

# Statistical analysis and design

Carlos Óscar Sorzano (CNB, CSIC)

# Contents

1. Why this course?
2. The basics of Statistics
3. Understanding the logic behind statistical inference.
4. Sample size calculation
5. Software for sample size
6. Linear models
7. Some specificities



1

Why this course?

Why this course?

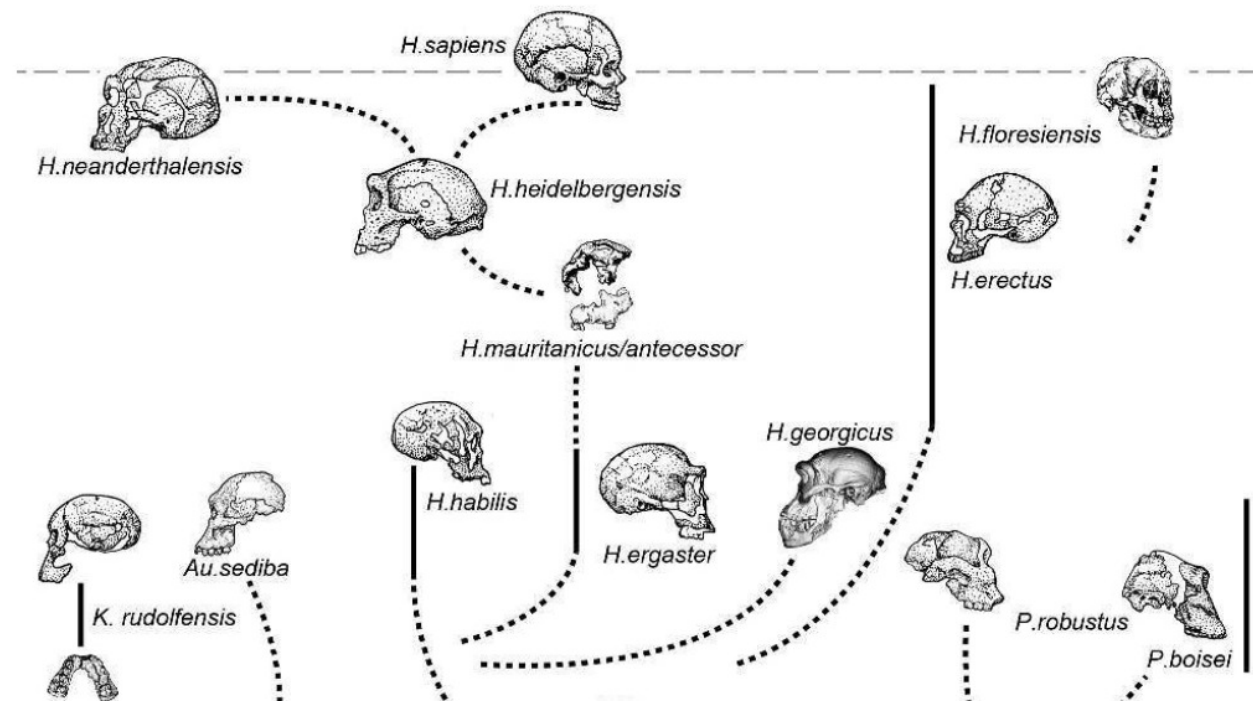
**“In God we trust.  
All others must  
bring data”.**

W. Edwards Deming





## Statistics is not intuitive.



Our evolutionary pressure was not on solving statistical problems ...  
so Statistics normally escapes from our intuition.

# Statistics is not intuitive

We don't realize that coincidences are common.

- We rank the grades of people in a class and study the characteristics of the people in the top 5. We realize that they are all scorpio, so we conclude that being born in November gives people an academic advantage.



We cannot conclude anything *a posteriori*. A different story is having the hypothesis that being scorpio gives an academic advantage, and verifying the hypothesis by analyzing the data from grades. Otherwise we may have found any other characteristic amongst the top 5 (being girls, wearing jeans, coming to school by bus, ...).

## Statistics is not intuitive

We don't naturally do Bayesian calculations.

- HIV affects 0.1% of blood donors. The antibody test correctly identifies 99% of infected samples, but it also incorrectly concludes that 1% of the noninfected samples have HIV. When this test identifies a problematic sample, what is the chance that it effectively has HIV?



If we have 100,000 donors, on average, only 100 ( $=0.1\%$ ) of them will have HIV. If we apply the test to these patients, 99 of them will be correctly identified (and 1 will escape). Of the remaining 99,900 donors (not having HIV), the test will be positive on 999 of them ( $=1\%$ ). Of the  $99+999=1,098$  positive tests, only 99 of them are HIV carriers. That is, **the chance of being HIV carrier if the test is positive is only  $99/1,098=9\%$ .**

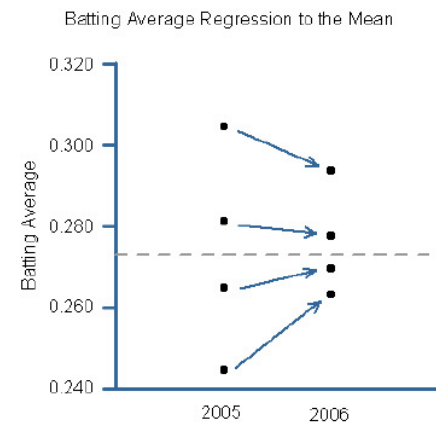
# Statistics is not intuitive



We are fooled by extreme values and regression to the mean.

- An athlete performs this season extremely well. Then he appears on the cover of Sports Illustrated. And next year, he performs worse than last season. **Appearing in Sports Illustrated brings bad luck to athletes!!**

**But we ignore that:** The athlete's performance may not have changed. Last season's performance may be an extreme from this distribution. Next draw from this distribution will most likely be from a more "central" region of the distribution.

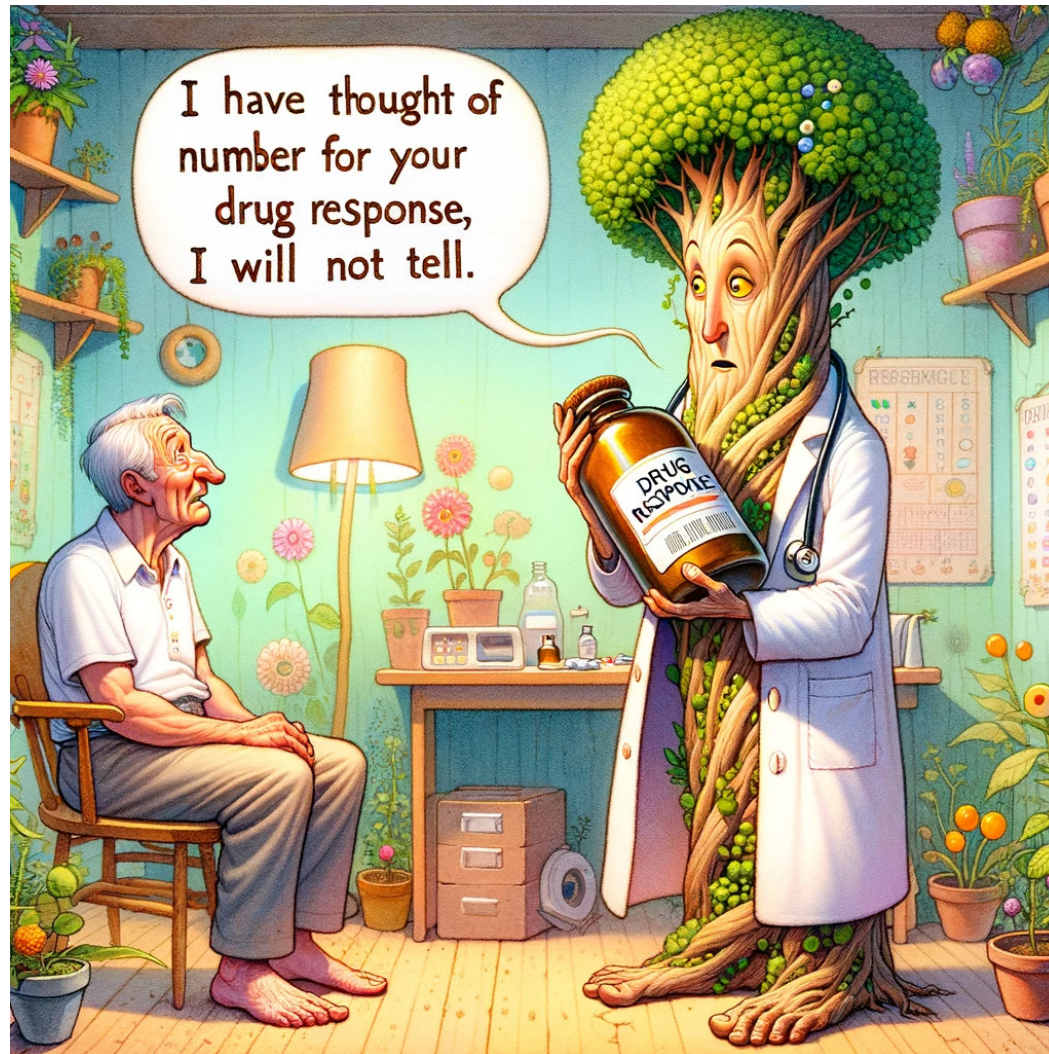




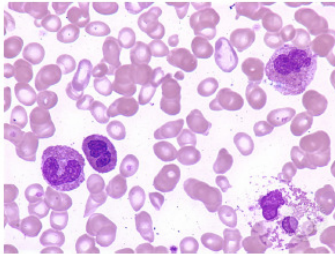
## 2

## The basics





## Discrete vs continuous variables



**Discrete data:** Number of eosinophils per microscopy field: 0, 1, 2, ...

**Continuous data:** pH of viable eosinophils: 6.00, 6.01, 6.02, ..., 7.49, 7.50

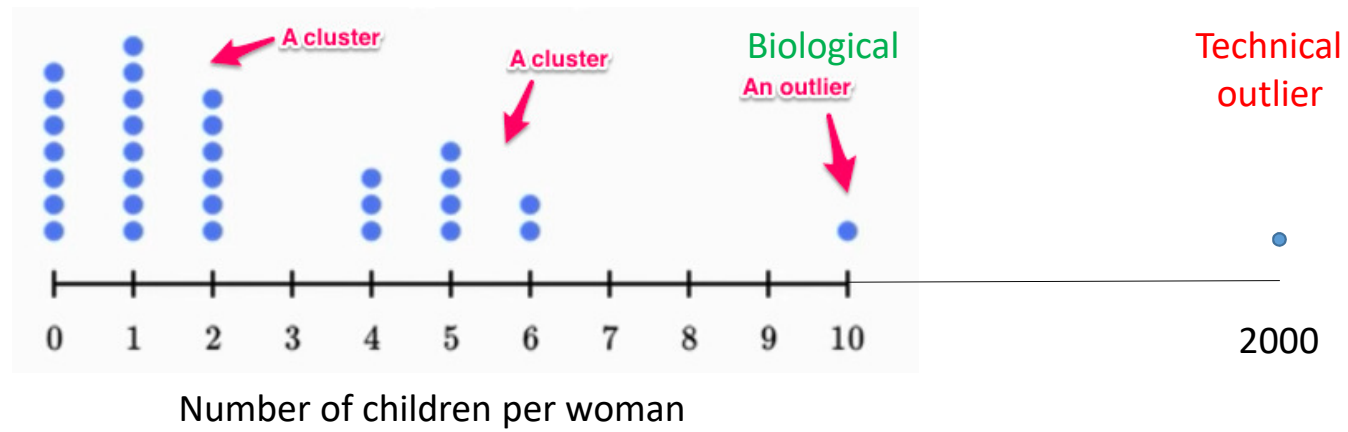
Assume that we measure the temperature to 9 people and get the data:

37.0, 36.0, 37.1, 37.1, 36.2, 37.3, 37.0, 37.0, 36.1

We may calculate a measure of centrality:

- **Mean:**  $\hat{\mu} = \frac{37.0+36.0+37.1+37.1+36.2+37.3+37.0+37.0+36.1}{9} = 36.76$
- **Median:**  $\hat{\mu} = (36.0, 36.1, 36.2, 37.0, 37.0, 37.0, 37.1, 37.1, 37.3) = 37.0$
- **Trimmed mean:**  $\hat{\mu} = \frac{\cancel{36.0}+36.1+36.2+37.0+37.0+37.0+37.1+37.1+\cancel{37.3}}{7} = 36.79$
- **Geometric mean:**  
 $\hat{\mu} = \exp\left(\frac{\log 36.0 + \log 36.1 + \log 36.2 + \log 37.0 + \log 37.0 + \log 37.0 + \log 37.1 + \log 37.1 + \log 37.3}{9}\right) = 36.75$
- **Harmonic mean:**  $\hat{\mu} = \frac{1}{\frac{1}{37.0} + \frac{1}{36.0} + \frac{1}{37.1} + \frac{1}{37.1} + \frac{1}{36.2} + \frac{1}{37.3} + \frac{1}{37.0} + \frac{1}{37.0} + \frac{1}{36.1}} = 36.75$
- **Mode:**  $\hat{\mu} = (36.0, 36.1, 36.2, 37.0, 37.0, 37.0, 37.1, 37.1, 37.3) = 37.0$

# Outliers





## Reference population

- World population
- Europeans
- Spanish
- Spanish women
- Spanish women aged between 20-40
- Spanish women aged between 20-40 with vitiligo
- Spanish women aged between 20-40 with vitiligo and with stressful jobs

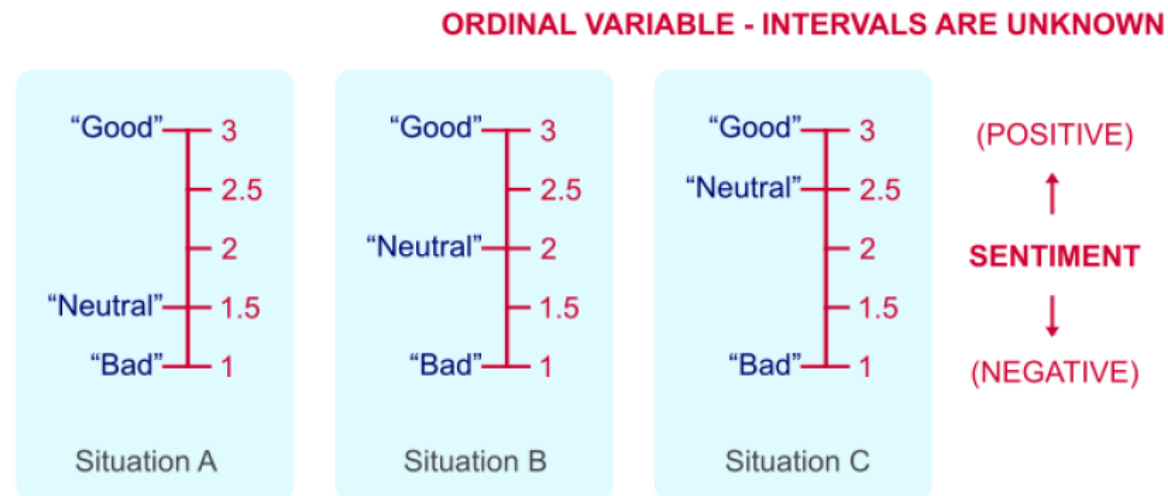
Skin levels of tyrosinase, TRP-1, TRP-2 and melanin



## Different measures of centrality

- **Mean**: Average of the input samples. The best for **normal** variables (heights, volumes, weights, ...)
- **Median**: Half the samples are below this value, and half the samples are above this value.
- **Trimmed mean**: Average removing the lowest and highest values. Robust to outliers.
- **Geometric mean**: Average in the logarithmic scale. The best for **log-normal** variables (number of cells, gene expression, ...)
- **Harmonic mean**: Average in the inverse scale. The best for speeds.
- **Mode**: The most frequent value (it does not necessarily be in the middle of the distribution).

## Types of variables: Ordinal variables



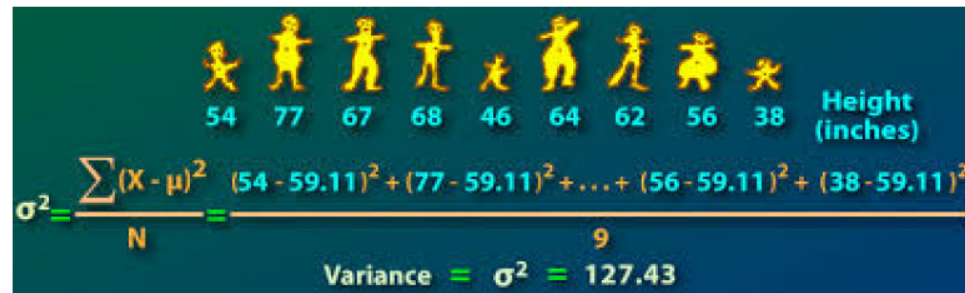
- Ordinal variables only express a relative rank between variables.
- Differences or ratios are meaningless.

## Types of variables: Categorical variables



- Categorical variables represent **labels** (male, female; no, yes; false, true; red, green, blue, ...; cat, dog, horse, ...)
- No mathematical operation is allowed even if they are encoded as numbers (0, 1, ...)

## Variability and bias



Variability may have different sources:

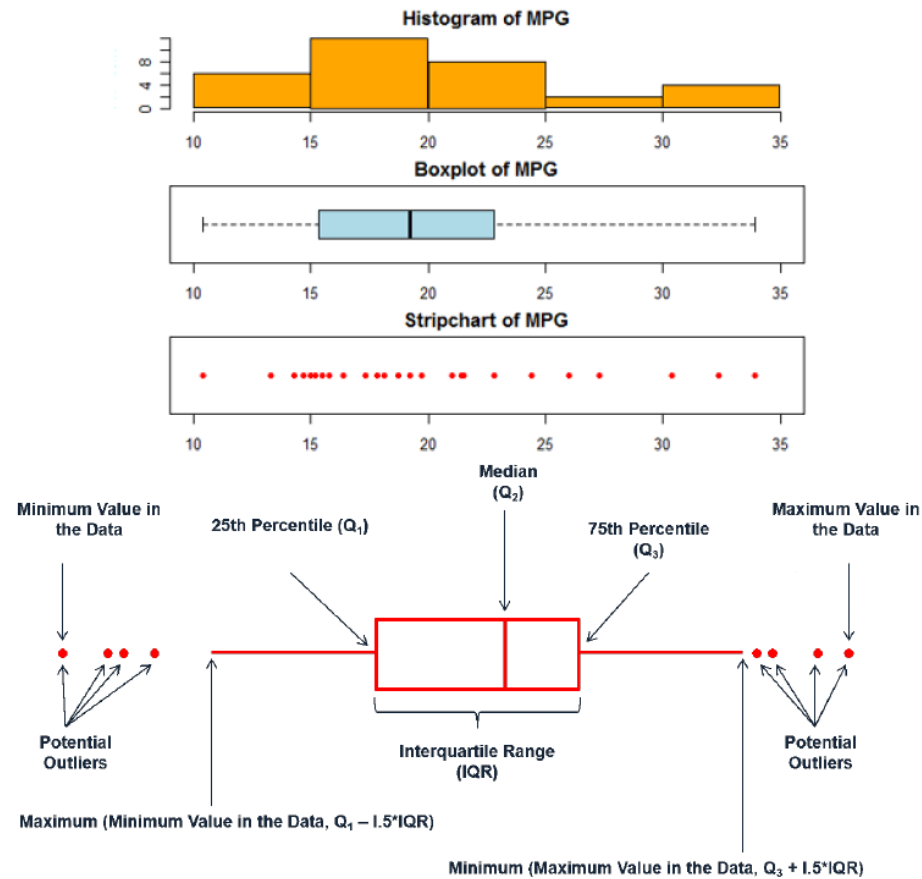
- **Biological:** There is an intrinsic variability associated to individuals.
- **Experimental random errors:** Reading (e.g. height) is subject to measurement errors (normally assumed to be Gaussian, but not necessarily)

Bias may have different sources:

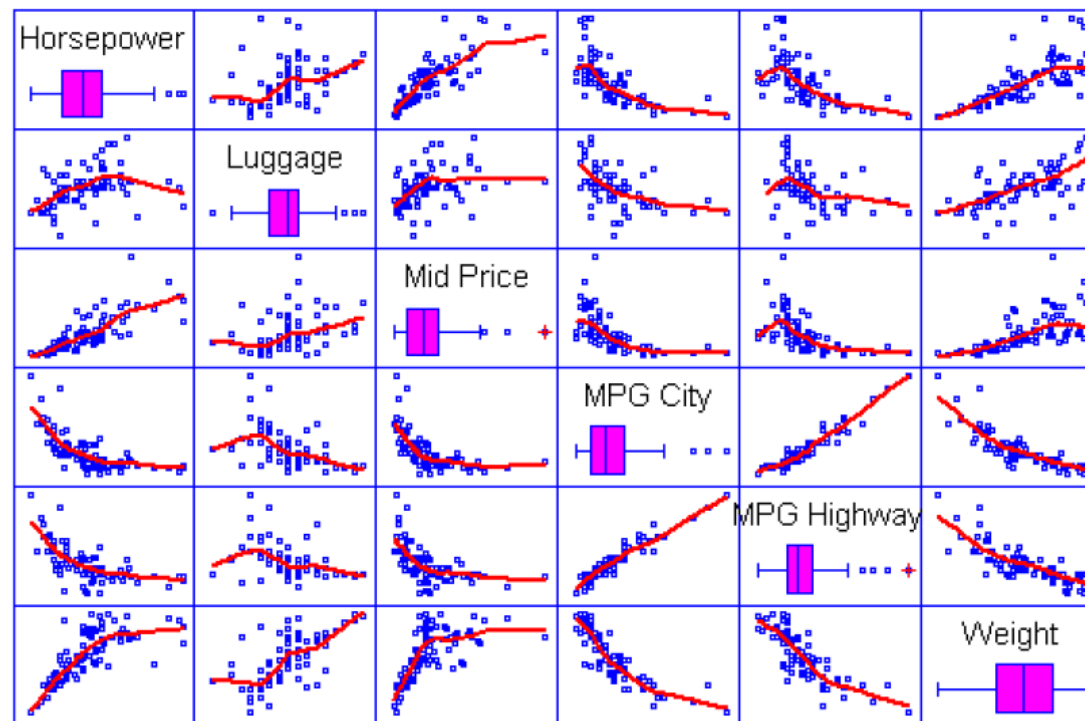
- **Systematic errors:** The instrument is wrongly used by the experimenter (zero offset, calibration, scale factors, ...), defective instruments, software bugs, ...

Bias data is not accurate.

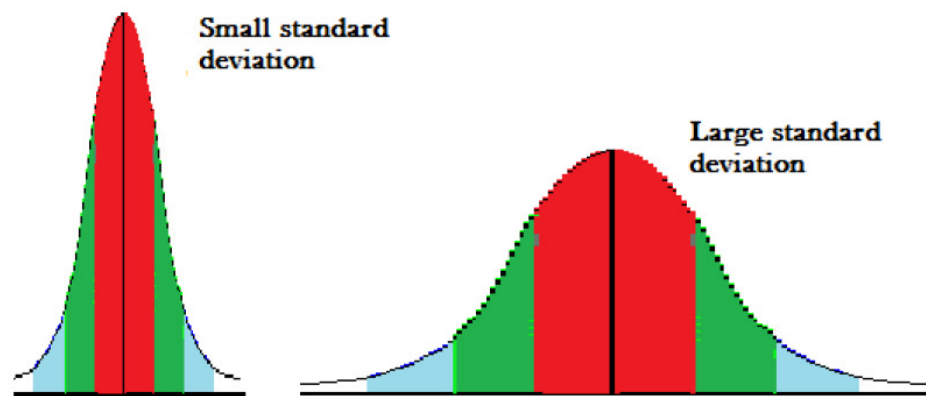
## Plots: 1D Scatter plots, histograms and boxplots



## Plots: 2D Scatter plots



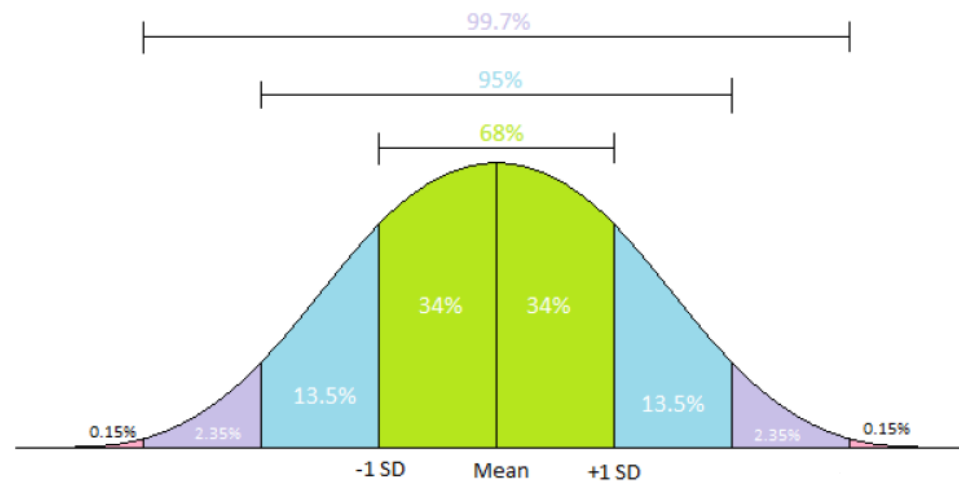
## Standard Deviation



The standard deviation (SD) expresses how samples differ from the average. For example, the average human temperature is  $36.82^{\circ}\text{C}$  with a SD of  $0.41^{\circ}\text{C}$ .

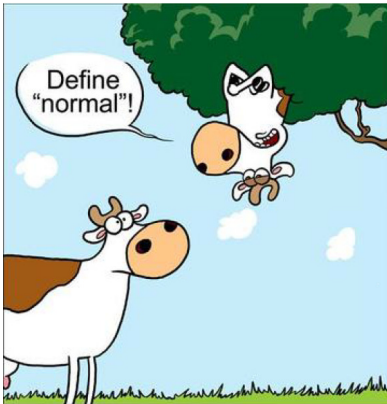


# Standard Deviation



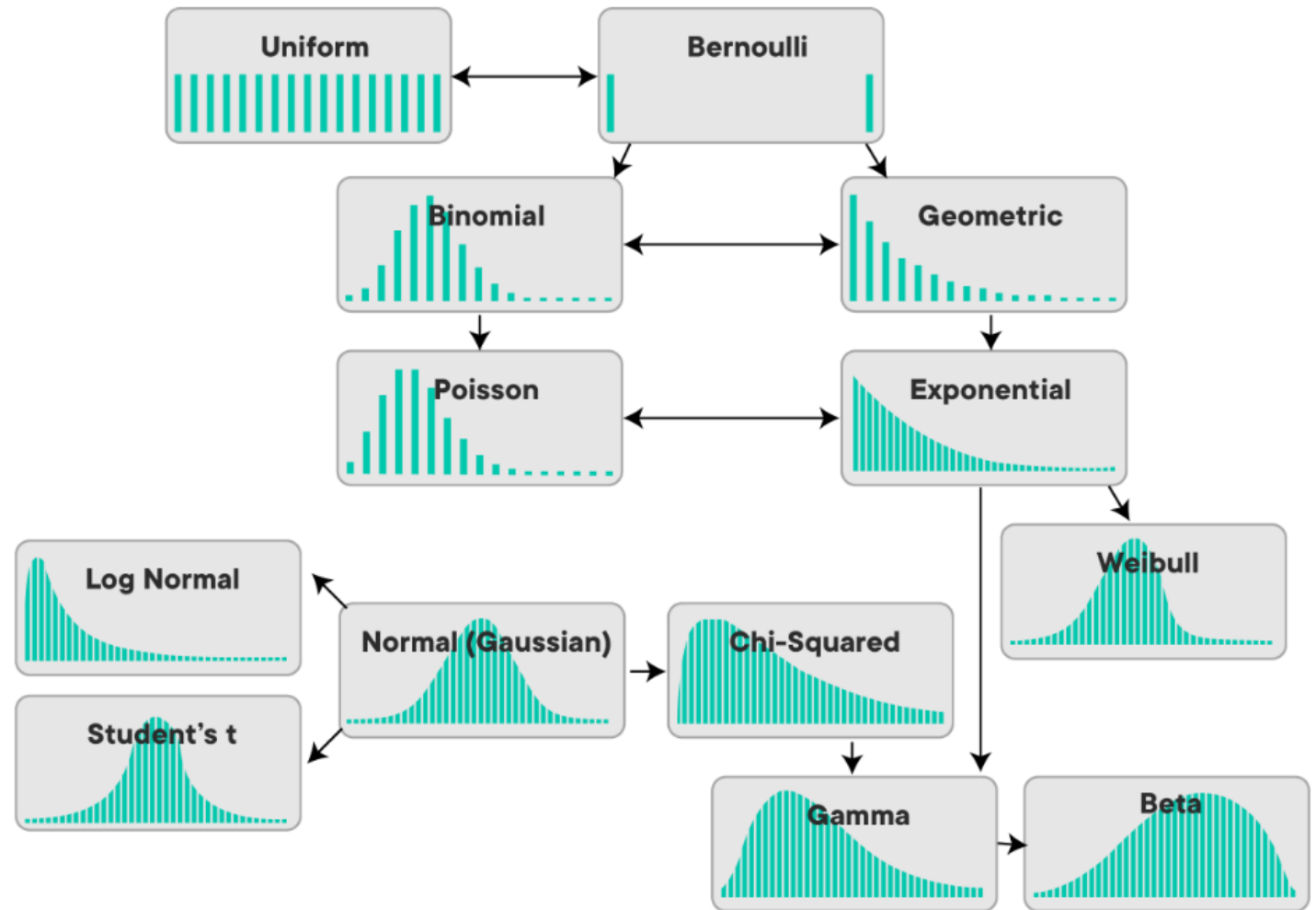
About 68% of the samples normally fall between  $\pm 1SD$ .  
About 95% of the samples normally fall between  $\pm 2SD$ .

# Distributions



Distribution	Examples
Gaussian, Normal	Height, BMI, Blood pressure
Log-normal	Length of hospital stays, Concentration of a chemical in blood, Viral load in patients
Binomial	Number of patients responding to a treatment, Incidence of a genetic trait, Success/failure of surgical procedures
Poisson	Number of new cases of a disease in a time period, Count of bacteria in a sample, Number of mutations in a DNA sequence
Exponential	Time until relapse of a disease, Survival time after a critical diagnosis, Time to infection after exposure
Chi-squared	Genetic linkage analysis, Analysis of vaccine adverse effects
Gamma	Time until failure of a biological system, Survival times in cancer research, Time between successive neuronal spikes

# Distributions



## Standard Deviation

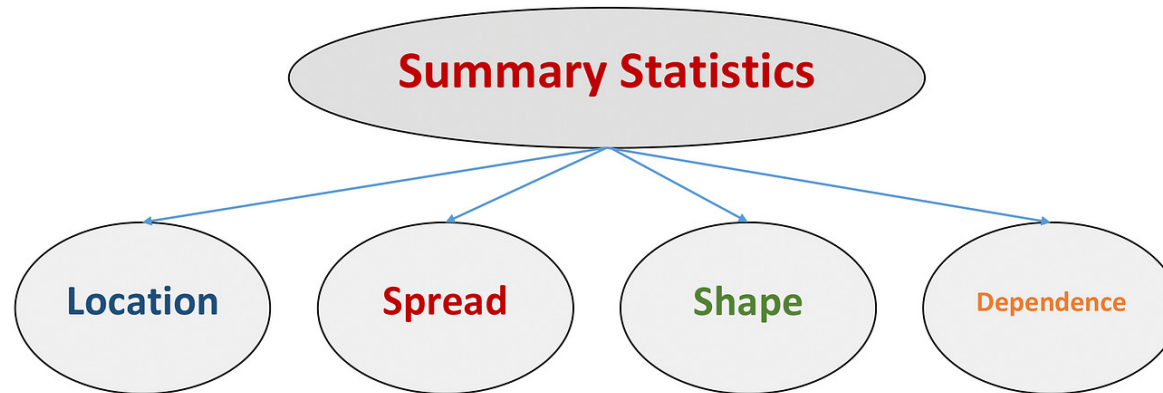
The **sample mean and standard deviation** are calculated as

$$\begin{aligned}\hat{\mu} &= \frac{1}{N} \sum_{i=1}^N x_i \\ \hat{\sigma} &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2\end{aligned}$$

Note that **sample variance** is the square of the standard deviation,  $\hat{\sigma}^2$ .

Means and standard deviations are sensitive to outliers. The equivalent robust estimates are the **median and median absolute deviation (MAD)**

$$\begin{aligned}med &= \text{med}(x_i) \\ mad &= \text{med}(|x_i - med|)\end{aligned}$$



**1) Mean**

- a) **Pythagorean Mean**
  - i) Arithmetic Mean
  - ii) Geometric Mean
  - iii) Harmonic Mean
- b) **Weighted Mean**
- c) **Truncated Mean**
- d) **Interquartile Mean**

**2) Mode**

**3) Median**

**1) Standard Deviation**

- 2) Variance**
- 3) Range**
- 4) Interquartile Range**
- 5) Absolute deviation**
- 6) Mean Absolute difference**
- 7) Distance Standard Deviation**

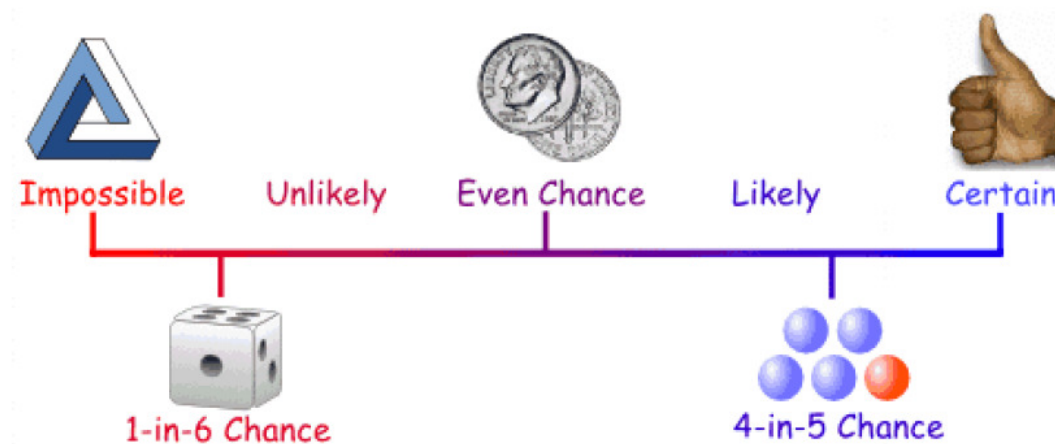
**1) Skewness**

- 2) Kurtosis**
- 3) Distance Skewness**

**1) Covariance**

- 2) Pearson Correlation**
- 3) Kendall Correlation**
- 4) Spearman Correlation**

# Probability



Probability is a number between 0 and 1 (=100%) that expresses our certainty about the occurrence of an event.

We may arrive to this probability by: 1) a model, or 2) by gathering data.

## Probability as a prediction from a model



We may establish a model for understanding the world:

- Each ovum has an X chromosome and none has a Y chromosome.
- Half the sperm have an X chromosome and the other half have a Y chromosome.
- Only one sperm will fertilize the ovum.
- Each sperm has an equal chance of fertilizing the ovum.
- If the winning sperm has a Y chromosome, then the embryo will be XY (boy).
- If the winning sperm has a X chromosome, then the embryo will be XX (girl).
- 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 boy or a girl.

## Probability based on data

In 2012, 51.7% of all babies born in the world were boys.

For a particular pregnant woman, the probability of having a boy is 51.7% ( $=0.517$ ).

If we take a group of 1000 pregnant women, we would expect to observe on average 517 male fetuses and 483 female fetuses.



This **does not mean** that if we take 1000 pregnant women, we **should observe** 517 male fetuses and 483 female fetuses.

It **means that** if we take **many (many) groups** of 1000 pregnant women, **and we average** the number of male and female fetuses of all these groups, as the number of groups go to infinity, **the average of male fetuses will approach** to 517 and the average of female fetuses will approach to 483.



# Understanding the assumptions of probability

Since in 2012 we have observed 51.7% of babies to be male, the probability of a new born being male is 51.7%.  
*Is that correct?* It is if:



- If the probabilities from the past can be used to predict the future. There is no change of the probability over the years.
- There is no change of the probability along the year (the male probability in January is the same as in July).
- There is no change of the probability along the race (the male probability for Africans is the same as for Asians).
- There is no change of the probability along region (the male probability in China is the same as in Japan).

## Well-defined probabilities (probability of what?)



Pierre Simon Laplace

$$Probability = \frac{\text{Positive results}}{\text{All possible outcomes}}$$

In our example

$$0.517 = \frac{\# \text{Male new borns}}{\# \text{All new borns}}$$

A lab test for VIH is 98% accurate.



**What does it mean?** With this information alone it is meaningless because it is an undefined probability. We don't know which are the positive cases and all possible outcomes!!

## Well-defined probabilities (probability of what?)

Interpretation 1: [Sensitivity](#).

Numerator: Correctly identified VIH cases in a group of people with VIH.

Denominator: Number of tested people (all of them had VIH).

Interpretation 2: [Specificity](#).

Numerator: Correctly identified non-VIH cases in a group of people not having VIH.

Denominator: Number of tested people (none of them had VIH).



Interpretation 3: [Predictive value of positive test](#).

Numerator: Correctly identified VIH cases.

Denominator: Number of people whose result with this test was positive.

Interpretation 4: [Predictive value of negative test](#).

Numerator: Correctly identified non-VIH cases.

Denominator: Number of people whose result with this test was negative.

## Conditional probabilities (probability of what?)

$$p(A|B(\text{given})) \neq p(B|A(\text{given}))$$



Thomas Bayes

- The probability that a Statistics book (given) is boring is not the same as the probability of a boring book (given) being about Statistics.

$$p(\text{boring}|\text{Statistics}) \neq p(\text{Statistics}|\text{boring})$$

- The probability that someone with abdominal pain (given) has appendicitis is not the same as the probability of someone with appendicitis (given) having abdominal pain.

$$p(\text{appendicitis}|\text{pain}) \neq p(\text{pain}|\text{appendicitis})$$

## Conditional probabilities (probability of what?)

$$p(A|B) \neq p(B|A)$$



Thomas Bayes

- The probability that a heroin addict (given) first used marijuana is not the same as the probability of a marijuana user (given) will later become addicted to heroin

$$p(\text{marijuana}|\text{heroin}) \neq p(\text{heroin}|\text{marijuana})$$

- The probability of a study for which the null hypothesis is true (given) having a p-value smaller than 0.05 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)

$$p(pval < 0.05|H_0) \neq p(H_0|pval < 0.05)$$

## Odds is different from probability



The odds is a ratio between two probabilities

- The odds of being a boy is

$$O = \frac{p(\text{boy})}{p(\text{girl})} = \frac{0.517}{0.483} = 1.07$$

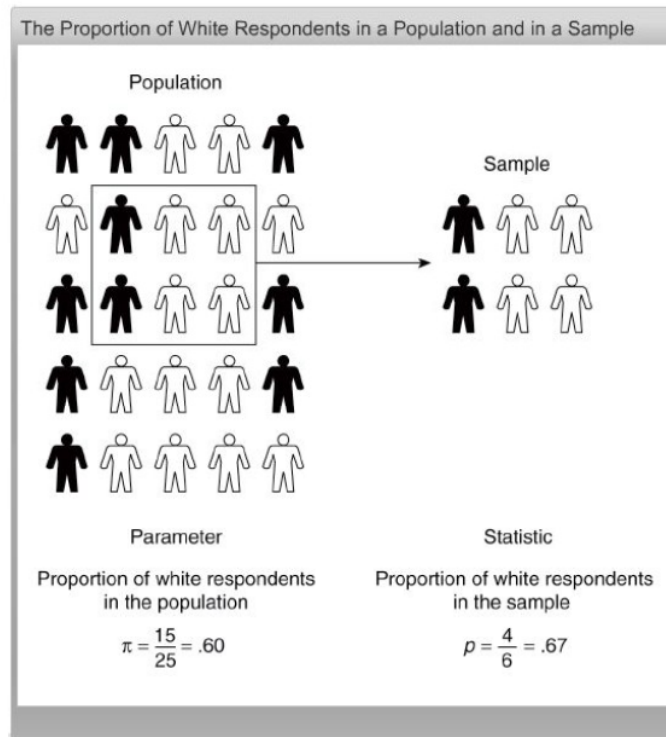
- The odds of developing a lung cancer if you smoke is 10 times larger than if you don't smoke.

$$O = 10 = \frac{p(\text{lung cancer}|\text{smoke})}{p(\text{lung cancer}|\text{don't smoke})} \Rightarrow$$

$$p(\text{lung cancer}|\text{smoke}) = 10p(\text{lung cancer}|\text{don't smoke})$$

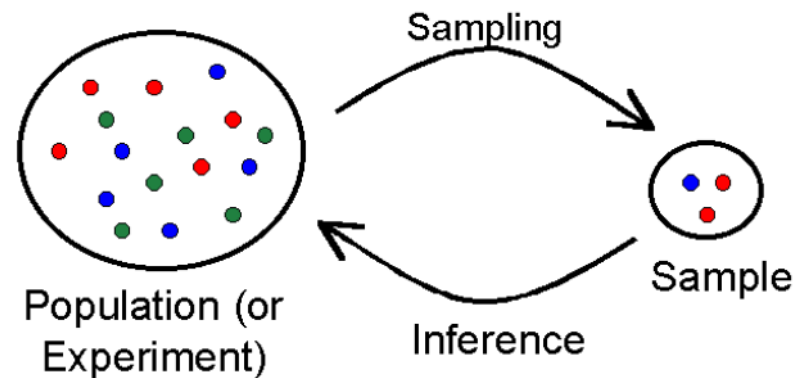


## From a sample to the population



From our calculations ([statistics](#)) performed on our sample we want to infer ([inference](#)) the true population parameters. In Biostatistics, we normally assume that our sample is small (<10%) than the population (normally considered to be [infinite](#)).

## Random sampling error



**Random sampling error.** Just by chance your sample might have a higher (or lower) mean/proportion/variance/correlation than that of the population.

Random sampling error decreases with the sample size.

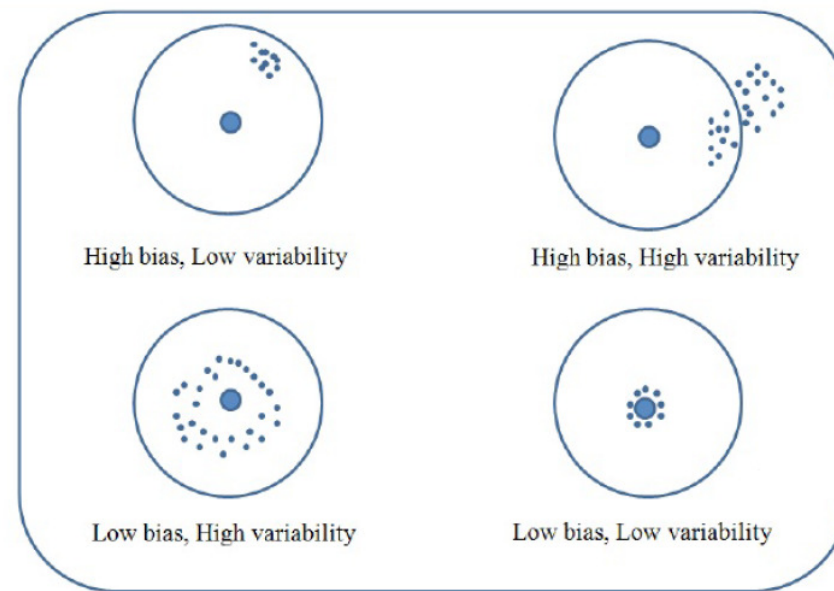


## Systematic errors

- **Non-response bias:** Individuals who do not respond to a call to participate in research studies behave differently from those who do respond.
- **Selection bias:** Studies performed in a hospital are not representative from the general population. The admissibility criteria may not represent the population.
- **Publicity bias:** Some individuals refer themselves to the investigator following publicity of the study (they have a particular interest in the disease being studied).
- **Healthy worker bias:** Voluntaries in studies may be particularly healthier as they are concerned about their own health and are predisposed to follow medical advice.
- **Overcoverage:** Including data from outside the population.
- **Undercoverage:** Sampling does not cover the whole population.
- **Measurement error:** Respondents fail to understand a question.
- **Processing error:** Mistakes in data coding.
- **Information bias:** Systematic misclassification of subjects.
- **Confounding:** The effect of one variable is mixed up with the effect of another variable (e.g., assessing the effect of smoking on lung cancer, but the average ages of the smoking and non-smoking groups are very different).

## Bias and variance

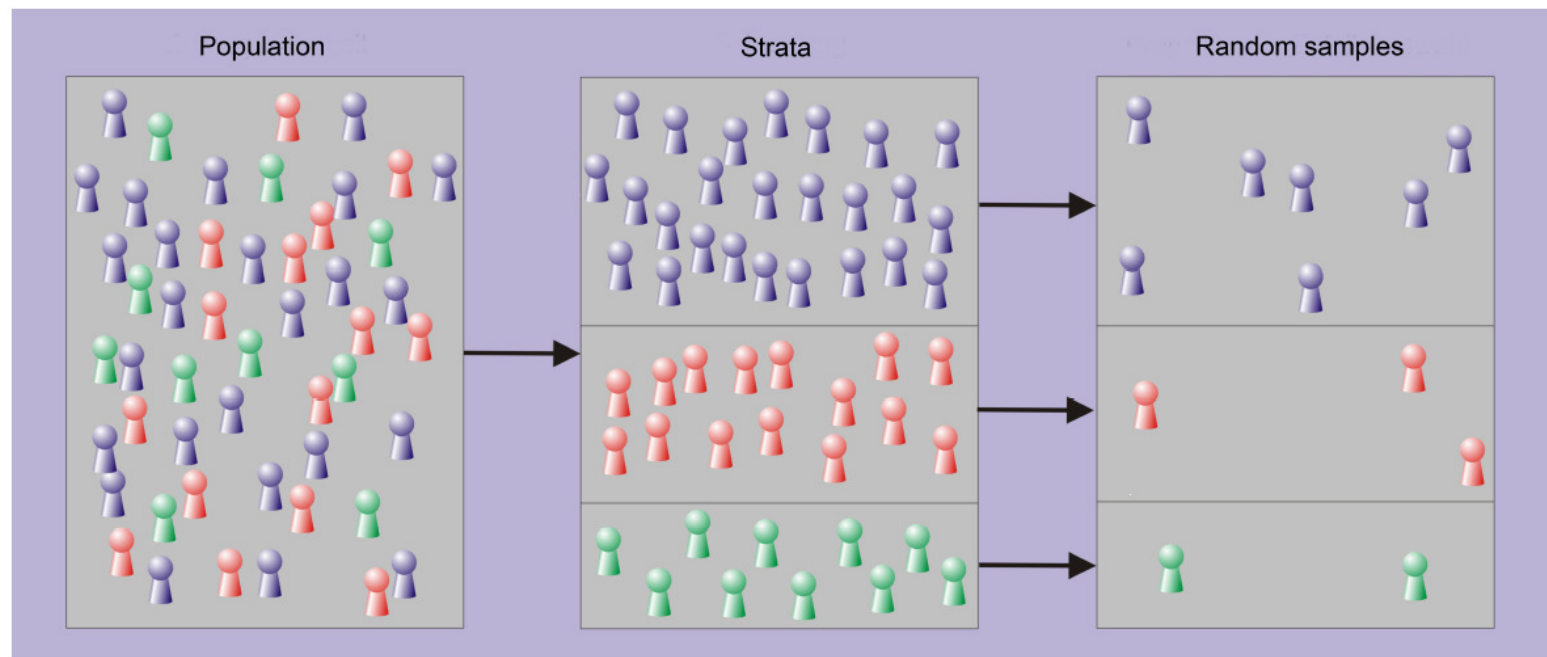
Figure 5. Bias and variability



D. Figueredo, et al. When is statistical significance not significant? Braz. Polit. Sci. Rev. 7 (2013)

Bias invalidate inference.

# Stratified sampling



Stratified sampling helps undercoverage.



# 3

Understanding the logic  
behind Statistical Inference

# Statistical hypothesis testing



Statistical hypothesis testing helps to automate decision making:

- In a pilot experiment, we must decide whether to proceed to further experimentation with this drug.
- At Phase II, we must decide whether to go to Phase III.
- At production quality control, we must decide if a batch can be released.

## Innocent until proven guilty

- A juror starts with the **presumption of innocence** of the defendant.
  - A juror bases his decision only on **factual evidence** presented at the trial and should not consider any other information (e.g., newspaper stories).
  - A juror reaches the **verdict of guilty** when the evidence is inconsistent with the assumption of innocence.
  - Otherwise, the juror reaches the verdict of **non-guilty**.
  - If the juror is not convinced, he can say "**I'm not sure**".
- A scientist starts with the presumption that the **null hypothesis** "there is no difference" is true.
  - A scientist bases his decision only on **data from one experiment**, without considering what other experiments have concluded.
  - A scientist reaches the conclusion of statistical **significant difference** when the p-value is small enough to make the null hypothesis very unlikely.
  - Otherwise, the scientist reaches the conclusion of **non-significantly different**.
  - If the scientist is not sure, he can **collect more data**.

## Some concepts

**Table 7-2 Type I and Type II Errors**

		True State of Nature	
		The null hypothesis is true	The null hypothesis is false
Decision	We decide to reject the null hypothesis	Type I error (rejecting true null hypothesis) <b>FP</b>	Correct decision <b>TP</b>
	We fail to reject the null hypothesis	Correct decision <b>TN</b>	Type II error (failing to reject a false null hypothesis) <b>FN</b>

## Some concepts

**Figure 4: Type I and Type II Errors in Hypothesis Testing**

<i>Decision</i>	<i>True Condition</i>	
	$H_0$ is true	$H_0$ is false
Do not reject $H_0$	Correct decision	Incorrect decision <b>Type II error</b>
Reject $H_0$	Incorrect decision <b>Type I error</b> Significance level, $\alpha$ , = P(Type I error)	Correct decision Power of the test = $1 - P(\text{Type II error})$

<b>Symbols</b>	<b>Phrase</b>	<b>p-value</b>
ns	Not significant	$p > 0.05$
*	Significant	$p < 0.05$
**	Highly significant	$p < 0.01$
***	Extremely significant	$p < 0.001$



# Basics of statistical inference

## Research hypothesis:

The new vaccine reduces the number of infected animals in a population.

$$H_0 : \pi \geq \pi_0 \quad \text{One-tail test}$$

$$H_1 : \pi < \pi_0$$

## Research hypothesis:

The new drug increases survival for patients with this disease in the next 5 years.

$$H_0 : S \leq S_0 \quad \text{One-tail test}$$

$$H_1 : S > S_0$$

## Research hypothesis:

The new machine does not produce tablets with the prescribed concentration

$$H_0 : c = c_0 \quad \text{Two-tail test}$$

$$H_1 : c \neq c_0$$

## Basics of statistical inference

- You CAN reject the null hypothesis and accept the alternative hypothesis
- You CAN fail to reject the null hypothesis because, there is not sufficient evidence to reject it
- You CANNOT accept the null hypothesis and reject the alternative because you would need to measure absolutely all elements (for instance, all hypertense patients).

It's like in **legal trials**:

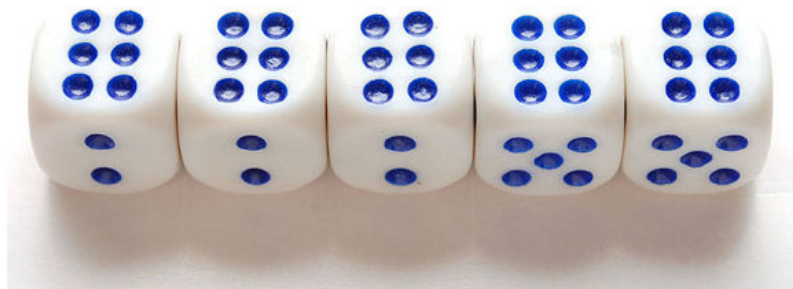
- The null hypothesis is the innocence of the defendant.
- You CAN reject his innocence based on proofs (always with a certain risk).
- You CAN fail to reject his innocence.
- You CANNOT prove his innocence (you would need absolutely all facts)



# Basics of statistical inference

The goal of hypothesis testing is to disprove the null hypothesis! We do this by proving that if the null hypothesis were true, then there would be a very low probability of observing the sample we have actually observed.

However, there is always the risk that we have been unlucky with our sample, this is our **confidence level** (the **p-value** is also related to this risk: the lower the p-value, the lower the risk).



## Not significant results



The other day [Michael Jordan](#) and me shot baskets. He shot 7 straight free throws. I hit 3 and missed 4. Being a statistician, I rushed to the sideline, calculated the p-value by Fisher's exact test which resulted to be 0.07. That meant, **there was no statistically significant difference between Michael Jordan and me!!!**

A **high p-value does not make the null hypothesis true**. It may be that the experiment was **not large enough**.

## Some mistakes

- **Stargazing**: Considering results in a paper only important if they have 1, 2, 3, ... stars. p-values are not as reproducible as CIs, and they only mean at showing 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.  $4.9999 \neq 5.0000$
- **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.

# CIs and hypothesis testing

These two techniques are based on the same theory

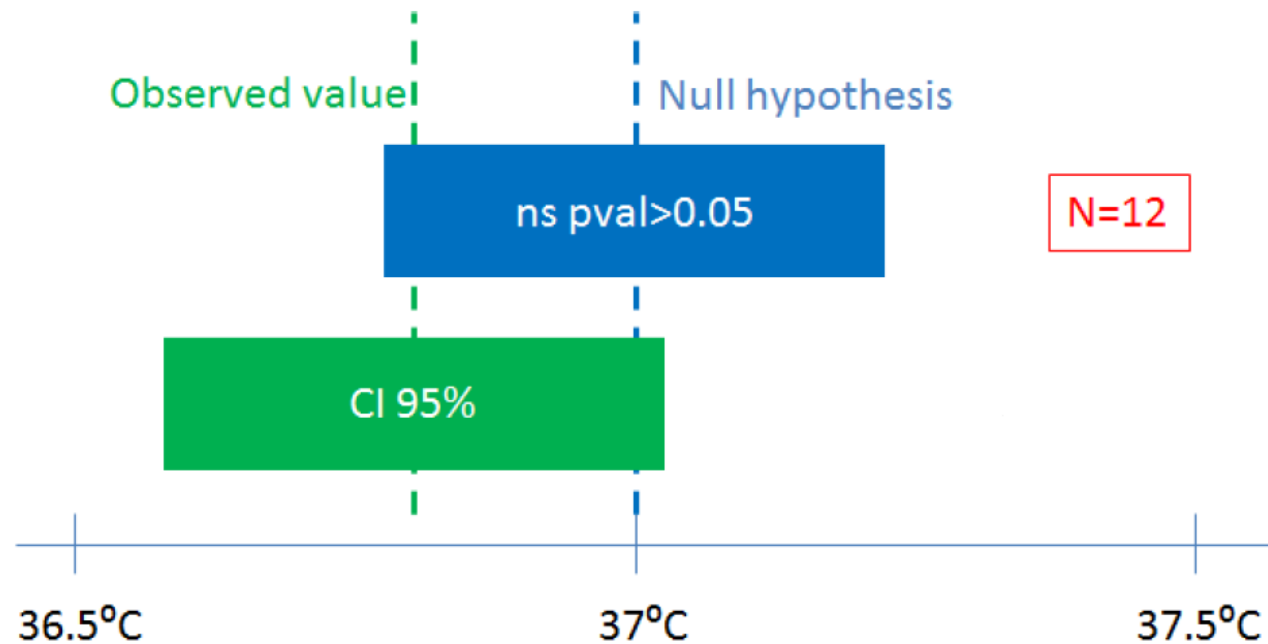
- CIs compute a range that 95% of the time will contain the population value (given some assumptions).
- Hypothesis testing computes a range that you can be 95% sure would contain the experimental results if the null hypothesis were true. Any result within this range is considered not statistically significant, and any result outside this range is considered statistically significant.

Remember

- If the 95% CI does not contain the value of the null hypothesis, then the result must be statistically significant (with  $p < 0.05$ ).
- If the 95% CI does contain the value of the null hypothesis, then the result is not statistically significant (with  $p < 0.05$ ).

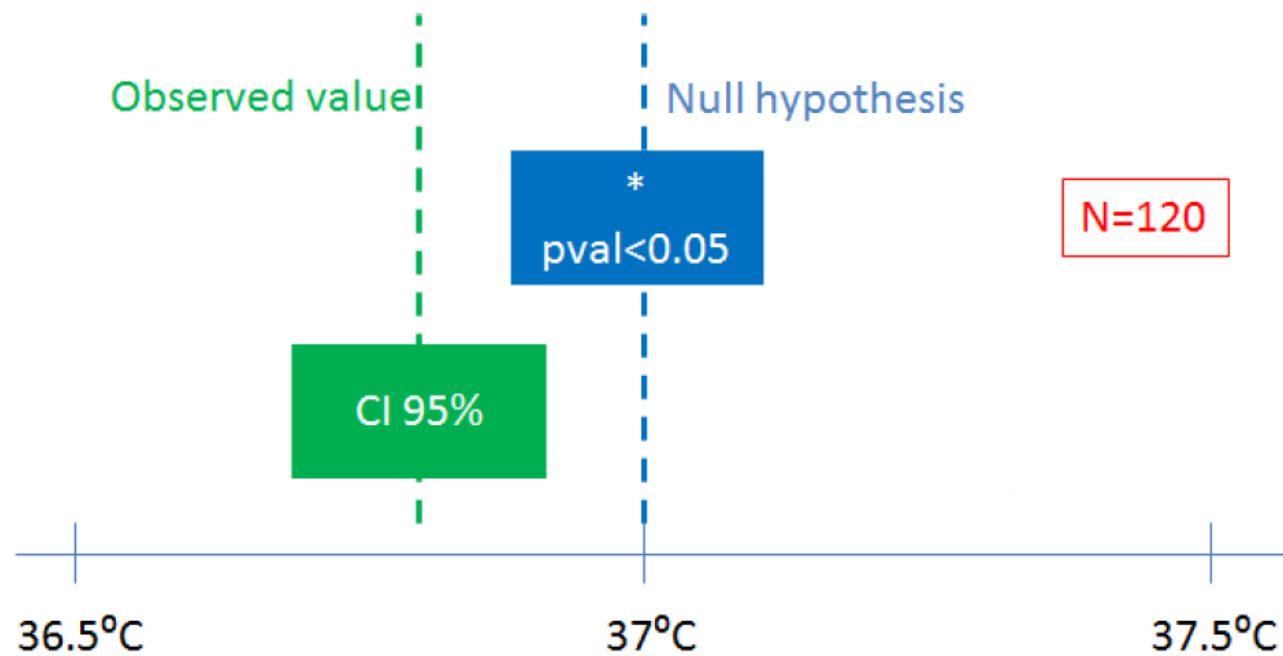
## CIs and hypothesis testing

With  $N = 12$  measurements we observe some difference between the average observed temperature and the reference (null) value ( $37^{\circ}\text{C}$ ). However, this result is **not significant**



## CIs and hypothesis testing

With  $N = 120$  measurements the result becomes significant





## Statistical significance does not imply relevance

We compare the responding proportion in a control and treatment group

<b>Sample size per group</b>	<b>Control</b>	<b>Responding</b>	<b>pval</b>	<b>CI 95%</b>
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).

## 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.

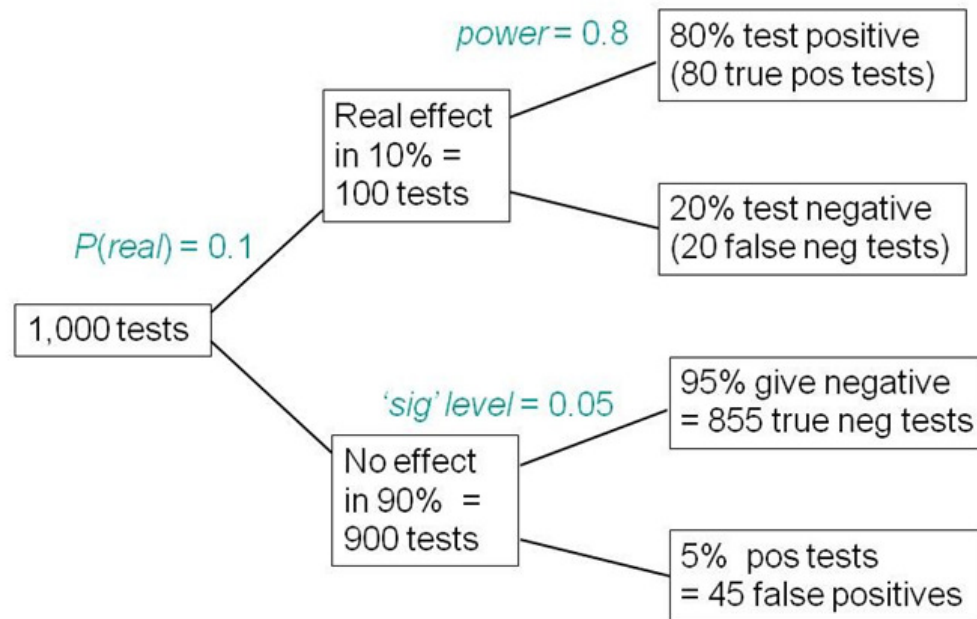
The null hypothesis is that the risk of adverse outcome is the same in both groups. The relative risk is 1.01 (=4.98/4.91) and has a **95% confidence interval [0.88,1.17]** and the **p-value is 0.86**.

Possible interpretations:

- 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 the increase because ultrasounds are harmful to the fetus?**

# Statistical power

## Significance tests



Total number of positive tests = 80 + 45

False discovery rate (proportion of false positives)  $\frac{45}{45 + 80} = 36$  percent (NOT 5%)

## Significance level, power and false discovery rate

	Reject $H_0$	Do not reject $H_0$	
$H_0$ is true	$A = 45$	$B = 855$	$A + B = 900$
$H_0$ is false	$C = 80$	$D = 20$	$C + D = 100$
	$A + C = 125$	$B + D = 875$	$A + B + C + D = 1000$

### Significance level

$$\alpha = \frac{A}{A + B} = \frac{45}{900} = 0.05$$

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?

## Significance level, power and false discovery rate

	Reject $H_0$	Do not reject $H_0$	
$H_0$ is true	$A = 45$	$B = 855$	$A + B = 900$
$H_0$ is false	$C = 80$	$D = 20$	$C + D = 100$
	$A + C = 125$	$B + D = 875$	$A + B + C + D = 1000$

Power

$$1 - \beta = \frac{C}{C + D} = \frac{80}{100} = 0.80$$

$$\beta = \frac{D}{C + D} = \frac{20}{100} = 0.20$$

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?

## Significance level, power and false discovery rate

	Reject $H_0$	Do not reject $H_0$	
$H_0$ is true	$A = 45$	$B = 855$	$A + B = 900$
$H_0$ is false	$C = 80$	$D = 20$	$C + D = 100$
	$A + C = 125$	$B + D = 875$	$A + B + C + D = 1000$

### False Discovery Rate

$$FDR = \frac{A}{A + C} = \frac{45}{125} = 0.36$$

FDR 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?

# Significance level, power and false discovery rate

Significance level, statistical power and FDR depend on the **sample size**, the **effect size** and the **population variance**.



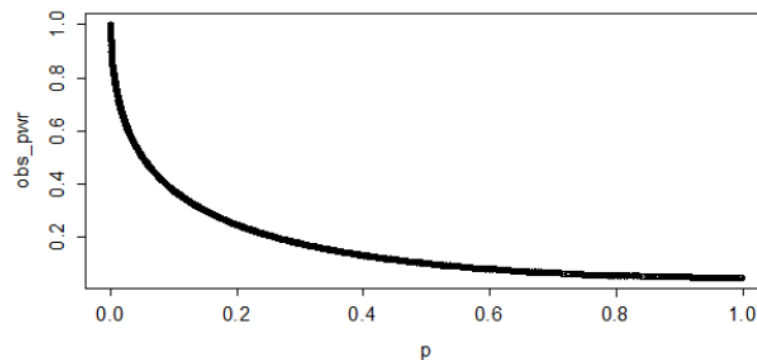
You send your child into the basement to find a tool. He comes back and says “It isn’t there”. What do you conclude? Is the tool there ( $H_0$ ) or not ( $H_1$ )?

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**)

## Post-hoc power analysis (Don't)

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.

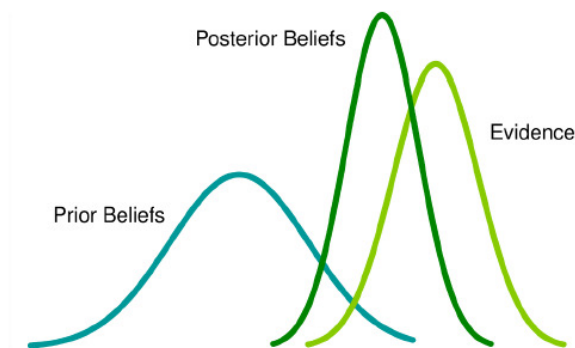


Unfortunately, post-hoc power is simply another way of reporting the p-value. There is a close relationship between the observed power and the observed p-value. If you want to look at your experiment retrospectively, look at the CI.



## Informal accounting for prior probabilities

- Experiment 1: The experiment makes biological sense and the p-value is 0.04. I would tend to believe that  $H_0$  is false and that the data confirms my alternative hypothesis.
- Experiment 2: The experiment does not make biological sense and the p-value is 0.04. I would tend to believe that  $H_0$  is true and that the observations are significant just by chance.
- Experiment 3: The experiment does not make biological sense and the p-value is 0.0000004. Although, for me, the experiment goes against my biological knowledge, the data evidence is so strong that probably  $H_0$  is false and I have to revise my knowledge base.  
(*Extraordinary claims require extraordinary proofs (Carl Sagan)*).



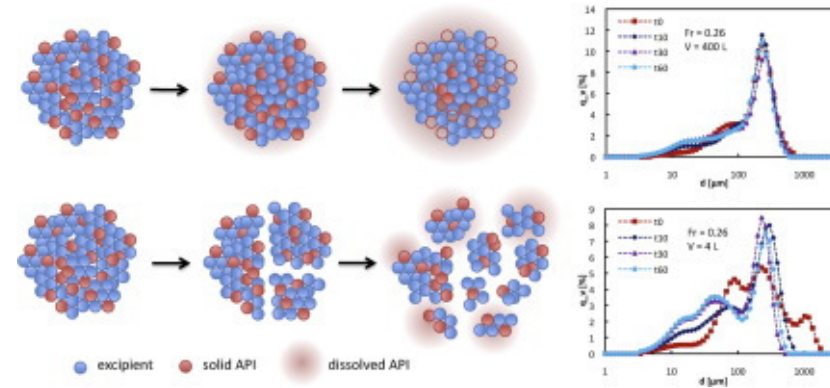
# •4

## Sample size calculation

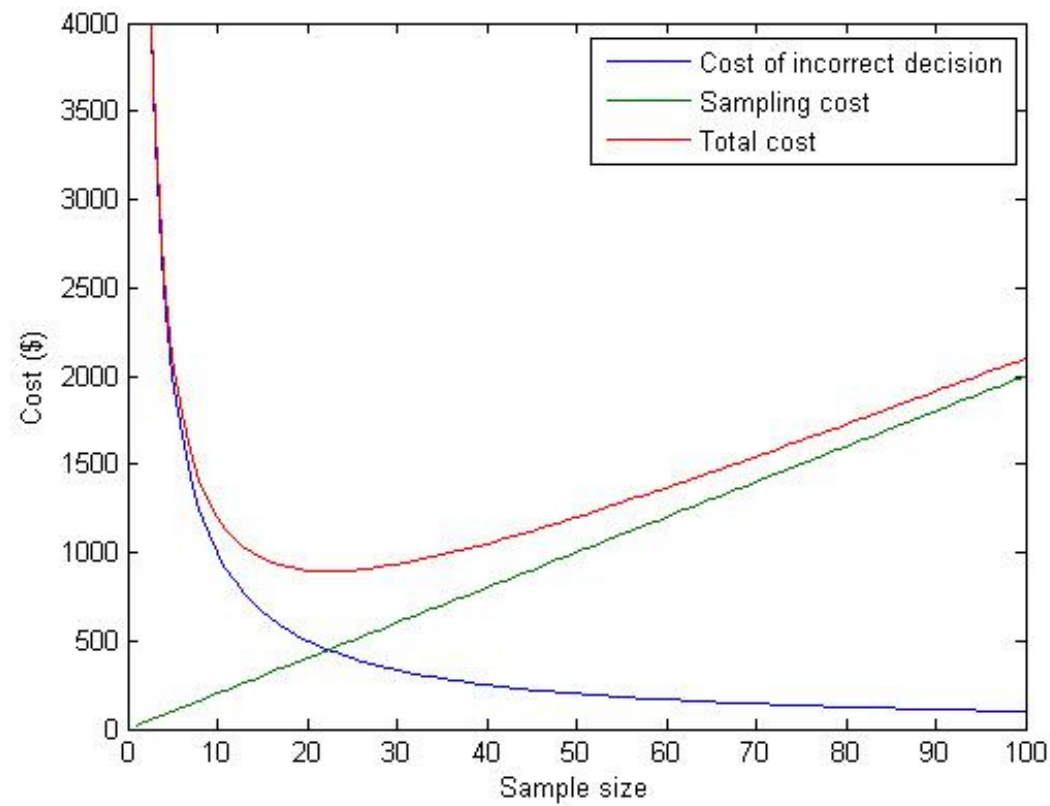
$N = \dots$

# Drug research

- We want to **determine the differences** between the dissolution in two delivery vehicles. How many samples do we need to observe to show that there is a **statistically significant** difference of at least 20% between any two groups. The nominal value in the reference drug is 40  $\mu\text{g}/\text{min}/\text{cm}^2$ . The standard deviation is 7. We want to have a statistical power of 90% and a confidence level of 95%.
- Using **too many or too few** samples is a loss of time and money.
- We must use **what it takes**; in this example, 18 per group.

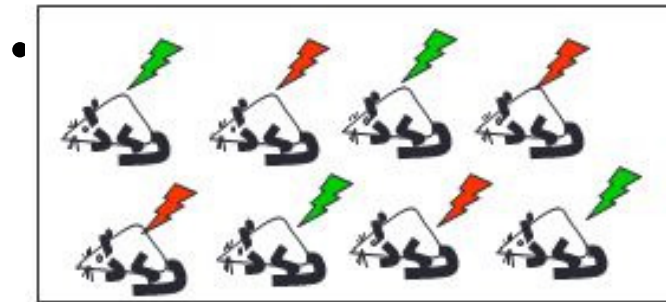


## Sample size

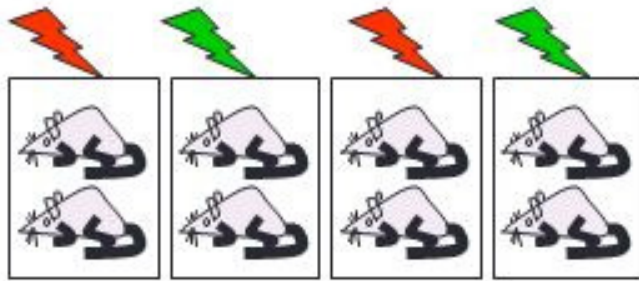


## Experimental unit

- The experimental unit is the smallest fraction of the experimental material where we can change the treatment.

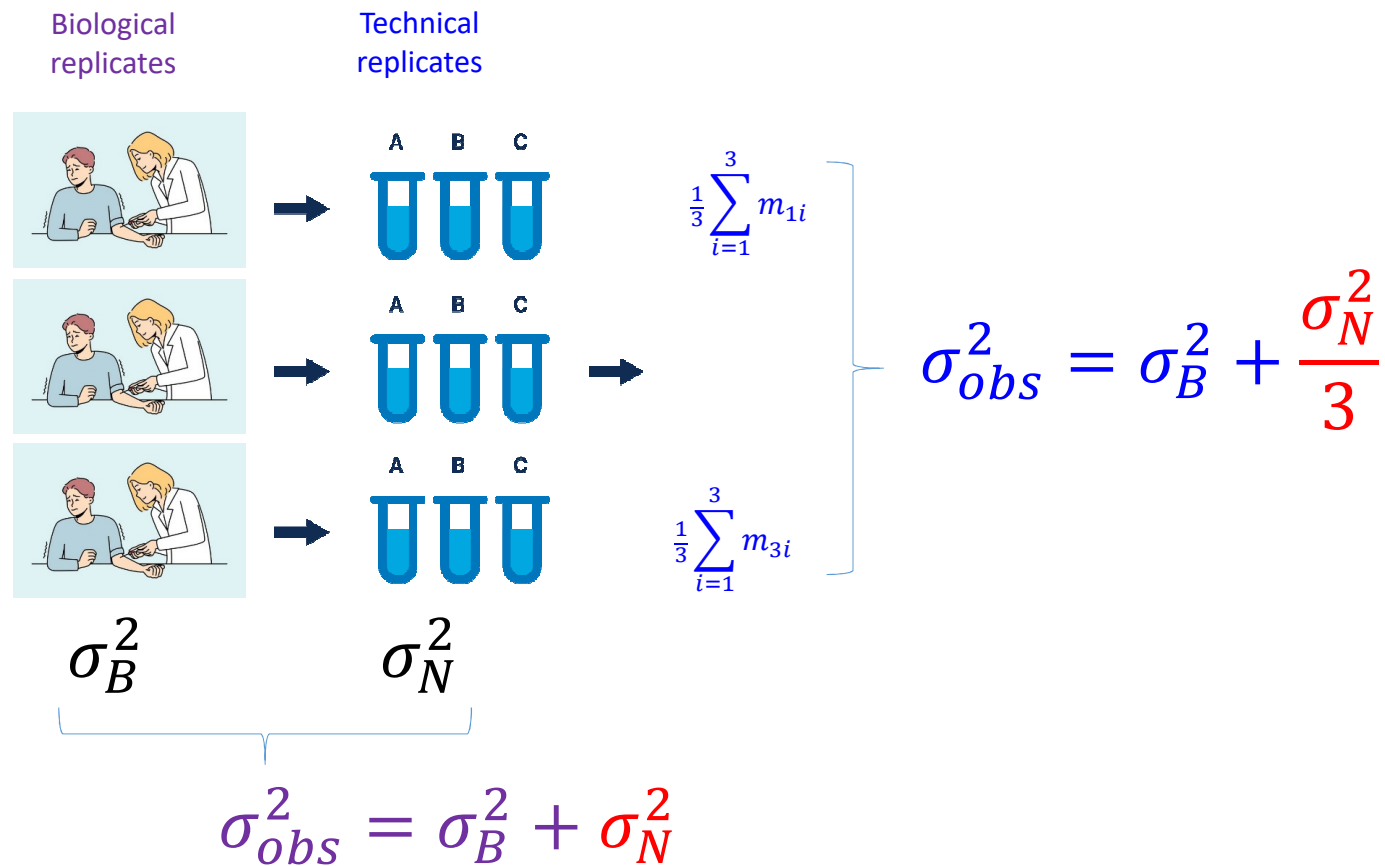


Treatment is injected

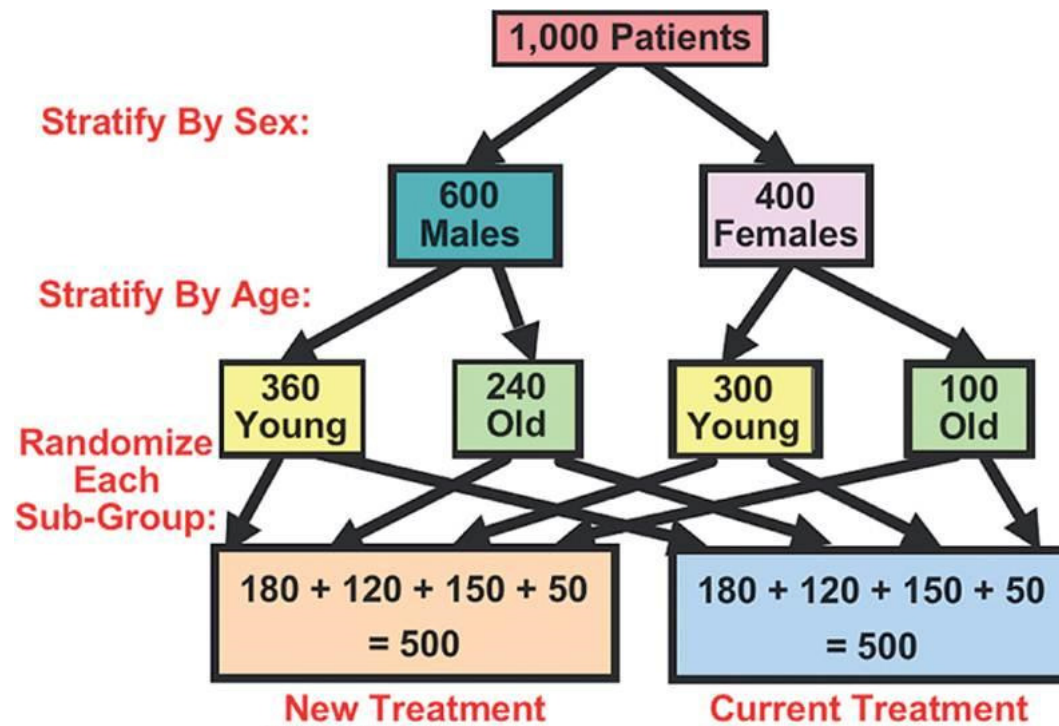


Treatment is in the water

# Experimental units

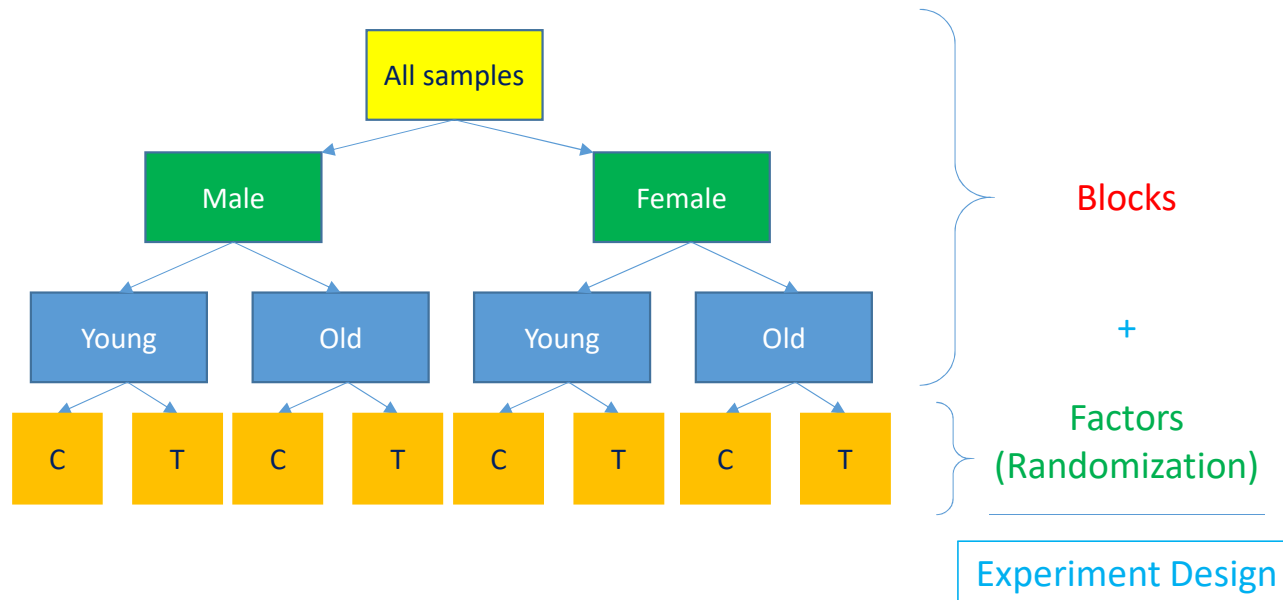


# Randomization and blocking



Gordis: Epidemiology, 4th Edition.  
Copyright © 2008 by Saunders, an imprint of Elsevier, Inc. All rights reserved

# Experiment design



Blocks			Treatments	
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



# Randomization and blocking

## Blocking time

20 females, 20 males. 20 treated, 20 controls. We can only process 4 animals/day → 10 days

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T

T = treated, C = control, pink = female, blue = male



# Randomization and blocking

## Blocking time

20 females, 20 males. 20 treated, 20 controls. We can only process 4 animals/day → 10 days

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
T	T	T	T	T	C	T	T	C	T
C	T	T	T	T	C	C	C	T	C
C	C	C	T	T	C	C	T	C	C
T	C	C	C	C	C	T	C	T	T

T = treated, C = control, pink = female, blue = male



# Randomization and blocking

## Blocking time

20 females, 20 males. 20 treated, 20 controls. We can only process 4 animals/day → 10 days

Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	T	T	C	T	C	C	T	C	T
T	T	C	C	C	T	T	T	C	C
C	C	T	T	C	C	T	C	T	C
T	C	C	T	T	T	C	C	T	T

T = treated, C = control, pink = female, blue = male



# Randomization and blocking

## Randomized block design

3) **Randomize** the rest.

2) **Block** what you cannot.

1) **Control** what you can.

- Randomize the position in the shelf.
- Randomize feeding order.
- Randomize treatment time.
- Randomize treatment order.
- Randomize ...

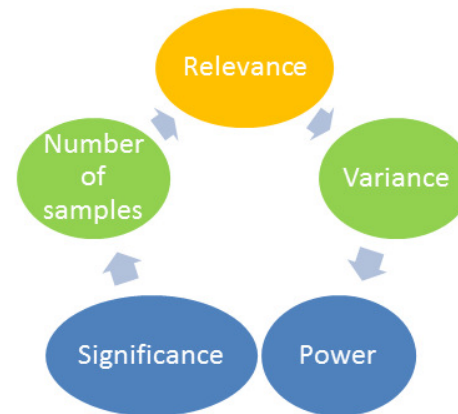


# Sample size calculation

## Unfounded fear



- **Fear**: A statistical design of the experiment will require “thousands” of mice.
- **Reality**: A statistical design of the experiment relates



# Sample size calculation

## How many mice do I need for my experiment?



It depends on:

- Experimental constraints (How the data is collected)
  - I need a minimum amount of material
  - Some mice die before I can measure
  - I cannot handle more than 100 mice at a time
  - Sometimes mice move while I'm injecting
  - ...
- Statistical constraints (How the data will be used)
  - I will perform a comparison to a control group
  - I will compare the mean of the two groups
  - The data is normally distributed
  - This is the first experiment ever!
  - ...

# Sample size calculation

## Stage 1: Arrival



# Sample size calculation

## Stage 1: Arrival





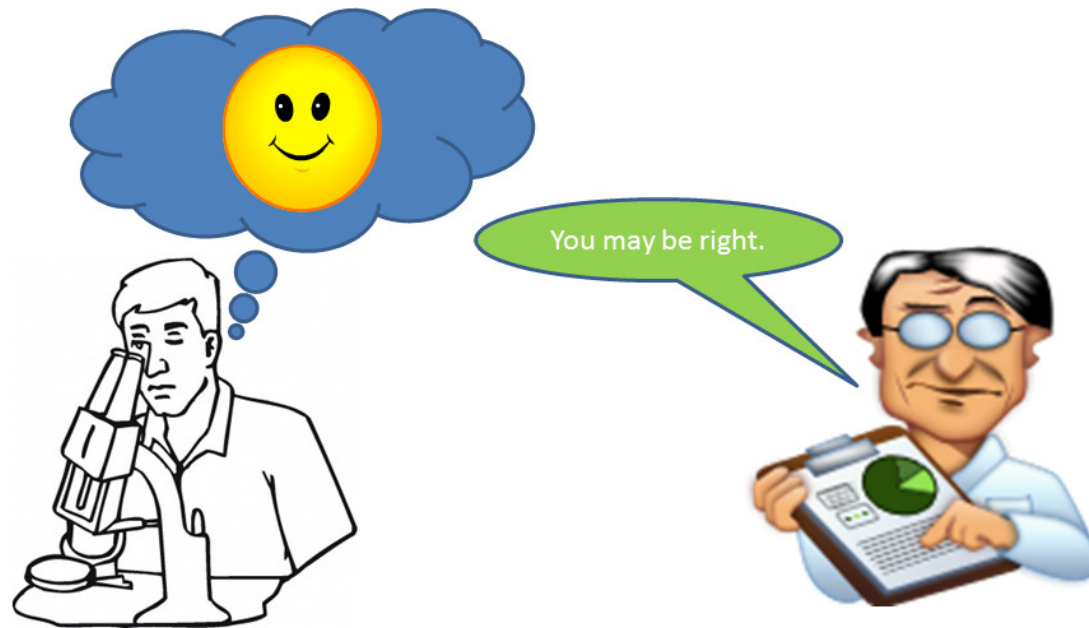
# Sample size calculation

## Stage 1: Arrival



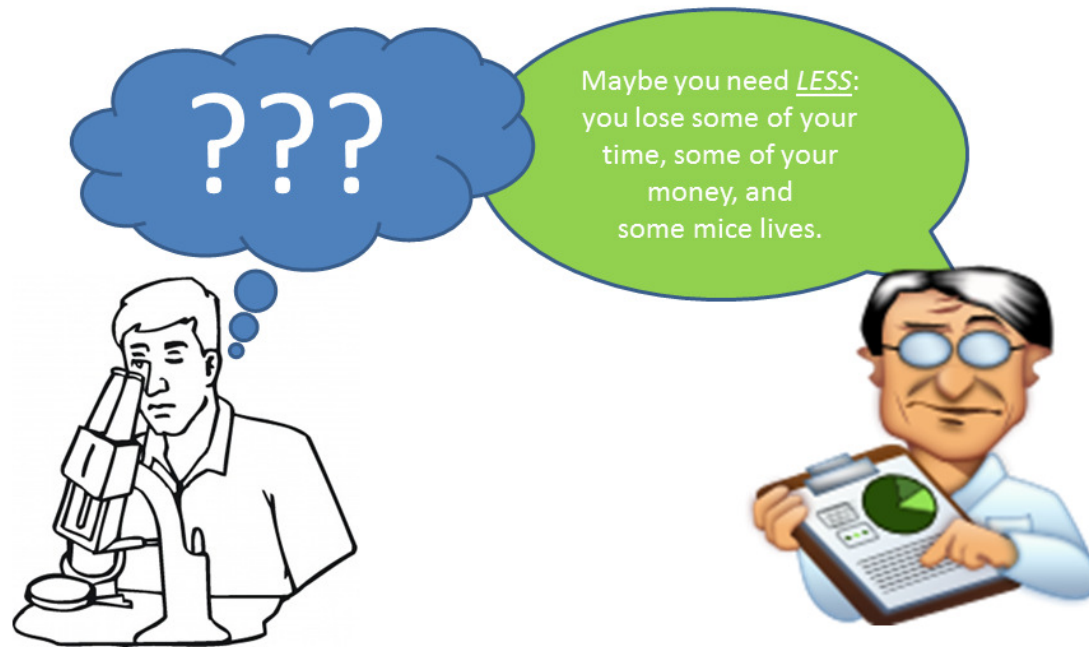
# Sample size calculation

## Stage 2: Value of Experimental Design



# Sample size calculation

## Stage 2: Value of Experimental Design



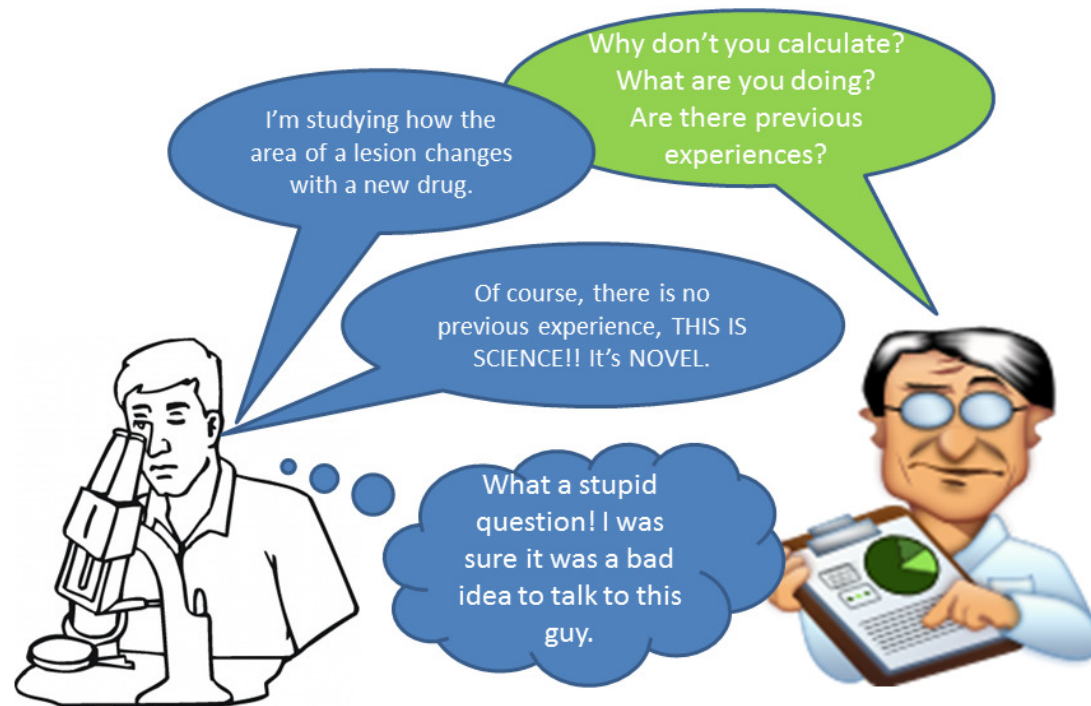
# Sample size calculation

## Stage 2: Value of Experimental Design



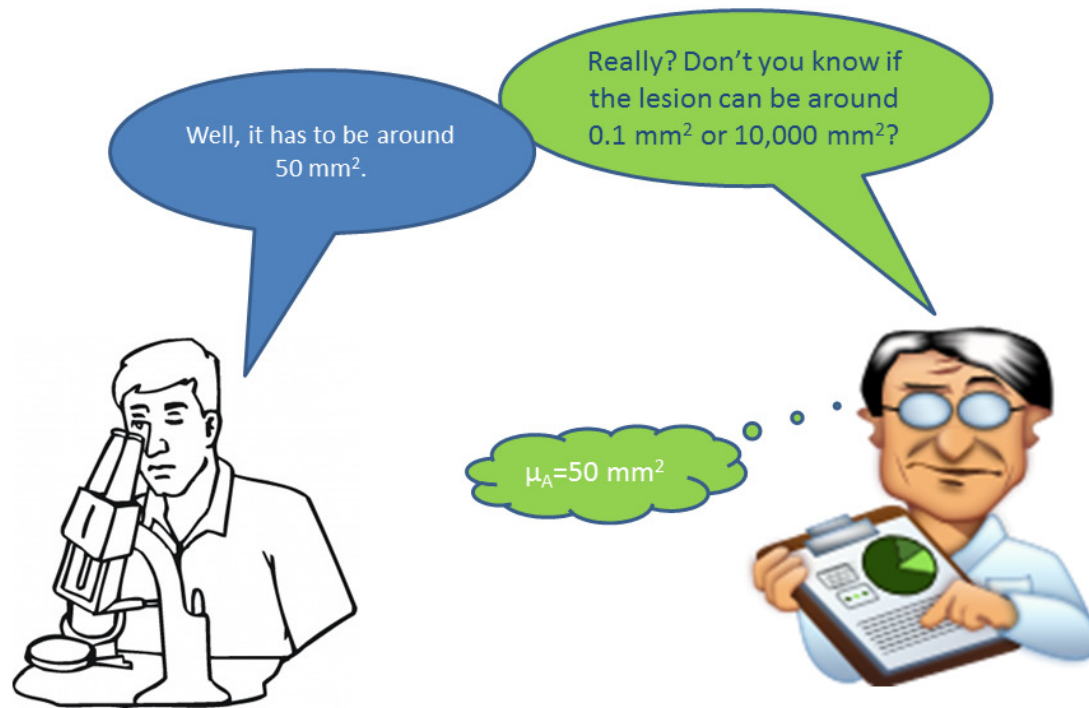
# Sample size calculation

## Stage 3: Experimental Design



# Sample size calculation

## Stage 4: Gathering prior information



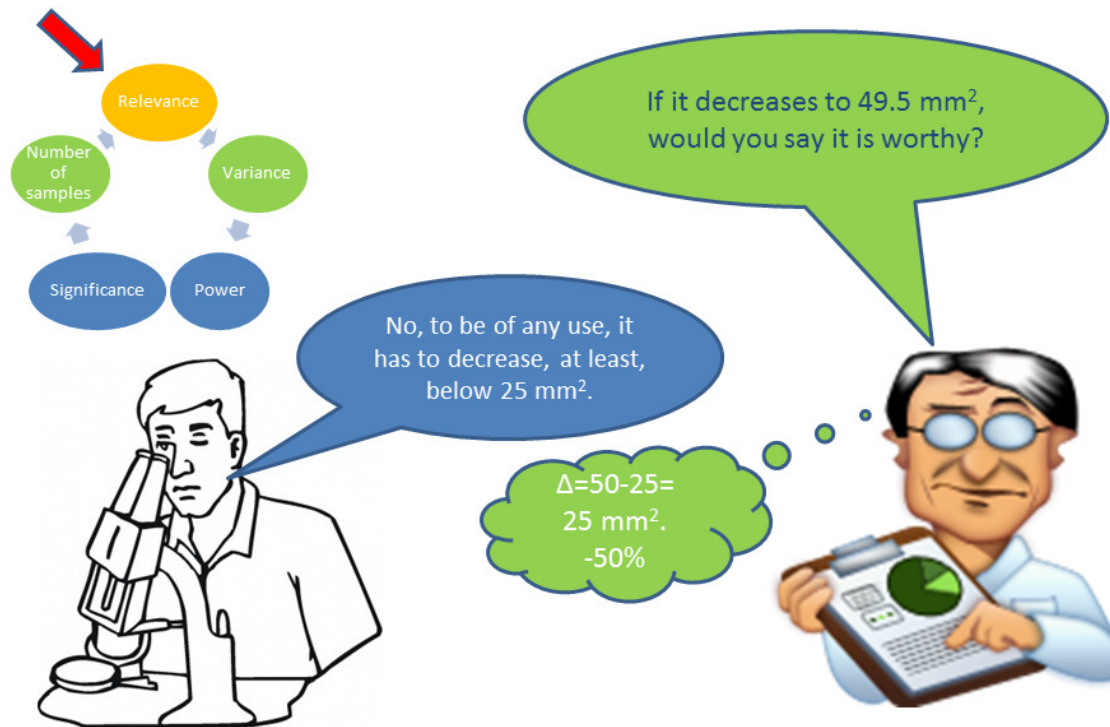
# Sample size calculation

## Stage 5: Setting a target for success



# Sample size calculation

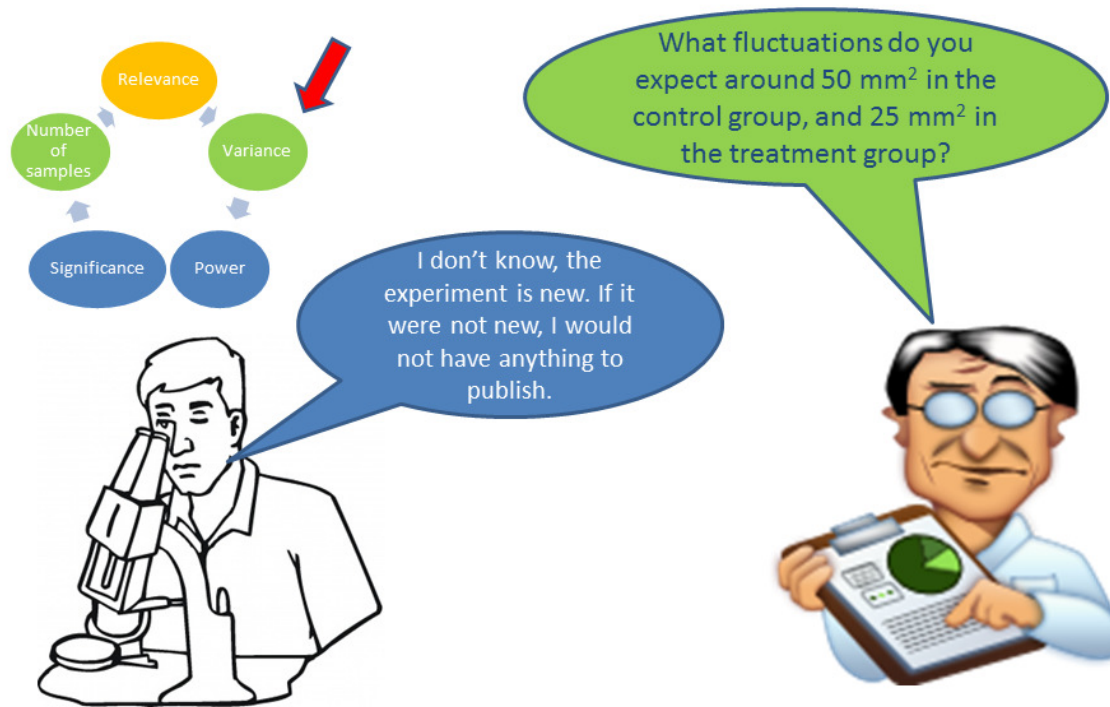
## Stage 5: Setting a target for success





# Sample size calculation

## Stage 6: Gathering variance information



# Sample size calculation

## Stage 6: Gathering variance information



But you must know something a priori. When you measure the control group, do you expect values like 50.00, 49.96, 50.14, 50.14, 49.76 ( $\sigma^2=0.1$ ) or values like 58.15, 8.27, 31.96, 21.86, 129.75 ( $\sigma^2=1000$ )? Both have a mean of 50, but you understand detecting a change of 25 in the second case is more difficult. The more difficult it is to detect the change, the more mice you'll need.



# Sample size calculation

## Stage 6: Gathering variance information



# Sample size calculation

## Stage 6: Gathering variance information



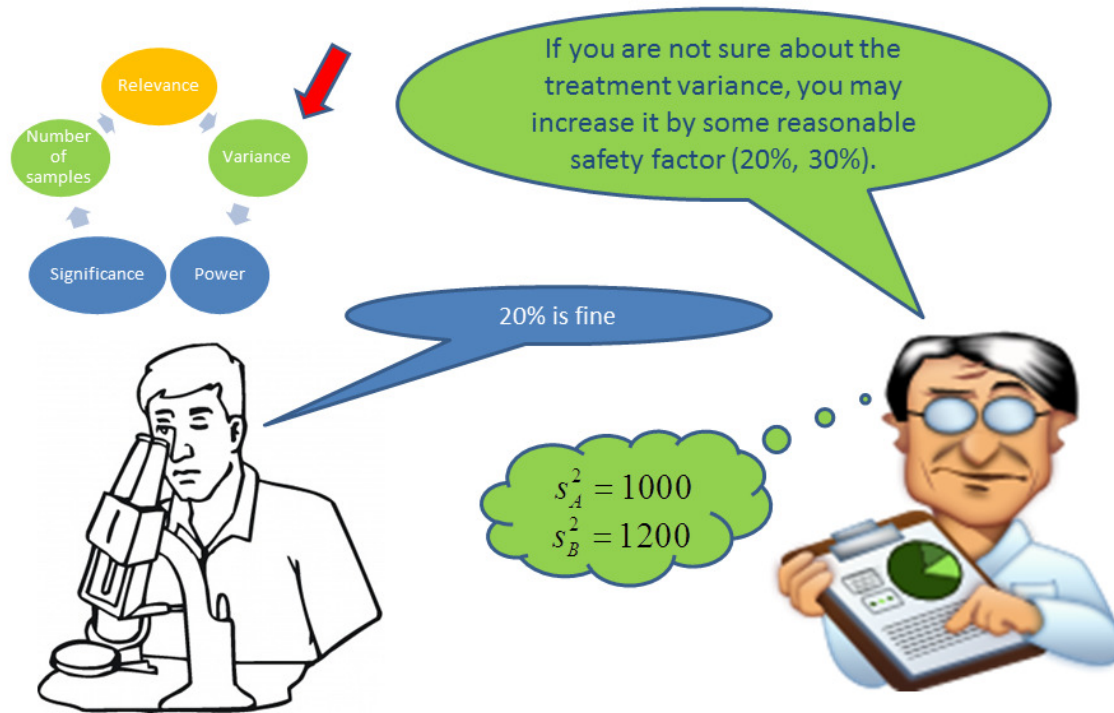
Well, ... there is a paper where they perform something similar to my experiment, but with a different drug. They report a variance of 1000. And for the treatment, it is less clear, but we may assume we will have similar fluctuations.

$$s_A^2 = s_B^2 = 1000$$



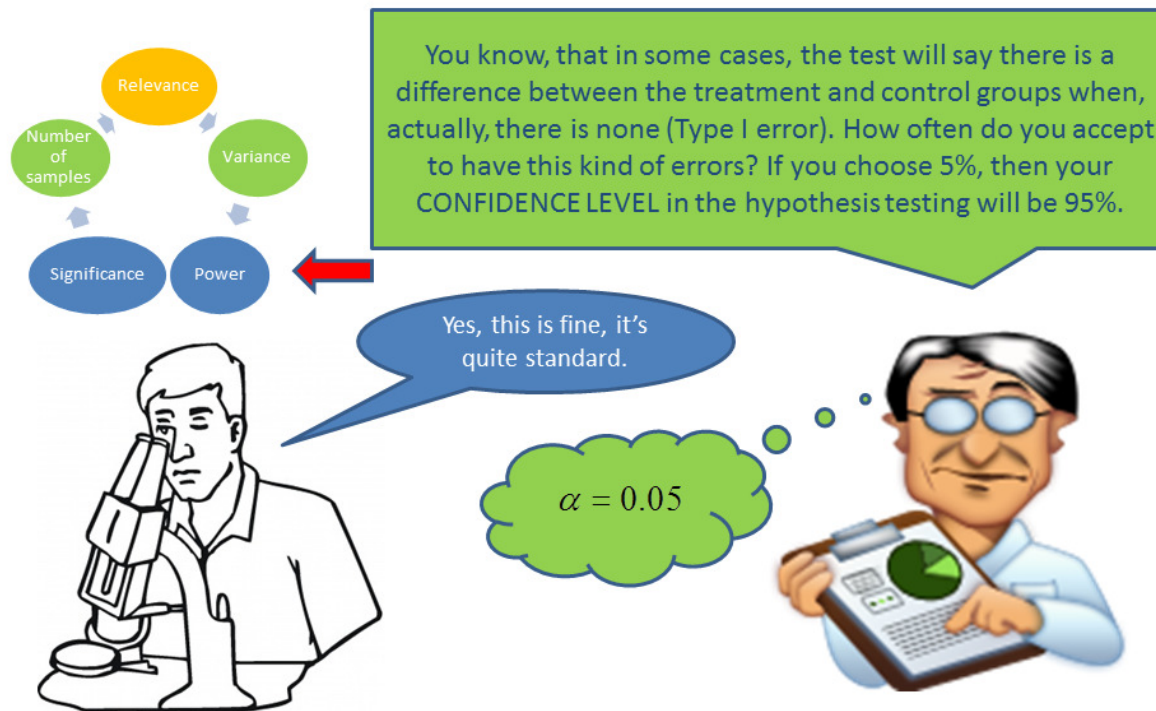
# Sample size calculation

## Stage 6: Gathering variance information



# Sample size calculation

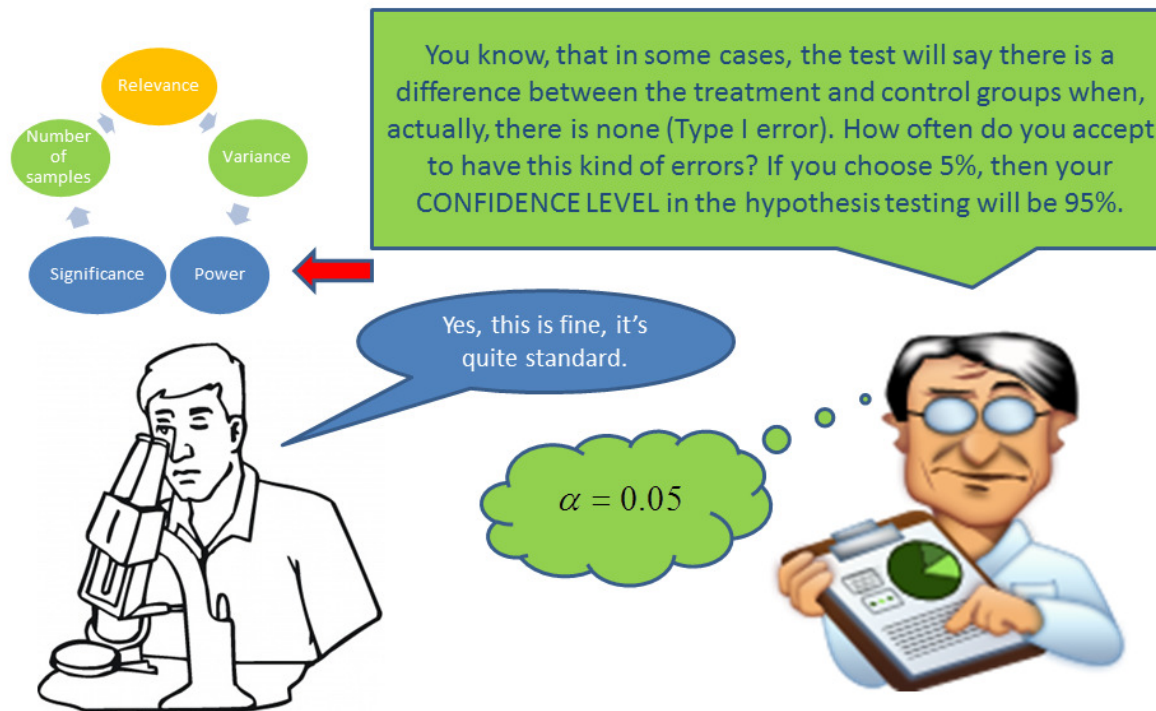
## Stage 7: Setting error tolerances





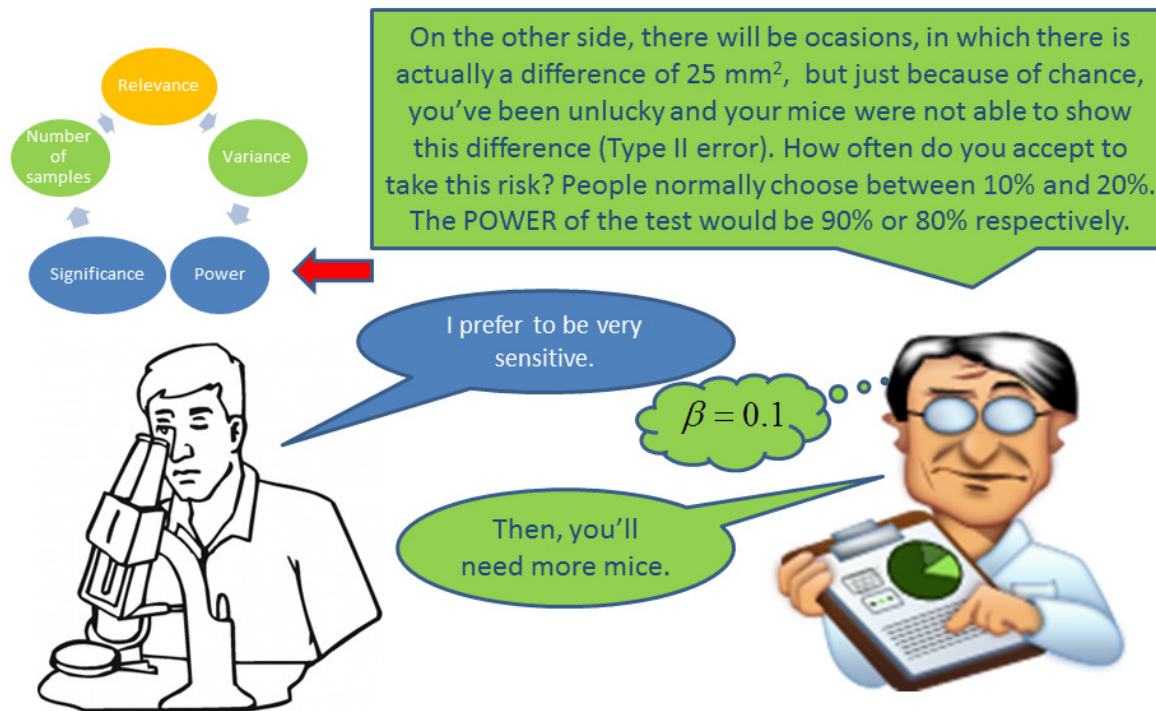
# Sample size calculation

## Stage 7: Setting error tolerances



# Sample size calculation

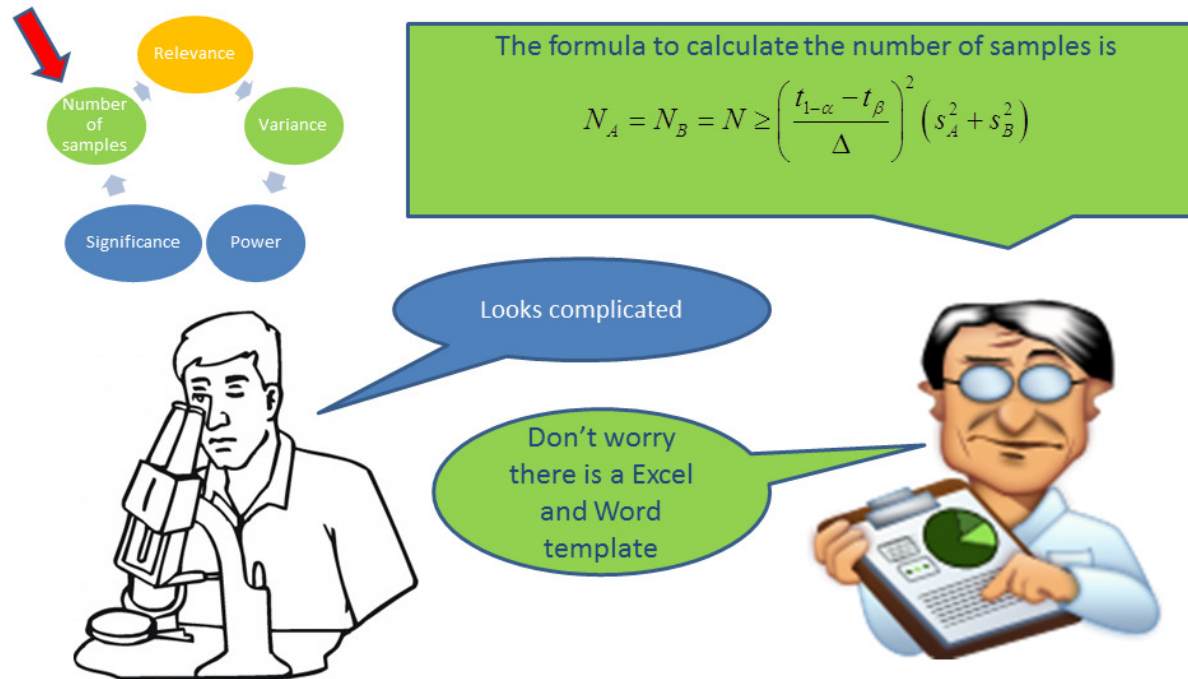
## Stage 7: Setting error tolerances





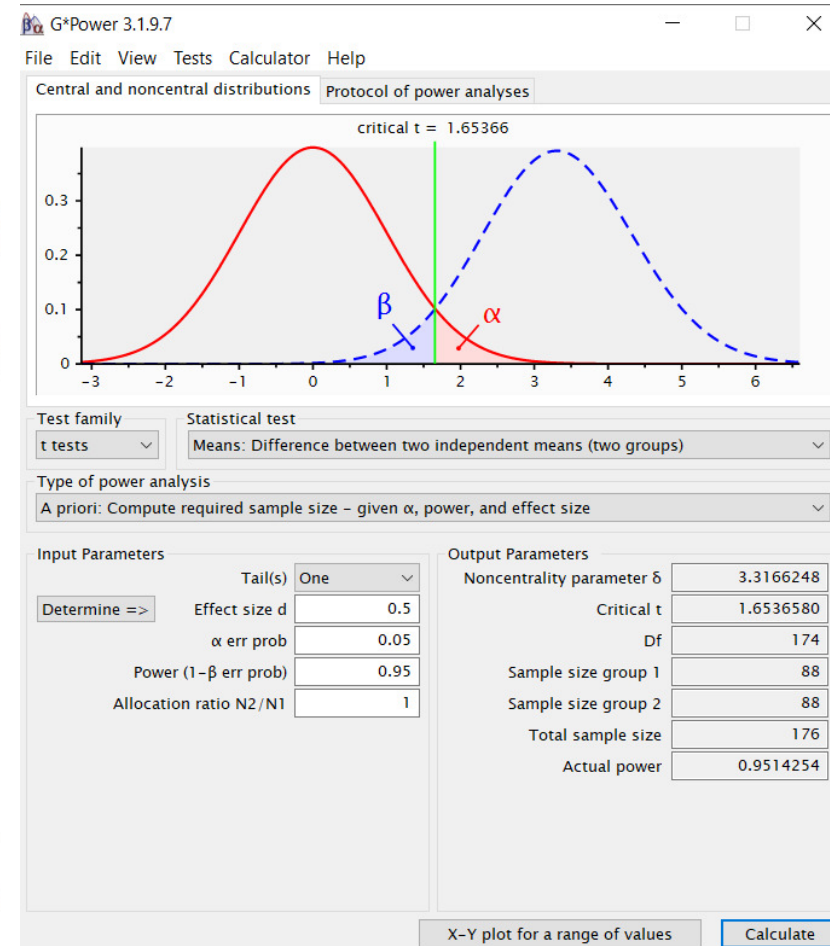
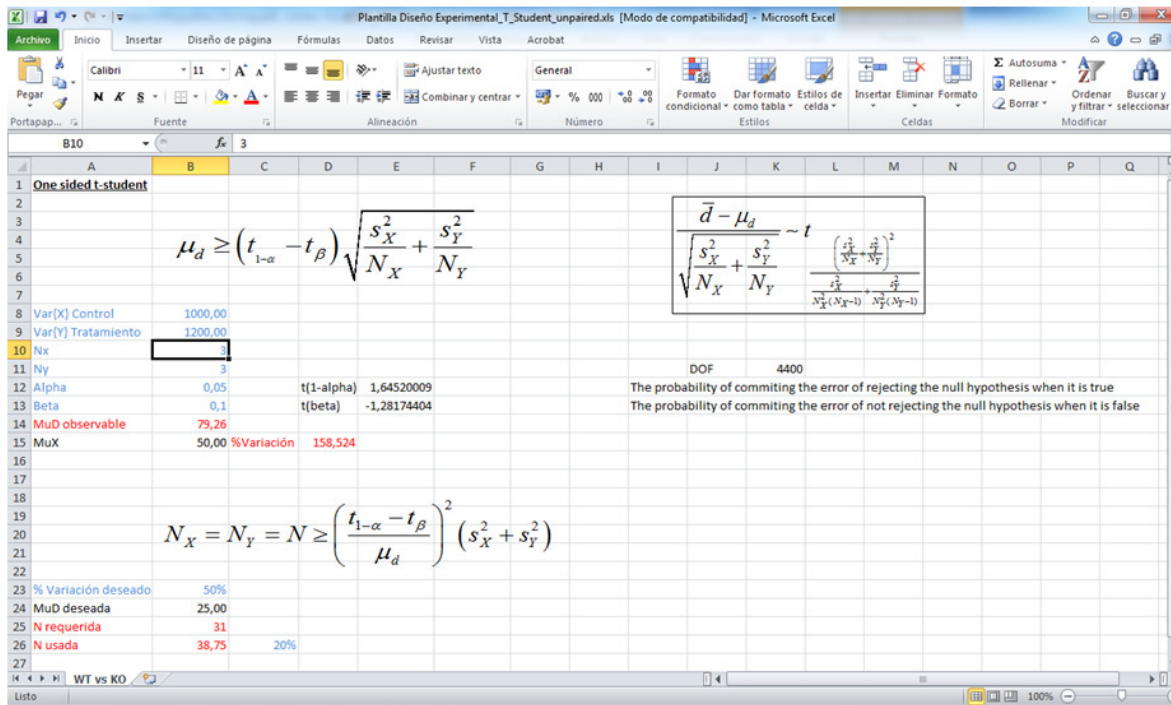
# Sample size calculation

## Stage 8: Calculating the number of mice



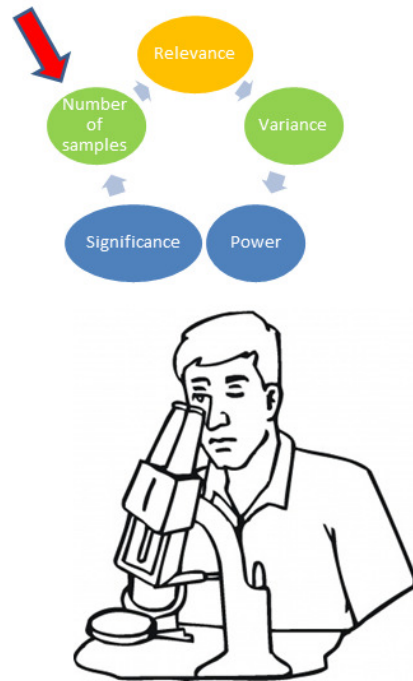
# Sample size calculation

## Stage 8: Calculating the number of mice



# Sample size calculation

## Stage 8: Calculating the number of mice

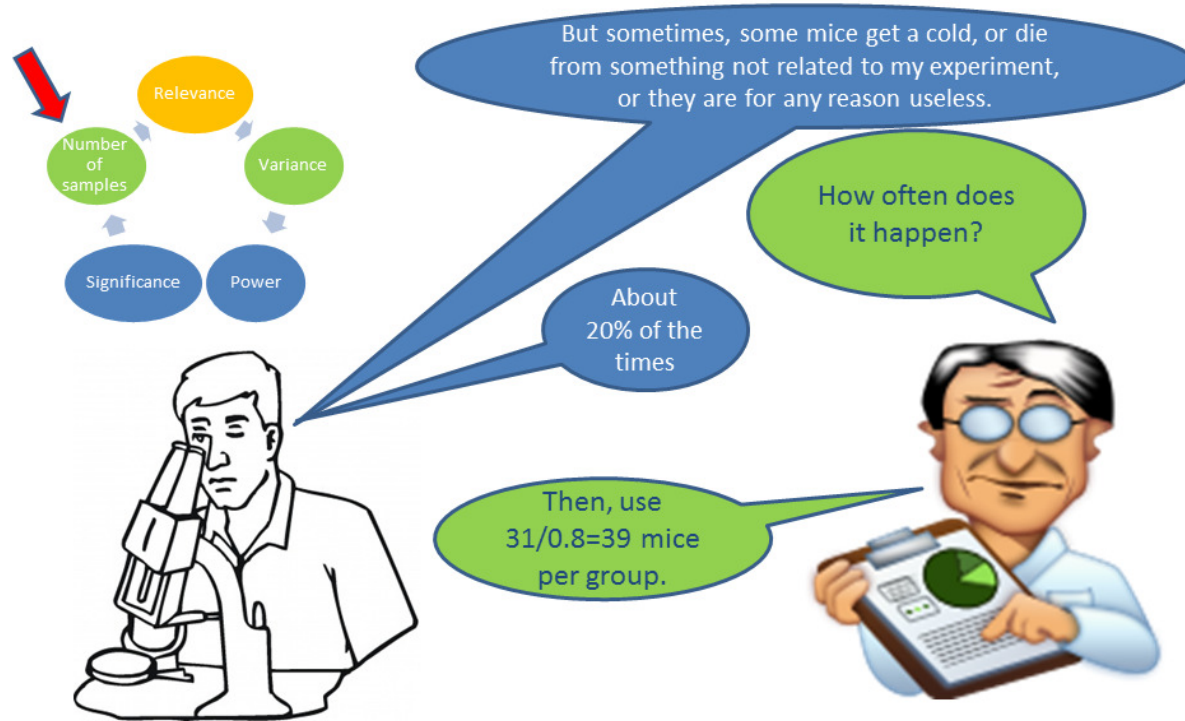


Good news, you  
only need 31  
mice in each  
group.



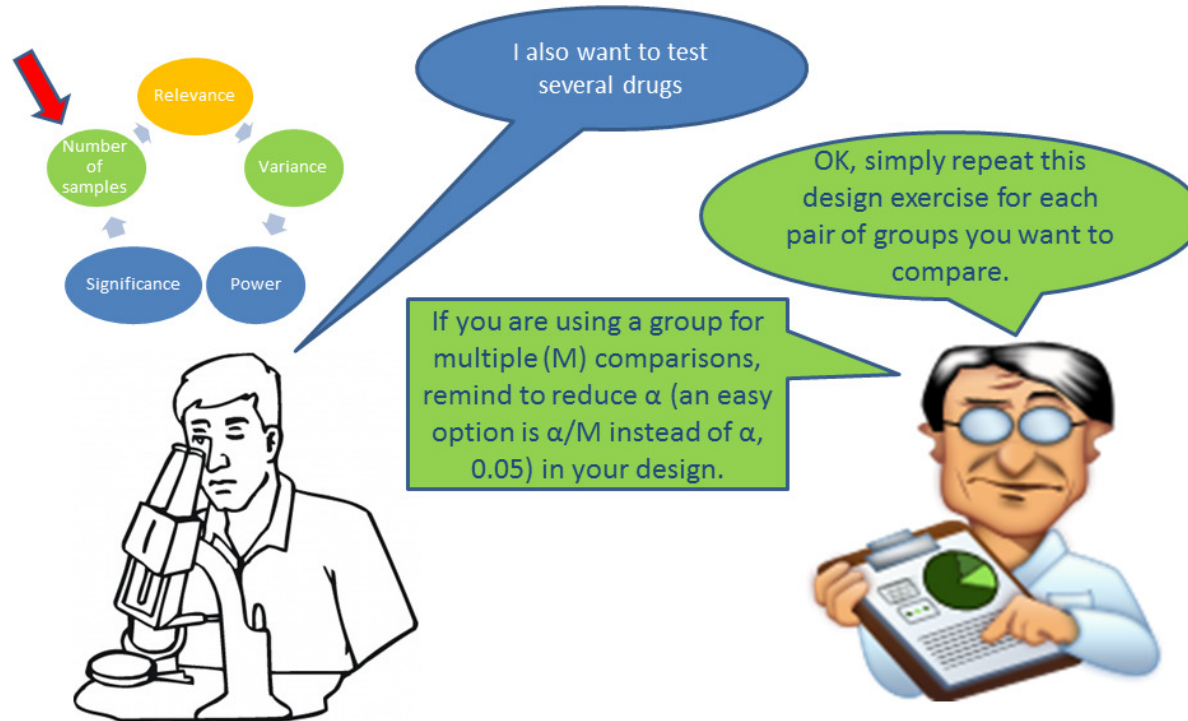
# Sample size calculation

## Stage 8: Calculating the number of mice



# Sample size calculation

## Stage 8: Calculating the number of mice



# Sample size calculation

## Stage 8: Calculating the number of mice





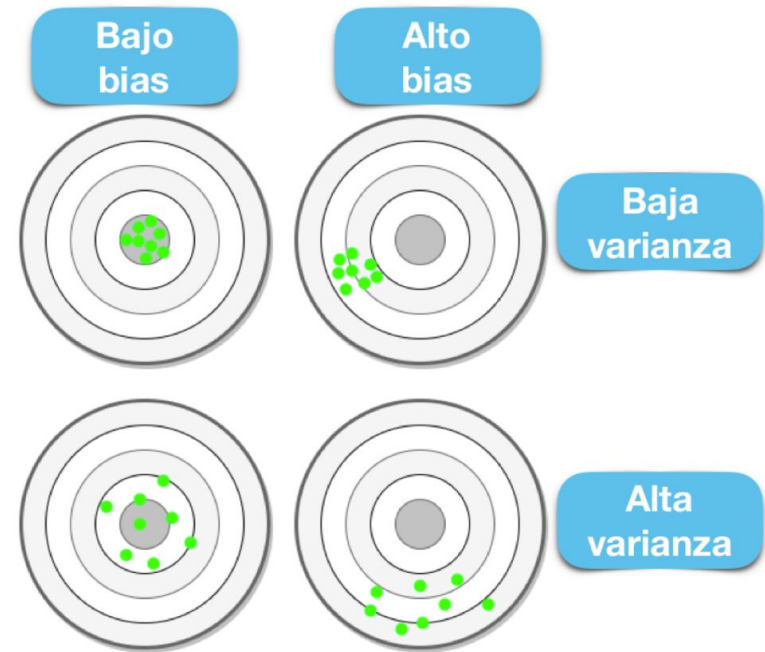
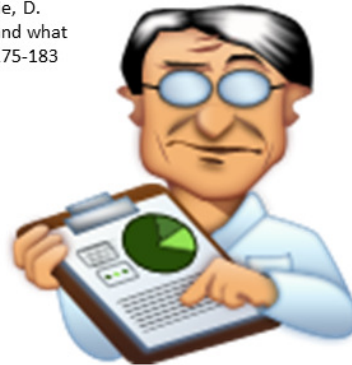
# Sample size calculation

## Replicates

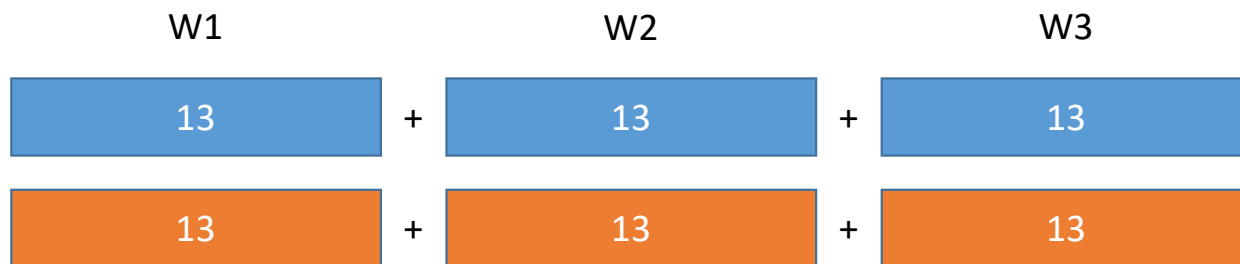
Should I replicate the experiment to be sure of the result? I would like to repeat it 3 times.

That depends on the p-value you get after analyzing the data. If you get  $p\text{-value}=\alpha$ , there is 50% chances that if you replicate the experiment you cannot reject the null hypothesis.  $P\text{-value}=\alpha$  is a suggestion of an interesting result, but not a definitive result. A  $p\text{-value}=\alpha/10$  has a probability of 80% of getting the same result if you replicate it.

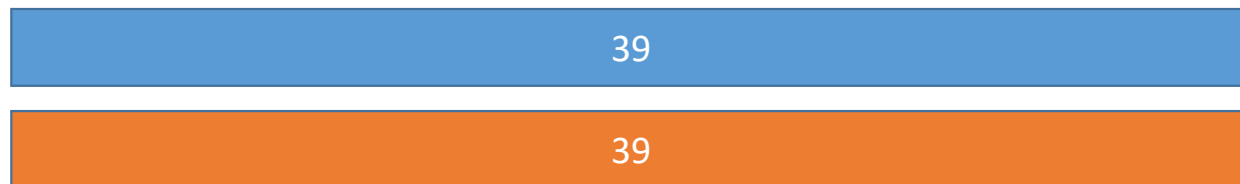
Greenwald, A. G.; Gonzalez, R.; Harris, R. J. & Guthrie, D.  
Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology*, **1996**, 33, 175-183



## Blocks: mini-experiments



	SS	df	MS=SS/df
Treatment	SS	1	MS
Block	SSB	2	MSB
Error	SSE	74	MSE
Total	SST	77	



	SS	df	MS=SS/df
Treatment	SS	1	MS
Error	SSE	76	MSE
Total	SST	77	



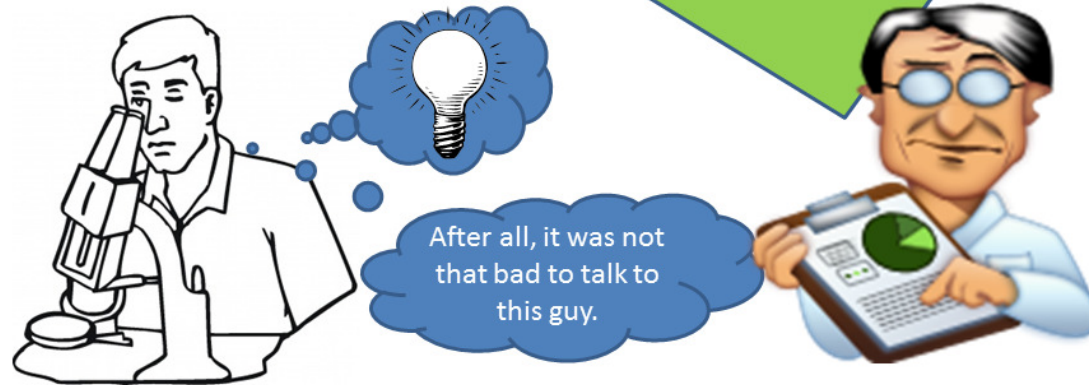
# Sample size calculation

## Summarizing

Designing your experiment made you think about:

1. What is a RELEVANT experiment result.
2. What is reasonable to EXPECT from the experiment.
3. Which is the BALANCE I have chosen in this experiment among sensitivity, errors and sample size.
4. HOW MANY mice you need because of statistical aspects and how many because of experimental aspects.

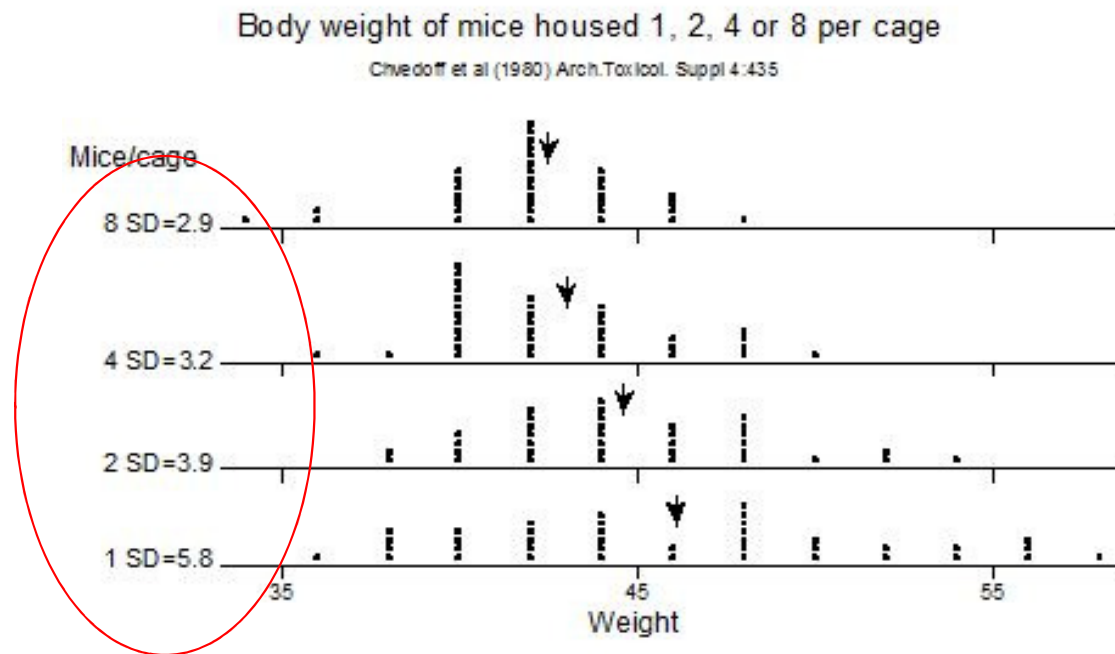
And you are guaranteed to SAVE your MONEY, TIME and mice LIVES.



# Variance reduction

## Change the experimental conditions

Chvedoff, M. et al. (1980). Effects on mice of numbers of animal per cage: an 18-month study. (preliminary results). Archives of Toxicology, Supplement 4:435-438



# Variance reduction

## Genetics of Mouse Behavior: Interactions with Laboratory Environment

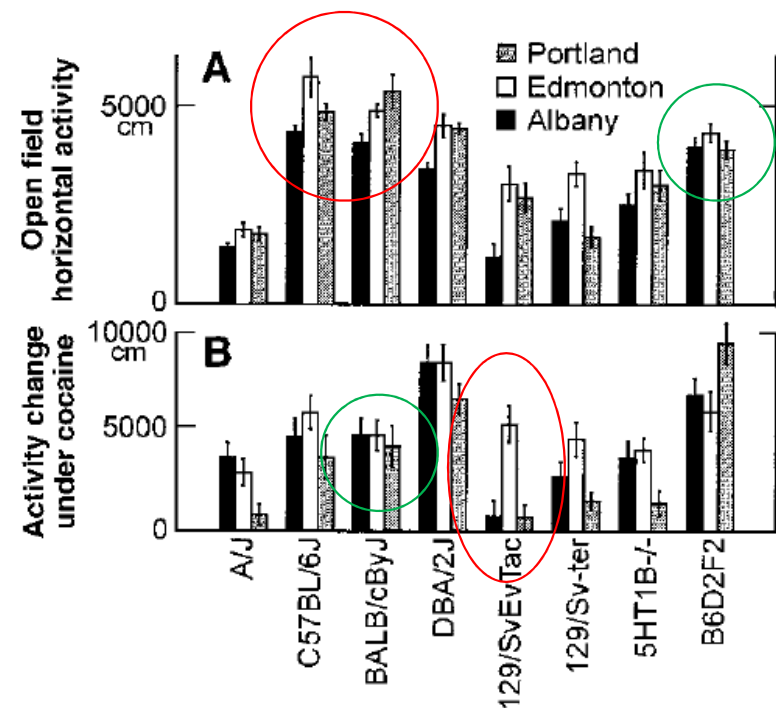
John C. Crabbe,<sup>1\*</sup> Douglas Wahlsten,<sup>2</sup> Bruce C. Dudek<sup>3</sup>

Strains of mice that show characteristic patterns of behavior are critical for research in neurobehavioral genetics. Possible confounding influences of the laboratory environment were studied in several inbred strains and one null mutant by simultaneous testing in three laboratories on a battery of six behaviors. Apparatus, test protocols, and many environmental variables were rigorously equated. Strains differed markedly in all behaviors, and despite standardization, there were systematic differences in behavior across labs. For some tests, the magnitude of genetic differences depended upon the specific testing lab. Thus, experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory.

1. Same research team
2. Same inbred strains
3. Equally calibrated apparatus
4. Equated husbandry
5. Same testing protocols
6. Same age
7. Same starting time
8. Same protocol order

But **significantly different results**

Crabbe, J. C.; Wahlsten, D. & Dudek, B. C. Genetics of mouse behavior: interactions with laboratory environment. *Science*, 1999, 284, 1670-1672.



## Basics of statistical inference



An engineer works for MyPharma. He knows that the manufacture of each tablet has a standard deviation of 1 mg. (the manufacturing process can be approximated by a Gaussian). Knowing this, he sets the machine to a target amount of 250 mg. In a routine check with 20 tablets, he measures an average of 250.66 mg. Is it possible that the machine is malfunctioning?

- Step 1: Define the hypotheses

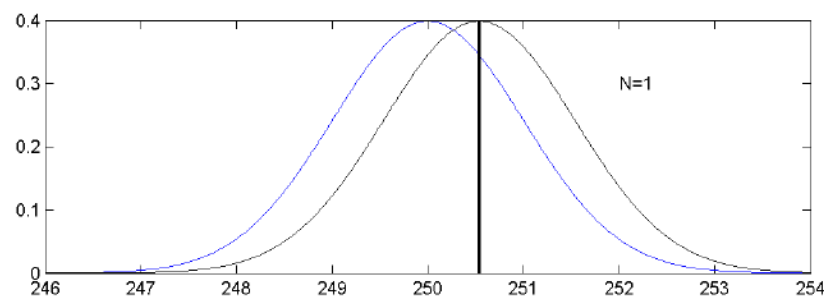
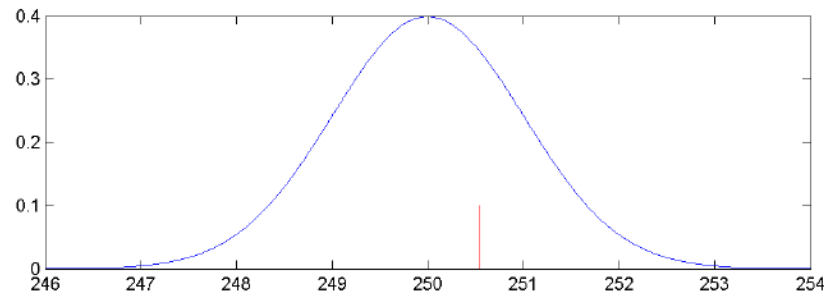
$$H_0 : \mu = 250$$

$$H_1 : \mu \neq 250$$

## Basics of statistical inference

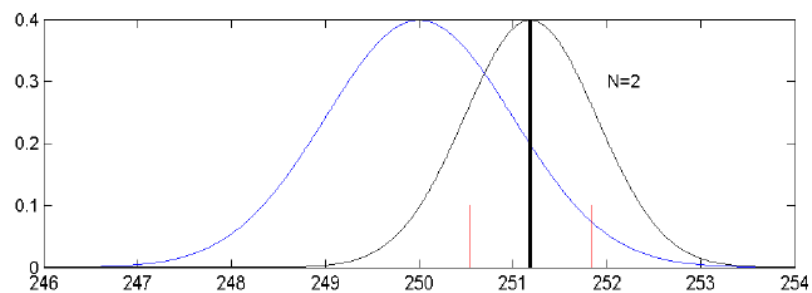
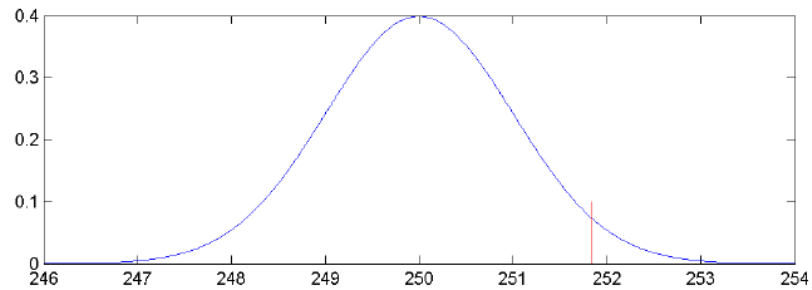
$$E\{x_1\} = \mu, \text{Var}\{x_1\} = \sigma^2$$

$$\hat{\mu} = x_1 \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \sigma^2$$



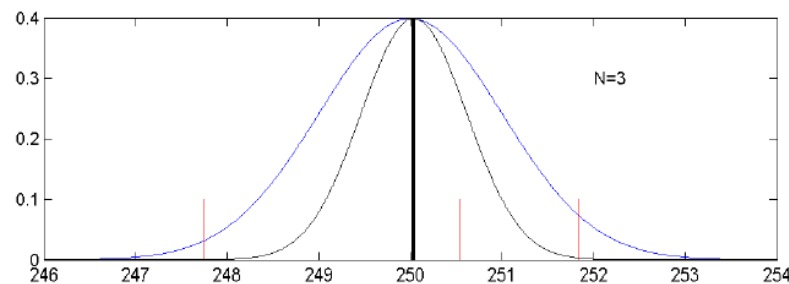
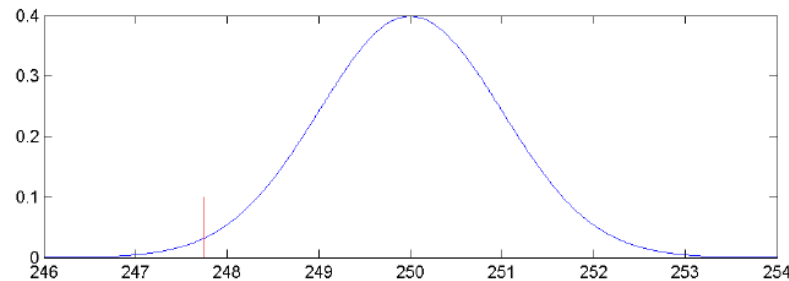
## Basics of statistical inference

$$E\{x_2\} = \mu, \text{Var}\{x_2\} = \sigma^2$$
$$\hat{\mu} = \frac{x_1 + x_2}{2} \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \frac{\sigma^2}{2}$$



## Basics of statistical inference

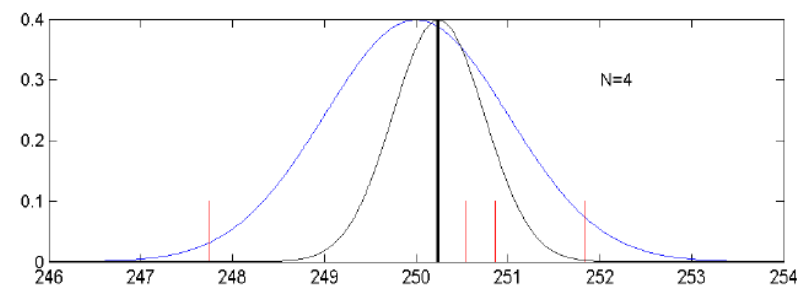
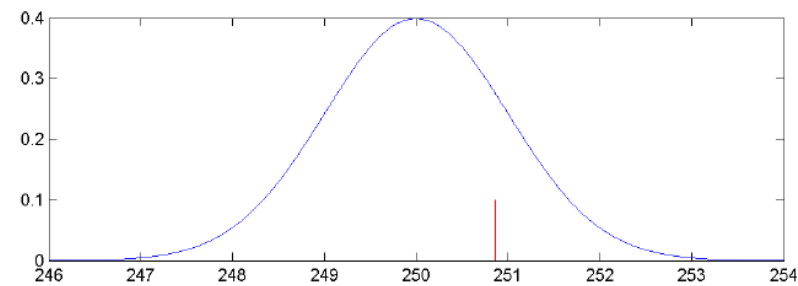
$$E\{x_3\} = \mu, \text{Var}\{x_3\} = \sigma^2$$
$$\hat{\mu} = \frac{x_1 + x_2 + x_3}{3} \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \frac{\sigma^2}{3}$$



## Basics of statistical inference

$$E\{x_4\} = \mu, \text{Var}\{x_4\} = \sigma^2$$

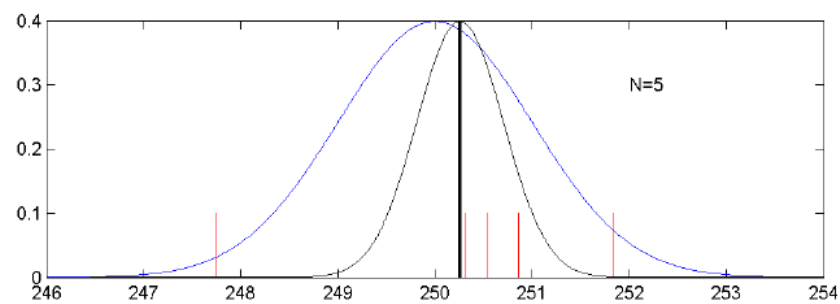
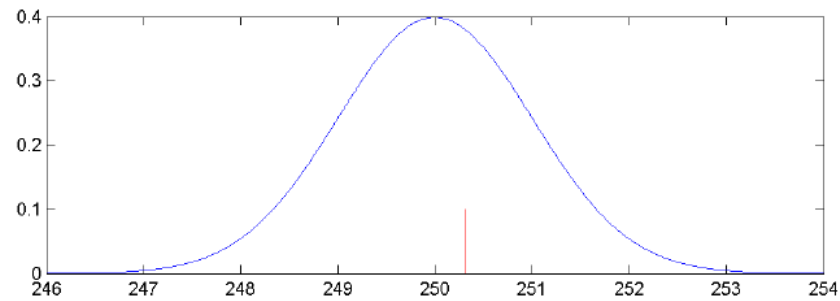
$$\hat{\mu} = \frac{x_1 + x_2 + x_3 + x_4}{4} \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \frac{\sigma^2}{4}$$





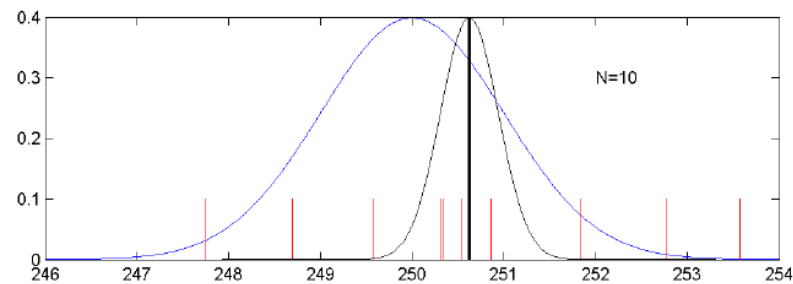
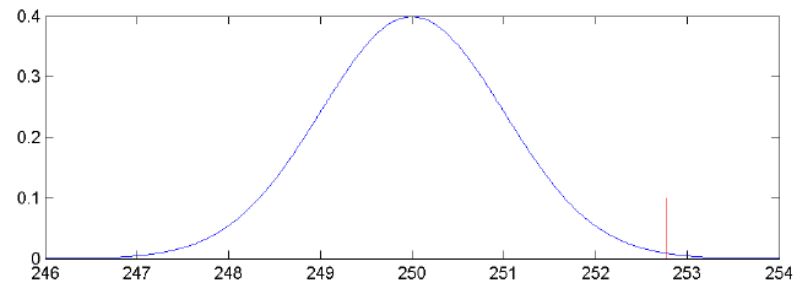
## Basics of statistical inference

$$\hat{\mu} = \frac{1}{5} \sum_{i=1}^5 x_i \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \frac{\sigma^2}{5}$$



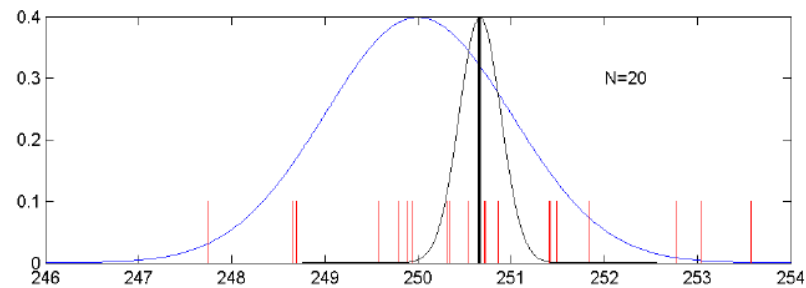
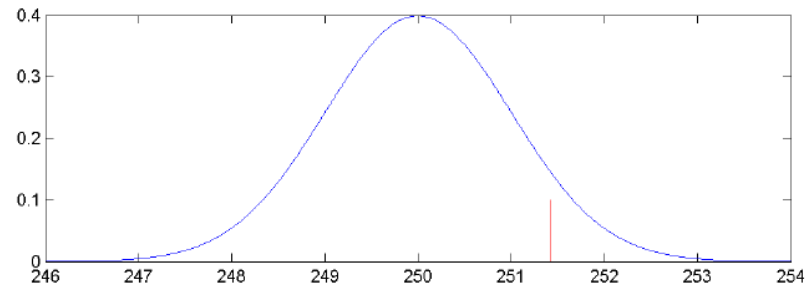
## Basics of statistical inference

$$\hat{\mu} = \frac{1}{10} \sum_{i=1}^{10} x_i \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \frac{\sigma^2}{10}$$



## Basics of statistical inference

$$\hat{\mu} = \frac{1}{20} \sum_{i=1}^{20} x_i \Rightarrow E\{\hat{\mu}\} = \mu, \text{Var}\{\hat{\mu}\} = \frac{\sigma^2}{20}$$



## Basics of statistical inference

- Step 2: Find the distribution of a suitable statistic if  $H_0$  is true

$$x_i \sim N(\mu, \sigma^2) \Rightarrow \hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right) \Rightarrow Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1)$$

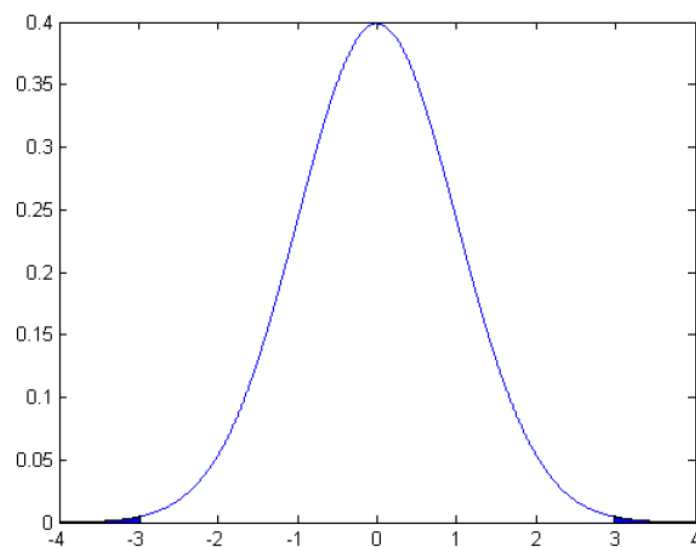
- Step 3: Plug-in the observed data if  $H_0$  is true

$$z = \frac{250.66 - 250}{\frac{1}{\sqrt{20}}} = 2.9721$$

## Basics of statistical inference

- Step 4: Calculate the p-value Probability of observing a value as extreme as this one if  $H_0$  is true.

$$\begin{aligned} \text{p-value} &= \Pr\{|Z| > 2.9721\} = \Pr\{Z < -2.9721\} + \Pr\{Z > 2.9721\} \\ &= 0.0030 = 0.3\% \end{aligned}$$



## Basics of statistical inference

- Step 5: Reject or not reject  $H_0$

$$p - \text{value} = 0.003(**) < 0.05 \Rightarrow \text{Reject } H_0$$

$p < 0.05$	*
$p < 0.01$	**
$p < 0.001$	***

## Sample size determination: Confidence level

- ① Step 1: Define the null hypothesis

$$H_0 : \mu = 250$$

- ② Step 2: Distribution under the null hypothesis

$$Z = \frac{\tilde{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{\Delta}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1)$$

- ③ Step 3: Plug-in observed data

$$z = 2.9721$$

- ④ Step 4: Calculate p-value

$$\Pr\{|Z| > 2.9721\} = 0.3\%$$

- ⑤ Step 5: Decide on  $H_0$

$$0.3\% < 5\% \Rightarrow \text{Reject}$$

- ① Step 1: Define the minimum meaningful difference

$$\Delta = 0.5(\text{mg})$$

- ② Step 2: Determine population variance

$$\sigma^2 = 1^2(\text{mg}^2)$$

- ③ Step 3: Determine significance and statistic threshold

$$\alpha = 0.05 \Rightarrow \Pr\{|Z| > 1.96\} = 0.05$$

- ④ Step 4: Solve for  $N$

$$\frac{\Delta}{\frac{\sigma}{\sqrt{N}}} > 1.96 \Rightarrow N > \left(\frac{1.96\sigma}{\Delta}\right)^2 = 15.4$$

## Sample size determination: Confidence level

Factors that affect sample size:

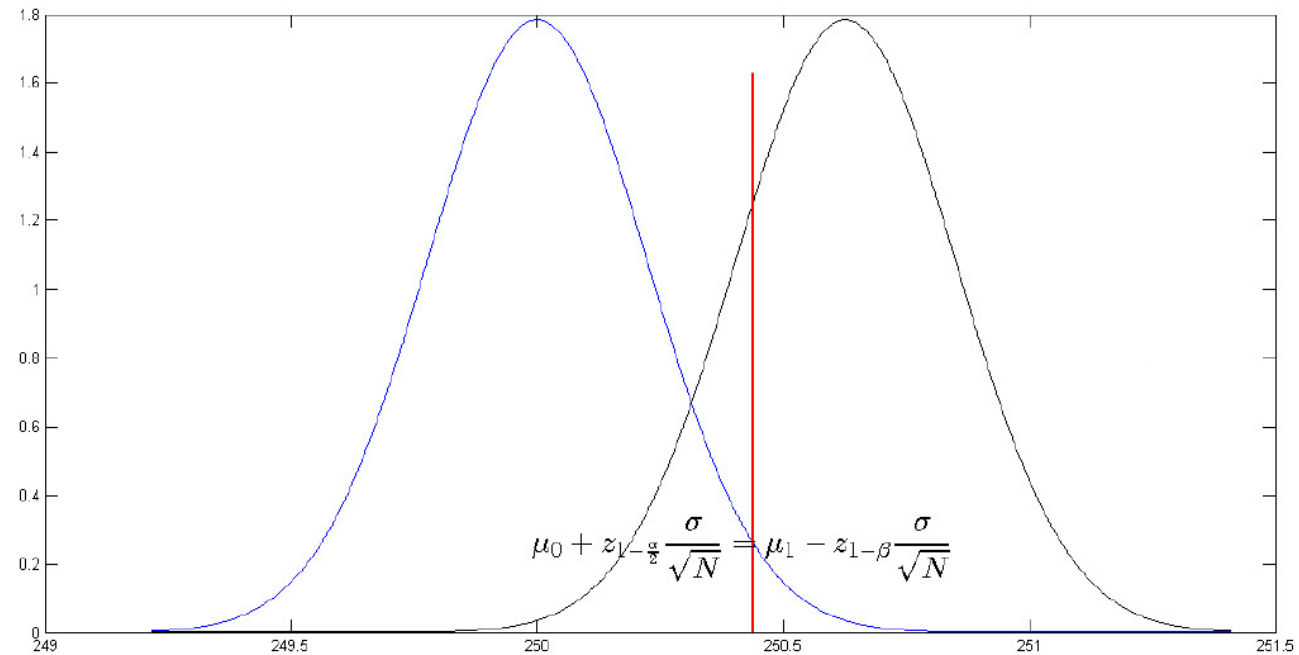
$$N > \left( \frac{z_{1-\frac{\alpha}{2}} \sigma}{\Delta} \right)^2 \quad (1)$$

- ① Confidence level:  $1 - \alpha \uparrow \Rightarrow z_{1-\frac{\alpha}{2}} \uparrow \Rightarrow N \uparrow$   
More confidence requires more samples.
- ② Sample variance:  $\sigma^2 \uparrow \Rightarrow N \uparrow$   
If the sample variance increases, it is more difficult to detect the difference  $\Delta$ .
- ③ Effect size:  $\Delta \downarrow \Rightarrow N \uparrow$   
If we want to detect more subtle differences, we need more samples.
- ④ One- or Two-sided test: Two-sided  $\Rightarrow N \uparrow$   
If the test is one-sided,  $z_{1-\frac{\alpha}{2}}$  should be replaced by  $z_{1-\alpha}$ , which is smaller.



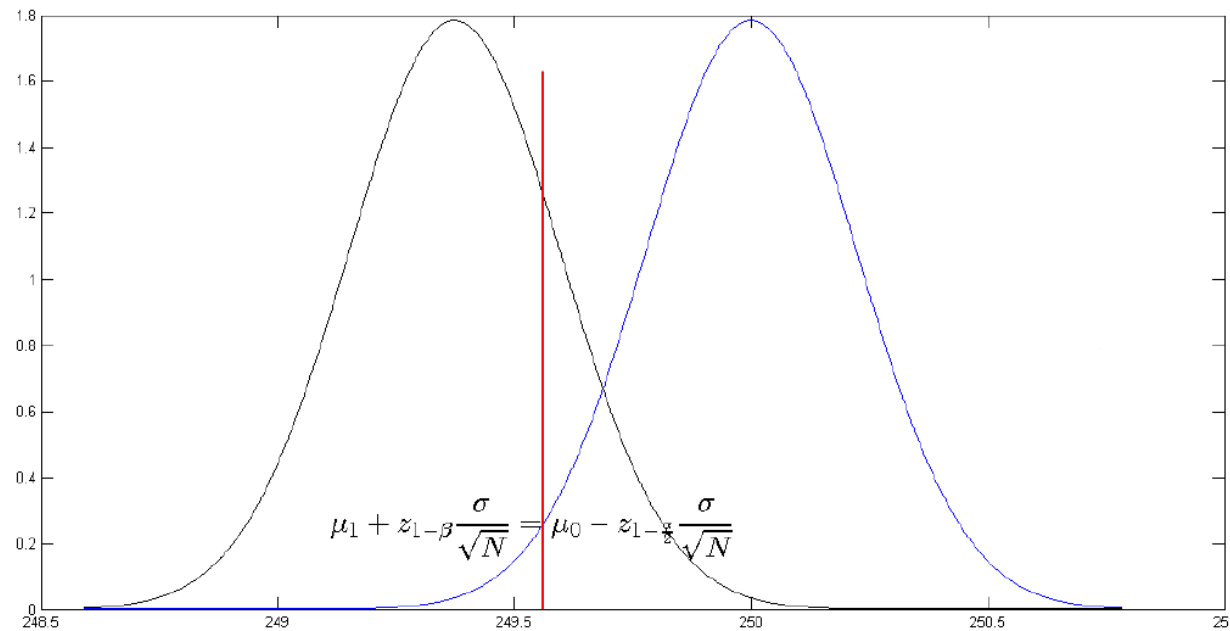
## Sample size determination: Test power (right)

$$\mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} < \mu_1 - z_{1-\beta} \frac{\sigma}{\sqrt{N}} \Rightarrow N > \left( \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sigma}{\mu_1 - \mu_0} \right)^2 = \left( \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sigma}{\Delta} \right)^2$$



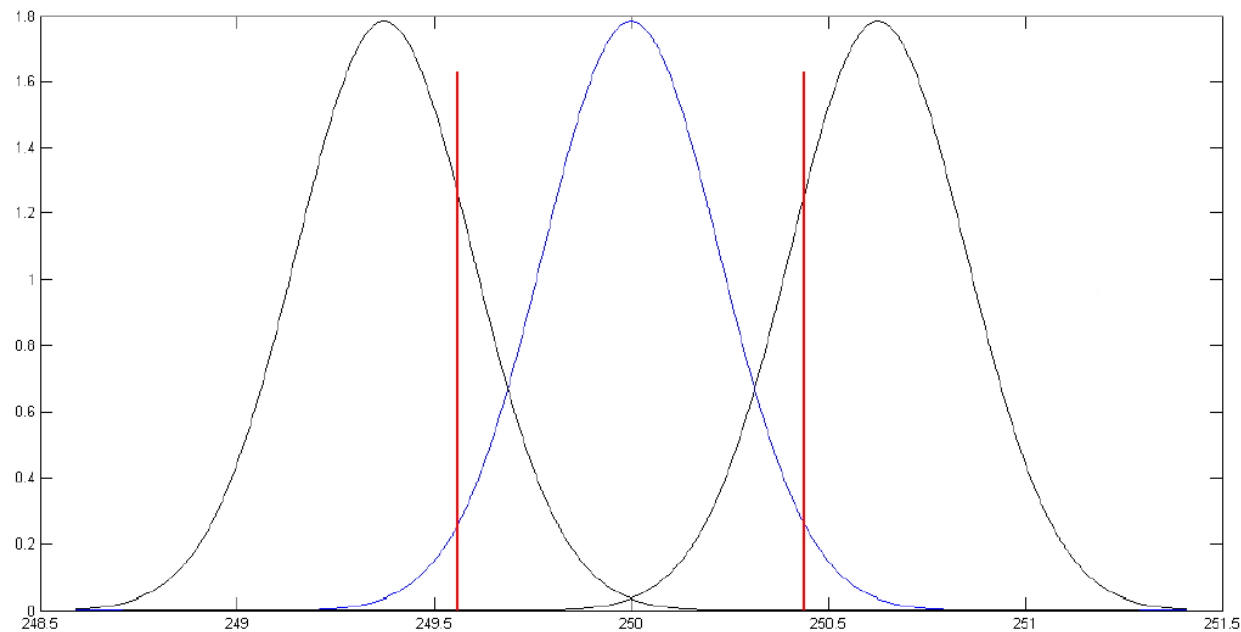
## Sample size determination: Test power (left)

$$\mu_1 + z_{1-\beta} \frac{\sigma}{\sqrt{N}} < \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \Rightarrow N > \left( \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sigma}{\mu_1 - \mu_0} \right)^2 = \left( \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sigma}{\Delta} \right)^2$$



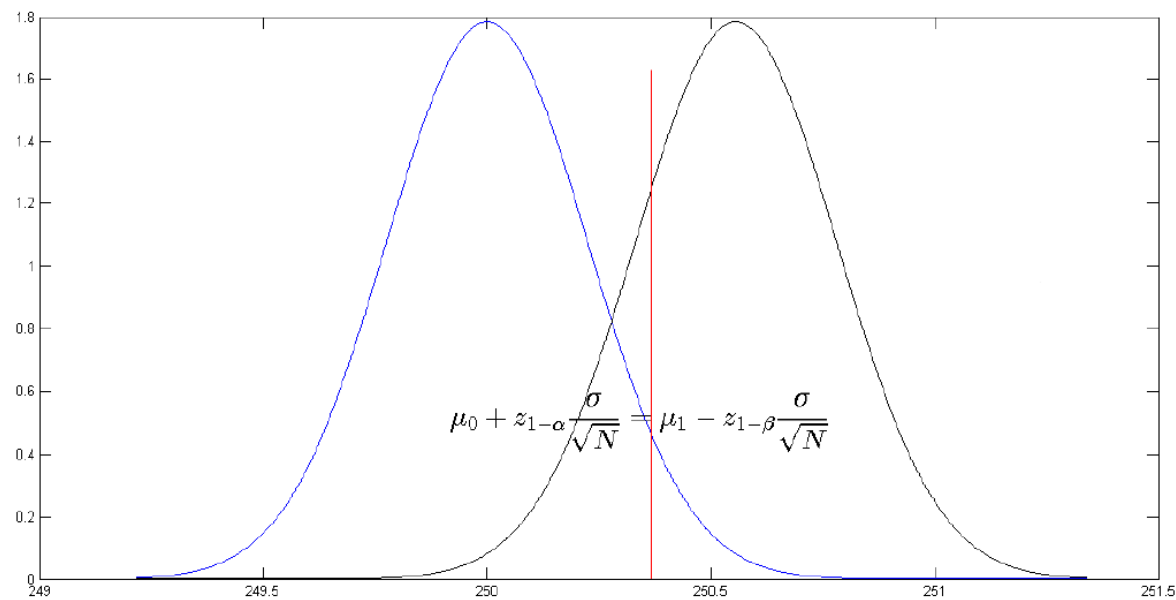
## Sample size determination: Test power (two-sided)

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \cup \mu > \mu_0 \end{array} \right\} \Rightarrow N > \left( \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sigma}{\Delta} \right)^2 \quad (2)$$



## Sample size determination: Test power (one-sided)

$$\left. \begin{array}{l} H_0 : \mu < \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \right\} \Rightarrow N > \left( \frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\Delta} \right)^2 \quad (3)$$



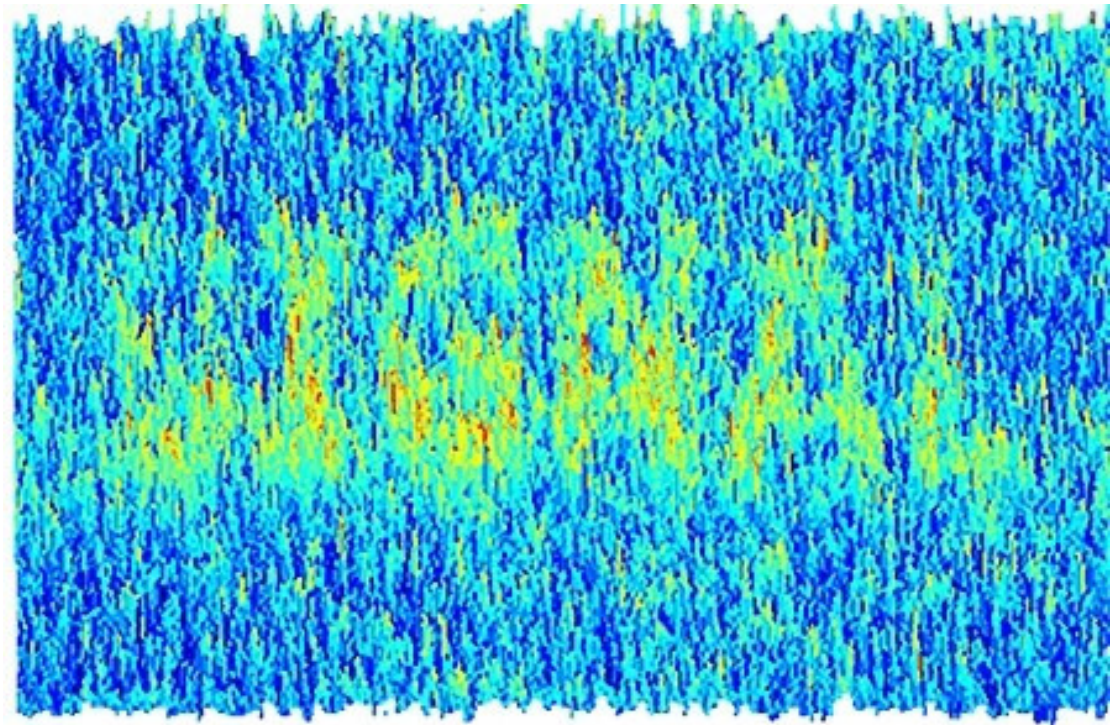
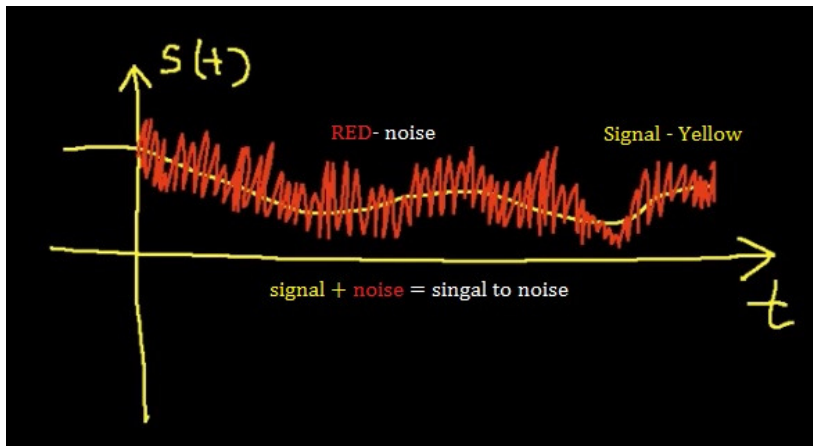
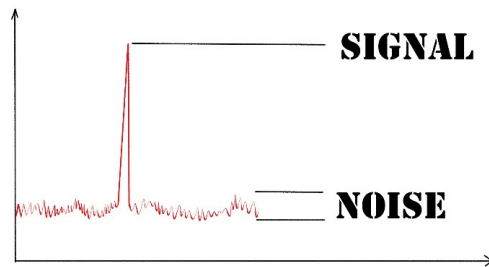
## Sample size determination: Confidence level+Test power

Factors that affect sample size:

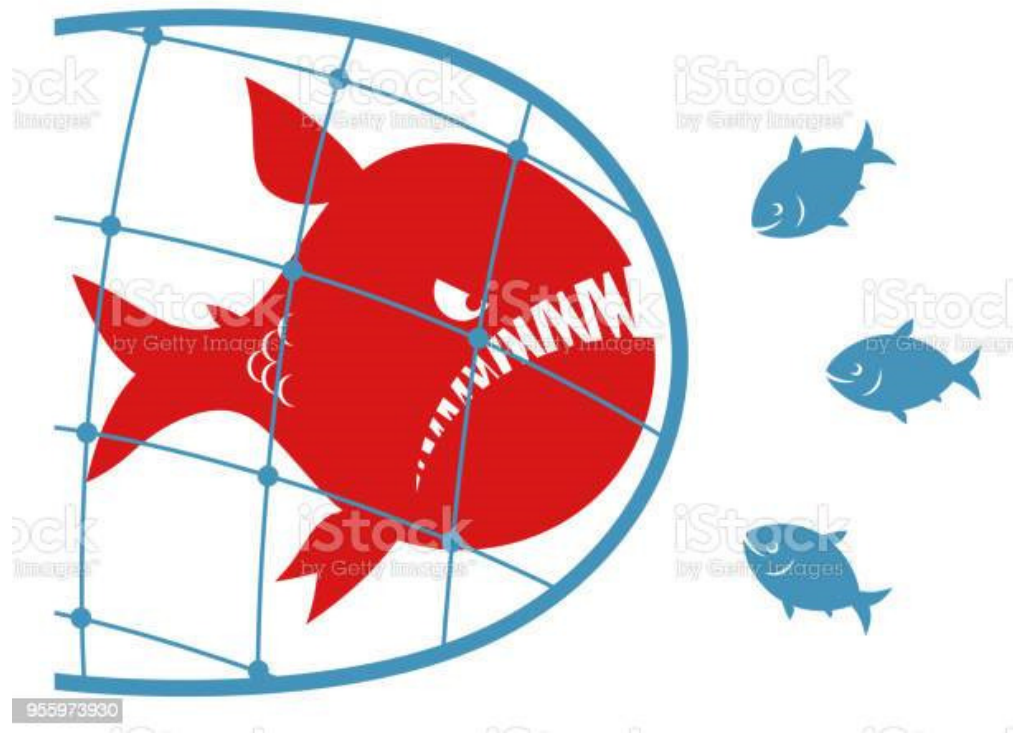
$$N > \left( \frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta/\sigma} \right)^2$$

- ① Confidence level:  $1 - \alpha \uparrow \Rightarrow z_{1-\frac{\alpha}{2}} \uparrow \Rightarrow N \uparrow$   
More confidence requires more samples.
- ② Population variance:  $\sigma^2 \uparrow \Rightarrow N \uparrow$   
If the population variance increases, it is more difficult to detect the difference  $\Delta$ .
- ③ Effect size:  $\Delta \downarrow \Rightarrow N \uparrow$   
If we want to detect more subtle differences, we need more samples.
- ④ One- or Two-sided test: Two-sided  $\Rightarrow N \uparrow$   
If the test is one-sided,  $z_{1-\frac{\alpha}{2}}$  should be replaced by  $z_{1-\alpha}$ , which is smaller.
- ⑤ Test power:  $1 - \beta \uparrow \Rightarrow N \uparrow$   
If we want to increase the power of the test, we need more samples.

# Signal-to-noise Ratio

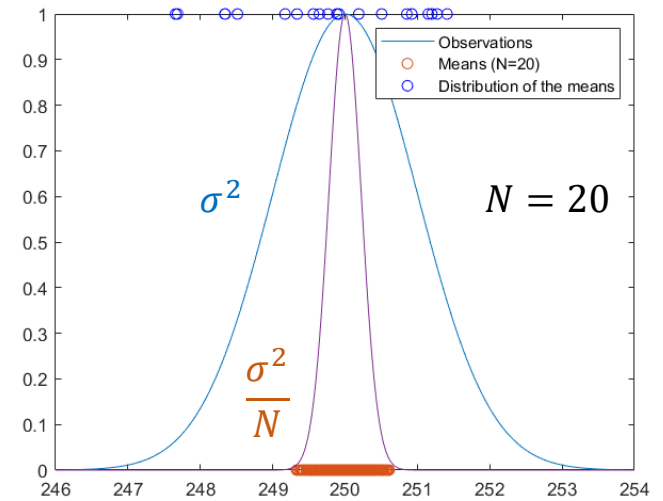
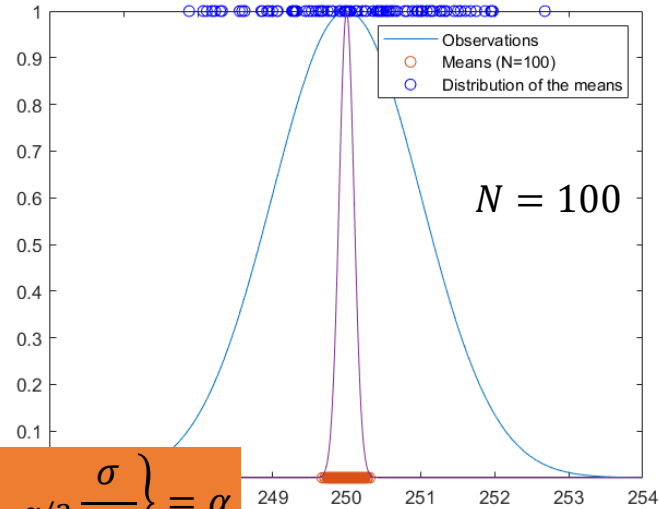


# Sample size calculation

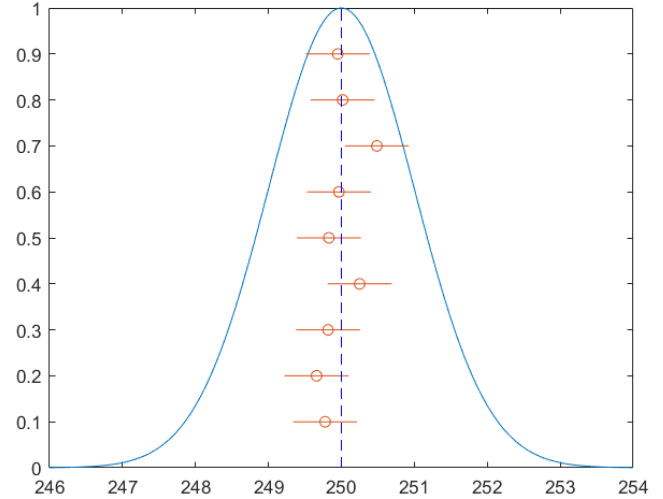
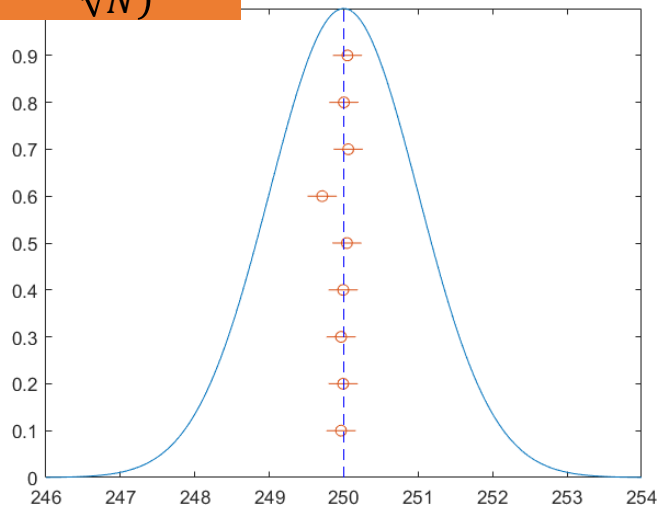


A researcher wanted to explore the submarine world. He used a net with squares of size 5x5 cm. After fishing thousands of wonderful creatures he came to the conclusion that in the Deep sea there are no creatures smaller than 5 cm.

## Confidence interval (Gaussian)

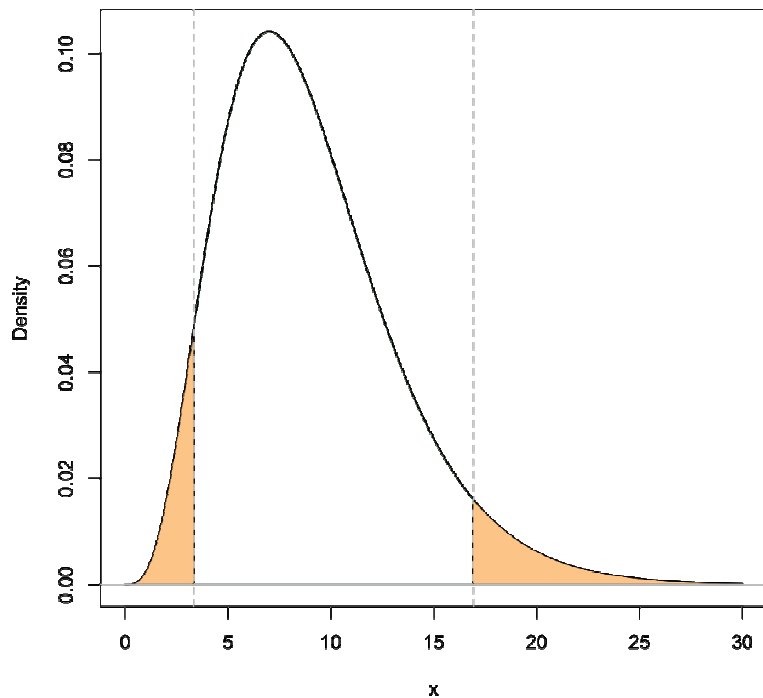


$$Pr \left\{ \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{N}} \right\} = \alpha$$

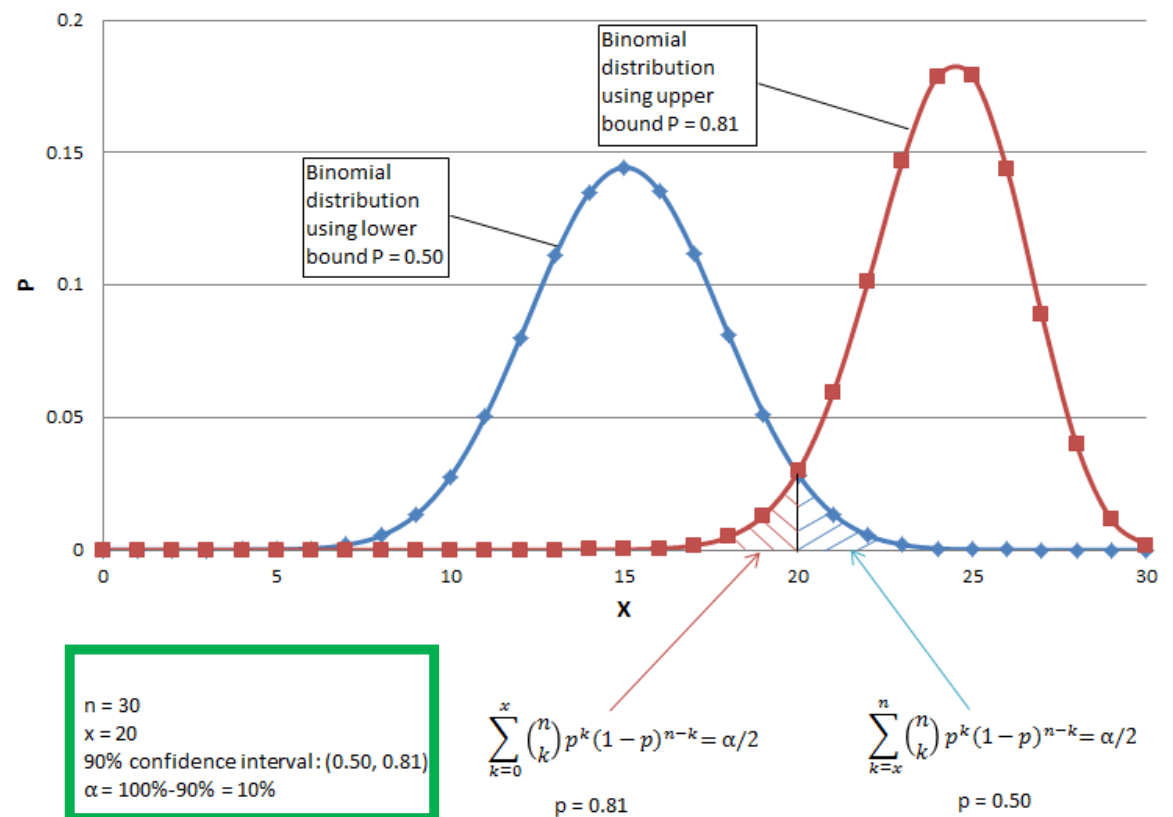




## Confidence interval (Snedecor's F)



## Exact confidence interval (Binomial)



## Confidence interval of a mean



Assume that we measure the temperature to 9 people and get the data:

37.0, 36.0, 37.1, 37.1, 36.2, 37.3, 37.0, 37.0, 36.1

The mean is  $36.76^{\circ}\text{C}$  and its 95% confidence interval  $[36.37, 37.14]^{\circ}\text{C}$ . This means that with probability 95%, this interval contains the true mean. Note that this interval is *symmetric* around 36.76.

The confidence interval is calculated as

$$\left[ \hat{\mu} - \frac{t_{1-\frac{\alpha}{2}, N-1} \hat{\sigma}}{\sqrt{N}}, \hat{\mu} + \frac{t_{1-\frac{\alpha}{2}, N-1} \hat{\sigma}}{\sqrt{N}} \right]$$

where  $t_{1-\frac{\alpha}{2}, N-1}$  is the  $1 - \frac{\alpha}{2}$  percentile of a Student's t distribution with  $N - 1$  degrees of freedom.

## Confidence interval of a standard deviation

Assume that we measure the temperature to 9 people and get the data:

37.0, 36.0, 37.1, 37.1, 36.2, 37.3, 37.0, 37.0, 36.1



The sample standard deviation is  $0.50^{\circ}\text{C}$  and its 95% confidence interval  $[0.34, 0.96]^{\circ}\text{C}$ . This means that with probability 95%, this interval contains the true standard deviation. Note that this interval is **not symmetric** around 0.50.

The confidence interval is calculated as

$$\left[ \hat{\sigma} \sqrt{\frac{N-1}{\chi_{1-\frac{\alpha}{2}, N-1}^2}}, \hat{\sigma} \sqrt{\frac{N-1}{\chi_{\frac{\alpha}{2}, N-1}^2}} \right]$$

where  $\chi_{1-\frac{\alpha}{2}, N-1}^2$  is the  $1 - \frac{\alpha}{2}$  percentile of a central  $\chi^2$  distribution with  $N - 1$  degrees of freedom.

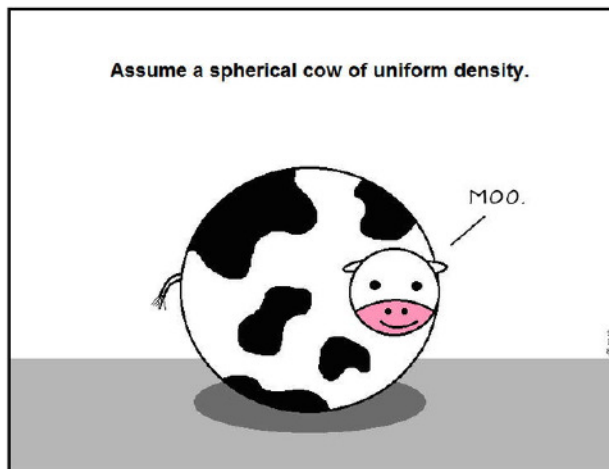
## Assumptions of confidence intervals

- **Random (representative) sample.** In clinical studies, patients are not randomly sampled from the patient population. They are included in the study because they were at the clinic at the right moment (**convenience sampling**). This assumption would also be violated if the body temperature is from people who joined the study because they suspected their body temperature was normally too high or too low (**voluntaries** in clinical studies are not random samples!)
- **Independent samples.** All subjects are sampled from the same population and independently selected from others. This assumption is violated if two siblings are included in the study, or if the same person is measured twice.
- **Accurate data.** Violated if the thermometer was not correctly placed or it was misread.
- **Population distribution.** Confidence intervals can only be constructed if the underlying, population distribution is known. The formulas in the previous slides are valid only for Gaussian populations.

# Properties of confidence intervals

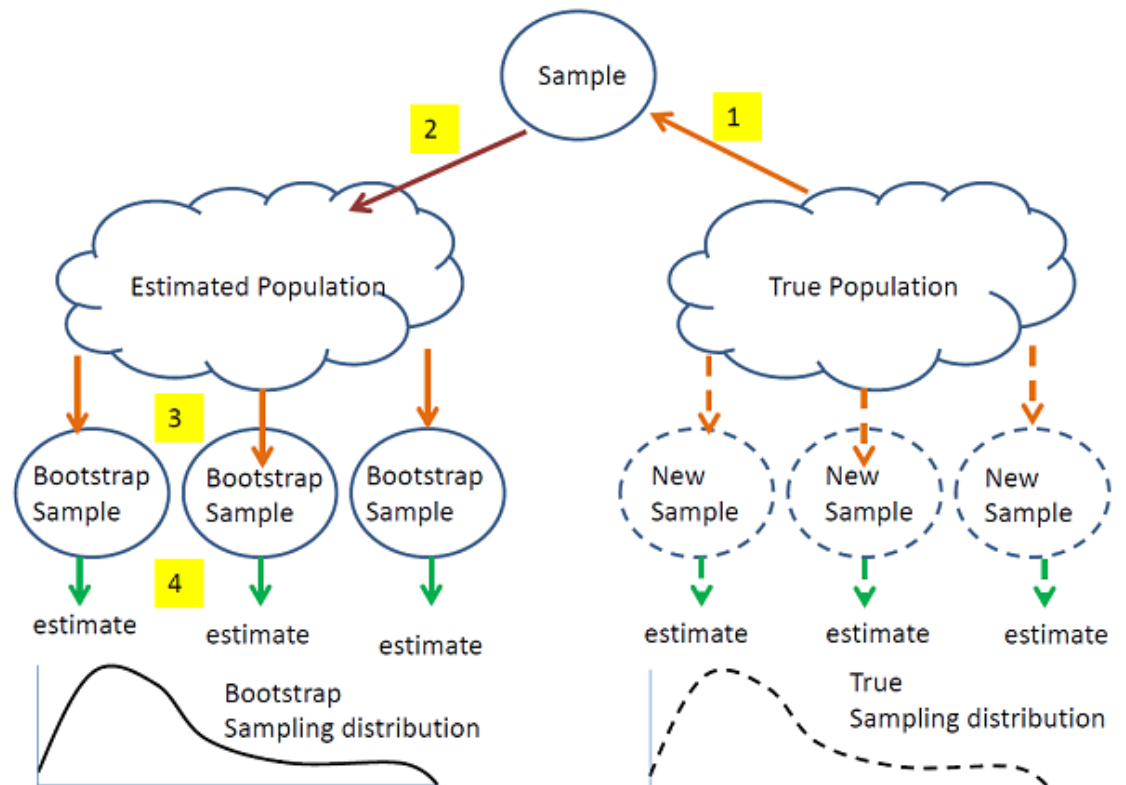
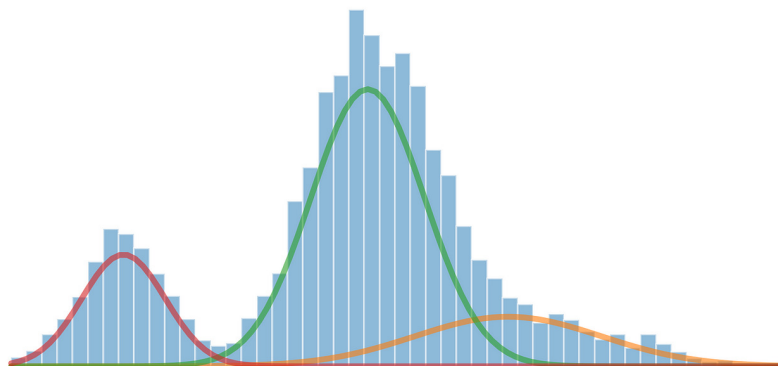
- **More samples.** The larger the experiment,  $N$ , the **narrower** the CI (we have less uncertainty about the underlying parameter).
- **More confidence.** The larger the confidence,  $1 - \alpha$ , the **wider** the CI (we need to enlarge it to be surer that it contains the true parameter).

What if the assumptions are **violated**?

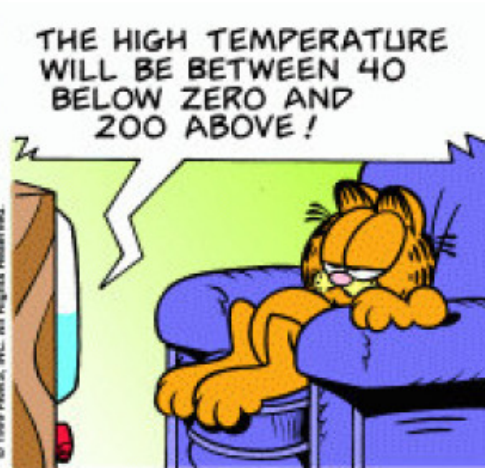


In many situations, these assumptions are not strictly true. Then, the CI may still be a **reasonable approximation of the range** of the underlying parameter (depending on the severity of the violation). But the confidence will, **for sure, not be** the one we think (95%).

## Confidence interval (Arbitrary)







## Non-parametric tests

In many occasions we do not know the distribution of the underlying data and non-parametric tests are used

Some Commonly Used Statistical Tests		
Normal theory based test	Corresponding nonparametric test	Purpose of test
$t$ test for independent samples	Mann-Whitney U test; Wilcoxon rank-sum test	Compares two independent samples
Paired $t$ test	Wilcoxon matched pairs signed-rank test	Examines a set of differences
Pearson correlation coefficient	Spearman rank correlation coefficient	Assesses the linear association between two variables.
One way analysis of variance ( $F$ test)	Kruskal-Wallis analysis of variance by ranks	Compares three or more groups
Two way analysis of variance	Friedman Two way analysis of variance	Compares groups classified by two different factors



## Non-parametric tests

The number of samples needed for a non-parametric test is larger than for a parametric one (because it throws away information, e.g., the sign test only uses the sign). The sample size must be increased by a factor that is inversely proportional to the “Asymptotic Relative Efficiency”:

$$N_{non-parametric} = \frac{N_{parametric}}{ARE} \quad (7)$$

Mann-Whitney U test	$3/\pi = 0.955$
Wilcoxon signed-rank test	$3/\pi = 0.955$
Spearman correlation test	0.91
Kruskal-Wallis test	0.864
Friedman ANOVA	$0.955J/(J + 1)$
If not in this table, use a conservative value	0.85

where  $J$  is the number of repeated measures.

## One sample: Test on proportion (from a large population)

### Example 17



The probability of suffering from De Quervain tenosynovitis (Blackberry finger) in the general population is  $p_0 = 0.01$ . It is suspected that amongst heavy smartphone users, this probability is larger. How many subjects (heavy smartphone users) do we need to study to determine if this is true? We want a power of 0.9 when  $p$  is at least 0.03.  $\alpha = 0.05$ .

Solution:

Our test is of the form

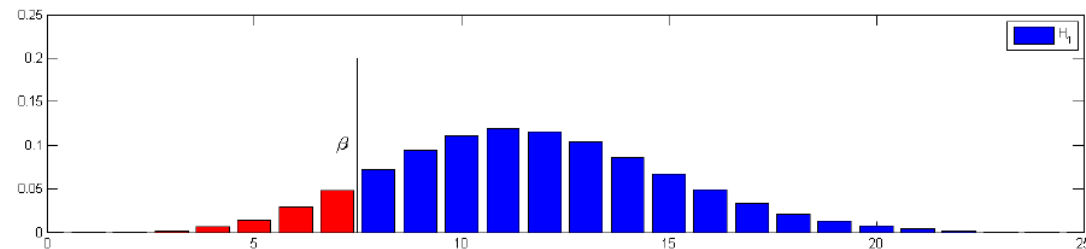
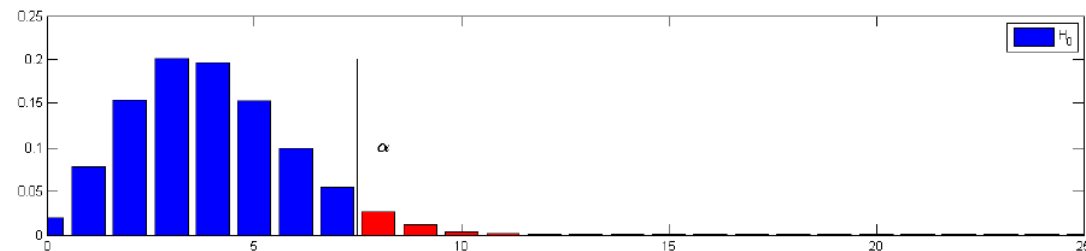
$$H_0 : p \leq p_0$$

$$H_A : p > p_0$$

## One sample: Test on proportion (from a large population)

We need to find  $N$  and  $X$  such that

$$\sum_{x=X+1}^{\infty} \binom{N}{x} p_0^x (1 - p_0)^{N-x} = \alpha \text{ and } \sum_{x=0}^X \binom{N}{x} p_1^x (1 - p_1)^{N-x} = \beta \quad (9)$$

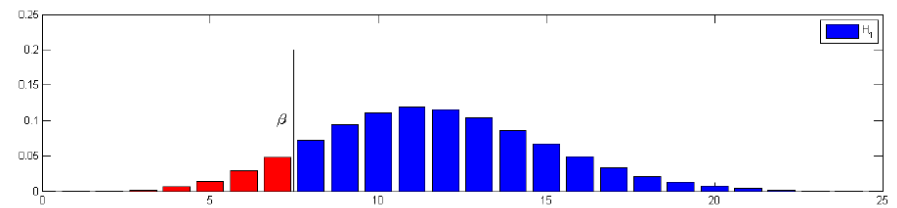
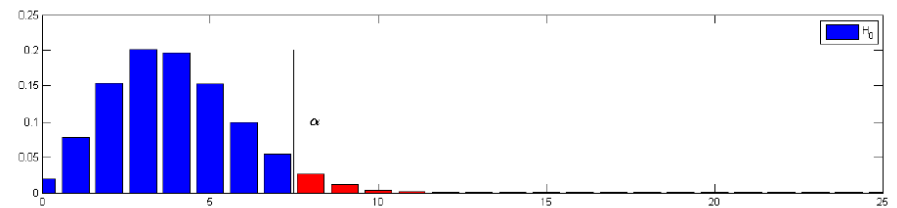
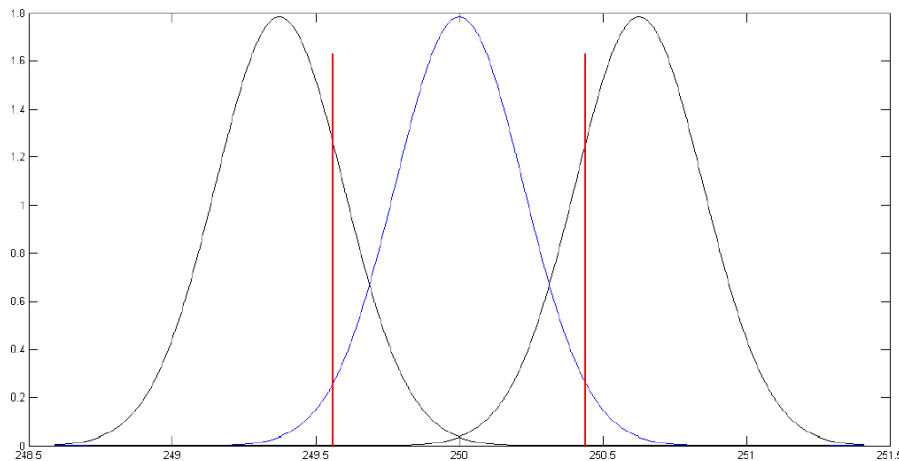


## Example (continued)

In our example  $p_0 = 0.01$  and  $p_1 = 0.03$ . The solution is

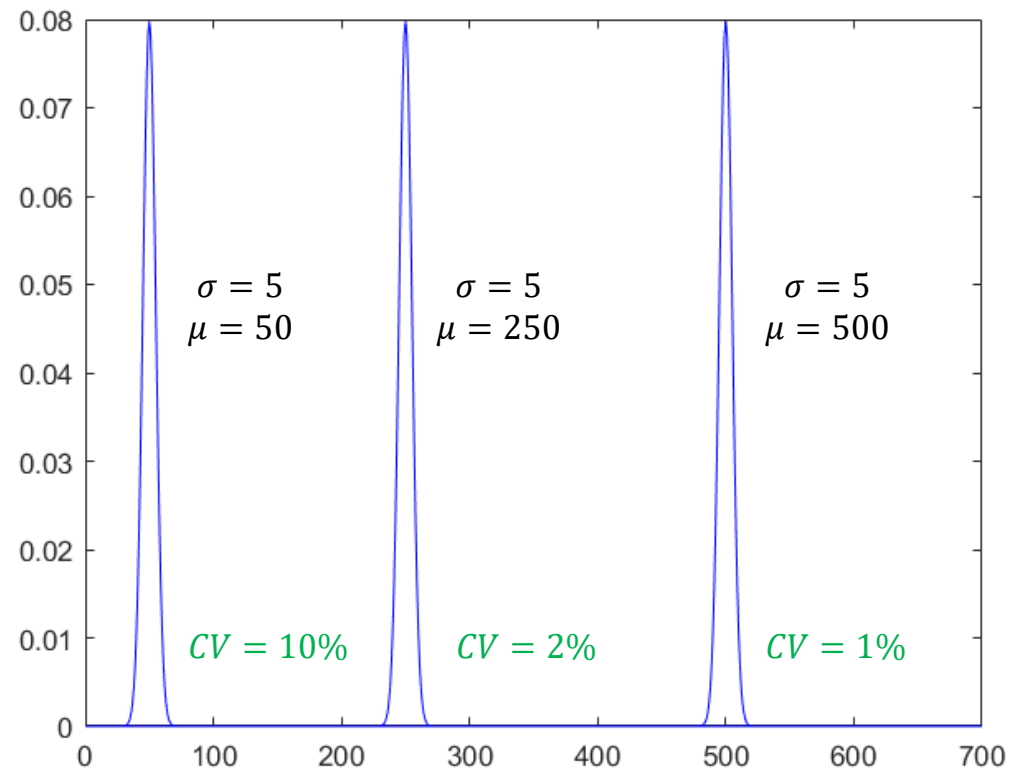
$$N = 390, X = 7$$

That is, we will evaluate 390 individuals. If the number of individuals with De Quervain syndrome is 7 or less, we cannot reject the null hypothesis ( $H_0 : p \leq p_0$ ).



## Coefficient of variation

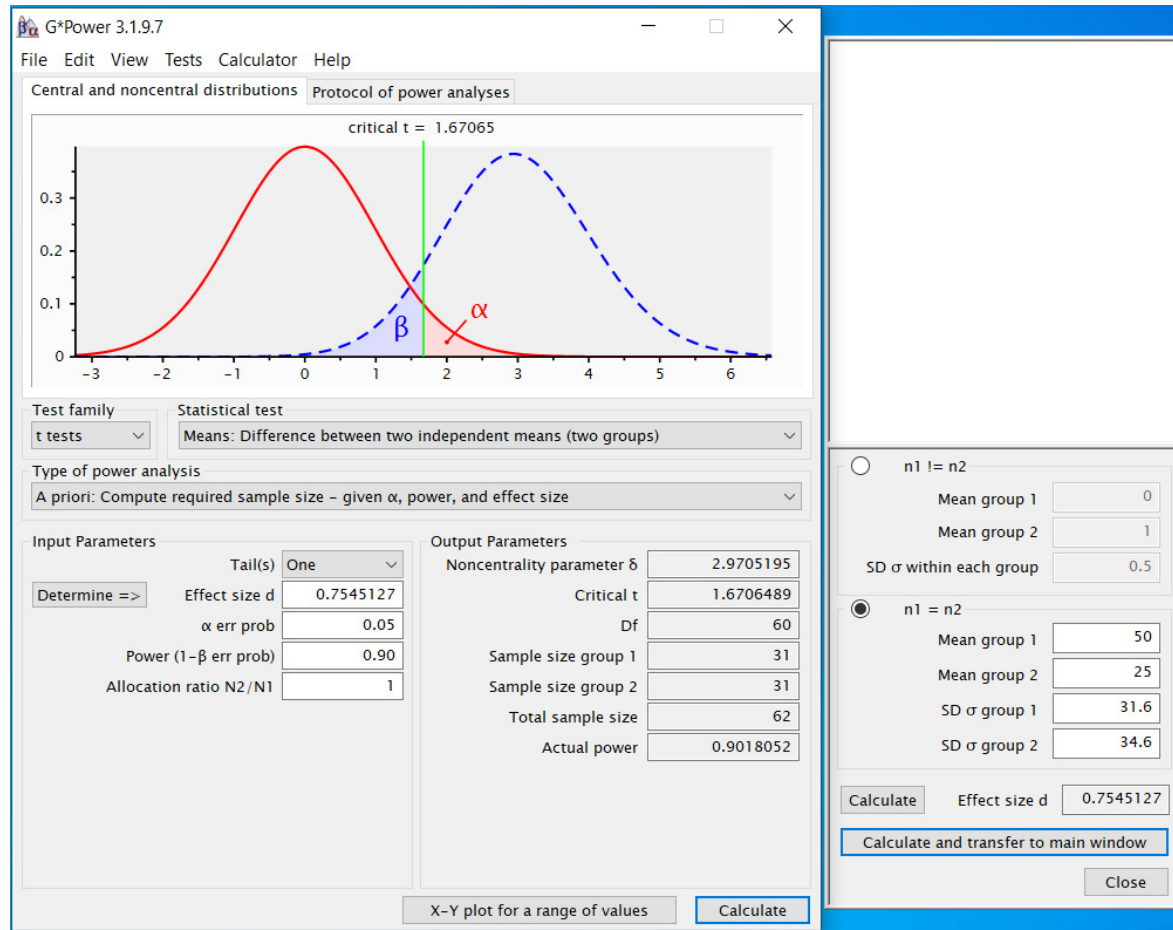
$$CV = \frac{\sigma}{\mu}$$



# •5

Software for sample size calculation

# Sample size calculation



# Sample size calculation

NCSS Pass

The screenshot displays the NCSS Pass software interface. The main window is titled "Two-Sample T-Tests Assuming Equal Variance (Enter Means)". The left sidebar shows a "Select a Procedure" menu with a tree structure under "Means" > "Two Independent Means". The main panel shows the "Design" tab with the following settings:

- Solve for:** Sample Size
- Test Direction:** Alternative Hypothesis:  $H_a: \mu_1 > \mu_2$
- Power and Alpha:** Power: 0,90; Alpha: 0,05
- Sample Size:** Group Allocation: Equal ( $N_1 = N_2$ )
- Means:**  $\mu_1$ : 0,8;  $\mu_2$ : 0,2
- Standard Deviation:**  $\sigma$ : 0,06666

The right sidebar contains a "Help Center" with links to documentation, examples, and validation, and an "Option Info" section for the "Standard Deviation Estimator".



# Sample size calculation

PASS Sample Size Software

[NCSS.com](http://NCSS.com)

## Chapter 422

# Two-Sample T-Tests Assuming Equal Variance (Enter Means)

### Introduction

This procedure provides sample size and power calculations for one- or two-sided two-sample t-tests when the variances of the two groups (populations) are assumed to be equal. This is the traditional two-sample t-test (Fisher, 1925). There are two PASS procedures for two-sample t-tests assuming equal variance. In this procedure, the assumed difference between means is specified by entering the means for the two groups and letting the software calculate the difference. If you wish to enter the difference directly, you can use the Two-Sample T-Tests Assuming Equal Variance (Enter Difference) procedure.

The design corresponding to this test procedure is sometimes referred to as a *parallel-groups* design. This design is used in situations such as the comparison of the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs.

There are several statistical tests available for the comparison of the center of two populations. This procedure is specific to the two-sample t-test assuming equal variance. You can examine the sections below to identify whether the assumptions and test statistic you intend to use in your study match those of this procedure, or if one of the other PASS procedures may be more suited to your situation.

### Other PASS Procedures for Comparing Two Means or Medians

Procedures in PASS are primarily built upon the testing methods, test statistic, and test assumptions that will be used when the analysis of the data is performed. You should check to identify that the test procedure described below in the Test Procedure section matches your intended procedure. If your assumptions or testing method are different, you may wish to use one of the other two-sample procedures available in PASS. These procedures are

PASS Sample Size Software

[NCSS.com](http://NCSS.com)

## Two-Sample T-Tests Assuming Equal Variance (Enter Means)

### Test Assumptions

When running a two-sample equal-variance t-test, the basic assumptions are that the distributions of the two populations are normal, and that the variances of the two distributions are the same. If those assumptions are not likely to be met, another testing procedure could be used, and the corresponding procedure in PASS should be used for sample size or power calculations.

### Test Procedure

If we assume that  $\mu_1$  and  $\mu_2$  represent the means of the two populations of interest, the null hypothesis for comparing the two means is  $H_0 : \mu_1 = \mu_2$ . The alternative hypothesis can be any one of

$$H_1 : \mu_1 \neq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

depending upon the desire of the researcher or the protocol instructions. A suitable Type I error probability ( $\alpha$ ) is chosen for the test, the data is collected, and a  $t$ -statistic is generated using the formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

This  $t$ -statistic follows a  $t$  distribution with  $n_1 + n_2 - 2$  degrees of freedom. The null hypothesis is rejected in favor of the alternative if,

for  $H_1 : \mu_1 \neq \mu_2$ ,

$$t < t_{\alpha/2} \text{ or } t > t_{1-\alpha/2},$$

for  $H_1 : \mu_1 > \mu_2$ ,

$$t > t_{1-\alpha},$$

or, for  $H_1 : \mu_1 < \mu_2$ ,

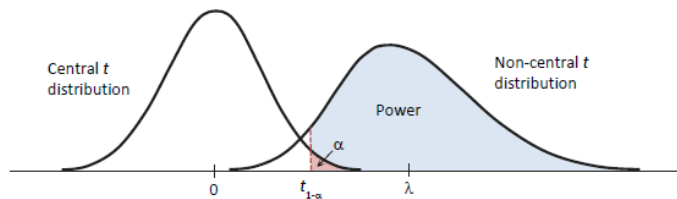
$$t < t_{\alpha}.$$

# Sample size calculation

## Power Calculation

This section describes the procedure for computing the power from  $n_1$  and  $n_2$ ,  $\alpha$ , the assumed  $\mu_1$  and  $\mu_2$ , and the assumed common standard deviation,  $\sigma_1 = \sigma_2 = \sigma$ . Two good references for these methods are Julious (2010) and Chow, Shao, and Wang (2008).

The figure below gives a visual representation for the calculation of power for a one-sided test.



If we call the assumed difference between the means,  $\delta = \mu_1 - \mu_2$ , the steps for calculating the power are as follows:

1. Find  $t_{1-\alpha}$  based on the central- $t$  distribution with degrees of freedom,

$$df = n_1 + n_2 - 2.$$

2. Calculate the non-centrality parameter:

$$\lambda = \frac{\delta}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

3. Calculate the power as the probability that the test statistic  $t$  is greater than  $t_{1-\alpha}$  under the non-central- $t$  distribution with non-centrality parameter  $\lambda$ :

$$Power = \Pr_{\text{non-central-}t}(t > t_{1-\alpha} \mid df = n_1 + n_2 - 2, \lambda).$$

The algorithms for calculating power for the opposite direction and the two-sided hypotheses are analogous to this method.

When solving for something other than power, PASS uses this same power calculation formulation, but performs a search to determine that parameter.

## Example 1 – Finding the Sample Size

Researchers wish to compare two types of local anesthesia to determine whether there is a difference in time to loss of pain. Subjects will be randomized to treatment, the treatment will be administered, and the time to loss of pain measured. The anticipated time to loss of pain for one of the types of anesthesia is 9 minutes. The researchers would like to generate a sample size for the study with 90% power to reject the null hypothesis of equal loss-of-pain time if the true difference is at least 2 minutes. How many participants are needed to achieve 90% power at significance levels of 0.01 and 0.05?

Past experiments of this type have had standard deviations in the range of 1 to 5 minutes. It is anticipated that the standard deviation of the two groups will be equal.

It is unknown which treatment has lower time to loss of pain, so a two-sided test will be used.

# •6

## Linear models

## Completely Randomized Design

### Example 0



We are testing a new drug (X 325mg) for blood pressure versus a placebo on 1000 people. We divide the group of people in two equal groups of 500 people. Each person will be randomly assigned to the treatment or the placebo.

$y_{11}$	$y_{21}$
$y_{12}$	$y_{22}$
...	...
$y_{1,500}$	$y_{2,500}$

- $y_{1.}, y_{2.}$ : Means of each one of the groups
- $y_{..}$ : Overall mean

# Completely Randomized Design

The data (blood pressure) is supposed to be generated as

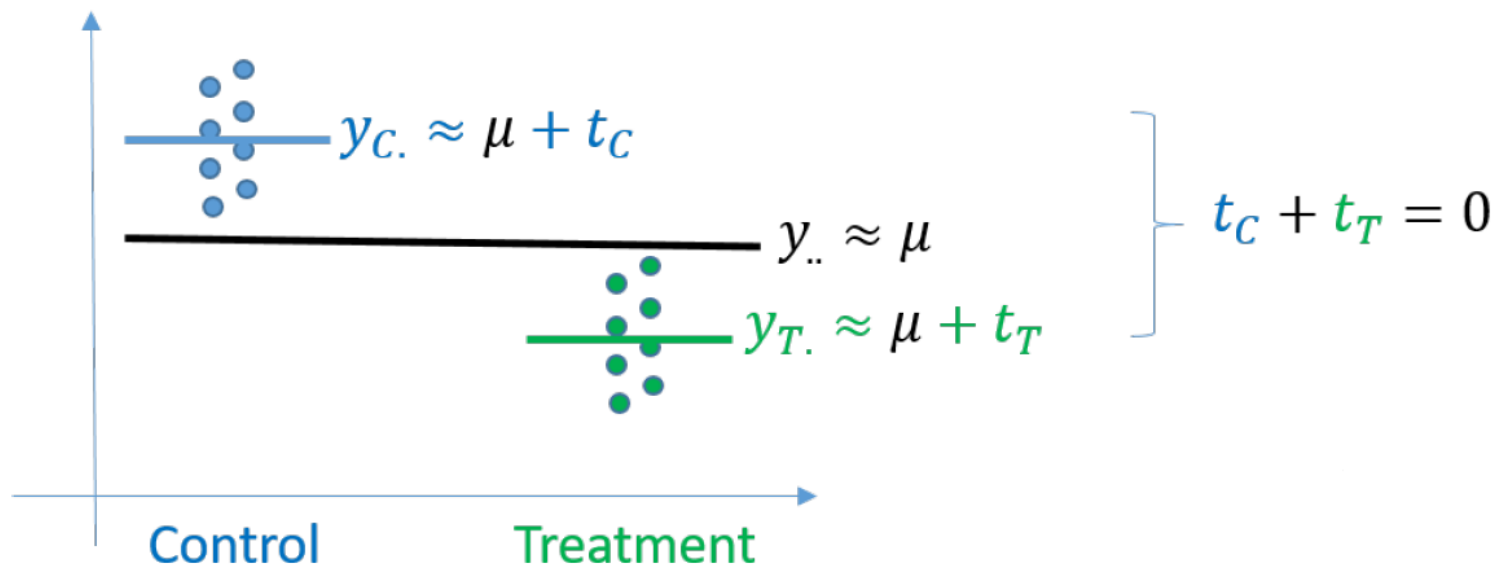
$$y_{jk} = \mu + t_j + \epsilon_{jk}$$

- $\mu$  is the average blood pressure of the whole population.
- $t_1$  and  $t_2$  are the effects of the drug ( $t_1$ ) and the placebo ( $t_2$ ). It must be

$$\sum_j t_j = 0$$

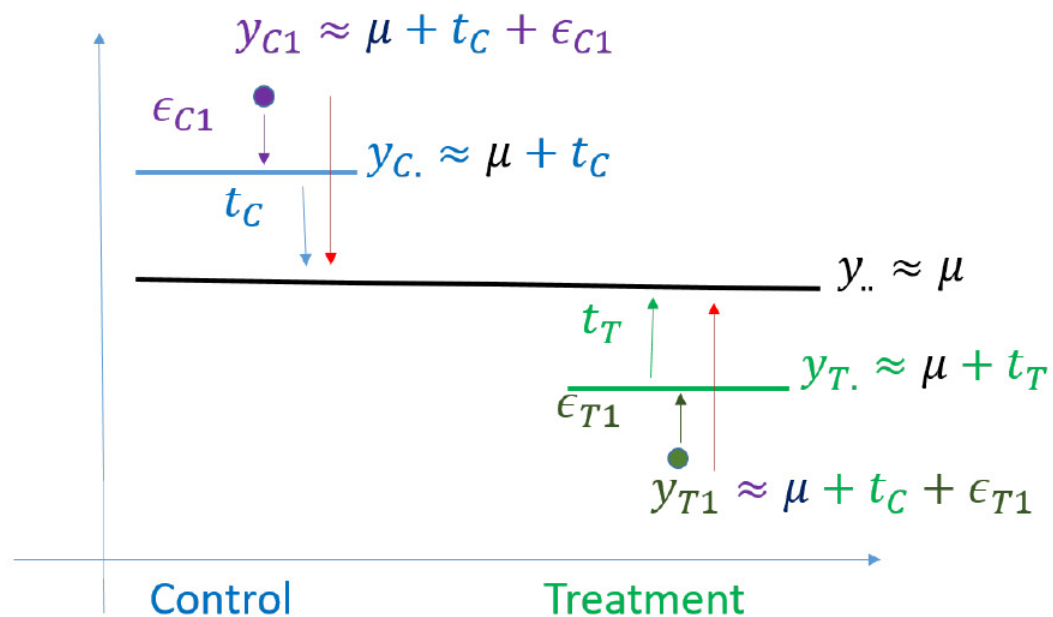
- $y_{jk}$  is the measurement observed for the  $k$ -th individual who has been given treatment  $j$ .
- $\epsilon_{jk}$  is the part of the observed measurement that cannot be explained by the average and the treatment.

## Completely Randomized Design



## Completely Randomized Design

$$\underbrace{\sum_j (y_{Cj} - y_{..})^2 + \sum_j (y_{Tj} - y_{..})^2}_{\text{Total variation}} = \underbrace{\sum_j (t_C)^2 + \sum_j (t_T)^2}_{\text{Control/Treatment}} + \underbrace{\sum_j (\epsilon_{Cj})^2 + \sum_j (\epsilon_{Tj})^2}_{\text{Noise}}$$



## Completely Randomized Design

Normally this is presented in a table

Source	Sum of Squares (SS)	Degrees of freedom (df)	Mean squares (MS=SS/df)
Treatments	$SS_T = \sum_{jk} (y_{j.} - y_{..})^2$	$t - 1$	$MS_T = \frac{SS_T}{df_t}$
Residuals	$SS_\epsilon = \sum_{jk} (y_{jk} - y_{j.})^2$	$\sum_j (n_j - 1) = n - t$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	$SS = \sum_{jk} (y_{jk} - y_{..})^2$	$n - 1$	

If the residuals are normally distributed, then the Linear Model checks whether the treatments have a significant contribution explaining the variance through a F-Snedecor statistic with  $t - 1$  and  $\sum_j (n_j - 1)$  degrees of freedom.

$$F = \frac{MS_T}{MS_\epsilon}$$



# Completely Randomized Design

## Example 1

Let us assume that the table in our case is

Source	SS	df	MS=SS/df
Treatments	256.88	1	256.88
Residuals	13600.28	998	13.61
Total	13857.16	999	



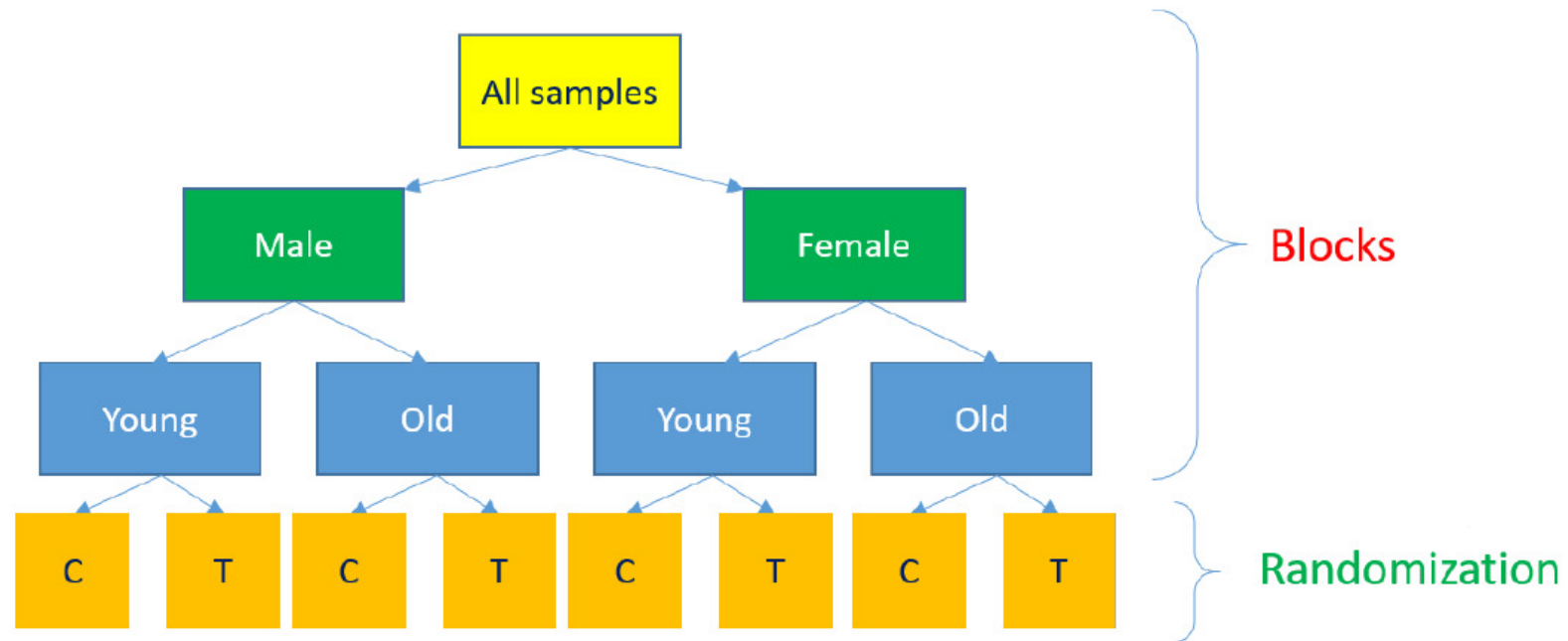
Note

$$\begin{aligned} 13857.16 &= 256.88 + 13600.28 \\ 999 &= 1 + 998 \end{aligned}$$

In this case

$$F = \frac{256.88}{13.61} = 18.87 \gg 3.85 = F_{0.95,1,998}$$

# Randomized Complete Block Design



## Randomized Complete Block Design

The data (blood pressure) is supposed to be generated as

$$y_{ijk} = \mu + b_i + t_j + \epsilon_{ijk}$$

- $\mu$  is the average blood pressure of the whole population.
- $b_1$  and  $b_2$  are the differences in blood pressure between men ( $b_1$ ) and women ( $b_2$ ), the blocks. It must be

$$\sum_i b_i = 0$$

- $t_1$  and  $t_2$  are the effects of the drug ( $t_1$ ) and the placebo ( $t_2$ ). It must be

$$\sum_j t_j = 0$$

- $y_{ijk}$  is the measurement observed for the  $k$ -th individual of the  $i$ -th block who has been given treatment  $j$ .
- $\epsilon_{ijk}$  is the part of the observed measurement that cannot be explained by the average, block and treatment.

## Randomized Complete Block Design

The table of the linear model becomes

Source	SS	df	MS=SS/df
Blocks	$SS_B$	$b - 1$	$MS_B = \frac{SS_B}{df_B}$
Treatments	$SS_T$	$t - 1$	$MS_T = \frac{SS_T}{df_T}$
Residuals	$SS_\epsilon$	$n - b - t + 1$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	SS	$n - 1$	

If the residuals are Gaussian, we may test whether the contribution of the blocks or treatments are significant through the same F-Snedecor as before (pay attention to use the corresponding degrees of freedom).

## Randomized Complete Block Design

### Example 2

Let us assume that in our case it becomes

Source	SS	df	MS=SS/df
Blocks	1500.04	1	1500.04
Treatments	256.88	1	256.88
Residuals	12100.24	997	12.13
Total	13857.16	999	



Note

$$\begin{aligned} 13857.16 &= 1500.04 + 256.88 + 12100.24 \\ 999 &= 1 + 1 + 997 \end{aligned}$$

In this case

$$F = \frac{256.88}{12.13} = 21.17 \gg 3.85 = F_{0.95,1,997}$$

# Factorial Design

We are measuring the effect of a treatment on a number of animals as a function of **age** and **sex**. These are called **factors**, and their different values are called **levels**. For each combination we have  $N = 6$  animals. The numbers below show the average of each one of the groups.

$$Y = \mu + t_{group} + \epsilon$$

All:	5	$\mu$	
Group 1: young, male	7	$= 5 + 2$	} $t_{group}$
Group 2: young, female	5	$= 5 + 0$	
Group 3: old, male	5	$= 5 + 0$	
Group 4: old, female	3	$= 5 - 2$	

# Factorial Design

However, we could have analyzed the data differently gaining **more insight** into the influence of each factor. This kind of analysis is called **main effects**.

$$Y = \mu + t_{group} + \epsilon$$

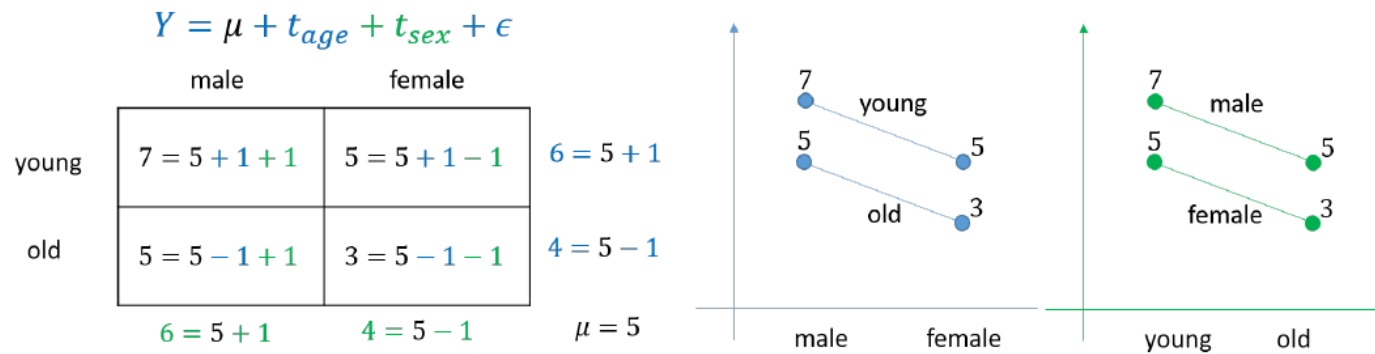
All:	5	$\mu$	
Group 1: young, male	7	$= 5 + 2$	} $t_{group}$
Group 2: young, female	5	$= 5 + 0$	
Group 3: old, male	5	$= 5 + 0$	
Group 4: old, female	3	$= 5 - 2$	

$$Y = \mu + t_{age} + t_{sex} + \epsilon$$

	male	female	
young	7 = 5 + 1 + 1	5 = 5 + 1 - 1	$t_{young} = 1$
old	5 = 5 - 1 + 1	3 = 5 - 1 - 1	$t_{old} = -1$
	$t_{male} = 1$	$t_{female} = -1$	$\mu = 5$

# Factorial Design

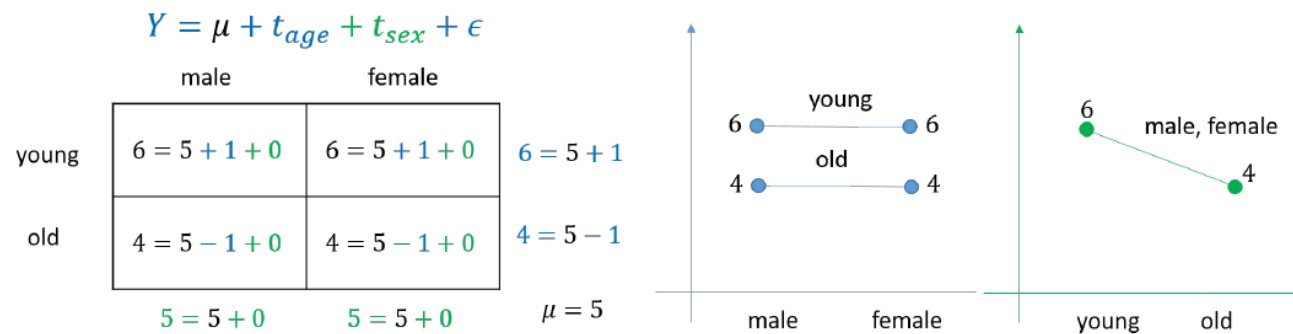
We may arrange the response graphically. Note the fact that the two lines are **parallel**.





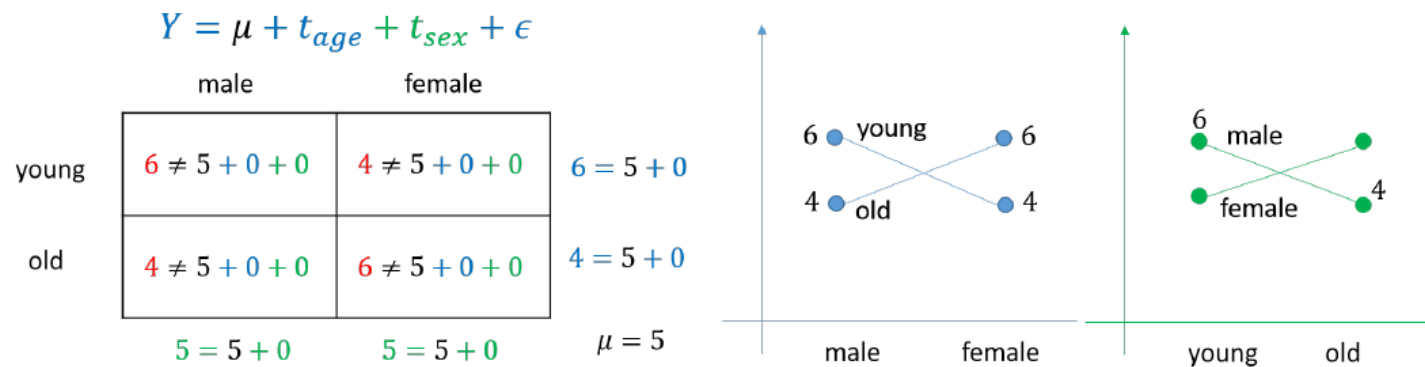
# Factorial Design

In the following example, only one of the factors has an effect. The lines are still **parallel** or **coincident**.



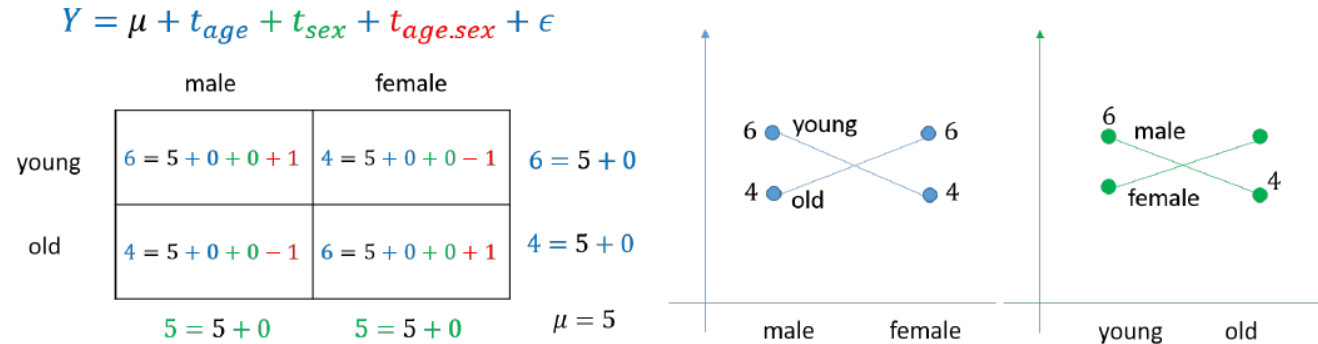
# Factorial Design

Main effects alone are not able to explain the data. Lines are not parallel anymore.



# Factorial Design

We need to add interactions to be able to explain the data. **Interaction effects** exist when differences on one factor depend on the level you are on another factor. The interactions are **between factors** and **not between levels**.



## Factorial Design

The analysis table may be represented as

Source	SS	df	MS=SS/df
$P$ main effects	$SS_P$	$p - 1$	$MS_P = \frac{SS_P}{df_P}$
$Q$ main effects	$SS_Q$	$q - 1$	$MS_Q = \frac{SS_Q}{df_Q}$
$PQ$ interactions	$SS_{PQ}$	$(p - 1)(q - 1)$	$MS_{PQ} = \frac{SS_{PQ}}{df_{PQ}}$
Residuals	$SS_\epsilon$	$n - pq$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	$SS$	$n - 1$	

# Factorial Design

## Example 4

Assume that we have resources for 24 observations and we assume that there is no interaction between factors

$$y_{ijkl} = \mu + s_i + m_j + h_k + \epsilon_{ijkl}$$

Three different experiment designs are considered:

- ① One variable changes at a time
  - (Frogs, Dry, NoHormone) vs (Toad, Dry, NoHormone): 4 animals each
  - (Frogs, Dry, NoHormone) vs (Frogs, Wet, NoHormone): 4 animals each
  - (Frogs, Dry, NoHormone) vs (Frogs, Dry, Hormone): 4 animals each
- ② Do not repeat (Frogs, Dry, NoHormone) in each comparison:
  - (Frogs, Dry, NoHormone): 6 animals
  - (Toads, Dry, NoHormone): 6 animals
  - (Frogs, Wet, NoHormone): 6 animals
  - (Frogs, Dry, Hormone): 6 animals
- ③ Factorial design (all possible combinations) with 3 animals each.

# Factorial Design

## Example 3

We are testing water uptake by amphibia. Frogs and toads (species factor  $S$ ) are kept in moist or dry conditions before the experiment (moisture factor  $M$ ) and half of the animals are injected with a mammalian water balance hormone (hormone factor  $H$ ). A full factorial experiment is performed with 2 animals per treatment combination (cell).



$$y_{ijkl} = \mu + s_i + m_j + h_k + (sm)_{ij} + (sh)_{ik} + (mh)_{jk} + \epsilon_{ijkl}$$

Source	SS	df	MS		Source	SS	df	MS
Species	515.06	1			Species	515.06	1	
Moisture	471.33	1			Moisture	471.33	1	
Hormone	218.01	1			Hormone	218.01	1	
SM	39.50	1			SH	165.12	1	
SH	165.12	1			Lack of fit	140.71	3	46.90
MH	57.73	1			Error	276.05	8	$s^2 = 34.51$
SMH	43.43	1			Total	1786.33	15	
Error	276.05	8	$s^2 = 34.51$					
Total	1786.33	15						

## Factorial designs and single replicates

High-order interactions can be assimilated to the error, and single replicate factorial designs may be conceived.

### Example 5



We are interested in the survival of *Salmonella typhimurium* under 3 experimental factors: 3 levels of sorbic acid (=Factor *S*), 6 levels of water activity (=Factor *A*), and 3 levels of pH (=Factor *P*). The data will be the log (density/ml) measured after 7 days after treatment started.

We have  $3 \times 6 \times 3 = 54$  treatments, and we will use a single replicate for each treatment.

## Factorial designs and single replicates

### Example 5(continued)

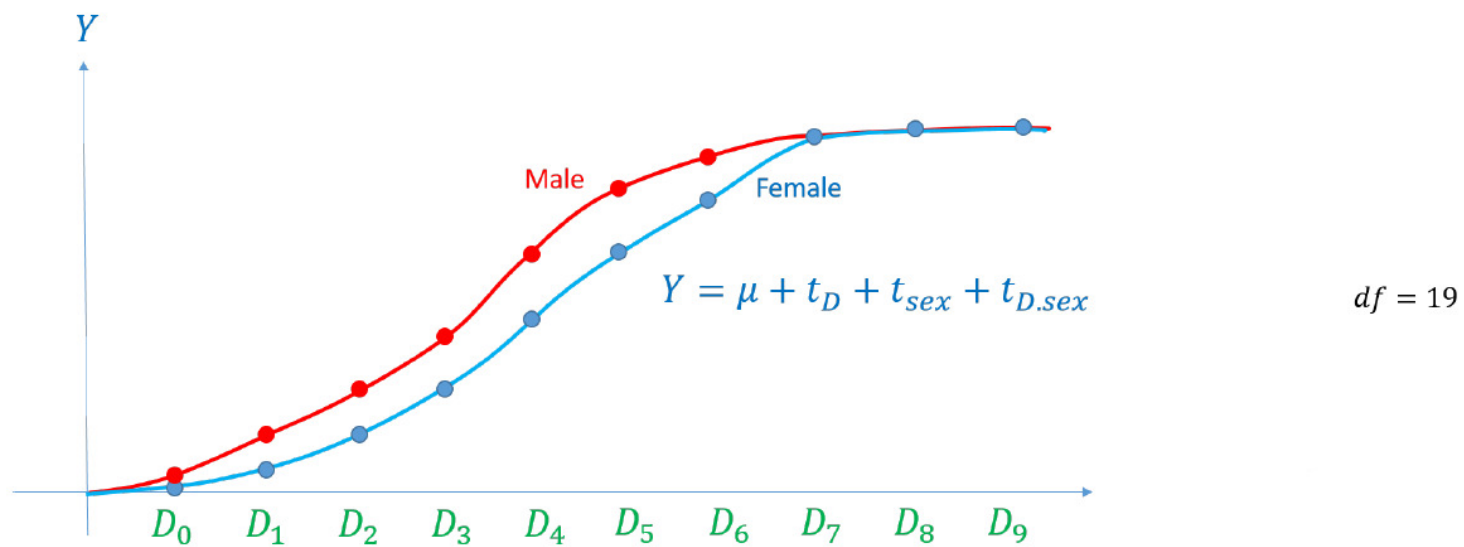
The data analysis table would be

	SS	df	MS	F
Water activity ( <i>A</i> )	81.57	5=(6-1)	16.31	473 > $F_{0.95,5,20}$
Sorbic acid ( <i>S</i> )	2.76	2=(3-1)	1.38	40 > $F_{0.95,5,20}$
pH ( <i>P</i> )	0.01	2=(3-1)	0.01	0.2 < $F_{0.95,2,20}$
<i>AS</i>	1.32	10=(6-1)(3-1)	0.13	3.8 > $F_{0.95,10,20}$
<i>AP</i>	0.45	10=(6-1)(3-1)	0.04	1.3 < $F_{0.95,10,20}$
<i>SP</i>	0.23	4=(3-1)(3-1)	0.06	1.7 < $F_{0.95,4,20}$
<i>ASP</i> ≈ Error	0.69	20=(6-1)(3-1)(3-1)	0.03	
Total	87.03	53		

The problem with single replicate, factorial designs is that 1) it is difficult to use blocking, 2) due to the lack of replication, there is no possibility to construct an unbiased estimate of the noise.

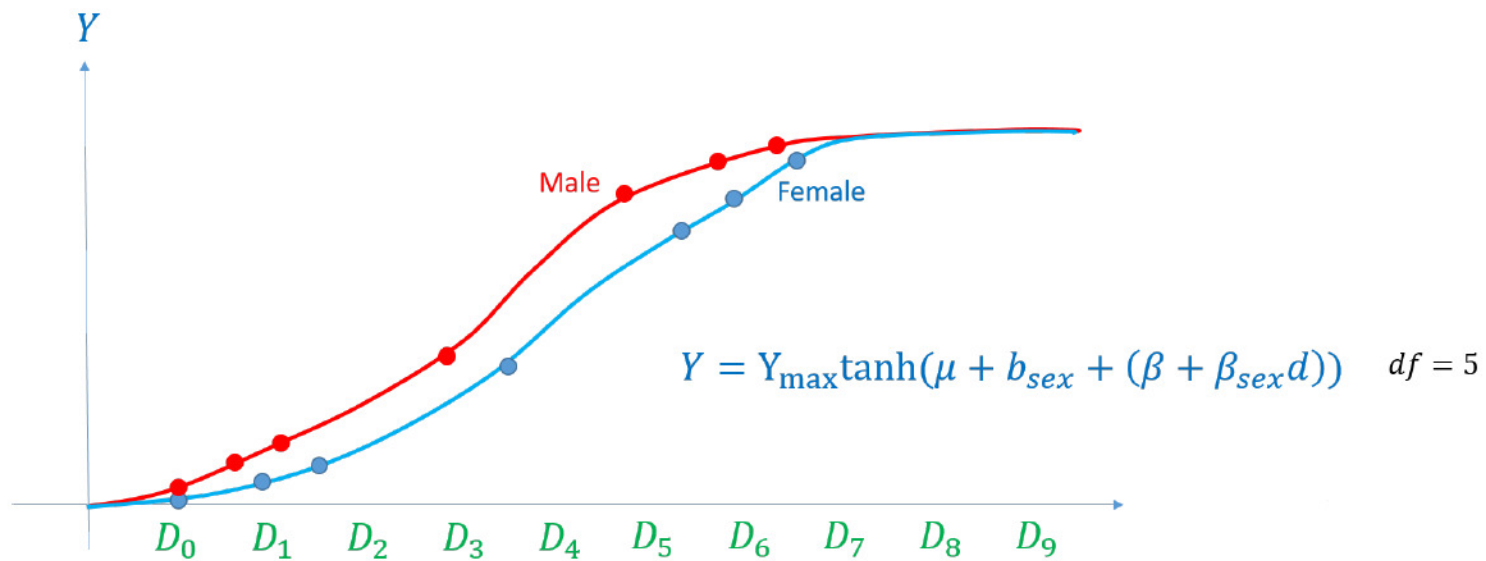


## Regression Design



Doses can be analyzed as a 2-way ANOVA, although we will need **more samples**.

## Regression Design



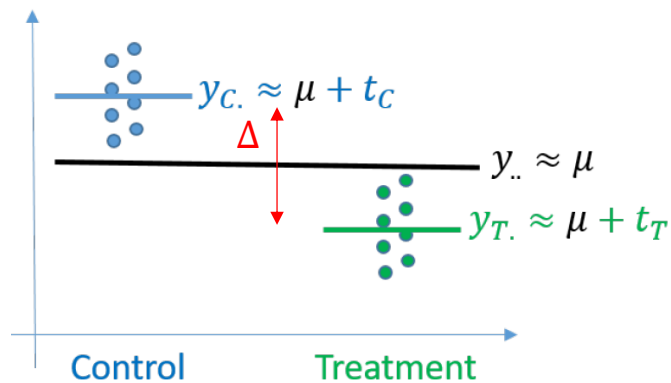
Doses can be analyzed as a regression, with fewer samples and located in different positions.



•7

Some specificities

# Confidence intervals for the difference of two independent means



$$\hat{\Delta} = y_{C.} - y_{T.}$$

$$y_{C.} \sim N\left(\mu_C, \frac{\sigma_C^2}{N_C}\right) \quad y_{T.} \sim N\left(\mu_T, \frac{\sigma_T^2}{N_T}\right)$$

$$\Delta \sim N\left(\mu_C - \mu_T, \frac{\sigma_C^2}{N_C} + \frac{\sigma_T^2}{N_T}\right)$$

$$CI_{\alpha} = \left[ \hat{\Delta} \pm t_{1-\frac{\alpha}{2}, N_C+N_T-2} S_{\Delta} \right]$$

Pooled variance

$$\sigma_C^2 = \sigma_T^2$$

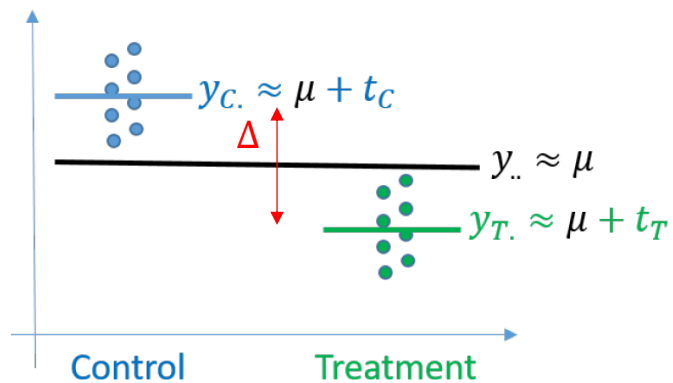
$$s_{\Delta}^2 = \frac{(N_C - 1)s_C^2 + (N_T - 1)s_T^2}{N_C + N_T - 2}$$

Non-pooled variance

$$\sigma_C^2 \neq \sigma_T^2$$

$$s_{\Delta}^2 = \frac{s_C^2}{N_C} + \frac{s_T^2}{N_T}$$

# Confidence intervals for the difference of two **dependent** means



$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^N \Delta_i$$

$$\Delta_i \sim N(\mu_{\Delta}, \sigma_{\Delta}^2)$$

$$CI_{\alpha} = \left[ \hat{\Delta} \pm t_{1-\frac{\alpha}{2}, N-1} s_{\Delta} \right]$$

	Pretest	Posttest	Delta
	74	98	24
	98	100	2
	85	98	13
	68	90	22
	79	90	11
	52	91	39
	80	84	4
	78	85	7
	81	93	12
Mean	77.2	92.1	14.9
StdDev	11.8	5.4	11.0
StdDev Pooled			4.32
StdDev Non-Pooled			9.15
StdDev Dependent			11.0
df Pooled	16	t 0.975	2.12
df Non-Pooled	16		2.12
df Dependent	8		2.31
CI		5.7	24.0
		-4.5	34.3
		-10.5	40.3

## Confidence intervals for the difference of two independent proportions

	Control	Treatment
Response	A	B
Non-response	C	D

$$\hat{p}_C = \frac{A}{N_C} \quad \hat{p}_T = \frac{B}{N_T}$$

$$\hat{\Delta} = \hat{p}_C - \hat{p}_T$$

Exact solutions are complicated

Approximate solution

$$CI_{\alpha} = \left[ \hat{\Delta} \pm z_{1-\frac{\alpha}{2}} s_{\Delta} \right]$$

$$s_{\Delta}^2 = \frac{\hat{p}_C(1 - \hat{p}_C)}{N_C} + \frac{\hat{p}_T(1 - \hat{p}_T)}{N_T}$$

# Confidence intervals for the difference of two dependent proportions

Same blood sample treated with drug A or drug B  
There are  $N$  blood samples.

	Response B	Non-response B
Response A	A	B
Non-response A	C	D

$$\hat{p}_A = \frac{A + B}{N} \quad \hat{p}_B = \frac{A + C}{N}$$

$$\hat{\Delta} = \hat{p}_A - \hat{p}_B$$

Exact solutions are complicated

Approximate solution (Wald)

$$CI_{\alpha} = \left[ \hat{\Delta} \pm z_{1-\frac{\alpha}{2}} s_{\Delta} \right]$$

$$s_{\Delta}^2 = \frac{(A + D)(B + C) + 4BC}{N^3}$$

	Response B	Non-Response B			
Response A	10	3	0.77		
Non-Response A	5	8	0.38		
	0.67	0.27			
Delta (Independent)	0.39		CI	0.04	0.75
Delta (Dependent)	0.08			0.05	0.10
StdDev Delta (Independent)	0.18				
StdDev Delta (Dependent)	0.01				

## Risk and related measures

	Control	Treatment	
Event	A	B	$N_E = A + B$
Non-event	C	D	$N_{NE} = C + D$
	$N_C = A + C$	$N_T = B + D$	

Rate in control group	$p(E C) = \frac{A}{N_C} = \frac{A}{A + C}$	Odds in control group	$O(E C) = \frac{p(E C)}{1 - p(E C)} = \frac{p(E C)}{p(NE C)} = \frac{A}{C}$
Rate in treated group	$p(E T) = \frac{B}{N_T} = \frac{B}{B + D}$	Odds in treated group	$O(E T) = \frac{p(E T)}{1 - p(E T)} = \frac{p(E T)}{p(NE T)} = \frac{B}{D}$
Absolute risk reduction	$ARR = p(E C) - p(E T)$	Number needed to Treat	$NTT = \frac{1}{ARR}$
Relative risk reduction	$RRR = \frac{ARR}{p(E C)}$		
Risk ratio	$RR = \frac{p(E C)}{p(E T)}$	Odds ratio	$OR = \frac{O(E C)}{O(E T)}$



## Risk and related measures

2000 men with high cholesterol without coronary heart disease were studied over 5 years. Half of them were given statins for the 5 years. We counted number of deaths from coronary disease within this period

	Control	Treatment	
Event	79	55	134
Non-event	921	945	1866
	1000	1000	2000

Rate in control group  $p(E|C) = \frac{79}{1000} = 7.9\%$

Odds in control group  $O(E|C) = \frac{7.9\%}{92.1\%} = 0.086$

Rate in treated group  $p(E|T) = \frac{55}{1000} = 5.5\%$

Odds in treated group  $O(E|T) = \frac{5.5\%}{94.5\%} = 0.058$

Absolute risk reduction  $ARR = 7.9 - 5.5 = 2.4\%$

Number needed to Treat  $NTT = \frac{1}{2.4\%} = 41.7$

Relative risk reduction  $RRR = \frac{2.4\%}{7.9\%} = 30.4\%$

Risk ratio  $RR = \frac{7.9}{5.5} = 1.4$

Odds ratio  $OR = \frac{0.086}{0.058} = 1.5$

## Risk and related measures

2000 men with high cholesterol without coronary heart disease were studied over 5 years. Half of them were given statins for the 5 years. We counted number of deaths from coronary disease within this period

	Control	Treatment	
Event	79	55	134
Non-event	921	945	1866
	1000	1000	2000

Absolute risk reduction  $ARR = 7.9 - 5.5 = 2.4\%$

You need to treat 100 men with high cholesterol for 5 years with statins to prevent 2.4 of them from dying from coronary disease.

Number needed to Treat  $NTT = \frac{1}{2.4\%} = 41.7$

You need to treat 41.7 men with high cholesterol for 5 years with statins to prevent 1 of them from dying from coronary disease. This is the technically preferred way.

Relative risk reduction  $RRR = \frac{2.4\%}{7.9\%} = 30.4\%$

This is the preferred way to present by media.

## Risk and related measures (rare events)

Study of thromboembolic events in reproductive-age women on oral contraceptives.

	Control	Treatment	
Event	1	7	8
Non-event	99999	99993	199992
	100000	100000	200000

Absolute risk reduction  $ARR = 0.007 - 0.001 = 0.006\%$

You need to stop oral contraceptives in 100 women to prevent 0.006 of them from a thromboembolic event.

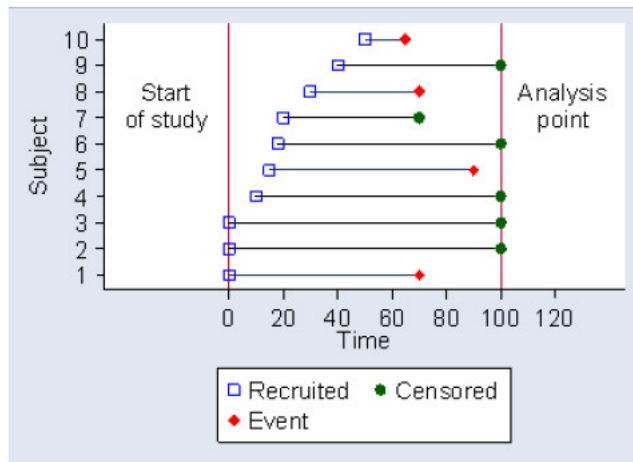
Number needed to Treat  $NTT = \frac{1}{0.006\%} = 16667$

You need to stop oral contraceptives on 16667 women to prevent 1 of them from a thromboembolic event. This is the technically preferred way.

Relative risk reduction  $RRR = \frac{0.006\%}{0.007\%} = 86.4\%$

This is the preferred way to present by media.

## Survival data (Life table analysis)



Survival data measures the **time to a well-defined event** such as

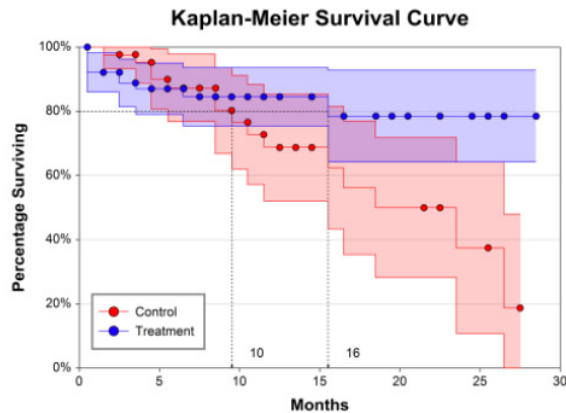
- ... death
- ... occlusion of a vascular graft
- ... first metastasis
- ... rejection of a transplanted kidney

Data is **censored**

- ... when we stop observing the subject at the end of the study.
- ... if they cease to collaborate.
- ... if they die from a different reason from that of the experiment.

# Kaplan-Meier analysis

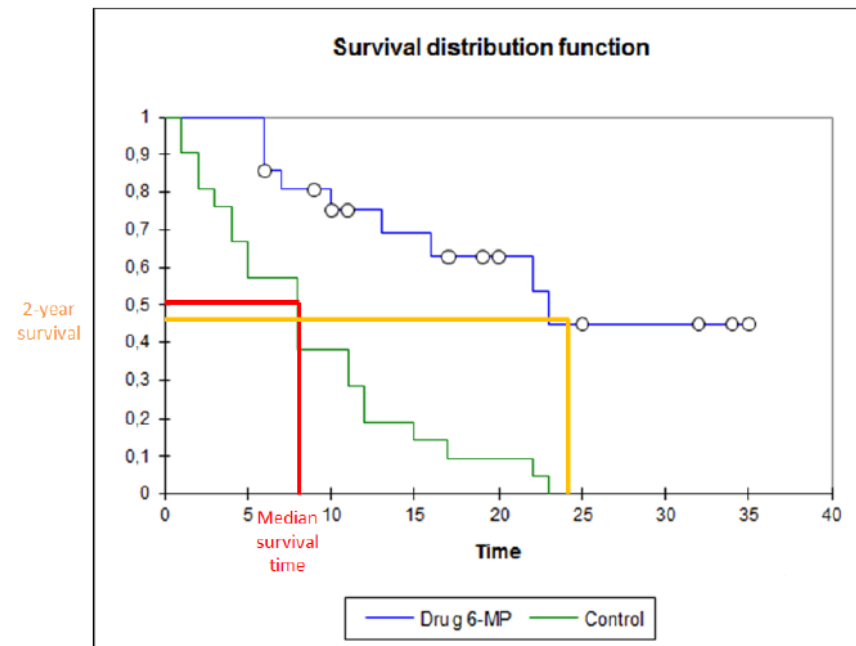
Time Period	At Risk	Became Unavailable (Censored)	Died	Survived	Kaplan-Meier Survival Probability Estimate
Year 1	100	3	5	95	$(95/100)=0.95$
Year 2	92	3	10	82	$(95/100) \times (82/92)=0.8467$
Year 3	79	3	15	64	$(95/100) \times (82/92) \times (64/79)=0.70$
Year 4	61	3	20	41	$(95/100) \times (82/92) \times (64/79) \times (41/61)=0.4611$
Year 5	38	3	25	13	$(95/100) \times (82/92) \times (64/79) \times (41/61) \times (13/38)=0.1577$



In this plot, red and blue points indicate censored data.

At each point in time we may create a confidence interval as shown in the figure.

## Survival summary



We may summarize survival data through:

- Median survival time (50% of the samples still survive)
- Two-year survival (survival proportion at a given time)

## Assumptions

- **Random sample.** So that the sample is representative from the population.
- **Independent subjects.** If the study pools from two different hospitals, each hospital with different average survival, then the proportion of individuals from each hospital will distort the survival curve.

If the studied disease has a genetic component, including family members in one treatment group distorts the survival curve.

- **Entry criteria are consistent.** If the study lasts for years, the enrollment criteria cannot change over time. For instance, cancer patients are enrolled at their first metastasis, but over the years new technology allows for earlier diagnosis.
- **End point is consistent.** In a cancer study, do we count deaths from car accidents as deaths? Counting or not counting makes sense, but the decision has to be taken before the study.
- **Average survival does not change over time.** If the nature of the disease changes over time (e.g., a rapidly evolving infectious pathogen), then results are difficult to interpret. If the treatment (including supportive care) changes over time, ...

# Assumptions

- **Starting time clearly defined.** For instance, the first hospital admission. Do not rely on the patient remembering when he first had symptoms.

Do we remove patients that they before they could start treatment? This leads to bias, especially if one treatment can start immediately (medication), but the other requires preparation or scheduling (surgery). Most study follow a policy of **intention to treat**.

- **Censoring is unrelated to survival.** If some patients dropout the study because they feel too sick or they thought the treatment was not useful, then the censored data is related to the disease progression or response to therapy and the analysis is invalid. In these cases it is recommended to analyze the data censoring the dropouts and excluding them. If the results of both analyses coincide, then the result is clear. If they do not coincide, then the study results are ambiguous.



# Survival, failure, and hazard curves

Survival curve

$$S(t) = P(T > t)$$

Failure curve

$$F(t) = P(T \leq t)$$

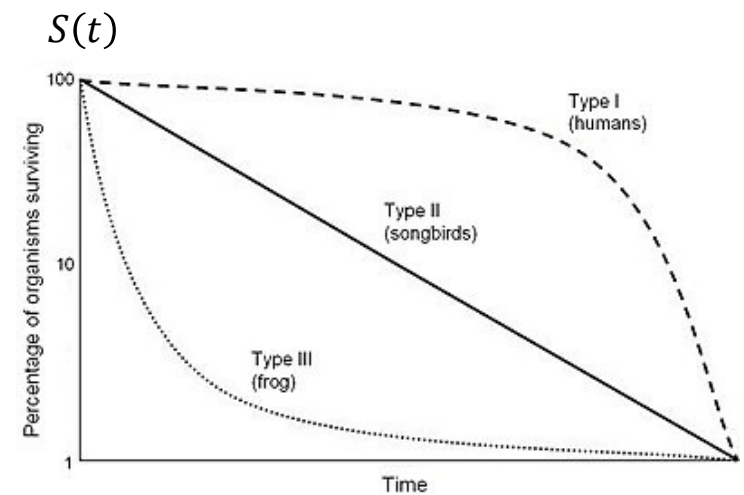
Hazard curve

“instantaneous  
probability of the  
event”

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$$
$$= \frac{dF(t)}{dt} \frac{1}{S(t)} = - \frac{d(\log S(t))}{dt}$$

Cumulative hazard  
curve

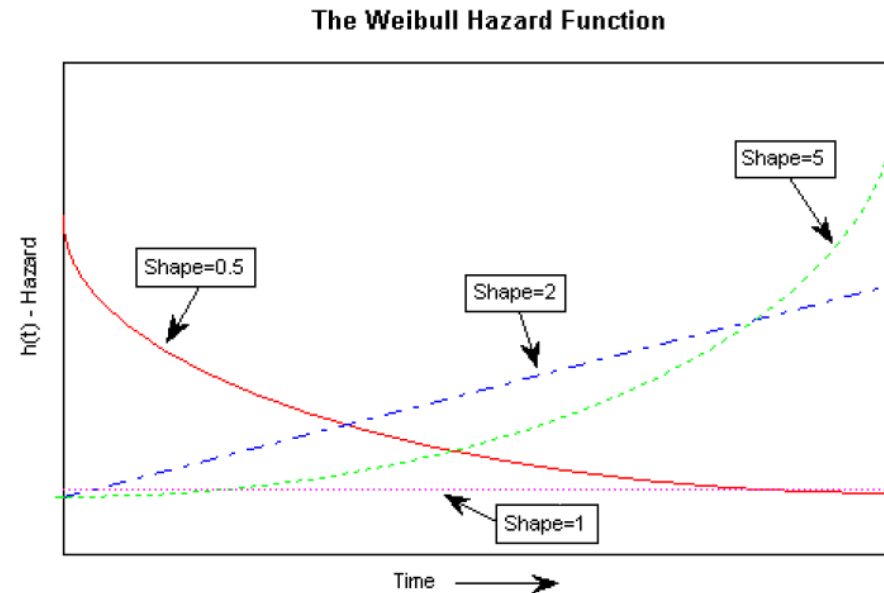
$$H(t) = \int_{-\infty}^t h(\tau) d\tau = -\log(S(t))$$



## One sample: Weibull survival

We may flexibilize the constant failure rate to a more general hazard function

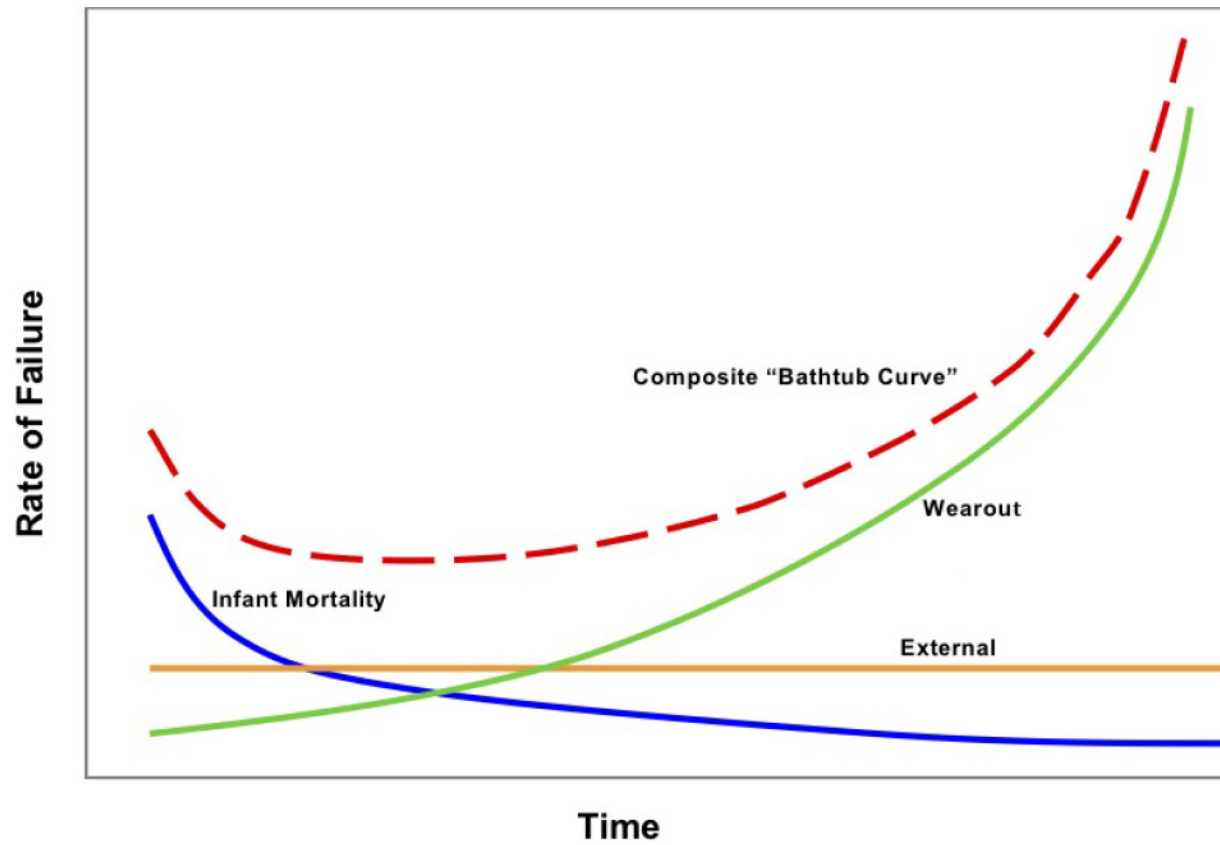
$$\begin{aligned}h(t) &= \frac{\beta}{\mu} \left(\frac{t}{\mu}\right)^{\beta-1} \\f(T) &= \frac{\beta}{\mu} \left(\frac{t}{\mu}\right)^{\beta} e^{-\left(\frac{t}{\mu}\right)^{\beta}} \\F(T) &= 1 - e^{-\left(\frac{t}{\mu}\right)^{\beta}} \\S(t) &= e^{-\left(\frac{t}{\mu}\right)^{\beta}}\end{aligned}$$



$\beta$  is a shape parameter,  $\mu$  a scale parameter and the mean survival time (MST) is

$$MST = \mu \Gamma \left( 1 + \frac{1}{\beta} \right)$$

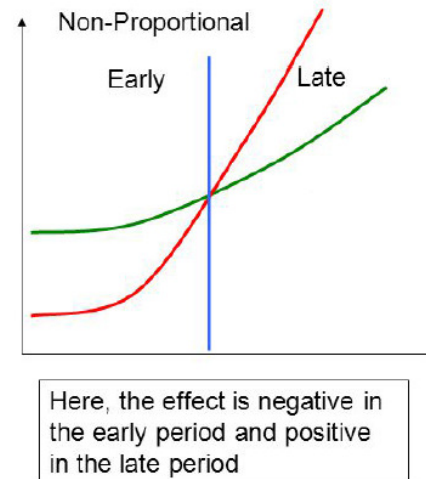
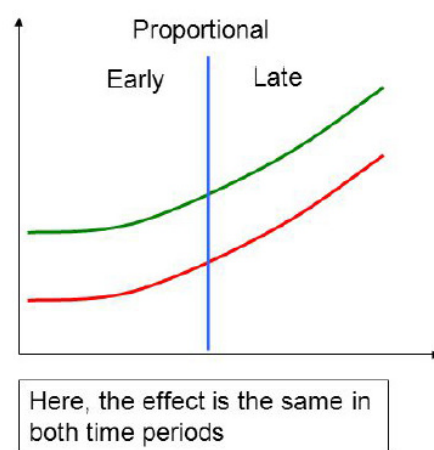
## One sample: Weibull survival



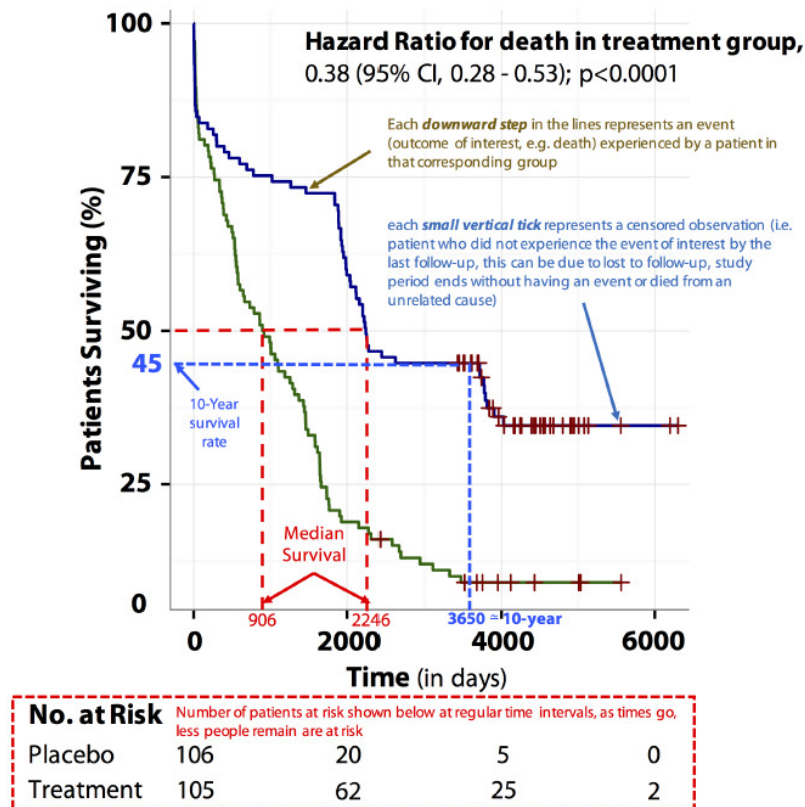
## Assumptions of Survival Analysis

When comparing two survival curves, additionally

- **Proportional hazards.** Hazard is the slope of the survival curve. The hazard ratio compares the hazard of both treatments, most tests assume that this ratio is constant over time and differences are simply due to random sampling. This assumption is violated when hazard changes over time. For instance, comparing surgery (high initial risk, lower later risk) with medical therapy (less initial risk, higher later risk).



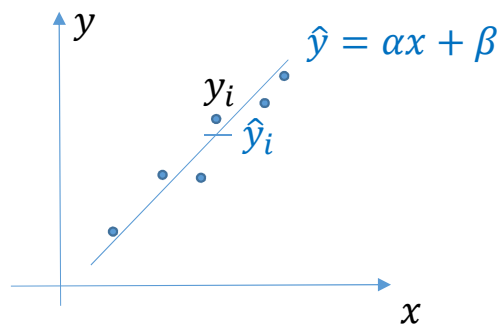
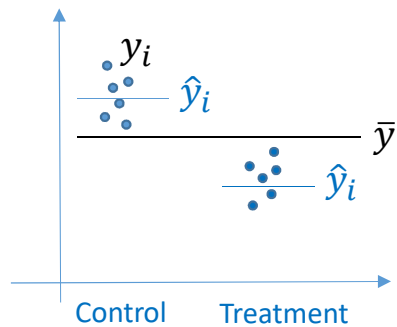
# Survival Analysis



If the proportional hazard assumption is accepted, you may use a **Hazard Ratio analysis** (related to Cox model). In this example the death hazard in one of the groups is 0.38 lower than in the other group. **The log-rank method or Mantel-Cox method** calculates a p-value under this assumption

If the hazard is constant over time, then we may also use the **Ratio of median survival times** (RMST, related to an exponential decay). In this example,  $RMST = \frac{906}{2206} = 0.41$ .

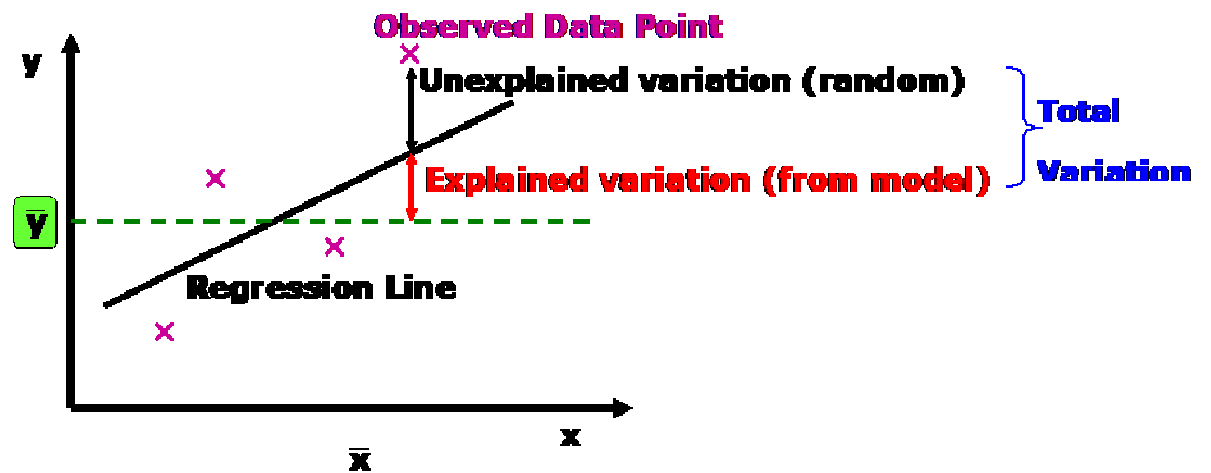
# Regression



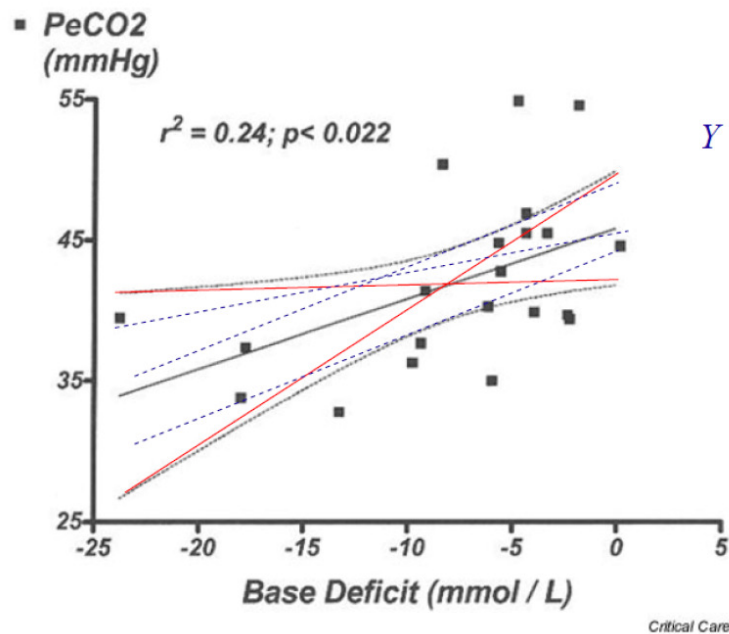
$$SS_{total} = SS_{model} + SS_{error}$$

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}} \quad \text{Coefficient of determination}$$



# Regression



$$Y = [40, 45] + [0.05, 0.45]X$$

We got a certain regression line but the true regression line lies within this region with a 95% confidence.

$$\hat{\alpha} = \bar{y} - (\hat{\beta} \bar{x}),$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\alpha \in [\hat{\alpha} - s_{\hat{\alpha}} t_{n-2}^*, \hat{\alpha} + s_{\hat{\alpha}} t_{n-2}^*] \quad s_{\hat{\alpha}} = s_{\hat{\beta}} \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$$

$$\beta \in [\hat{\beta} - s_{\hat{\beta}} t_{n-2}^*, \hat{\beta} + s_{\hat{\beta}} t_{n-2}^*] \quad s_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\hat{\alpha} = 0.859, \quad \hat{\beta} = -1.817.$$

The 95% confidence intervals for these estimates are

$$\alpha \in [0.76, 0.96], \quad \beta \in [-2.06, -1.58].$$

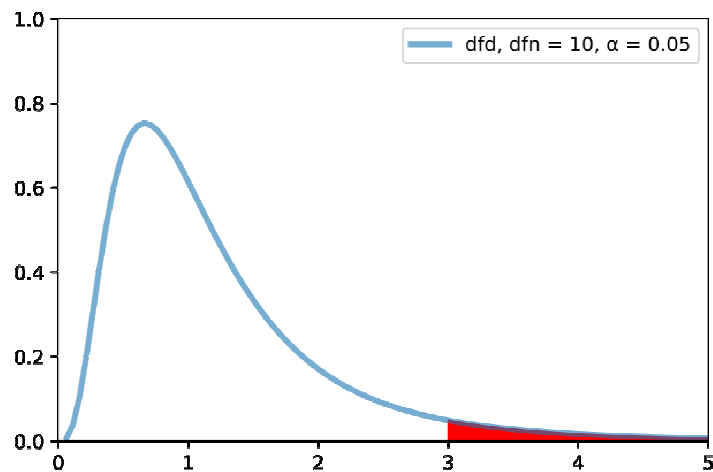
# Regression

## Overall test

$$H_0: y = \mu_y \neq f(x)$$

$$H_a: y = f(x)$$

$$F = \frac{\frac{SS_{model}}{k}}{\frac{SS_{error}}{N-k-1}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{N-k-1}}$$



## Nested models

$$1 - R_{adj}^2 = (1 - R^2) \frac{N - 1}{N - k - 1}$$

$$y = \beta_0$$

$$y = \beta_0 + \beta_1 x$$

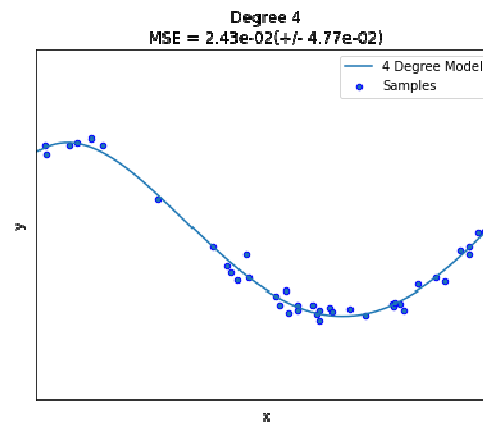
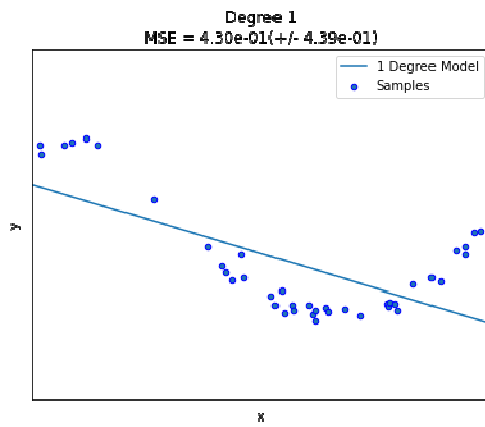
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

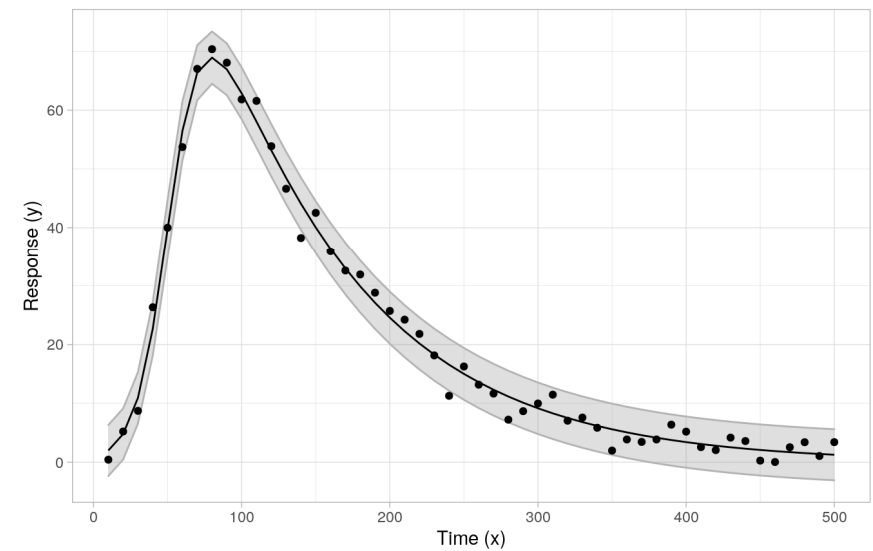


# Regression

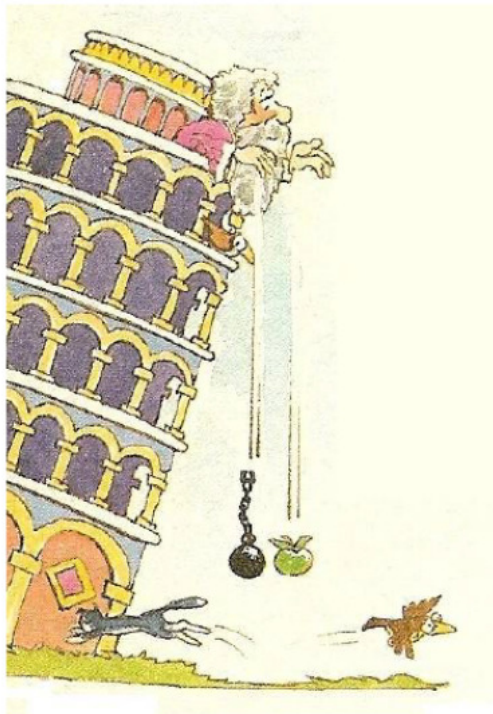
$$MSE = \frac{SS_{Total}}{N}$$



Approximate 95%-prediction intervals individual responses  
Infected compartment SIR model



# Regression models



We climb to a couple of towers (one with a height of 30 meters and another one with 60 meters), let a ball fall 10 times and measure the time it takes to reach the floor. We know there is an error in our time measurement that is assumed to be normal, with zero mean, and a standard deviation of 0.2s.

Which of the following regression models are valid?



$$h(t) = a_0 + a_1 t + a_1 t^2 + \varepsilon$$

$$h(t) = a_0 + a_{\frac{1}{2}} \sqrt{t} + a_1 t + a_1 t^2 + \varepsilon$$

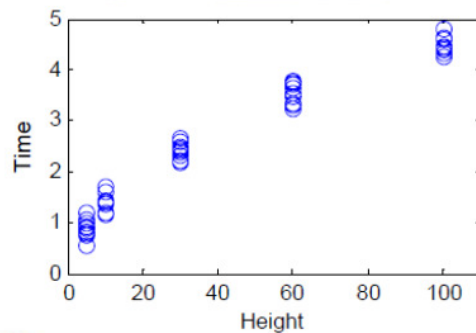
$$h(t) = a_0 + a_{\frac{1}{2}} \sqrt{t} + a_1 t + a_2 t^2 + \varepsilon$$

$$t(h) = a_0 + a_1 h + a_2 h^2 + \varepsilon$$

$$t(h) = a_0 + a_{\frac{1}{2}} \sqrt{h} + a_1 h + a_1 h^2 + \varepsilon$$

$$t(h) = a_0 + a_{\frac{1}{2}} \sqrt{h} + a_1 h + a_2 h^2 + \varepsilon$$

# Regression models



We climb to a few towers (with heights of 5, 10, 20, 30, 60 and 100 meters), let a ball fall 15 times and measure the time it takes to reach the floor. We know there is an error in our time measurement that is assumed to be normal, with zero mean, and a standard deviation of 0.2s.

$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + a_1h + a_2h^2 + \varepsilon$$

$$t(h) = -0.33 + 0.62\sqrt{h} + 0.02h + 0h^2 + \varepsilon \quad R^2 = 0.9773$$

$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + a_1h + \varepsilon$$

$$t(h) = -0.15 + 0.51\sqrt{h} + 0h + \varepsilon \quad R^2 = 0.9772$$

$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + \varepsilon$$

$$t(h) = 0 + 0.45\sqrt{h} + \varepsilon \quad R^2 = 0.9766$$

$$t(h) = a_{\frac{1}{2}}\sqrt{h} + \varepsilon$$

$$t(h) = 0.45\sqrt{h} + \varepsilon$$

← This is the true model!!!

$$R^2 = 0.9766$$

# Regression models



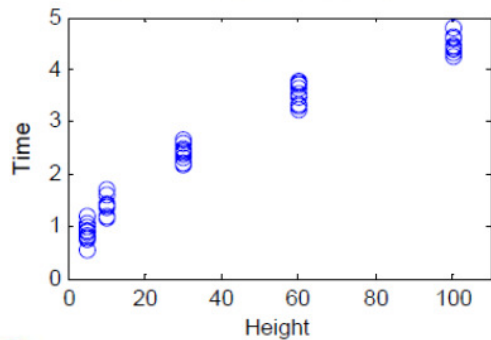
We climb to a few towers (with heights of 5, 10, 20, 30, 60 and 100 meters), let a ball fall 15 times and measure the time it takes to reach the floor. We know there is an error in our time measurement that is assumed to be normal, with zero mean, and a standard deviation of 0.2s.

$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + a_1h + a_2h^2 + \varepsilon$$

$$t(h) = -0.33 + 0.62\sqrt{h} + 0.02h + 0h^2 + \varepsilon \quad R^2 = 0.9773$$

$$t(h) = \cancel{[-0.90, 0.23]} + [0.30, 0.93]\sqrt{h}$$

$$+ \cancel{[-0.06, 0.02]}h + \cancel{[-0.00, 0.00]}h^2 + \varepsilon$$



# Regression



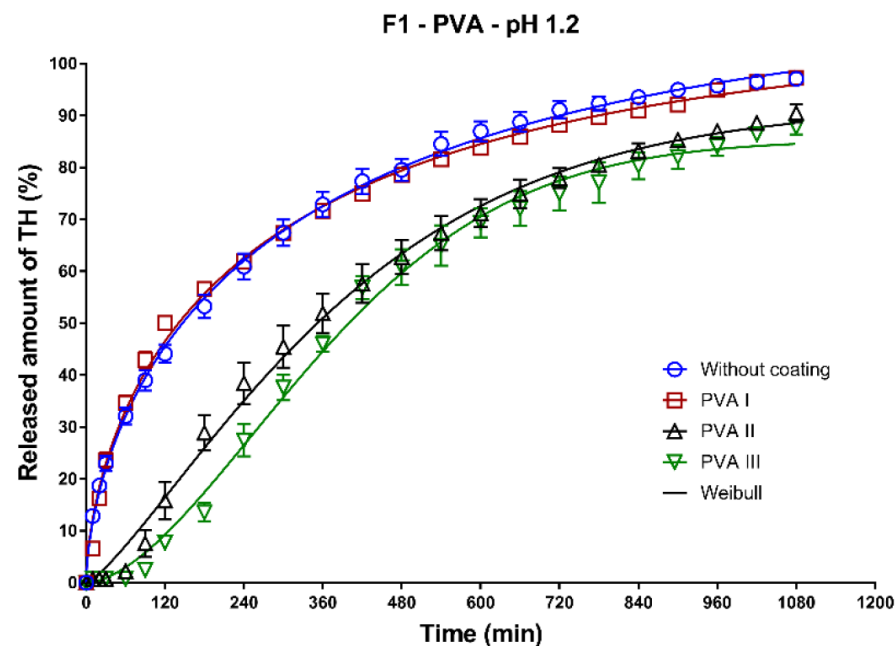
Article

## 3D-Printed Coating of Extended-Release Matrix Tablets: Effective Tool for Prevention of Alcohol-Induced Dose Dumping Effect

Barbora Skalická <sup>1</sup>, Kevin Matzick <sup>1,\*</sup>, Alena Komersová <sup>1</sup>, Roman Svoboda <sup>1</sup>, Martin Bartoš <sup>2</sup> and Luděk Hromádko <sup>3</sup>

$$M_{t(l)} = M_{\infty}(1 - \exp(-k_1 t))$$

Weibull 
$$M_{t(l)} = M_{\infty} \left( 1 - \exp(-k_w t^{\beta}) \right)$$



Weibull Model					
	$(k_w \pm SD) \times 10^3$ (min <sup>-β</sup> )	$A_{\infty} \pm SD$ (%)	$\beta \pm SD$	ASS	$R^2$
Without coating	25.57 ± 1.57	113.6 ± 2.89	0.63 ± 0.01	112	0.9974
PVA I	29.06 ± 2.13	107.8 ± 2.84	0.62 ± 0.02	153	0.9962
PVA II	0.42 ± 0.12	91.9 ± 1.98	1.29 ± 0.05	311	0.9940
PVA III	0.04 ± 0.01	85.3 ± 1.25	1.69 ± 0.07	313	0.9940

# Regression

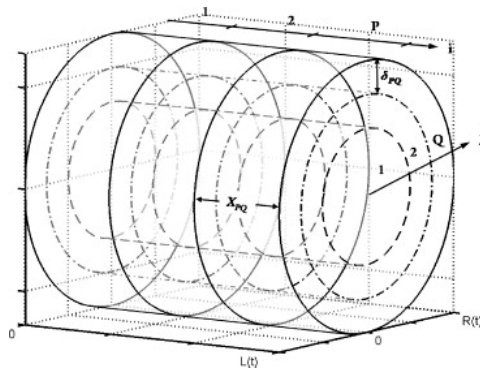


Journal of Controlled Release  
Volume 113, Issue 3, 20 July 2006, Pages 216-225



## A model for the drug release from a polymer matrix tablet—effects of swelling and dissolution

Per Borgquist<sup>a</sup>, Anna Körner<sup>b</sup>, Lennart Piculell<sup>b</sup>, Anette Larsson<sup>c</sup>, Anders Axelsson<sup>a</sup>



### 3.1. Polymer dissolution

The matrix dissolution of the drug-loaded polymer tablet is modelled in the same manner as in the previously presented pure polymer case [27]. The dissolution of a boundary volume (a volume in contact with the bulk phase) can be expressed by using a dissolution coefficient,  $k_d^p$ :

$$\frac{dm_{ij}^p}{dt} = -k_d^p A_{ij}^B \rho^p y_d^p, \quad (3)$$

where  $A_{ij}^B$  is the area available for dissolution (axial and/or radial) and  $y_d^p$  is the polymer volume fraction at the boundary.

### 3.2. Water mass balance

The unsteady-state water mass balance over a finite volume ( $i, j$ ), taking into account diffusion and convection fluxes and polymer dissolution can be written [27]:

$$\begin{aligned} \frac{V_{ij}}{y_{ij}^p} \frac{dy_{ij}^w}{dt} - A_{ij}^{es,out} y_{i,j+1}^w \sum_{k=1}^j \frac{d\delta_{ik}}{dt} + A_{ij}^{es,in} y_{ij}^w \sum_{k=1}^{j-1} \frac{d\delta_{ik}}{dt} - A_{ij}^{cs} y_{i+1,j}^w \sum_{k=1}^i \frac{dX_{kj}}{dt} \\ + A_{ij}^{cs} y_{ij}^w \sum_{k=1}^{i-1} \frac{dX_{kj}}{dt} = -y_{ij}^w A_{ij}^B \Phi_{ij} + A_{ij}^{es,out} N_{i,j+1 \rightarrow j}^{Dw} - A_{ij}^{es,in} N_{i,j \rightarrow j-1}^{Dw} + A_{ij}^{cs} N_{i+1 \rightarrow i,j}^{Dw} \\ - A_{ij}^{cs} N_{i \rightarrow i-1,j}^{Dw}, \end{aligned} \quad (4)$$

where the contribution due to dissolution,  $\Phi$ , is:

$$\Phi_{ij} = 0, \quad (5)$$

# Regression

## Rate laws: Michaelis-Menten

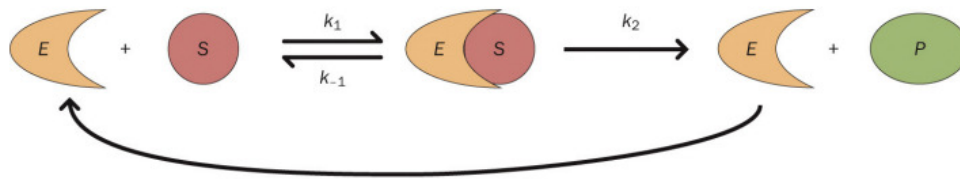


Figure 8.2 A First Course in Systems Biology 2e (© Garland Science 2018)

$$\dot{S} = -k_1SE + k_{-1}(ES)$$

$$(\dot{ES}) = k_1SE - (k_{-1} + k_2)(ES)$$

$$\dot{P} = k_2(ES)$$

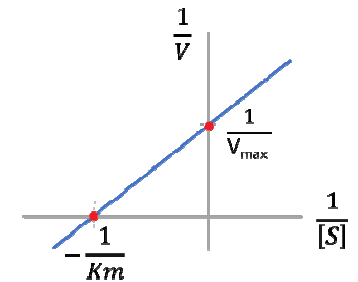
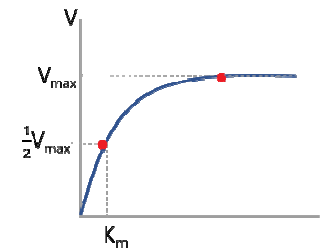
Quasi-Steady-State Assumption (QSSA)  $S \gg E$   $(\dot{ES}) = 0 \Rightarrow \frac{SE}{(ES)} = K_M = \frac{k_{-1} + k_2}{k_1}$

ES dissociation (at equilibrium)  $\frac{SE}{(ES)} = k_d = \frac{k_{-1}}{k_1} \approx K_M$

E occupancy  $\frac{(ES)}{E_{tot}} = \frac{(ES)}{E + (ES)} = \frac{SE/K_M}{E + SE/K_M} = \frac{S}{K_M + S}$

$$\dot{P} \equiv v_P = k_2 E_{tot} \frac{S}{K_M + S} = \frac{V_{Max} S}{K_M + S}$$

## 반응속도론



$$V = \frac{v_{max}[S]}{K_m + [S]} = \frac{k_{cat}[E_T][S]}{k_m + [S]}$$

$$\frac{1}{V} = \frac{K_m}{V_{max}} \cdot \frac{1}{[S]} + \frac{1}{V_{max}}$$



## Logistic regression

We try to predict a binary variable (0 or 1) from other binary or continuous variables.

$$Obese = f(Residence, Age, Education, Smoking, Married, LowIncome)$$

- Residence: binary (0=rural, 1=urban)
- Age: continuous (years)
- Education: continuous (years)
- Smoking: binary (0=No, 1=Yes)
- Married: binary (0=No, 1=Yes)
- LowIncome: binary (0=No, 1=Yes)

We will rather predict the probability of obese taking the value 1.



## Logistic regression

Remind the relationship between probability and odds (ratio of the probability of something happening vs. not happening)

$$OR = \frac{p}{1-p} = \text{logit}(p)$$

We will transform the problem into

$$OR_{Obese} = OR_0 OR_{Residence} OR_{Age} OR_{Education} OR_{Smoking} OR_{Married} OR_{LowIncome}$$

Taking logarithms

$$\text{logit}(p_{Obese}) = \beta_0 + \beta_{Residence} Residence + \beta_{Age} Age + \beta_{Education} Education + \dots$$

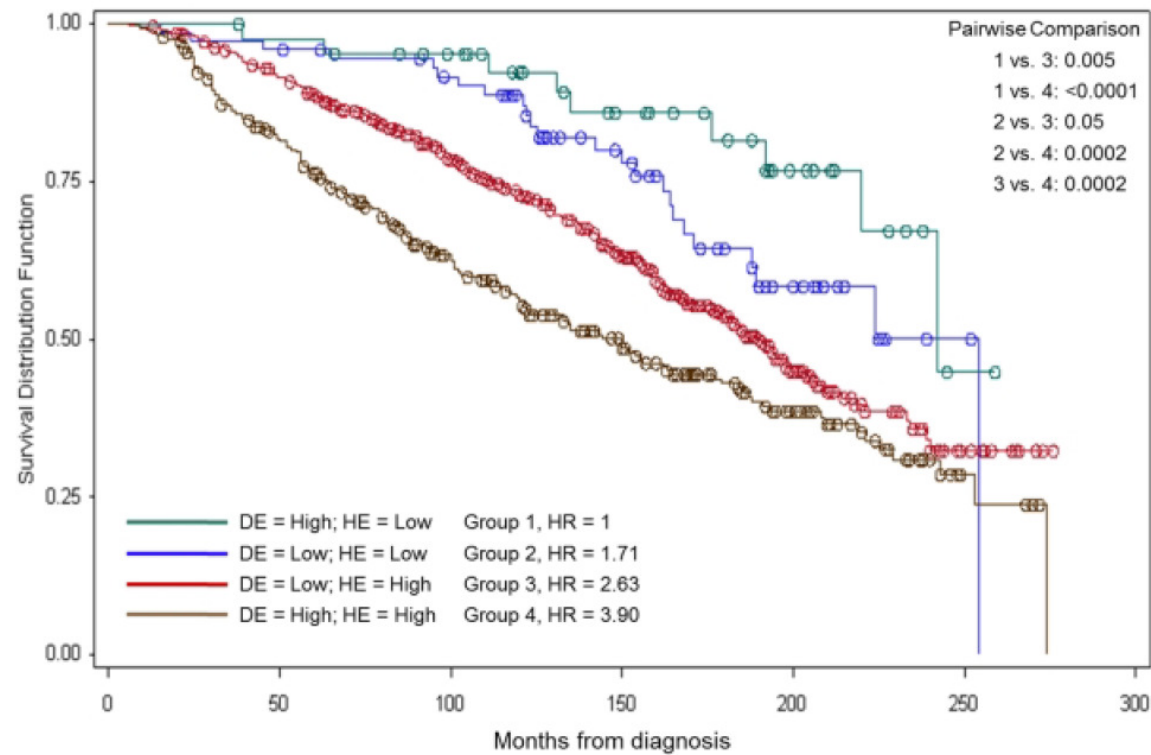
$p_{Obese}$  is the probability of being obese.

## Logistic regression

We may **interpret** the  $\beta$ s in the **standard way** (if the CI includes 0, then that term is not significant) or **in terms of OR**. Example:

- Residence:  $\beta_{Residence} = 0.3218 \Rightarrow \exp(0.3218) = 2.13$ , that is a person living in a urban environment has 2.13 times the odds of being obese than someone living in a rural environment.
- Age:  $\beta_{Age} = 0.0086 \Rightarrow \exp(0.0086) = 1.02$ , for every year, there is an odds ratio increase by a factor 1.02.

## Proportional hazards (Cox) regression



## Proportional hazards (Cox) regression

Remember that the **hazard** is related to the slope of the survival curve ( $\lambda(t) = -\frac{S'(t)}{S(t)}$ ). The proportional hazards model proposes

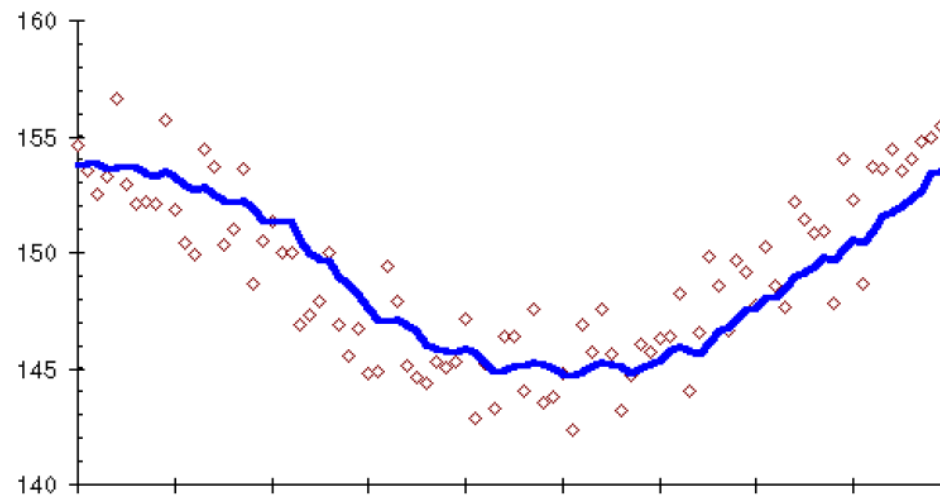
$$\lambda = \exp(\beta_0 + \beta_{HE}HE + \beta_{DE}DE)$$

Taking logarithms

$$\log(\lambda) = \beta_0 + \beta_{HE}HE + \beta_{DE}DE$$

## Common mistakes

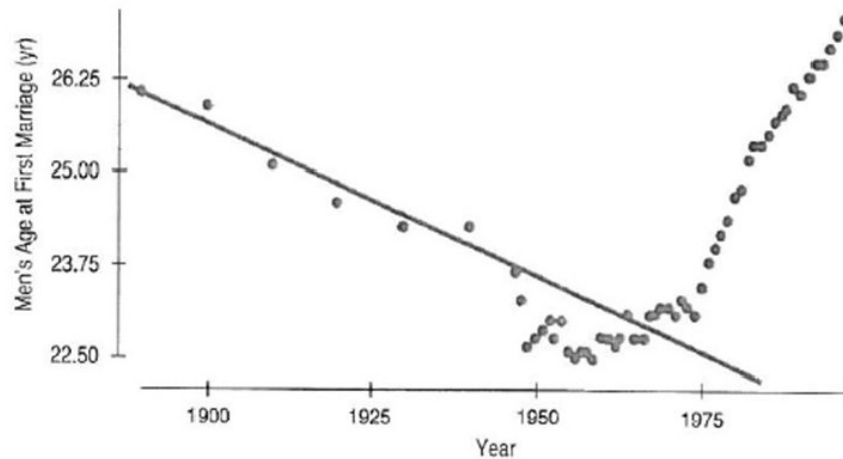
- Fitting smoothed/moving average data.
  - Smoothing the data artificially increases the  $R^2$  and reduces the p-value.
  - Smoothing can artificially create trends where there is no relationship.
  - Smoothing violates the assumption of data independence.



## Common mistakes

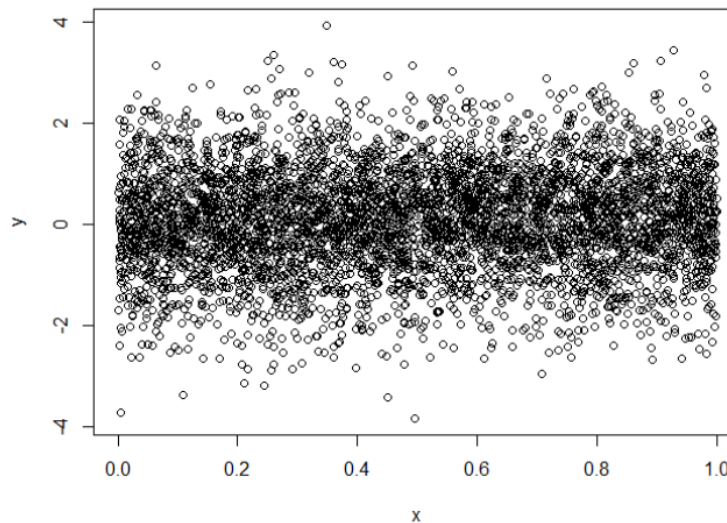
- **Extrapolating beyond the data.** Models are valid only within the range of observed  $X$  values. Extrapolation beyond this range is at the user's own risk.

### Extrapolation



## Common mistakes

- **Overinterpreting a small p-value.** A small p-value indicates that the model fits the data better than a constant. However, this is not enough to be a good model. A linear model ( $y = a + bx$ ) of the data in the figure below has a p-value of 0.000105 (very significant), but  $R^2 = 0.003005$ , that is, the model does not explain even 0.5% of the observed variance.



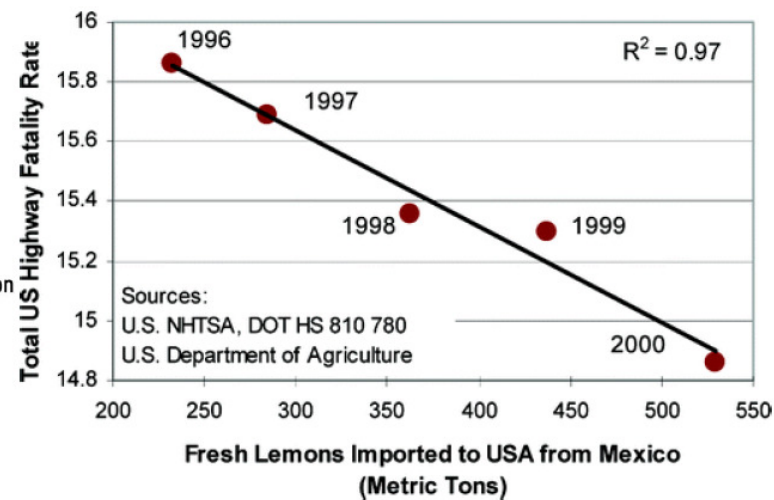
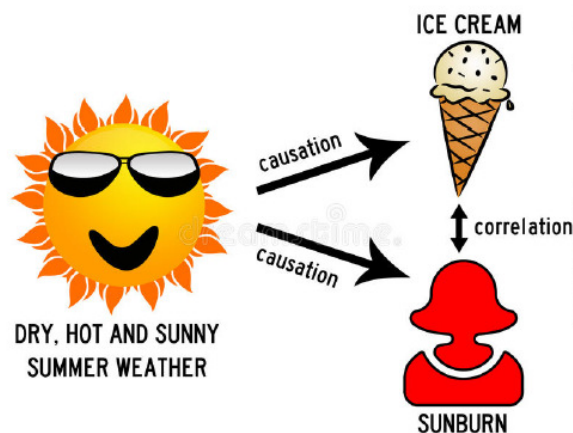
## Regression does not imply causation

Assume we perform an experiment and discover that there is a relationship between lead concentration in blood and kidney function (measured by creatinine clearance).

$$CrCl = 101[mL/min] - 9.51 \log C_{Pb}[\mu g/L]$$

Can we assess that lead exposure causes kidney malfunctioning?

No, it could be the opposite. Kidney malfunctioning causes lead raise in blood.



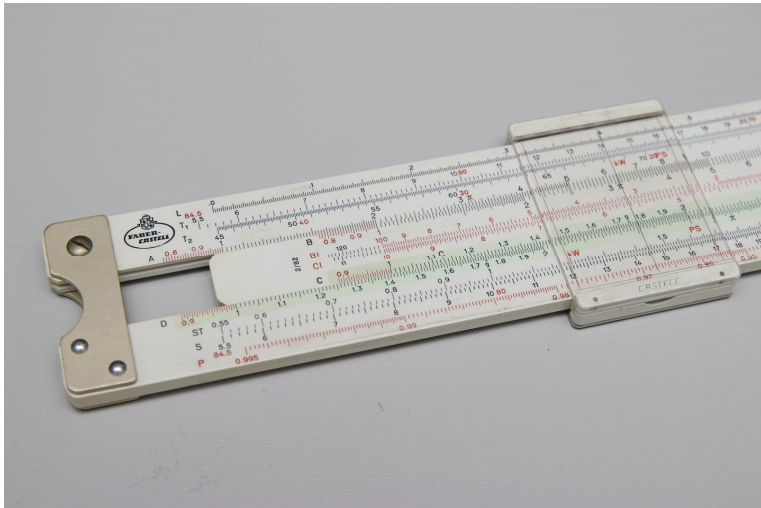




# Conclusions

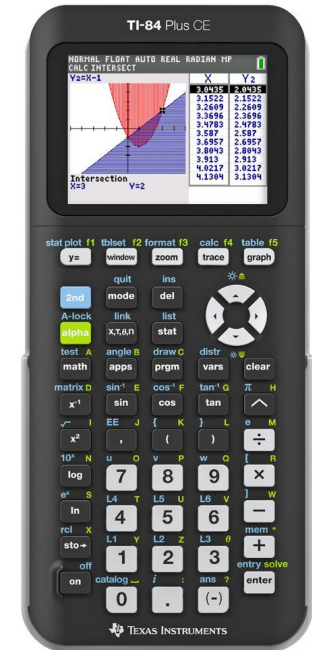
# Conclusions

- If you think that education is expensive, try ignorance.
- If you think that using Statistics is difficult, try not to use it.

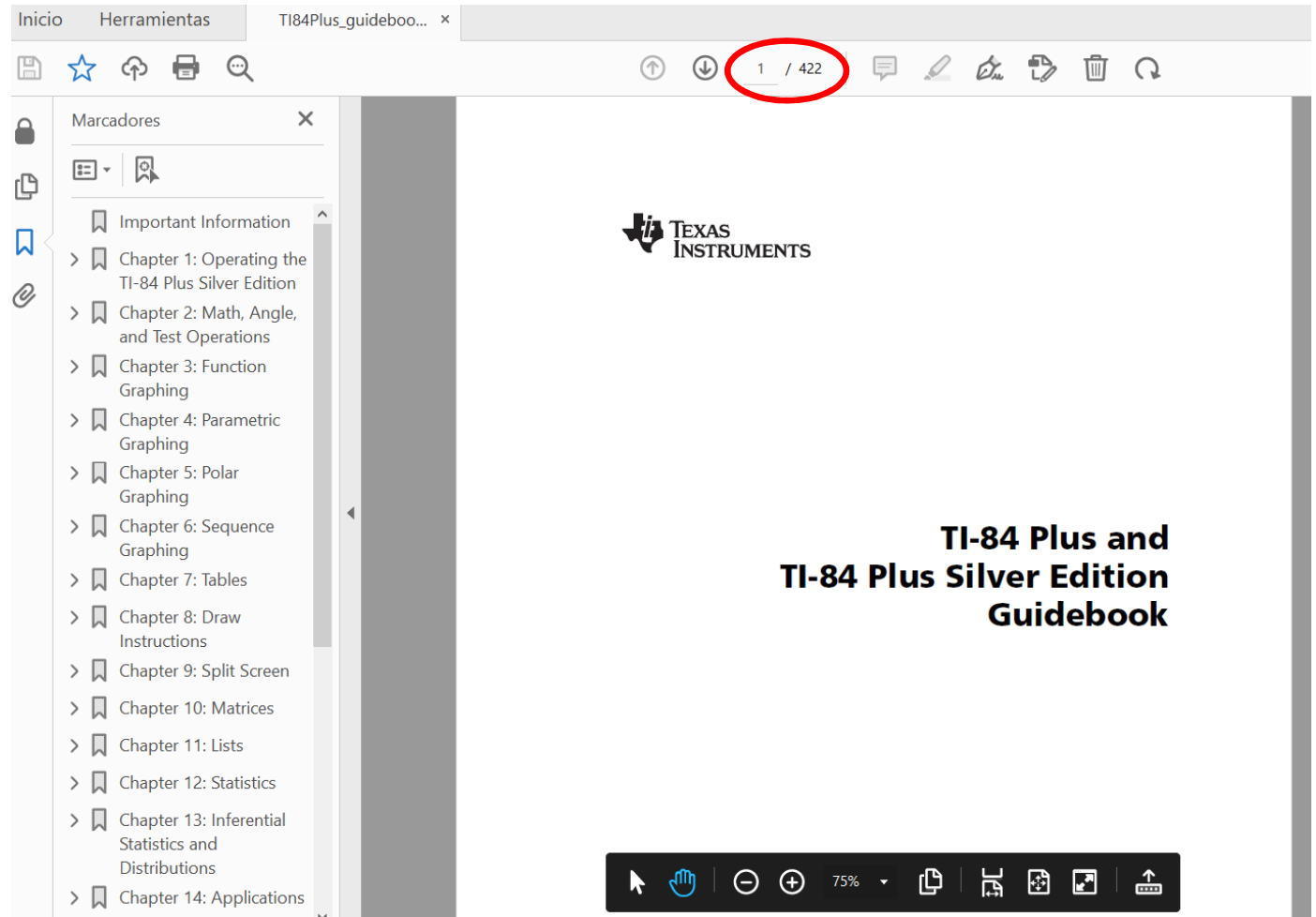
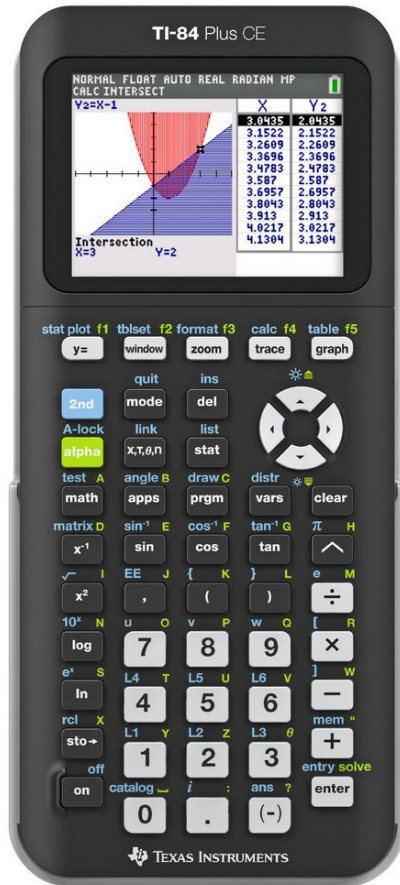


$$10^{2.8699} = 741.1$$

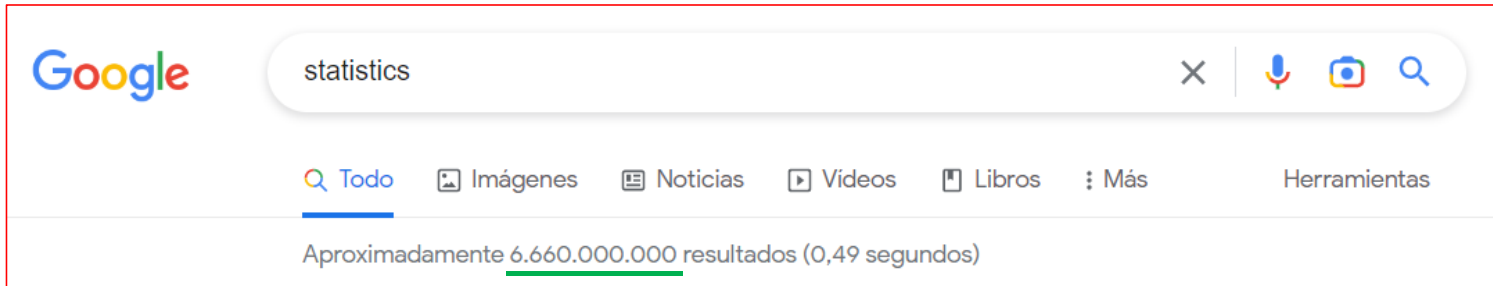
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	2	3	4	5	6	7	8	9
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	4	5	6	7	8	9	10
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	3	4	5	6	7	8	9	10	11
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6903	4	5	6	7	8	9	10	11	12
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7064	5	6	7	8	9	10	11	12	13
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	6	7	8	9	10	11	12	13	14
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	7	8	9	10	11	12	13	14	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	8	9	10	11	12	13	14	15	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	9	10	11	12	13	14	15	16	17
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	10	11	12	13	14	15	16	17	18



# Conclusions

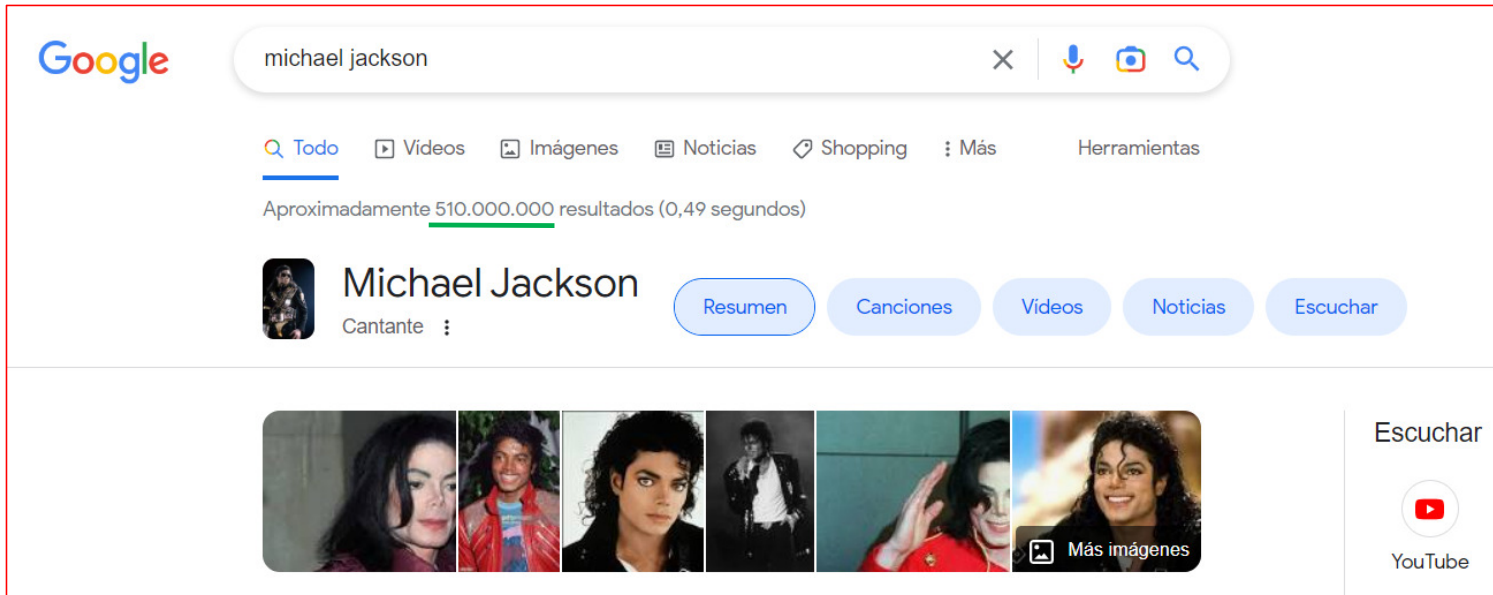


# Conclusions



Google search results for the query "statistics". The search bar shows the query and icons for voice search, image search, and a magnifying glass. Below the search bar, navigation links include "Todo", "Imágenes", "Noticias", "Videos", "Libros", "Más", and "Herramientas". The results section indicates "Aproximadamente 6.660.000.000 resultados (0,49 segundos)".

x13



Google search results for the query "michael jackson". The search bar shows the query and icons for voice search, image search, and a magnifying glass. Below the search bar, navigation links include "Todo", "Videos", "Imágenes", "Noticias", "Shopping", "Más", and "Herramientas". The results section indicates "Aproximadamente 510.000.000 resultados (0,49 segundos)".

**Michael Jackson**  
Cantante

Buttons: Resumen, Canciones, Videos, Noticias, Escuchar

Image carousel showing various photos of Michael Jackson. A button labeled "Más imágenes" is visible.

Escuchar  
YouTube

# Questions



# Bibliography

- C.O.S. Sorzano. [Statistical experiment design for animal research](https://osf.io/e9s25). OSF Preprints: osf.io/e9s25 (2023)
- <http://i2pc.es/coss/Articulos/Sorzano2023.pdf>

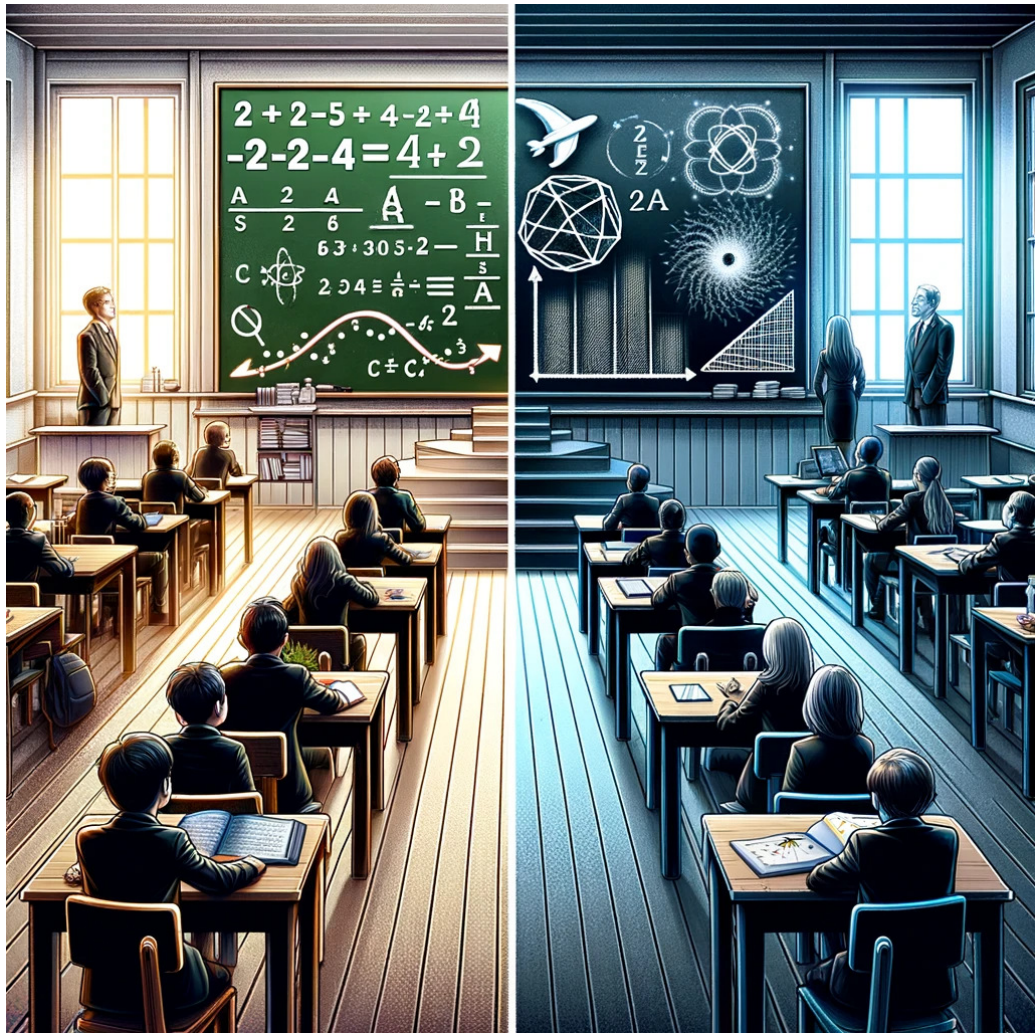
<b>1</b>	<b>Why do we need a statistical experiment design?</b>	<b>7</b>
1.1	Pilot, exploratory and confirmatory experiments . . . . .	14
1.2	Independence between individuals: experimental units . . . . .	17
1.3	Avoiding bias: blocking, randomization and blinding . . . . .	20
1.4	Reducing variance: variable and population selection, experimental conditions, averaging, and blocking . . . . .	27
1.4.1	Variable selection . . . . .	28
1.4.2	Population selection . . . . .	28
1.4.3	Experimental conditions . . . . .	29
1.4.4	Population scope, outliers and lack of independence . . . . .	30
1.4.5	Averaging and pooling . . . . .	34
1.4.6	Blocking . . . . .	38
1.4.7	Paired samples . . . . .	41
1.4.8	Blocking and randomization . . . . .	42
1.5	Automating decision making: hypothesis testing . . . . .	43
1.5.1	An intuitive introduction to hypothesis testing . . . . .	46
1.5.2	Statistical power and confidence . . . . .	49
1.5.3	Multiple testing . . . . .	53
1.5.4	A worked example . . . . .	55
1.6	A primer in sample size calculations . . . . .	60
<b>2</b>	<b>Sample size calculations</b>	<b>69</b>
2.1	Sample size for the mean . . . . .	70
2.1.1	Hypothesis test on the mean of one sample when the variance is known . . . . .	70

## Bibliography

- C.O.S. Sorzano. Statistical design for animal research (2023).
- M. Festing, P. Overend, M. Cortina, M. Berdoy. The design of animal experiments (2016).
- S.T. Bate, R.A. Clark. The Design and Statistical Analysis of Animal Experiments (2014).
- Sample size: <https://www.youtube.com/playlist?list=PLQjWlcrmtc4KtxXMj4byAsZlF9gwWgTXH>
- Experiment design: [https://www.youtube.com/playlist?list=PLQjWlcrmtc4LMu47i\\_elxGYPLEwdrQ6ey](https://www.youtube.com/playlist?list=PLQjWlcrmtc4LMu47i_elxGYPLEwdrQ6ey)
- General statistics: <https://www.youtube.com/playlist?list=PLQjWlcrmtc4JUvzoJvloLA9wXWkgYuS9I>



Course





# Course

