these two designs. Both allow the estimation second order terms, but the CCF requires 15 samples and the Box-Behnken only 13. This difference increases as the number of factors,  $k$ , grows. Although, they are not seen, both designs have center points. The following table shows some of the treatments for the Box-Behnken design of  $k = 5$  factors, note that in each of the rows there are exactly two treatments

that are different from 0. For comparison purposes, the Box-Behnken design of  $k = 5$  factors require 46 runs, while the  $3^k$  full factorial requires 243.

$X'_1$	$X'_2$	$X'_3$	$X'_4$	$X'_5$
0	+	0	0	+
0	0	-	0	-
-	-	0	0	0
0	+	+	0	0
-	0	-	0	0

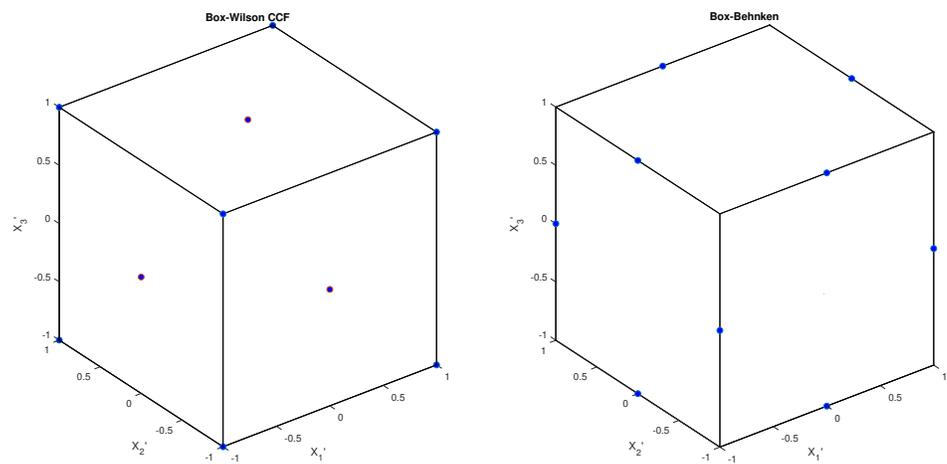


Figure 5.16: Comparison of Box-Wilson CCF and Box-Behnken designs.

Blocking is more difficult with response surface designs and only certain combinations are allowed. For instance, CCI designs allow blocking a variable with two levels; CCC also allows blocking with 2 or 3 levels; Box-Behnken designs allow blocking only in limited circumstances; and CCF does not allow blocking. However, we remember that we can follow a custom design by optimizing some property (see optimality criteria above). Then, we can easily include the block as one parameter more. Fig. 5.17 shows a D-optimal design of two control variables with three blocks (for instance, the researcher performing the experiment).

#### Important remarks

148. Blocking a variable is difficult in response surface designs. It can be done through optimality criteria, but the blocking pattern does not follow any particularly symmetric structure.

### 5.2.12 Mixture designs

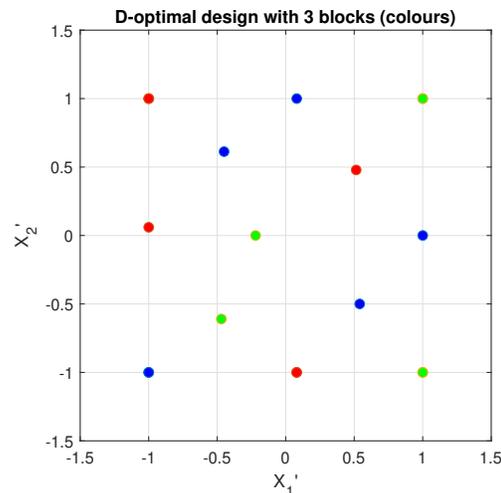


Figure 5.17: Example of D-optimal design for  $k = 2$  variables and 3 blocks.

**Design summary.** Similarly to response surface designs, mixture designs also address regressions of the type  $Y = f(X_1, X_2, \dots, X_k)$ , where  $Y$  is a variable of interest and  $X_1, X_2, \dots$  are the fractions of the mixture made of compound 1, compound 2, ... These designs are similar to surface response designs, only that there is an extra constraint that all control variables must add up to 1.

- **Example 138:** We are interested in preparing a feed for laboratory animals that maximizes the density of the bones. We have three ingredients for the feed, and we want to determine the optimal fraction of the three ingredients we must use. Our variable of interest,  $Y$ , is the density of the bones, that is supposed to be a function of the fraction of the three ingredients:

$$Y = f(X_1, X_2, X_3)$$

For instance, if we use 50% of ingredient 1, 30% of ingredient 2, and 20% of ingredient 3, then we would have  $X_1 = 0.5$ ,  $X_2 = 0.3$ , and  $X_3 = 0.2$ . These three fractions are constrained to add up to 1

$$X_1 + X_2 + X_3 = 1$$

Our analysis can be very similar to the one we have followed for the Surface response designs in the previous section. For instance, we may assume a linear model for the response with second order interactions

$$\begin{aligned}
 Y = & \beta_0 && \text{overall mean} \\
 & +\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 && \text{main effects} \\
 & +\beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{22} X_2^2 + \beta_{23} X_2 X_3 + \beta_{33} X_3^2 && \text{2nd order interactions} \\
 & +\varepsilon && \text{residual}
 \end{aligned}
 \tag{5.38}$$

All possible combinations of the mixture form a simplex. A simplex of  $k - 1$  dimensions is defined as the set of all points with coordinates  $(X_1, X_2, \dots, X_k)$  such that  $0 \leq X_i \leq 1$  and

$$X_1 + X_2 + \dots + X_k = 1$$

For instance, the simplex of 2 dimensions is given by 3 points and its shape is an equilateral triangle as the one shown in Fig. 5.18. The vertices of the simplex are given by the pure ingredients (a mixture in which only one of the ingredients is used). Points inside the triangle represent different mixtures. In the figure we show a few mixtures. For instance, the mixture in which all the ingredients are equally used  $X_1 = X_2 = X_3 = 1/3$  is called the barycenter of the triangle. The coordinates of any of the points,  $X$ , in the simplex (also called the barycentric coordinates) is a set of numbers  $(\lambda_1, \lambda_2, \dots, \lambda_k)$  such that

$$X = \lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_k X_k$$

and

$$\lambda_1 + \lambda_2 + \dots + \lambda_k = 1$$

These barycentric coordinates are the ones represented in parenthesis in Fig. 5.18.

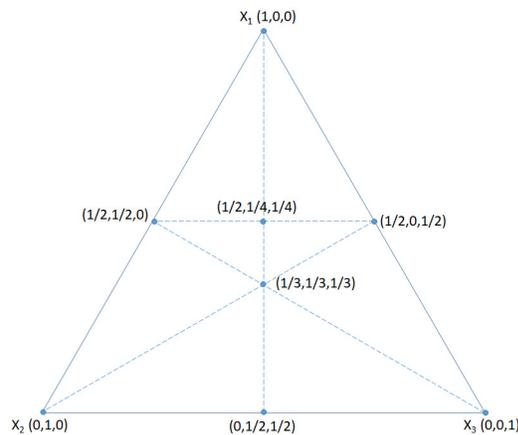


Figure 5.18: Simplex in 2 dimensions defined by 3 points  $(X_1, X_2, X_3)$ .

If we have  $k$  factors to explore, we can perform a *simplex*  $(k, m)$  design by using all combinations of the levels  $x_i = 0, \frac{1}{m}, \frac{2}{m}, \dots, 1$  such that the mixture constraint of adding up to 1 is fulfilled. For instance, with  $k = 3$  ingredients and  $m = 2$ , our design would be

$X_1$	$X_2$	$X_3$
1	0	0
1/2	1/2	0
1/2	0	1/2
0	1	0
0	1/2	1/2
0	0	1

For a  $(k, m)$ -simplex design, the number of runs is

$$N = \frac{(k + m - 1)!}{m!(k - 1)!}$$

Consequently, this design has only 6 runs ( $=4!/(2!2!)$ ), while for a quadratic model such as the one in Eq. 5.38, with 11 parameters, we need to replicate some of the points or augment the design to have more sampling points. As we did with the Response surface designs, we may repeat some of the combinations in order to monitor the stability of the process. In this way, we may augment the previous design to 14 runs as shown in this table and in Fig. 5.19.

$X_1$	$X_2$	$X_3$
0	0	1
1/2	1/2	0
1/3	1/3	1/3
0	1	0
1	0	0
1/2	0	1/2
2/3	1/6	1/6
1/6	2/3	1/6
0	0	1
1/2	1/2	0
1/6	1/6	2/3
0	1/2	1/2
0	1	0
1	0	0

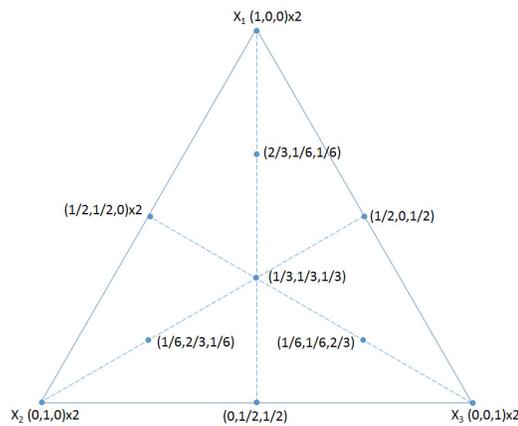


Figure 5.19: Mixture design, those points with replicated twice have been indicated by  $\times 2$ .

We may also perform constrained designs. For instance, we may want to perform mixtures in which  $X_1$  and  $X_2$  are constrained to be less than 50%. Then, our sampling

points must lay in the unshaded area of Fig. 5.20. The D-optimal sampling points are shown on the bottom of the same figure.

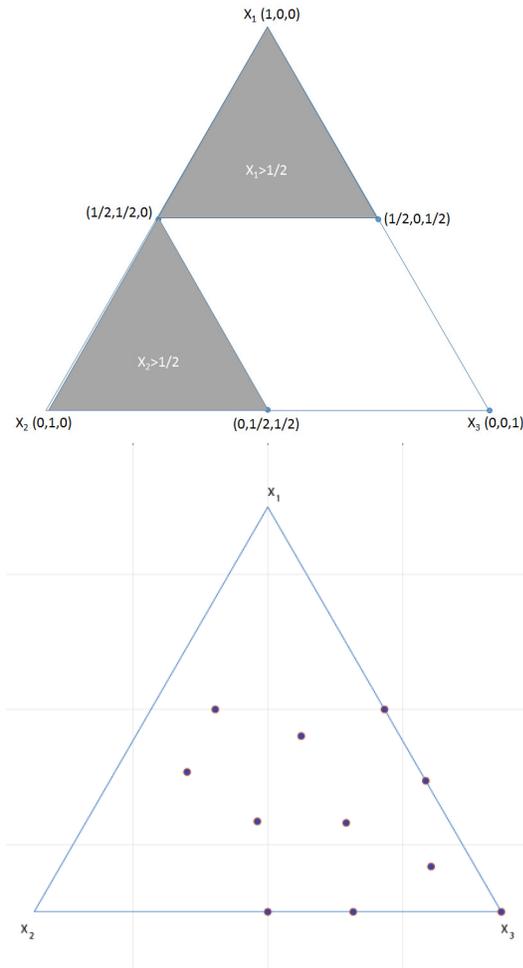


Figure 5.20: Top: Shaded areas correspond to mixtures in which  $X_1 > 1/2$  or  $X_2 > 1/2$ . Bottom: The D-optimal sampling points for the unshaded area.

### 5.3 Design selection guide

In the following paragraphs we provide some guidance on how to choose a suitable experimental design. The guide is not meant to be the ultimate word in the selection and other criteria may also be valid. However, we find it very useful for practitioners. The approach followed in this guide refers to several sections of the book depending on multiple criteria. Do not forget that the experiment designs explained in this chapter

must be combined with randomization so that we avoid biases caused by uncontrolled factors.

Criterion 1. What kind of study are you performing?

- Screening: It requires the least number of runs, but also gives the least amount of information. Very low resolution designs are employed (see fractional factorial designs in Sec 5.2.5). Only the main effects can be estimated. It requires a second experiment focusing on those factors that have been identified to be relevant. For continuous factors, the complexity of the fitted model (main effects, second order interactions, etc.) depend on the number of sampling points (see response surface designs in Sec. 5.2.11).
- Characterization: Requires more runs than screening and less than optimization. It can estimate models up to second order interactions (see fractional factorial designs in Sec 5.2.5 and response surface designs in Sec. 5.2.11).
- Optimization: This is the most informative design, but also requiring the most number of runs. It should be used with at most 5 factors, because the number of runs grows exponentially. For discrete factors, these are the standard factorial designs (Secs. 5.1.6 and 5.2.4). We may not need to perform all the treatment combinations if we are not going to estimate very high order interactions (see fractional designs in Secs. 5.1.7 and 5.2.5).

Criterion 2. What kind of variables do you have?

- Discrete factors: with two (*e.g.*, yes/no; absent/present) or more levels (*e.g.*, drug A/B/C or placebo). These are the standard factorial designs (Secs. 5.1.6 and 5.2.4).
  - If you have a few factors (*e.g.*, drug dose; drug dose and stomach state), you may want a Completely Randomized Design (Sec. 5.1.1). Actually, if we are not going to estimate all the high order interactions we can reduce the experimental effort by performing only a fraction of the whole list of experiments (Secs. 5.1.7 and 5.2.5).
  - If there are variables to block (*e.g.*, animal sex, day of the experiment, researcher performing it), you may want to use a Randomized Block Design (Sec. 5.1.3). If the block size does not allow for replicating all the treatments in all blocks, you will need an incomplete or imbalanced block design (Sec. 5.1.7). Latin squares is a quick way of analyzing one factor and two nuisance factors (Sec. 5.2.1). If we concatenate several experiments of this kind, you may use a Graeco-Latin square (Sec. 5.2.2).
  - Using covariates (*e.g.*, the laboratory temperature or time of the day at which the experiment is perform) can reduce the variability of the experiment (Sec. 5.1.4).

- If the discrete levels come from the discretization of a continuous variable (*e.g.*, three levels of a drug dose), you may want to explore a regression design (Sec. 5.1.2) or a response surface design with a single factor (Sec. 5.2.11). If the animal can be its own control, you may use a cross-over design (Sec. 5.2.3). They also serve as a way to block time as a nuisance factor.
- Continuous factors: these are continuous variables that can be independently set (*e.g.*, concentration of one, two or more solvents for an ointment; temperature and humidity of the experiment). These are the response surface designs (Sec. 5.2.11). You may also want to explore Regression designs (Sec. 5.1.2). Although in this book they have been presented in separate contexts, they are essentially the same thing.
- Mixture: these are continuous variables that not all of them can be independently set, because they are fractions of a total mixture, their sum must add up to one. These are the mixture designs (Sec. 5.2.12).
- Combined mixture+factors: These are combined designs in which the optimal mixture is sought for a number discrete or continuous factors. These designs have not been explicitly introduced in this book, but we have all the pieces to build them up.

Criterion 3. Is any of the factors hard to change?

In Sec. 5.2.6 we saw split-plot designs in which of the some factors cannot be easily changed (*e.g.*, the oven temperature, the room of an animal house, the individual in a cross-over design). A single level of these hard-to-change factors receive many combinations of the other easier-to-change factors. Repeated measures designs in which the same animal is measured several times is also treated as a split-plot design.

## 5.4 Sample size for designed experiments

Along the chapter we have presented the methodology to design an experiment so that it optimally allocates animals to different groups in order to eliminate the impact of biases, minimize the variance of group comparisons, and introduce other information known about each animal such as covariates. In this section we show how to calculate the sample size for these designs.

### 5.4.1 Sample size for completely randomized and randomized block designs

Consider Example 98. In that example we were testing the effect of two drug doses and a control on the cholesterol level of animals. We showed that the analysis of the data for this experiment, assuming we had already performed it with 10 animals per group, led to the ANOVA table reproduced below

Source	SS	df	MS
Treatments	31252	2	15626
Residuals	30600	27	1133
Total	61852	29	

We explained that the decision on whether all treatments had the same effect or not was taken of the ratio between the Mean Squares (MS) of the treatments and the Mean Squares of the residuals

$$f = \frac{15626}{1133} = 13.79$$

which, if the null hypothesis were true, it would be distributed as an Snedecor's F with 2 and 27 degrees of freedom. In this case, the probability of observing such a large  $f$ , or larger, if the null hypothesis was true was only  $6.18 \cdot 10^{-5}$  (this is the p-value of the hypothesis test) and we rejected the hypothesis that the treatments had no effect.

We remind that the experiment was performed with 10 animals per group. But how did we arrive to this number? Sample size calculations are performed before the experiment is done, so we need to assume something about the posterior behaviour of our measurements. In this case, control animals are supposed to have a cholesterol level around 250 mg/dL and a standard deviation of 30 mg/dL.

Under the Completely Randomized Design (CRD), the observed data is presumed to follow the ANOVA model

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

and the analysis of the data results in the following ANOVA table (see Sec. 5.1.1)

Source	Sum of squares	Degrees of freedom	Mean squares ( $MS = SS/df$ )
Treatments	$SS_{\alpha} = \sum_{ij} (y_i - y_{..})^2$	$T - 1$	$MS_{\alpha} = SS_{\alpha}/(T - 1)$
Residuals	$SS_{\varepsilon} = \sum_{ij} (y_{ij} - (\mu + \alpha_i))^2$	$N - T$	$MS_{\varepsilon} = SS_{\varepsilon}/(N - T)$
Total	$SS_T = \sum_{ij} (y_{ij} - y_{..})^2$	$N - 1$	

The ANOVA test is based on the  $F_{\alpha}$  statistic which under the  $H_0$  hypothesis is distributed as a central Snedecor's F with  $df_{\alpha}$  and  $df_{\varepsilon}$  degrees of freedom. Under  $H_a$  it is distributed as a noncentral Snedecor's F with the same degrees of freedom and non-centrality parameter

$$\phi = \frac{SS_{\alpha}}{\sigma_{\varepsilon}^2}$$

The principles of sample size calculation were already introduced in Sec. 1.6. We must specify a confidence level,  $1 - \alpha$ . Then, this confidence level will result in a critical value for an statistic, in this example  $f$ , such that if the observed  $f$  exceeds the critical value  $f_{1-\alpha}$ , then we reject the null hypothesis. We must also specify the statistical power we want to have to detect a specific departure from the null hypothesis. The statistical power is one minus the probability of failing to reject the null hypothesis when the alternative hypothesis is true. This probability depends on the sample size and the statistical power constraint gives us a sample size design equation.

This methodology applied to the particular case of the CRD results in the following reasoning. Let us assume we have  $T$  treatments, and we will employ  $n$  animals per group. The total number of animals will be  $N = nT$  and the corresponding number of degrees of freedom will be  $nT - 1$ . The degrees of freedom “consumed” by the treatments will be  $T - 1$  and the remaining degrees of freedom,  $nT - T$  go to the residuals. Consequently, the Snedecor’s  $F$  of interest is of  $T - 1$  and  $nT - T$  degrees of freedom. For the alternative hypothesis, we need to calculate the non-centrality parameter which is given by

$$\phi = \frac{SS_{\alpha}}{\sigma_{\varepsilon}^2} = \frac{\sum_{ij} (y_i - y_{..})^2}{\sigma_{\varepsilon}^2} = \frac{\sum_i n\alpha_i^2}{\sigma_{\varepsilon}^2} = n \frac{\sum_i \alpha_i^2}{\sigma_{\varepsilon}^2}$$

The sample size calculation for CRD is traditionally performed by hypothesizing a possible result for which we already want to have a given statistical power. In the cholesterol example, let us assume that we want detect with a statistical power of 90% those deviations of at least 20% from the nominal level, that is, if any of the groups departs more than 50 mg/dL ( $=0.2 \cdot 250$ ). We had three groups (control and two doses), let us refer to the mean in each one of the groups as  $\mu_1$ ,  $\mu_2$  and  $\mu_3$ , respectively. If all groups have the same number of individuals, then we can calculate the main effects as

$$\begin{aligned}\mu &= \frac{1}{T} \sum_{i=1}^T \mu_i \\ \alpha_i &= \mu_i - \mu\end{aligned}$$

- Example 139: In our example, we would have, for instance

$$\mu_1 = 250, \mu_2 = 250, \mu_3 = 200$$

Actually, it does not matter if the deviation is positive or negative, and in which of the groups it occurs. Then

$$\begin{aligned}\mu &= \frac{1}{3} (250 + 250 + 200) = 233.33 \\ \alpha_1 &= 250 - 233.33 = 16.67 \\ \alpha_2 &= 250 - 233.33 = 16.67 \\ \alpha_3 &= 200 - 233.33 = -33.33\end{aligned}$$

The corresponding non-centrality parameter would be

$$\phi = n \frac{16.67^2 + 16.67^2 + (-33.33)^2}{30^2} = 1.85n$$

The sample size design must find a number of samples per group such that the probability of rejecting the null hypothesis when this is true is  $\alpha$  and the probability of not rejecting it when it is false is  $\beta$  (see Fig. 5.21). In general, we must satisfy

$$\boxed{F_{1-\alpha, df_{\alpha}, df_{\varepsilon}} = F_{\beta, \phi, df_{\alpha}, df_{\varepsilon}}} \quad (5.39)$$

and in particular, with the variables introduced so far

$$\boxed{F_{1-\alpha, T-1, Tn-T} = F_{\beta, \phi, T-1, Tn-T}} \quad (5.40)$$

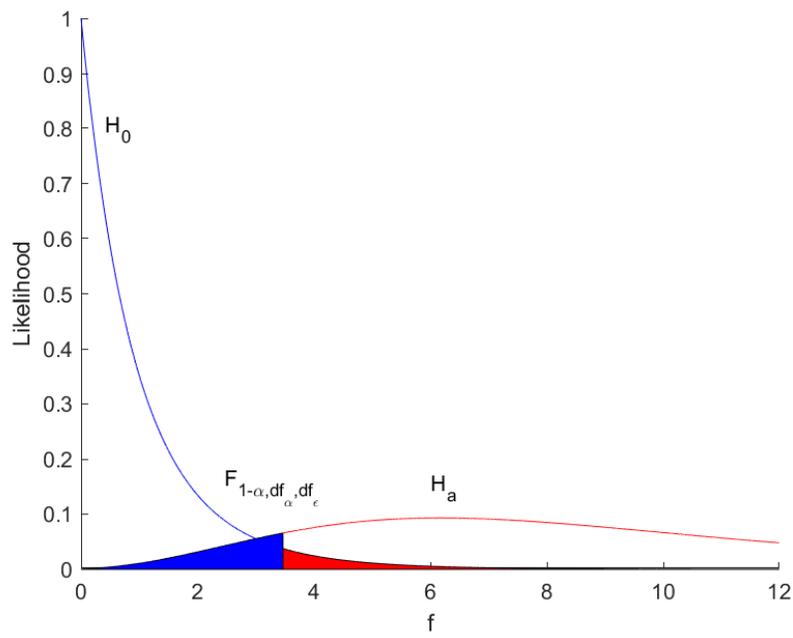


Figure 5.21: The red shaded area is the probability of observing an  $f$  statistic larger than the critical value  $F_{1-\alpha, df_a, df_e}$  if the null hypothesis is true (this area should be  $\alpha$ ). The blue shaded is the probability of not rejecting the null hypothesis when the alternative hypothesis is true (this area should be  $\beta$ ).

- Example 139 (continued): In our example, we must find  $n$  such that

$$F_{0.95,2,3n-3} = F_{0.1,1.85n,2,3n-3}$$

This is satisfied for  $n = 8$ . We see that by taking  $n = 10$  animals per group we have used more animals than necessary for our research goals.

A similar design, although from a different perspective, would have been obtained from the calculations in Sec. 4.1.6.

#### Important remarks

149. Eq. 5.39 is also valid for randomized block experiments and experiments with covariates. The only difference is the number of degrees of freedom left for the residuals and the variance of the residuals (which is reduced by the introduction of the blocks and the covariates if they really explain part of the variability of the data). The non-centrality parameter only depends on the main effects and the effect size we want to be able to detect.
150. For randomized block experiments, we need to remind that  $n$  is the number of samples for a given treatment. If there is a block with two levels, then each of the two levels will receive  $n/2$  samples.

## 5.4.2 Sample size for factorial designs

The sample size for a full factorial design follows the same principles as for the standard completely randomized designs. Eq. 5.39 is still valid as long as we correctly account for the number of degrees of freedom. For several factors, we need to perform a sample size calculation for each one of the factors. In the following example we illustrate these ideas.

- Example 140: Following with our example of the previous section about the design of the diet to control cholesterol, let us assume that we have 5 different nutritional compositions (factor  $P$ ), 3 different levels of fiber content (factor  $Q$ ) and 4 different presentations (factor  $R$ ). We want to be able to detect a change of 20% in the cholesterol level with a statistical power of 90%. How many animals do we need per group (defined as the combination of  $P$ ,  $Q$ , and  $R$ )?

When we analyze the data the ANOVA table will be of the following form (note that samples have now four indices, three for the factors and one for the animal within that combination of factor levels). We assume that there are  $n$  animals per combination

Source	Sum of squares	Degrees of freedom
Treatments $P$	$SS_{\alpha^{(P)}} = \sum_{ijkl} (\alpha_i^{(P)})^2 = nQR \sum_i (\alpha_i^{(P)})^2$	$P - 1$
Treatments $Q$	$SS_{\alpha^{(Q)}} = \sum_{ijkl} (\alpha_j^{(Q)})^2 = nPR \sum_j (\alpha_j^{(Q)})^2$	$Q - 1$
Treatments $R$	$SS_{\alpha^{(R)}} = \sum_{ijkl} (\alpha_k^{(R)})^2 = nPQ \sum_k (\alpha_k^{(R)})^2$	$R - 1$
Interactions $PQ$	$SS_{\alpha^{(PQ)}} = \sum_{ijkl} (\alpha_{ij}^{(PQ)})^2 = nR \sum_{ij} (\alpha_{ij}^{(PQ)})^2$	$(P-1)(Q-1)$
Interactions $PR$	$SS_{\alpha^{(PR)}} = \sum_{ijkl} (\alpha_{ik}^{(PR)})^2 = nQ \sum_{ik} (\alpha_{ik}^{(PR)})^2$	$(P-1)(R-1)$
Interactions $QR$	$SS_{\alpha^{(QR)}} = \sum_{ijkl} (\alpha_{jk}^{(QR)})^2 = nP \sum_{jk} (\alpha_{jk}^{(QR)})^2$	$(Q-1)(R-1)$
Interactions $PQR$	$SS_{\alpha^{(PQR)}} = \sum_{ijkl} (\alpha_{ijk}^{(PQR)})^2 = n \sum_{ijk} (\alpha_{ijk}^{(PQR)})^2$	$(P-1)(Q-1)(R-1)$
Residuals	$\sum_{ijkl} \varepsilon_{ijkl}^2$	$N - PQR$
Total	$SS_T = \sum_{ijkl} (y_{ijkl} - y_{\dots})^2$	$N - 1$

In general, given  $T$  levels within a factor, a change  $\Delta$  in one of the classes results in main effects of the form  $\alpha_i = \frac{1}{T} \Delta$  for all levels except for one that will be  $\alpha_T = -\frac{T-1}{T} \Delta$ . The corresponding sum of squares is

$$\sum_{i=1}^T \alpha_i^2 = \frac{T-1}{T} \Delta^2.$$

For different factors the associated non-centrality parameters are

$$\begin{aligned} \phi_P &= \frac{SS_{\alpha^{(P)}}}{\sigma_\varepsilon^2} = \frac{nQR \frac{P-1}{P} \Delta^2}{\sigma_\varepsilon^2} \\ \phi_Q &= \frac{SS_{\alpha^{(Q)}}}{\sigma_\varepsilon^2} = \frac{nPR \frac{Q-1}{Q} \Delta^2}{\sigma_\varepsilon^2} \\ \phi_R &= \frac{SS_{\alpha^{(R)}}}{\sigma_\varepsilon^2} = \frac{nPQ \frac{R-1}{R} \Delta^2}{\sigma_\varepsilon^2} \end{aligned}$$

and  $n$  must be such that the statistical confidence and power requirements are met for all factors

$$\begin{aligned} F_{1-\alpha, P-1, N-PQR} &= F_{\beta, \phi_P, P-1, N-PQR} \\ F_{1-\alpha, Q-1, N-PQR} &= F_{\beta, \phi_Q, Q-1, N-PQR} \\ F_{1-\alpha, R-1, N-PQR} &= F_{\beta, \phi_R, R-1, N-PQR} \end{aligned} \quad (5.41)$$

- Example 140 (continued): In our example, we have  $P = 5$ ,  $Q = 3$ ,  $R = 4$ . Then,

$$\begin{aligned} \phi_P &= \frac{n \cdot 3 \cdot 4 \cdot \frac{4}{5} \cdot 50^2}{30^2} = 26.67n \\ \phi_Q &= \frac{n \cdot 5 \cdot 4 \cdot \frac{2}{3} \cdot 50^2}{30^2} = 37.04n \\ \phi_R &= \frac{n \cdot 5 \cdot 3 \cdot \frac{3}{4} \cdot 50^2}{30^2} = 31.25n \end{aligned}$$

The we must find  $n$  such that

$$\begin{aligned} F_{0.95,4,60n-60} &= F_{\beta,26.67n,4,60n-60} \\ F_{0.95,2,60n-60} &= F_{\beta,37.04n,2,60n-60} \\ F_{0.95,3,60n-60} &= F_{\beta,31.25n,3,60n-60} \end{aligned}$$

It suffices with  $n = 2$  samples per group. We cannot do it with  $n = 1$  because we would not have degrees of freedom available for the residuals. Actually, with  $n = 2$  we have a statistical power much larger than 90%, and we will be able to detect smaller changes.

#### Important remarks

151. We have presented an example with a full factorial design in which the sample size calculation has been performed considering only the main effects. However, it would be straightforward to base the design in the interactions of order two, three, ... or to consider a factorial design in which the interactions are not estimated. A key point is that we need to foresee before doing the experiment which will be the variance of the residuals.

- Example 141: Following with the previous example, we have a total of  $PQR = 5 \cdot 3 \cdot 4 = 60$  different treatments, with 2 animals per group. That makes a total of  $N = 120$  animals. They are housed in cages of 4 animals. If we think that the cages may have some influence, we can use them as blocks. We will need 30 cages, and we will spend 29 degrees of freedom in estimating their effects. If our model contains second order interactions, then the number of degrees of freedom required for the model are:  $4 + 2 + 3 = 9$  for the main effects and  $4 \cdot 2 + 4 \cdot 3 + 2 \cdot 3 = 26$  for the second order interactions. The total number of degrees of freedom required for the model will be  $29 + 9 + 26 = 64$ . We have 120 animals (119 degrees of freedom), so that leaves us with  $119 - 64 = 55$  degrees of freedom for the residuals. With this many number of spare degrees of freedom, we are still capable of blocking other variables like the researcher performing the experiment, animal sex, and age.

## Appendix. Mathematical introduction to experiment design

The theory of statistical experimental design is very much linked to the one of linear regression. For this reason, we will start our presentation of the topic by presenting linear regression, and later we will make use of its properties to design our experiments.

### Linear models

Let us assume that we have a response (dependent) variable  $y$  that we want to explain as a function of some input (independent) variable  $x$  (we will later expand the model

to multiple independent variables). In general, the dependence will be through some parameters  $\boldsymbol{\beta}$  that define a function between the two variables:

$$y = f(\mathbf{x}; \boldsymbol{\beta}) \quad (5.42)$$

This function is called a model. For instance,

$$\begin{aligned} (1) \quad & y = \beta_0 + \beta_1 x \\ (2) \quad & y = \beta_0 + \beta_1 x + \beta_2 x^2 \\ (3) \quad & y = \beta_0 + \beta_1^2 x \\ (4) \quad & y = \frac{\beta_0}{\beta_1 + \beta_2 x} \end{aligned}$$

The model is linear in  $\boldsymbol{\beta}$  if the following condition holds:

$$f(\mathbf{x}; a\boldsymbol{\beta}_1 + b\boldsymbol{\beta}_2) = af(\mathbf{x}; \boldsymbol{\beta}_1) + bf(\mathbf{x}; \boldsymbol{\beta}_2) \quad (5.43)$$

In the examples above, Models 1 and 2 are linear and Models 3 and 4 are not.

### Bias and variance

The goal of regression is to estimate the parameters  $\boldsymbol{\beta}$  of the model given pairs of observations  $\{(\mathbf{x}_i, y_i)\}$  where  $i = 1, 2, \dots, N$ . We will assume that these observations are generated through the model

$$y_i = f(\mathbf{x}_i; \boldsymbol{\beta}) + \varepsilon_i \quad (5.44)$$

where  $\varepsilon_i$  is some random variation with zero mean (later we will further assume that it is normally distributed, although we do not need this at the moment).

Our estimate of the parameters will be  $\hat{\boldsymbol{\beta}}$ , and for a given predictor variable,  $x_i$ , our predicted value will be

$$\hat{y}_i = f(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \quad (5.45)$$

The difference between the observed and the predicted values is the *residual* of that observation

$$e_i = y_i - \hat{y}_i \quad (5.46)$$

Our estimate of the model parameters,  $\hat{\boldsymbol{\beta}}$ , is obtained by some procedure that tries to optimize some objective function. For instance, *Ordinary Least Squares (OLS)*, minimizes the sum of the squares of the residuals

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_i e_i^2 \quad (5.47)$$

Let us consider the procedure to estimate  $\boldsymbol{\beta}$ , we define the *bias* of the procedure

$$B\{\hat{\boldsymbol{\beta}}\} = \mathbb{E}\{\hat{\boldsymbol{\beta}}\} - \boldsymbol{\beta} \quad (5.48)$$

The *Mean Square Error (MSE)* is defined as

$$\text{MSE}\{\hat{\boldsymbol{\beta}}\} = \mathbb{E}\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\} \quad (5.49)$$

The *Standard Error* is the square root of the MSE. Finally, the variance of the estimator is

$$\text{Cov}\{\hat{\boldsymbol{\beta}}\} = \mathbb{E}\{(\hat{\boldsymbol{\beta}} - \mathbb{E}\{\hat{\boldsymbol{\beta}}\})(\hat{\boldsymbol{\beta}} - \mathbb{E}\{\hat{\boldsymbol{\beta}}\})^T\} \quad (5.50)$$

An important result is that

$$\text{MSE}\{\hat{\boldsymbol{\beta}}\} = (\mathbf{B}\{\hat{\boldsymbol{\beta}}\})^2 + \text{Trace}\{\text{Cov}\{\hat{\boldsymbol{\beta}}\}\} \quad (5.51)$$

Among all possible estimators of  $\boldsymbol{\beta}$ , we are normally interested in one that is unbiased and has minimum variance (MVUE, Minimum Variance Unbiased Estimator). In general, there could be multiple MVUE estimators.

### Linear estimators

$\hat{\boldsymbol{\beta}}$  is a *linear estimator* of  $\boldsymbol{\beta}$  if it can be written as a linear combination of the predictor variables

$$\hat{\boldsymbol{\beta}} = \sum_i a_i x_i \quad (5.52)$$

Linear estimators having the MVUE property is called BLUE (Best Linear Unbiased Estimator). If  $\mathbb{E}\varepsilon_i = 0$ , and  $\text{Cov}\{\varepsilon_i, \varepsilon_j\} = 0$  for all  $i$  and  $j$  (that is, the residuals are uncorrelated), and  $\text{Var}\{\varepsilon_i\} = \sigma^2 < \infty$  (that is, residuals are homocedastic), then the Gauss-Markov theorem states that the OLS is the only BLUE estimator, that is, the linear estimate given by OLS,  $\hat{\boldsymbol{\beta}}$ , is unique.

Linear models can be written in a matrix form. For instance, let us assume that we presume that the model is of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Given a set of measurements, we could write all the equations implied by all observations as the matrix equation

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_N & x_N^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

We may simplify the notation using vectors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.53)$$

$\mathbf{X}$  is called the design matrix. Let  $P$  be the dimension of the vector  $\boldsymbol{\beta}$ , that is, the number of parameters to estimate.

A similar model is obtained in multiple regression. For instance, let us assume that we have two predictors,  $x_1$  and  $x_2$ , and our model is of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \varepsilon$$

Then, the equation system corresponding to all our measurements would be

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}^2 & x_{11}x_{12} & x_{12}^2 \\ 1 & x_{21} & x_{22} & x_{21}^2 & x_{21}x_{22} & x_{22}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & x_{N1}^2 & x_{N1}x_{N2} & x_{N2}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{12} \\ \beta_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

The standard assumptions in regression are

$$\begin{aligned} (1) \quad \mathbb{E}\{\boldsymbol{\varepsilon}\} &= \mathbf{0} \\ (2) \quad \text{Var}\{\boldsymbol{\varepsilon}\} &= \sigma^2 I_N \end{aligned} \quad (5.54)$$

where  $I_N$  is the identity matrix of size  $N \times N$ . Additionally, we may assume the Gaussianity of the noise perturbations

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_N) \quad (5.55)$$

The OLS estimate of the model parameters is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} \quad (5.56)$$

This estimate is unbiased

$$\mathbb{E}\{\hat{\boldsymbol{\beta}}\} = \boldsymbol{\beta} \quad (5.57)$$

and its variance is

$$\text{Var}\{\hat{\boldsymbol{\beta}}\} = \sigma^2 (X^T X)^{-1} \quad (5.58)$$

If we assume that the noise,  $\boldsymbol{\varepsilon}$ , is a multivariate Gaussian (Eq. 5.55), then the estimate in Eq. 5.56 is also the maximum likelihood solution of the regression problem, that is, the estimate that maximizes the likelihood of observing these observation pairs

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} f_{\boldsymbol{\varepsilon}}(\mathbf{y} - X\boldsymbol{\beta}) = \arg \min_{\boldsymbol{\beta}} \sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \quad (5.59)$$

where  $f_{\boldsymbol{\varepsilon}}$  is the multivariate probability density function of the noise (the multivariate Gaussian) and in the second equality we have made use of the fact that residuals are independent of each other, they all have the same variance, and that maximizing the likelihood is equivalent to minimizing its minus logarithm. The vector  $\mathbf{x}_i^T$  is the  $i$ -th row of the  $X$  matrix (Eq. 5.53).

Given the OLS solution of the model parameters, our prediction for the observed values would be

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} = H\mathbf{y} \quad (5.60)$$

where we have defined  $H = X(X^T X)^{-1} X^T$ . This matrix is referred to as the hat matrix and it projects the vector  $\mathbf{y}$  onto the column space of the matrix  $X$ . It is easily seen that  $H$  must be idempotent ( $H^2 = H$ ), because once a vector is projected onto a subspace, projecting it again onto the same subspace will not change the vector.

The vector of residuals can be calculated as

$$\mathbf{e} = (I_N - H)\mathbf{y} \quad (5.61)$$

That is,  $I_N - H$  is the projection matrix onto the subspace orthogonal to the column space of  $X$ . The mean and variance of  $\mathbf{e}$  are

$$\begin{aligned} \mathbb{E}\{\mathbf{e}\} &= \mathbf{0} \\ \text{Var}\{\mathbf{e}\} &= \sigma^2(I_N - H) \end{aligned} \quad (5.62)$$

Given the observations, an unbiased estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{N - P} \mathbf{y}^T (I_N - H) \mathbf{y} \quad (5.63)$$

We also may define the  $N \times N$  matrix

$$J = \mathbf{1}_N (\mathbf{1}_N^T \mathbf{1}_N)^{-1} \mathbf{1}_N^T = \frac{1}{N} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix} \quad (5.64)$$

where  $\mathbf{1}_N$  is a column vector of 1's.  $J$  is the projection matrix onto the column space of  $\mathbf{1}_N$ .

Let us define  $\bar{y}$  as the average of all the  $y_i$  measurements (a property of the OLS solution is that the mean of the  $y_i$  measurements is equal to the mean of the  $\hat{y}_i$  predictions). Then,

$$J\mathbf{y} = \bar{y}\mathbf{1}_N$$

$J$ ,  $H - J$ ,  $I_N - H$ , and  $I_N - J$  are also idempotent. Additionally,  $HJ = J$ ,  $HX = X$ ,  $(I_N - H)X = (I_N - H)H = [0]$  and  $H$  and  $J$  are both symmetric matrices ( $H^T = H$  and  $J^T = J$ ). The column space of  $H$  is the same as the column space of  $X$ . Finally, the rank of  $H$  is  $P$ , the rank of  $J$  is 1, the rank of  $I_N - J$  is  $N - 1$ , the rank of  $H - J$  is  $P - 1$ , and the rank of  $I_N - H$  is  $N - P$ . All these ranks are related to the numbers of degrees of freedom that appear in the ANOVA table.

### Partition of the total variation

Given the set of observations,  $\{(x_i, y_i)\}$ , we define the total variation as

$$SS_{total} = \sum_i (y_i - \bar{y})^2. \quad (5.65)$$

Using the matrix properties of  $H$  and  $J$  stated above, we may show that

$$SS_{total} = \mathbf{y}^T (I_N - J) \mathbf{y} \quad (5.66)$$

We may also define the regression variation as

$$SS_{regr} = \sum_i (\hat{y}_i - \bar{y})^2 = \mathbf{y}^T (H - J) \mathbf{y} \quad (5.67)$$

Finally, we define the residual variation as

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 = \mathbf{y}^T (I_N - H) \mathbf{y} \quad (5.68)$$

It can be shown that

$$SS_{total} = SS_{regr} + SS_{res} \quad (5.69)$$

It can also be shown that

$$\begin{aligned} \mathbb{E}\{SS_{res}\} &= \sigma^2(N - P) \\ \mathbb{E}\{SS_{regr}\} &= (P - 1)\sigma^2 + (\mathbf{X}\boldsymbol{\beta})^T (I_N - J)(\mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (5.70)$$

A consequence of these expectations is that an unbiased estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{SS_{res}}{N - P} = \frac{\|\mathbf{e}\|^2}{N - P} \quad (5.71)$$

Even more general is to substitute  $P$  by the rank of  $X$ , in case that  $X^T X$  is not invertible.

Another known property is that  $SS_{res}$  must decrease with the addition of each new predictor, that is, with increasing  $P$ .

With these definitions, the coefficient of determination is defined as

$$R^2 = \frac{SS_{regr}}{SS_{total}} = 1 - \frac{SS_{res}}{SS_{total}} \quad (5.72)$$

If this number is 1, then we have made a perfect regression in which all residuals are zero.  $R^2$  can be understood as a function of the Signal-to-Noise Ratio (SNR)

$$R^2 = \frac{SS_{regr}/SS_{res}}{SS_{regr}/SS_{res} + 1} = \frac{SNR}{SNR + 1} \quad (5.73)$$

The higher the SNR, the closer  $R^2$  gets to 1.

### Centering and scaling the models

Let us consider a regression on two predictors ( $x_1$  and  $x_2$ ) to a set of observations  $\{x_{i1}, x_{i2}, y_i\}$  with  $i = 1, 2, \dots, N$ . To keep things simple, we can consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

We may center the predictors by subtracting their means ( $\bar{x}_j = \frac{1}{N} \sum_i x_{ij}$ )

$$\begin{aligned} \tilde{x}_{i1} &= x_{i1} - \bar{x}_1 \\ \tilde{x}_{i2} &= x_{i2} - \bar{x}_2 \end{aligned}$$

Then, we could construct a new model with the centered variables

$$y_i = \beta'_0 + \beta'_1 \tilde{x}_{i1} + \beta'_2 \tilde{x}_{i2} + \varepsilon_i$$

It would be easy to go from one model to the other

$$\begin{aligned}\beta_0 &= \beta'_0 - \beta'_1 \bar{x}_1 - \beta'_2 \bar{x}_2 \\ \beta_1 &= \beta'_1 \\ \beta_2 &= \beta'_2\end{aligned}$$

Considering that the range of the predictors might be very different, we may also use the sample standard deviation of each one of the predictor ( $s_j^2 = \frac{1}{N-1} \frac{1}{N} \sum_i (x_{ij} - \bar{x}_j)^2$ )

$$\begin{aligned}\tilde{x}_{i1} &= \frac{x_{i1} - \bar{x}_1}{s_1} \\ \tilde{x}_{i2} &= \frac{x_{i2} - \bar{x}_2}{s_2}\end{aligned}$$

$$y_i = \beta''_0 + \beta''_1 \tilde{x}_{i1} + \beta''_2 \tilde{x}_{i2} + \varepsilon_i$$

Again, we can recover the coefficients of the original model

$$\begin{aligned}\beta_0 &= \beta''_0 - \beta''_1 \bar{x}_1 / s_1 - \beta''_2 \bar{x}_2 / s_2 \\ \beta_1 &= \beta''_1 / s_1 \\ \beta_2 &= \beta''_2 / s_2\end{aligned}$$

As the OLS solution of the regression problem is unique, all these models with centered and standardized predictors are equivalent to each other in the sense that they all have the same  $H$  matrices and consequently the same predictions,  $\hat{y}_i$ .

However, there is an advantage in working with centered and standardized predictors. As we saw in Eq. 5.56, finding the model parameters requires the inversion of the matrix  $X^T X$ . This is the matrix

$$X^T X = \begin{pmatrix} N & N\bar{x}_1 & N\bar{x}_2 \\ N\bar{x}_1 & \mathbf{x}_1^T \mathbf{x}_1 & \mathbf{x}_1^T \mathbf{x}_2 \\ N\bar{x}_2 & \mathbf{x}_2^T \mathbf{x}_1 & \mathbf{x}_2^T \mathbf{x}_2 \end{pmatrix}$$

where the vectors  $\mathbf{x}_j$  collect all the  $x_{ij}$  predictors. For the centered predictors, the  $2 \times 2$  lower-right matrix is  $N - 1$  times the sample estimate of the covariance matrix; and for the standardized it is  $N - 1$  times the sample estimate of the correlation matrix. In any case, this matrix is positive semidefinite meaning that all its eigenvalues are 0 or larger than 0.

The condition number of a matrix is the ratio between its largest and smallest singular values (or eigenvalues, if the matrix is square as is the case):

$$\kappa\{X^T X\} = \frac{\lambda_{max}}{\lambda_{min}} \quad (5.74)$$

If the condition number is very large, the matrix is said to be ill-conditioned. We will not give here a formal definition of a matrix being  $\varepsilon$ -ill conditioned, but it suffices to say that a matrix is  $\varepsilon$ -ill conditioned if it has at least an eigenvalue smaller than  $\varepsilon$ . If a matrix is  $\varepsilon$ -ill conditioned, then its condition number is larger than  $1/\varepsilon$ .

In our regression setup, the inversion of this ill-conditioned matrix will result in a large instability of the estimate of the model parameters,  $\hat{\boldsymbol{\beta}}$ . For seeing that, let us

consider the eigenvalues of the matrix  $X^T X$ ,  $\lambda_p$  ( $p = 1, 2, \dots, P$ ), and study the behaviour of

$$\mathbb{E}\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\} = \text{Trace}\{\text{Var}\{\hat{\boldsymbol{\beta}}\}\} = \sigma^2 \text{Trace}\{(X^T X)^{-1}\} = \sigma^2 \sum_{p=1}^P \frac{1}{\lambda_p} \quad (5.75)$$

It is easy to see that the raw predictors may have more differences in eigenvalues (larger condition number) due to the different centers and scales, while the standardized predictors will have the smaller difference possible between eigenvalues (smaller condition number).

### Multicollinearity and Variance Inflation Factor

In the extreme, the matrix  $X^T X$  may not be invertible. This happens if one or several of the predictors are linearly dependent on the remaining predictors. Then, the column space of the matrix  $X$  is not of dimension  $P$ , but  $P - K$  if there are  $K$  linearly dependent predictors. That case is called multicollinearity and the determinant of  $X^T X$  will be 0. In practice, it may be that the determinant is not 0, but very close to 0, meaning that there is no strict linear dependence, but at least one of the predictor is almost linearly dependent on the rest. To detect problems of collinearity let us consider  $X^T X$  for the standardized predictor model (we reason above that all models are equivalent in terms of their predictive power)

$$X^T X = \begin{pmatrix} N & \mathbf{0}^T \\ \mathbf{0} & (X^*)^T X^* \end{pmatrix} \Rightarrow (X^T X)^{-1} = \begin{pmatrix} \frac{1}{N} & \mathbf{0}^T \\ \mathbf{0} & ((X^*)^T X^*)^{-1} \end{pmatrix}$$

where  $X^*$  are the columns corresponding to the predictors excluding  $\beta_0$ . We know that

$$(X^*)^T X^* = (N - 1)R \Rightarrow ((X^*)^T X^*)^{-1} = \frac{1}{N - 1}R^{-1} \quad (5.76)$$

where  $R$  is the sample correlation matrix. In the ideal case, all predictors are uncorrelated and  $R = I_{p-1}$ . Then, the variance of  $\hat{\beta}_p$  is  $\sigma^2$ .

Let us analyze the effect of some correlation. For doing so, let us consider the first predictor  $x_1$  with respect to the rest. We split the  $X^*$  matrix in two parts:  $\mathbf{x}_1$  (the vector corresponding to  $x_1$ ) and the rest of columns  $X_1^*$ :

$$X^* = (\mathbf{x}_1 \quad X_1^*)$$

Then,

$$\begin{aligned} (X^*)^T X^* &= \begin{pmatrix} \mathbf{x}_1^T \mathbf{x}_1 & \mathbf{x}_1^T X_1^* \\ (X_1^*)^T \mathbf{x}_1 & (X_1^*)^T X_1^* \end{pmatrix} \Rightarrow \\ ((X^*)^T X^*)^{-1} &= \begin{pmatrix} (\mathbf{x}_1^T \mathbf{x}_1 - \mathbf{x}_1^T X_1^* ((X_1^*)^T X_1^*)^{-1} (X_1^*)^T \mathbf{x}_1)^{-1} & \dots \\ \dots & \dots \end{pmatrix} \end{aligned}$$

That is, the variance of  $\beta_1$  is now

$$\text{var}\{\hat{\beta}_1\} = (\mathbf{x}_1^T \mathbf{x}_1 - \mathbf{x}_1^T X_1^* ((X_1^*)^T X_1^*)^{-1} (X_1^*)^T \mathbf{x}_1)^{-1} \sigma^2$$

The factor in front of  $\sigma^2$  is always larger than 1, and it is called the Variance Inflation Factor (VIF). It reflects the increase of uncertainty in the estimate of  $\beta_1$  due to linear correlations between  $x_1$  and the rest of predictors. We could extend this calculation to any other variable. Typically, we would prefer to have VIFs below 5. A common source of correlation is working with non-standardized predictors (the fact that they all have a mean different from zero induce some correlation between them). It can be shown that

$$VIF = \frac{1}{1 - R_i^2} \quad (5.77)$$

where  $R_i^2$  is the coefficient of determination of the linear regression between the predictor  $x_i$  and the rest of predictors.

### Underfitting and overfitting

**Underfitting** Let us assume that the true model of some observation requires some predictors  $x_1$  and  $x_2$ . However, we do not know that it depends on  $x_2$  and we fit a model only depending on  $x_1$ . Then,

- the MSE is biased: it is made of  $\sigma^2$  plus some positive value that depends on all the variability due to  $x_2$  that we have not explained.
- the estimate of  $\beta_1$  is also biased: that is, part of the information of  $x_2$  is transferred to the estimate of the coefficient of  $x_1$ . How much is transferred depends on the linear correlation between  $x_1$  and  $x_2$ .

**Overfitting** Let us assume that the true model of some observation only depends on  $x_1$ . However, we fit a model that also makes use of a predictor  $x_2$  (on which  $y$  does not depend on). Then,

- the estimate of  $\beta_1$  has more variance: that is, part of the information of  $x_1$  is transferred to the estimate of the coefficient of  $x_2$ . How much is transferred depends on the linear correlation between  $x_1$  and  $x_2$ .

### Discrete predictors

Let us imagine that we are predicting a variable  $y$  as a function of  $x_1$ . But, the dependence of  $y$  on  $x_1$  may change depending on another variable  $\tilde{x}_2$

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \varepsilon & \tilde{x}_2 = A \\ \beta'_0 + \beta'_1 x_1 + \varepsilon & \tilde{x}_2 = B \end{cases}$$

$\tilde{x}_2$  is a variable that only takes the values A or B (which are pure symbols without any numerical value, for instance, male and female, or young and old). We may use an indicator variable that takes the values 0 or 1 depending on whether  $\tilde{x}_2$  takes the value A or B:

$$x_{2A} = I(\tilde{x}_2 = A) = \begin{cases} 1 & \tilde{x}_2 = A \\ 0 & \text{otherwise} \end{cases}$$

This is referred to as an indicator variable, and we can integrate into the fitting model in the standard way

$$\begin{aligned} y &= (\beta_0 + \beta_2 x_{2A}) + (\beta_1 + \beta_3 x_{2A})x_1 + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_{2A} + \beta_3 x_1 x_{2A} + \varepsilon \end{aligned}$$

$\beta_2$  accounts for a change of intercept between  $x_2 = 0$  and  $x_2 = 1$ , while  $\beta_3$  accounts for a change of slope, or equivalently, the interaction between  $x_1$  and  $\tilde{x}_2$ , that is, the effect of  $x_1$  depends on the value of  $\tilde{x}_2$ .

Let assume now that  $\tilde{x}_2$  can take three possible values: A, B, or C. Instead of one indicator variable, we can define 2:

$$\begin{aligned} x_{2A} &= I(\tilde{x}_2 = A) \\ x_{2B} &= I(\tilde{x}_2 = B) \end{aligned}$$

The translation of  $\tilde{x}_2$  into the different variables is

$\tilde{x}_2$	$x_{2A}$	$x_{2B}$
A	1	0
B	0	1
C	0	0

As we will see, we do not need to include a third indicator variable to account for  $\tilde{x}_2 = C$ . Let us construct the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_{2A} + \beta_3 x_{2A} x_1 + \beta_4 x_{2B} + \beta_5 x_{2B} x_1 + \varepsilon$$

Under the different values of  $\tilde{x}_2$ , this model simplifies to

$$y = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1 + \varepsilon & \tilde{x}_2 = A \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1 + \varepsilon & \tilde{x}_2 = B \\ \beta_0 + \beta_1 x_1 + \varepsilon & \tilde{x}_2 = C \end{cases}$$

This is a general feature of linear models. Given a discrete variable with  $L$  levels, we only need to estimate  $L$  parameters. For instance, in ANOVA we assume that we have  $L$  treatments and a number of observations for each one of the treatments. In this way,  $y_{ij}$  is the observation of the  $j$ -th individual in the  $i$ -th group of the treatments.

We could use a *means model*

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{5.78}$$

For example we have two measurements per treatment and three treatments, then the equation system would be

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

Alternatively, we could have used the *effects model*

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (5.79)$$

Then, the equation system is

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

Finally, we could have used a *multiple linear regression model with indicator variables*

$$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{ij} \quad (5.80)$$

The corresponding equation system would be

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

Finally, the *ANOVA model* is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (5.81)$$

with the restriction

$$\sum_i \alpha_i = 0$$

The equation system corresponding to this model is The corresponding equation system would be

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

The rank of the  $X$  matrix of all these equation systems is 3, meaning that we can only fit 3 model parameters. This implies that the effects model, with 4 model parameters, is not uniquely identifiable, that is, there are multiple values of the model parameters all of them leading to the same relationship between the predicted and predictor variables (this statement is related to the non-uniqueness of the generalized inverse mentioned in the next session). All the other models are equivalent in terms of prediction power because the column space of the three models is the same. However, they are not all equally determined: the condition number of the means model is 1, while the condition

numbers of the multiple linear regression model with indicators and the one of ANOVA are 3. That is, the means model is the one with less variance in the estimate of its parameters.

The matrix  $H$  of all these models is

$$H = \frac{1}{2} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Interestingly, the predicted values of all these models for each one of the observations are the average of the samples of each treatment

$$\hat{\mathbf{y}} = H\mathbf{y} = \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{1.} \\ \bar{y}_{2.} \\ \bar{y}_{2.} \\ \bar{y}_{3.} \\ \bar{y}_{3.} \end{pmatrix}$$

which is a very sensible prediction when all treatments have the same number of observations. A common feature of  $H$  in ANOVA designs is that it is a block diagonal matrix, as the one above, although not all blocks necessarily have the same value.

In this way, we see that there is not a single way to deal with discrete predictors. ANOVA, ANCOVA, and linear regression are all linear regression models that can be set under the same modelling framework. Most importantly, the system matrix  $X$  plays a central role as  $(X^T X)^{-1}$  as it determines the variance of each one of the parameters. It also determines if the model can be estimated or not, if  $X^T X$  cannot be inverted.

### Fixed vs. Random effects

In the fixed-effects ANOVA, the model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

with the constraint  $\sum_i \alpha_i = 0$  to make the estimation unique. The  $\alpha_i$ 's are fixed, although unknown, numbers.

In the random-effects ANOVA, the model is the same, but the  $\alpha_i$ 's are supposed to be realizations of a random variable

$$\alpha_i \sim N(0, \sigma_A^2)$$

We may write the system matrix  $X$  as

$$X = (\mathbf{1}_N \quad X^*)$$

where  $X^*$  contains the part corresponding to the effects. It can easily be shown that

$$\begin{aligned}\mathbb{E}\{\mathbf{y}\} &= \boldsymbol{\mu}\mathbf{1}_N \\ \text{Var}\{\mathbf{y}\} &= \sigma_A^2 X^*(X^*)^T + \sigma_\varepsilon^2 I_N\end{aligned}\quad (5.82)$$

The partition of the total variation is exactly the same as in the general regression case. For the details on how to estimate  $\sigma_A^2$  and  $\sigma_\varepsilon^2$  and the number of degrees of freedom associated to the different sum of squares, the reader is referred to Sec. 5.2.8.

### Experiment design

One the experiment is performed, it is assumed that it will be analyzed through some form of linear regression (Eq. 5.53). We here reproduce the main results from the previous sections for convenience

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.83)$$

The OLS estimate of the model parameters is

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} \quad (5.84)$$

The variance of this estimate is

$$\text{Var}\{\hat{\boldsymbol{\beta}}\} = \sigma^2 (X^T X)^{-1} \quad (5.85)$$

And the predictions

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} = H\mathbf{y} \quad (5.86)$$

It might be that  $X^T X$  cannot be inverted (because its determinant is 0), then we may use one of its generalized inverse or pseudo-inverse matrices (the matrix  $A^-$  is a generalized inverse of the matrix  $A$  if  $AA^-A = A$ , this matrix is not unique and there are several methods to calculate them). The properties of the matrices  $H$  and  $I_N - H$  with the inverse of  $X^T X$  or its generalized inverse are the same (they are idempotent, symmetric, they span the column space of  $X$  or the complementary of the column space of  $X$ , etc.)

Experiment design aims at defining the matrix  $X$  such that  $X^T X$  has some useful properties. For instance: 1) the model parameters can be uniquely determined if and only if the rank of  $X^T X$  equals the dimension of  $\boldsymbol{\beta}$ ; 2) the variance (uncertainty) of the estimates of  $\boldsymbol{\beta}$  depends on the condition number of  $X^T X$ . We may minimize its trace ( $A$ -optimality), its determinant ( $D$ -optimality), its maximum eigenvalue ( $E$ -optimality), etc.

# Bibliography

- Aarts E, Dolan CV, Verhage M, van der Sluis S (2015) Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC neuroscience* 16(1):1–15
- Adcock C (1997) Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46(2):261–283
- Afsarinejad K (1983) Balanced repeated measurements designs. *Biometrika* 70:199–204
- Albers C, Lakens D (2018) When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *J experimental social psychology* 74:187–195
- Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews Genetics* 7:55–65
- Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454
- Barton D, HogenEsch H, Weih F (2000) Mice lacking the transcription factor *relb* develop t cell-dependent skin lesions similar to human atopic dermatitis. *European J of immunology* 30(8):2323–2332
- Bate S, Karp NA (2014) A common control group-optimising the experiment design to maximise sensitivity. *PLoS One* 9:e114,872
- Beal J (2017) Biochemical complexity drives log-normal variation in genetic expression. *Engineering Biology* 1(1):55–60
- Bebarta V, Luyten D, Heard K (2003) Emergency medicine animal research: does use of randomization and blinding affect the results? *Academic Emergency Medicine* 10(6):684–687
- Begley CG, Ioannidis JPA (2015) Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research* 116:116–126
- Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, et al (2018) Redefine statistical significance. *Nature Human Behaviour* 2(1):6

- Bennett CM, Wolford GL, Miller MB (2009) The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience* 4(4):417–422
- Bert B, Heintz C, Chmielewska J, Schwarz F, Grune B, Hensel A, Greiner M, Schönfelder G (2019) Refining animal research: The animal study registry. *PLoS biology* 17:e3000463
- Boos DD, Stefanski LA (2011) P-value precision and reproducibility. *The American Statistician* 65(4):213–221
- Button KS, Ioannidis J, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* 14:365–376
- Chia R, Achilli F, Festing MFW, Fisher EMC (2005) The origins and uses of mouse outbred stocks. *Nature genetics* 37:1181–1186
- Chow SC, Shao J, Wang H (2008) *Sample size calculations in clinical research*, 2nd edn. Chapman & Hall, CRC
- Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. *Nature genetics* 32 Suppl:490–495
- Chvedoff M, Clarke MR, Faccini JM, Irisarri E, Monro AM (1980) Effects on mice of numbers of animals per cage: an 18-month study (preliminary results). *Archives of toxicology Supplement* 4:435–438
- Clark TS, Linzer DA (2015) Should I use fixed or random effects? *Political science research and methods* 3(2):399–408
- Clayton JA (2016) Studying both sexes: a guiding principle for biomedicine. *FASEB J* 30:519–524
- Cohen J (1994) The earth is round ( $p < .05$ ). *American Psychologist* 49:997–1003
- Crabbe JC, Wahlsten D, Dudek BC (1999) Genetics of mouse behavior: interactions with laboratory environment. *Science* 284:1670–1672
- Cumming G (2008) Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science* 3(4):286–300
- Dobson AJ, Barnett A (2008) *An introduction to generalized linear models*. CRC press
- Doncaster CP, Davey A (2007) *Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences*. Cambridge Univ. Press
- Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG (2006) A gentle introduction to imputation of missing values. *J Clinical Epidemiology* 59(10):1087–1091

- Ekwall B, Barile FA, Castano A, Clemedson C, Clothier RH, Dierickx P, Ekwall B, Ferro M, Fiskesj G, Garza-Ocañas L, Gomez-Lechon MJ, Golden M, Hall T, Iso-maa B, Kahru A, Kerszman G, Kristen U, Kunimoto M, Krenlampi S, Lewan L, Loukianov A, Ohno T, Persoone G, Romert L, Sawyer TW, Shrivastava R, Segner H, Stamatii A, Tanaka N, Valentino M, Walum E, Zucco F (1998) MEIC evaluation of acute systemic toxicity: Part VI. the prediction of human toxicity by rodent LD50 values and results from 61 in vitro methods. *Alternatives to laboratory animals* : ATLA 26 Suppl 2:617–658
- Ellenberg J (2014) *How not to be wrong. The power of mathematical thinking.* Penguin books
- Festing MFW (2003) Principles: The need for better experimental design. *Trends in Pharmacological Sciences* 24:341–345
- Festing MFW, Altman DG (2002) Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR journal* 43(4):244–258
- Fleming TR (1982) One-sample multiple testing procedure for phase ii clinical trials. *Biometrics* 38:143–151
- Fox J (2015) *Applied regression analysis and generalized linear models*, 3rd edn. SAGE Publications
- Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. *PLoS biology* 13:e1002165
- Giesbrecht FG, Gumpertz ML (2004) *Planning, construction, and statistical analysis of comparative experiments.* Wiley and Sons, Inc.
- Gomes DG (2022) Should i use fixed effects or random effects when i have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ* 10:e12794
- Gore KH, Stanley PJ (2005) An illustration that statistical design mitigates environmental variation and ensures unambiguous study conclusions. *Animal Welfare* 14(4):361–365
- Greenwald AG, Gonzalez R, Harris RJ, Guthrie D (1996) Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology* 33(2):175–183
- Guerra C, Collado M, Navas C, Schuhmacher AJ, Hernández-Porras I, Cañamero M, Rodriguez-Justo M, Serrano M, Barbacid M (2011) Pancreatitis-induced inflammation contributes to pancreatic cancer by inhibiting oncogene-induced senescence. *Cancer cell* 19(6):728–739
- Halsey LG (2019) The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology letters* 15:20190174

- Hartman H, Wang Y, Schroeder Jr HW, Cui X (2018) Absorbance summation: A novel approach for analyzing high-throughput elisa data in the absence of a standard. *PLoS One* 13(6):e0198,528
- Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. *PLoS biology* 13:e1002,106
- Higgins JP, Green S (2011) *Cochrane handbook for systematic reviews of interventions*, vol 4. John Wiley & Sons
- Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A (2019) Moving beyond p values: data analysis with estimation graphics. *Nature methods* 16:565–566
- Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 55(1):19–24
- Hooijmans CR, Rovers MM, de Vries RBM, Leenaars M, Ritskes-Hoitinga M, Langendam MW (2014) Syrcle’s risk of bias tool for animal studies. *BMC medical research methodology* 14:43
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS medicine* 2:e124
- Jay Jr GE (1955) Variation in response of various mouse strains to hexobarbital (evipal). *Proc Soc Experimental Biology and Medicine* 90(2):378–380
- Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8(1):118–127
- Kalliokoski O, Jacobsen KR, Teilmann AC, Hau J, Abelson KS (2012) Quantitative effects of diet on fecal corticosterone metabolites in two strains of laboratory mice. *In vivo* 26:213–221
- Kanji GK (2006) *100 Statistical tests*. Sage Publications
- Karp NA, Segonds-Pichon A, Gerdin AKB, Ramirez-Solis R, White JK (2012) The fallacy of ratio correction to address confounding factors. *Laboratory animals* 46:245–252, DOI 10.1258/la.2012.012003
- Karpen SC (2017) p value problems. *American J pharmaceutical education* 81:6570
- Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. *Biostatistics* 2:183–201
- Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS one* 4:e7824
- Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) Improving bio-science research reporting: the arrive guidelines for reporting animal research. *PLoS biology* 8:e1000,412

- Klein CJ, Budiman T, Homberg JR, Verma D, Keijer J, van Schothorst EM (2022) Measuring locomotor activity and behavioral aspects of rodents living in the home-cage. *Frontiers in Behavioral Neuroscience* 16
- Lakens D (2015) On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ* 3:e1142
- Lazic SE (2010) The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC neuroscience* 11:5
- Lazic SE, Clarke-Williams CJ, Munaf? MR (2018) What exactly is 'n' in cell culture and animal experiments? *PLoS biology* 16:e2005,282
- Lee HS, Kim SK, Han JB, Choi HM, Park JH, Kim EC, Choi MS, An HJ, Um JY, Kim HM, et al (2006) Inhibitory effects of *rumex japonicus* houtt. on the development of atopic dermatitis-like skin lesions in nc/nga mice. *British J of Dermatology* 155(1):33–38
- Leist M, Hartung T (2013) Inflammatory findings on species extrapolations: humans are definitely no 70-kg mice. *Archives of toxicology* 87:563–567
- Limpert E, Staehl W, Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51:341–352
- Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PMW, Buchan A, Buchan A, van der Worp HB, Traystman RJ, Minematsu K, Donnan GA, Howells DW (2009) Reprint: Good laboratory practice: preventing introduction of bias at the bench. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism* 29:221–223
- Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, Hirst T, Hemblade R, Bahr Z, Nunes-Fonseca C, Potluru A, Thomson A, Baginskaitė J, Baginskaitė J, Egan K, Vesterinen H, Currie GL, Churilov L, Howells DW, Sena ES (2015) Risk of bias in reports of in vivo research: A focus for improvement. *PLoS biology* 13:e1002,273
- Mathews P (2010) *Sample size calculations*. Mathews Malnar and Bailey, Inc.
- McCance I (1995) Assessment of statistical procedures used in papers in the Australian veterinary journal. *Australian veterinary J* 72(9):322–329
- McDonagh M, Peterson K, Raina P, Chang S, Shekelle P (2013) Avoiding bias in selecting studies. *Methods guide for effectiveness and comparative effectiveness reviews*
- Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis J (2017) A manifesto for reproducible science. *Nature human behaviour* 1:1–9
- van der Naald M, Chamuleau SAJ, Menon JML, de Leeuw W, de Haan JJ, Duncker DJ, Wever KE (2021) A 3-year evaluation of preclinicaltrials.eu reveals room for improvement in preregistration of animal studies. *PLoS biology* 19:e3001,397

- Nájera JL, Gómez CE, García-Arriaza J, Sorzano CO, Esteban M (2010) Insertion of vaccinia virus c7l host range gene into nyvac-b genome potentiates immune responses against hiv-1 antigens. *PLoS One* 5(6):e11,406
- Nieuwenhuis S, Forstmann BU, Wagenmakers EJ (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience* 14:1105–1107
- O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. *Biometrics* pp 549–556
- Pearl DL (2014) Making the most of clustered data in laboratory animal research using multi-level models. *ILAR journal* 55(3):486–492
- Pease ME, Hammond JC, Quigley HA (2006) Manometric calibration and comparison of tonolab and tonopen tonometers in rats with experimental glaucoma and in normal mice. *J of glaucoma* 15:512–519
- Pinkert CA (2014) *Transgenic animal technology*. Elsevier
- Price R, Bethune R, Massey L (2020) Problem with p values: why p values do not tell you if your treatment is likely to work. *Postgraduate medical journal* 96:1–3
- Quackenbush J (2002) Microarray data normalization and transformation. *Nature genetics* 32 Suppl:496–501
- ter Riet G, Korevaar DA, Leenaars M, Sterk PJ, Van Noorden CJF, Bouter LM, Lutter R, Elferink RPO, Hooft L (2012) Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. *PLoS one* 7:e43,404
- Rodrigues I, Sanches J, Bioucas-Dias J (2008) Denoising of medical images corrupted by poisson noise. In: *Proc. 15th IEEE Int. Conf. Image Processing*, pp 1756–1759
- Sackett DL, et al (1979) Bias in analytic research. *J Chron Dis* 1979 32:51–63
- Schulz MA, Schmalbach B, Brugger P, Witt K (2012) Analysing humanly generated random number sequences: a pattern-based approach. *PLoS one* 7:e41,531
- Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, Clark A, Cuthill IC, Dirnagl U, et al (2020) The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *J Cerebral Blood Flow & Metabolism* 40:1769–1777
- Sheskin DJ (2004) *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC
- Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22:1359–1366

- Simon R (1989) Optimal two-stage designs for phase ii clinical trials. *Controlled clinical trials* 10:1–10
- Simonsohn U, Nelso L, Simmons JP (2014) P-curve: A key to the file-drawer. *J Experimental psychology: General* 143:534–547
- Simsek B, Firat MZ (2011) Application of multilevel analysis in animal sciences. *Applied Mathematics and Computation* 218(3):1067–1071
- Smith AJ, Clutton RE, Lilley E, Hansen KEA, Brattelid T (2018) Prepare: guidelines for planning animal research and testing. *Laboratory animals* 52(2):135–141
- Smith RJ (2020)  $P > .05$ : The incorrect interpretation of “not significant” results is a significant problem. *American J physical anthropology* 172:521–527
- Snijders TA, Bosker RJ (2011) *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage
- Sorzano COS, Tabas-Madrid D, Núñez F, Fernández-Criado C, Naranjo A (2017) Sample size for pilot studies and precision driven experiments. *arXiv preprint arXiv:170700222*
- Sorzano COS, Tabas-Madrid D, Núñez F, Fernández-Criado C, Naranjo A (2018) Sample size for pilot studies and precision driven experiments. *arXiv p 1707.00222*
- Sullivan LM, Weinberg J, Keaney JF (2016) Common statistical pitfalls in basic science research. *Journal of the American Heart Association* 5
- Thompson SK (2012) *Sampling*. John Wiley & Sons
- Vidgen B, Yasseri T (2016) P-values: misunderstood and misused. *Frontiers in Physics* 4:6
- Warton DI, Hui FK (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92(1):3–10
- Wassmer G, Brannath W (2018) *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer
- Waters JC (2009) Accuracy and precision in quantitative fluorescence microscopy. *The Journal of cell biology* 185:1135–1148
- Weichbrod RH, Thompson GAH, Norton JN (2017) *Management of animal care and use programs in research, education, and testing*. CRC Press
- Yoshida M, Miyasaka Y, Ohuchida K, Okumura T, Zheng B, Torata N, Fujita H, Nabae T, Manabe T, Shimamoto M, Ohtsuka T, Mizumoto K, Nakamura M (2016) Calpain inhibitor calpeptin suppresses pancreatic cancer by disrupting cancer-stromal interactions in a mouse xenograft model. *Cancer science* 107:1443–1452, DOI 10.1111/cas.13024

- Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L (2015) The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. *J evidence-based medicine* 8(1):2–10
- Zwillinger D (1996) *CRC Standard Mathematical Tables and Formulae*, 30th edn. CRC Press, Boca Raton (Florida)

# Index

- ANOVA
  - hierarchical, 301
  - multiway, 293, 307
  - one way, 284
  - two way, 293, 307
- Bias, 28
- Bioequivalence, 193
- Data imputation, 38
- Data snooping, 22, 145, 163, 347
- Errors
  - Benjamini-Hochberg correction, 70
  - bias, 28
  - Bonferroni correction, 69
  - False Discovery Rate, FDR, 69
  - false negative, 63
  - false positive, 63
  - multiple testing, 67
  - outliers, 44
  - Sidak correction, 70
  - type I, 63
  - type II, 63
- Experiment design, 283
  - balanced incomplete block, 330
  - completely randomized, 284
  - criss-cross, 355
  - cross-over, 337
  - cyclic, 333
  - factorial, 307, 342
  - fixed effects, 359
  - fractional factorial, 347
  - Graeco-Latin squares, 337
  - hierarchical, 356
  - imbalanced, 327
  - incomplete, 324
  - incomplete cross-over, 342
  - Latin squares, 334
  - lattice, 333
  - least squares, 318
  - loop, 366
  - mixed effects, 359
  - mixture, 374
  - multilevel model, 365
  - nested, 355, 356
  - non-orthogonal, 318
  - Plackett-Burman, 353
  - random effects, 359
  - randomized block, 293
  - randomized complete block, 295
  - regression, 290
  - repeated measures, 356
  - resolution, 351
  - response surface, 368
  - screening, 353
  - selection guide, 378
  - single replicate factorial, 317
  - split-plot, 354
  - split-unit, 354
  - Taguchi, 353
  - Youden squares, 334
- Experiment efficiency, 324
- Experimental unit, 25
- Factorial design
  - between-factor, 355
  - within-factor, 355
- Hypothesis testing, 56
  - alternative hypothesis, 57, 77
  - effect size, 59
  - non-parametric, 80
  - null hypothesis, 57, 75

- one-tail tests, 58
  - p-value, 63, 142, 146
  - parametric, 80
  - significance test, 58
  - statistical confidence, 63
  - statistical power, 63
  - superiority tests, 58
  - two-tails test, 58
- Intention to treat, 38
- Linear models, 284, 290, 293, 300, 307, 318
- Metaanalyses, 29
- Noise model
  - additive, 47
  - multiplicative, 51
- Observational unit, 27
- Outliers, 44
- p-value, 63, 142, 146
- Pseudoreplication, 40
- Sample independence, 45
- Sample size, 75
  - adaptive, 271
  - ANOVA, 189, 304, 381
  - Asymptotic Relative Efficiency, ARE, 80
  - Cohen's  $\kappa$ , 249
  - completely randomized design, 381
  - confidence interval, 185, 200, 204
  - contingency tables, 209
  - correlations, 242
  - count data, 229
  - designed experiments, 380
  - exponential survival, 251, 255, 266
  - factorial design, 384
  - Gaussian survival, 253, 256, 260
  - intraclass correlation, 245
  - log-rank test for survival, 267
  - Mead's resource equation, 191
  - mean, 182, 183, 185, 186, 188, 193
  - mean survival time, 251, 261
  - pilot experiments, 268
  - Poisson counts, 234
  - proportion, 196, 200, 202, 204, 206, 207, 209, 211, 212
  - randomized block design, 381
  - reestimation, 280
  - regression, 213
    - Cox, 225
    - linear, 217, 219
    - logistic, 221
    - Poisson, 229
  - survival analysis, 251
  - survival rate, 259, 264
  - survival time percentile, 255
  - unequal group sizes, 192
  - variance, 238
  - Weibull survival, 252, 256
- Statistical inference, *see also* Hypothesis testing
- Statistical pitfalls, 137
- Survival analysis, 225
- Technical replicate, 40
- Test guide
  - $\Gamma$  survival, 177
  - $\Omega^2$ , 176
  - $\chi^2$ , 165, 172, 174–176
  - $\chi^2$  goodness-of-fit, 171, 174
  - ANCOVA, 166–168
  - Anderson-Darling, 165
  - ANOVA, 167, 168, 176
  - ANOVA nested, 168
  - Bartlett, 167, 168
  - binomial sign, 167, 172, 174
  - bootstrap, 172
  - Bowker, 174
  - Brown-Forsythe, 167, 168
  - canonical correlation, 170
  - Chauvenet, 166
  - clustering, 171
  - Cochran, 175
  - Cochran's C, 168
  - Cochran's Q, 175
  - Cochran-Mantel-Haenszel, 175
  - coefficient of determination  $R^2$ , 176
  - Cohen's  $\kappa$ , 176

- Cohen's d, 176
- Cohen's f, 176
- Cohen's g, 176
- conjoint analysis, 170
- contingency coefficient, 176
- correspondence analysis, 171
- Coupon's collector, 174
- Cox proportional hazards, 177
- Cramér-von Mises, 165
- Cramer's  $\phi$ , 176
- D'Agostino-Pearson, 165
- difference sign, 166
- dimensionality reduction, 170
- discriminant analysis, 170
- Dixon's Q, 166
- Duncan, 167
- Dunnett, 167
- Durbin-Watson, 169
- exact goodness-of-fit, 174
- exponential survival, 177
- exponential-logarithmic survival, 177
- factor analysis, 170
- Fisher's cumulant, 165
- Fisher's exact, 170, 174, 176
- frequency, 174
- Friedman, 168, 172, 175
- G, 174, 175
- gap, 174
- Gart, 174
- Goodman and Kruskal's  $\gamma$ , 176
- Grubbs, 166
- Hartley's  $F_{max}$ , 167, 168
- Hotelling's  $T^2$ , 169, 170
- Hsu, 167
- Hurdle, 173
- intraclass, 175
- jackknife, 172
- Jarque-Bera, 165
- Jonckheere-Terpstra, 167, 168, 172
- Kendall, 175
- Kendall's  $\tau$ , 176
- Kolmogorov-Smirnov, 165, 167, 172
- Kruskal-Wallis, 167, 168, 172
- kurtosis, 165
- Lawley-Hotelling, 170
- Least significance Difference, 167
- Levene, 167, 168
- Lilliefors, 165
- Link-Wallace, 167
- log-logistic survival, 177
- log-rank, 177
- logistic regression, 170
- MANCOVA, 170
- Mann-Whitney U, 167, 171
- MANOVA, 169, 170
- MANOVA repeated, 169
- maximum, 174
- McNemar, 170, 174
- mean-square successive difference, 166
- median, 167, 168
- Mood's median, 167
- Moses, 167, 172
- multiple correlation, 176
- multiple regression, 169
- negative binomial, 172
- Newman-Keuls, 167
- non-negative matrix factorization, 170
- normality, 165
- odds ratio, 176
- ordered logistic regression, 168, 171
- Page, 168, 172
- partial correlation, 176
- Pearson's correlation, 176
- Peirce, 166
- permutation, 166, 171, 172
- Pillai-Bartlett, 170
- Poisson, 172, 173
- Poker, 174
- principal component analysis, 170
- randomness, 166, 174
- rank correlation, 166, 174, 175
- rank sum, 167
- repeated measures logistic regression, 175
- Roy's greatest root, 170
- run test, 166
- run test on successive differences, 166
- Sandler's A, 167
- Scheffe, 167
- semipartial correlation, 176
- sequential, 177
- serial correlation, 166

- Shapiro-Wilk, 165
  - Siegel-Tukey, 167, 172
  - sign, 166
  - skewness, 165
  - Snedecor's F, 167, 169, 173
  - Spearman's rank-order correlation, 176
  - Steel, 168
  - structural equation modelling, 170
  - Stuart-Maxwell, 174
  - Student's t, 165–167, 169
  - survival trees, 177
  - Tukey's Honestly Significant Difference, 167
  - Tukey-Duckworth, 167
  - Tukey-Kramer, 167
  - turning point, 166
  - van der Waerden, 167, 168, 172
  - w/s, 165
  - Weibull survival, 177
  - Wilcoxon inversion, 167
  - Wilcoxon matched-pairs signed-rank, 167, 172
  - Wilcoxon signed-rank, 166, 171
  - Wilcoxon-Mann-Whitney, 166, 167, 174
  - Wilks'  $\Lambda$ , 170
  - Yule's Q, 176
  - z, 165–167, 169, 173, 174
  - zero-inflated count model, 173
  - zero-truncated count model, 173
- Unit
- experimental, 25
  - observational, 27
- Variance reduction, 39
- bias, 28
  - blinding, 38
  - blocking, 32, 52, 55, 293
  - covariates, 54, 300
  - paired samples, 54
  - randomization, 35, 55
  - stratified randomization, 37