Chapter 0. Statistics and probability

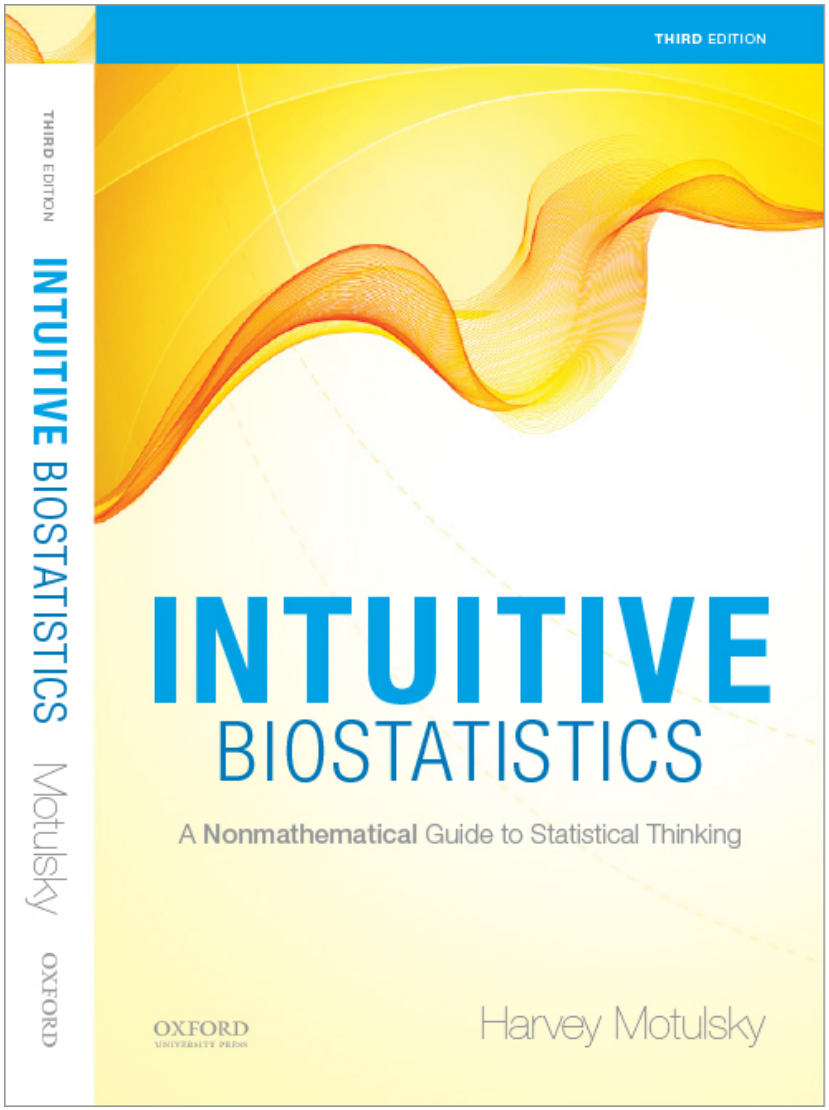
C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

August 13, 2017

Outline

References



Harvey Motulsky. Intuitive biostatistics. Oxford Univ. Press (2014)

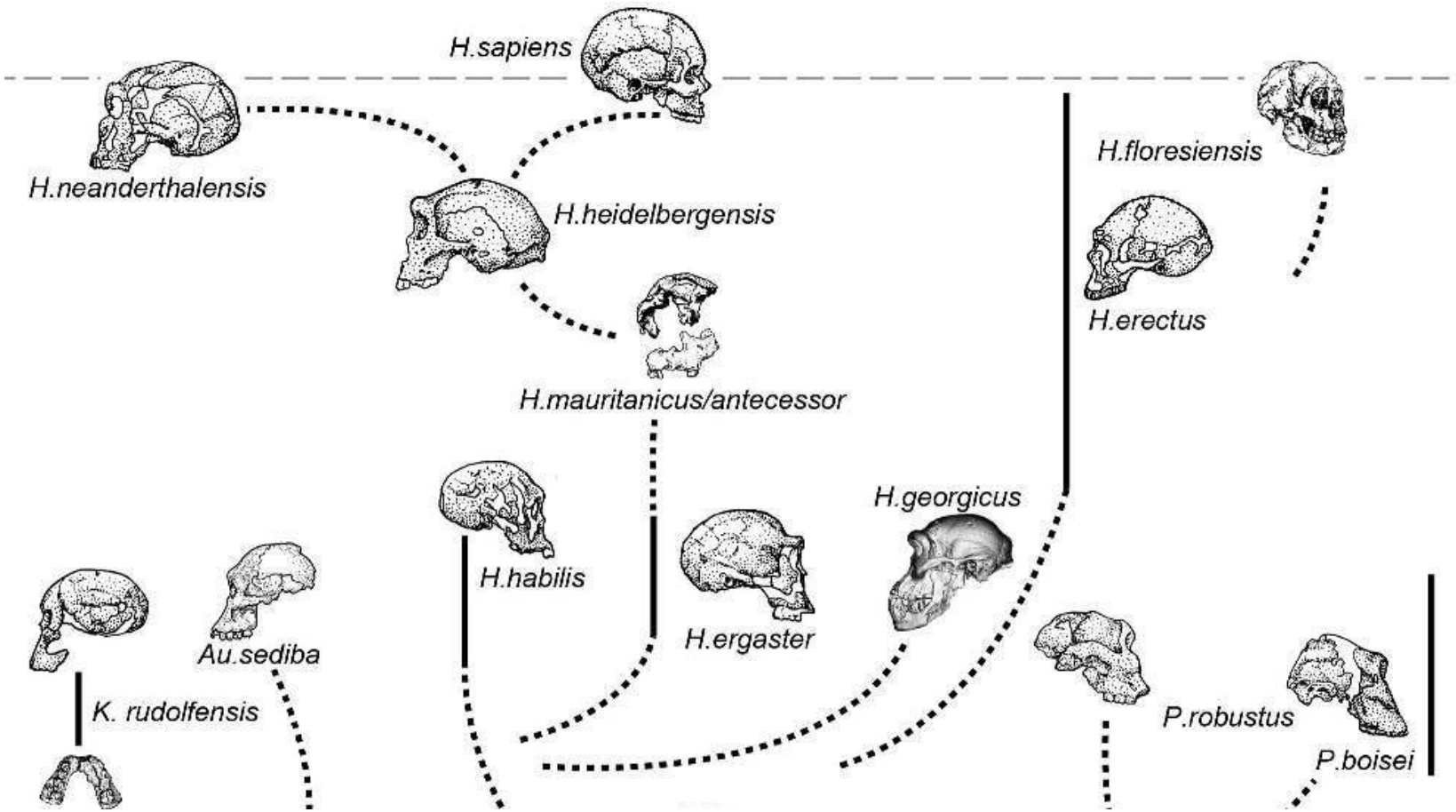
Outline

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 tend to jump to conclusions. A 4-years little child may think all doctors are female simply because the 4 doctors she has met are women.

From a small sample we cannot generalize to the whole population.

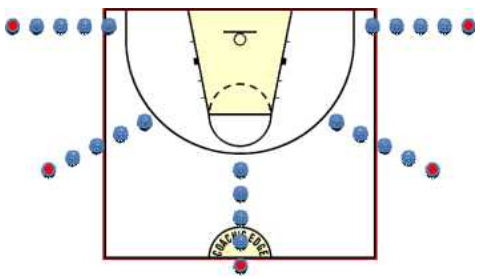
Statistics is not intuitive

We tend to be overconfident.

- Most people to have more common sense than the average person and drive better than the average.
- Most drugs tested do not help ... Of course, this applies to other people's experiments. My experiment has a large probability of succeeding (I carefully designed it).

Scientists need statistical methods to quantify confidence on the results.

Statistics is not intuitive



We see patterns in random data.

- We tend to see winning or losing streaks, but the true probability is 0.5 and all shots are independent.
  0 1 1 0 0 0 1 0 1 0 1 0 1 1 1 1 1 0 0 0 1 0 1 0 1

But this is mental bias. Statistical rigor is needed to avoid being fooled by apparent patterns.

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 had found any other characteristic amongst the top 5 (being girls, wearing jeans, coming to school by bus, ...).

Statistics is not intuitive

We find it hard to combine probabilities.

- Behind one door there is a fancy new car, and you must choose just one door. The host chooses one of the other two doors and shows that there is no car behind it. He offers you to change your choice to the remaining door. Should you change your mind and choose the other door? Most people think it does not matter, the door you chose either contains the car or not, so there is 50% of chances of getting the car.

Let's analyze the game. There are two possible situations:

- Case A: I originally chose the right door ($p = 1/3$). If I change, there is a loss of 1 car (-1)
- Case B: I originally chose the wrong door ($p = 2/3$). If I change, there is a gain of 1 car ($+1$)

On average

$$\mathrm{E}\{change\} = \frac{1}{3}(-1) + \frac{2}{3}(+1) = \frac{1}{3}$$

So changing the door increases my chances of winning the car.

Statistics is not intuitive

We find hard to match long run probabilities with single shots.

- Changing doors is beneficial only if we are to play this game a large (infinite) number of times. For a single shot, expectations do not help.

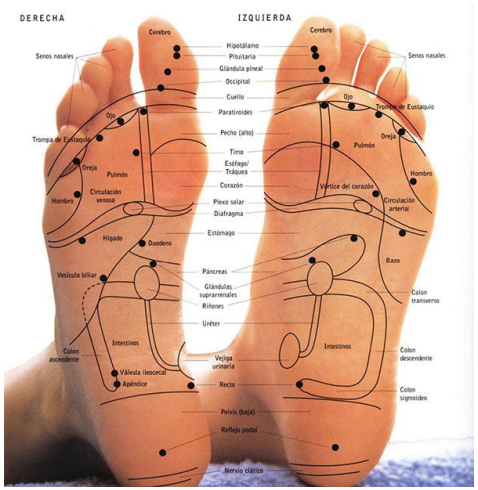
Statistics is not intuitive

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



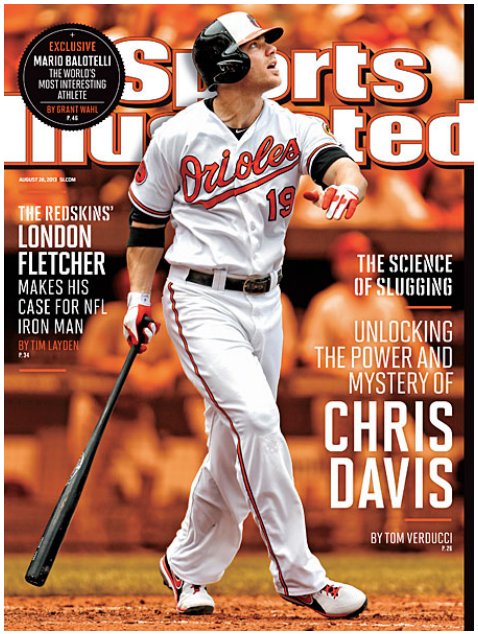
We tend to ignore alternative explanations.

- We are studying the effect of acupuncture on osteoarthritis. Patients with sever arthritis pain are treated with acupuncture. They are asked to rate their pain before and after treatment and there is a (statistically significant) decrease in the pain. So acupuncture must have worked, right?

But we ignore that:

- If the patients believe in the therapist and treatment, this belief may reduce pain (placebo effect).
- Patients may want to be polite and tell the experimenter what she wants to hear.
- During the acupuncture session, the therapist talks to the patients and he may recommend a change in the aspirin dose, exercise, nutritional supplements, ...
- The experimenter may remove data from the study, those for which acupuncture did not work because these patients have a different kind of arthritis, they had to climb stairs because the elevator was not working, ...
- Patients go to the therapist when they are feeling really bad, so they can only improve along the day.

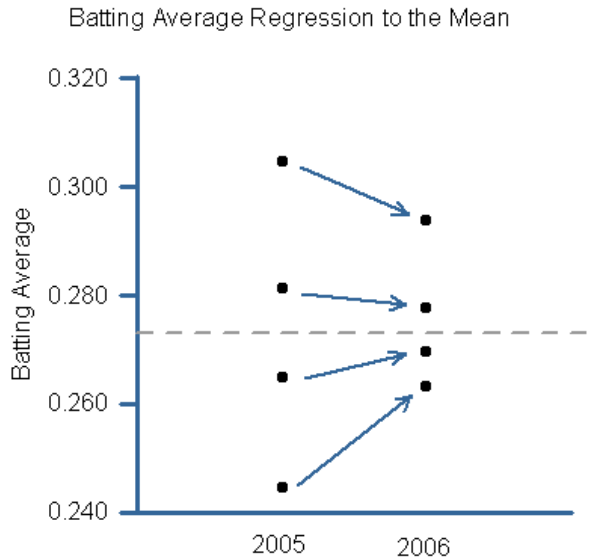
Statistics is not intuitive



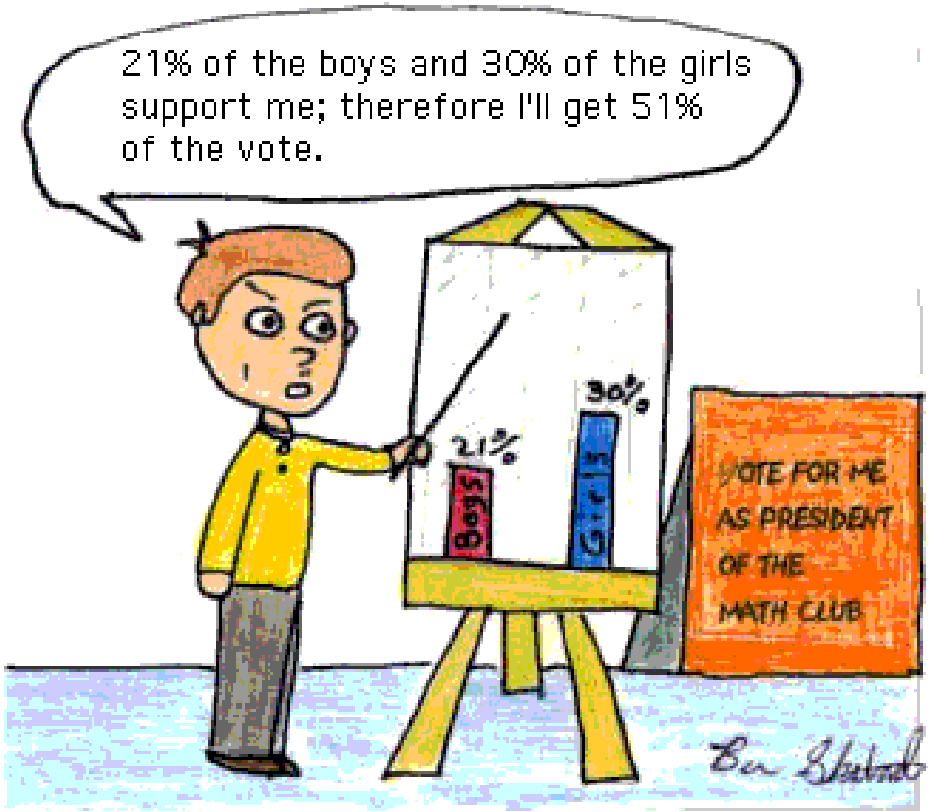
We are fooled by extreme values and regression to the mean.

- 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 athelete'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.

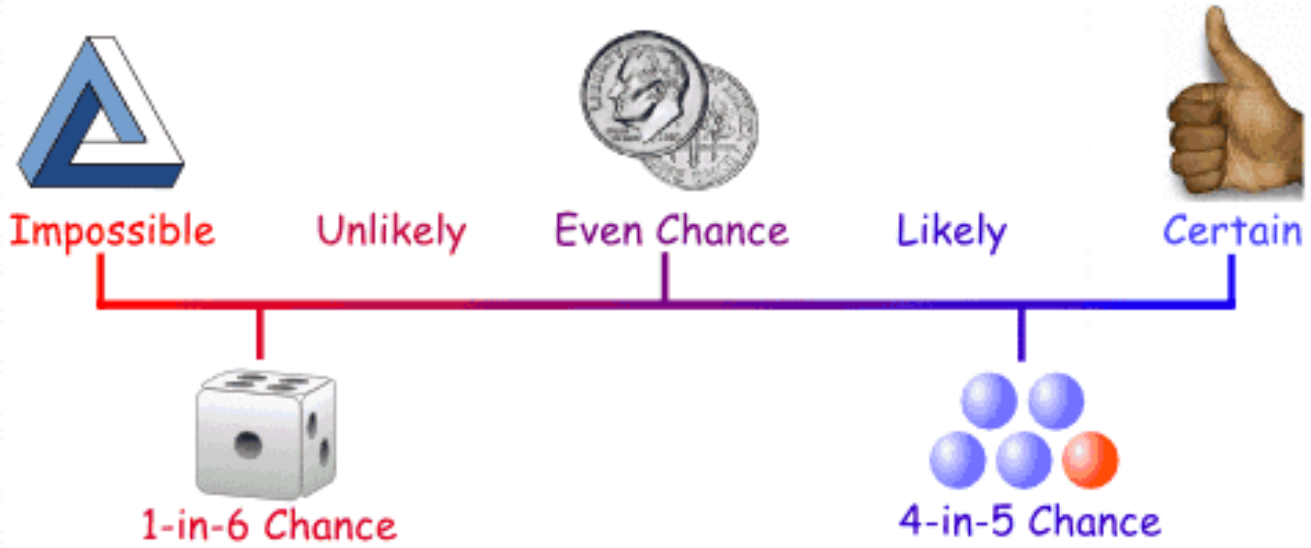


Statistics is not intuitive



Outline

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?)

Thomas Bayes

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

- 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?)

Thomas Bayes

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

- 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

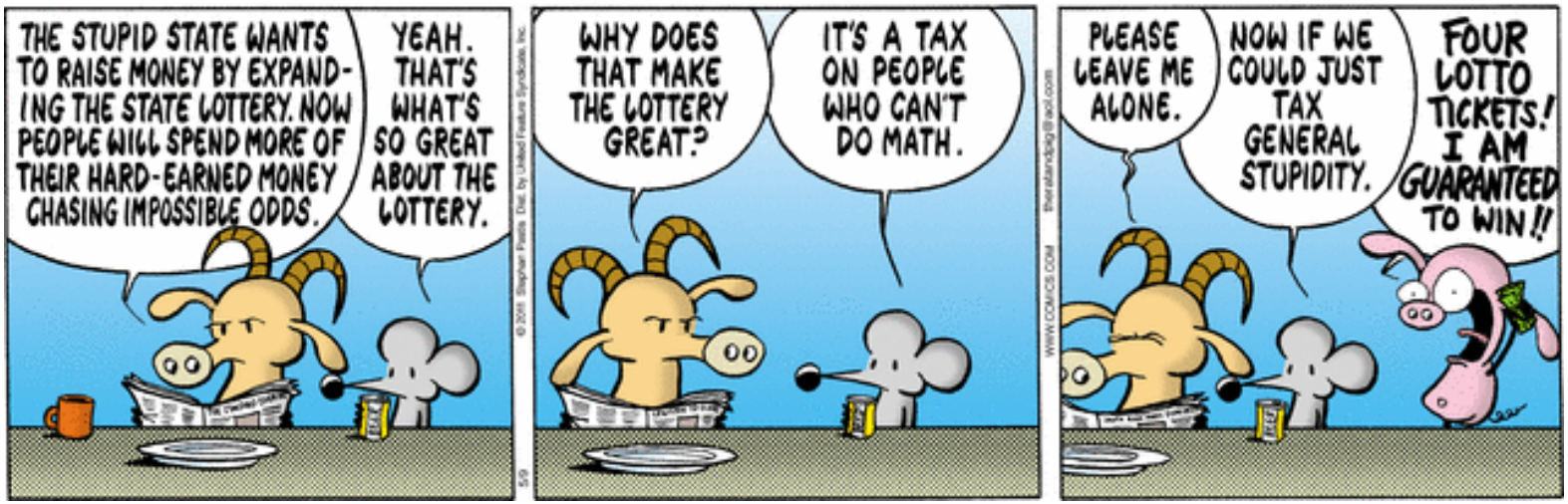
$$O = \frac{p(boy)}{p(girl)} = \frac{0.517}{0.483} = 1.07$$

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

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

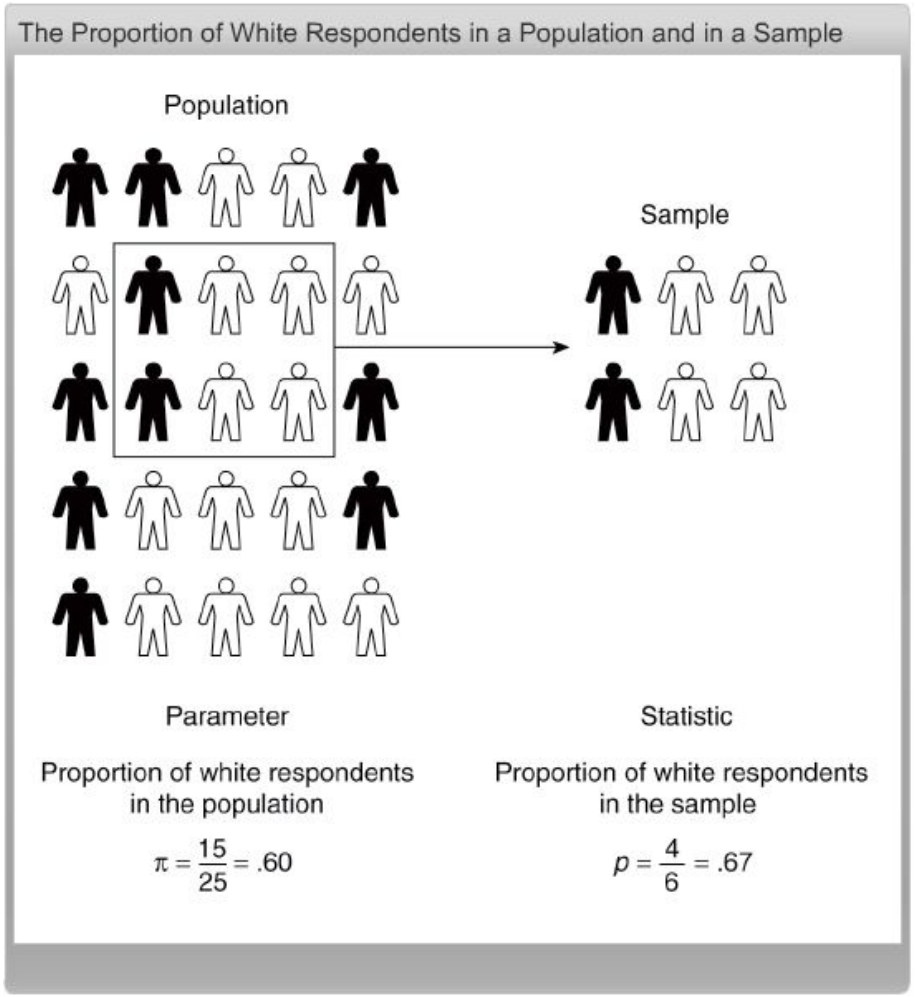
$$p(\text{lung cancer}|\text{smoke}) = 10p(\text{lung cancer}|\text{don't smoke})$$

Probability



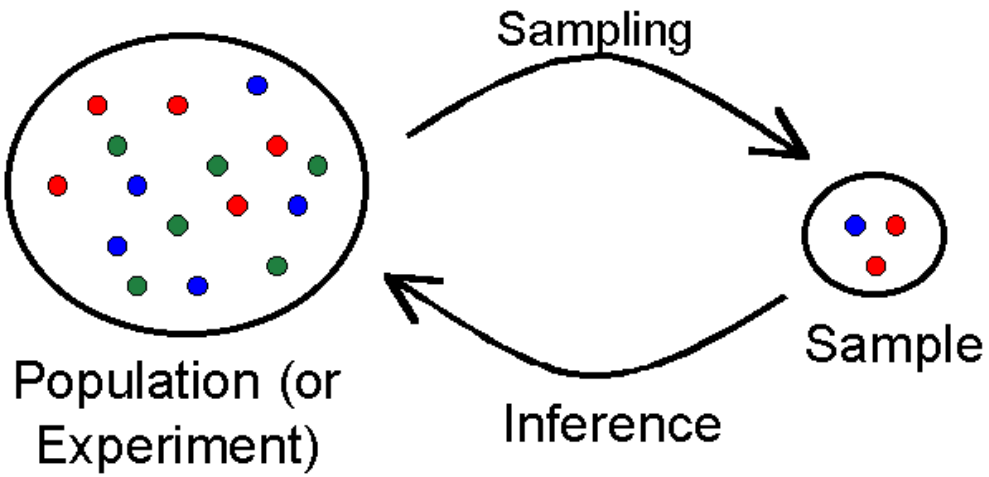
Outline

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.

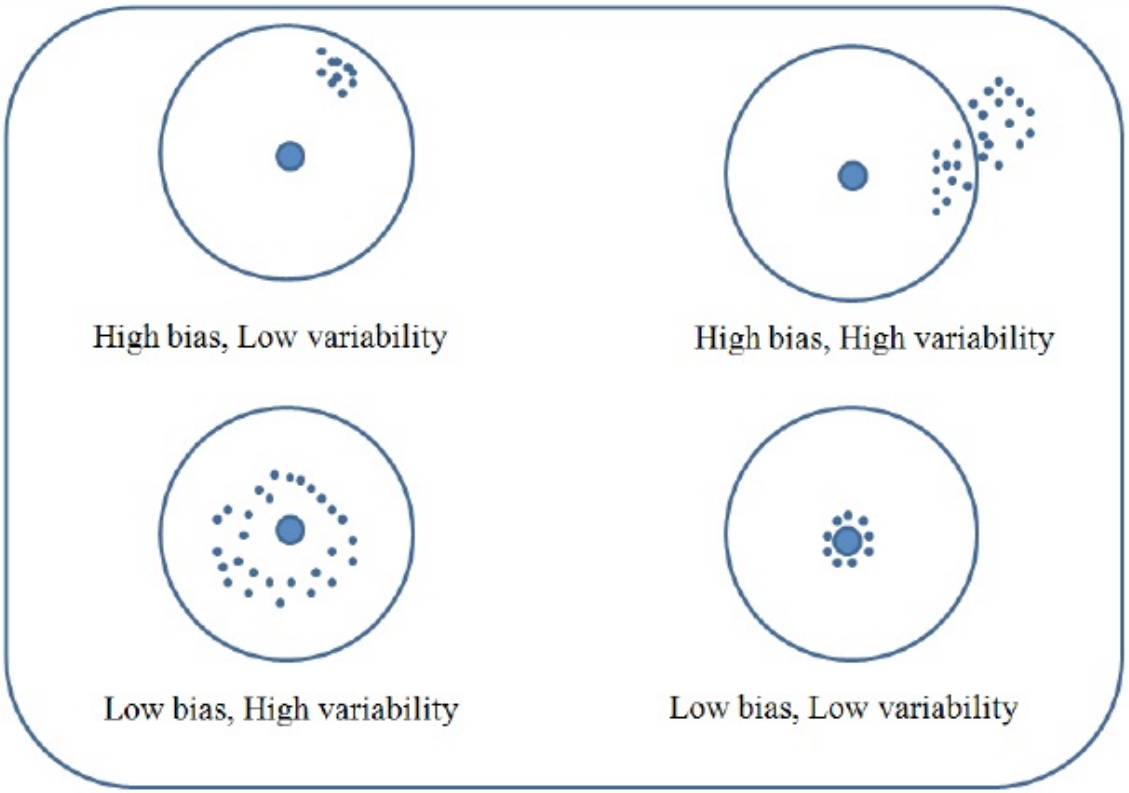
Random sampling error decreases with the sample size.

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

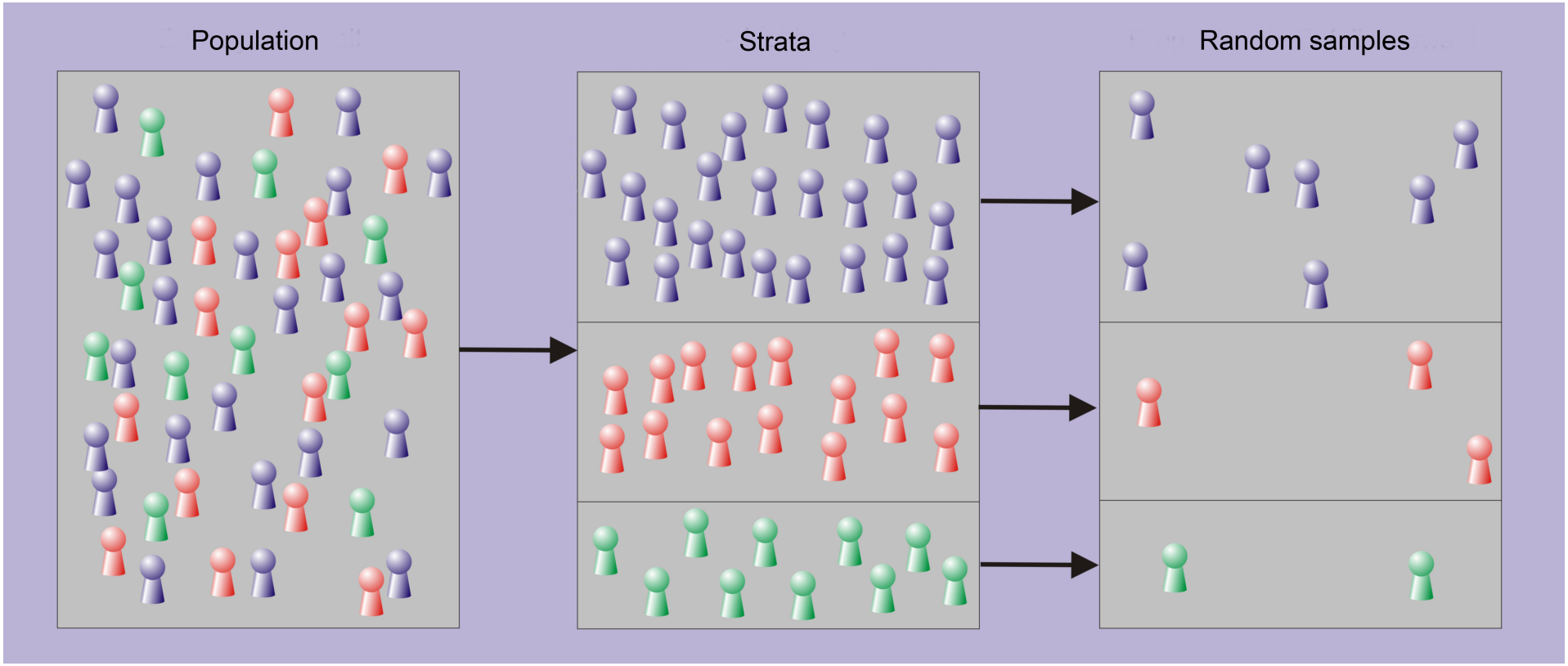
Figure 5. Bias and variability



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

Bias invalidate inference.

Stratified sampling



Stratified sampling helps undercoverage.

Random sampling

Outline

Chapter 1. Confidence Intervals

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

August 13, 2017

Outline

Outline

From population to sample ("simulation")

Binomial distribution. The probability of observing r independent successes out of N, each has a probability p.

- If you flip a fair coin ($p = 0.5$) 10 times ($= N$), what is the chance of observing exactly 7 heads ($= r$).
- If the probability of getting an infection after a surgical operation is 5%, what is the chance that 10 of the next 30 patients will get an infection?

Cumulative binomial distribution. The probability of observing r or more independent successes out of N, each has a probability p.

- What is the chance that 10 or more of the next 30 patients will get an infection?

Negative binomial distribution. The probability of observing q independent successes before observing r failures, each has a probability p.

- What is the chance of observing 30 non-infected patients before we observe the first infection?

From sample to (estimated) proportion ("experimentation")

Premature babies born at Johns Hopkins Hospital between 1990-1993:

- 0/29 babies born at Week 22 survived after 6 months (0/29=0%).
- 31/39 babies born at Week 25 survived after 6 months (31/39=79.5%).

Is the true survival proportion exactly 0 and 79.5%?
Probably, not.
Can we give a confidence interval that contains the true proportion with probability 95%
Yes [0,13.9]% and [64.3,89.5]%.
These intervals only account for sampling error (not for bias).

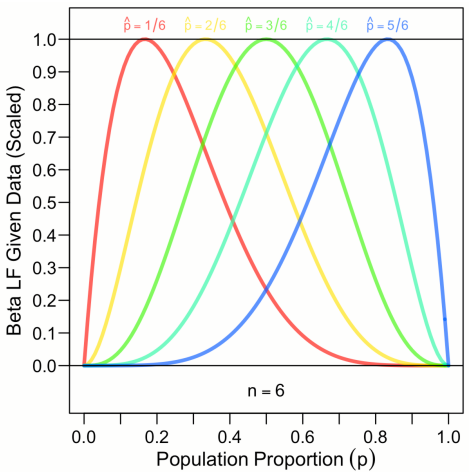
Assumptions

- Random (or representative) sample. 1) Other than chance there is no systematic difference between the newborns at Johns Hopkins Hospital and the general population newborns (that is, we assume there is no difference in nutrition of the mothers before giving birth, medical care to the newborns, hygienic state of hospital, ...). 2) This proportion is at some particular conditions (location, time, medical knowledge) and it can only be used to predict the outcome at the same conditions (a change in location (Africa), time (20 years later), ...) will most likely have different underlying parameters.

Assumptions

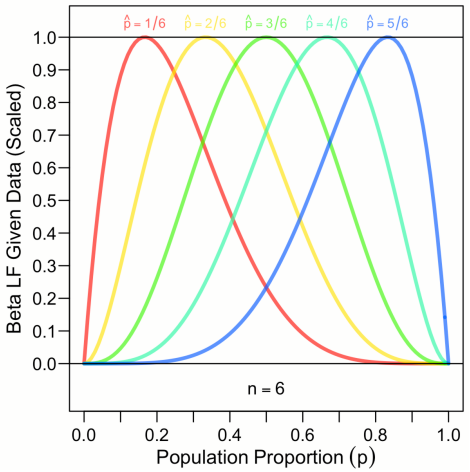
- Independent observations. Twins are not independent (they share genetic and environmental factors), or if deaths are caused by a hospital infection that affect some newborns.
- Accurate data. If the doctors know that 6-month survival is to be tracked, they may make heroic efforts to bring a 5-month old baby a few days more so that he accounts (even if he dies a few days after 6 months).

What is confidence?



- The true population proportion lies or lies not in the 95% CI. But there is no way to know if it does or not.
- If we repeat the experiment (calculating the CI) many (many) times, in 95% of the occasions, our CI contain the true population proportion (altough we don't know which ones).
- 95% is the probability that our CI contains the true proportion.
- If the true parameter is outside our CI, it is due to bad luck with our samples (sampling error). This occurs in 5% of the cases.
- There is nothing special about 95% (except tradition). Lower confidence results in narrower CI.
- Actually, the confidence is on our procedure to construct intervals, not about this particular interval.

What is confidence?



- 95% is not the probability that the true proportion is in our CI.
- A 95% CI does not mean that 95% of the sample data falls within this interval.
- A 95% CI does not mean that with probability 95% if we repeat the experiment, the estimated proportion falls within this interval.

How to calculate the confidence interval for the Binomial distribution?

- Clopper-Pearson exact formula

$$\frac{r}{r+(N-r+1)F_{1-\frac{\alpha}{2};2(N-r+1),2r}} \leq p \leq \frac{(r+1)F_{1-\frac{\alpha}{2};2(r+1),2(N-r)}}{(N-r)+(r+1)F_{1-\frac{\alpha}{2};2(r+1),2(N-r)}}$$

r is the number of observed success, N is the total number of samples. (F is Snedecor's F distribution)

- Approximated by the modified Wald formula

$$\begin{aligned} p' &= \frac{r+0.5z^2_{1-\frac{\alpha}{2}}}{N+z^2_{1-\frac{\alpha}{2}}} \overset{95\%}{\approx} \frac{r+2}{N+4} \\ W &= z_{1-\frac{\alpha}{2}}\sqrt{\frac{p'(1-p')}{N+z^2_{1-\frac{\alpha}{2}}}} \overset{95\%}{\approx} 2\sqrt{\frac{p'(1-p')}{N+4}} \end{aligned}$$

$$p' - W \leq p \leq p' + W$$

How to calculate the confidence interval for the Binomial distribution?

- More approximations

$$\hat{p} = \frac{0}{N} \overset{95\%}{\Rightarrow} \boxed{0 \leq p \leq \frac{3}{N}}$$

$$\hat{p} = \frac{1}{N} \overset{95\%}{\Rightarrow} \boxed{0 \leq p \leq \frac{5}{N}}$$

$$\hat{p} = \frac{2}{N} \overset{95\%}{\Rightarrow} \boxed{0 \leq p \leq \frac{7}{N}}$$

Common mistake

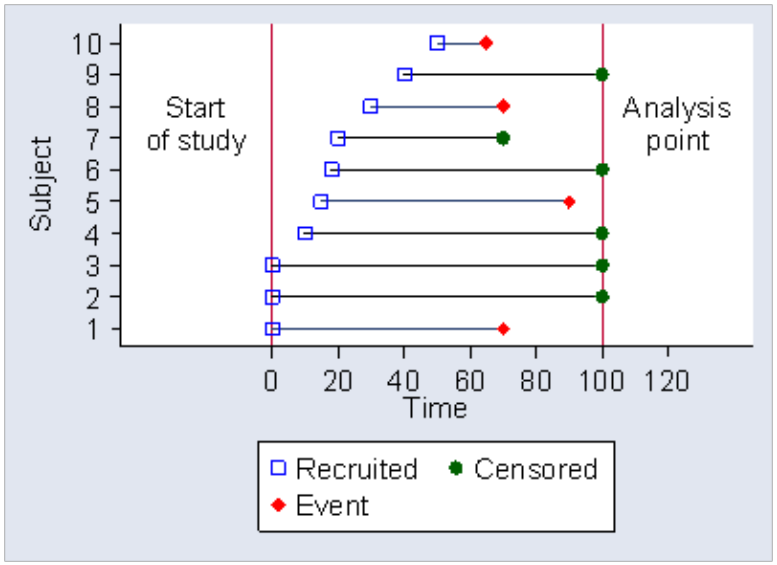
Giving an antipyretic to mice with fever makes their temperature drop from 39.5°C to 37°C. That is a temperature drop of 6.3%.

This percentage is not a probability of the occurrence of an event.

It is the change of a continuous variable. So the confidence interval calculated in this section does not apply.

Outline

Survival data



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

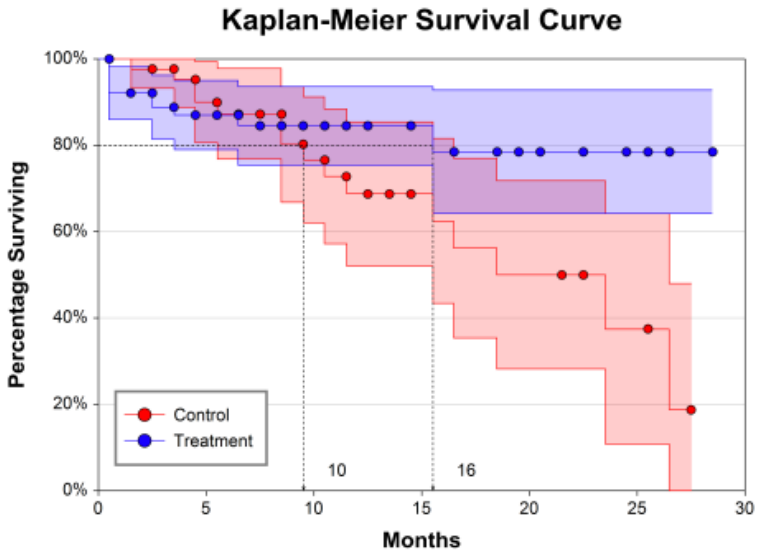
- ... death
- ... occlusion of a vascular graft
- ... first metastasis
- ... rejection of a transplanted kidney

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

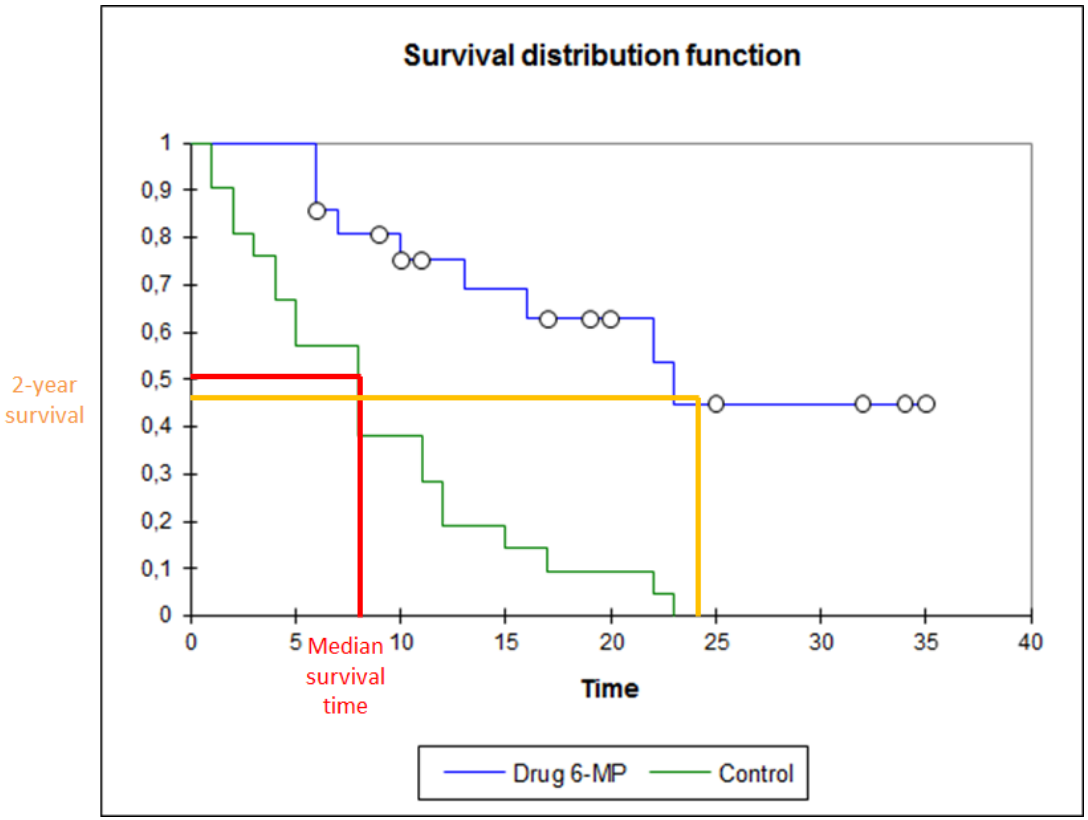
Time Period	At Risk	Became Unavailable (Censored)	Died	Survived	Kaplan-Meier Survival Probability Estimate
Year 1	100	3	5	95	(95/100)=0.95
Year 2	92	3	10	82	(95/100)x(82/92)=0.8467
Year 3	79	3	15	64	(95/100)x(82/92)x(64/79)=0.70
Year 4	61	3	20	41	(95/100)x(82/92)x(64/79)x(41/61)=0.4611
Year 5	38	3	25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

Outline

Counted data

Poisson distributed counts (number of events occurring independently of the time since the last event and at a fixed rate). The variance of Poisson is equal to its mean.

- Babies born in an obstretics ward each day.
- Number of eosinophils see in one microscope field.
- Number of radioactive disintegrations detected by a scintillator in 1 minute.

Binomial distributed counts (number of Bernouilli successes, each with a fixed probability, occurring independently)

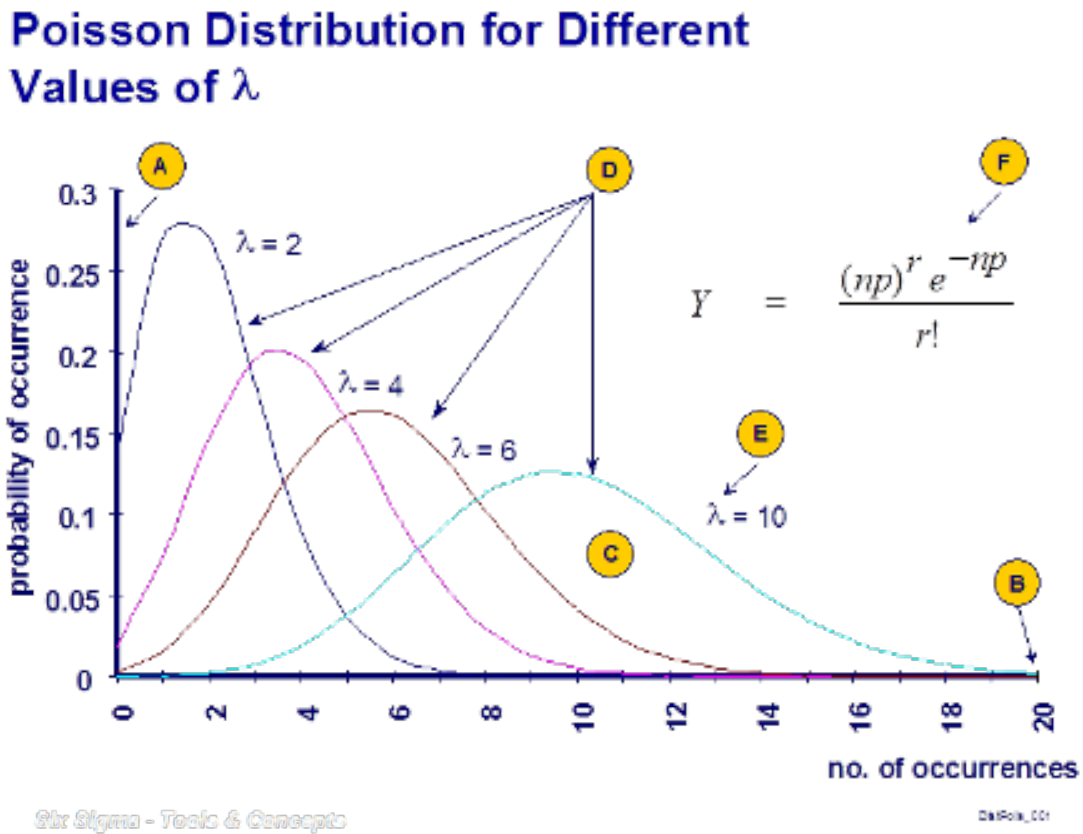
- Number of heads in 50 coin flips.
- Number of left-handed and right-handed in a sample.
- Number of male and female in a sample.

Counted data

Negative binomial distributed counts (number of Bernouilli successes, each with a fixed probability, occurring independently before r failures are observed). This distribution is used for overdispersed Poisson counts (the variance is larger than its mean). Negative binomial has two parameters (p and r) that can be adjusted to the observed data.

- Number of parasites in a blood specimen.
- Number of alcoholic drinks taken over a period of time.
- Incidence rate of mastitis in cattle.
- Annual counts of tropical cyclones in the North Atlantic.

Poisson distribution



Distribution examples for event rates of $\lambda = 2, 4, 6, 10$ counts/min, counts/field, counts/period, ...

Assumptions of Poisson distribution

- The event is clearly defined (the cell type in a microscope field is sometimes difficult to determine).
- Each event occurs randomly, independently of other events (baby twins in an obstretics ward violate this assumption, parasite aggregations in blood samples too)
- The average rate does not change over time.
- Each event is counted only once (in a study of airplanes close to collide, researchers asked pilots and copilots how many times they were about to collide with another plane; some events were counted twice because the pilots and copilots of the two planes were interviewed separately).

Confidence interval. Larger samples are better.

- You carefully dissect 1 bagel and find 10 raisins. If raisins do not aggregate and the recipe does not change over time, the 95% confidence interval is between 4.8 and 18.4 raisins per bagel.
- You carefully dissect 10 bagels and find 9, 7, 13, 12, 10, 9, 11, 9, 10 and 10 raisins. A total of 100 raisins in 10 bagels (an average of 10 raisins per bagel). For 100 objects counted, the 95% confidence interval is from 81.36 to 121.63. If we divide by 10, then the confidence interval for the number of raisins per bagel is from 8.1 to 12.2. (A much smaller confidence interval.)

Confidence interval. Larger samples are better.

If we observe C counts, then ...

- ... the exact confidence interval is

$$\frac{1}{2}\chi^2_{\frac{\alpha}{2},2C} \leq \lambda \leq \frac{1}{2}\chi^2_{1-\frac{\alpha}{2},2(C+1)}$$

- ... an approximated confidence interval is

$$C - z_{1-\frac{\alpha}{2}}\sqrt{C} \leq \lambda \leq C + z_{1-\frac{\alpha}{2}}\sqrt{C}$$

Remind that for $\alpha = 0.05$, $z_{1-\frac{\alpha}{2}} = 1.96$

Confidence interval

Outline

Chapter 2. Continuous variables

C.O.S. Sorzano
coss@cnb.csic.es

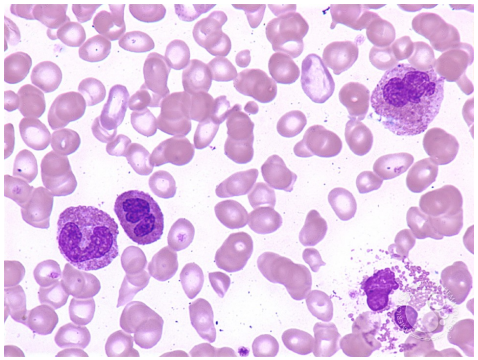
National Center of Biotechnology (CSIC)

August 13, 2017

Outline

Outline

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:

$$37.0,\ 36.0,\ 37.1,\ 37.1,\ 36.2,\ 37.3,\ 37.0,\ 37.0,\ 36.1$$

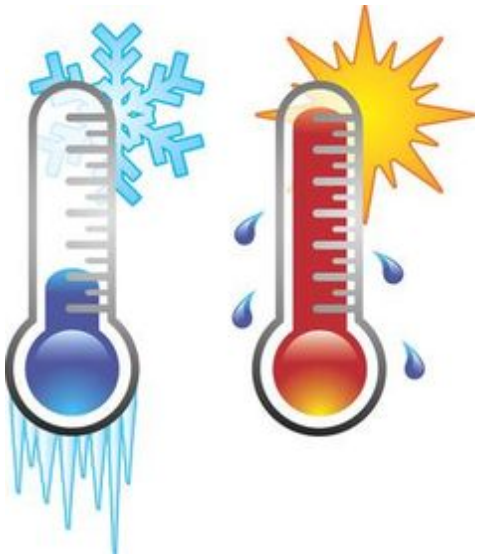
We may calculate a measure of centrality:

- 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: $\hat{\mu} = \frac{1}{\frac{\frac{1}{37.0}+\frac{1}{36.0}+\frac{1}{37.1}+\frac{1}{37.1}+\frac{1}{36.2}+\frac{1}{37.3}+\frac{1}{37.0}+\frac{1}{37.0}+\frac{1}{36.1}}{9}} = 36.75$
- Mode: $\hat{\mu} = (36.0, 36.1, 36.2, 37.0, 37.0, 37.0, 37.1, 37.1, 37.3) = 37.0$

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: Interval variables

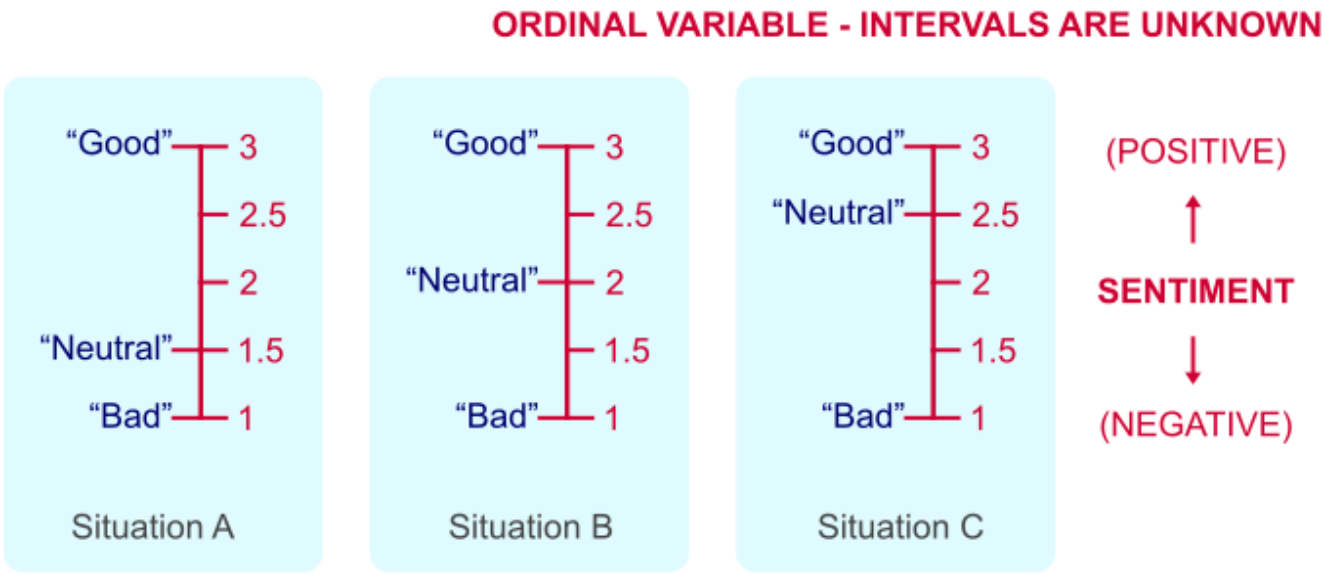


- A change of an interval (e.g. 1°) is the same all along the interval).
- The zero reference may be arbitrary (Celsius or Fahrenheit degrees).
- If the temperature of an object is 20° and of another object 30°, the temperature of the second object is not 50% larger (in Fahrenheit scale, the percentage would be different).
- Calculating differences between two interval variables makes sense.
- Calculating ratios between two interval variables does not make sense.

Types of variables: Ratio variables

- The zero reference is not arbitrary (height, weight, enzyme activity, temperature in Kelvin).
- It makes sense calculating the ratio between two values. A weight of 4 grams is twice the weight of 2 grams.
- Calculating differences and ratios between two ratio variables makes sense.

Types of variables: Ordinal variables

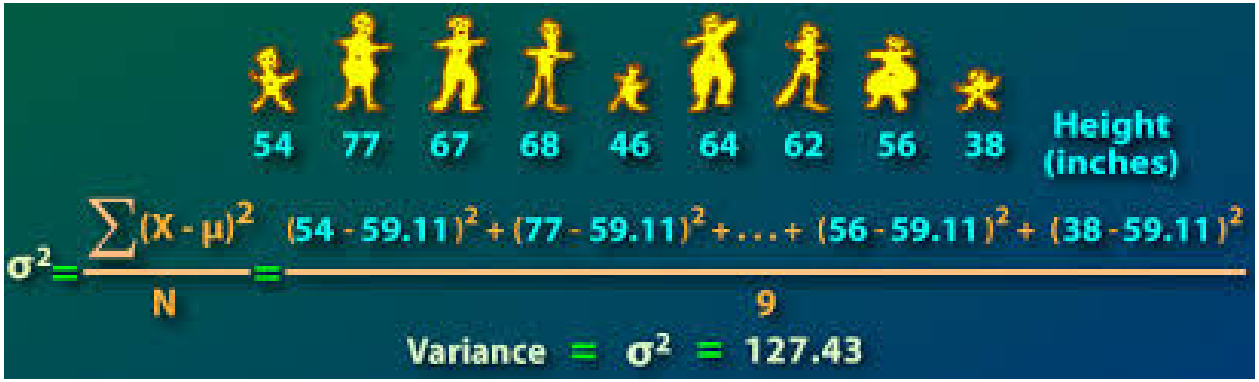


- Ordinal variables only express a relative rank between variables.
- Differences or ratios are meaningless.

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.

Variability and bias

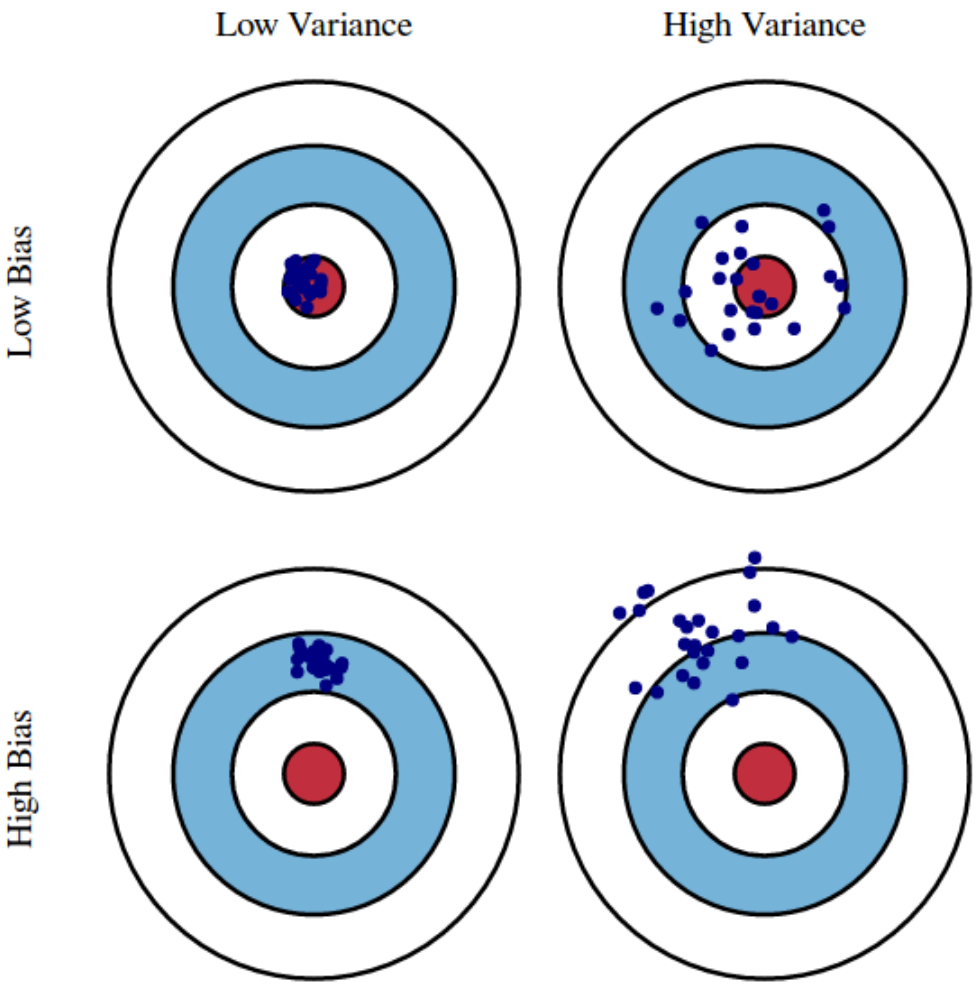
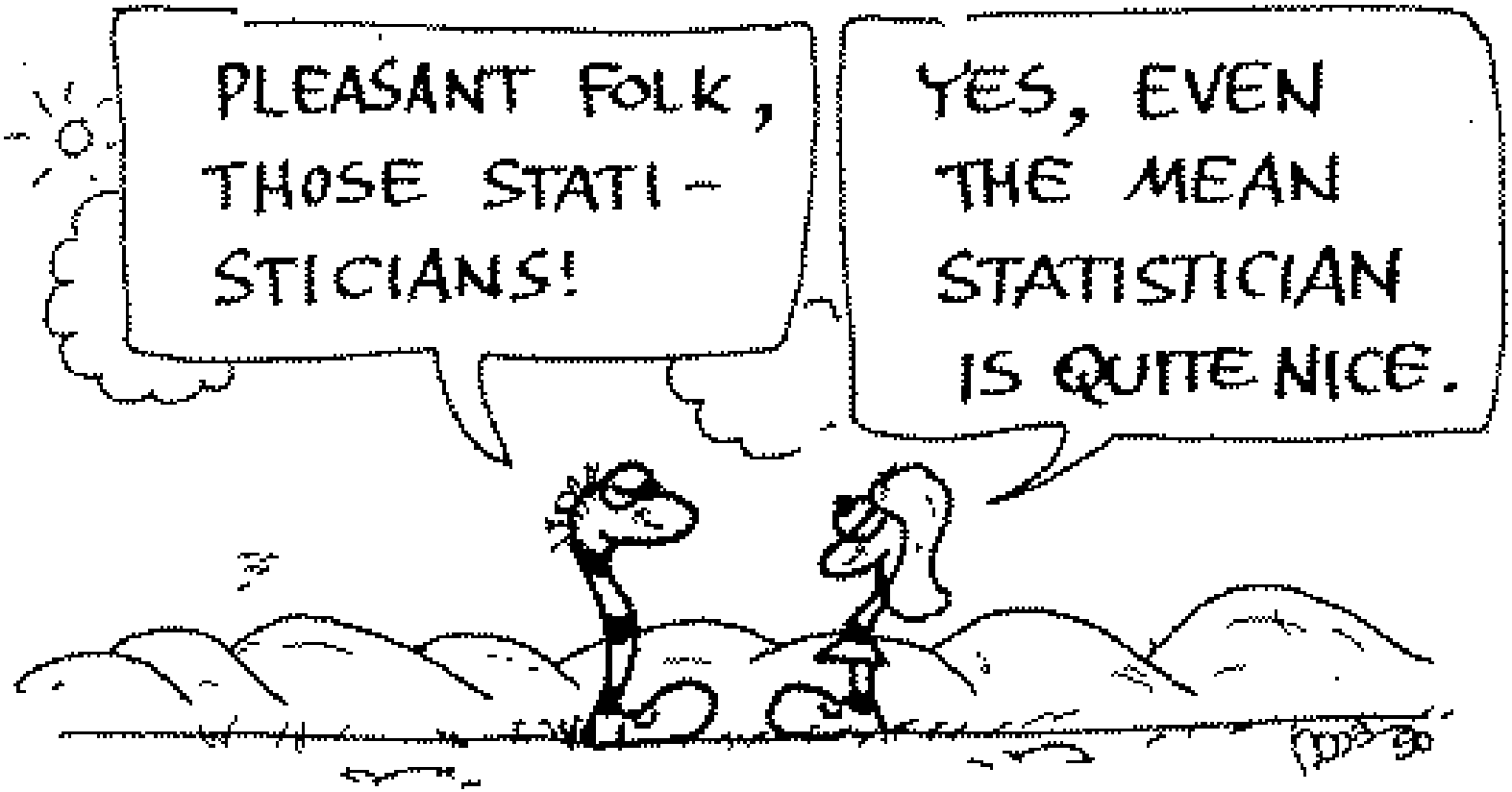


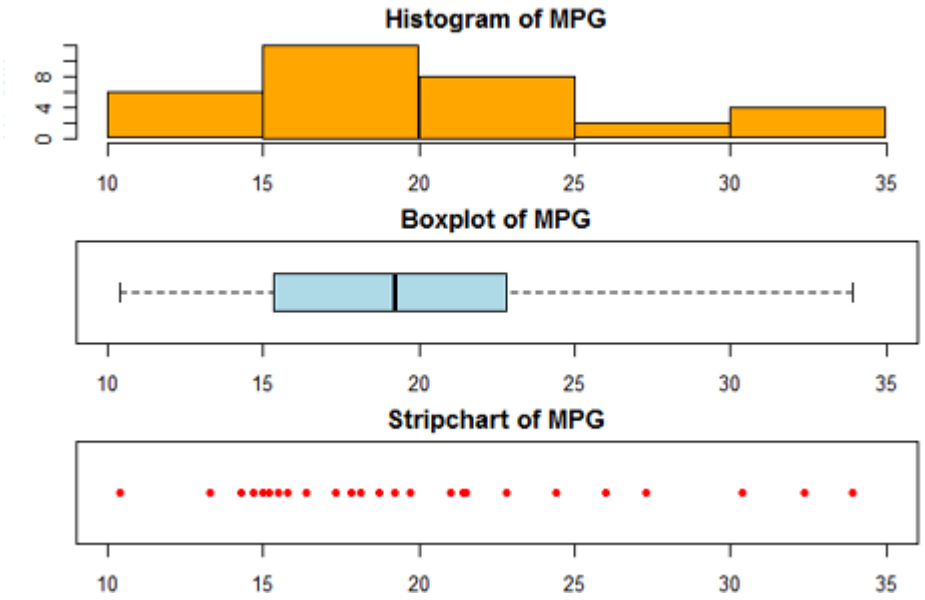
Fig. 1 Graphical illustration of bias and variance.

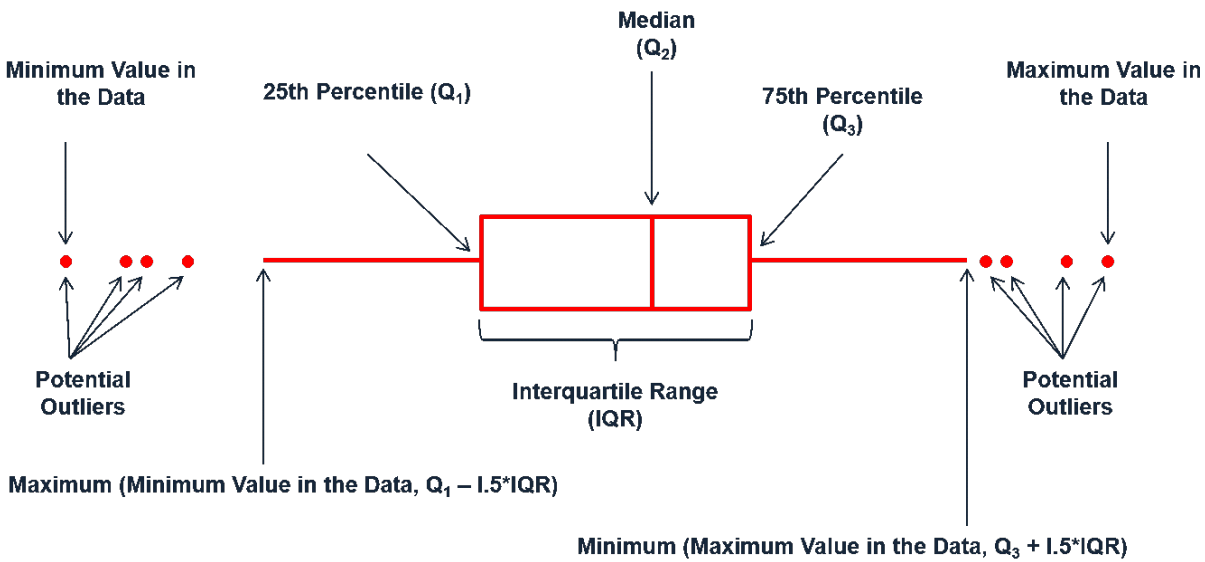
Mean


PLEASANT FOLK, THOSE STATI-STICIANS!
YES, EVEN THE MEAN STATISTICIAN IS QUITE NICE.

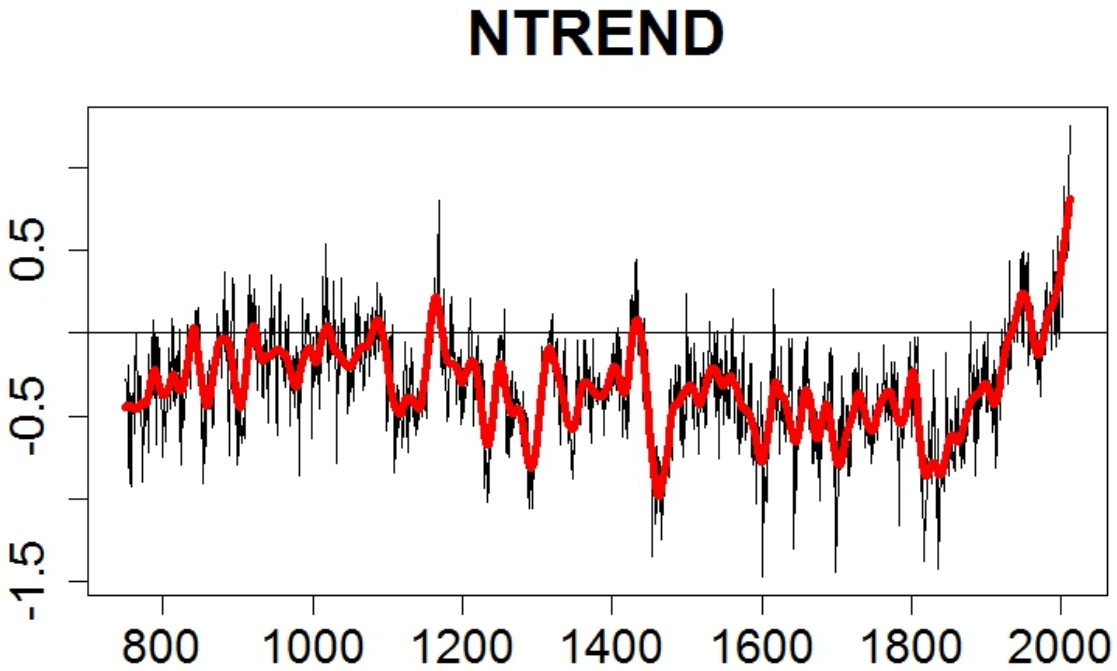
Outline

Plots: 1D Scatter plots, histograms and boxplots



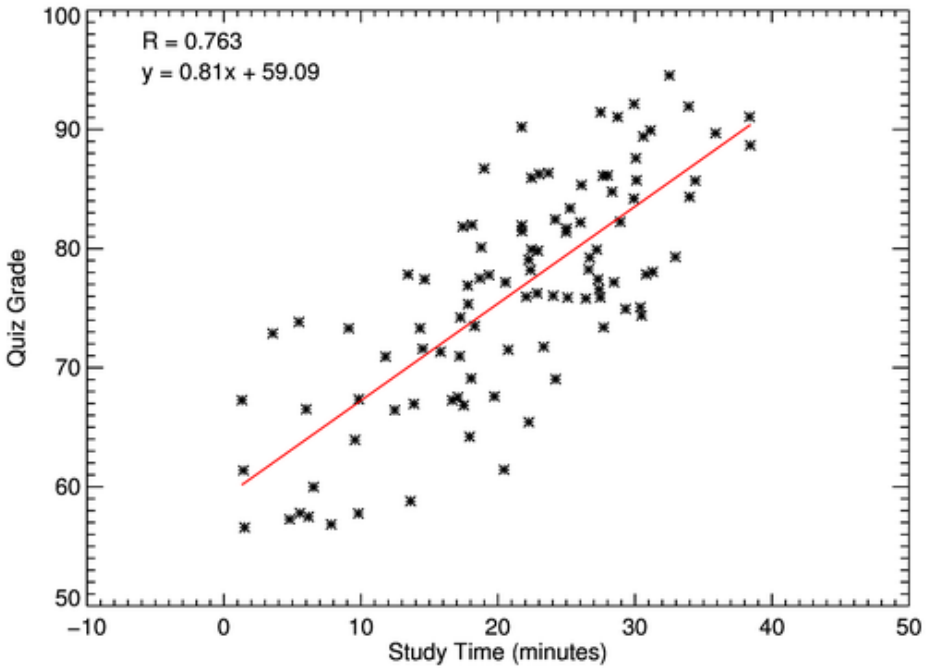


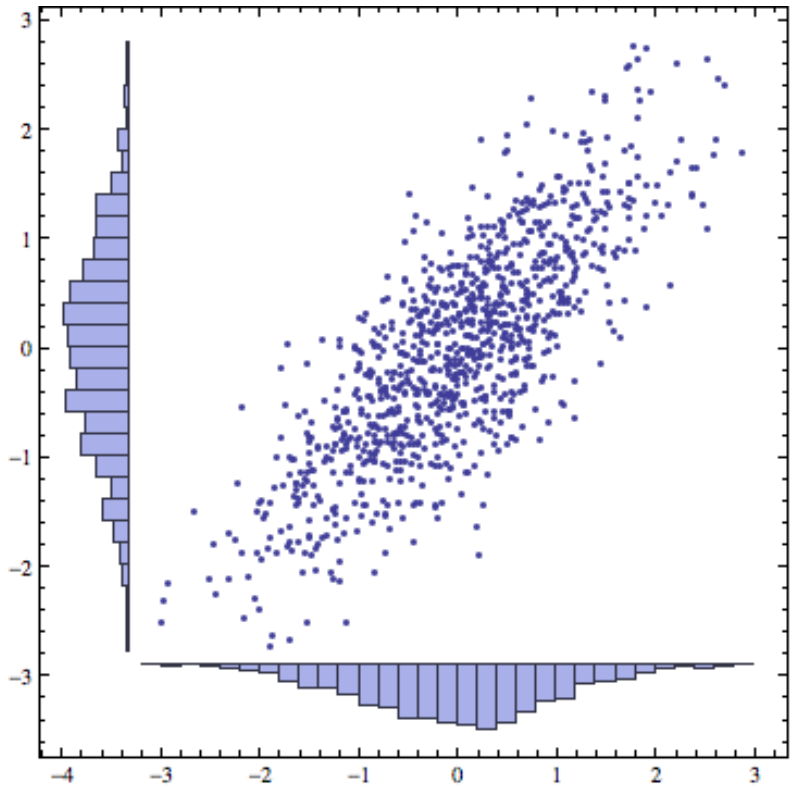
Plots: Time plots and data smoothing



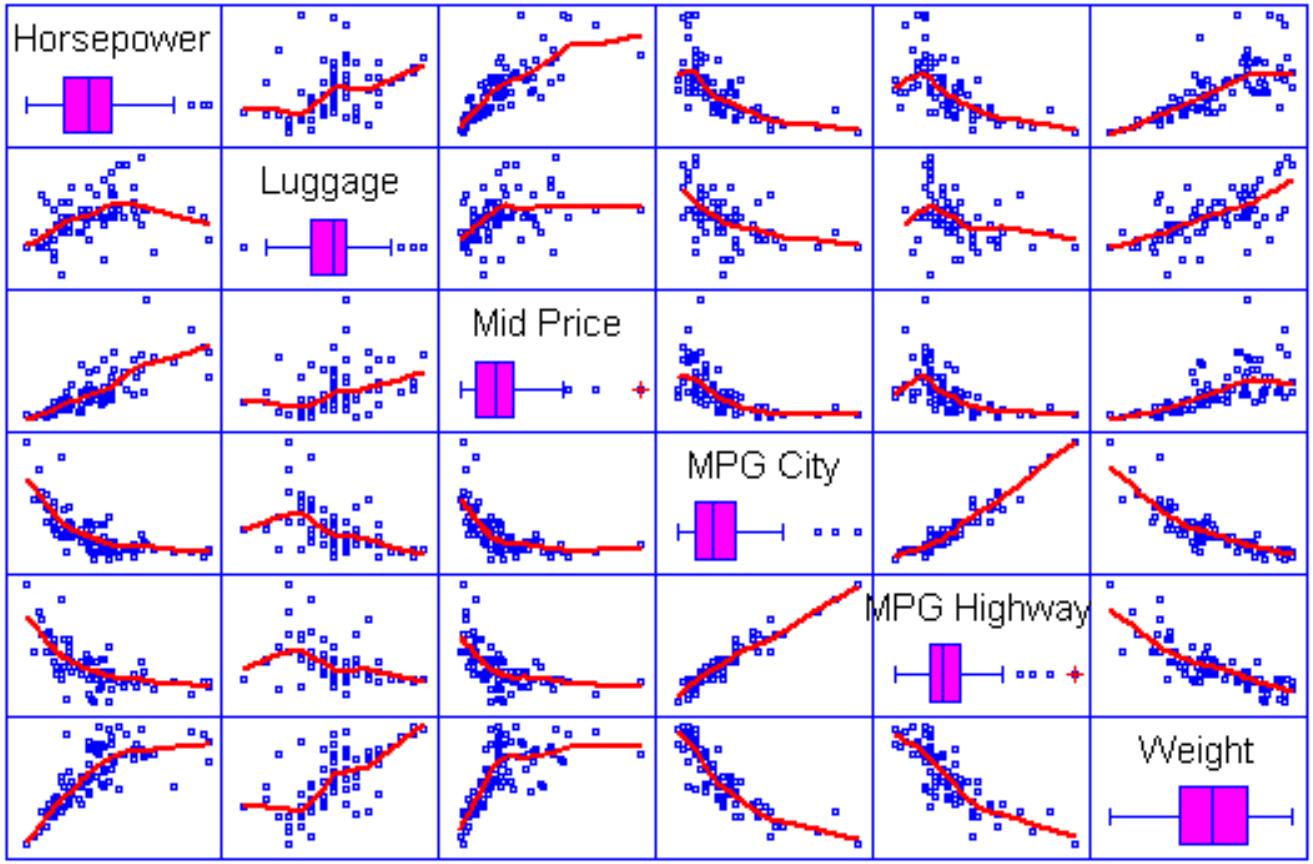
Typical data smoothers are splines and LOESS.

Plots: 2D Scatter plots and 1D histograms

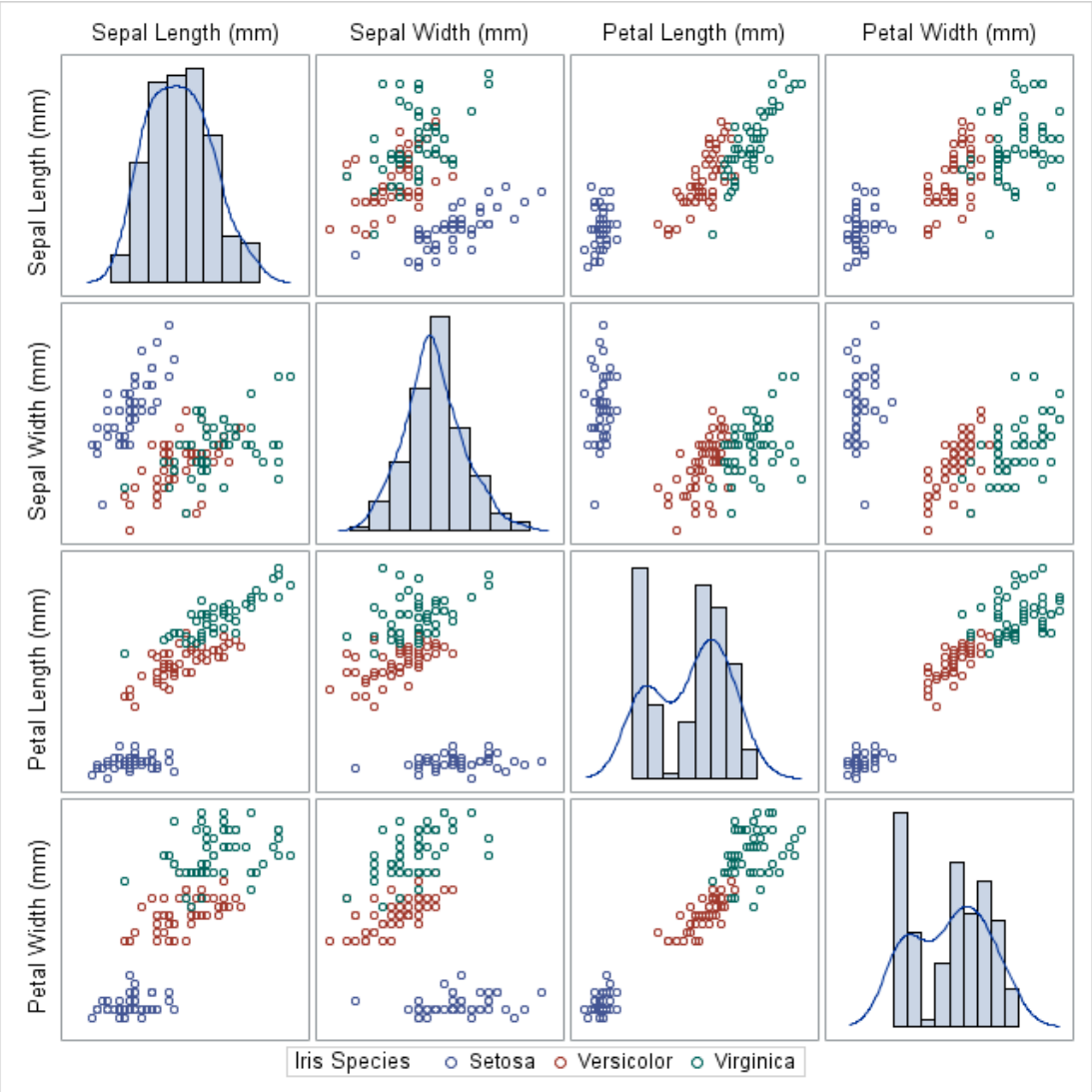




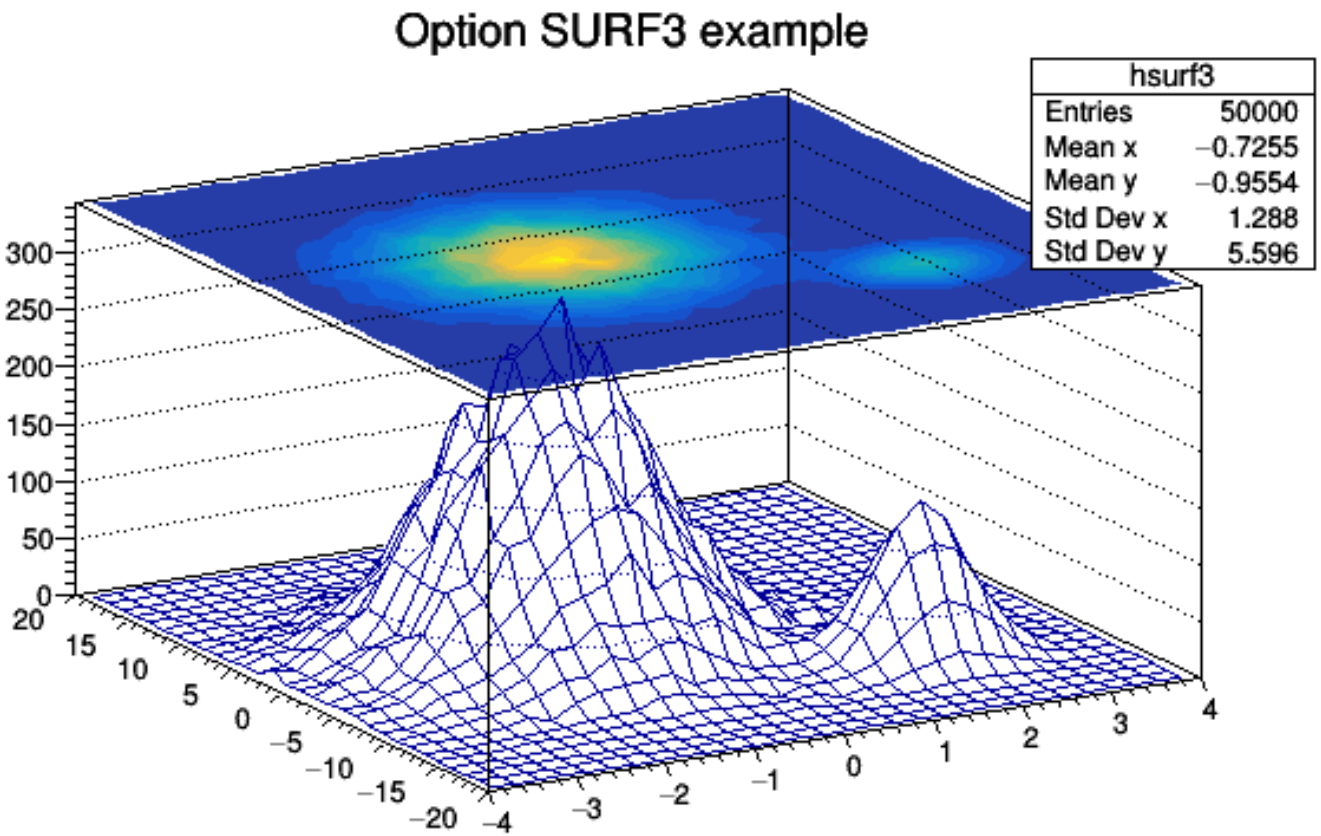
Plots: 2D Scatter plots



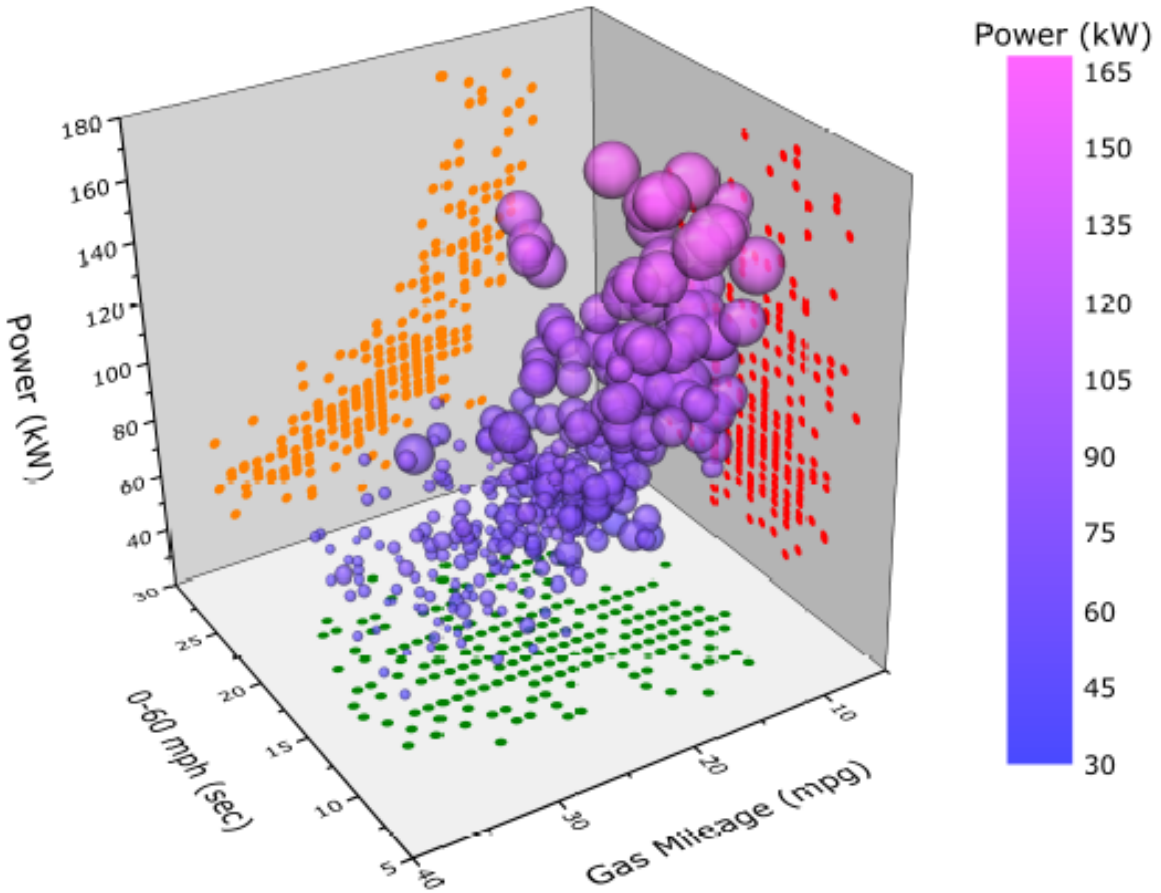
Plots: 2D Scatter plots



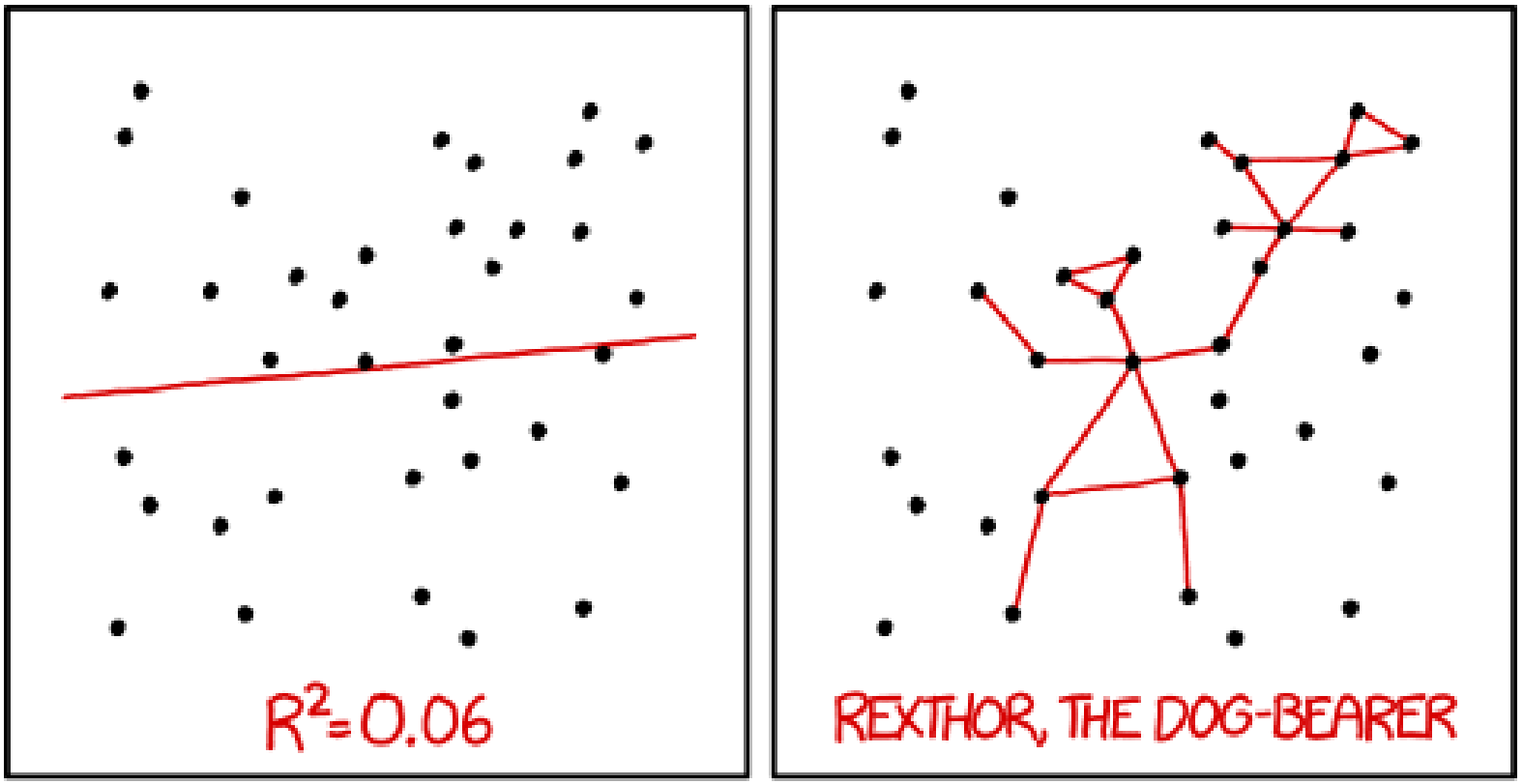
Plots: 2D histograms



Plots: 3D Scatter plots



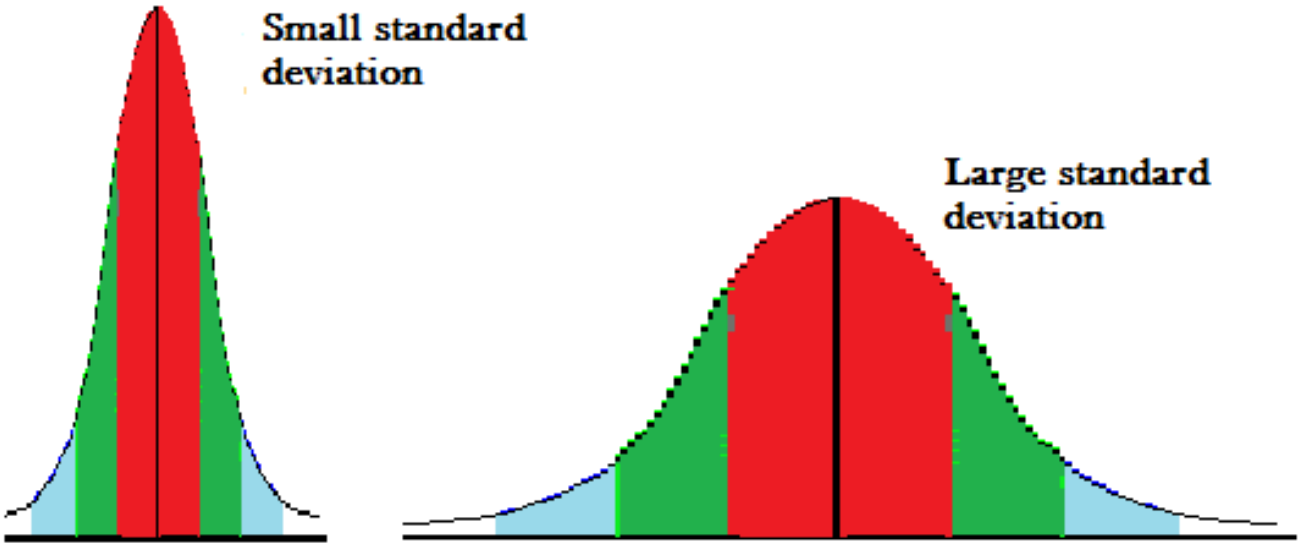
Scatter plots



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

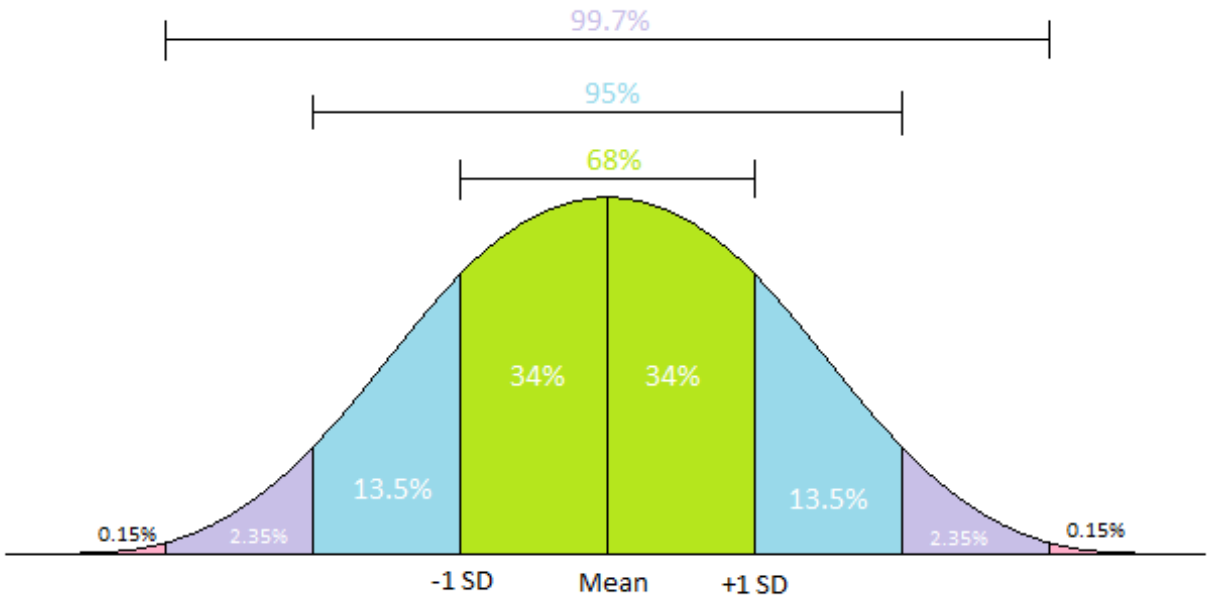
Outline

Standard Deviation



The standard deviation (SD) expresses how samples differ from the average. For example, the average human temperature is 36.82°C with a SD of 0.41°C.

Standard Deviation



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

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, $\hat{\sigma}^2$.

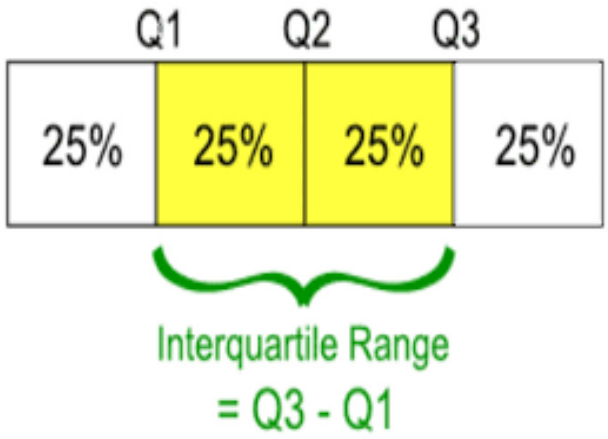
Means and standard deviations are sensitive to outliers. The equivalent robust estimates are the median and median absolute deviation (MAD)

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

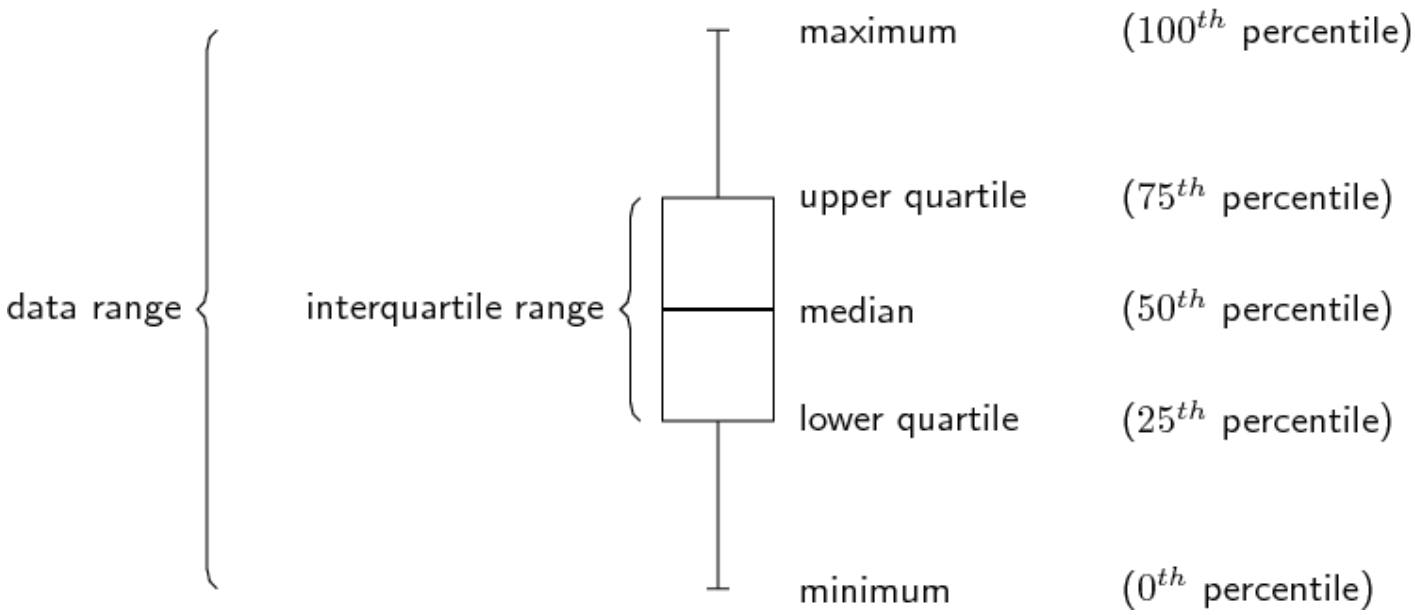
Standard Deviation

- N is the number of independent samples (biological replicates).
- Technical replicates (measuring the same individual multiple times) does not give independent samples.
- Multiple measurements from the same individual ($n = 1$ experiments) is representative of the samples obtained from that person, not from the whole population.
- The SD from 100 samples is approximately the same as the SD from 1000 samples. The SD quantifies the underlying variability, as long as the sample is large enough to gain some precision, the SD estimates should not change with the sample size.

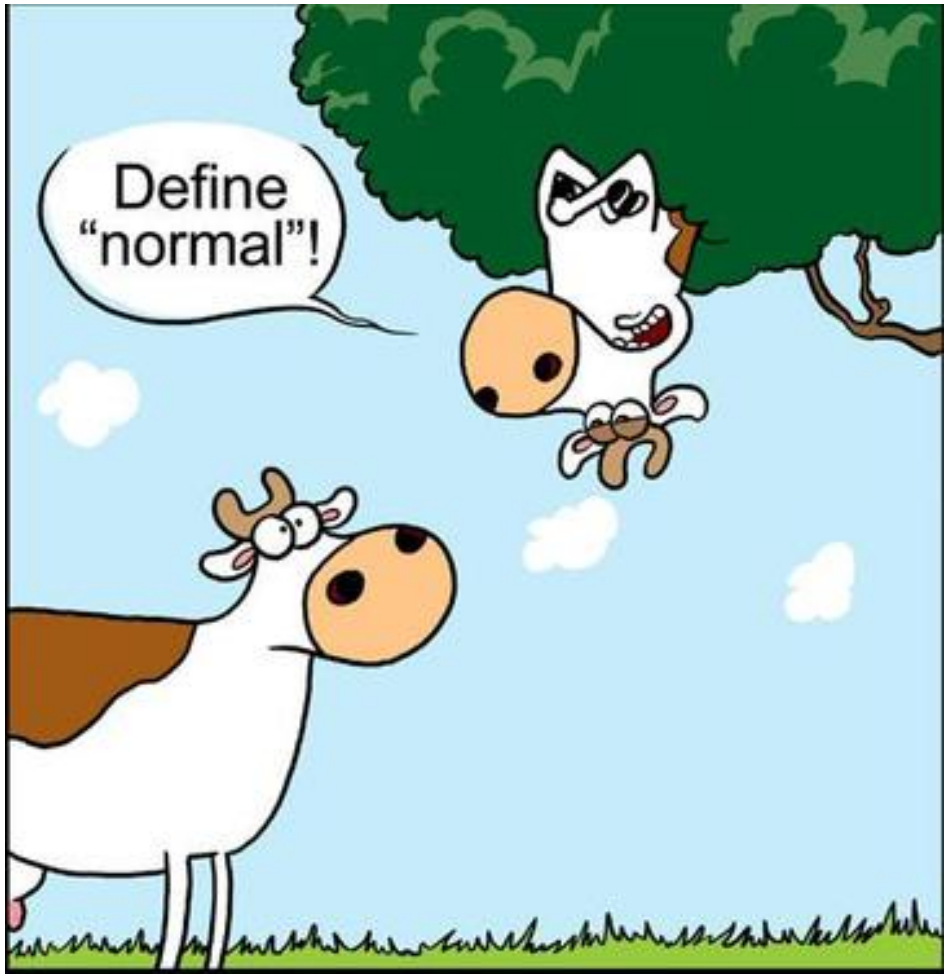
Histogram summary



The interquartile range shows the difference in the central 50% of the data.
The 5-number summary shows a quick look summary of the histogram.

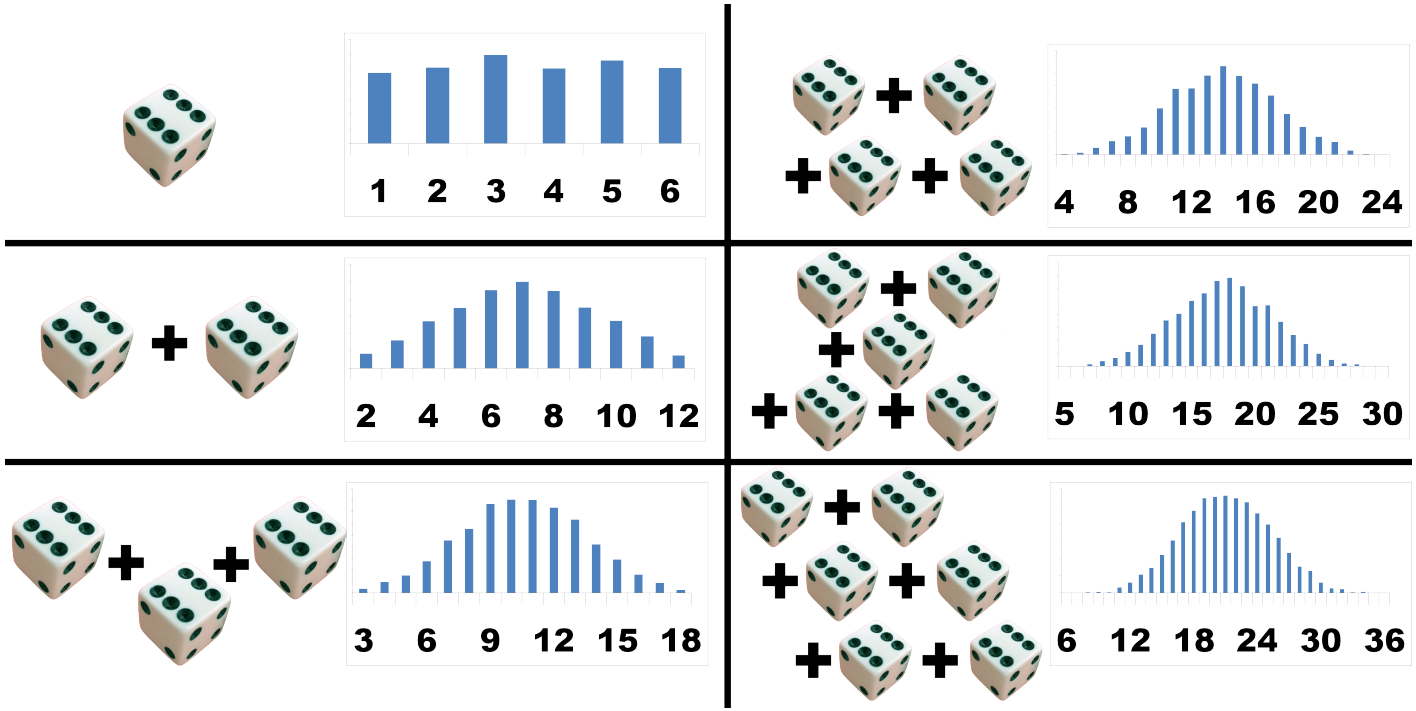


Variability



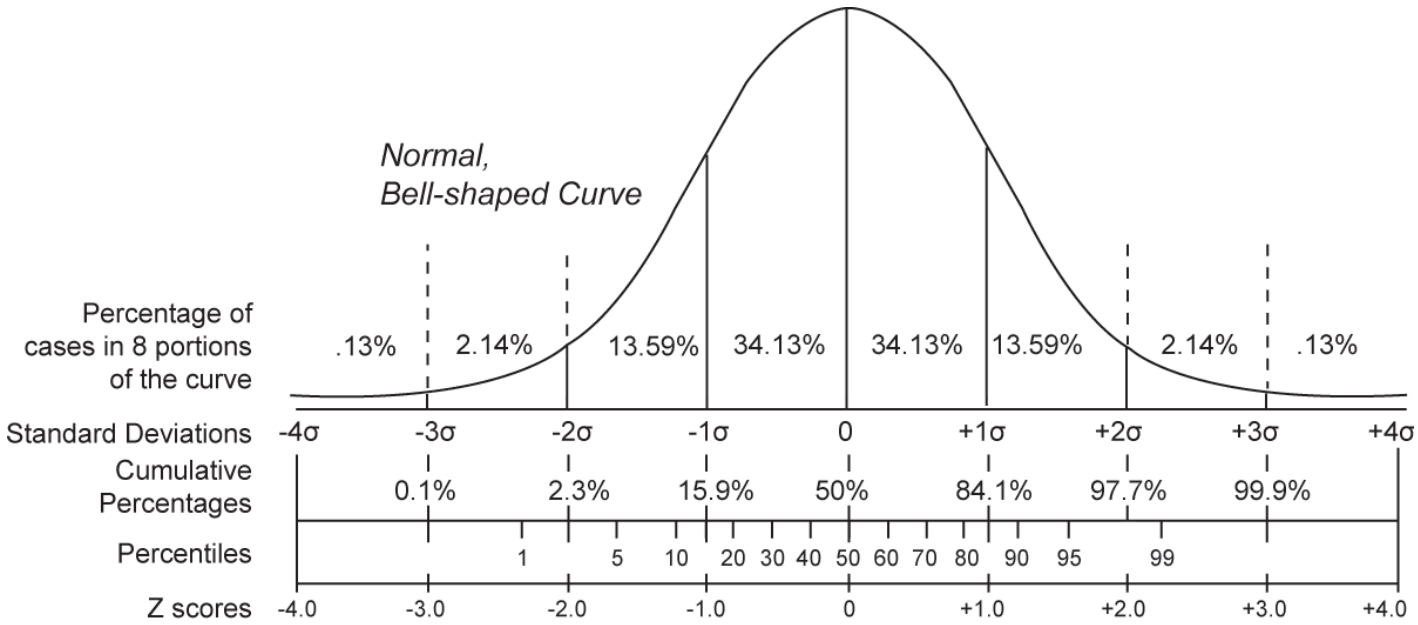
Outline

Gaussian distribution



The Gaussian is the limit distribution of many additive random variables (central limit theorem). Variations in experiments may be caused by many factors at the same time: imprecise weighing of reagents, imprecise pipetting, the random nature of radioactive decay, nonhomogeneous suspensions of cells, ...

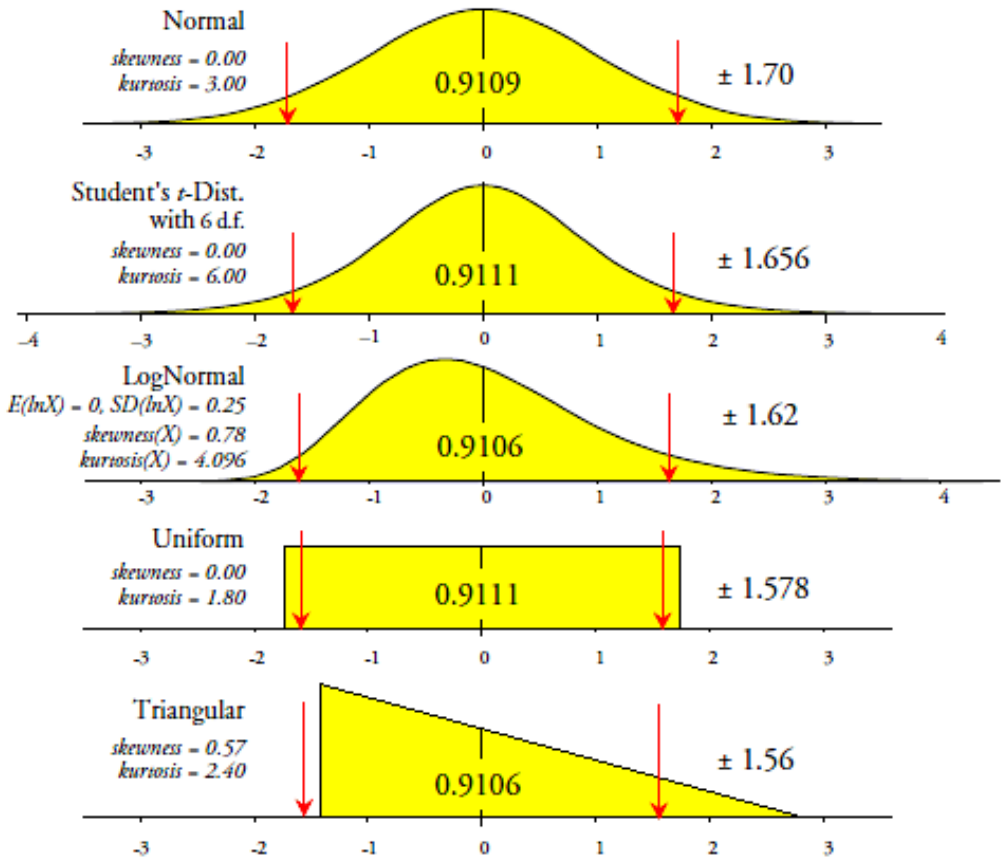
Gaussian distribution



This is the "probability density function". It indicates the likelihood of each value. On this curve we can measure percentiles (the 95% percentile is a value such that 95% is a value such that 95% of the values drawn from this distribution occur below this value). We may normalize a value through its z-score

$$z = \frac{x - \mu}{\sigma}$$

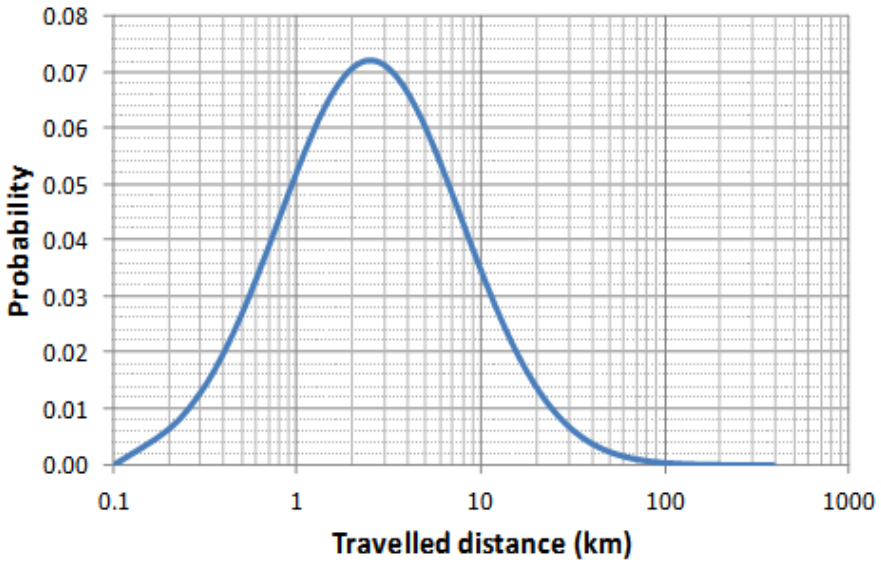
Not everything is Gaussian (Normal)

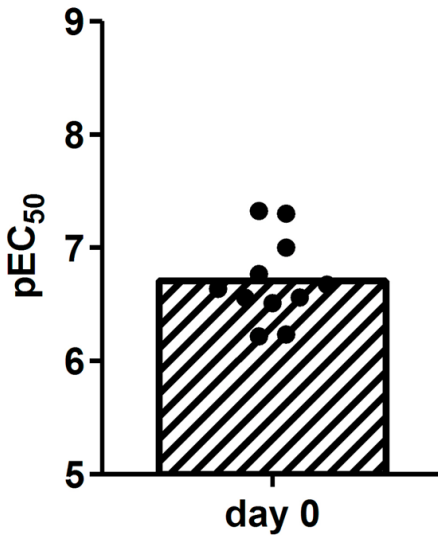


Many other distributions exist. Even there are variables for which there are no closed-form distributions. In the example above, the distributions have been normalized to have 0 mean and standard deviation 1. The red arrows indicate a central interval with 91.1% of the population.

Log-normal distribution

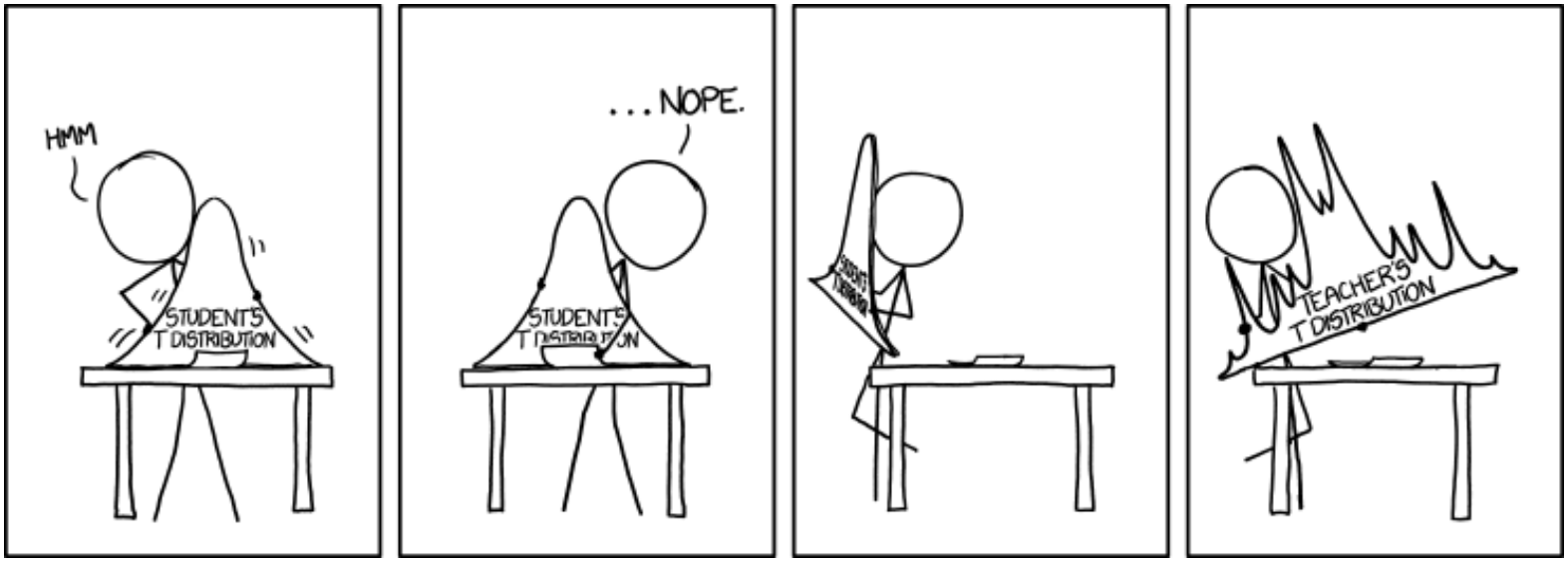
The logarithm of our measurements are Gaussian. This is typical of cell counts and concentrations (e.g., EC50, concentration to achieve 50% of the effect, pEC50 is its logarithm).





This distribution is well suited for variables that act as multiplicative rather than additive. They are equally likely to double their value or cur it in half. These variables should be used in the logarithmic scale and treated, then, normally.

Mean


HMM
STUDENT'S T DISTRIBUTION
...NOPE.
STUDENT'S T DISTRIBUTION
TEACHER'S T DISTRIBUTION

Outline

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°C and its 95% confidence interval [36.37,37.14]°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°C and its 95% confidence interval [0.34,0.96]°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

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

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

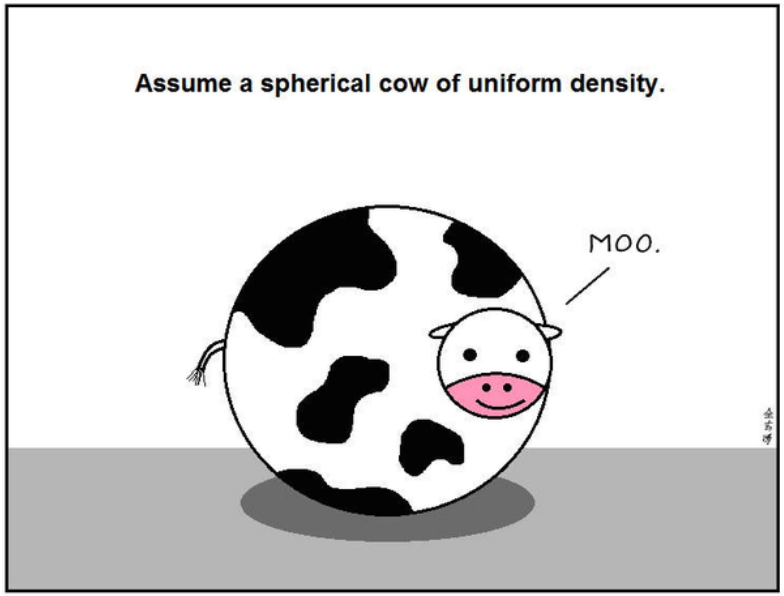
Assumptions of confidence intervals

- 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 joint 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

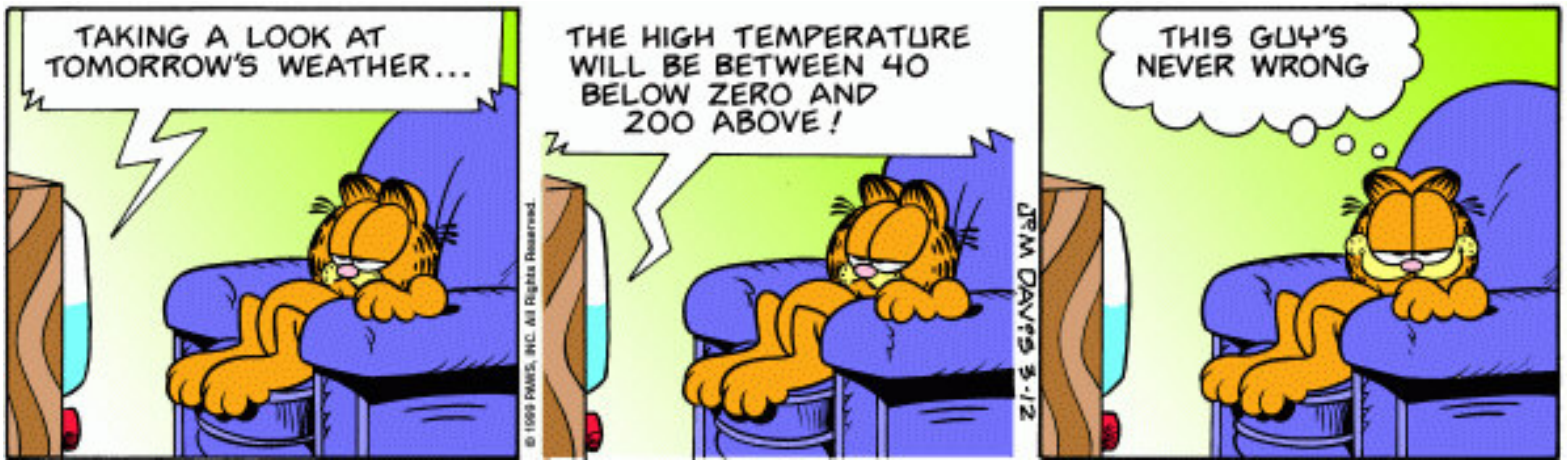
- More samples. The larger the experiment, N, the narrower the CI (we have less uncertainty about the underlying parameter).
- More confidence. The larger the confidence, $1 - \alpha$, the wider the CI (we need to enlarge it to be surer that it contains the true parameter).

What if the assumptions are violated?



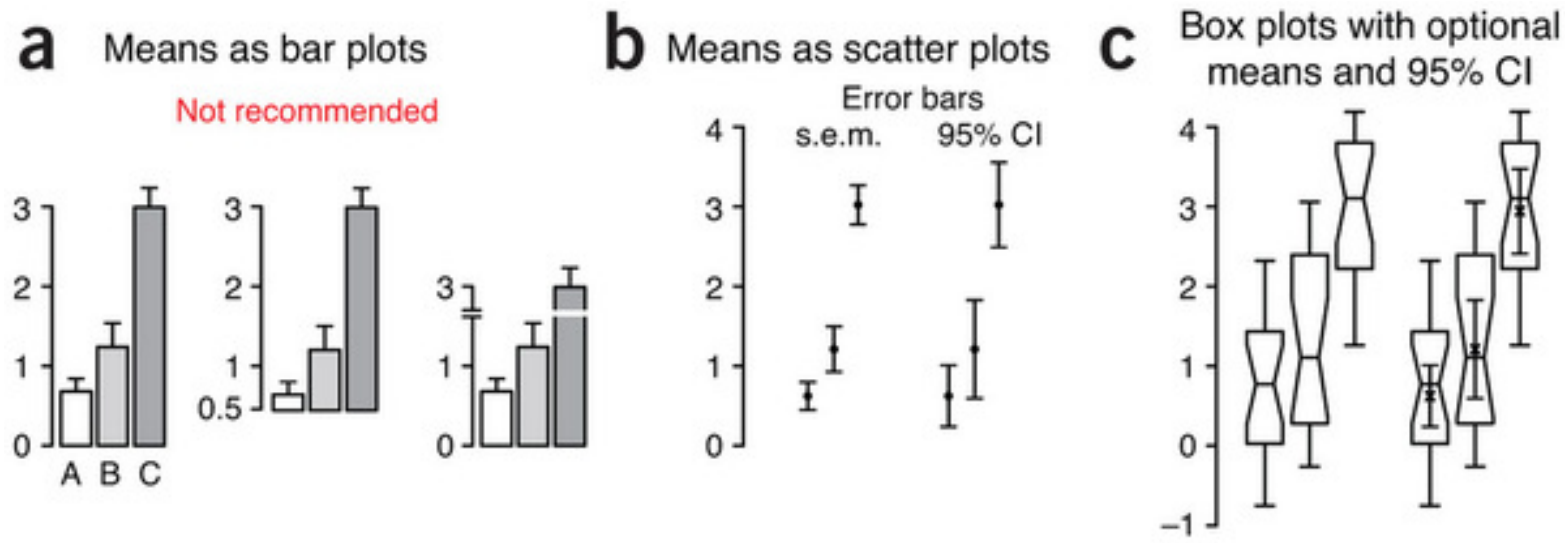
In many situations, these assumptions are not strictly true. Then, the CI may still be a reasonable approximation of the range of the underlying parameter (depending on the severity of the violation). But the confidence will, for sure, not be the one we think (95%).

Confidence Intervals



Outline

CIs vs SD vs SEM



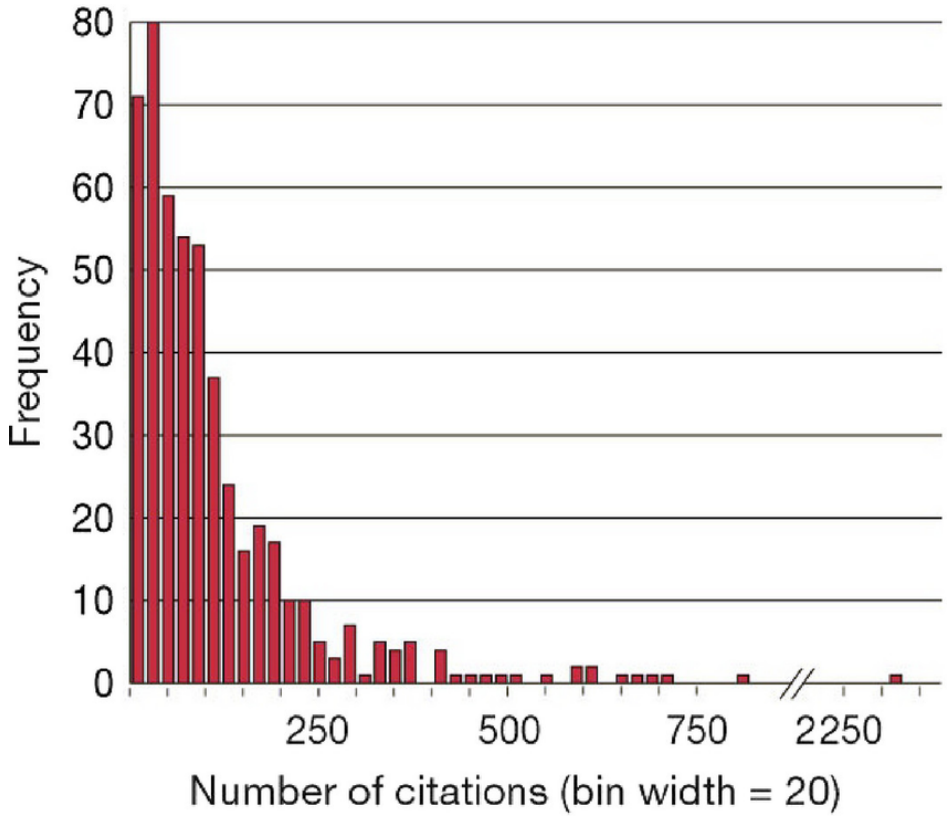
- SD is the standard deviation of the population
- SEM (standard error of the mean) is the standard deviation of our estimate of the mean

$$SEM = \frac{SD}{\sqrt{N}}$$

- CI is about our estimate of the mean

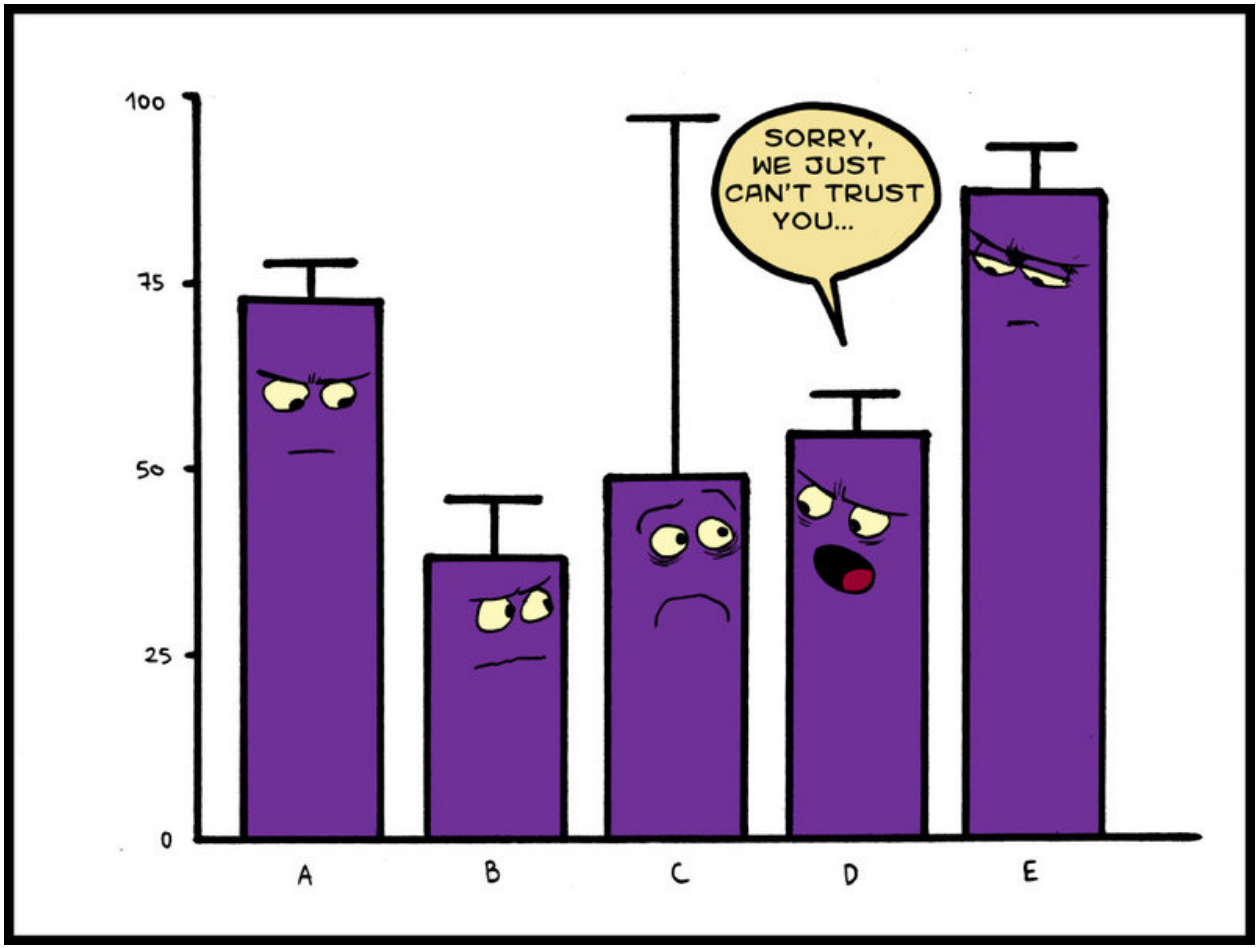
SEM and CIs show how accurate is our estimate of the mean, not how variable is our data. To show the variability of the data show a boxplot or a histogram.

Boxplots better than $\hat{\mu} \pm SD$



$\hat{\mu} \pm SD$ implicitly imply that the distribution is symmetric (Gaussian), although this is not always the case as in this example of the number of citations of 500 Nature papers.

Mean



Outline

Practice

- Compute the mean, median, SD, percentiles of a population.
- Compute the SEM of the mean estimate.
- Compute the 95% CI of the mean estimate.
- Represent a boxplot of the population.
- Represent a histogram of the population.
- Multiple scatter plots of several variables.

Outline

Chapter 3. P-values and statistical significance

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

August 13, 2017

Outline

Outline

What is a p-value through examples

We flip a coin 20 times and observe 16 heads and 4 tails. Is it a fair coin?. On average, for a fair coin we would have expected 10 heads and 10 tails. 16 heads is too far from 10, can we say the coin is not fair?

This table shows the probability (%) of observing different number of heads in 20 flips.

0 or 20	0.000	1 or 19	0.002	2 or 18	0.018	3 or 17	0.109	4 or 16	0.462
5 or 15	1.479	6 or 14	3.696	7 or 13	7.393	8 or 12	12.013	9 or 11	16.018
10	17.620								

We should be suspicious on the coin if we observe a result as strange as 16 heads or even stranger. These results are 0, 1, 2, 3, 4, 16, 17, 18, 19, 20 heads.

$$\begin{aligned} \text{p} - \text{val} &= \Pr\{0\} + \Pr\{1\} + \Pr\{2\} + \Pr\{3\} + \Pr\{4\} + \\ &\quad \Pr\{16\} + \Pr\{17\} + \Pr\{18\} + \Pr\{19\} + \Pr\{20\} \\ &= 1.18\% = 0.0118 \end{aligned}$$

The p-value, that is, if the coin is fair, the probability of observing a result as extreme or more as the actually observed (16 heads) is 1.18%. That is, just by coincidence, a fair coin flipped 20 times (and repeated this experiment many times) could have 0, 1, 2, 3, 4 or 16, 17, 18, 19, 20 heads in the run in 1.18% of the cases. Now it is your turn to decide. In this single run, we observed 16 heads, would you say the coin is fair?

What is a p-value through examples

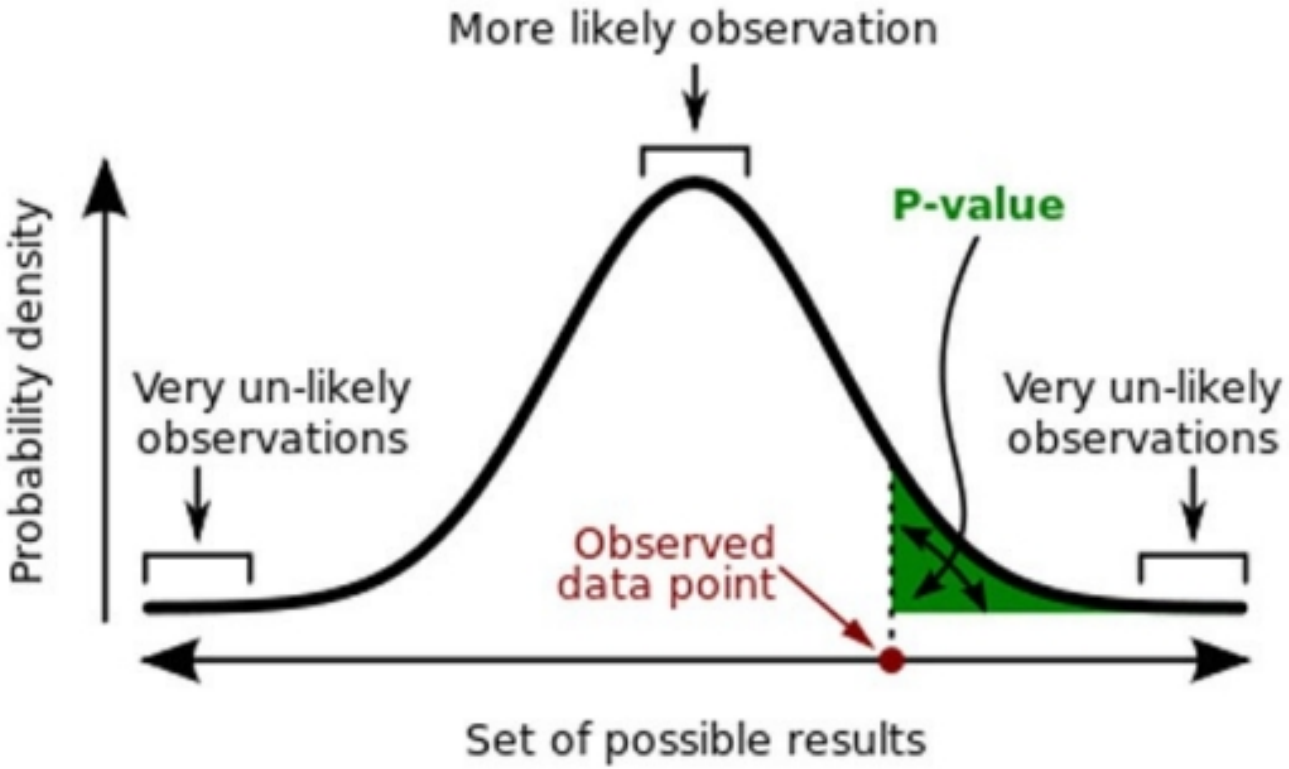
We randomly assigned 972 surgical patients to receive an antibiotic ointment or an ointment without an active medication.

Infections occurred in 6.6% of the patients who received the antibiotic and in 11.0% of the patients who received the inactive ointment. That is, the risk of infection was 67% higher in the inactive ointment group ($\frac{11}{6.6} = 1.666$).

If we assume that the risk of infection is the same in both groups and that the antibiotic ointment is not helping to prevent infections, what is the probability of observing a result as extreme as this one or more?

$$\mathrm{p-val} = 1\% = 0.01$$

What is a p-value through examples



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

The null and alternative hypotheses

- If the coin is fair, then ...
- If the the risk of infection is the same in both groups and that the antibiotic ointment is not helping, then ...

The null hypothesis is the state of affairs we want to disprove. The alternative hypothesis is the opposite (what we want to prove):

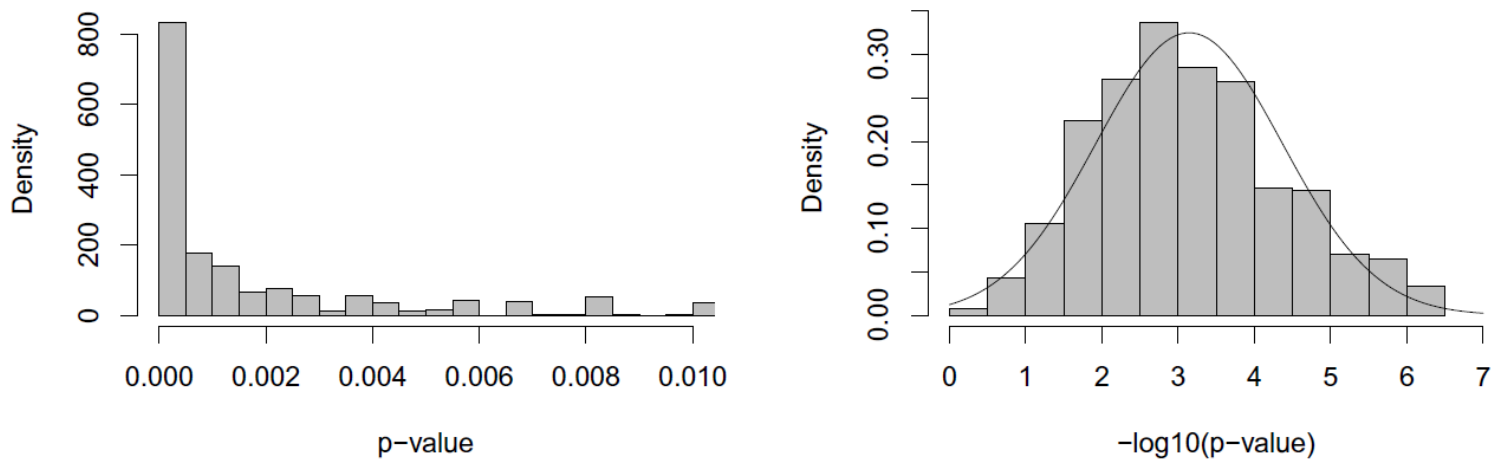
- The coin is not fair.
- The risk of infection is not the same in both groups or the antibiotic ointment is helping.

p-values are random variables

Miller (1986)gave data comparing partial thromboplastin times for patients whose blood clots were dissolved (R=recanalized) and for those whose clots were not dissolved (NR).

- R: 41 86 90 74 146 57 62 78 55 105 46 94 26 101 72 119 88
- NR: 34 23 36 25 35 23 87 48

The (bootstrap) distribution of p-values would be



That is, with this data (or similar) we could have obtained a p-value going from 0.000 to more than 0.01. Its logarithm is approximately normal.

Boos and Stefanski. p-value precision and reproducibility. The American Statistician, 65:4, 213-221 (2011).

p-values and replicability

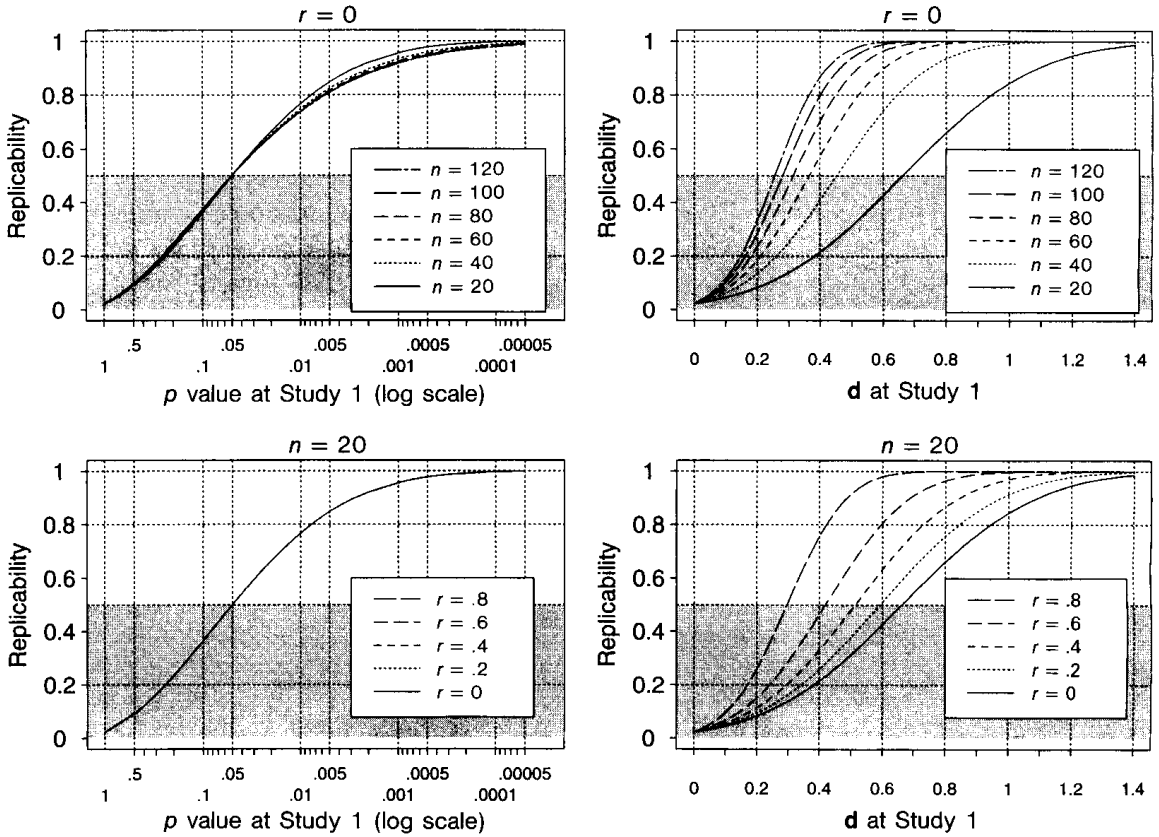
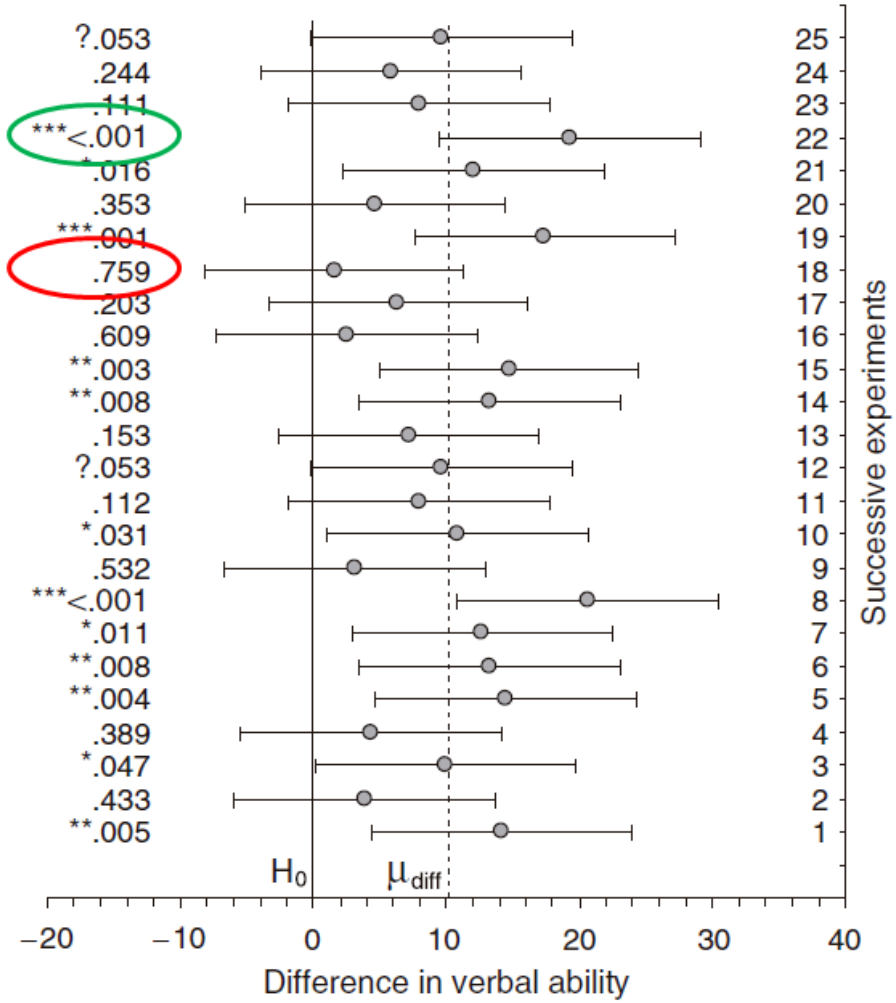


Figure 1. Estimated replicability as a function of p value (log scale) and effect size (**d**), assuming a two-treatment repeated-measures design with selected samples sizes (ns) and selected levels of between-treatment population correlation (r) of the dependent variable. The shaded portion of each plot gives values that are not properly replicabilities because the initial result is not a null hypothesis rejection in these regions. These regions of the plot show the probability of obtaining a null hypothesis rejection (at $\alpha = .05$, two-tailed) from a second study, contingent on the p value of a first study that did not reject the null hypothesis.

If we are in the limit of the p-value decision, there is a high chance (50%) that a replicate may result in a non-significant difference.

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

p-values and replicability



This is simulated data, all experiments have a mean difference of 10 and a SD=20 (effect size=10/20=0.5). However, some of the experiments are highly significant and some others are not. The p-value range from <0.001 to 0.76!!

Better use CIs, rather than p-values.

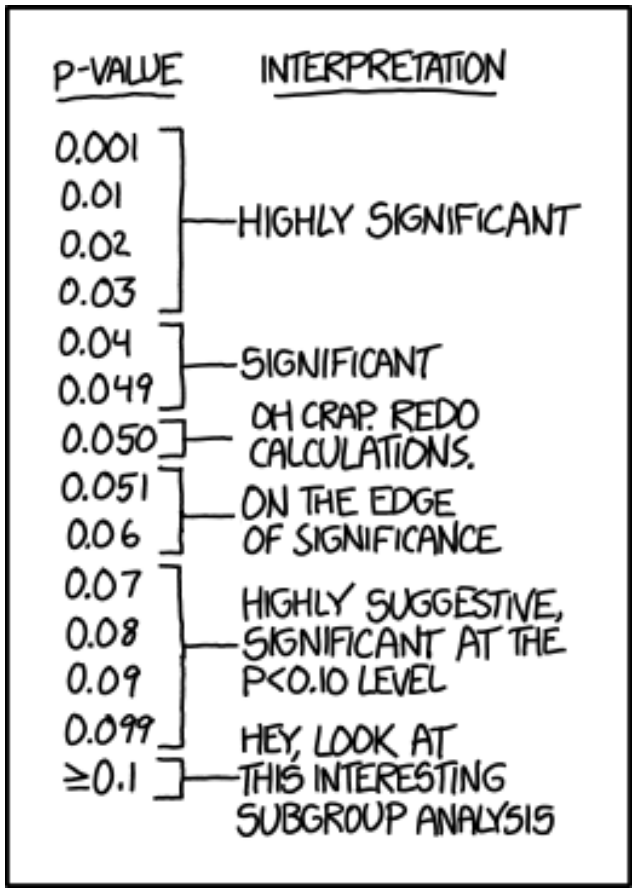
Cumming, G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science, 2008, 3, 286-300

Common mistakes on the p-value

Some researchers compared the mean in two groups (treatment and control) and found the p-value to be 0.03.

- If the two population means were identical (null hypothesis), there is a 3% chance of observing a difference as large as you observed (or larger).
- Random sampling from identical populations would lead to a difference smaller than what you observed in 97% of the experiments, and larger than you observed in 3% of the experiments.
- There is a 97% chance that there is a real difference between the two populations and 3% chance that the difference is a random coincidence.
- The p-value is the probability that the result is due to sampling error.
- The p-value is the probability that the null hypothesis is true.
- The probability that the alternative hypothesis is true is not $1 - pval$.
- The probability that the experiment will hold up when repeated is not $1 - pval$.
- A high p-value does not prove that the null hypothesis is true.

P-values


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Outline

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	True State of Nature: The null hypothesis is false
Decision	We decide to reject the null hypothesis	Type I error (rejecting a true null hypothesis)	Correct decision
	We fail to reject the null hypothesis	Correct decision	Type II error (failing to reject a false null hypothesis)

Some concepts

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

Decision	*True Condition*	
	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Incorrect decision **Type II error**
Reject H_0	Incorrect decision **Type I error** Significance level, α, = P(Type I error)	Correct decision Power of the test = 1 – P(Type II error)

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

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.
- 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

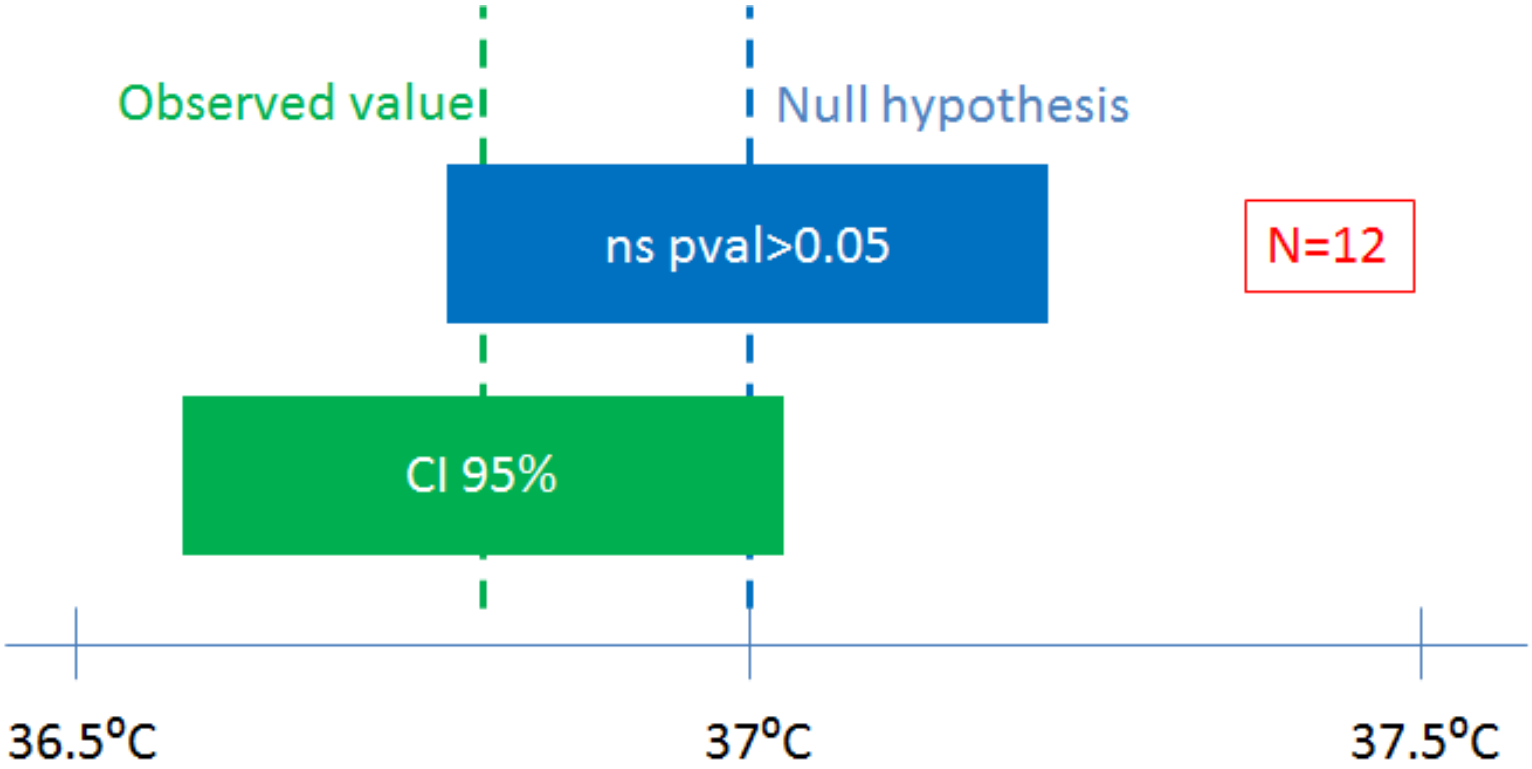
- CIs compute a range that 95% of the time will contain the population value (given some assumptions).
- Hypothesis testing computes a range that you can be 95% sure would contain the experimental results if the null hypothesis were true. Any result within this range is considered not statistically significant, 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 $p < 0.05$).

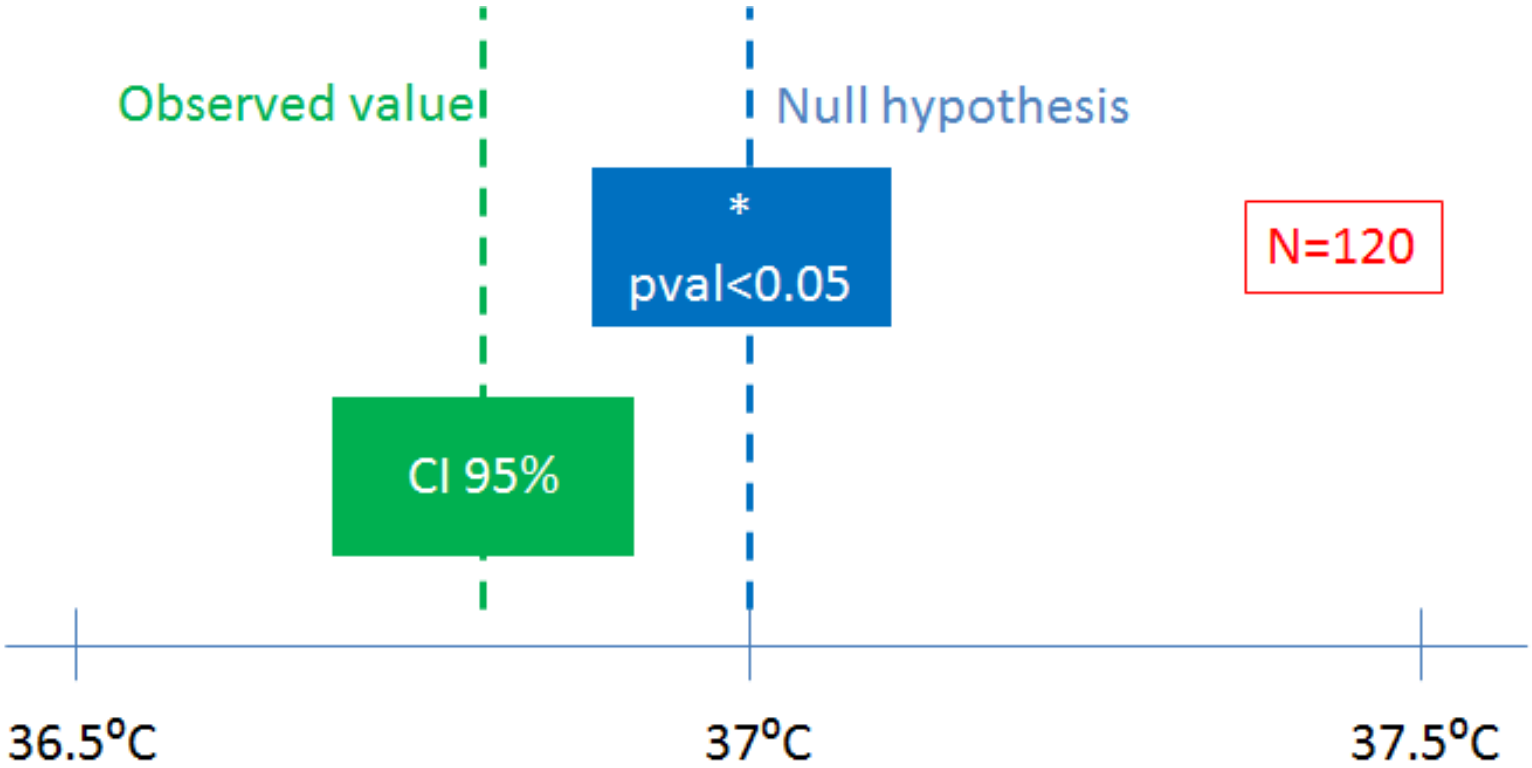
CIs and hypothesis testing

With $N = 12$ measurements we observe some difference between the average observed temperature and the reference (null) value (37°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

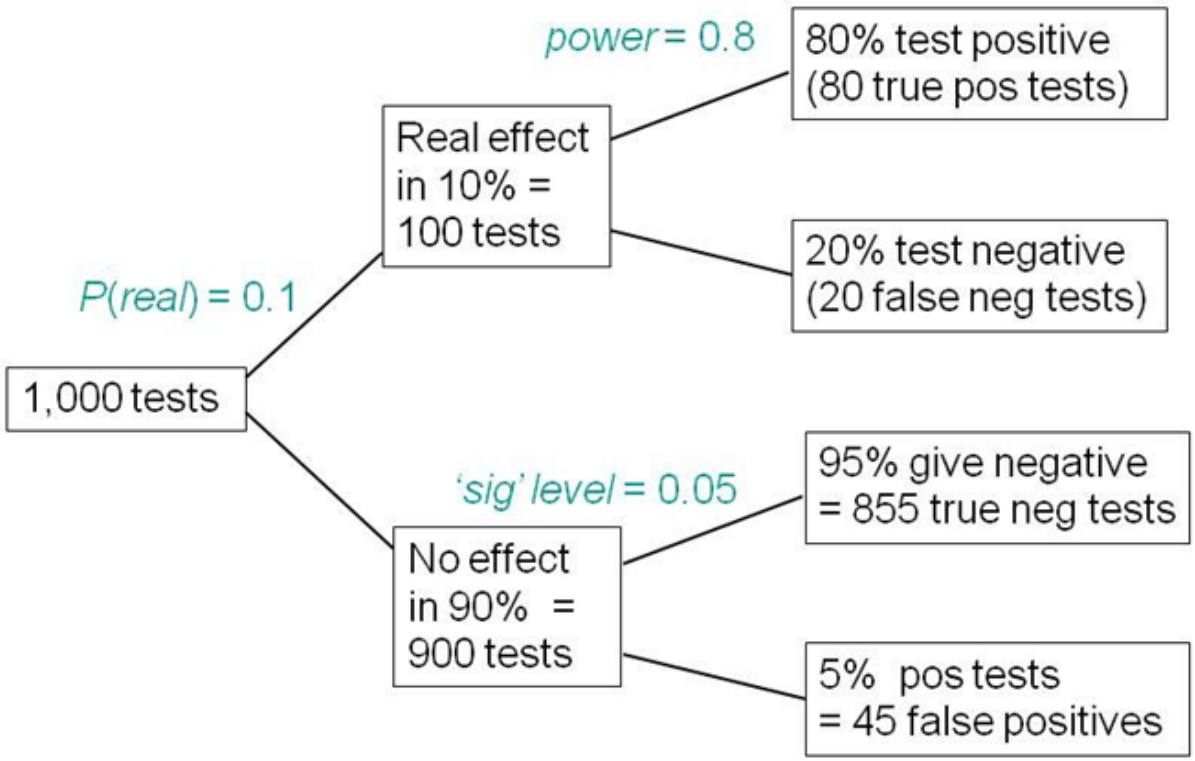
We compare the responding proportion in a control and treatment group

Sample size per group	Control	Responding	pval	CI 95%
10	10%	80.0%	0.006	[44.39,97.48]%
100	10%	26.0%	0.006	[17.74,35.73]%
1000	10%	14.1%	0.006	[12.00,16.41]%
10000	10%	11.2%	0.006	[10.59,11.83]%

They all have the save p-value, but their relevance are rather different (e.g., the last one is seldom interesting, the effect is too small).

Significance level, power and false discovery rate

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

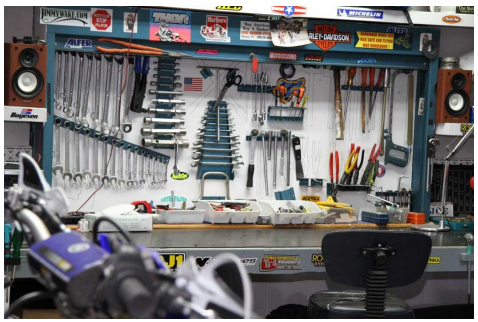
$$FDR = \frac{A}{A + C} = \frac{45}{125} = 0.36$$

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.

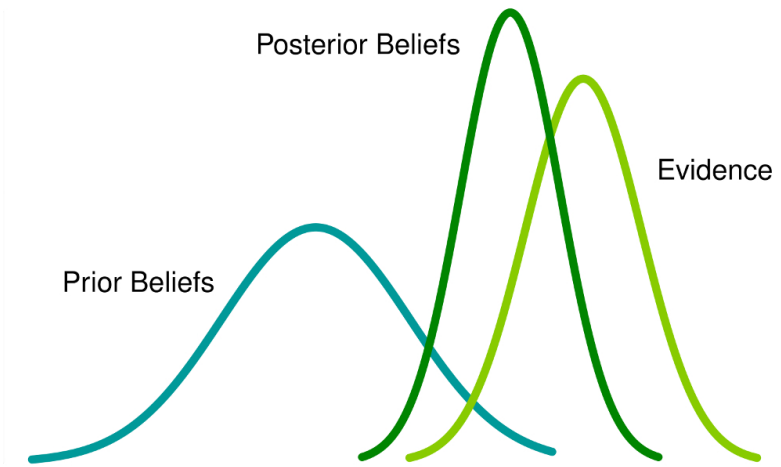


You send your child into the basement to find a tool. He comes back and says "It isn't there". What do you conclude? Is the tool there (H_0) or not (H_1)?
Your conclusion depends on:

- 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)

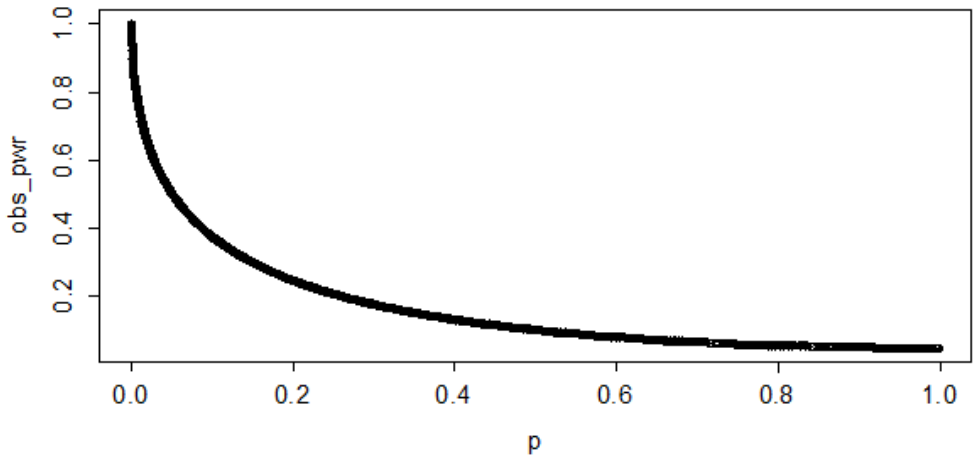
Informal accounting for prior probabilities

- 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)*).



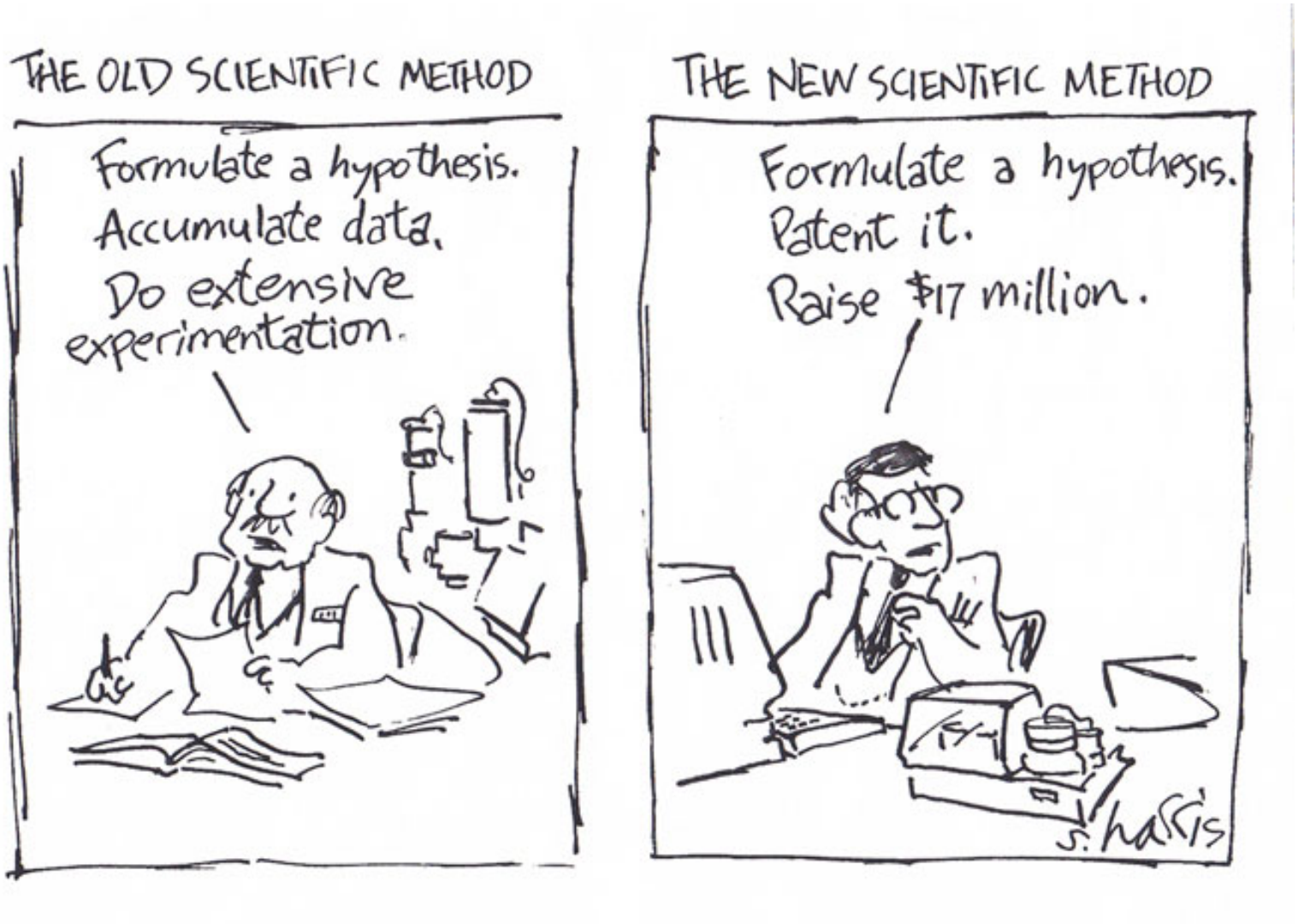
Post-hoc power analysis (Don't)

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.



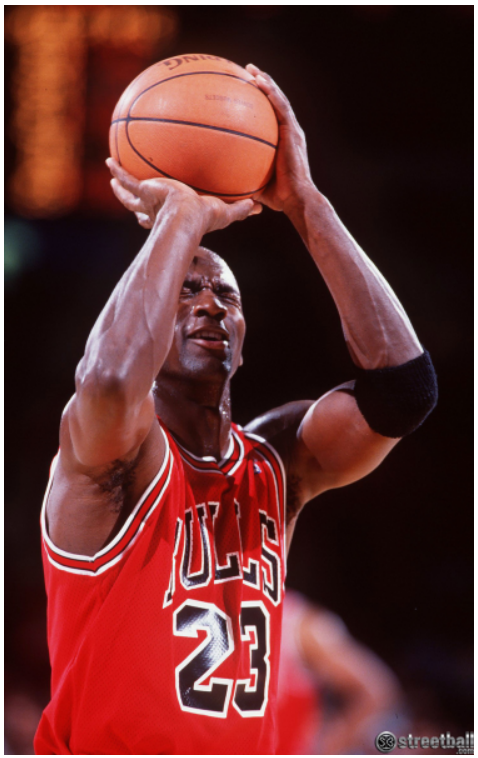
Unfortunately, post-hoc power is simply another way of reporting the p-value. There is a close relationship between the observed power and the observed p-value. If you want to look at your experiment retrospectively, look at the CI.

Hypothesis testing



Outline

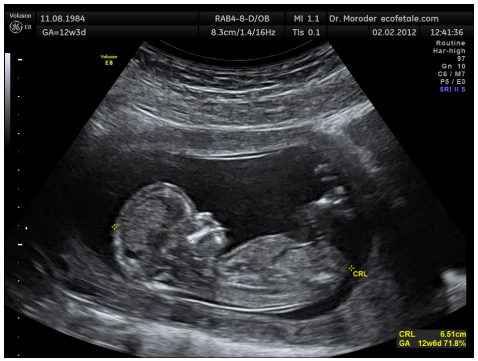
Not significant results



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.

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?

P-values


Believe me..! P value greater than
0.05 indicates chance of your
drowning is not significant.

Outline

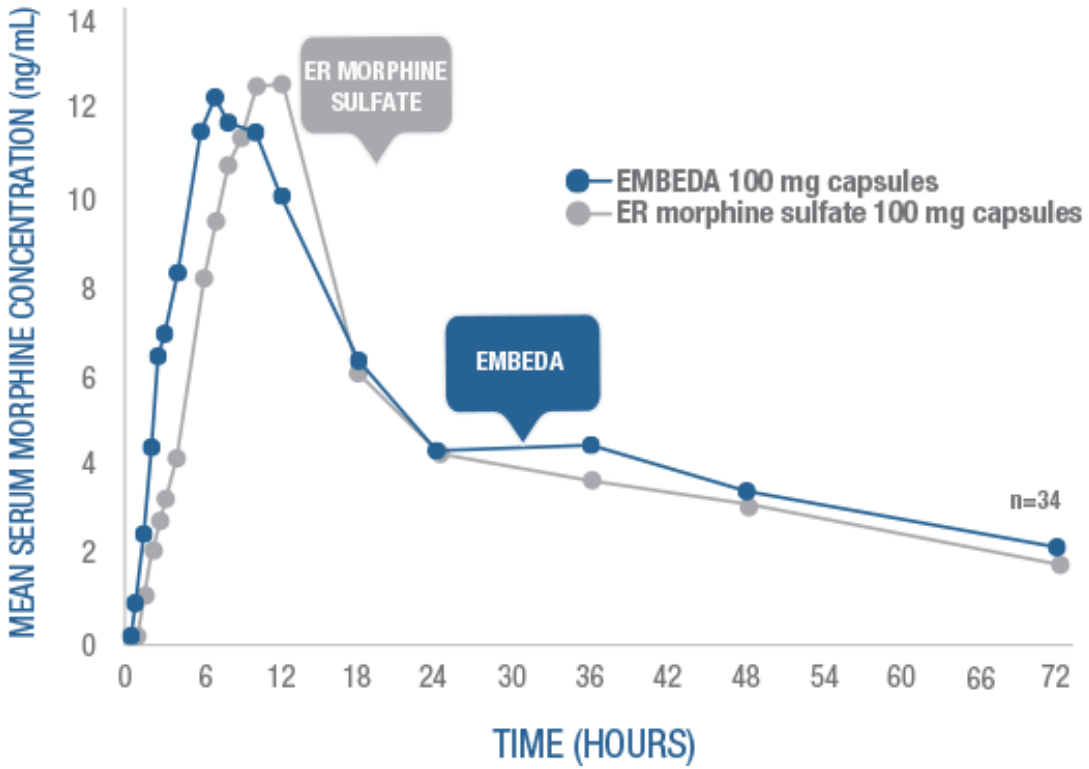
Significance vs. equivalence tests

- Significance tests:

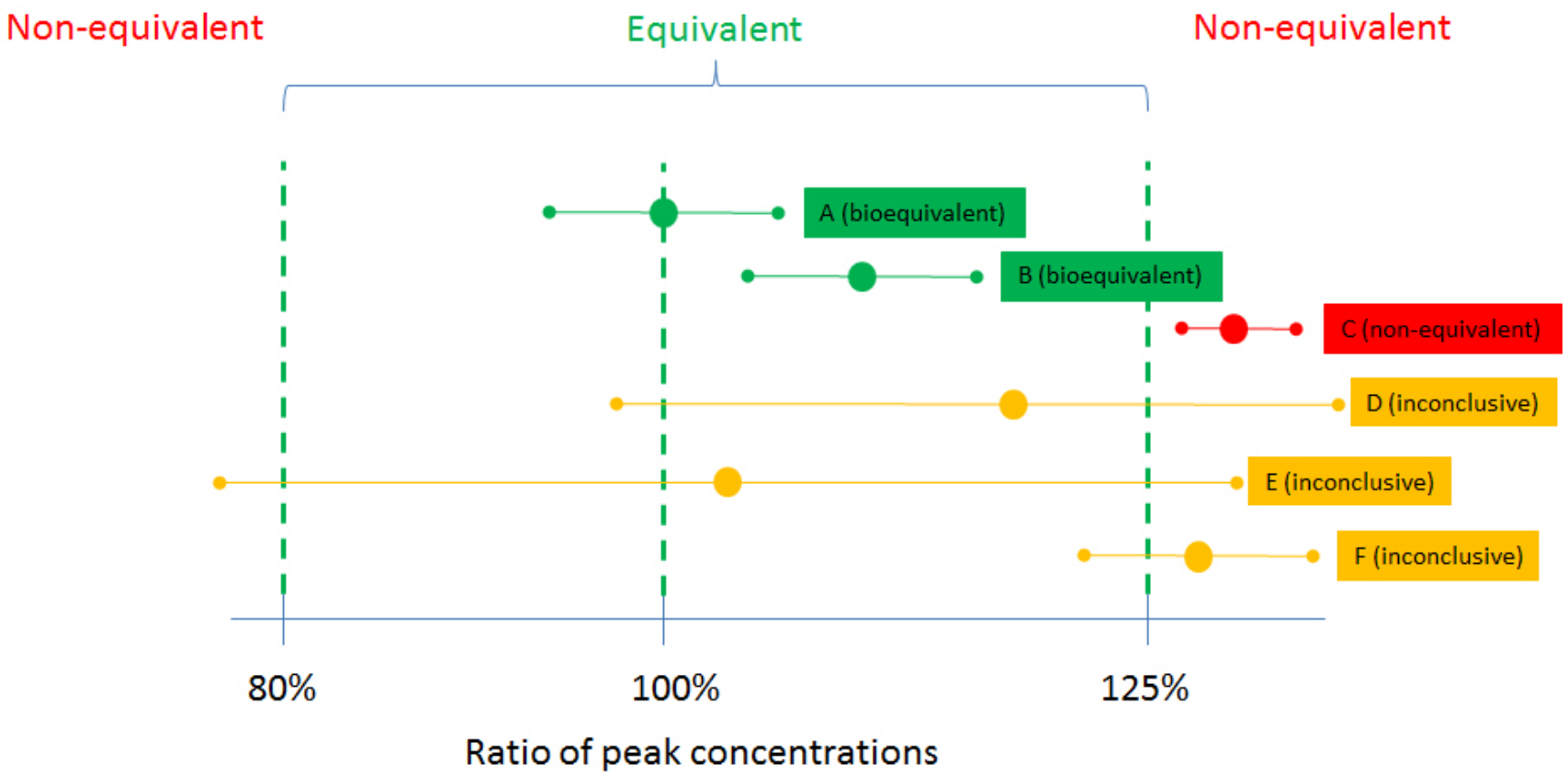
$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

- Equivalence tests:

$$H_0 : \mu_1 \neq \mu_2$$
$$H_A : \mu_1 = \mu_2$$



Use CI to determine equivalence



To verify bioequivalence check that the 95% CI is within the bioequivalent area. Standard significance testing is not valid. Equivalence tests translate into two significance tests ($H_0 : R < 0.8$ and $H_0 > 1.25$).

Statistical truth

"Well, I'll be damned if I'll defend to the death your right to say something that's statistically incorrect."

Outline

Chapter 4. Statistical assumptions

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

August 12, 2017

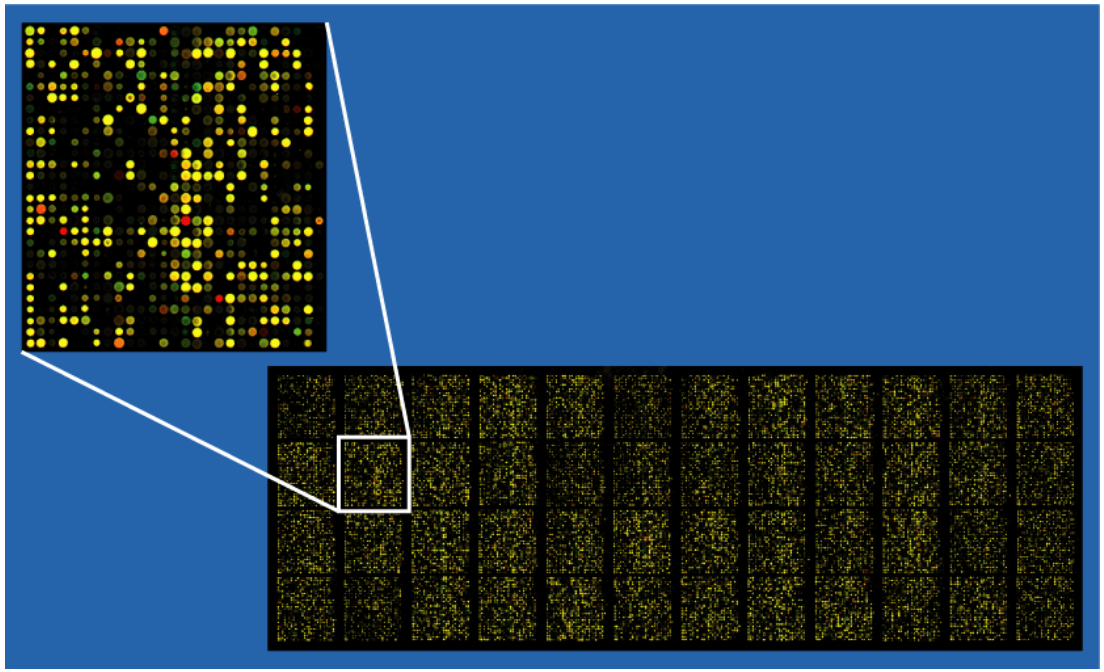
Outline

Outline

4. Statistical assumptions
   - Small number of tests
   - Gaussian distribution
   - A single population (no outliers)

Small number of tests

If the fishing expedition catches a boot, the fishermen should throw it back and not claim they were fishing for boots. (James Mills)



In an experiment with DNA microarrays, 20.000 genes are tested for association with disease, condition, etc. If the confidence level is 0.05, on average, even if no gene is related to the disease, 1.000 genes will be identified as related to it.

Small number of tests

Corrections for multiple, K, comparisons:

- Bonferroni: Lower the confidence level to α/K. E.g., a gene is identified as related to the disease if its p-value is below

$$\frac{0.05}{20000} = 2.50e-6$$

  The family-wise confidence level is still 0.05. Bonferroni is sometimes too conservative.

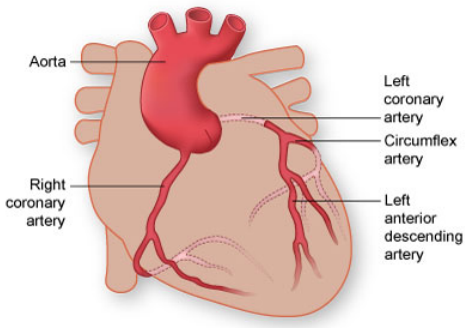
- Benjamini and Hochberg FDR: Sort the p-values in ascending order. E.g., the i-th gene is identified as related to the disease if its p-value is smaller than $i\frac{\alpha}{K}$. For example,

p-value	i	Threshold	Significant
1e-9	1	$2.50e-6$	Yes
1e-8	2	$5.00e-6$	Yes
1e-7	3	$7.50e-6$	Yes
1e-6	4	$1.00e-5$	Yes
1e-5	5	$1.25e-5$	Yes
1e-4	6	$1.50e-5$	No
1e-3	7	$1.75e-5$	No
...	...	...	No

Multiple subgroups

We study the effect of drugs A and B on the survival time of patients with coronary heart disease. We may analyze the data as

- Comparison between 2 groups: A and B
- Comparison between 2 groups within subgroups of patients depending on the number of arteries with disease, the ventricle contraction and ECG findings

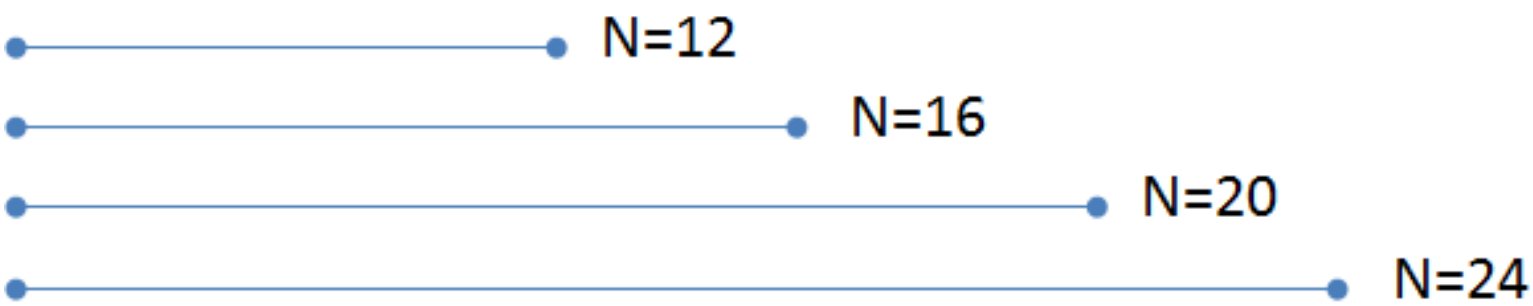


#Arteries	Ventricle	ECG
1	Normal	Normal
1	Normal	Abnormal
1	Abnormal	Normal
1	Abnormal	Abnormal
2	Normal	Normal
2	Normal	Abnormal
2	Abnormal	Normal
2	Abnormal	Abnormal

The number of comparisons starts to be high for not making any multiple testing correction and we may find a small p-value in one of the groups just by chance.

Multiple sample sizes

If you run an experiment and the result is not quite significant, it is tempting to add a few more samples to the experiment and test again. You repeat this procedure until the result is significant



The problem with this approach is that you keep collecting data only when the result is not statistically significant and stop when the result is statistically significant. If the experiment were continued a little bit longer, you might be back to not statistically significant, but you will never know because you stopped.

Lookup sequential data analysis for a rigorous way of carrying out an experiment by adding samples (the trick is to change the significance level at each comparison).

Multiple geographical areas

5 children in a particular school developed leukemia last year. Is that a coincidence? or does the clustering of cases suggest the presence of an environmental toxin that caused the disease?

What is the probability that 5 children in a particular school would all get leukemia this year?
We may estimate this probability if we know the overall incidence rate of leukemia in children and the number of children enrolled in the school (Binomial distribution). The probability is very low and parents are alarmed.

But you have asked the wrong question once you have observed the cluster. The school only came to your attention because of the cluster of cases. The right question is

What is the probability that 5 children in any school would all get leukemia this year?
You would have to define the geographical area to include and the number of schools, size of schools, ... but, this probability is much higher.

Multiple secondary outcomes

- When a clinical trial is designed, there must be some clearly defined primary outcomes (variables we care the most, and where the statistical analysis will be focused).
- During the clinical trial, we will measure many other variables. But they should be treated as secondary outcomes (they may strengthen the scientific argument of the primary outcomes and lead to new hypotheses to study). But, if you measure many secondary outcomes, you should expect some of them to be significantly different between groups just by chance (Type I error).
- If a better understanding of the disease is achieved during the clinical trial, we may change the primary outcomes, but without looking at the data first.
- You cannot change your primary outcomes after looking at the data and choosing those variables with lowest p-values.

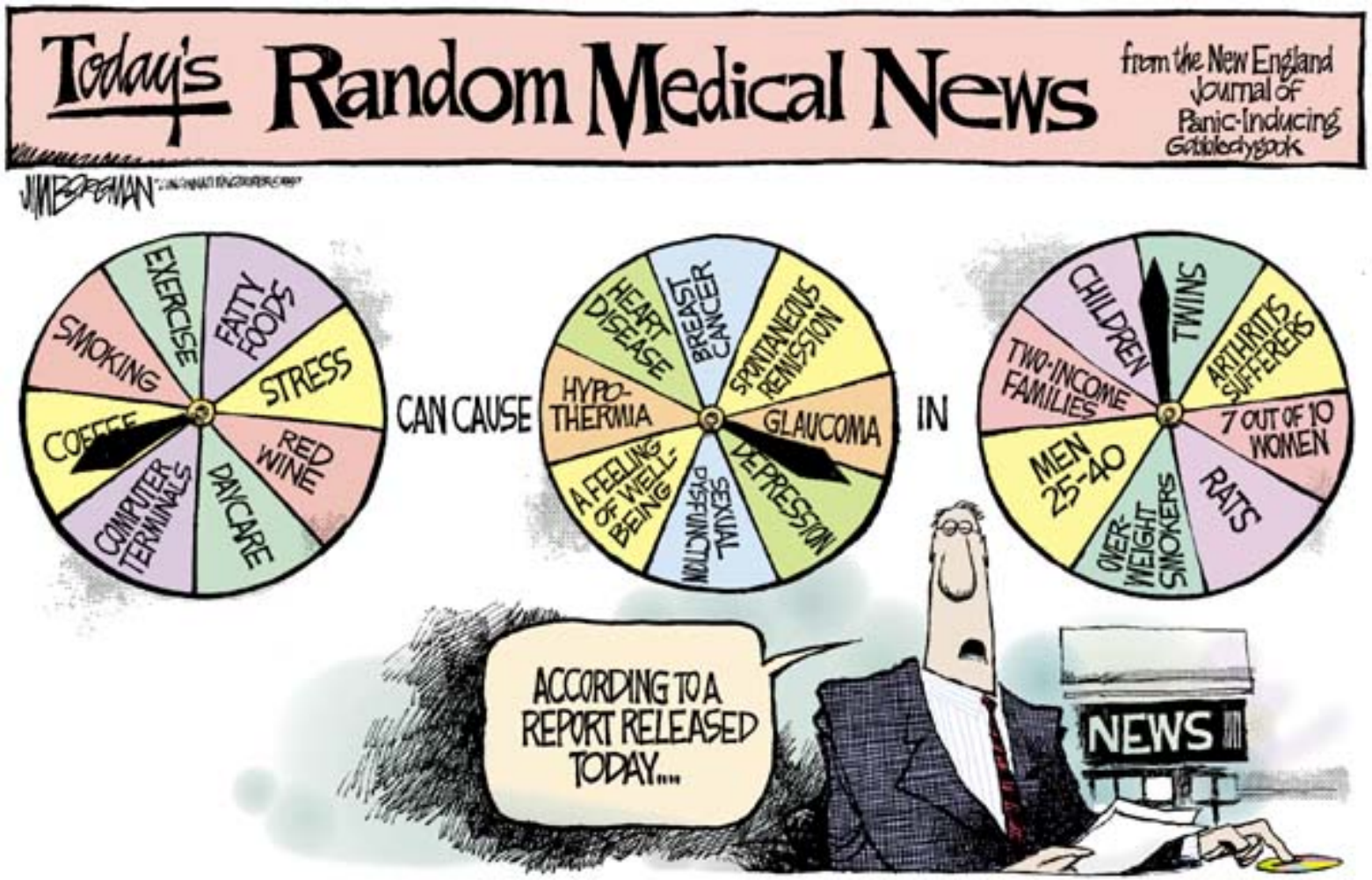
Multiple ...

You shall not ... after looking at the data

- ... decide the definition of groups (comparing drugs A and B in Pisces patients; if you inspect many different ways of grouping, some of them will be significant by chance)
- ... choose the important variables for regression (drug response as a function of mother's age and Real Madrid score that week; if you inspect many different combinations of predictors, some of them will be significant by chance)
- ... preprocess the data in multiple ways (smoothing, outlier rejection, logarithmic transformation, ...; if you inspect many different preprocessing schemes, some of them will be significant by chance)
- ... analyze the data in multiple ways (testing all analysis possibilities of a program; if you try many different hypothesis tests, some of them will be significant by chance)

Be skeptical of results obtained by data torture or p-hacking.

Multiple testing



Outline

4. Statistical assumptions
   - Small number of tests
   - Gaussian distribution
   - A single population (no outliers)

Gaussian distribution

Data Gaussianity is assumed by many parametrical tests:

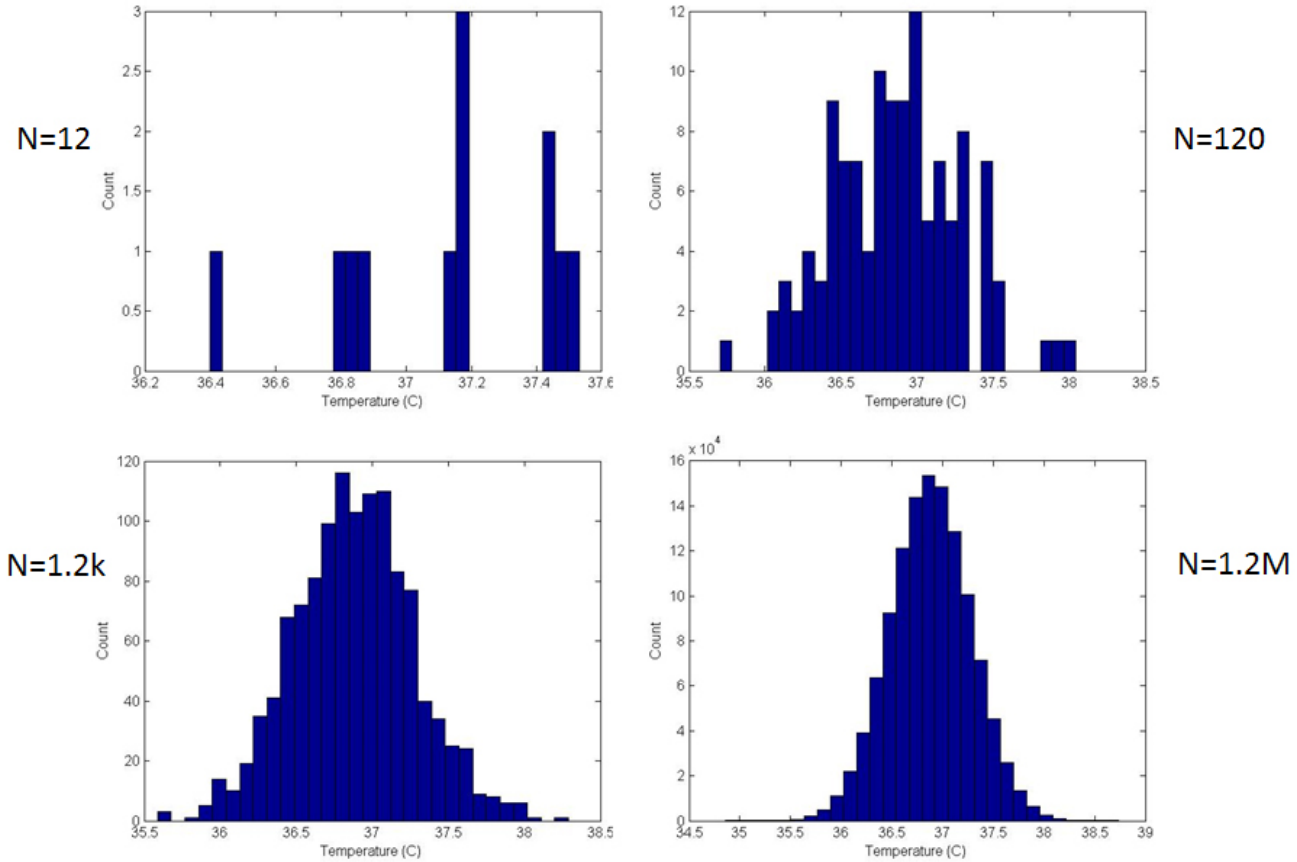
- Student's t tests (about means)
- ANOVA and Snedecor's F tests (about means and variances)
- χ^2 (depending on the source of data)
- Some tests and sample size formulas for proportions
- ...

Gaussian functions are easy to deal mathematically and they approximate well many processes (especially those that are the result of the sum of many contributions; Central Limit Theorem).

Gaussian distribution

The Gaussian is an idealization of a random process (e.g., it extends to infinity, but in practice blood pressures, weights, heights, concentrations, ... cannot). The question is whether real data is well approximated by a Gaussian distribution.

Gaussian data may not necessarily look Gaussian.



Normality tests

There are a number of statistical tests (D'Agostino-Pearson, Shapiro-Wilk, Kolmogorov-Smirnov, Darling-Anderson) that help to quantify if the data contradicts the Gaussian assumption.

- If the p-value is high, then the observed data is not incompatible with a Gaussian distribution.
- If the p-value is low, then the observed data contradicts the hypothesis that the data was drawn from a Gaussian distribution (null hypothesis).

If the data is not normal, you may:

- Transform it (taking logarithms from log-normal data (dillutions, number of cells, ...)).
- Identify and remove outliers.
- Switch to a non-parametric test that does not assume normality.
- Ignore small departures from the Gaussian ideal (many parametric tests are robust to mild violations).

Normality tests

Should I use a normality test to decide whether to perform a parametric or non-parametric analysis?

- Make sure that the data is not log-normal.
- Make sure that there are no outliers.
- The decision on parametric or non-parametric is most important with small sample sizes, but with small sample sizes most normality tests cannot show that the data is not Gaussian (high p-values). This gives a false confidence on the use of parametric analysis.
- Remind that in a long term analysis, data should be analyzed in the same way.
- Remind that non-parametric tests are less powerful than parametric ones.

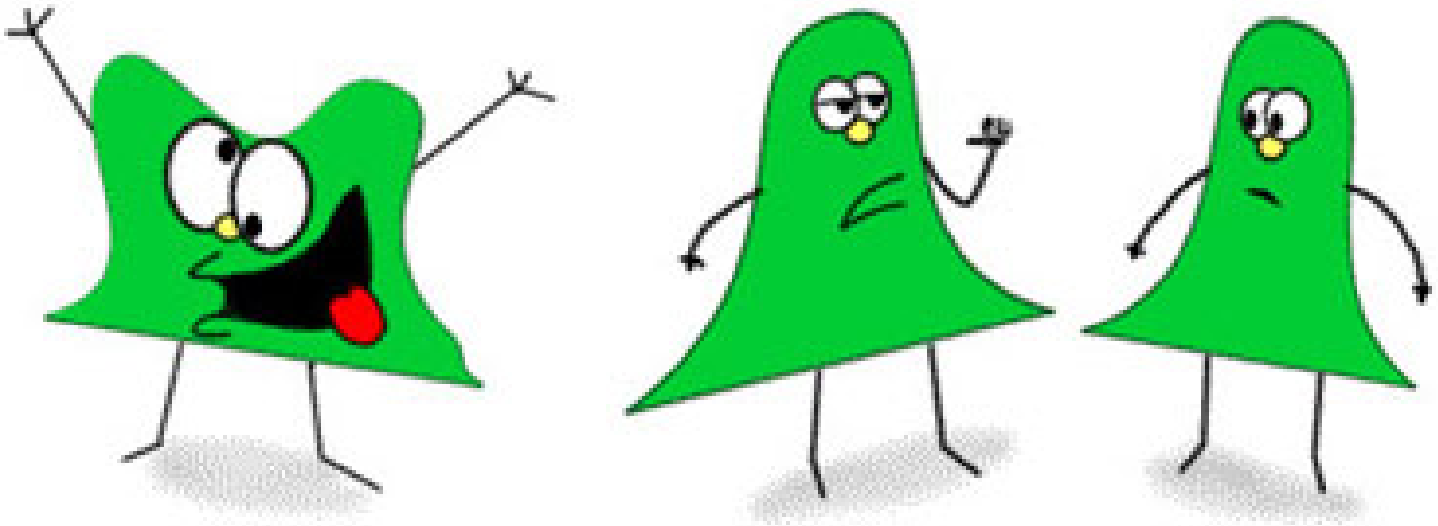
Overall, the decision to go parametric or not is hard and requires experience, thinking and perspective.

Non-parametric tests

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

To choose the right statistical test, visit this Statistical test selection guide.

normality



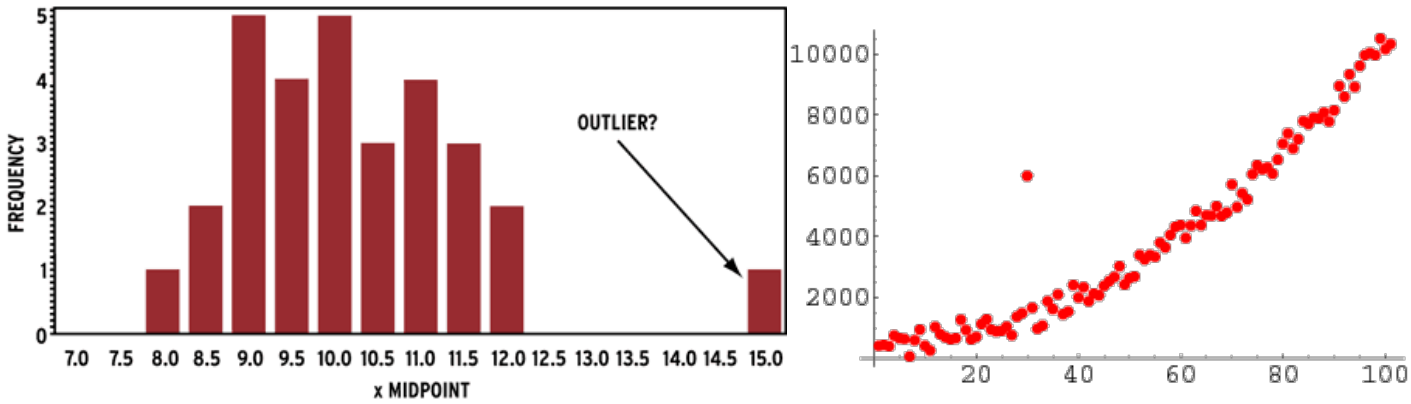
"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW...NORMAL!"

Outline

4. Statistical assumptions
   - Small number of tests
   - Gaussian distribution
   - A single population (no outliers)

A single population

Tests assume that the observed data come from a single population. Outliers seem to come from a different population.



It is an outlier if it comes from

- Invalid data (transposed digits, shifted decimal point, sensor blackout, ...)
- Experimental mistake (bad pipetting, a voltage spike, a hole in a filter)

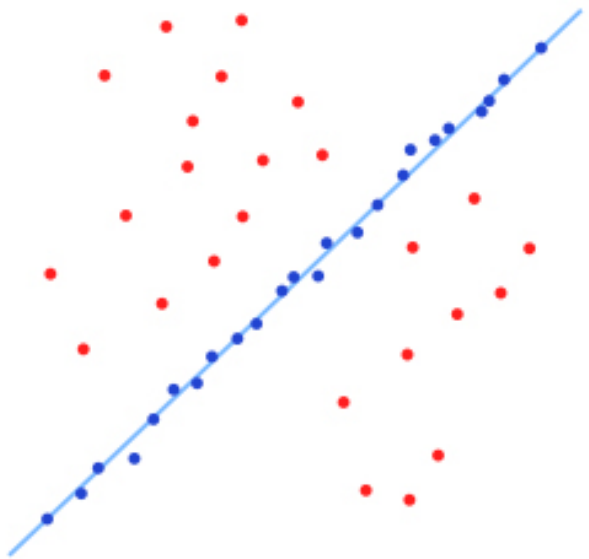
It is not an outlier if it comes from

- Random chance (just by chance some values are larger/smaller than rest)
- Biological diversity (the population is really variable)
- Invalid assumption (I assume it is normal, but it is log-normal)

Is it legitimate to remove outliers?

- Removing data because it does not fit our "expectations" is cheating.
- Leaving outliers may lead to invalid results, it is another way of "cheating".
- It is not cheating when the decision to remove an outlier is based on rules and methods established before the data was collected.

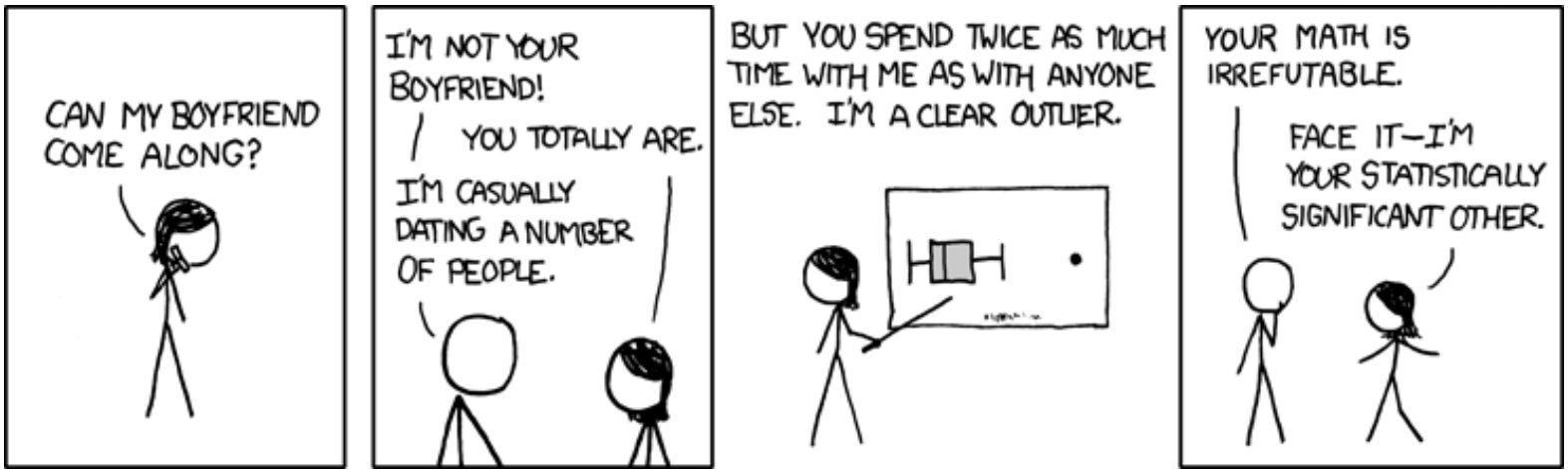
Alternatively, use robust statistics



Common mistakes with outliers

- Not realizing that the data is log-normal, instead of normal.
- Using a test designed to detect a single outlier when there are several outliers. Applying multiple times a test to detect a single outlier does a poor job.
- Eliminating outliers only when you don't get the results you want
- Truly eliminating outliers from your notebook.

Outliers


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

Outline

Chapter 5. Statistical tests

C.O.S. Sorzano
coss@cnb.csic.es

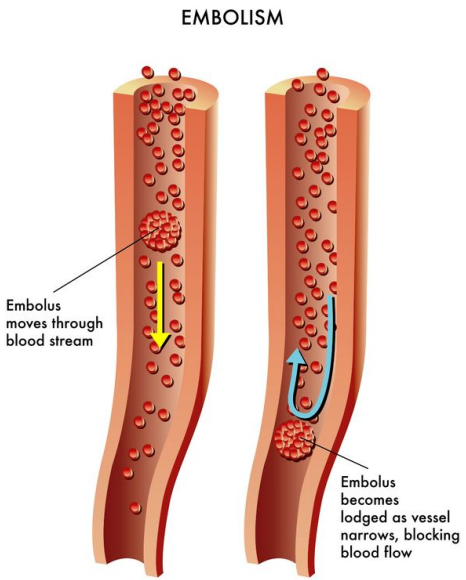
National Center of Biotechnology (CSIC)

August 15, 2017

Outline

Outline

Comparing two proportions



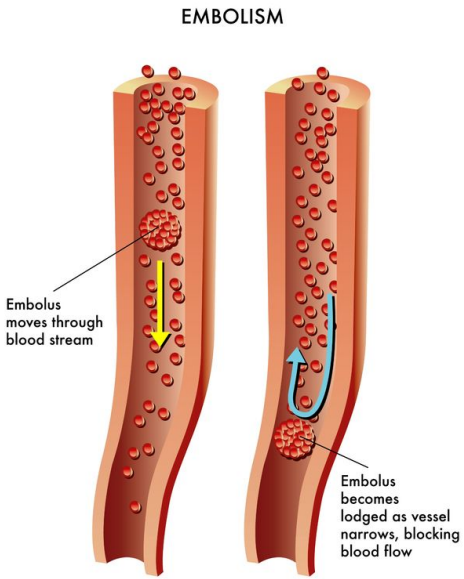
A research group tested whether apixaban 2.5 mg, twice a day reduced the recurrence of thromboembolism. They randomly assigned 1669 patients to a placebo or treatment and checked the number of thromboembolisms after 1 year. The results were (contigency table)

	Recurrence	No recurrence	Total
Placebo	73	756	829
Treatment	14	826	840
Total	87	1582	1669

Fisher's exact test showed that the proportion of recurrence in both groups was different with a p-value<0.0001. This p-value answers the question: if the null hypothesis were true ($H_0 : p_{apixaban} = H_{placebo}$, apixaban does not have any effect on the thromboembolism recurrence), what is the probability of observing 14 or less recurrences out of 840 patients?

Fisher's exact test with large samples is difficult to calculate and can be approximated by a χ^2 test.

Comparing two proportions



The 95% confidence intervals for the proportion of recurrence in each group were

- Placebo: p=73/829=8.8%, 95% CI=[7.0,10.9]
- Treatment: p=14/840=1.7%, 95% CI=[0.9,2.8]

The difference between 8.8 and 1.7 (=7.1%) is called the attributable risk.

The relative risk is the ratio between the two, RR=1.7/8.8=0.19 (95% CI=[0.11,0.33]). This means that the treatment reduces the risk of recurrence by a factor between 0.11 and 0.33.

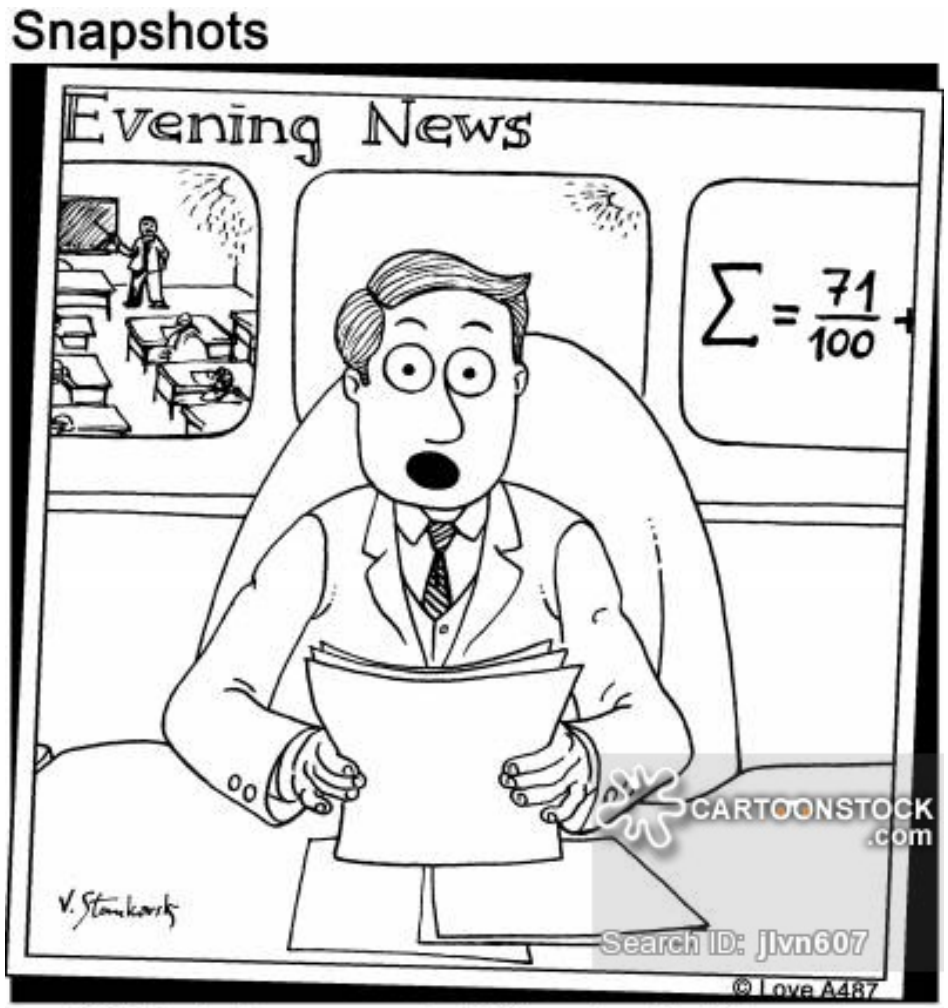
Secondary results:

- This reduction was similar by age and sex.
- Patients receiving the treatment did not have more bleeding than those with the placebo.

Assumptions

- Random sample: The sample is representative of the whole population. The subjects in the study were not randomly selected (they may come from the same hospital), but they were randomly assigned to receive the drug or placebo.
- Independent observations: Selecting one member of the population should not change the chance of selecting anyone else, and the results of one person is not correlated to the result of any other person in the study. This is violated if the study included several members of the same family.
- No difference between the two groups except treatment: The researchers had to show that there was no significant difference in terms of age, weight, sex, kidney function, etc.

Comparing two proportions

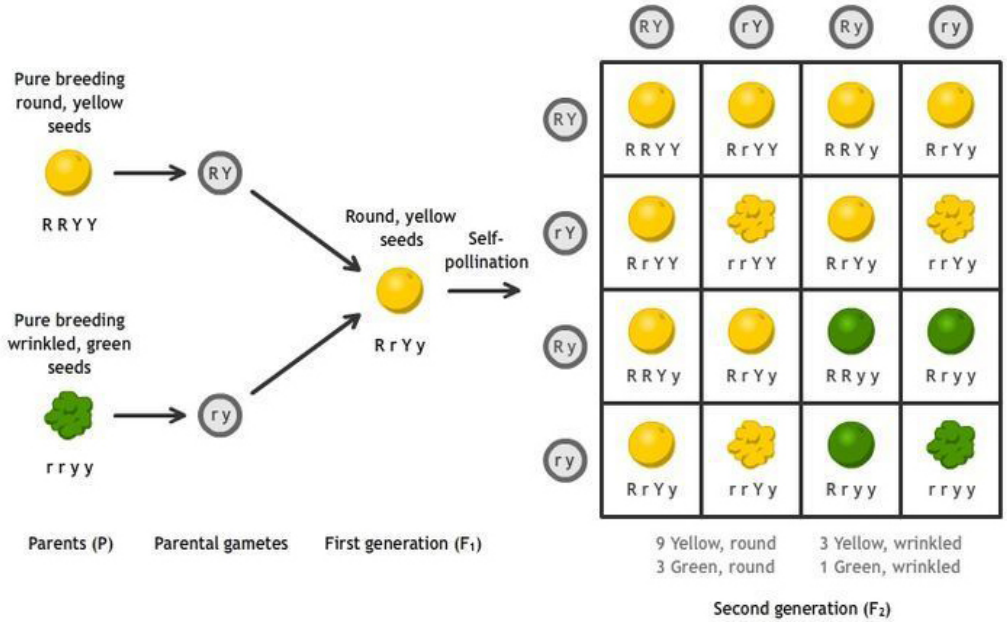


"Of students surveyed, 64% prefer English and 32% prefer math. The fact that these numbers do not add up to 100 may help explain why."

Outline

Comparing proportions to theoretical values

Mendel studied the shape and color of peas.



In an experiment with 556 peas, these are the observed and the expected results

Phenotype	# Observed	Expected proportion	# Expected
Round and yellow	315	9/16	312.75
Round and green	108	3/16	104.25
Wrinkle and yellow	101	3/16	104.25
Wrinkle and green	32	1/16	34.75

Does the observed data contradict our theory?

Comparing proportions to theoretical values

The correct way of analyzing a multinomial distribution (9/16,3/16,3/16,1/16) is by the exact test of goodness-of-fit. Its mathematics are relatively complicated and if the number of samples is large enough it can be approximated by χ^2 test of goodness-of-fit.

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

The number of degrees of freedom is $N - 1$ where N is the number of categories (in the peas example $N = 4$). In the example $X^2 = 0.470$ and the p-value is 0.93, so the observed data does not contradict our theory.

For $N = 2$, the exact test of goodness-of-fit becomes the binomial test.

Common mistakes

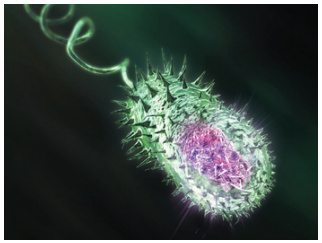
- Mixing the two approximations by χ^2. It is not the same comparing two groups than comparing one group to theory.
- Constructing observed values that are real counts but "normalized" counts (normalized to 1, 100 or 1000).

Comparing proportions to theoretical values

"I can't understand why the whole audience hated my Simpson's Paradox joke. I tried it on the men and the women in the crowd separately and each group loved it!"

Outline

Case-control studies



Some researchers investigated whether cholera was effective.

- Prospective approach: 1) Recruit unvaccinated people. 2) Randomly assign them to be vaccinated or not. 3) Follow both groups for many years to compare the incidence of cholera. But this study would require many years and would withhold the vaccine from many people.
- Retrospective approach (case-control): Pick cases and controls and annotate whether they had been vaccinated or not.

The results of a case-control study on cholera vaccination were

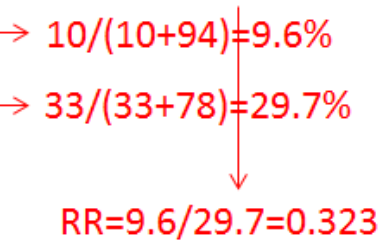
	Cases (cholera)	Controls
Vaccinated	10	94
Unvaccinated	33	78
Total	43	172

Fisher's exact test is still applicable. The null hypothesis is that the variables Vaccination and Disease state are independent. The corresponding (two-sided) p-value was 0.0003. The two-sided implies that we don't know whether cholera incidence will be higher or lower in the vaccinated group.

Case-control studies

Note that the number of controls is chosen by the researchers, so it does not make sense to calculate the relative risk (analysis by rows) as we did with the prospective study on apixaban.

	Cases (cholera)	Controls
Vaccinated	10	94
Unvaccinated	33	78
Total	43	172



Case-control studies

Instead we must calculate the odds ratio (analysis by columns)

	Cases (cholera)	Controls
Vaccinated	10	94
Unvaccinated	33	78
Total	43	172

Meaning that vaccinated people are 25% as likely to get cholera as unvaccinated people. The 95% CI for the Odds Ratio is [0.12,0.54].

The effectiveness of the vaccine is $1 - OR = 1 - 0.251 = 74.9\%$, and its 95% CI is calculated by subtracting the 95% CI of the OR from 1, that is,
$[1 - 0.54, 1 - 0.12] = [46\%, 88\%]$.

Problems of Case-control studies

In the cholera example, the researchers picked the controls visiting homes near the area of the patient, but ...

- Controls were picked because they were of the same gender and age as the subject. They could not determine if the vaccine was more or less effective depending on the gender and age.
- Patients knew they had cholera and may remember their vaccination more vividly than controls.
- Interviewers knew who had cholera and who did not. They may inadvertently pose the questions differently leading to different responses on vaccination.
- Patients may want to help researchers, while controls just want to finish the interview.
- Patients were chosen because they attended to the Cholera Treatment Center. Patients with mild symptoms did not seek medical attention.
- Controls were at home when the interviewer came. So the study was biased towards people who stay at home a lot.

Validation of Case-control studies

In the cholera example, the researchers performed a second case-control study to validate the cholera study.

- They tested if there was an association between cholera vaccination and other patients with bloody diarrhea not caused by cholera.
- Controls were chosen in the same way as in the 1st cholera study, so they shared the same biases.
- The OR was 0.64 (95% CI=[0.34,1.18]).

Because of these problems it is recommended to be skeptical of OR between 0.33 and 3, even between 0.25 and 4. Case-control studies can be trusted if they can be repeated and make sense biologically.

Common mistakes

- Confusing relative risk with odds ratio. The relationship is

$$RR = \frac{OR}{(1 - p_0) + p_0 OR}$$

being p_0 the prevalence of the disease (the fraction of the control group that has the disease).

- Entering normalized data instead of the actual counts of observed events.
- Trying to compute an OR when one of the four values is 0.

Case-control studies

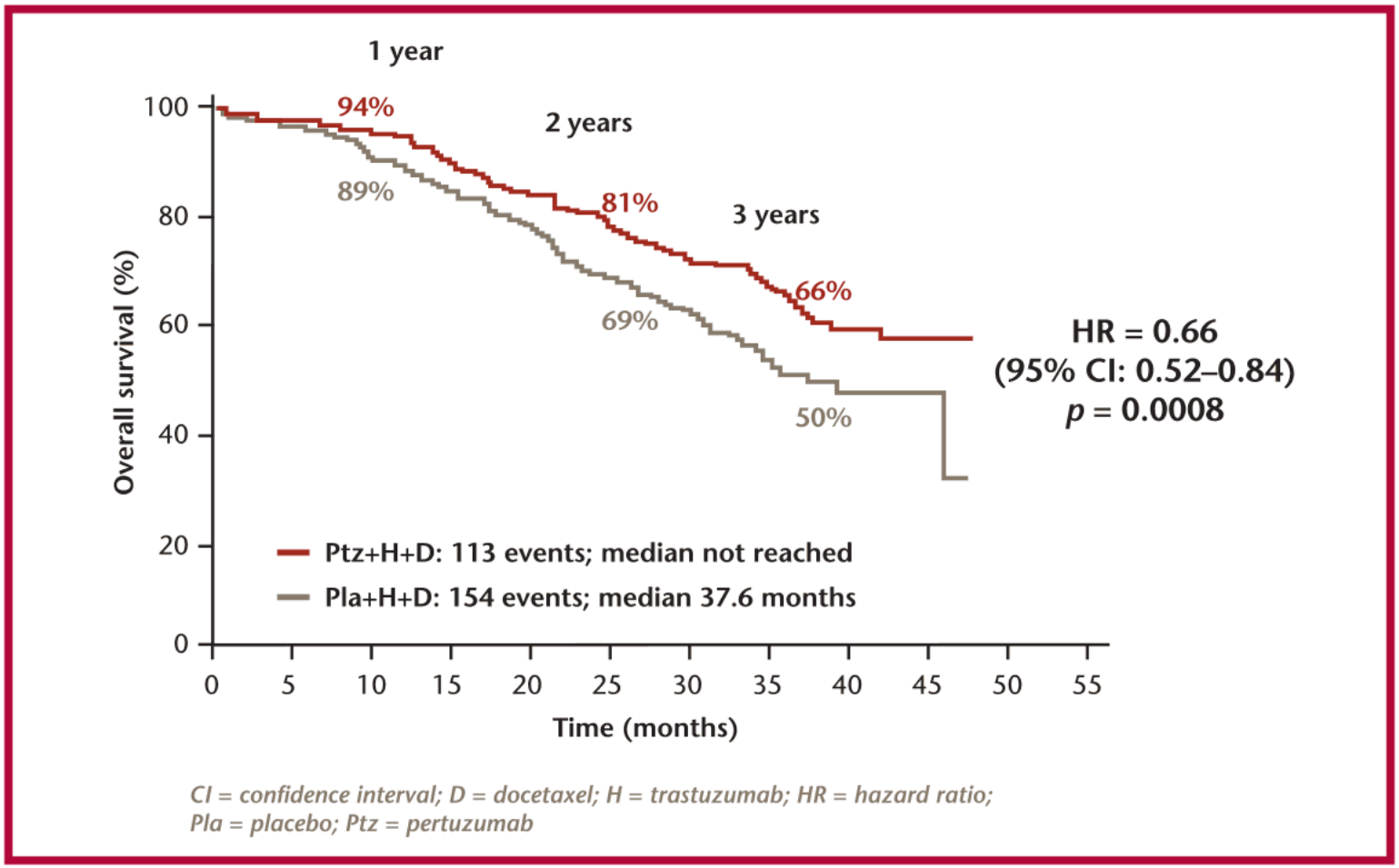
"Let's see...number of cheeseburgers eaten in a typical month? three...no, I'll put down four."

Outline

Survival data

Some researchers are comparing the survival time with two different treatments. The study spans 4 years.

Figure 1. Confirmatory overall survival analysis



CI = confidence interval; D = docetaxel; H = trastuzumab; HR = hazard ratio; Pla = placebo; Ptz = pertuzumab

Assumptions of Survival Analysis

- Random (or representative) sample.
- Independent subjects.
- Consistent entry criteria.
- Consistent definition of the end point.
- Clear definition of the starting point.
- Time of censoring is unrelated to survival.
- Average survival does not change during study.

Assumptions of Survival Analysis

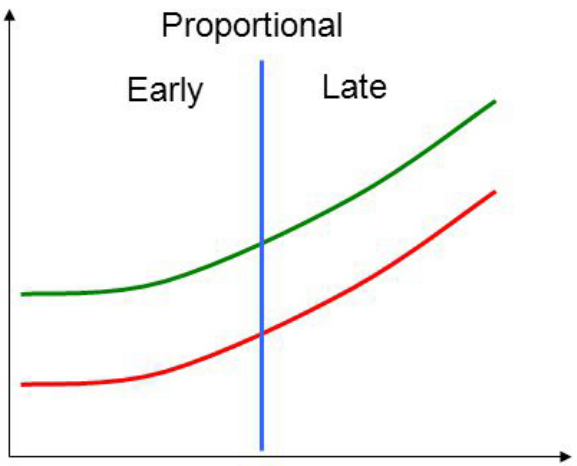
When comparing two survival curves, additionally

- Treatment groups are defined before data collection began. It is not valid to take a single group of patients, all equally treated, and split them in two subgroups depending on whether they response.
- Groups are defined consistently as data are accrued. If the study spans several years, the diagnostic groups must be defined consistently. For instance, we are comparing the survival of cancer patients with and without metastases. During the study a new scanner is acquired that is able of detecting metastases much earlier, so these patients are moved to the metastase group. The survival of the non-metastasic group improves, because the patients with small metastases are moved to the other group. But, the survival time of the metastase group also improves, because the new patients have much smaller metastases than they used to have with the old scanner. (Will Rogers phenomenon: "When the Okies left Oklahoma and moved to California, they raised the average intelligence in both states".)

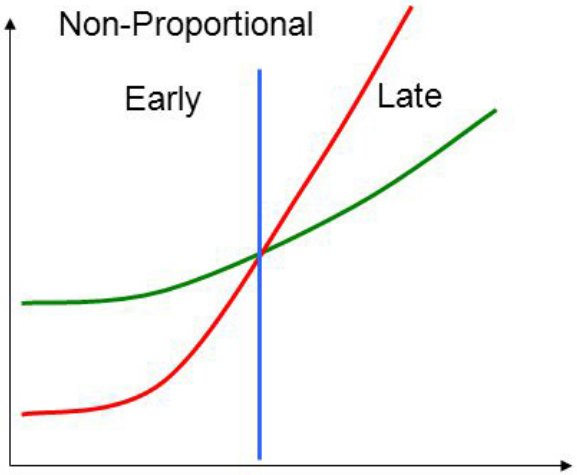
Assumptions of Survival Analysis

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).

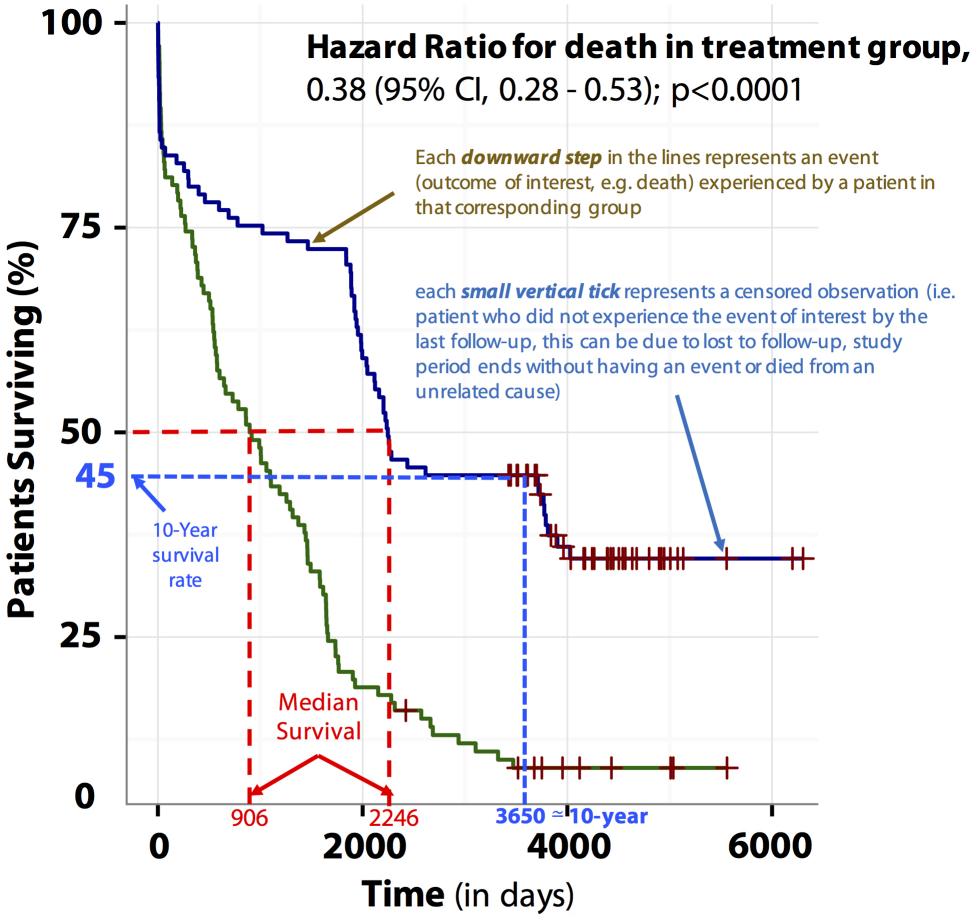


Here, the effect is the same in both time periods



Here, the effect is negative in the early period and positive in the late period

Survival Analysis

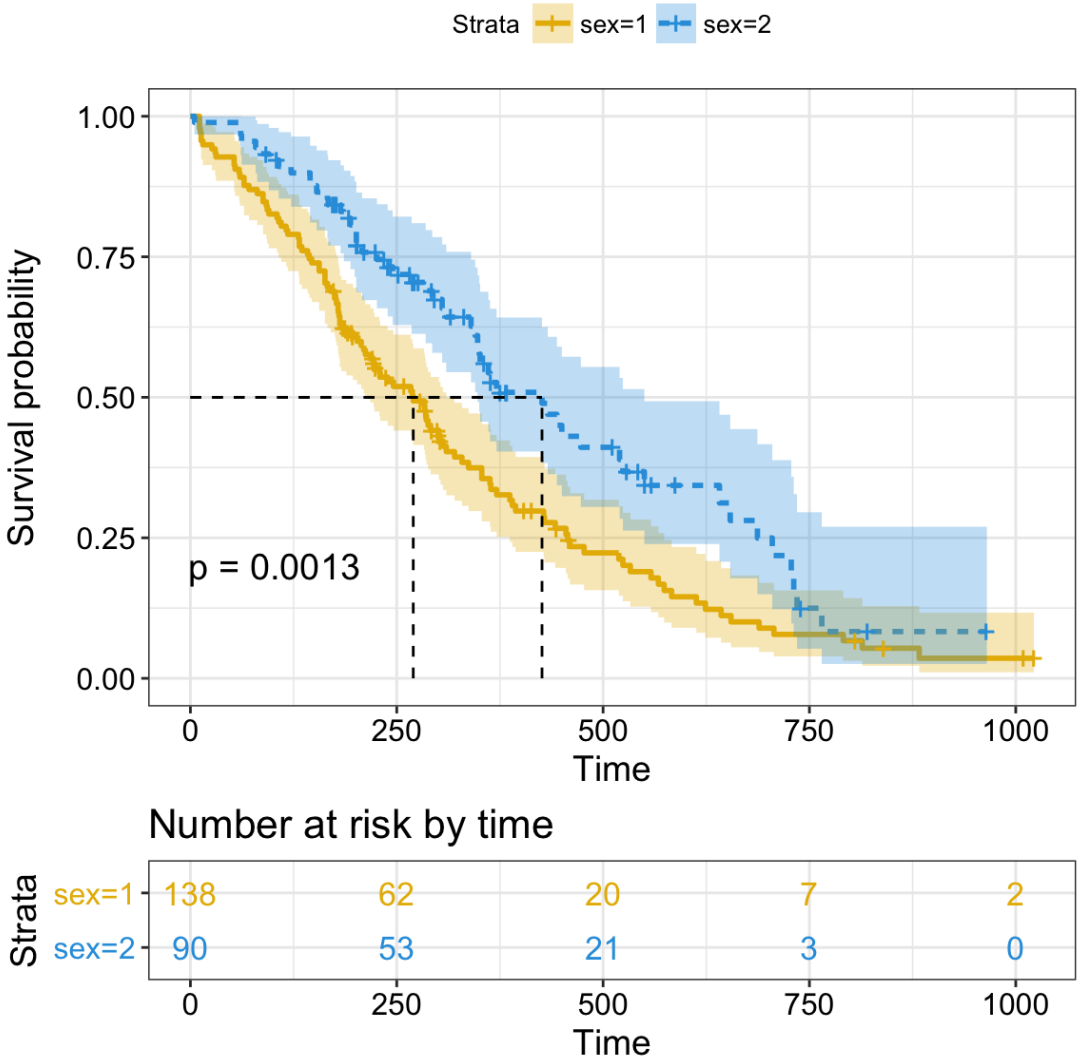


No. at Risk	Number of patients at risk shown below at regular time intervals, as times go, less people remain are at risk			
Placebo	106	20	5	0
Treatment	105	62	25	2

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, $RMST = \frac{906}{2206} = 0.41$.

Survival Analysis

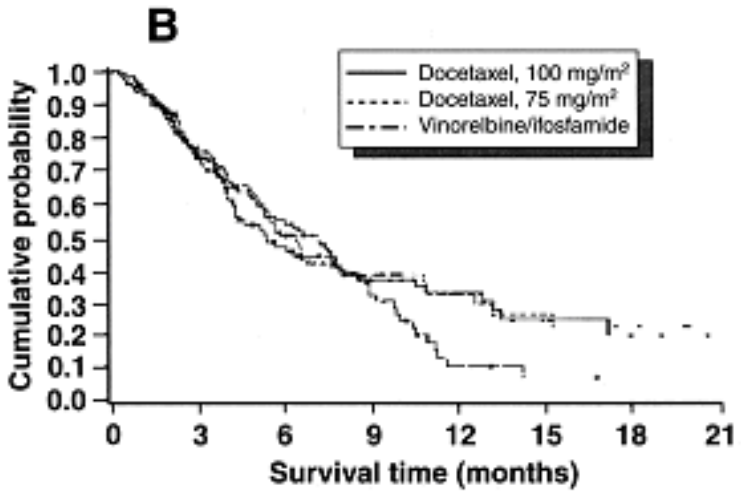


Survival bands help to decide at each time whether the two curves are significantly different.

Why Survival Analysis?

Is it not simpler to compare the mean or median survival times or five-year survival?

- Mean survival time: Cons:
  - It does not consider people that have not died or has been censored during the experiment.
  - The survival time is not Gaussian so you cannot construct useful confidence intervals.
- Median survival time or 5-year survival: They solve the problem of not all subjects dying. Cons:
  - They loose the information richness of the full survival curve. For instance, these treatments have similar median and survival at 9 months, but very different behavior after 9 months.



Intention to treat

Imagine a study that randomly assigns patients with severe angina to surgery or medical treatment. But some of the patients assigned to surgery die before being operated. Should we remove them from the analysis since they did not get the treatment? No. If we remove them from this group, we would be removing early deaths from one group (surgery), but not the other (medical treatment), and this would bias the results.

The Intention-to-Treat approach can be summarized as "analyze as randomized" even if

- Later the patient does not meet the entry criteria.
- The treatment was not given.
- They stopped the treatment for any reason.

We may also analyze the data with the intention-to-treat approach or removing the data from samples that did not receive the fully assigned treatment. If there is not a significant difference, we know that the dropouts did not affect the analysis.

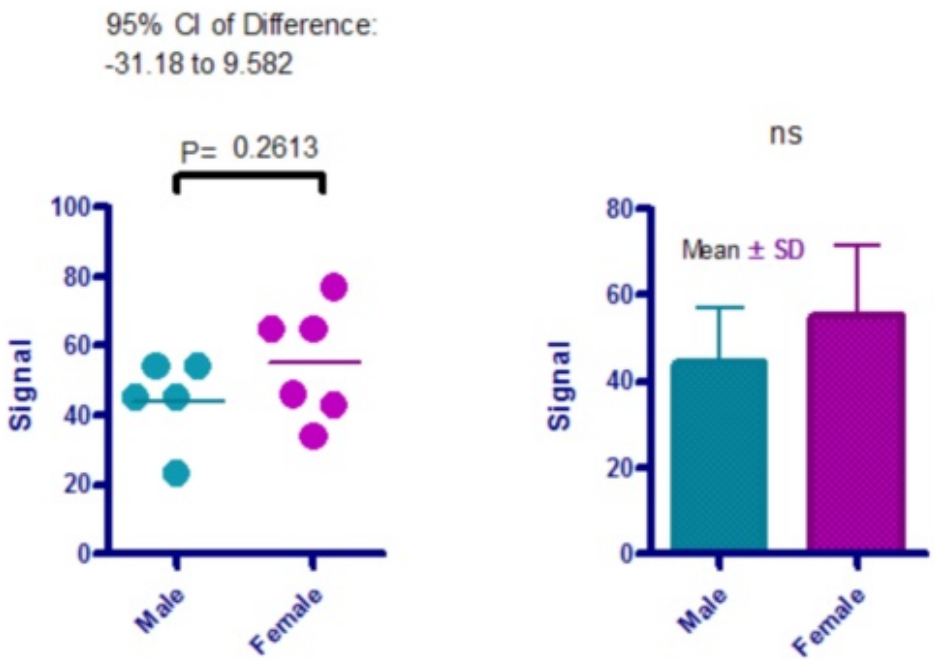
Survival data

'I'm not going to your mother's for Christmas until I know the death and survival rates'

Outline

Comparing two independent means

Some researchers are comparing a signal between two independent groups: male and female. In the example below, we see that they did not find a significant difference between both groups because the CI of the difference includes 0.0.



The fact that the mean±SD overlap is not enough for not being significantly different.

Comparing two independent means

The CI is constructed as

$$\left[\hat{d} - t_{1-\frac{\alpha}{2},df}\, s_d,\, \hat{d} + t_{1-\frac{\alpha}{2},df}\, s_d\right]$$

where

$$\begin{aligned} \hat{d} &= \bar{x}_1 - \bar{x}_2 \\ s_d &= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \\ df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \end{aligned}$$

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1}\sum_i (x_{1i} - \bar{x}_1)^2 \\ s_2^2 &= \frac{1}{n_2 - 1}\sum_i (x_{2i} - \bar{x}_2)^2 \end{aligned}$$

If we assume equal variances in both groups, the number of degrees of freedom simplifies to (this test is more powerful than the one with unequal variances)

$$df = n_1 + n_2 - 2$$

The CI is depends on

- Variability: The higher the scatter, the larger the CI.
- Sample size: The larger the sample size, the smaller the CI.
- Confidence: The larger the confidence, the larger the CI.

Comparing two independent means

The null hypothesis is that both data sets have been drawn from populations with the same mean. The p-value answers the question:

If the null hypothesis were true, what is the chance of randomly observing a difference as large or larger than that observed in this experiment?

The p-value depends on the variability, sample size and mean difference.
Assumptions:

- Random (or representative) data.
- Independent observations.
- Accurate data.
- Populations with Gaussian distributions.
- Equal standard deviation of the two populations.

Assumption on the same variance

It can be verified through another test for equal variance (F ratio or Snedecor's F). If it fails, we can

- Ignore the result. The t test is fairly robust to violations of this assumption as long as the sample size is large enough and equal in both groups.
- Emphasize that the two distributions are different, since at least the variance is different (its mean may or may not be different).
- Transform the data to reduce the variance (problem: data snooping).
- Running the unequal variance t test (Welch) or non-parametric (Mann-Whitney) (both are less powerful).
  - After checking that the variances are different (problem: data snooping).
  - Systematically.
  - Depending on some prior knowledge on the kind of experiment.

Common mistakes

- If the result is almost statistically significant, collect more data to increase the sample size and then recalculate the t test. Problem: multiple testing, and stopping when the difference becomes significant.
- If the experiment has three or more treatment groups, use the unpaired t test to compare two groups at a time. Use ANOVA and post-hoc analysis, instead.
- If the experiment has three or more treatment groups, compare the largest with the smallest means. Problem: data snooping, the decision of largest and smallest are taken looking at the data.
- If the p-value is larger than 0.05, try other tests to see whether they give a lower p-value. Problem: multiple testing.

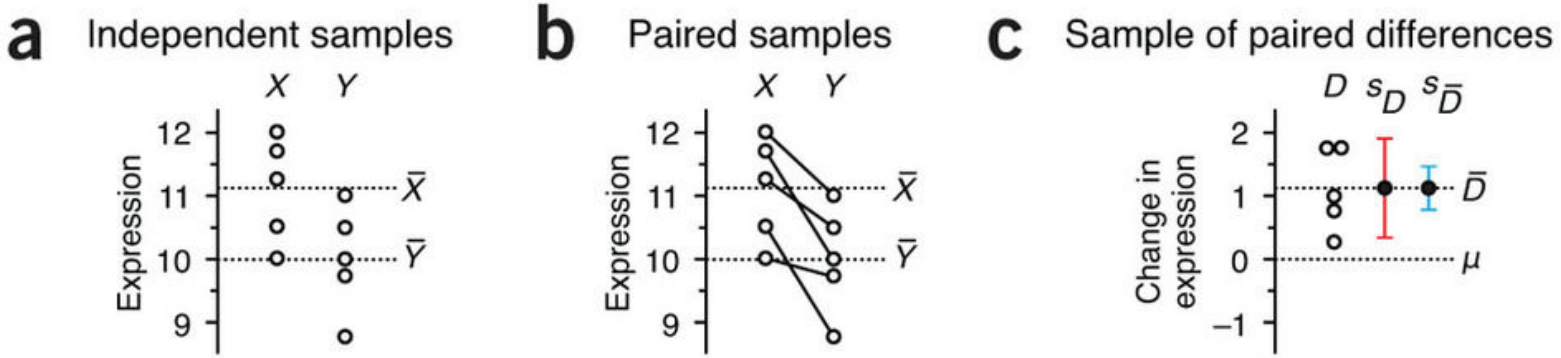
Comparing two groups

Outline

Comparing two paired means

- A variable is measured in each subject before and after an intervention.
- The left and right eyes of a person are treated with different eye drops.
- Subjects are recruited in pairs matched for age, postal code, or diagnosis.
- Twins or siblings are recruited as pairs.
- Each run in a laboratory has a control and treated preparation handled in parallel.

This data should be analyzed with paired t test or the non-parametric Mann-Whitney if the data is continuous, or McNemar's test if the data is binary.



The better control of other variables (genotype/phenotype, individual, environmental conditions, ...) helps to reduce the variance. They key is reducing each data from each pair to a single variable (difference estimate).

Comparing two paired means

The CI is constructed on the difference and it depends on the usual suspects (variability, sample size and confidence). If it includes 0, then the two groups are not significantly different. For instance Darwin measured the growth of 15 seeds of plants when they were cross-fertilized and self-fertilized. The difference in growth was 2.62 inches (the average height was 17.6 inches, $\sim$15%). The 95% CI was $[0.0037, 5.230]$. The result was significant (it does not include 0), but with very little margin.

The null hypothesis is that there is no difference in the two groups. The p-value answers the question if the null hypothesis were true, which would be the probability of observing a difference as large or larger than the one observed in this experiment. The p-value depends on the variability, sample size and mean difference. For Darwin's experiment, the p-value was 0.0497.

Comparing two paired means

How effective was the pairing?

You do the paired analysis because you expect that there is a strong correlation (positive) between the measurements in the two groups. To check this end, you may

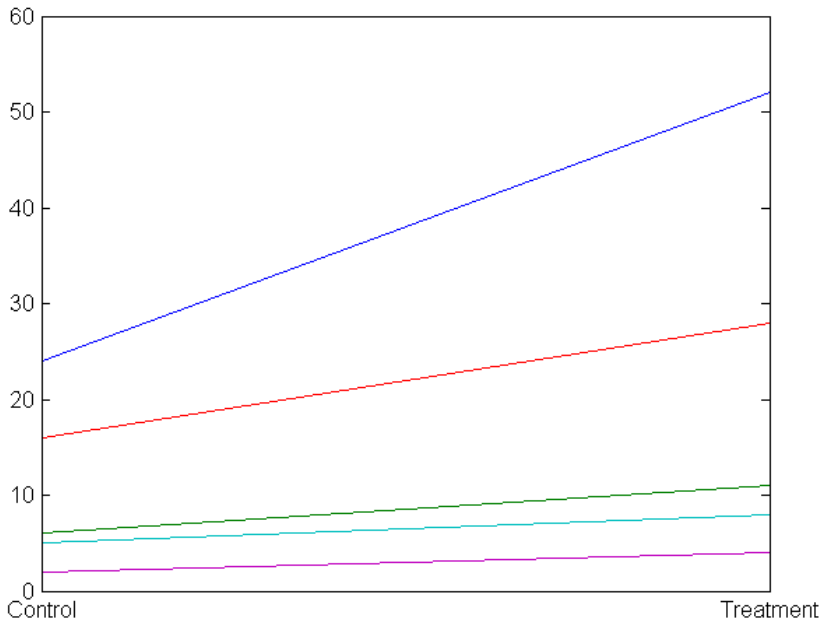
- Estimate the correlation coefficient, its CI and the p-value of the one-sided null hypothesis of no correlation.
  For Darwin's data ($r = -0.3348$, it is rare that both variables are negatively correlated), the one-sided p-value 0.1113 (not significant).
- Check the p-value if the data is analyzed unpaired.
  For Darwin's data, the p-value was 0.02

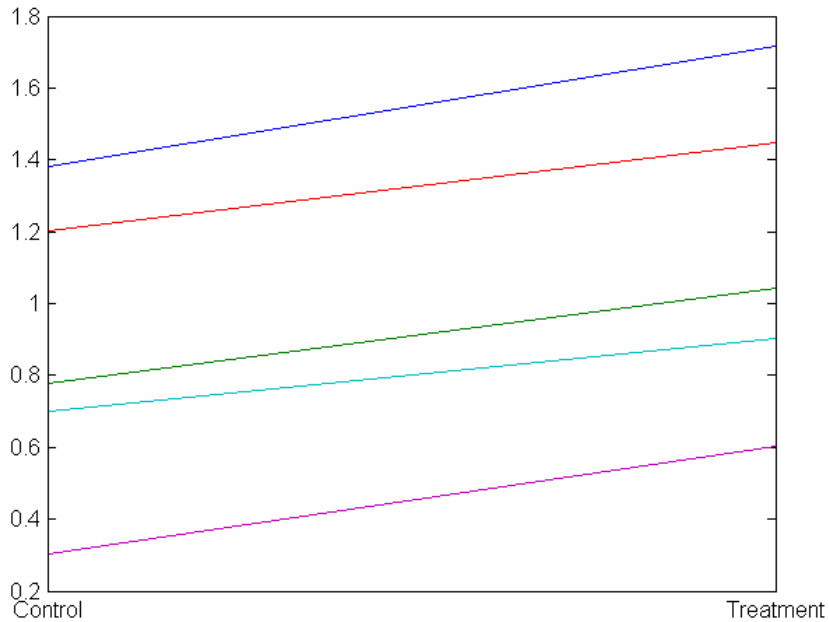
Assumptions

- Random (representative) samples.
- Pairs are independent of other pairs.
- The difference between matched values follow a Gaussian distribution.

Logarithmic transformations

Some variables may need to be transformed before being analyzed. For instance, enzymatic activity before and after a treatment with a drug has a multiplicative model. This can be easily seen in the plot below, after log-transforming the paired lines are "more parallel".





Paired binary variables (McNemar's test)

Examples

- Case-control studies where each case has a matching control (matched on age, gender, race, ...)
- Twins studies
- Before-after data, the outcome is absence or presence of some characteristic.

We want to know if a drug helps to reduce the prevalence of a symptom (e.g. running nose with a cold). For a number of people we measure the effect of the drug before and after treatment.

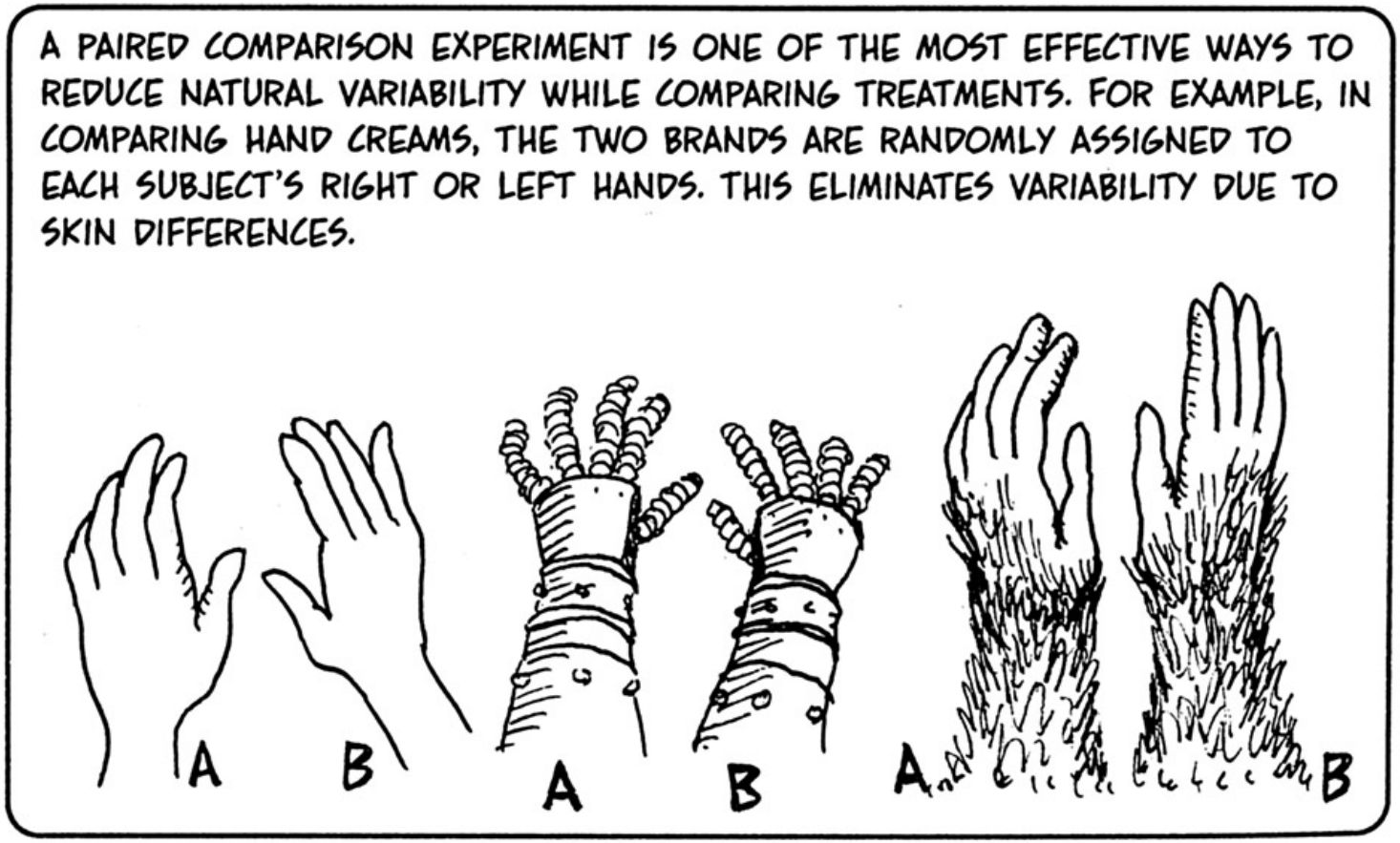
	After: present	After: absent
Before: present	A	B
After: absent	C	D

Measures are correlated because they are taken from the same subject. This is not a standard contigency table where counts in each cell are independent.

Common mistakes

- If the result is almost statistically significant, collect more data to increase the sample size and then recalculate the t test. Problem: multiple testing, and stopping when the difference becomes significant.
- If the experiment has three or more treatment groups, use the unpaired t test to compare two groups at a time. Use repeated measures ANOVA and post-hoc analysis, instead.
- Not log-transforming the data if it makes sense in this experiment (multiplicative model).
- Analyzing the absolute value of the differences instead of the differences.
- Deciding on the pairing only after seeing the data.

Comparing two paired means



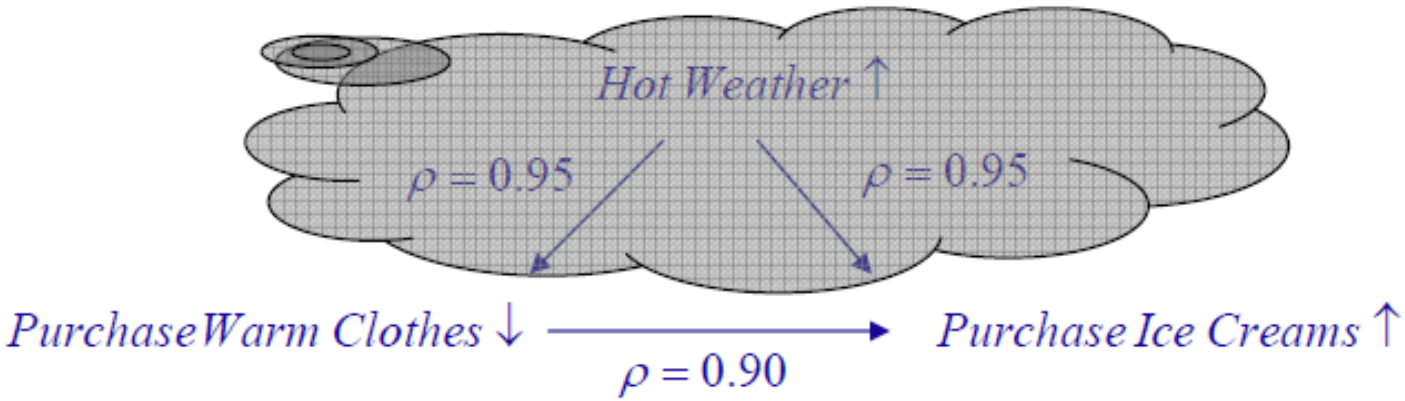
Outline

Calculating correlation

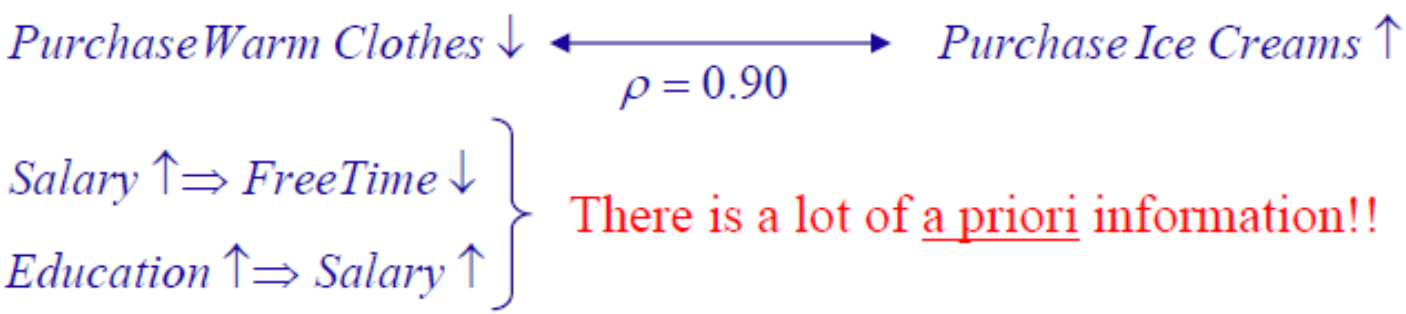
- Correlation coefficient: How much of Y can I explain given X?
  - Pearson's correlation coefficient: for continuous variables.
  - Kendall's rank correlation coefficient
  - Spearman's rank correlation coefficient
  - Coefficient of determination (R^2): when a model is available
- Multiple correlation coefficient: How much of Y can I explain given X_1 and X_2?
- Partial correlation coefficient: How much of Y can I explain given X_1 once I remove the variability of Y due to X_2?
- Part correlation coefficient: How much of Y can I explain given X_1 once I remove the variability of X_1 due to X_2?

Calculating correlation

Pitfall: Correlation means causation

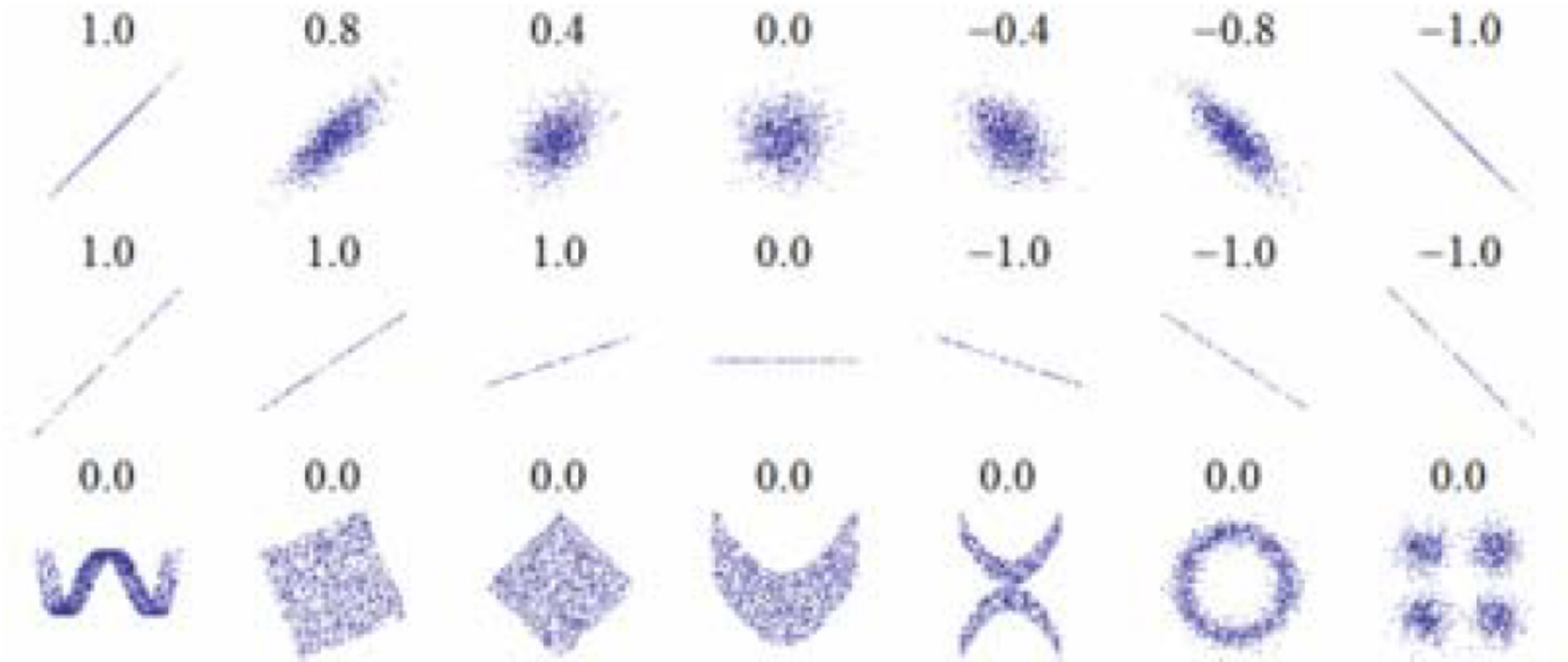


Correct: Correlation means linear covariation



Calculating correlation

Pitfall: Correlation measures all possible associations

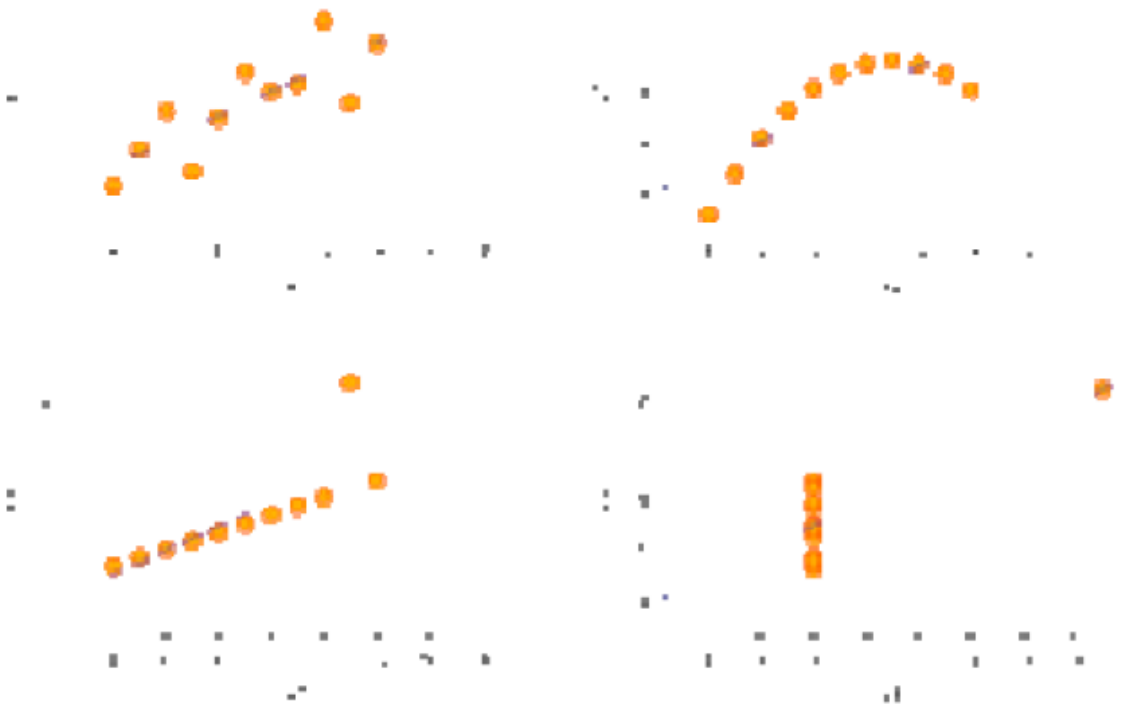


Correct: Correlation measures only linear associations

To measure non-linear associations the coefficient of determination is used (R^2)

Calculating correlation

Pitfall: Correlation summarizes well the relationship between two variables



$$\overline{y} = 7.5$$
$$s_Y = 4.12$$
$$y = 3 + 0.5x$$
$$r = 0.81$$

Correct: Visual inspection of the data structure is always needed

Calculating correlation

Is there any relationship between education and salary?

Person	Education	Salary $
A	3 (High)	70K
B	3 (High)	60K
C	2 (Medium)	40K
D	1 (Low)	20K

Pitfall: Compute the correlation between a categorical/ordinal variable and an interval variable.

Correct:

- Use ANOVA and the coefficient of determination
- Use Kendall or Spearman's rank correlation coefficient (valid only for ordinal, not categorical, variables)

Is there any relationship between education and salary?

Person	Education	Salary
A	3 (High)	3 (High)
B	3 (High)	3 (High)
C	2 (Medium)	2 (Medium)
D	1 (Low)	1 (Low)

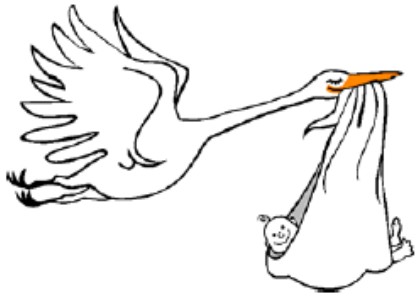
Pitfall: Compute the correlation between a two ordinal variables.

Correct:

Use Kendall or Spearman's rank correlation coefficient

Calculating correlation

Pitfall: Correlation between combinations with common variables



Village	#Women	#Babies	#Storks	#Babies/#Women	#Storks/#Women
VillageA	…				
VillageB	…				
VillageC	…				

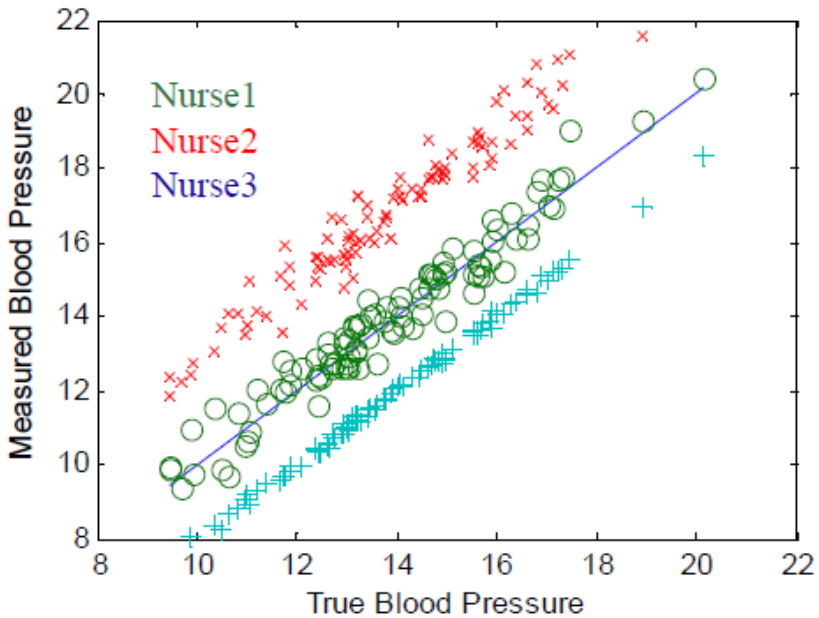
$$r_{BabiesPerWoman,StorkPerWoman} = 0.63!! \quad (p < 0.00001)$$

Calculating correlation

Pitfall: Correlation is invariant to changes in mean and variance

Three nurses take blood pressure from the same pool of patients:

- Nurse 1 takes the true value with some variance.
- Nurse 2 takes consistently larger values with the same variance as nurse 1.
- Nurse 3 takes consistently smaller values with much less variance than the other 2.



$$r_{Nurse1,Nurse2} = 0.95$$

$$r_{Nurse1,Nurse3} = 0.97$$

$$r_{Nurse2,Nurse3} = 0.97$$

All correlations are rather high (meaning high agreement) although the data is quite different

Calculating correlation

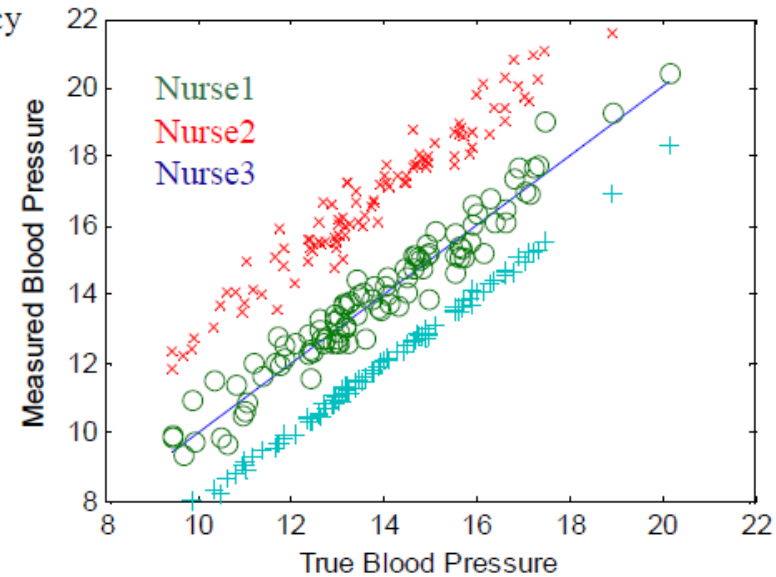
Solution: Assess agreement through bias, scale difference and accuracy

$$E\left\{(X_1 - X_2)^2\right\} = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2(1-\rho)\sigma_1\sigma_2$$

$$\frac{E\left\{(X_1 - X_2)^2\right\}}{2\sigma_1\sigma_2} = \underbrace{\frac{(\mu_1 - \mu_2)^2}{2\sigma_1\sigma_2}}_{\text{Normalized bias}} + \underbrace{\frac{(\sigma_1 - \sigma_2)^2}{2\sigma_1\sigma_2}}_{\text{Normalized scale difference}} + \underbrace{(1-\rho)}_{\text{Accuracy}}$$

Nurse1 vs. Nurse2	1.01	1e-5	0.05
Nurse1 vs. Nurse3	0.51	7e-4	0.03
Nurse2 vs. Nurse3	3.05	4e-4	0.03

Now we have separated the three different effects (mean shift, scale shift, correlation) while the correlation alone only accounted for one of them.



Calculating correlation

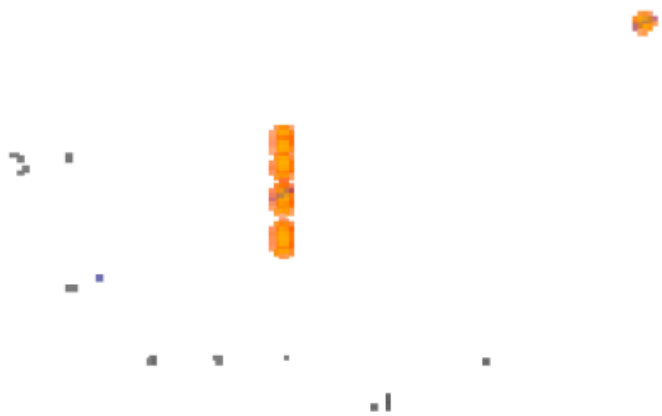
Pitfall: Summarize a regression experiment through correlation

Bivariate sampling: the experimenter does not control the X nor the Y, he only measures both.

Which is the relationship between height and weight in a random sample?
The experimenter cannot control any of the two variables

Regression sampling: the experimenter controls X but not Y, he measures Y

Which is the response of blood pressure to a certain drug dosis?
The experimenter decides the dosis to be tested.



$$r^2 = \frac{1}{1 + K^2 \frac{s^2_{Y|X}}{s^2_X}}$$

$K, s^2_{Y|X}$ Depend on the system he is measuring

s^2_X The experimenter controls the width of values tested.
By making the range of X large, we can have a correlation as closed to 1 as desired.

Calculating correlation

Outline

Chapter 6. Fitting models

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

August 20, 2017

Outline

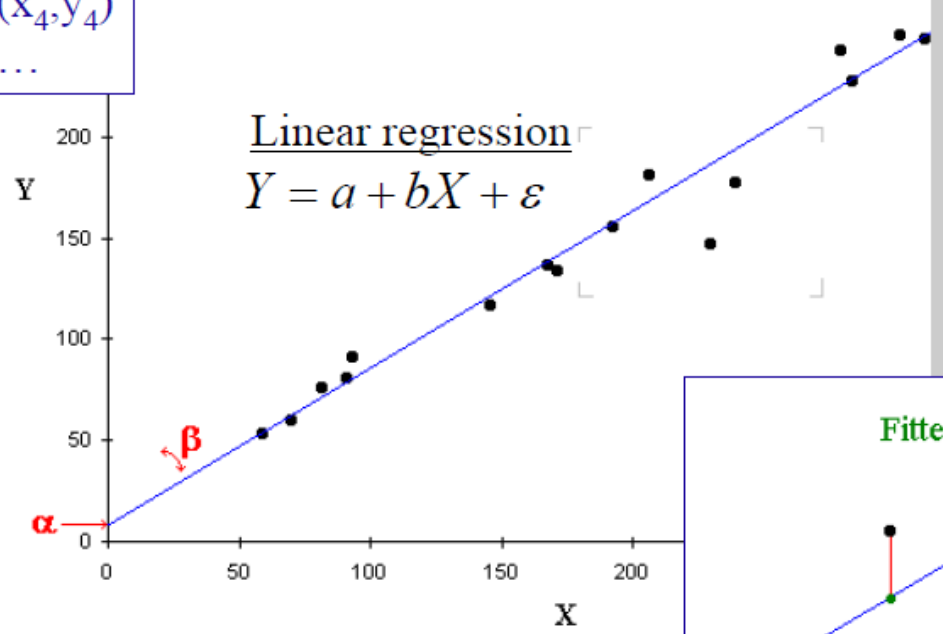
Outline

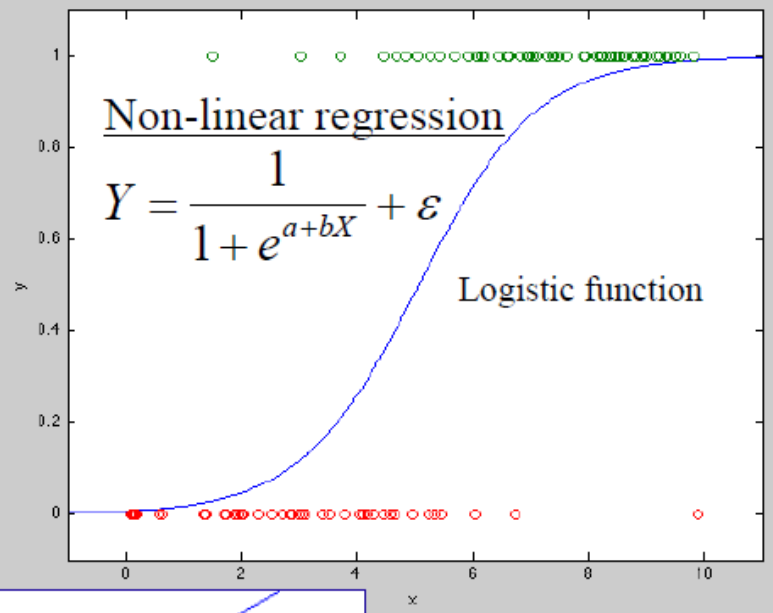
Regression models

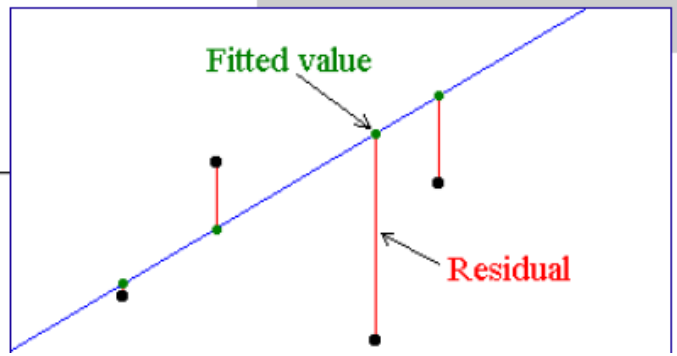
(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
(x_4, y_4)
...

$$Y = f(X) + \varepsilon$$

$$CarPrice = f(Age) + \varepsilon$$



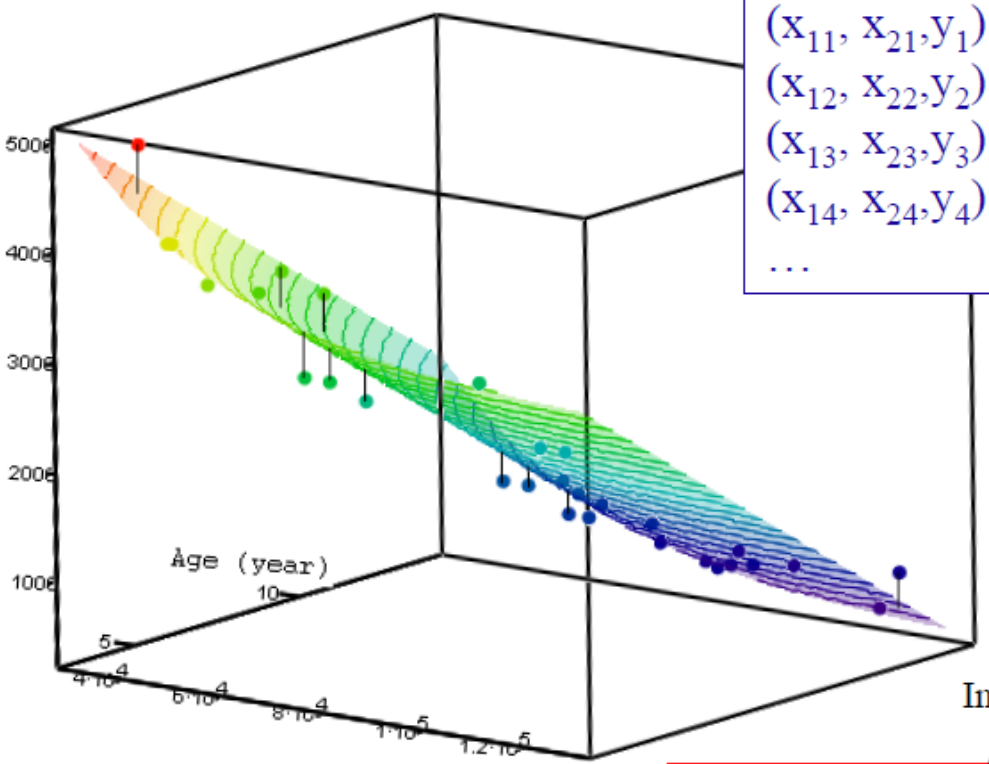




7.4 Regression as a model

$$Y = f(X_1, X_2) + \varepsilon$$

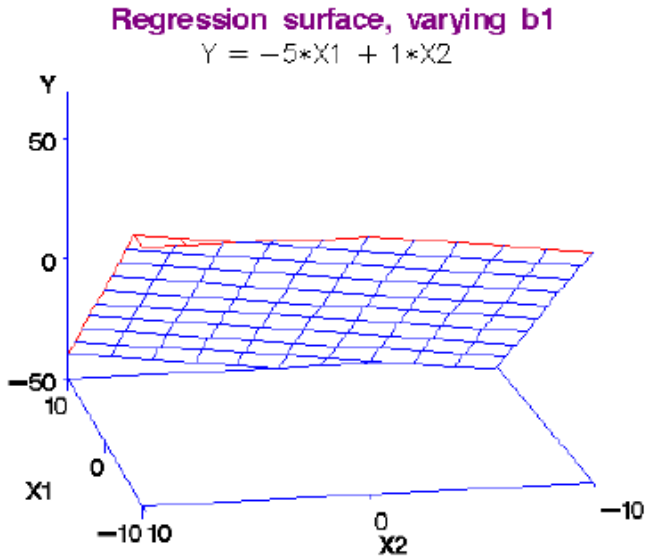
$$CarPrice = f(Age, Mileage) + \varepsilon$$

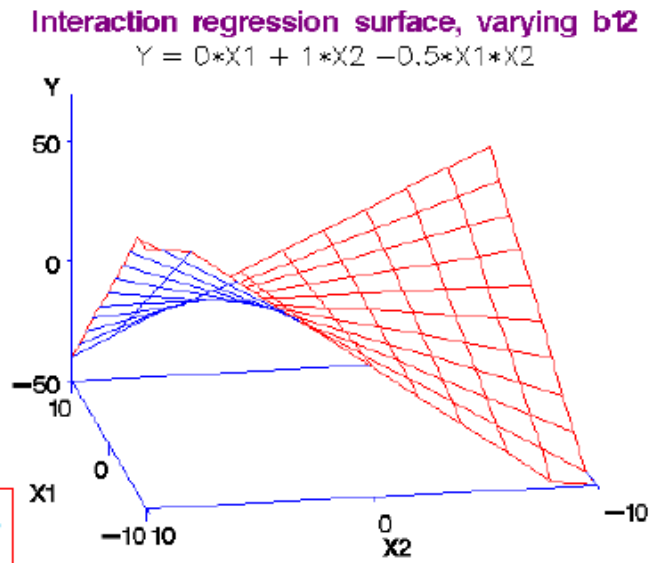


(x_{11}, x_{21}, y_1)
(x_{12}, x_{22}, y_2)
(x_{13}, x_{23}, y_3)
(x_{14}, x_{24}, y_4)
...

Interactions

$$Y = f(X_1, X_2, X_1X_2) + \varepsilon$$





Regression models

Coefficient of determination

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

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

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}}$$

The coefficient of determination represents the percentage of "unexplained" variance

Multiple correlation coefficient

$$r = \sqrt{R^2}$$

Partial correlation coefficient $\rho_{XY.\mathbf{Z}}$

The partial correlation coefficient of Y and X removing the effect of $(Z_1,\ldots,Z_p)$ is the correlation of the residuals of Y after linear multiple regression with $(Z_1,\ldots,Z_p)$ and the residuals of X after linear multiple regression with $(Z_1,\ldots,Z_p)$

What is the correlation between crop yield and temperature? Standard correlation
What is the correlation between crop yield and temperature holding rain fixed? Partial correlation

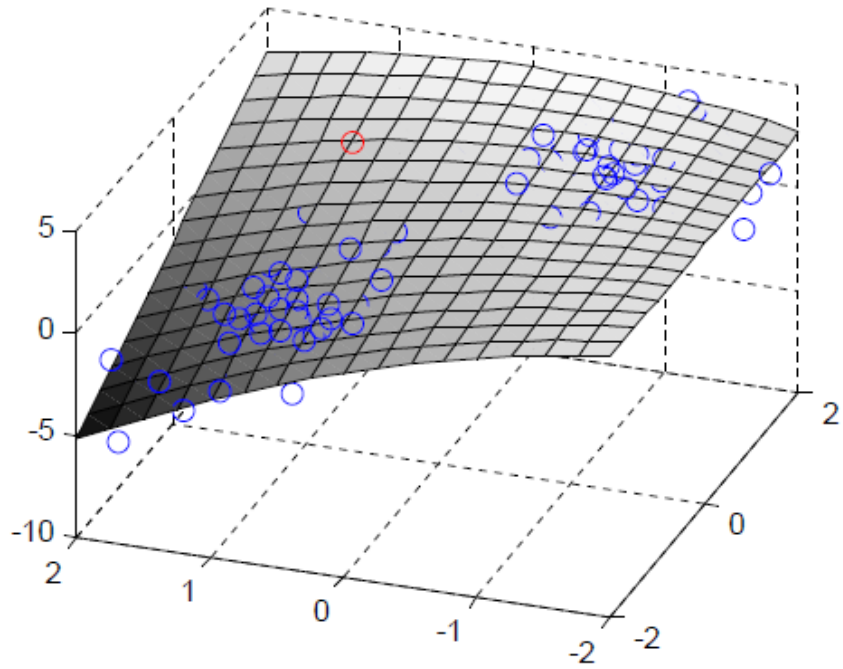
Regression models

Multiple regression

$$Y = f(X_1, X_2, ..., X_p) + \varepsilon$$

Predictors: continuous, discrete, categorical

Random variable with known distribution



Assumptions

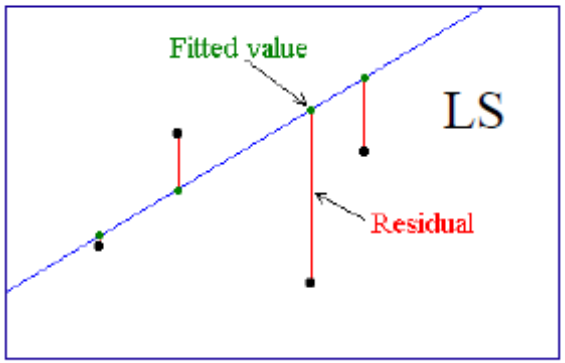
1. The sample is representative of your population
   - If you are to predict the price of a car of 1970 make sure that your "training sample" contained cars from that year.
   - If you are to predict the price of a car of 1970 with 100.000 miles, make sure that your "training sample" contained cars from that year and that mileage.

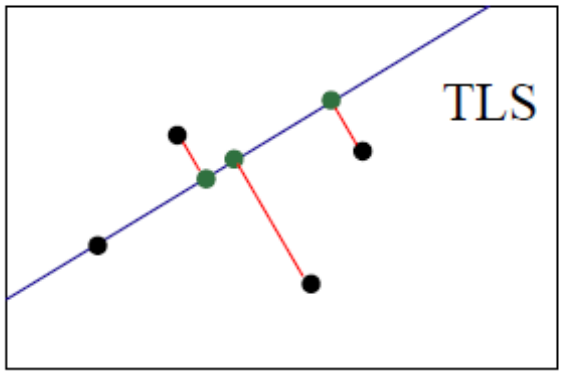
   Solution: Make sure that your predictor vector (X_1, …, X_p) is not an outlier of the "training sample".

Regression models

Assumptions

2. The dependent variable is noisy, but the predictors are not!!
   Solution: If the predictors are noisy, use a scheme like Total Least Squares





$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

$$Y = f(X_1 + \varepsilon_1, X_2 + \varepsilon_2, \ldots, X_p + \varepsilon_p) + \varepsilon$$

Zero-mean
Random variable
Systematic errors are not contemplated

Least Squares

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - f(\mathbf{X}_i, \boldsymbol{\beta})\right)^2$$

Total Least Squares

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left\| (\mathbf{X}_i, Y_i) - (\mathbf{X}_i, f(\mathbf{X}_i, \boldsymbol{\beta})) \right\|^2$$

Regression models

Assumptions

3. Predictors are linearly independent (i.e., no predictor can be expressed as a linear combination of the rest), although they can be correlated. If it happens, this is called multicollinearity.

$$PersonHeight = f(weightPounds, weightKilograms) + \varepsilon$$

Problem:

- Confidence intervals of the regression coefficients are very wide.
- Large changes in coefficients when a sample is added/deleted.
- Simply for predicting Y, multicollinearity is not a problem.

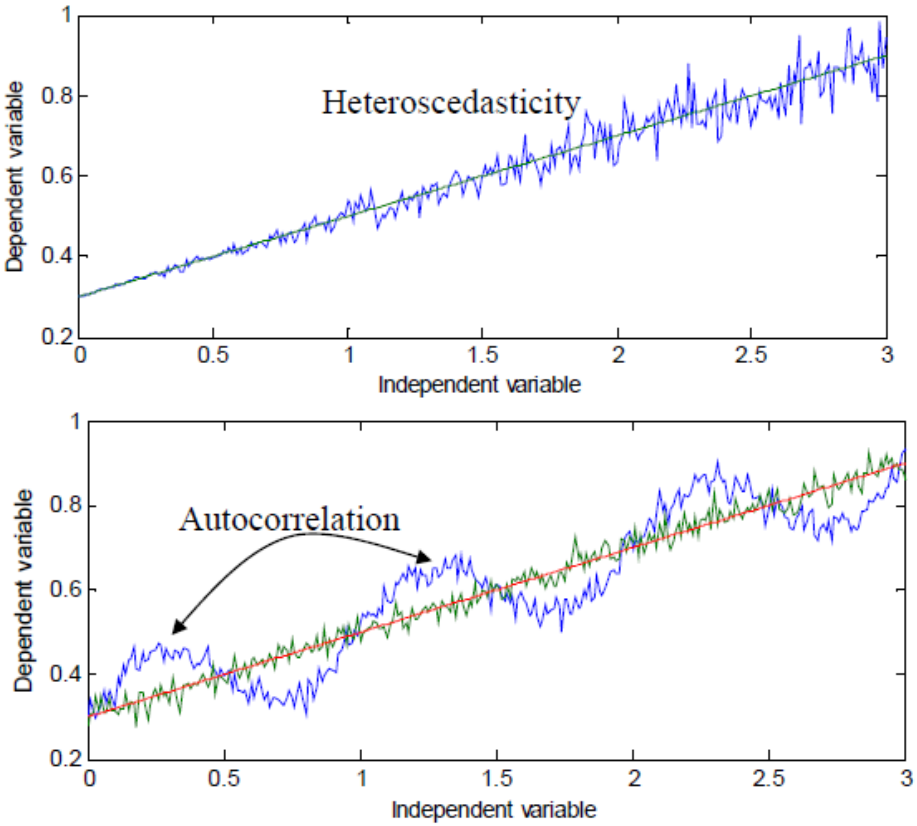
Solution:

- Understand the reason and remove it (usually it means, that several predictors are measuring essentially the same thing).
- Add more samples (if you have more predictors than observations you have multicollinearity for sure).
- Change your predictors to orthogonal predictors through PCA

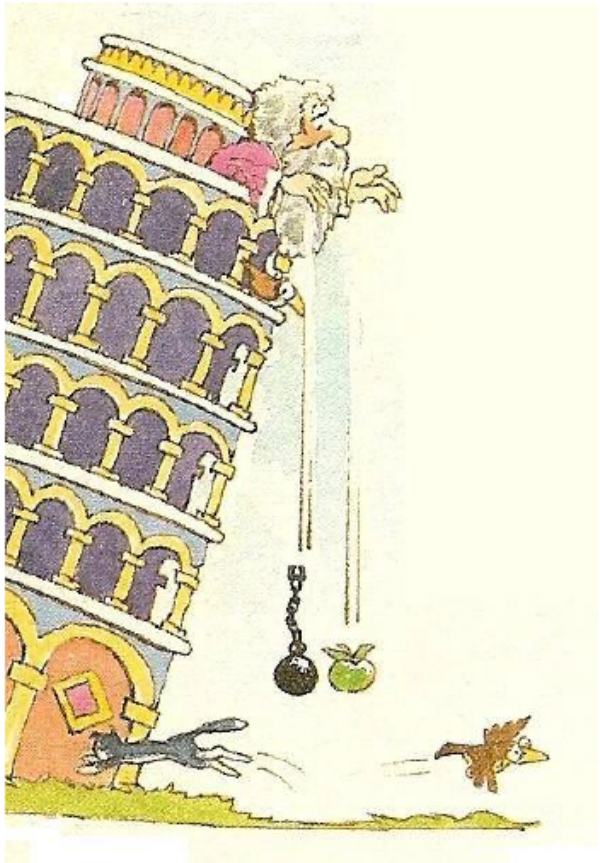
Regression models

Assumptions

4. The errors are homocedastic (i.e., they have the same error at all predictor values)
   Solution:
   - Transform the data (square root, log, …) to diminish this effects
   - Use Weighted Least Squares
5. The errors are uncorrelated to the predictors and to itself (i.e., the covariance matrix is diagonal).
   Solution:
   - Use Generalized Least Squares.
6. The errors follow a normal distribution.
   Solution:
   - Use Generalized Linear Models.



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?

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

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

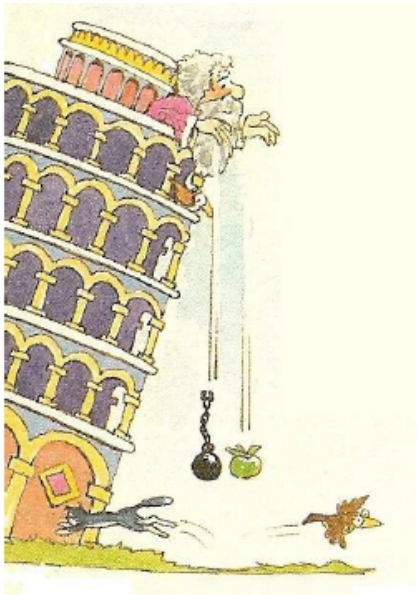
$$h(t) = a_0 + a_{\frac{1}{2}} \sqrt{t} + a_1 t + a_2 t^2 + \varepsilon$$

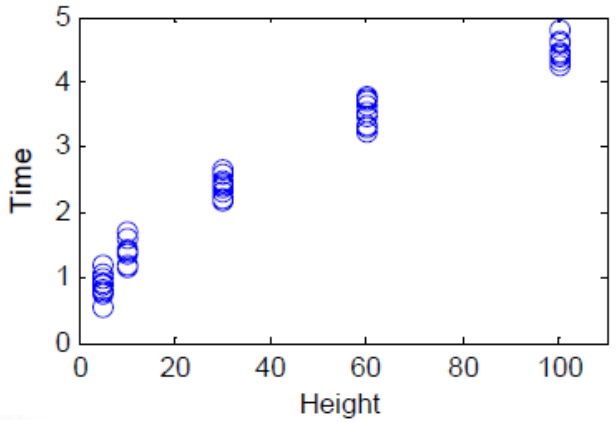
$$t(h) = a_0 + a_1 h + a_2 h^2 + \varepsilon$$

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

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

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.

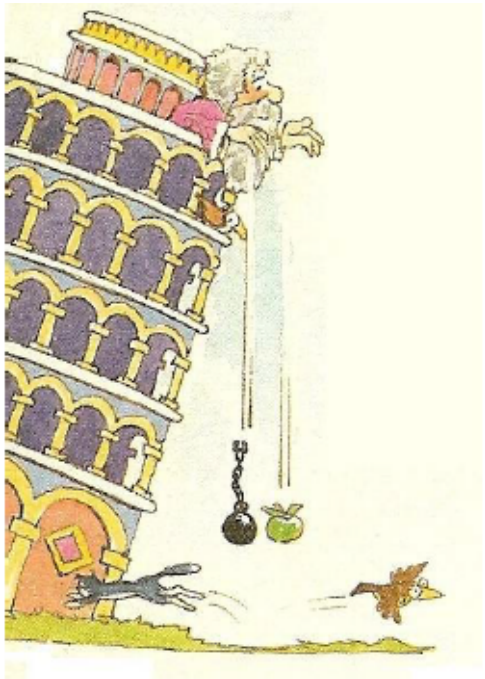
$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + a_1 h + a_2 h^2 + \varepsilon$$
$$t(h) = -0.33 + 0.62\sqrt{h} + 0.02h + 0h^2 + \varepsilon \quad R^2 = 0.9773$$

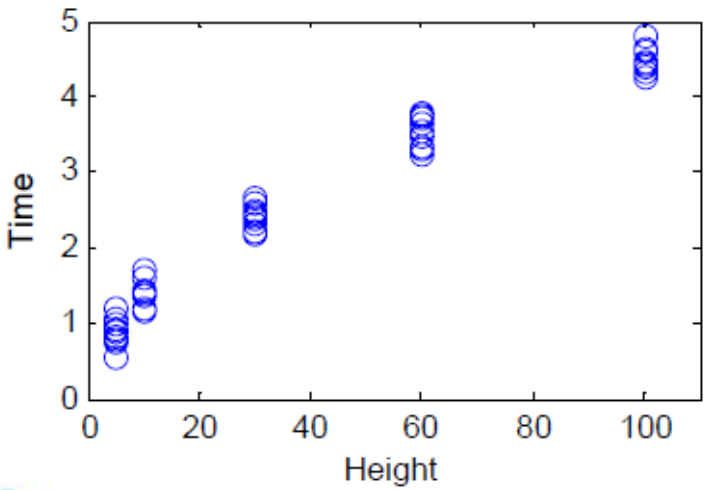
$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + a_1 h + \varepsilon$$
$$t(h) = -0.15 + 0.51\sqrt{h} + 0h + \varepsilon \quad R^2 = 0.9772$$

$$t(h) = a_0 + a_{\frac{1}{2}}\sqrt{h} + \varepsilon$$
$$t(h) = 0 + 0.45\sqrt{h} + \varepsilon \quad R^2 = 0.9766$$

$$t(h) = a_{\frac{1}{2}}\sqrt{h} + \varepsilon \longleftarrow \text{This is the true model!!!}$$
$$t(h) = 0.45\sqrt{h} + \varepsilon \quad R^2 = 0.9766$$

Regression models





Adjusted R: this is a way of reducing overfitting

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{N-1}{N-k-1}$$

$$t(h) = -0.33 + 0.62\sqrt{h} + 0.02h + 0h^2 + \varepsilon$$
$$R^2 = 0.9773 \quad R^2_{adjusted} = 0.9760$$

$$t(h) = -0.15 + 0.51\sqrt{h} + 0h + \varepsilon$$
$$R^2 = 0.9772 \quad R^2_{adjusted} = 0.9762$$

$$t(h) = 0 + 0.45\sqrt{h} + \varepsilon$$
$$R^2 = 0.9766 \quad R^2_{adjusted} = 0.9759$$

$$t(h) = 0.45\sqrt{h} + \varepsilon$$
$$R^2 = 0.9766 \quad R^2_{adjusted} = 0.9762$$

Regression models

Can we distinguish between important coefficients and non important coefficients?

Linear regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \begin{pmatrix} t_1 \\ t_2 \\ \dots \\ t_N \end{pmatrix} = \begin{pmatrix} 1 & \sqrt{h_1} & h_1 & h_1^2 \\ 1 & \sqrt{h_2} & h_2 & h_2^2 \\ \dots & \dots & \dots & \dots \\ 1 & \sqrt{h_N} & h_N & h_N^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_{\frac{1}{2}} \\ a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

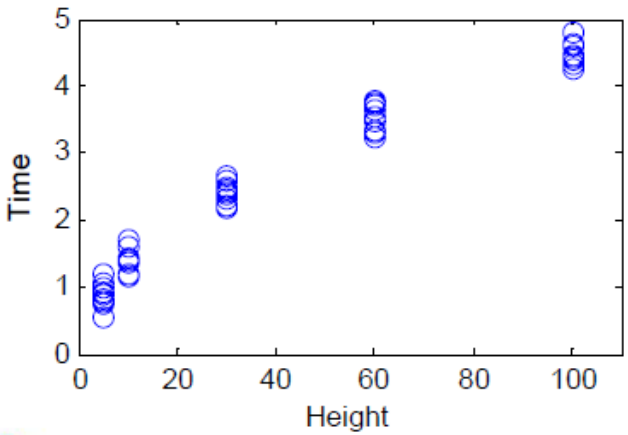
$$\boldsymbol{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} \longleftarrow \text{Regression coefficients}$$

$$\hat{\sigma}^2 = \frac{1}{N-k}\sum_{i=1}^{N}\left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i \rangle\right)^2 \longleftarrow \text{Unbiased variance of the residuals}$$

$$\hat{\sigma}_j^2 = \hat{\sigma}^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{jj} \longleftarrow \text{Unbiased variance of the j-th regression coefficient}$$

$$\beta_j \in \hat{\beta}_j + t_{1-\frac{\alpha}{2},N-k-1}\hat{\sigma}_j^2 \longleftarrow \text{Confidence interval for the j-th regression coefficient}$$

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.

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

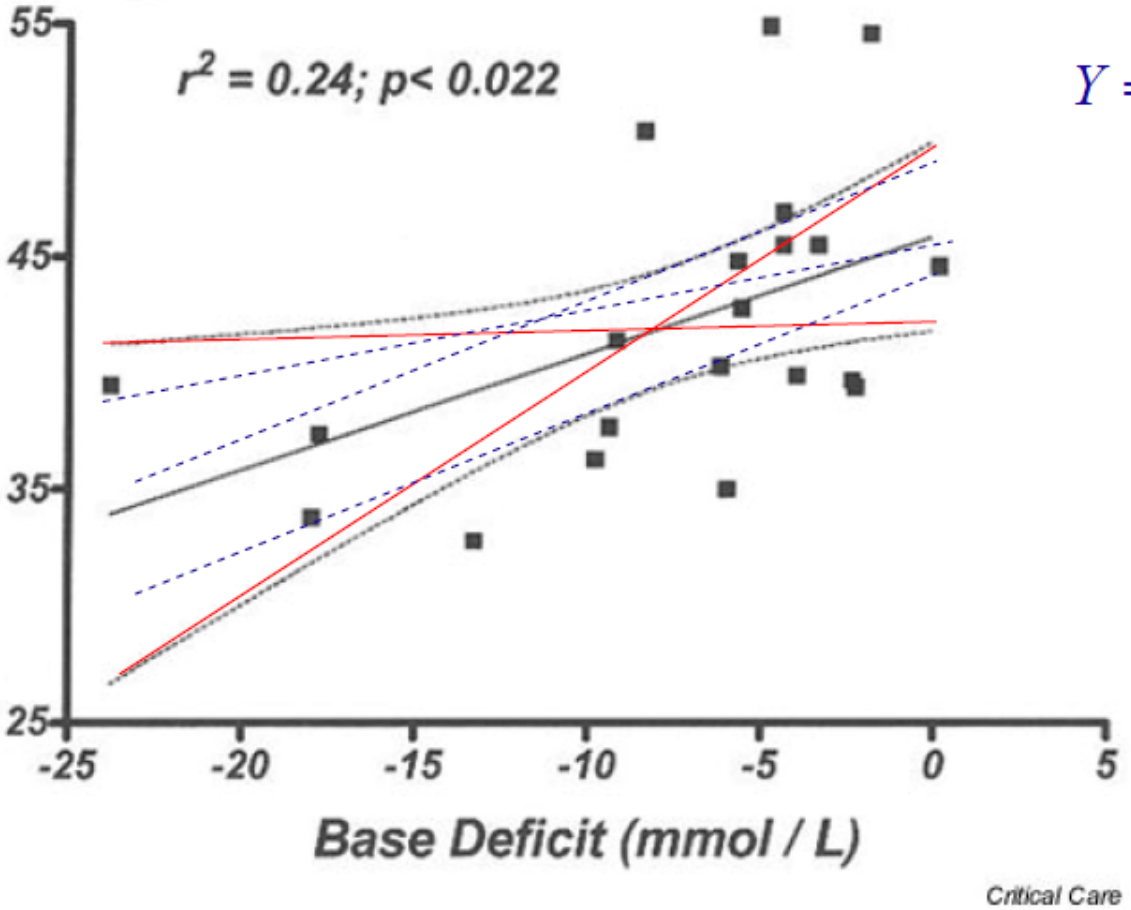
$$t(h) = -0.33 + 0.62\sqrt{h} + 0.02h + 0h^2 + \varepsilon \qquad R^2 = 0.9773$$

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

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

Regression models

- *PeCO2 (mmHg)*

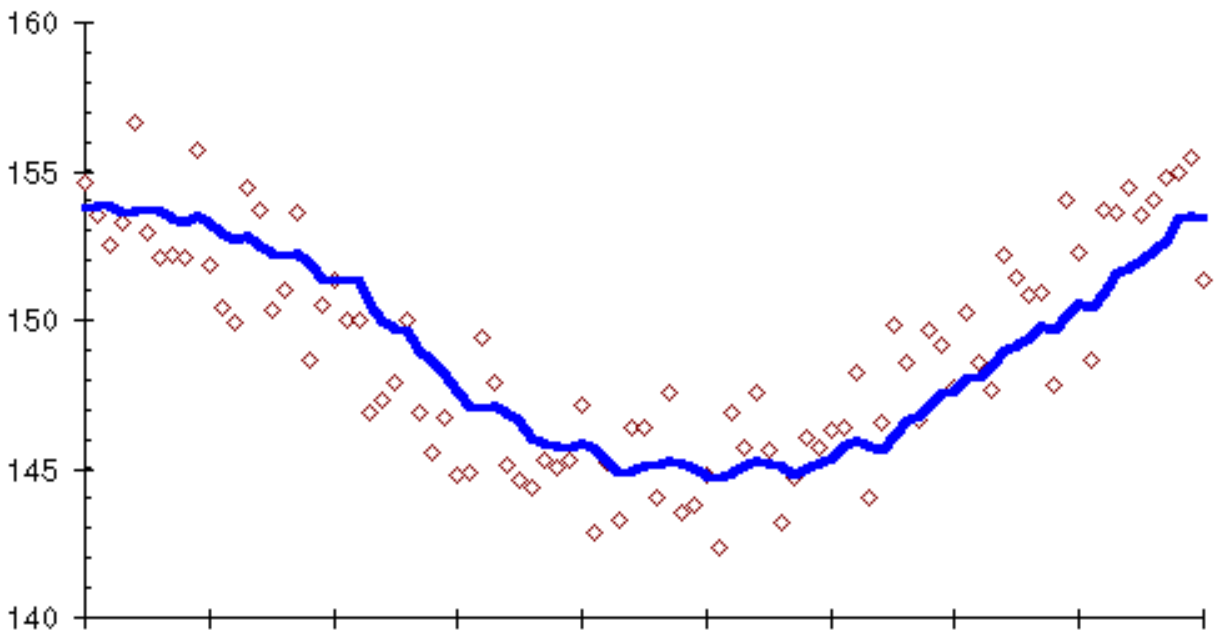


$$Y = [40, 45] + [0.05, 0.45]X$$

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

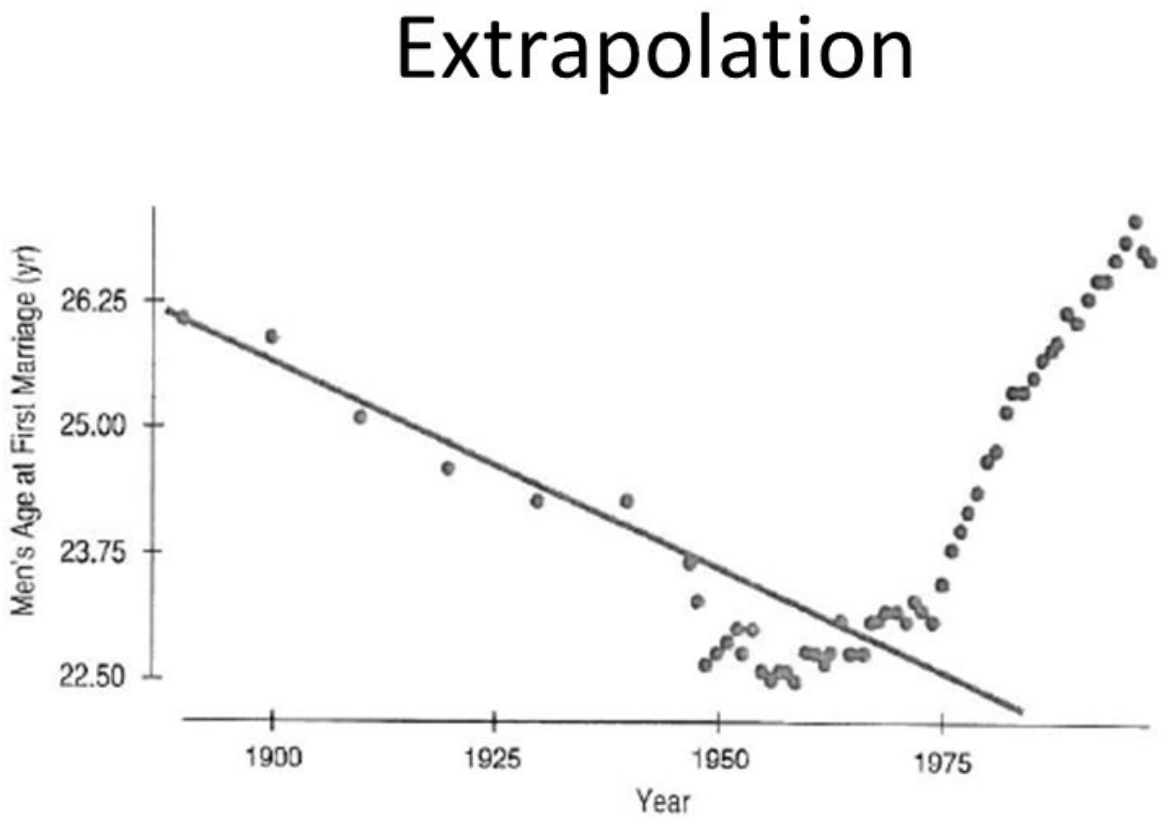
Common mistakes

- 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.



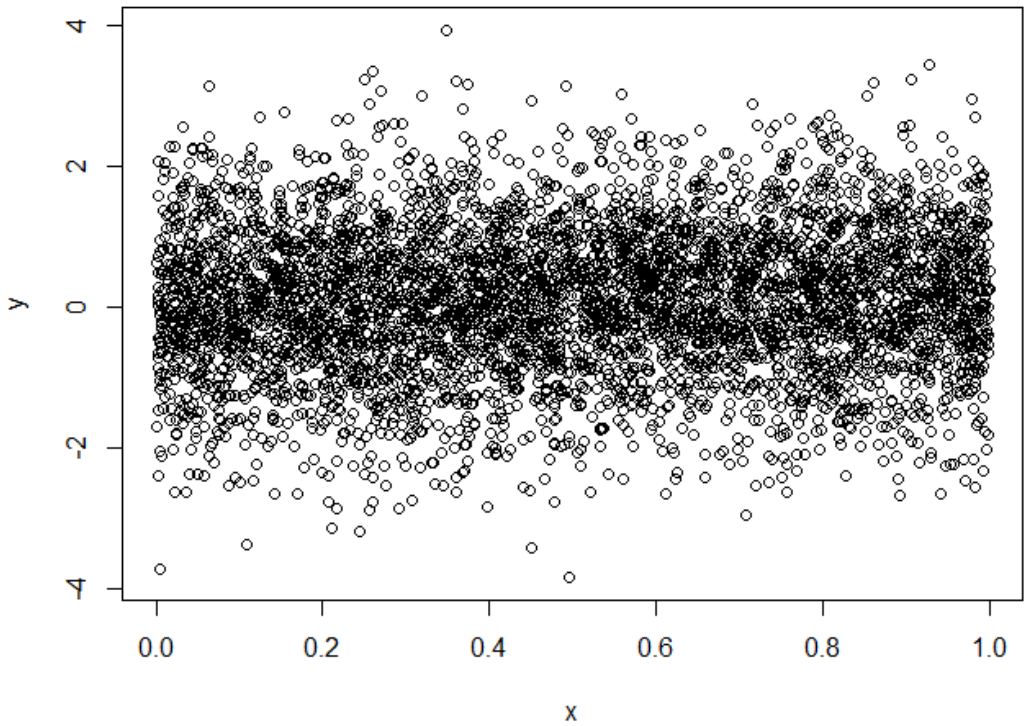
Common mistakes

- 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.

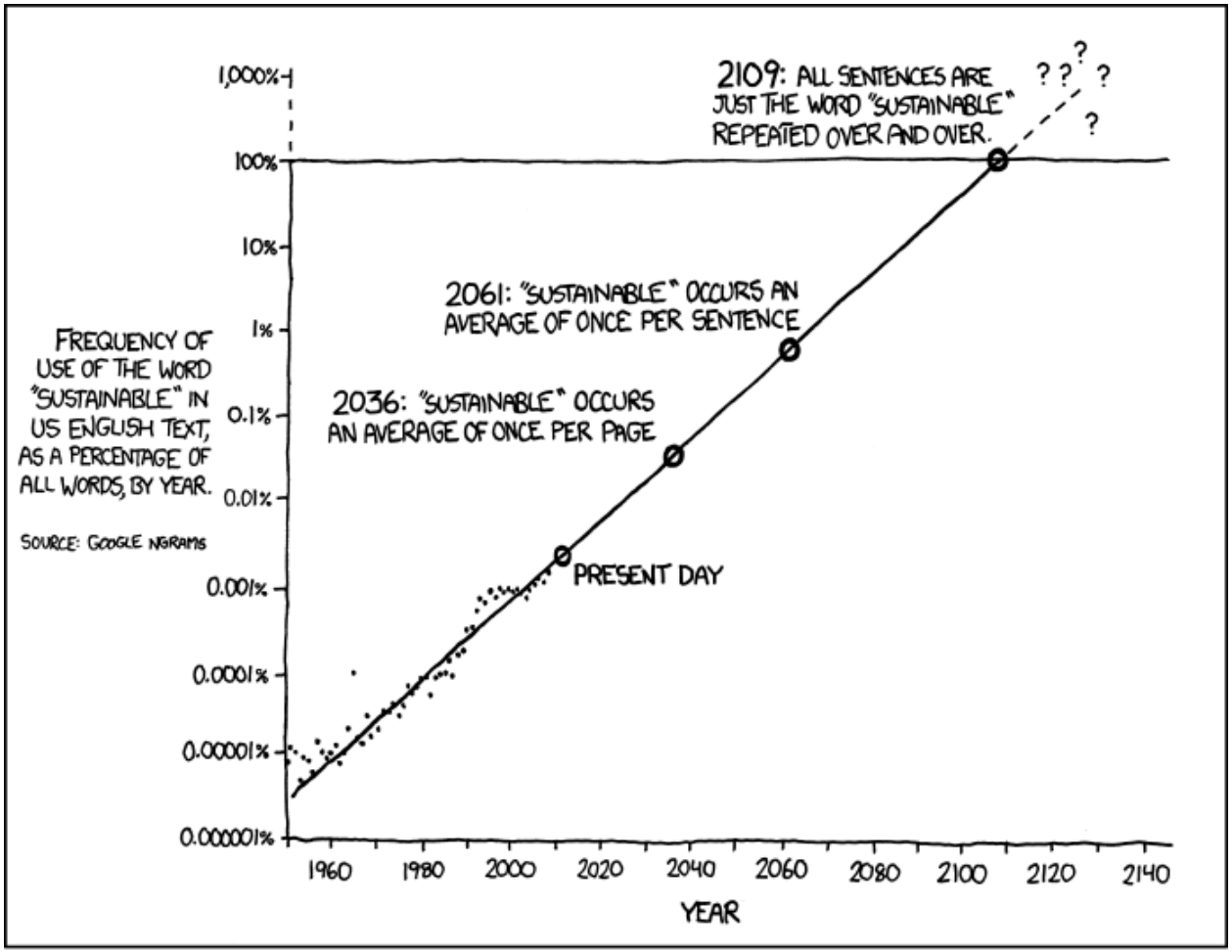


Common mistakes

- 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.



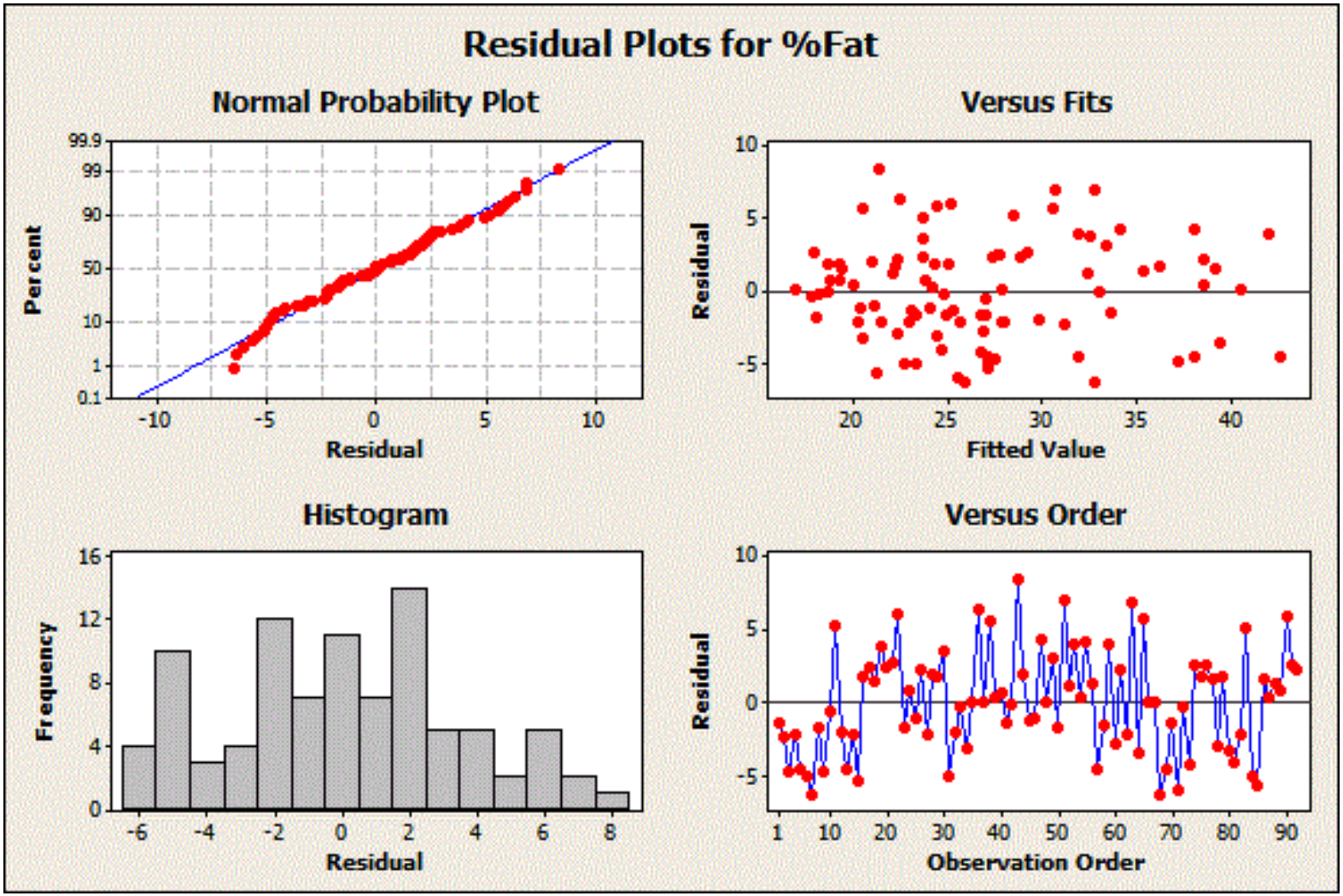
Regression models



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

Outline

Residual analysis



Residual analysis

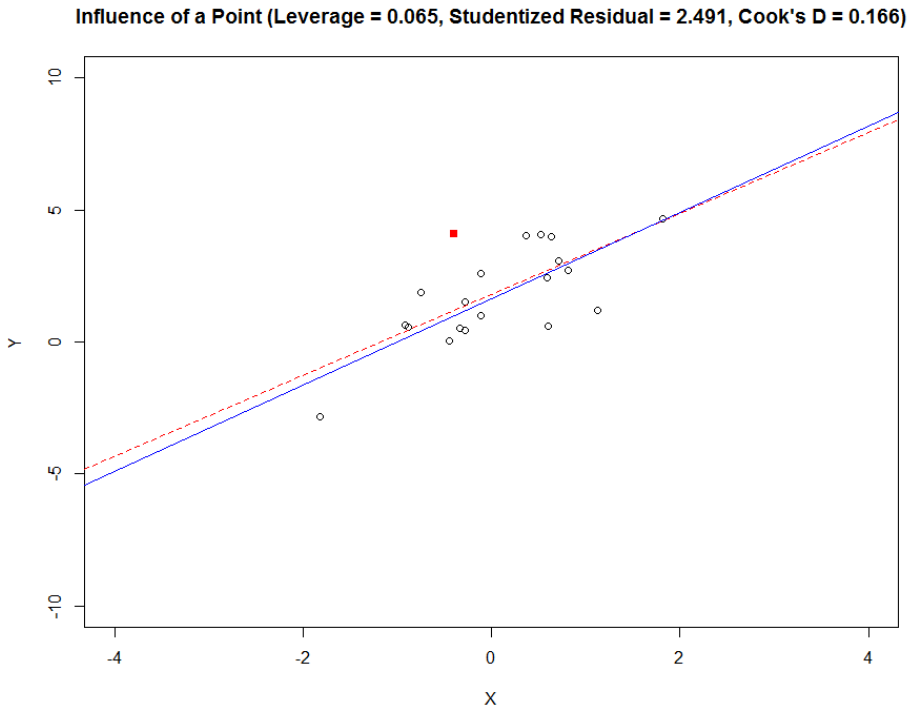
- Residuals should follow a Gaussian distribution (there should not be large outliers).
- Residuals should not depend on the predicted variable (Y).
- Residuals should not depend on the predictor variables (X).

If these conditions are not met

- Revise the data for outliers and very influential points.
- Revise the model, probably this model cannot explain well the data.

Influence of points

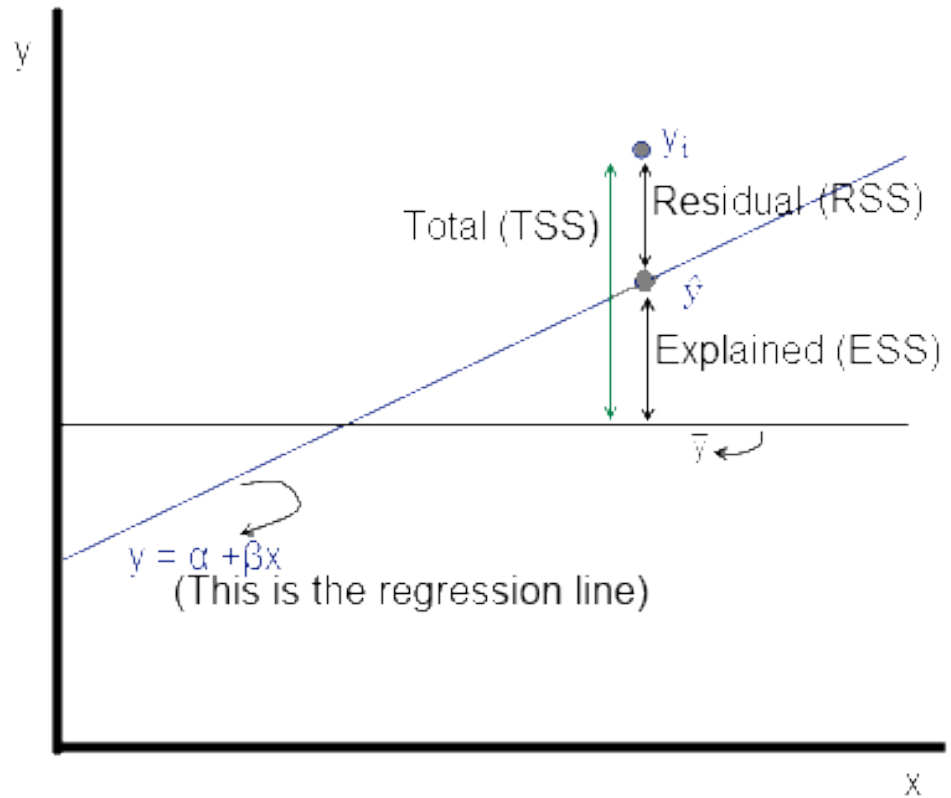
We may measure the influence of each individual point through different statistics.



- Leverage. How much x_i is an outlier? Between 0 (not an outlier) and 1 (totally outlying).
- Studentized residual. How much y_i is an outlier? In terms of the standard deviation of the residuals (0=not an outlier)
- Cook's D How much the regression would change if we remove that point? (0=no change)

Meaning of R^2

R^2 is the fraction of the total variance explained by the regression model. It is between 0 and 1, the larger the better.



$\hat{y}$ is the predicted value of y given x, using the equation $y = \alpha + \beta x$.

y_i is the actual observed value of y

$\bar{y}$ is the mean of y.

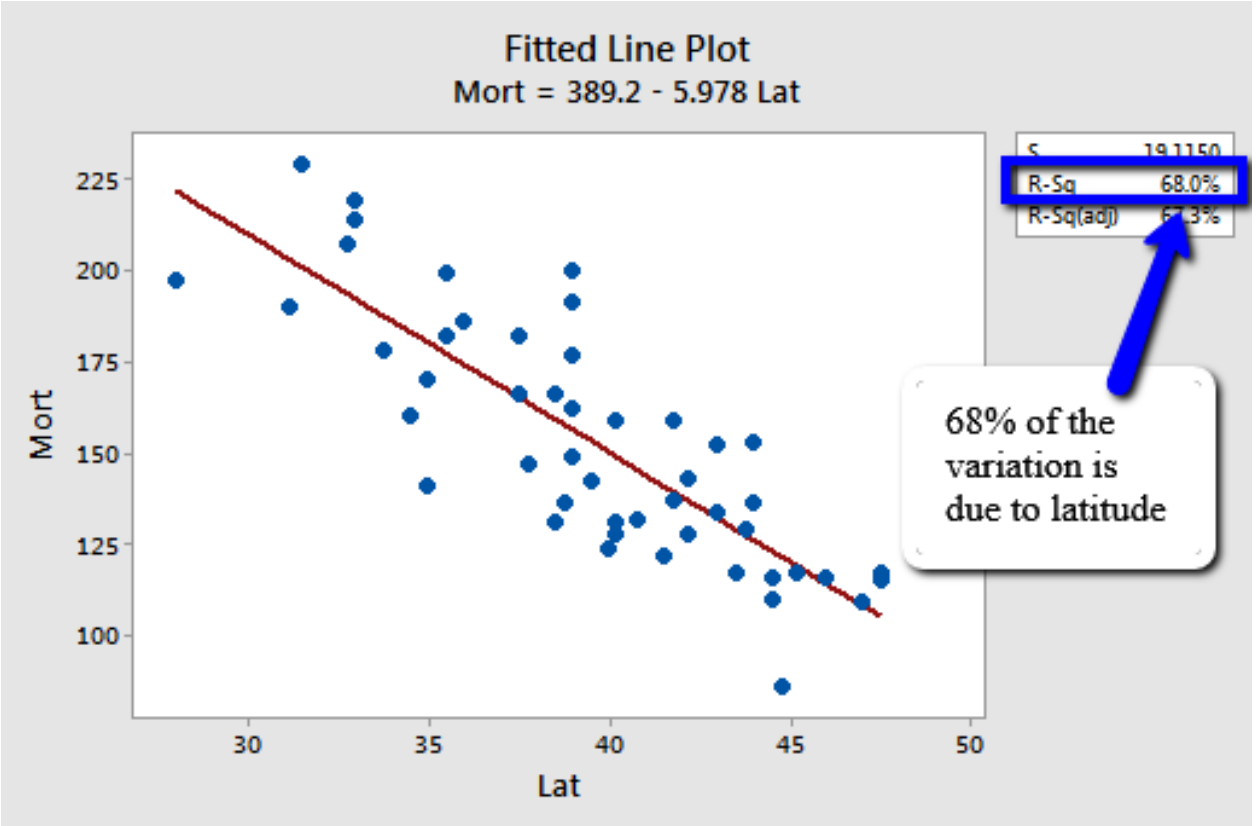
The distances that RSS, ESS and TSS represent are shown in the diagram to the left - but remember that the actual calculations are squares of these distances.

$$TSS = \Sigma (y_i - \bar{y})^2$$

$$RSS = \Sigma (y_i - \hat{y})^2$$

$$ESS = \Sigma (\hat{y} - \bar{y})^2$$

Meaning of R^2

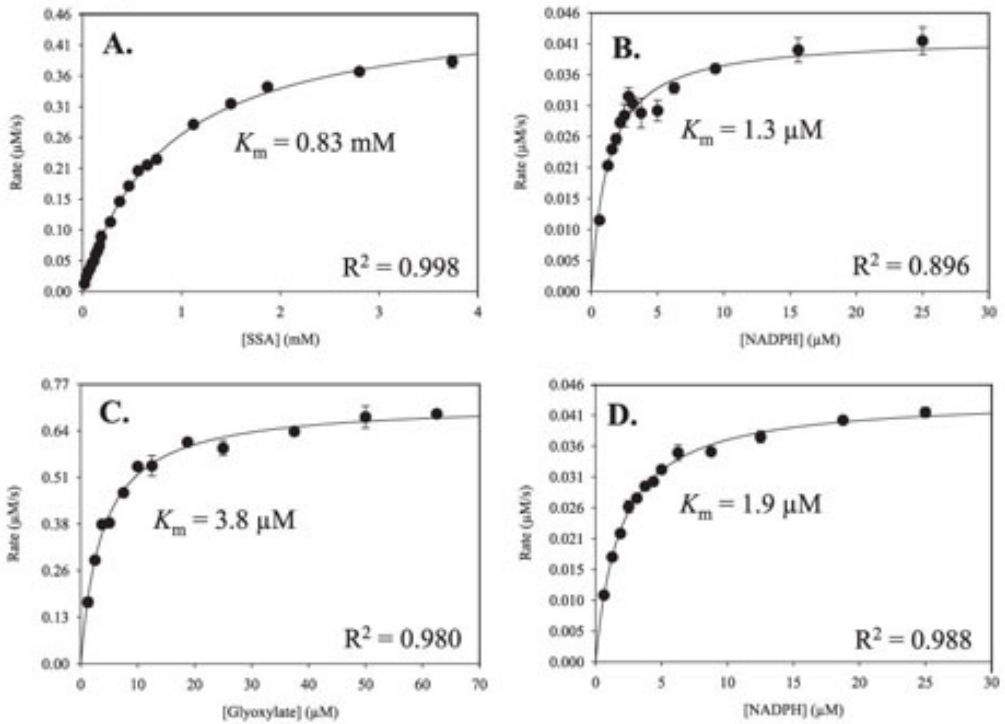


Nonlinear regression

Diagnostic tools are the same for nonlinear regression

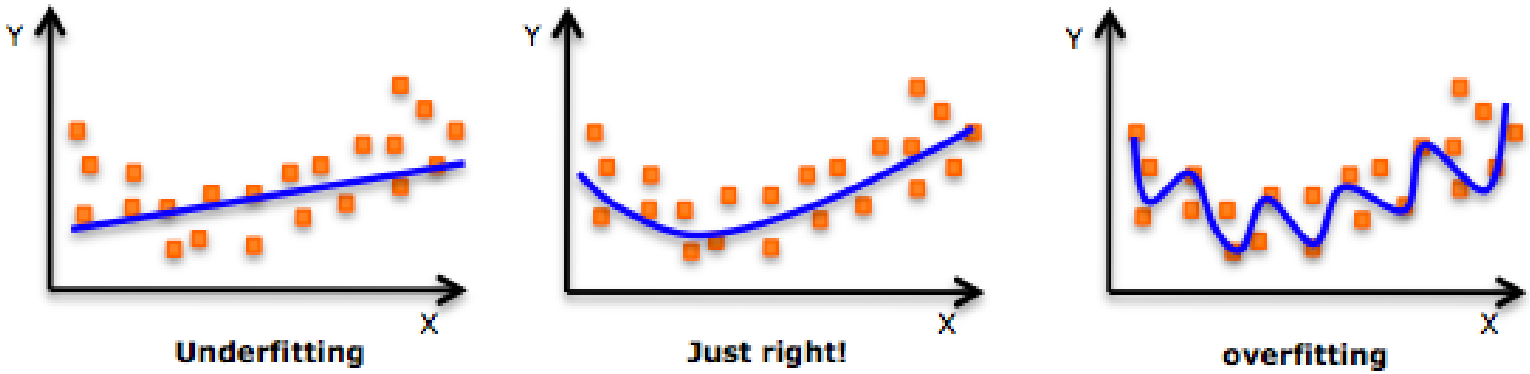
$$v = v_{max} \frac{[S]}{K_m + [S]}$$

Fig. 5. Kinetic characterization of succinic semialdehyde (A and B) and glyoxylate reductase (C and D) activities. Kinetic data from a single typical enzyme preparation were best fit by nonlinear regression analysis. Each datum represents the mean ± SE; where SE is not shown, it is within the symbol.



Choosing between models

Choosing the best model is not easy. R^2 is not enough because models with more parameters will tend to have lower R^2 simply by overfitting.



A solution is to penalize by the number of parameters (p) with respect to the number of samples (N), the R^2 is said to be adjusted

$$\begin{aligned} R^2 &= 1 - \frac{SS_{residuals}}{SS_{total}} \\ R^2_{adjusted} &= 1 - \frac{SS_{residuals}/(N-p)}{SS_{total}/(N-1)} = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \end{aligned}$$

There are other penalization schemes: Akaike's Information Criterion, Schwarz's Bayesian Information Criterion (BIC), Minimum Description Length (MDL), Mallow's C_p.

Choosing between models

In their standard forms, many of the expressions for these methods are thought for nested models ($y = a, y = a + bx, y = a + bx + cx^2, ...$). But this is not necessarily so. For nested models we can also use:

- Partial F test

$$F = \frac{\frac{SS^{reduced}_{residuals} - SS^{full}_{residuals}}{p_{full} - p_{reduced}}}{\frac{SS^{full}_{residuals}}{N - p_{full}}}$$

- AIC, BIC, MDL
- Likelihood ratio test
- Wald test
- Score (Lagrange multiplier) test

For non-nested models we can use:

- AIC
- Relative likelihood test

You cannot compare models fitted to different datasets.

Choosing between models

Common mistakes

- You cannot compare models fitted to different datasets.
- You cannot use techniques designed for nested model for non-nested models.
- You cannot compare models whose predictions are undistinguishable in the range of X.
- Fitting models that do not make scientific sense.
- Fitting lots of models and accepting the one that fits best. This is a kind of multiple testing.

Regression models

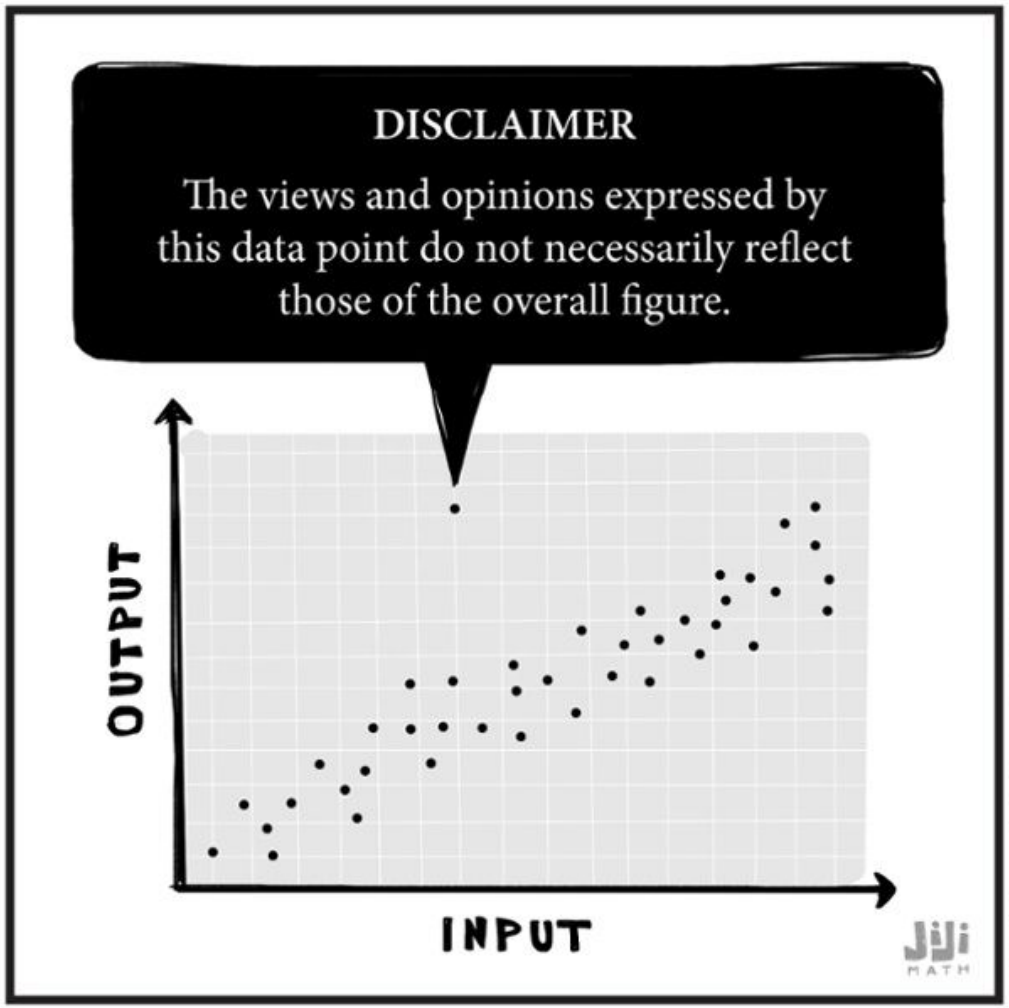
Now I have a model of how confidence and alcohol affect sociability.
How do I know if it is a valid model?

Validation strategies or Model selection

- K-Fold cross-validation: Train your model with part of your data (9/10), and use the rest of the data (1/10) to validate. Repeat this procedure 10 times.
- Leave-one-out: Build N different models, each time leave a different observation out of the training set, and use the new built model to predict it. Repeat this procedure with all observations.
- Bootstrap: Take a bootstrap sample. Build a model with them. Predict the values of the observations that were not in the training set. This estimate is biased but it can be corrected (0.632 correction)

Average the prediction error of each trial.

Regression models



Outline

Multiple regression

We are studying the relationship between several factors and kidney function, measured through creatining clearance. For women (second column) we have found

$$CrCl = 99 - 12.64 \log C_{Pb} - 0.05 Age - 0.006 Age^2 + 0.92 BMI - 7.56 Therapy$$

Table 3. Determinants of the Measured Creatinine Clearance Rate.

Variable*	Relation with Lead		Relation with Zinc Protoporphyrin	
	Men (N = 965)	Women (N = 1016)	Men (N = 965)	Women (N = 1016)
R^2	0.27	0.25	0.26	0.25
Intercept	86	99	66	79
Partial regression coefficient				
Log lead (μg/liter)	−9.51†	−12.64‡	—	—
Log zinc protoporphyrin (μg/g of hemoglobin)	—	—	−8.88†	−7.72§
Age (yr)	0.71‡	−0.05¶	0.43¶	−0.22¶
Age squared	−0.015‡	−0.006‡	−0.012‡	−0.005‡
Body-mass index	1.76‡	0.92‡	1.72‡	0.99‡
Log γ-glutamyl transpeptidase (U/liter)	−6.38†	NS¶	NS¶	NS¶
Diuretic therapy	−8.77†	−7.56‡	−8.48†	−8.04‡

*The linear and squared terms of age were tested simultaneously for entry into the regression model. Diuretic therapy was coded as 0 (not taking diuretic agents) or 1 (currently taking diuretic agents).

†P≤0.05. ‡P≤0.001. §P = 0.08. ¶NS denotes P not significant.

Staessen, J.A. et al. New England J. Medicine, 327: 151-156 (1992)

Multiple regression

- Diagnostic tools are also valid for multiple regression.
- Interactions are modeled by new variables merging both variables

$$Age \cdot BMI, \frac{BMI}{Age}, Age^2 \cdot BMI, Age^{\alpha} \cdot BMI, ...$$

The way of analyzing interactions can be tricky, and the fitting assumes that there is no other form of interaction other than that specified.

Multiple regression

What should the sample size be?

$$N_{observations} \geq \max\left\{50 + 8N_{predictors}, 104 + N_{predictors}\right\}$$

Cohen's effect size $f^2 = \frac{R^2}{1-R^2}$

$$\alpha = 0.05; \beta = 0.2; f^2 = 0.15$$

Medium

Determine minimum effect size for the regression coefficients (f^2), the confidence level ($1-\alpha$), the statistical power ($1-\beta$) and the number of predictors k.

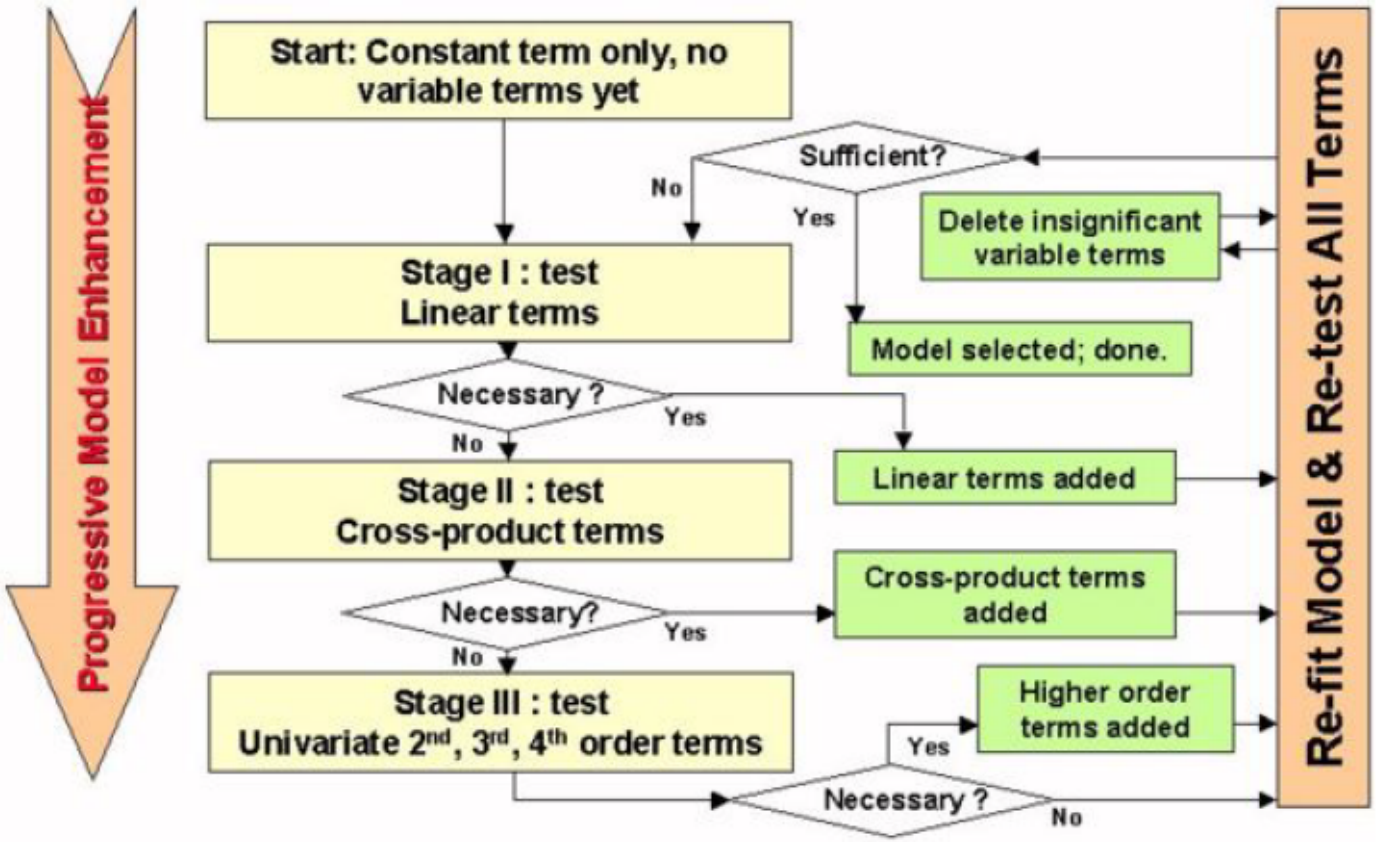
- Compute degrees of freedom of denominator $df = k + 1$
- Compute $F_{1-\alpha,k,df}$ with a central F distribution with k and df degrees of freedom
- Compute the centrality parameter $\lambda = f^2(k + df + 1)$
- Compute the current power with a noncentral F (NCF) $Power = \int_0^{F_{1-\alpha,k,df}} f_{NCF_{k,df,\lambda}}(x)dx$
- Repeat Steps 2-5 until the power is the desired one, and increase in each iteration the number of degrees of freedom of the denominator by 1.

Online calculator: http://www.danielsoper.com/statcalc/calc01.aspx

Other authors defend 10 – 20 observations per predictor.

Multiple regression

Stepwise forward regression: Add variables one or a group at a time while significant

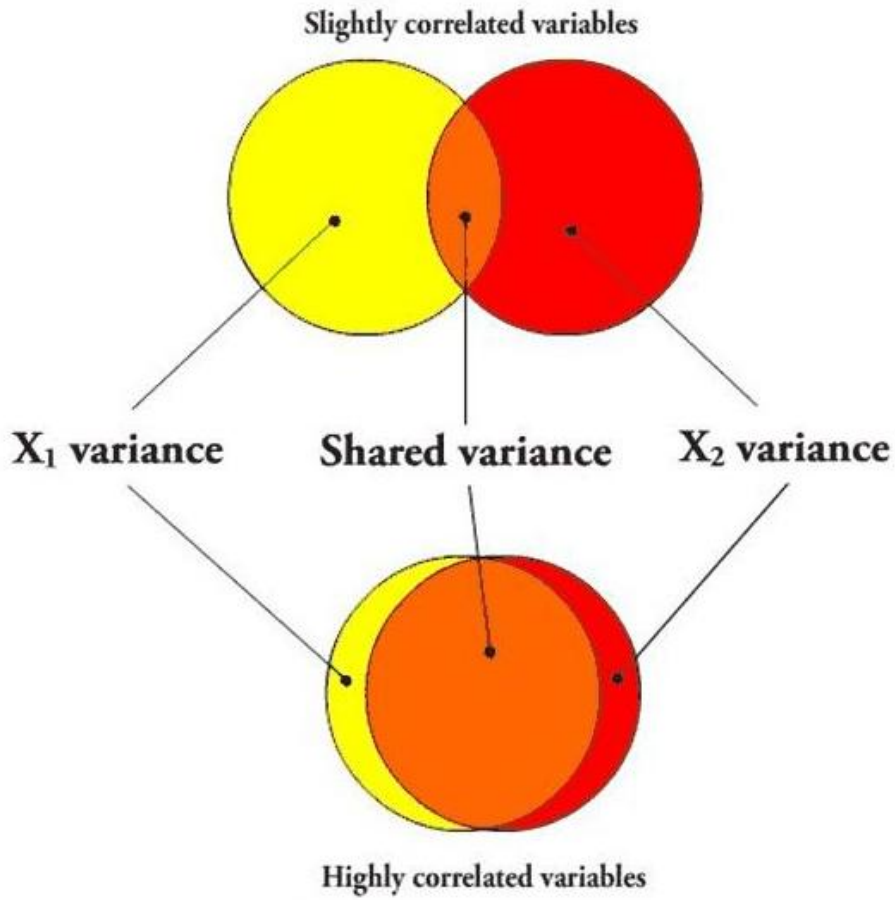


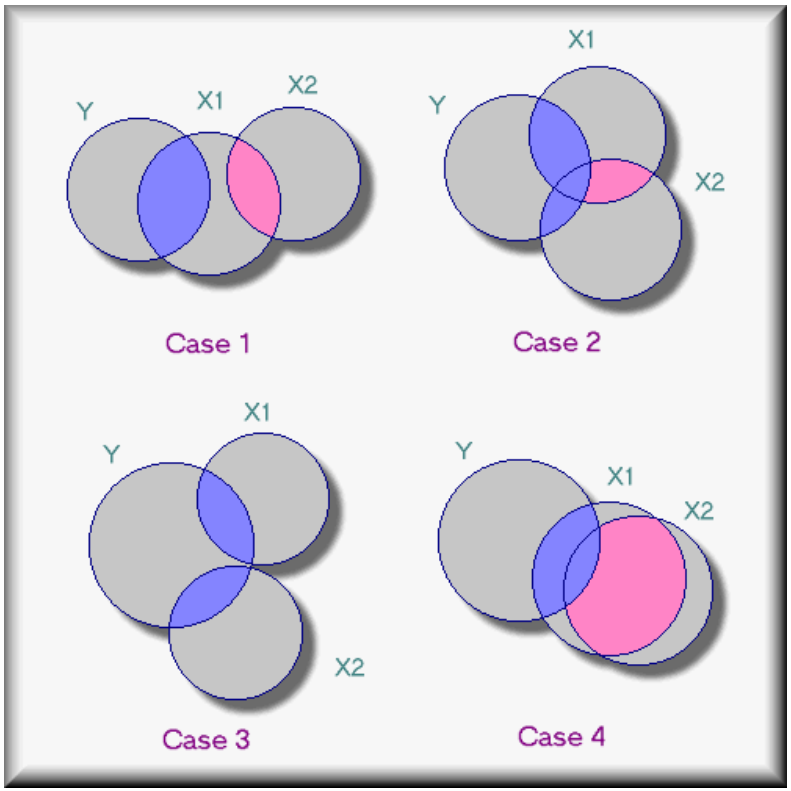
Stepwise backward regression: Remove variables one at a time while not significant

Be careful with R^2 inflation by multiple testing.

Multiple regression

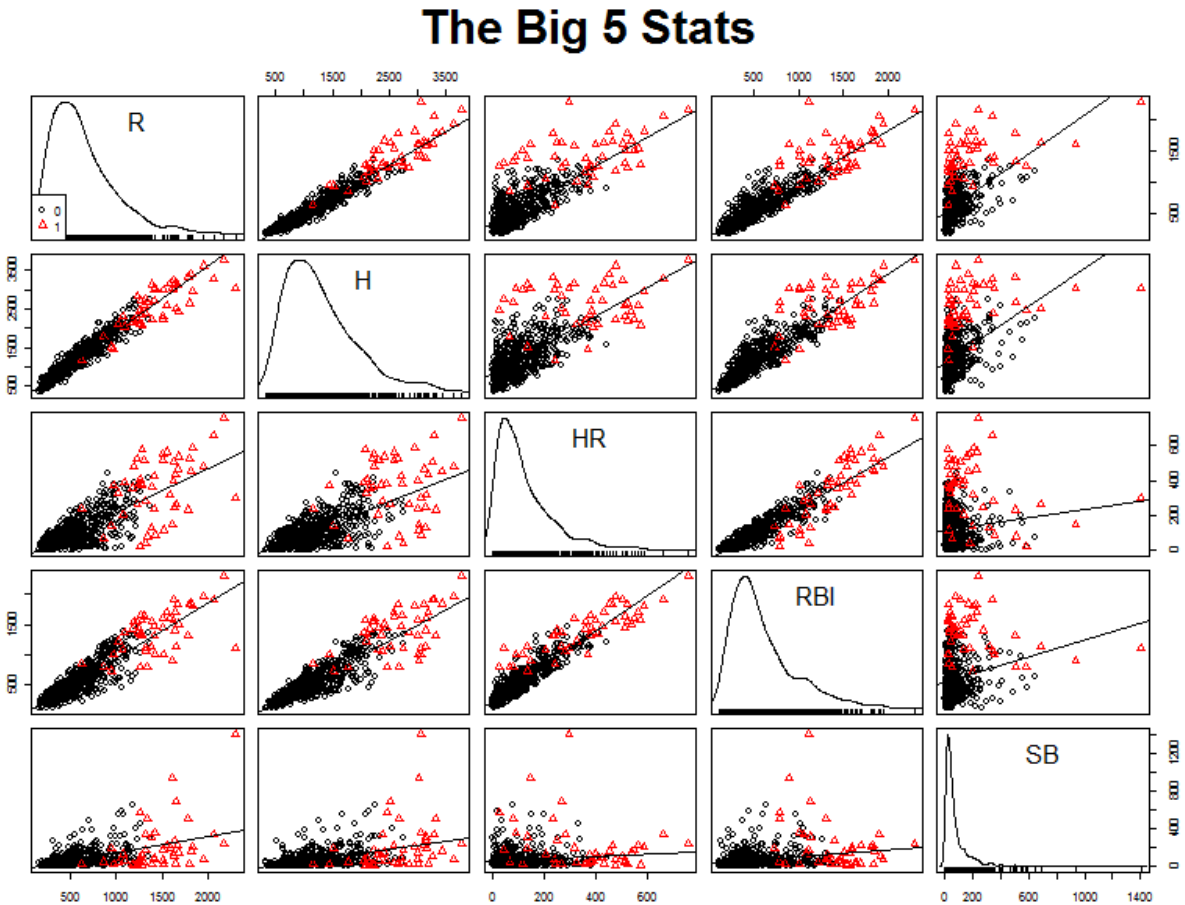
Be careful with Multicollinearity since it results in ill defined models (wide CI).
Solution: Partial Least Squares.





Multiple regression

Some simple forms of multicollinearity can be easily seen in scatterplots.



Outline

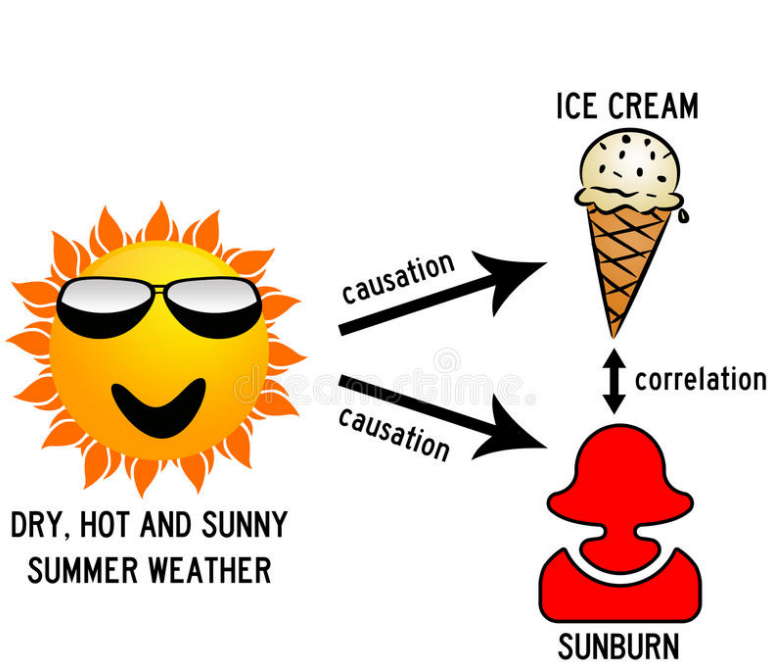
Regression does not imply causation

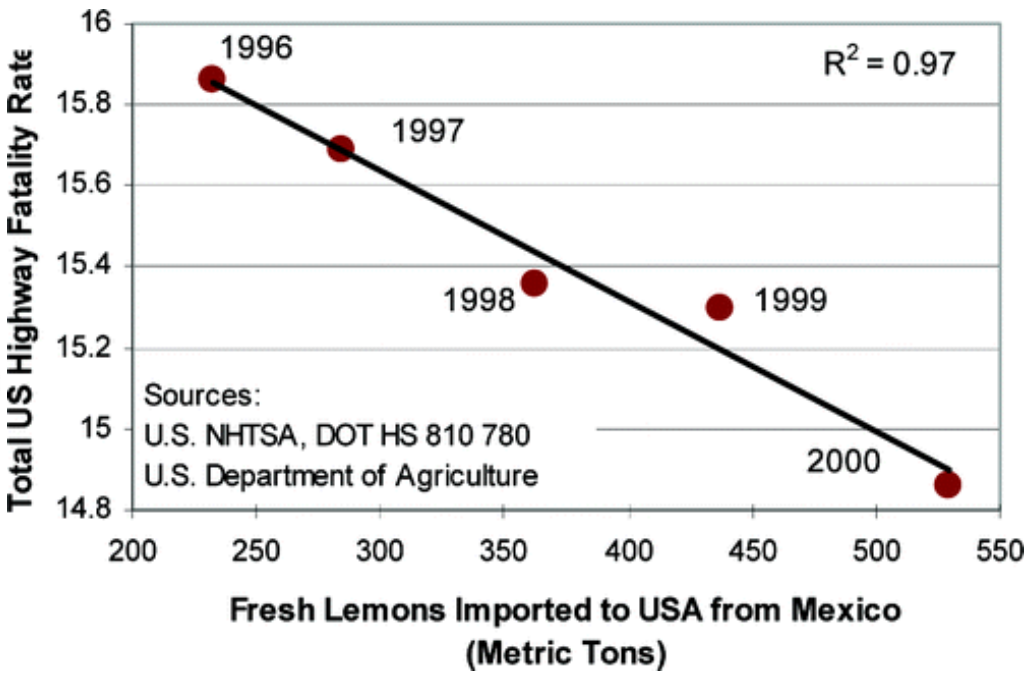
Assume we perform an experiment and discover that there is a relationship between lead concentration in blood and kidney function (measured by creatinine clearance).

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

Can we assess that lead exposure causes kidney malfunctioning?

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

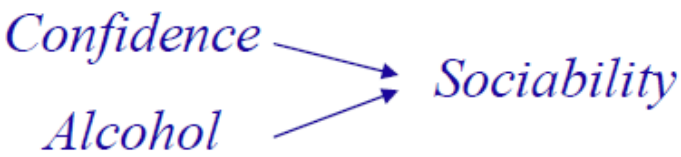




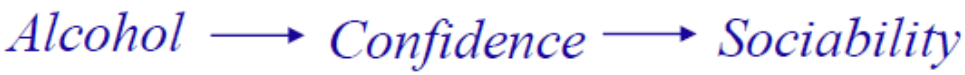
Checking for causation

Stepwise forward regression and causality

We want to measure what is the relationship between sociability, self-confidence and alcohol. There are two possible causal models:



Alcohol does not affect sociability by giving higher self-confidence.



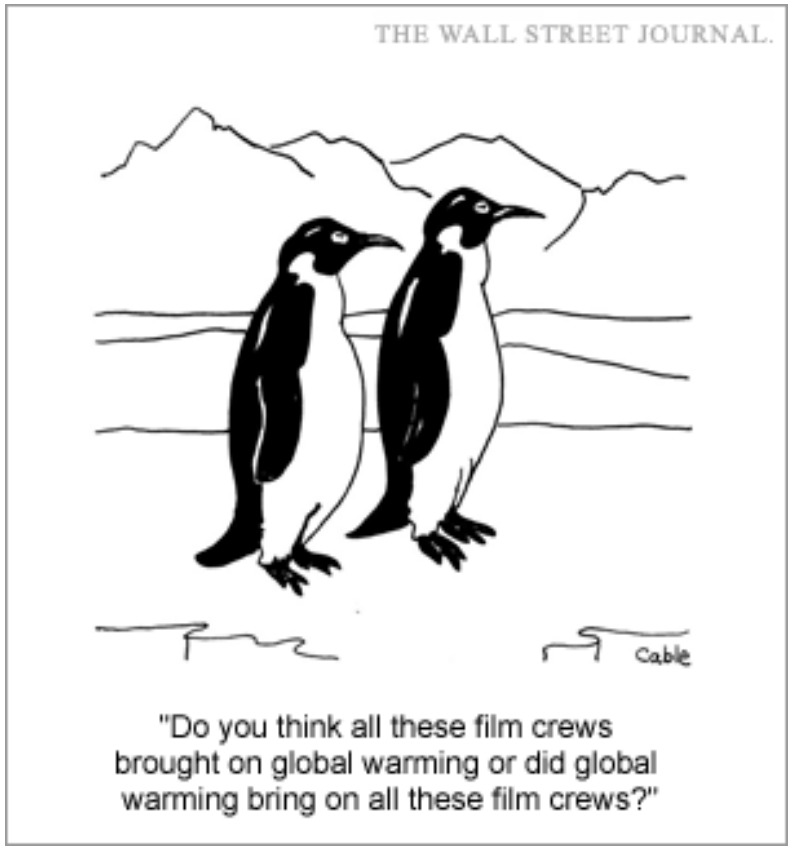
Alcohol affects sociability by giving higher self-confidence.

Let us build a regression of sociability (Y) as a function of confidence (X), and compute the residuals.

The regression of this residuals with alcohol should give nonsignificant coefficients in the second model and significant coefficients in the first model.

If we have significant coefficients, we reject model 2 but we cannot accept model 1!!!

Causation


THE WALL STREET JOURNAL.
Cable
"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

Outline

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)

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

We will transform the problem into

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

Taking logarithms

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

p_{Obese} is the probability of being obese.

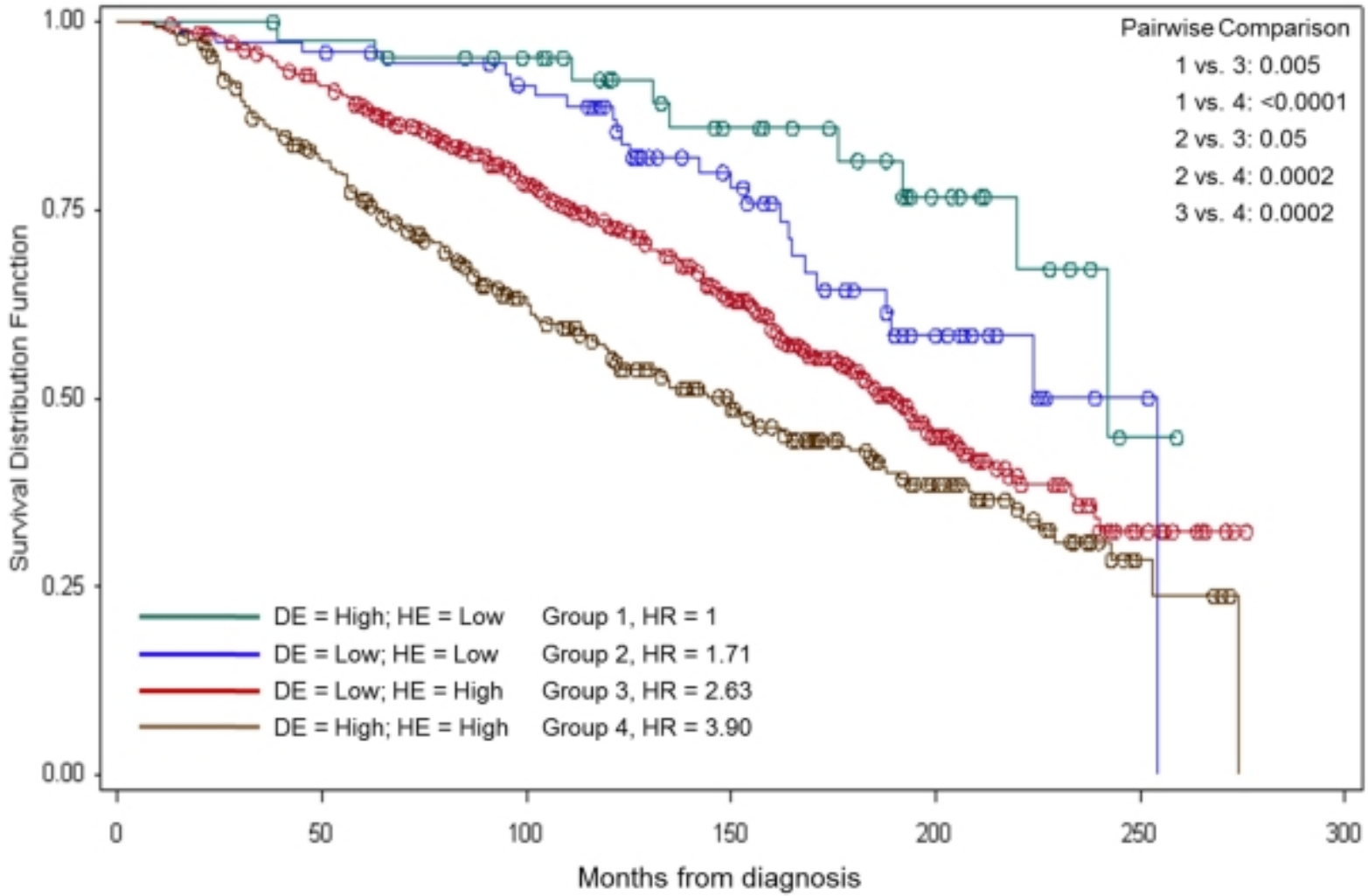
Logistic regression

We may interpret the βs 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.

Outline

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

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

Outline

7.3 ANOVA as a model

Factorial approach: factors are varied together

$$Mean = 1.61$$

$$Effect_{Plays} = \frac{2+2+1+4+3+2+1+1+2}{9} - 1.61 = 0.39$$

$$Effect_{DoesNotPlay} = \frac{2+0+1+1+3+2+2+0+0}{9} - 1.61 = -0.39$$

	4-4-2	4-3-3	4-2-3-1
Plays	2, 2, 1	4, 3, 2	1, 1, 2
Does not play	2, 0, 1	1, 3, 2	2, 0, 0

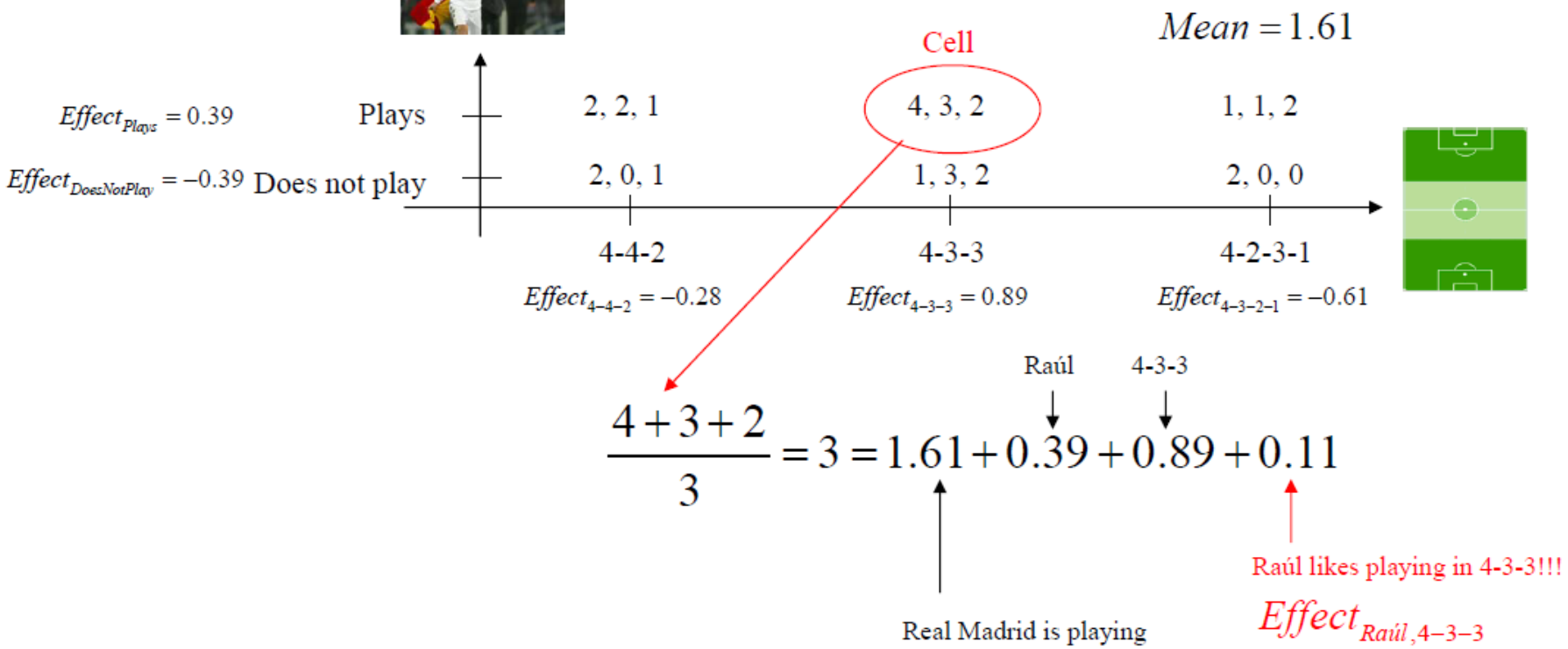
$$Effect_{4-4-2} = \frac{2+2+1+2+0+1}{6} - 1.61 = -0.28$$

$$Effect_{4-3-2-1} = \frac{1+1+2+2+0+0}{6} - 1.61 = -0.61$$

$$Effect_{4-4-2} = \frac{4+3+2+1+3+2}{6} - 1.61 = 0.89$$

7.3 ANOVA as a model

Factorial approach: factors are varied together



7.3 ANOVA as a model

Analysis of Variance: ANOVA

$Mean = 1.61$

$Effect_{Plays} = 0.39$ Plays: 2, 2, 1 | 4, 3, 2 | 1, 1, 2

$Effect_{DoesNotPlay} = -0.39$ Does not play: 2, 0, 1 | 1, 3, 2 | 2, 0, 0

4-4-2 | 4-3-3 | 4-2-3-1

$Effect_{4-4-2} = -0.28$ | $Effect_{4-3-3} = 0.89$ | $Effect_{4-3-2-1} = -0.61$

Raúl 4-3-3 Noise

$$4 = 1.61 + 0.39 + 0.89 + 0.11 + 1$$

Real Madrid is playing | Rául likes 4-3-3

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

7.3 ANOVA as a model

Analysis of Variance: ANOVA $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

	Variance	Degrees of freedom	
Mean	0	0	
Raúl effect (treatment)	"-0.39,+0.39"	1=a-1	
Strategy effect (treatment)	"-0.28,0.89,-0.61"	2=b-1	ab-1
Interactions Raúl-Strategy	"0.11,…"	2=(a-1)(b-1)	
Residual	"1,…"	12=N-1-(ab-1)=N-ab=ab(r-1)	
Total	"2,2,1,4,3,2,1,1,2, 2,0,1,2,3,2,2,0,0"	17=N-1	

r=number of replicates per cell
N=total number of experiments
a=number of different levels in treatment A
b=number of different levels in treatment B

7.3 ANOVA as a model

Single Measures:

1. If there are only two treatments is equivalent to two-sample t-test
2. If there are more, it is called between-subject ANOVA

Drug 1	Drug 2	Drug 3	Placebo
Group 1	Group2	Group3	Group4

Assumptions

1. Homogeneity of variance
2. Normality
3. Independence of observations

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

τ_j: Drug effect

X_{ij}: Individual score

Repeated Measures

1. If there are only two treatments is equivalent to paired-sample t-test
2. If there are more, it is called within-subject ANOVA

Drug 1	Drug 2	Drug 3	Placebo
Subj. 1	Subj. 1	Subj. 1	Subj. 1
Subj. 2	Subj. 2	Subj. 2	Subj. 2
Subj. 3	Subj. 3	Subj. 3	Subj. 3
….	….	….	…..

Assumptions

1. Homogeneity of variance
2. Normality
3. Homogeneity of correlation

$$X_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij}$$

π_i: Individual effect

7.3.1 What is ANOVA really?

ANOVA is a hypothesis test about means, not about variance!!

1-way ANOVA: $X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_K$$
$$H_1 : \exists i, j \mid \alpha_i \neq \alpha_j$$

There is no effect of the treatment

2-way ANOVA: $X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_K$$
$$H_1 : \exists i, j \mid \alpha_i \neq \alpha_j$$

There is no effect of the first treatment

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{K'}$$
$$H_1 : \exists i, j \mid \beta_i \neq \beta_j$$

There is no effect of the second treatment

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{KK'}$$
$$H_1 : \exists i, j, k, l \mid (\alpha\beta)_{ij} \neq (\alpha\beta)_{kl}$$

There is no interaction

7.3.1 What is ANOVA really?

How are the tests performed? F-tests

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$\mu = 1.61$

$\alpha_{Raul} = 0.39$ $\quad \alpha_{NoRaul} = -0.39$

$\beta_{4-4-2} = -0.28$ $\quad \beta_{4-3-3} = 0.89$ $\quad \beta_{4-3-2-1} = -0.61$

$(\alpha\beta)Raul._{4-4-2} = -0.05$ $\quad (\alpha\beta)Raul._{4-3-3} = 0.11$ $\quad (\alpha\beta)Raul._{4-3-2-1} = -0.06$

$(\alpha\beta)NoRaul._{4-4-2} = 0.05$ $\quad (\alpha\beta)NoRaul._{4-3-3} = -0.11$ $\quad (\alpha\beta)NoRaul._{4-3-2-1} = 0.06$

Source	SumSquare	df	MeanSquare	F	Prob>F	
Raúl	2.7222	1	2.7222	3.2667	0.0958	$F_{1,12}$
Strategy	7.4444	2	3.7222	4.4667	0.0355	$F_{2,12}$ → Significant at a confidence level of 95%
Interaction	0.1111	2	0.0556	0.0667	0.9359	$F_{2,12}$
Error	10.0000	12	0.8333			
Total	20.2778	17				

→ Also called within-groups

MSE: This value gives an idea of the goodness-of-fit

$$MS_i = \frac{SS_i}{df_i} \qquad F_i = \frac{MS_i}{MS_{Error}}$$

7.3.1 What is ANOVA really?

How are the different terms computed?

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

The sum of squares is computed using the information explained by the elements involved

SS Total

$$SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \hat{\mu})^2$$

SS explained by variable α

$$SS_{\alpha} = JK\sum_{i=1}^{I}(\bar{x}_{i.} - \hat{\mu})^2$$

SS explained by variable β

$$SS_{\beta} = IK\sum_{j=1}^{J}(\bar{x}_{.j} - \hat{\mu})^2$$

SS explained by interaction αβ

$$SS_{\alpha\beta} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2$$

SS Residuals

$$SS_{E} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha}\hat{\beta}_{ij}))^2$$

7.3.1 What is ANOVA really?

Index of effect size: "Coefficient of determination"

$$\eta_{\alpha}^2 = \frac{SS_{\alpha}}{SS_{total}} \qquad \eta_{\beta}^2 = \frac{SS_{\beta}}{SS_{total}} \qquad \eta_{\alpha\beta}^2 = \frac{SS_{\alpha\beta}}{SS_{total}}$$

Source	SumSquare	df	MeanSquare	F	Prob>F
Raúl	2.7222	1	2.7222	3.2667	0.0958
Strategy	7.4444	2	3.7222	4.4667	0.0355
Interaction	0.1111	2	0.0556	0.0667	0.9359
Error	10.0000	12	0.8333		
Total	20.2778	17			

$$\eta_{Raúl}^2 = \frac{SS_{Raúl}}{SS_{total}} = \frac{2.7222}{20.2778} = 0.13$$

$$\eta_{strategy}^2 = \frac{SS_{strategy}}{SS_{total}} = \frac{7.4444}{20.2778} = 0.37 \longrightarrow$$ This is the only one coming from a significant (95%) effect

$$\eta_{interaction}^2 = \frac{SS_{interaction}}{SS_{total}} = \frac{0.1111}{20.2778} = 0.01$$

7.3.2 What is ANCOVA?

Analysis of Covariance=Regression+ANOVA

Situation: We want to study the effect of vocabulary level in crossword solving performance. We form three groups of 10 people according to their vocabulary level (High, Medium, Low). The ANOVA summary table is as follows.

Source	SumSquare	df	MeanSquare	F	Prob>F
Vocabulary	50.00	2	25.00	13.5	8.6e-5
Error	50.00	27	1.85		
Total	100.00	29			

$F_{2,27}$

Situation: We discover that, whichever the group, age has a strong influence on the crossword performance. That is, part of the variability is explained by a covariate (the age) that we can measure but not control. Then, we try to explain (through linear regression) part of the performance.

Source	SumSquare	df	MeanSquare	F	Prob>F
Age	20.00	1	20.00	17.3	3.0e-4
Vocabulary	50.00	2	25.00	21.7	2.8e-6
Error	30.00	26	1.15		
Total	100.00	29			

$F_{1,26}$
$F_{2,26}$

7.3.3 How do I use them with pretest-posttest designs?

Situation: We want to study the effect of a new treatment on patients. We use a pretest to evaluate the state of the patient and a posttest to evaluate the improvement. We want to use a control group (which may not be different from the treatment group) so that the results can be generalized to the population

	Standard	New
Pretest	Subj. 1 Subj. 2 Subj. 3	Subj. A Subj. B Subj. C
Posttest	Subj. 1 Subj. 2 Subj. 3	Subj. A Subj. B Subj. C

Option A: Repeated measurements ANOVA

$$X_{ijk} = \mu + Person_i + Treatment_j + Time_k + \varepsilon_{ijk}$$

$$X_{ij} = PostTest_{ij} - PreTest_{ij} = \mu + Person_i + Treatment_j + \varepsilon_{ij}$$

These two models are equivalent. The latter is called Gain Scores ANOVA.

Option B: Repeated measurements ANCOVA

$$X_{ij} = PostTest_{ij} - PreTest_{ij} = \mu + a \cdot PreTest_{ij} + Person_i + Treatment_j + \varepsilon_{ij}$$

More powerful analysis!!

7.3.4 What are planned and post-hoc contrasts?

What is a contrast?

Situation: We want to diversify our investments in 4 different types of business: stocks, housing, fixed rate funds, and variable rate funds. We measure the interest in each of these businesses, and build a linear model

$$Rate = \mu + \alpha_{investment} + \varepsilon$$

We have the hypothesis that investing half of the money in stocks and housing, gives the same interest rate than investing in funds.

$$H_0 : \frac{\alpha_{stocks} + \alpha_{housing}}{2} - \frac{\alpha_{fixedFunds} + \alpha_{variableFunds}}{2} = 0$$

In general a contrast is:

$$H_0 : c_A\alpha_A + c_B\alpha_B + c_C\alpha_C + c_D\alpha_D = 0 \qquad H_0 : \langle \mathbf{c}, \boldsymbol{\alpha} \rangle = 0$$

$$c_A + c_B + c_C + c_D = 0 \qquad \langle \mathbf{c}, \mathbf{1} \rangle = 0$$

Two contrasts are orthogonal iff $\langle \mathbf{c}_1, \mathbf{c}_2 \rangle = 0$

$$H_0 : \frac{\alpha_{stocks} - \alpha_{housing}}{2} - \frac{\alpha_{fixedFunds} - \alpha_{variableFunds}}{2} = 0$$

is orthogonal to the previous contrast

7.3.4 What are planned and post-hoc contrasts?

Does this relate somehow to ANOVA?

Of course!! ANOVA hypothesis is that there is no effect

$$H_0 : \alpha_{stocks} = \alpha_{housing} = \alpha_{fixedFunds} = \alpha_{variableFunds}$$

If it is rejected, at least one of them is different from another one, but I don't know which pair!

I have to run post-hoc tests to detect which is the different pair. There are $\binom{4}{2} = 6$ pairs, and I have to test all of them

Pairwise comparisons

- Fisher's LSD
- Tukey's HSD
- Dunnett's Test
- Student-Neuman-Keuls test
- REGQW test

$$H_0 : \alpha_i = \alpha_j$$
$$H_1 : \alpha_i \neq \alpha_j$$

Any contrast:

- Scheffé Test
- Brown-Forsyth test

$$H_0 : \langle \mathbf{c}, \boldsymbol{\alpha} \rangle = 0$$
$$H_1 : \langle \mathbf{c}, \boldsymbol{\alpha} \rangle \neq 0$$

7.3.5 What are fixed-effects and random-effects?

Fixed effects: The experimenter controls the treatments applied to each individual

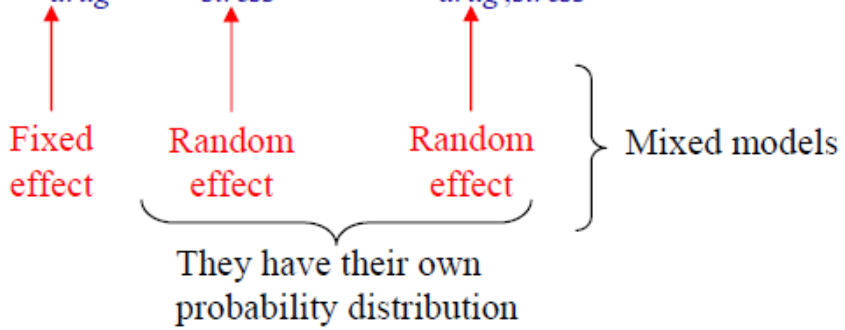
Situation: We want to compare the effect of 4 different drugs with respect to a placebo.

$$Improvement = \mu + \alpha_{drug} + \varepsilon$$

Random effects: The experimenter controls the treatments applied to each individual

Situation: We want to compare the effect of 4 different drugs with respect to a placebo, taking into account the stress that the patient has at work

$$Improvement = \mu + \alpha_{drug} + \beta_{stress} + (\alpha\beta)_{drug,stress} + \varepsilon$$



7.3.6 When should I use Multivariate ANOVA (MANOVA)?

What is MANOVA?

$$\textbf{Ability} = \mu + \alpha_{mathText} + \beta_{physicsText} + \gamma_{college} + \varepsilon$$

$$\textbf{Ability} = \begin{pmatrix} abilityMath \\ abilityPhysics \end{pmatrix}$$

$$H_0 : \alpha_{mathTextA} = \alpha_{mathTextB}$$
$$H_0 : \beta_{physicsTextA} = \beta_{physicsTextB}$$
$$H_0 : \gamma_{collegeA} = \gamma_{collegeB}$$

	Math Text A	Math Text B
Physics Text A College A	(9,9) (7,9) (10,6) (6,7)	(7,7) (4,5) (10,10) (9,9)
Physics Text A College B	(3,1) (5,5) (5,5) (5,5)	(6,7) (8,7) (8,8) (9,8)
Physics Text B College A	(2,8) (9,10) (10,10) (6,9)	(9,6) (5,4) (1,3) (8,8)
Physics Text B College B	(10,8) (7,5) (5,5) (6,5)	(6,6) (7,7) (8,3) (9,7)

When should I use MANOVA?

MANOVA is a very powerful technique but it is even more "omnibus" than ANOVA. If you expect correlation among the two dependent variables (ability in math, and ability in physics), then use MANOVA. If they are clearly independent, use two separate ANOVAs.

Outline

Chapter 0. Introduction to sample size calculations

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

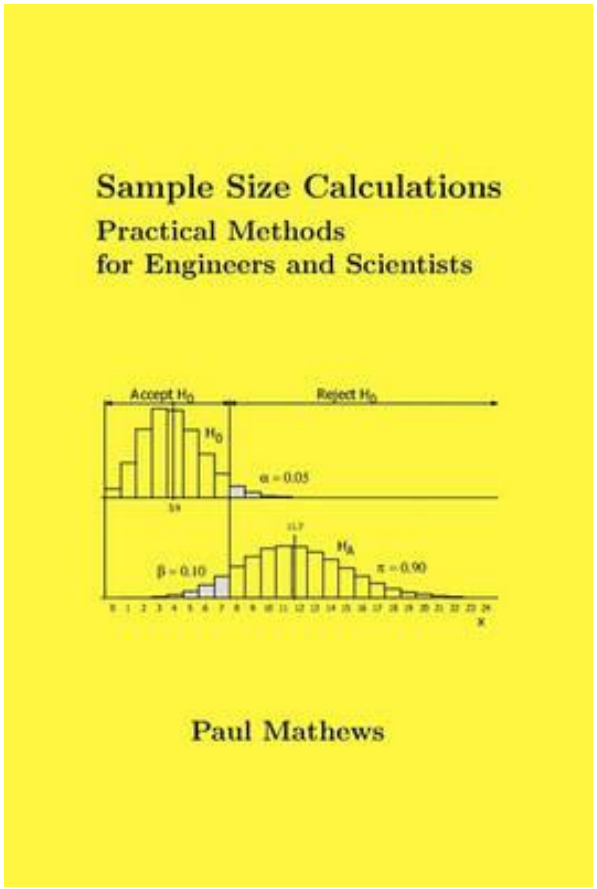
September 15, 2016

Outline

Sample size

$$N = \ldots$$

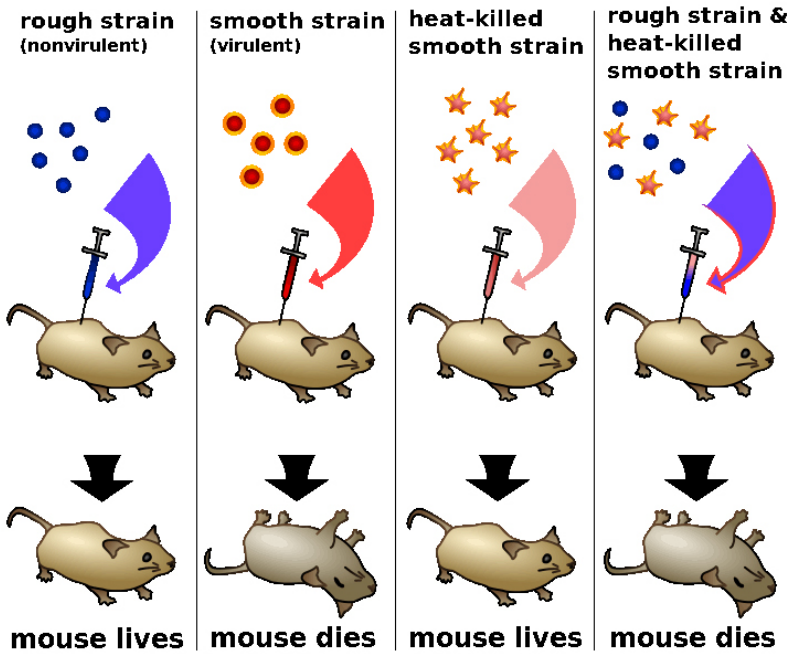
References



Paul Mathews. Sample size calculations. Practical methods for engineers and scientists. Mathews Malnar and Bailey, Inc. (2010)

Outline

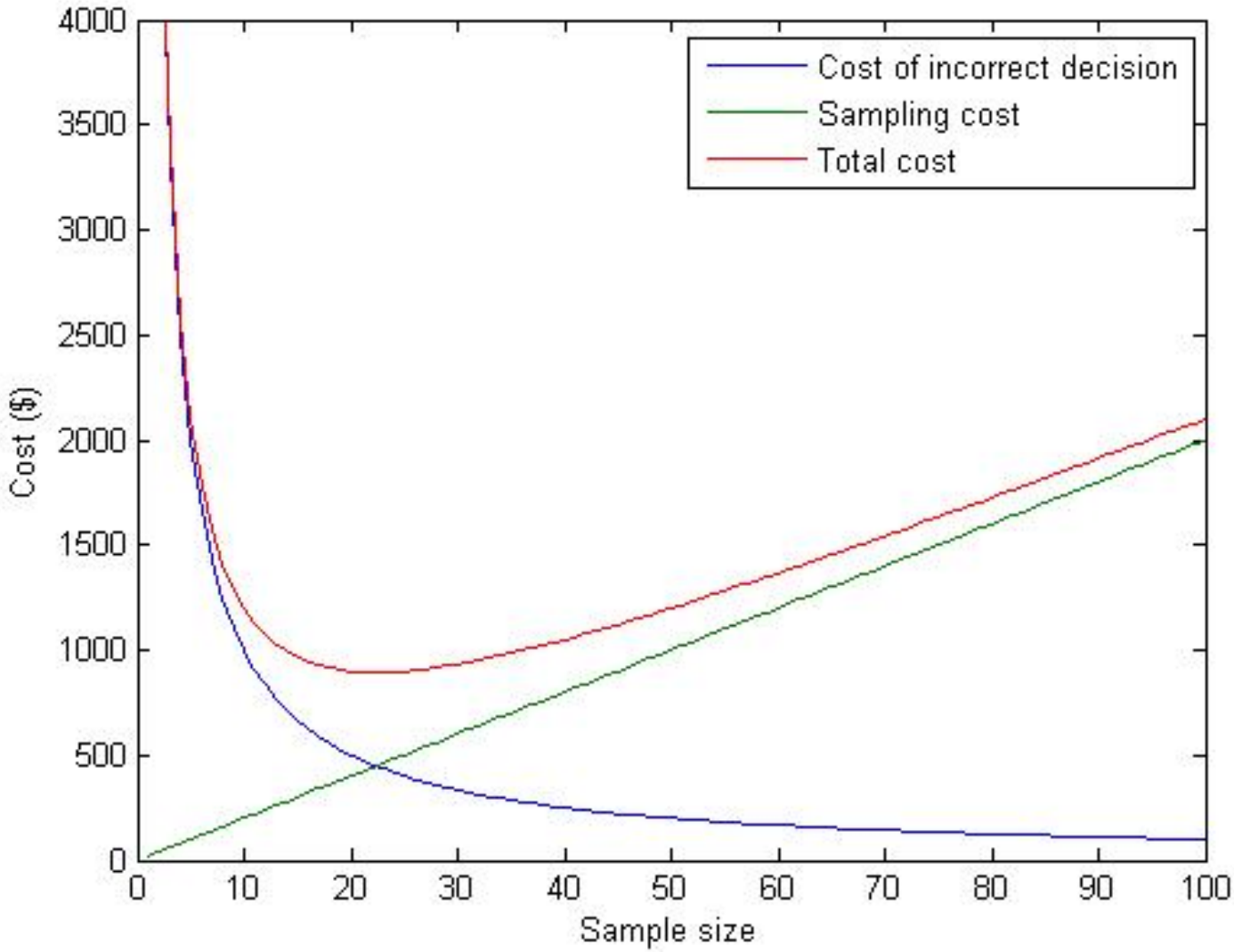
Why this course?



How many mice do I need to put in each group to show that a new vaccine is effective?

- Too few is a waste of time (=money) and money, it is unethical
- Too many is a waste of time (=money) and money, it is unethical
- I need just enough

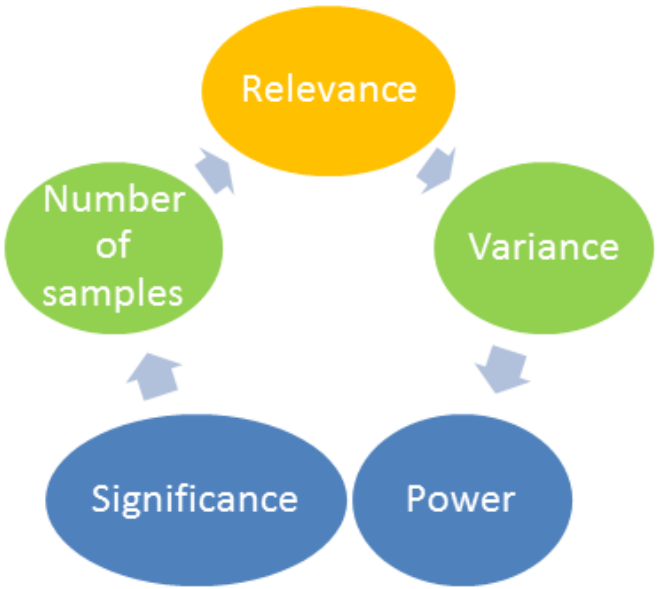
Why this course?



Why this course?

Unfounded fear

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



Why this course?

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!
  - ...

Why this course?

Why this course?

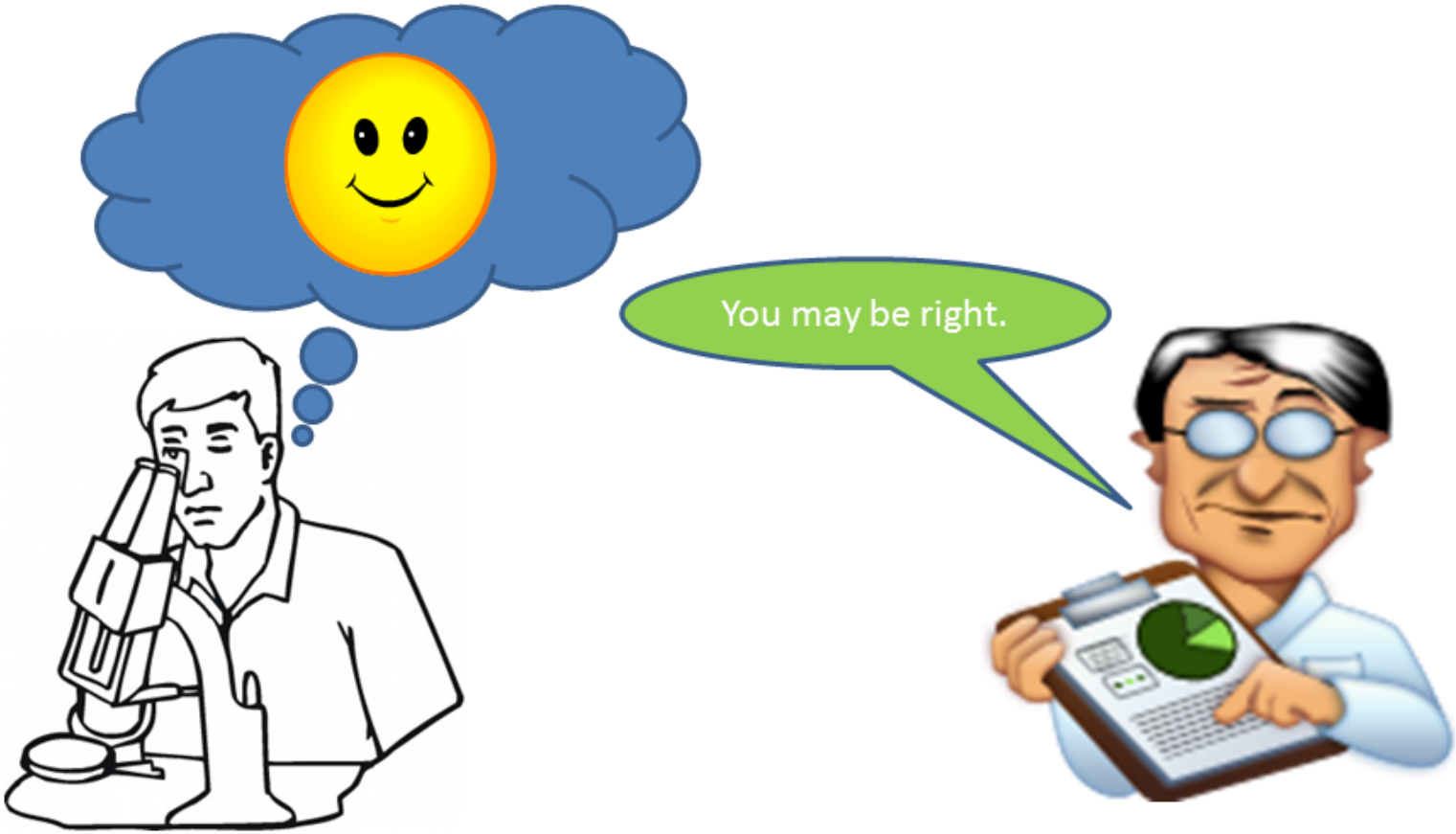
Stage 1: Arrival

Why this course?

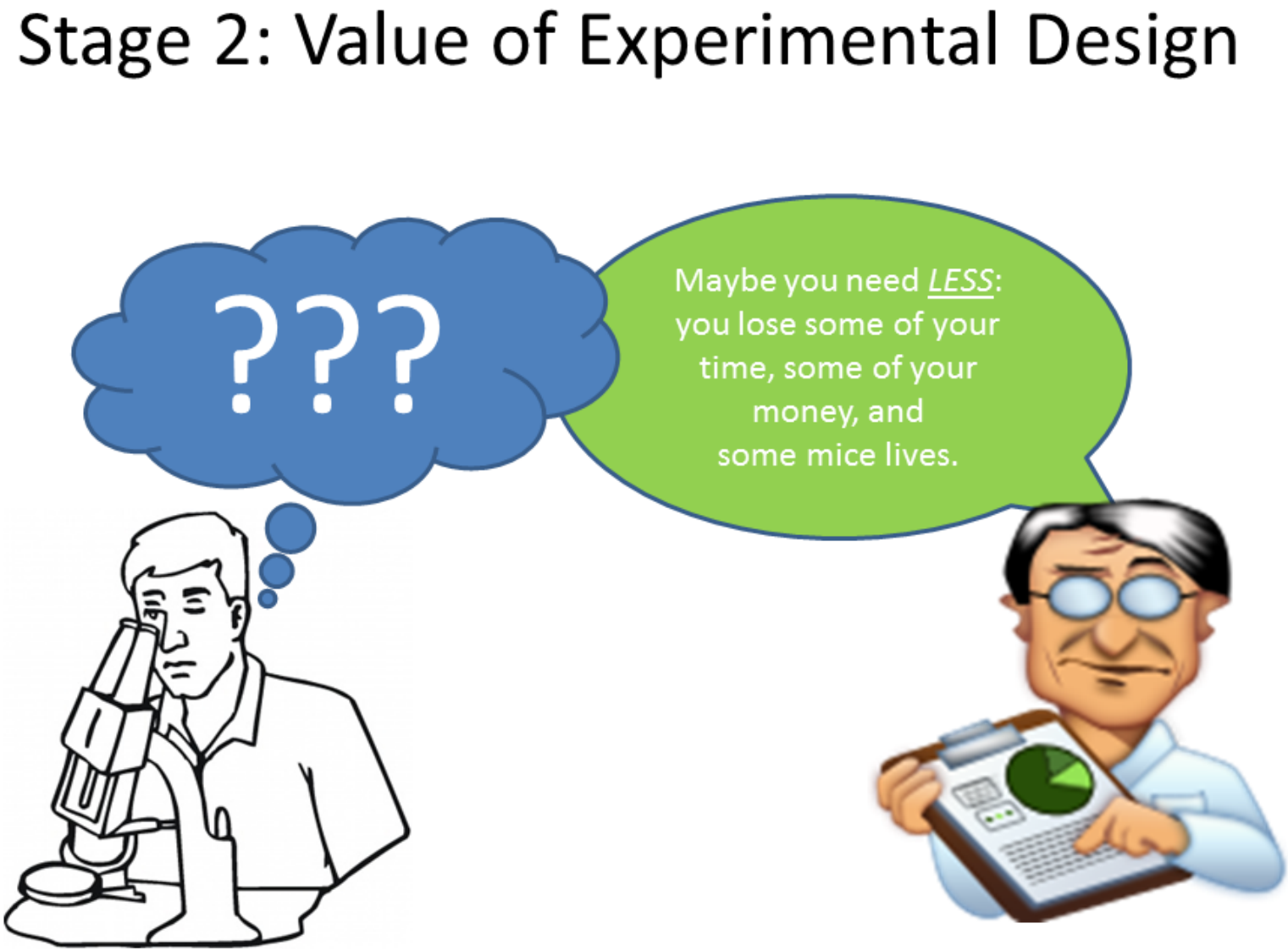
Stage 1: Arrival

Why this course?

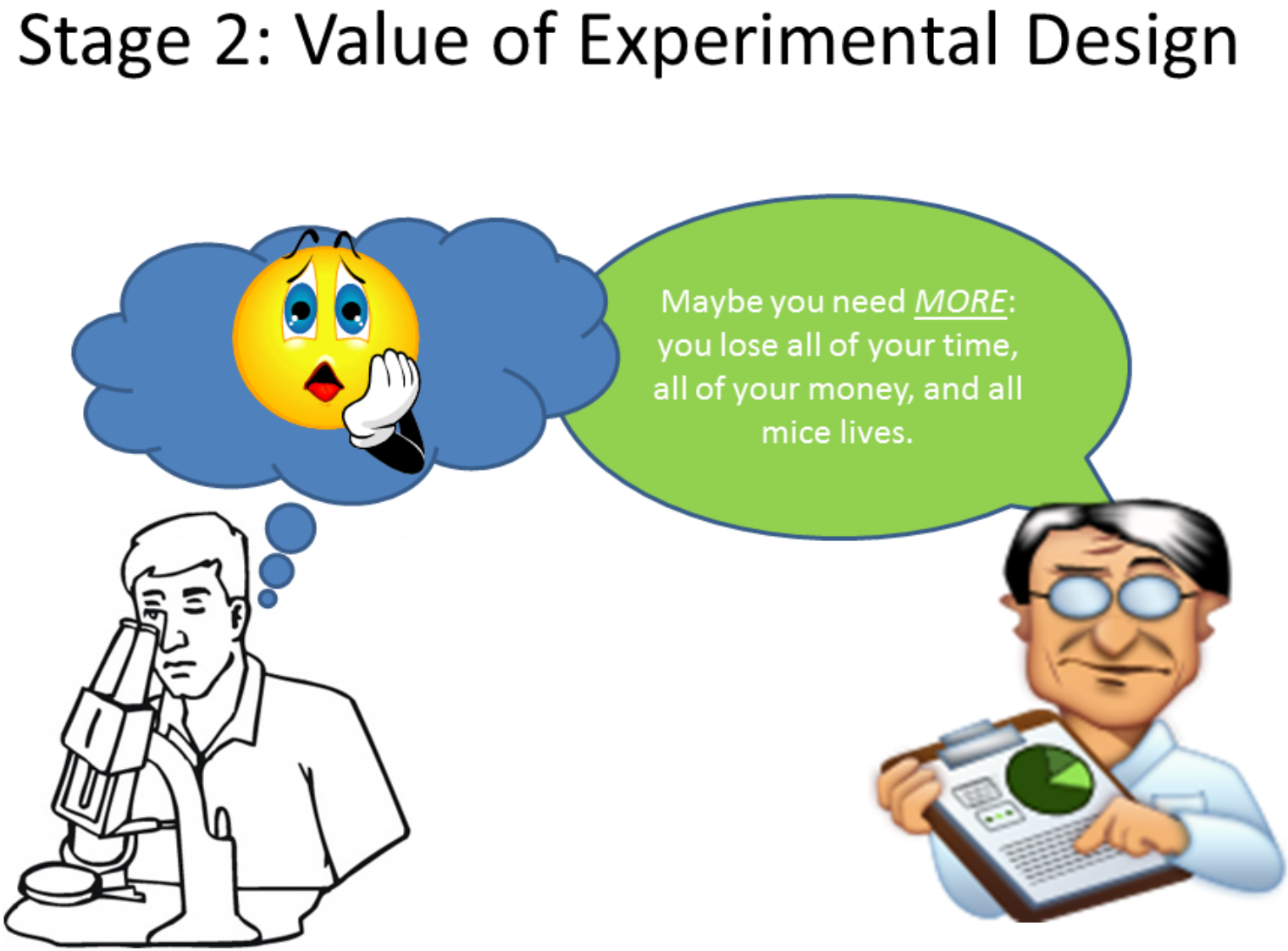
Stage 2: Value of Experimental Design



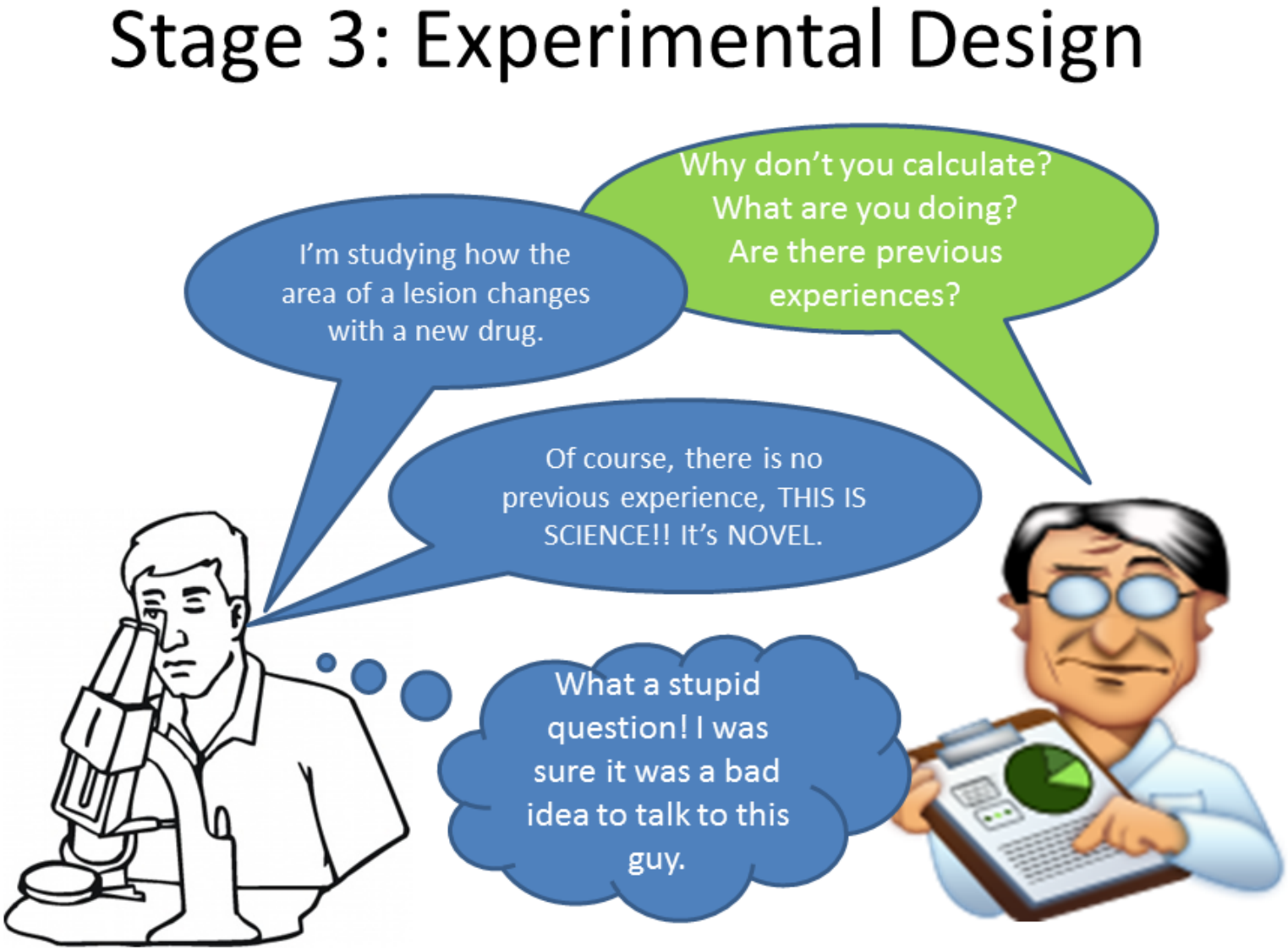
Why this course?



Why this course?

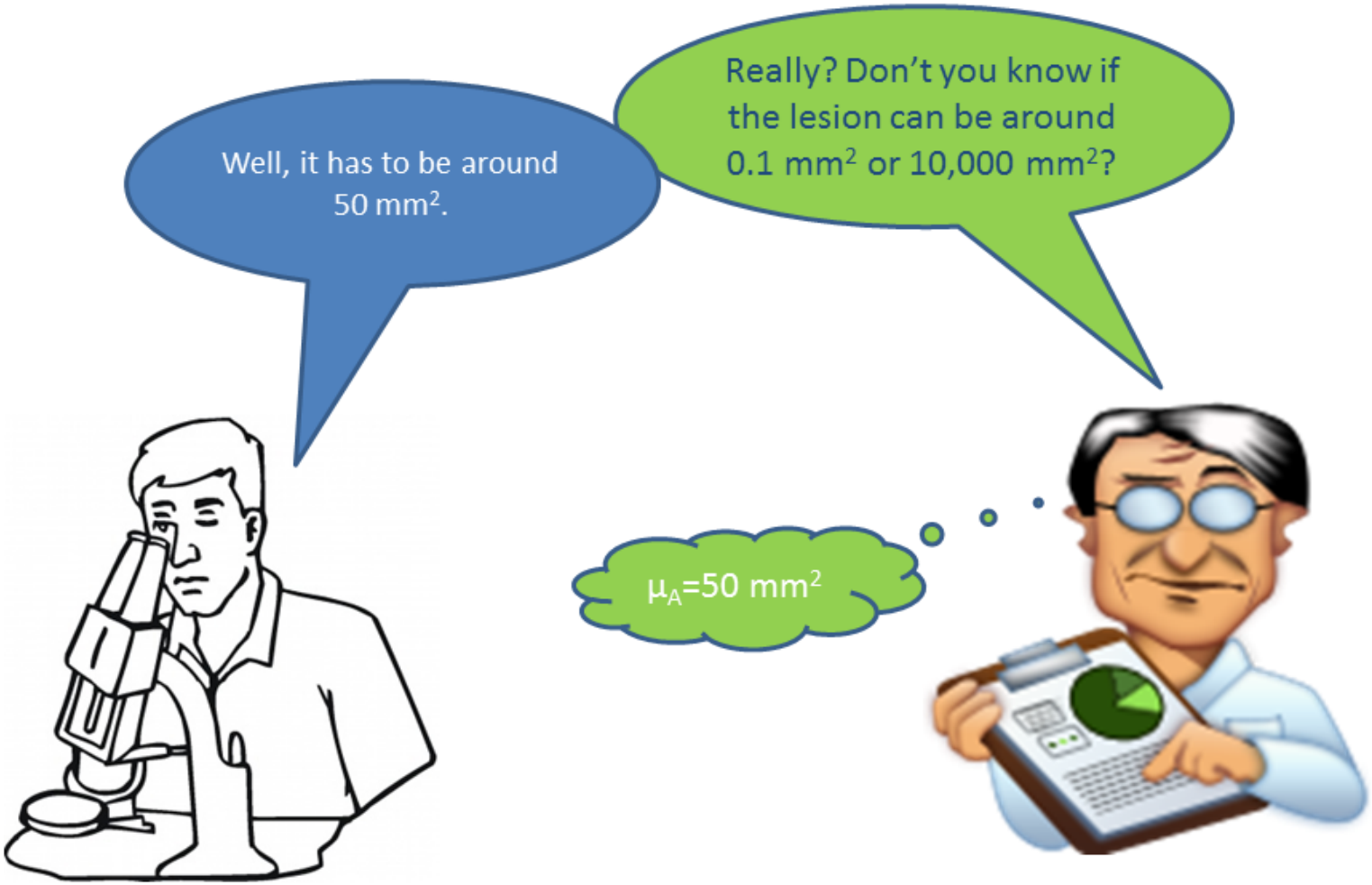


Why this course?



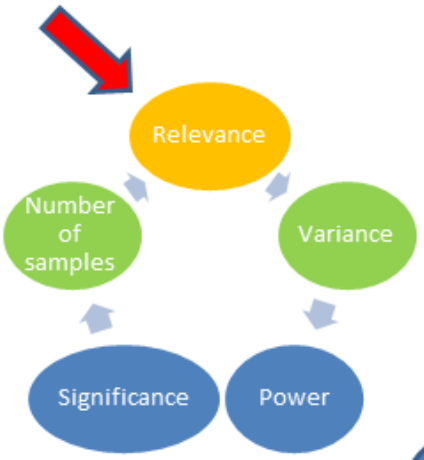
Why this course?

Stage 4: Gathering prior information



Why this course?

Stage 5: Setting a target for success

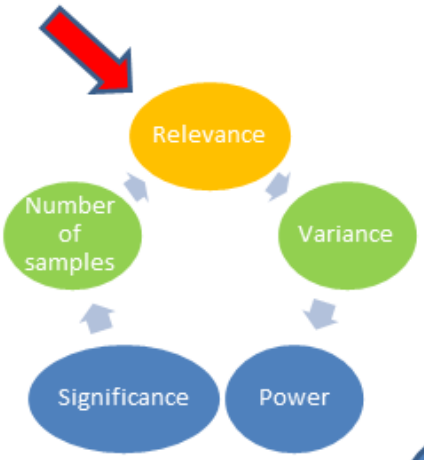


When would you say your drug is interesting?

I don't know, I just want to see what comes out from the experiment.

Why this course?

Stage 5: Setting a target for success



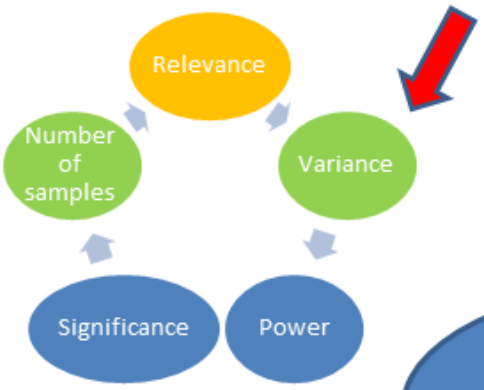
No, to be of any use, it has to decrease, at least, below 25 mm^2.

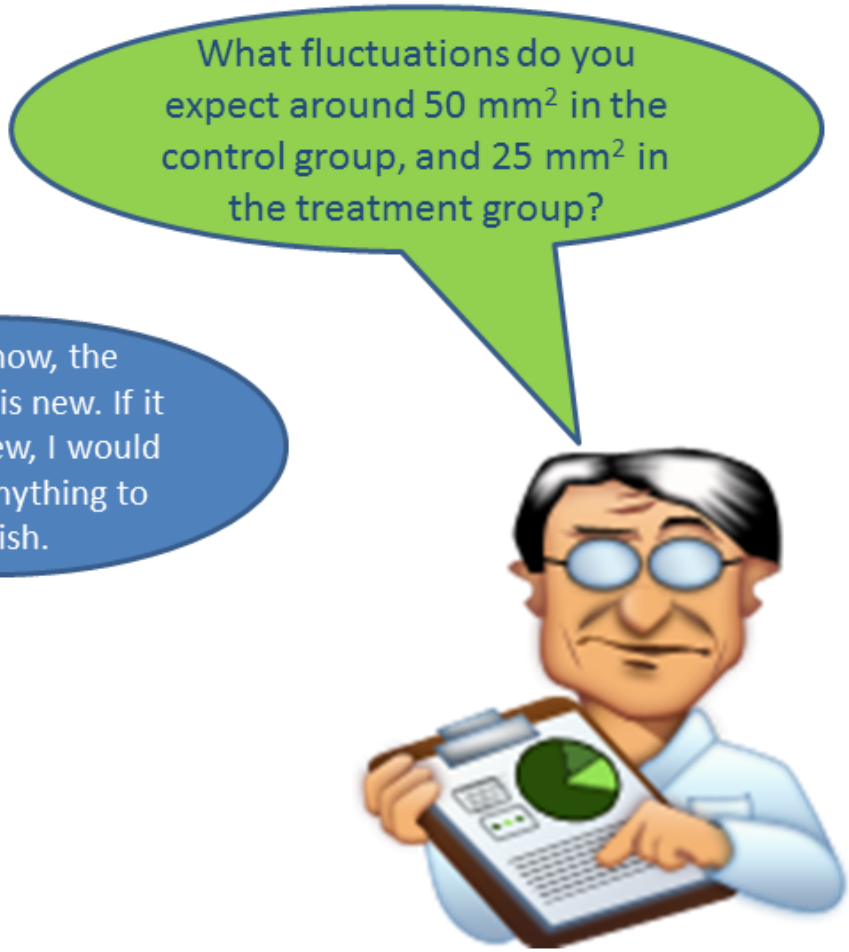
If it decreases to 49.5 mm^2, would you say it is worthy?

Δ=50-25= 25 mm^2. -50%

Why this course?

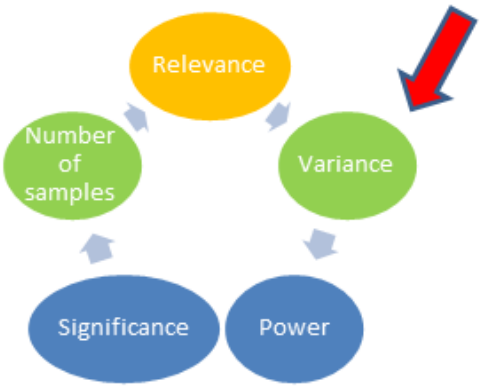
Stage 6: Gathering variance information



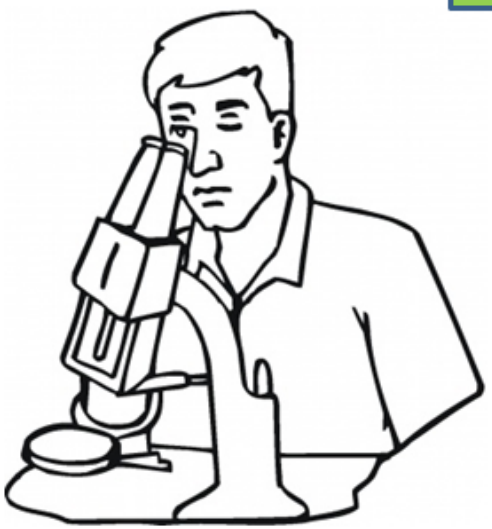


Why this course?

Stage 6: Gathering variance information

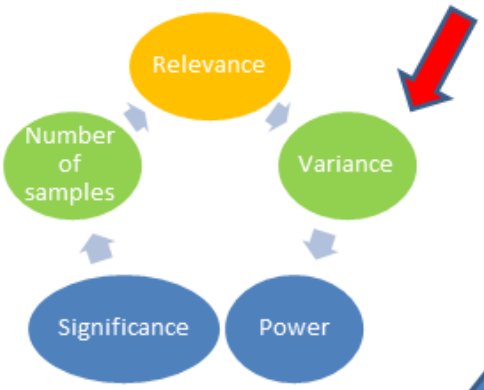


But you must know something a priori. When you measure the control group, do you expect values like
50.00, 49.96, 50.14, 50.14, 49.76 (σ^2=0.1)
or values like
58.15, 8.27, 31.96, 21.86, 129.75 (σ^2=1000)?
Both have a mean of 50, but you understand detecting a change of 25 in the second case is more difficult.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



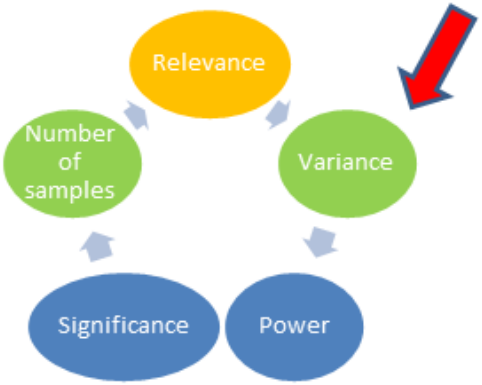
Why this course?

Stage 6: Gathering variance information



Why this course?

Stage 6: Gathering variance information

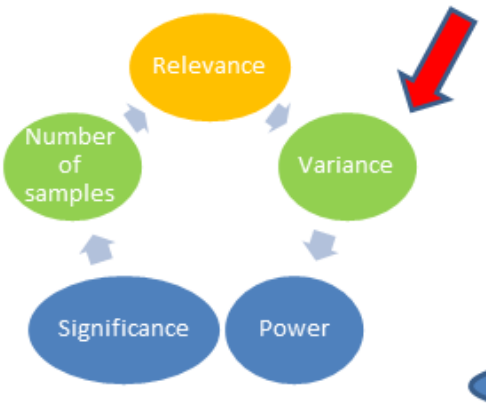


Well, ... there is a paper where they perform something similar to my experiment, but with a different drug. They report a variance of 1000. And for the treatment, it is less clear, but we may assume we will have similar fluctuations.

$s_A^2 = s_B^2 = 1000$

Why this course?

Stage 6: Gathering variance information



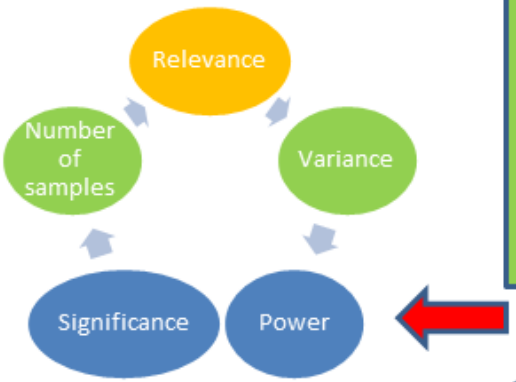
If you are not sure about the treatment variance, you may increase it by some reasonable safety factor (20%, 30%).

20% is fine

$s_A^2 = 1000$
$s_B^2 = 1200$

Why this course?

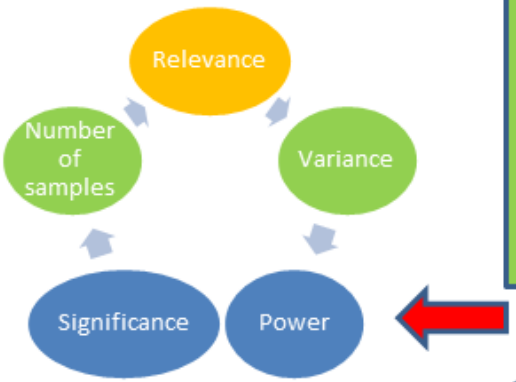
Stage 7: Setting error tolerances



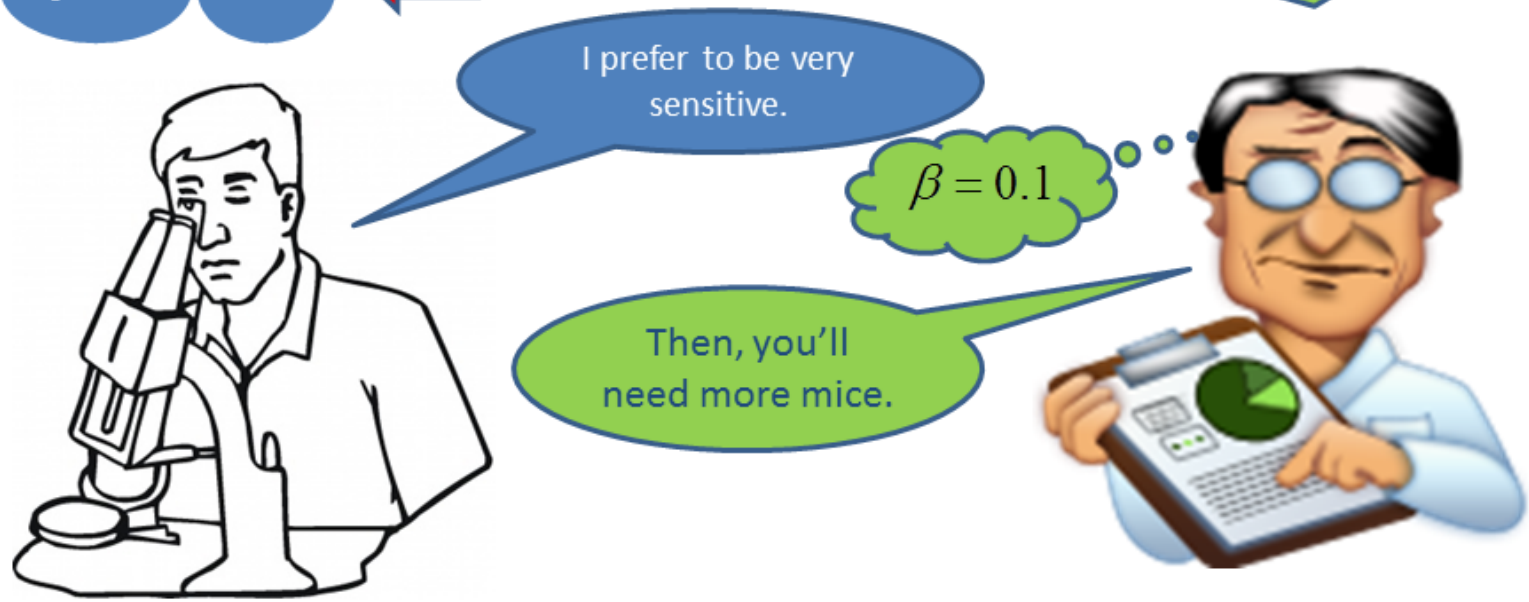
You know, that in some cases, the test will say there is a difference between the treatment and control groups when, actually, there is none (Type I error). How often do you accept to have this kind of errors? If you choose 5%, then your CONFIDENCE LEVEL in the hypothesis testing will be 95%.

Why this course?

Stage 7: Setting error tolerances

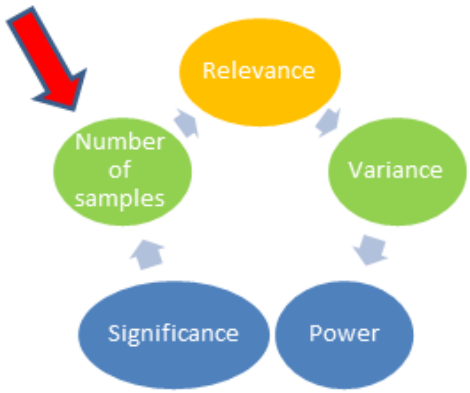


On the other side, there will be ocasions, in which there is actually a difference of 25 mm^2, but just because of chance, you've been unlucky and your mice were not able to show this difference (Type II error). How often do you accept to take this risk? People normally choose between 10% and 20%. The POWER of the test would be 90% or 80% respectively.



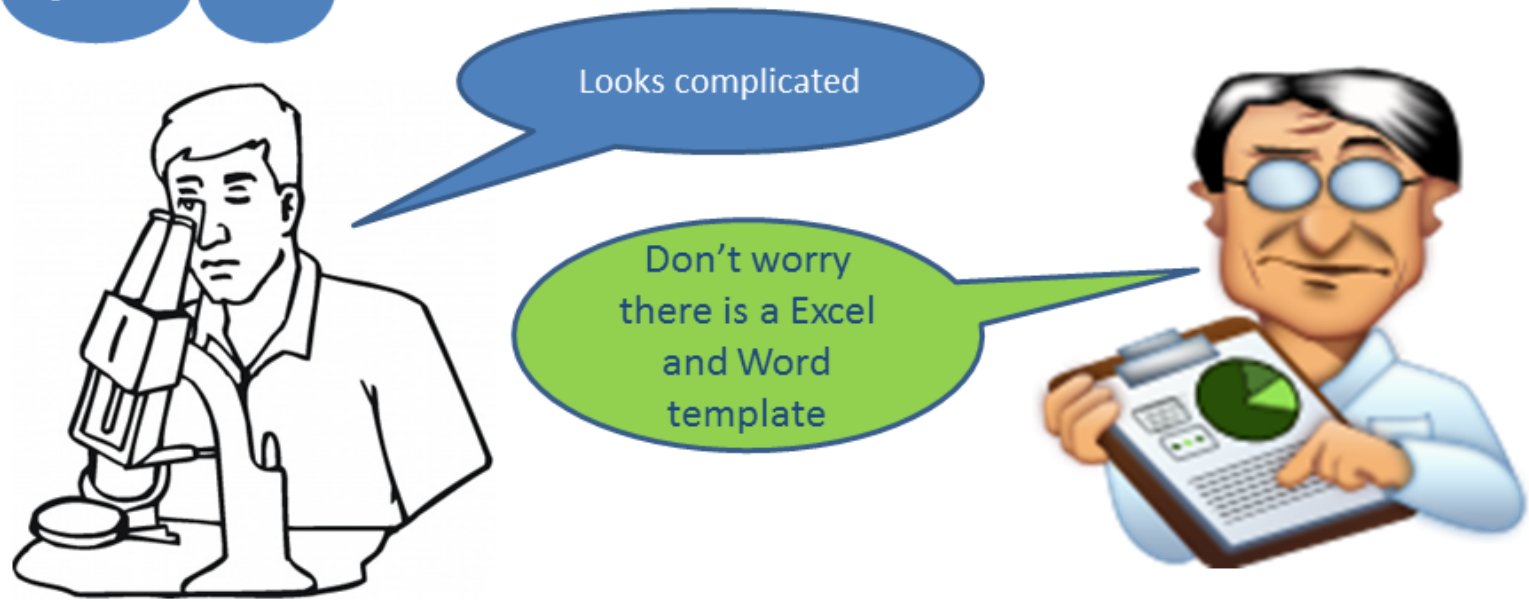
Why this course?

Stage 8: Calculating the number of mice



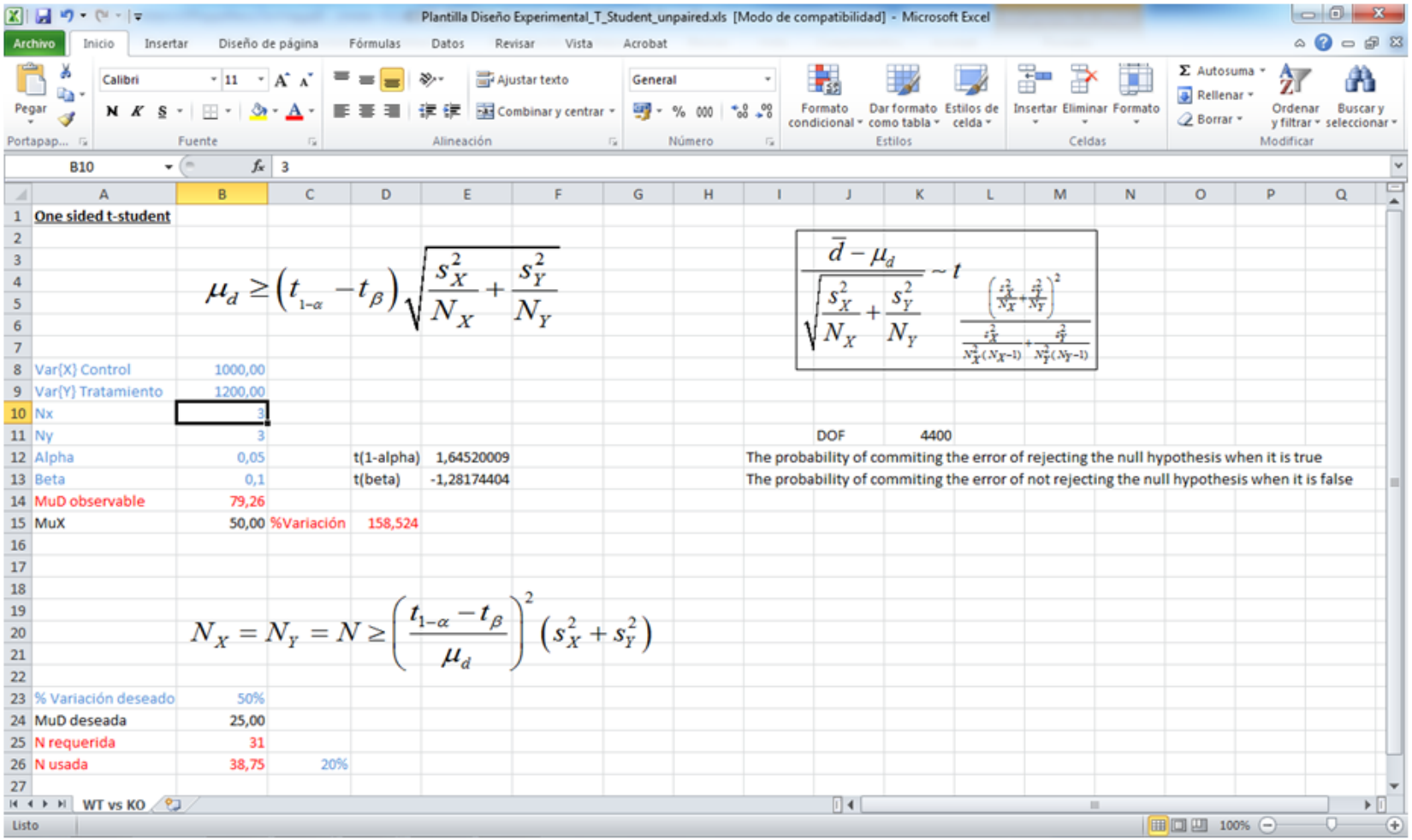
The formula to calculate the number of samples is

$$N_A = N_B = N \geq \left(\frac{t_{1-\alpha} - t_{\beta}}{\Delta} \right)^2 \left(s_A^2 + s_B^2 \right)$$



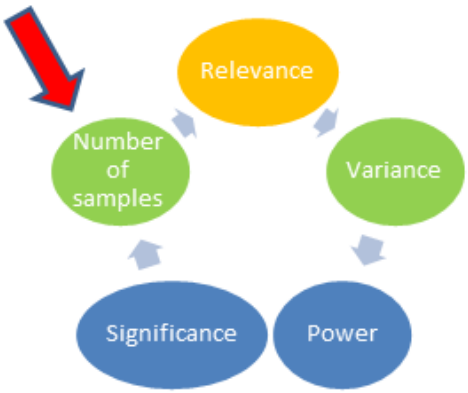
Why this course?

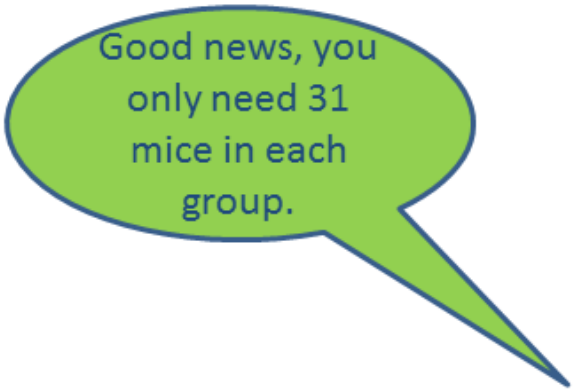
Stage 8: Calculating the number of mice

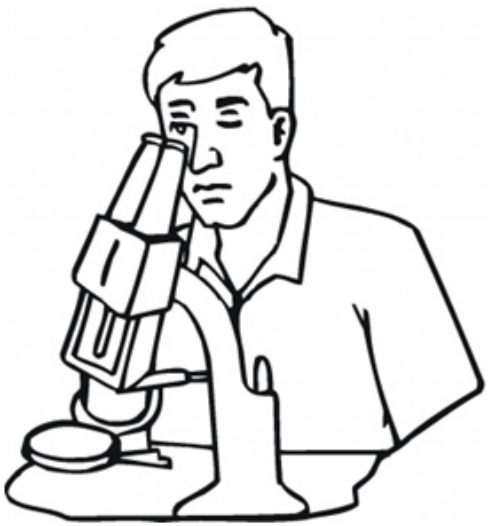


Why this course?

Stage 8: Calculating the number of mice

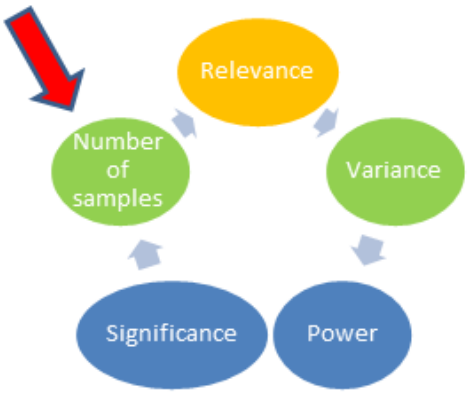






Why this course?

Stage 8: Calculating the number of mice



But sometimes, some mice get a cold, or die from something not related to my experiment, or they are for any reason useless.

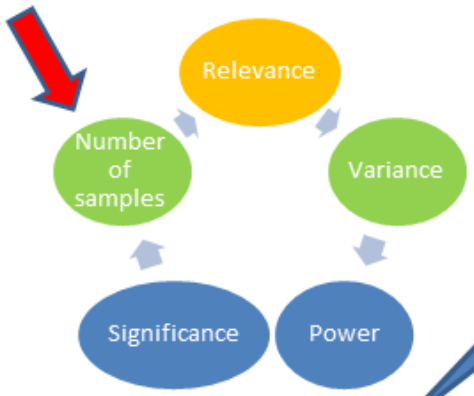
How often does it happen?

About 20% of the times

Then, use 31/0.8=39 mice per group.

Why this course?

Stage 8: Calculating the number of mice



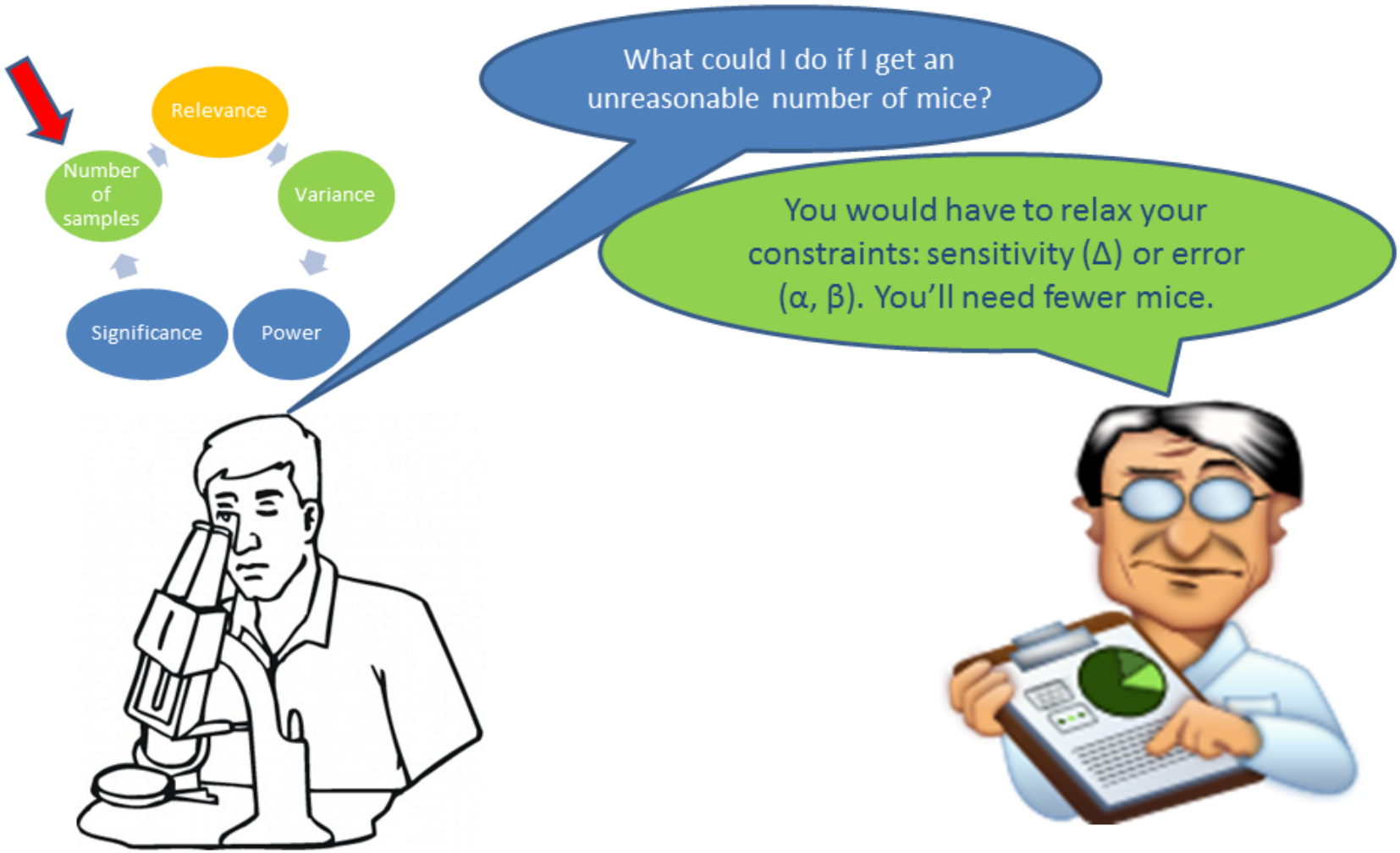
I also want to test several drugs

OK, simply repeat this design exercise for each pair of groups you want to compare.

If you are using a group for multiple (M) comparisons, remind to reduce α (an easy option is α/M instead of α, 0.05) in your design.

Why this course?

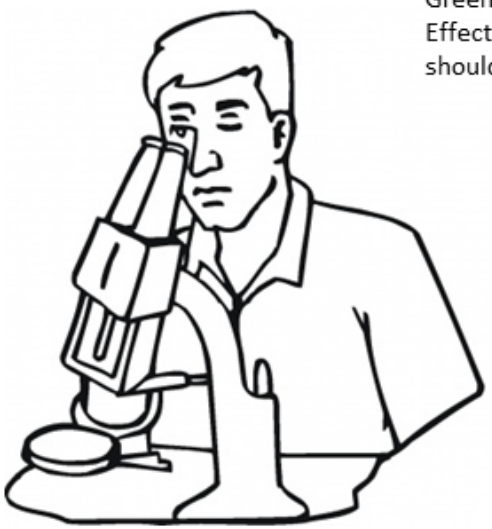
Stage 8: Calculating the number of mice



Why this course?

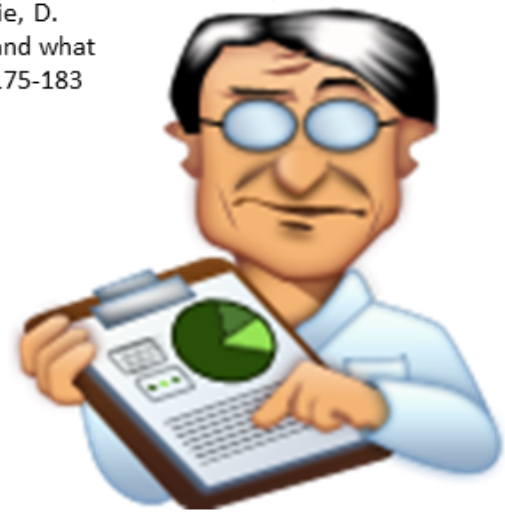
Replicates

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-value=α, there is 50% chances that if you replicate the experiment you cannot reject the null hypothesis. P-value=α is a suggestion of an interesting result, but not a definitive result. A p-value=α/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



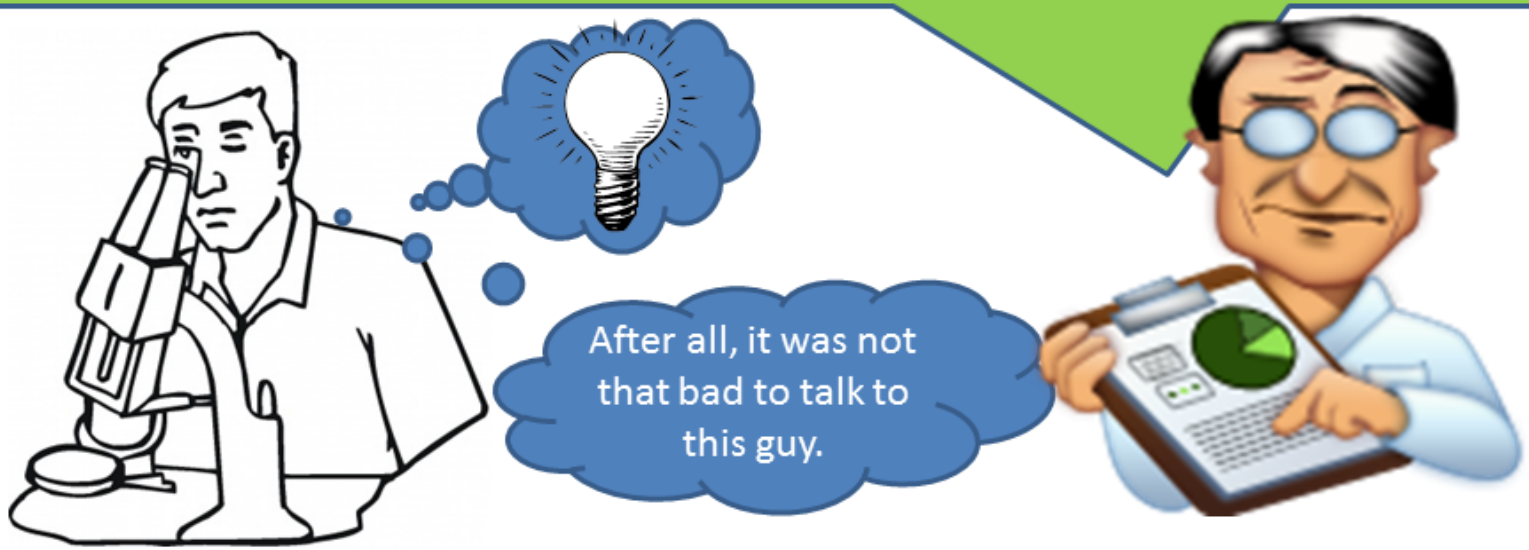
Why this course?

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.



Outline

Basics of statistical inference

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

$H_0: \quad \pi \geq \pi_0$ One-tail test
$H_1: \quad \pi < \pi_0$

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

$H_0: \quad S \leq S_0$ One-tail test
$H_1: \quad S > S_0$

Research hypothesis:
The new machine does not produce tablets with the prescribed concentration

$H_0: \quad c = c_0$ Two-tail test
$H_1: \quad c \neq c_0$

Basics of statistical inference

Research hypoteheses never use “All”, “Some”, “None” or “Not all”.

Research hypothesis:
All hypertense patients benefit from a new drug.
No hypertense patient benefits from a new drug.
Problem: We would have to measure absolutely **ALL** hypertense patients

Research hypothesis:
Not all hypertense patients benefit from a new drug.
Some hypertense patients benefit from a new drug.
Problem: Too imprecise, being true does not provide much information

Basics of statistical inference

Post-mortem analysis:

1. Design hypothesis research
2. Collect data
3. Hypothesis test

Safe analysis:

1. Design hypothesis research
2. Calculate number of samples
3. Collect data
4. Hypothesis test

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).

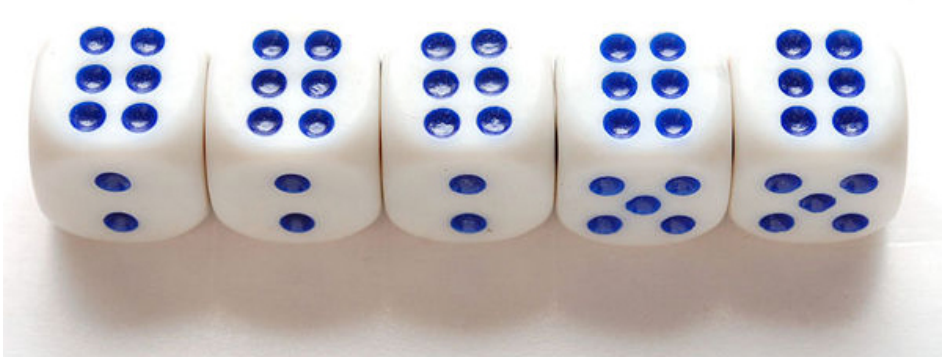
It's like in legal trials:

- 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

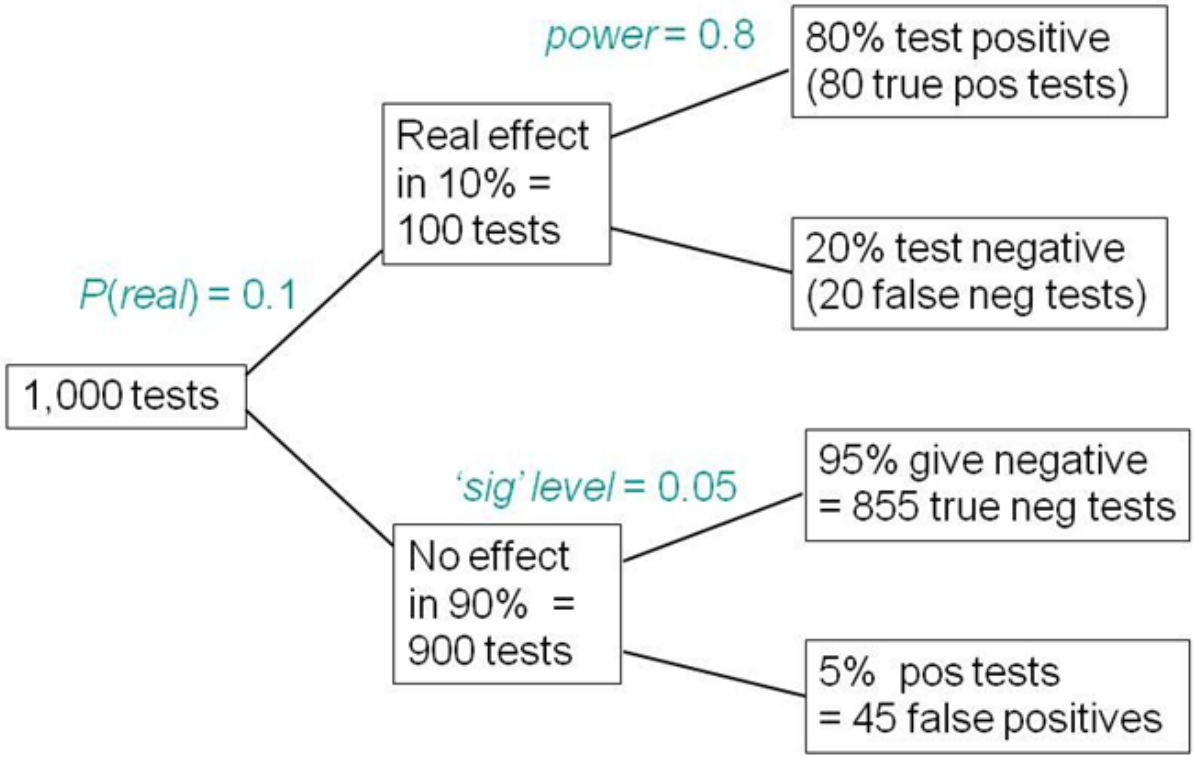
The goal of hypothesis testing is to disprove the null hypothesis! We do this by proving that if the null hypothesis were true, then there would be a very low probability of observing the sample we have actually observed.

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).



Basics of statistical inference: Significance and 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%)

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

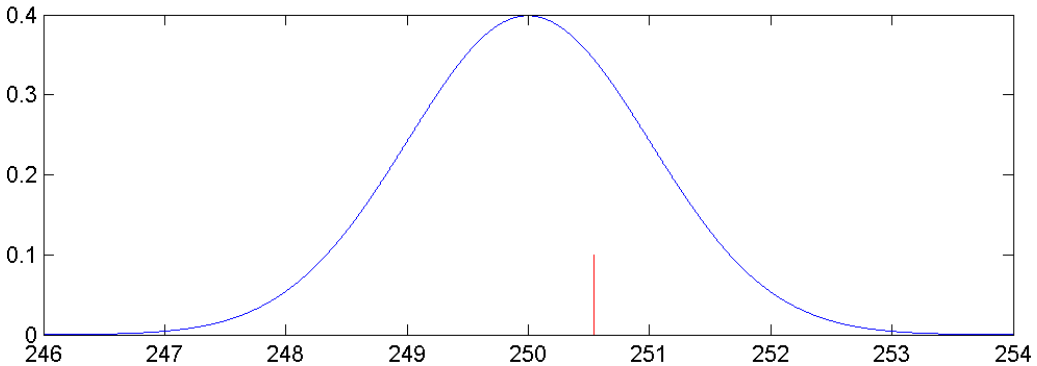
$$H_0 : \mu = 250$$

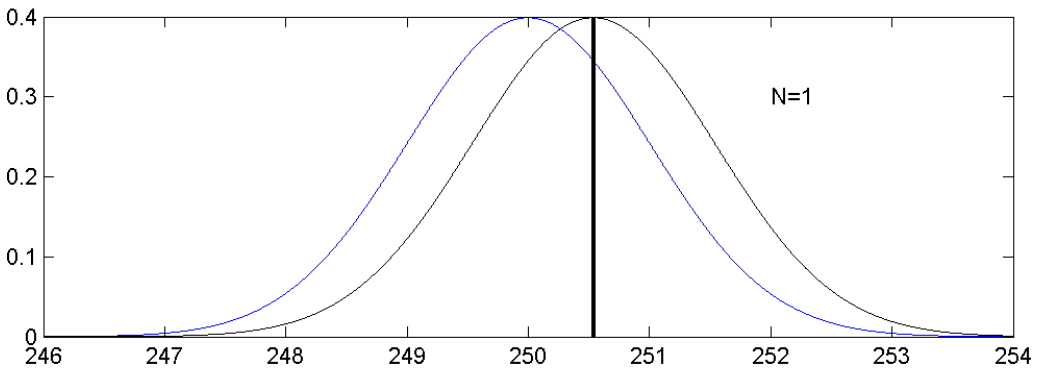
$$H_1 : \mu \neq 250$$

Basics of statistical inference

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

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

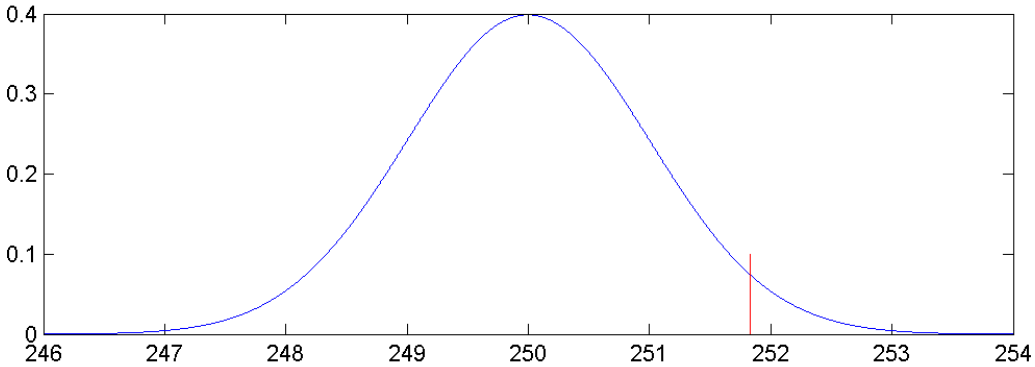


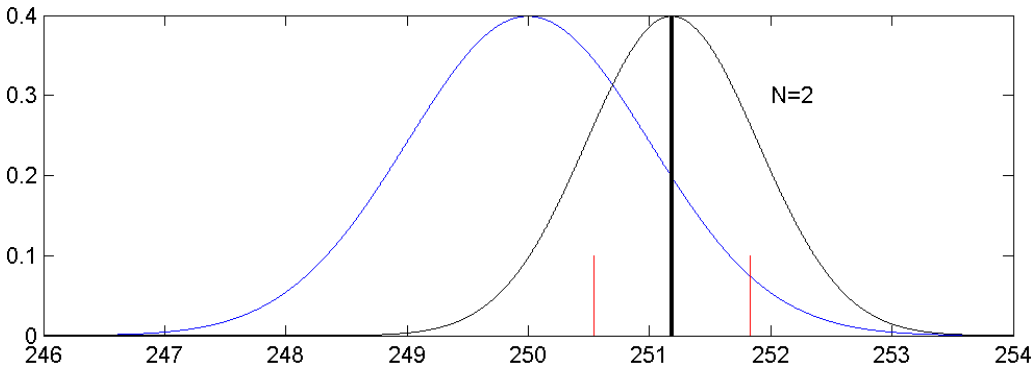


Basics of statistical inference

$$\mathrm{E}\{x_2\} = \mu, \mathrm{Var}\{x_2\} = \sigma^2$$

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

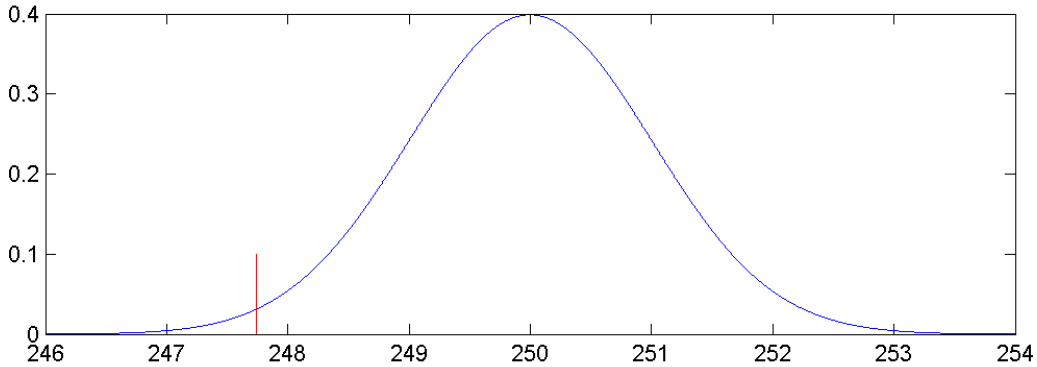


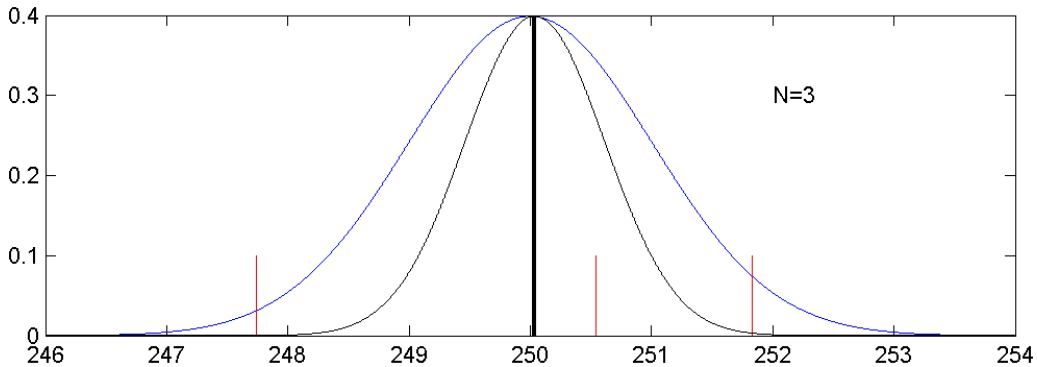


Basics of statistical inference

$$\mathrm{E}\{x_3\} = \mu, \mathrm{Var}\{x_3\} = \sigma^2$$

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

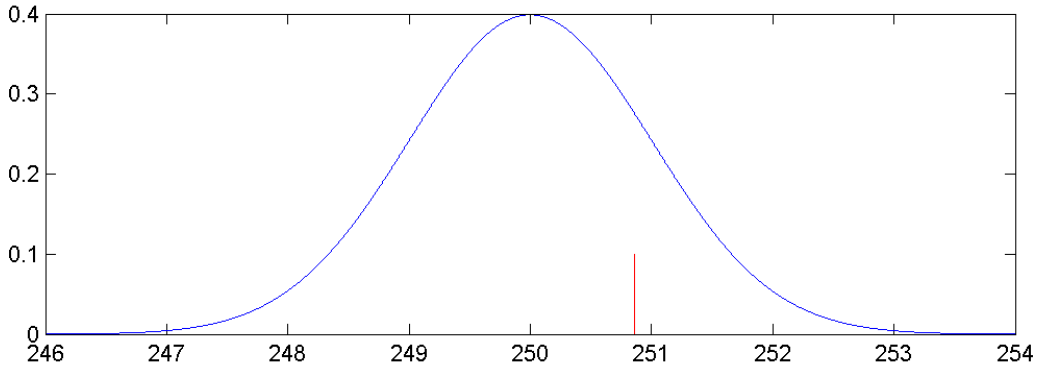


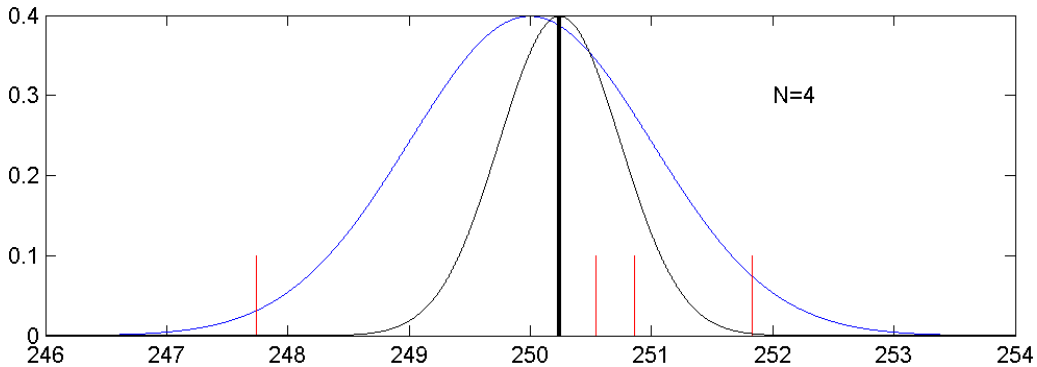


Basics of statistical inference

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

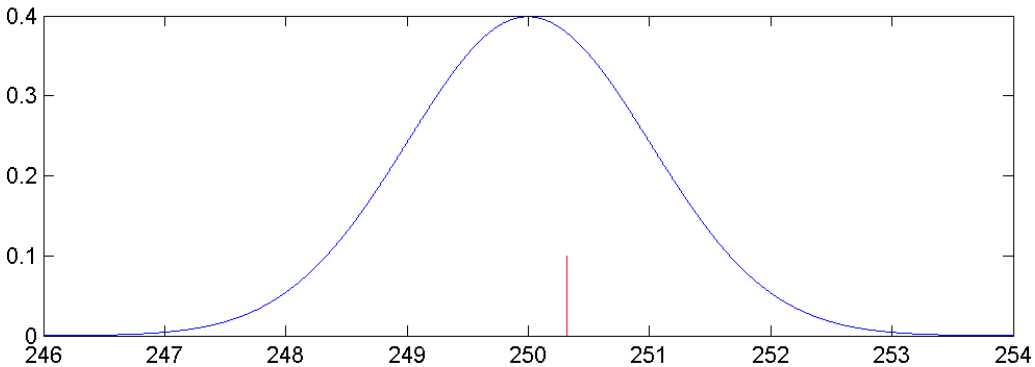
$$\hat{\mu} = \frac{x_1 + x_2 + x_3 + x_4}{4} \Rightarrow \mathrm{E}\{\hat{\mu}\} = \mu, \mathrm{Var}\{\hat{\mu}\} = \frac{\sigma^2}{4}$$

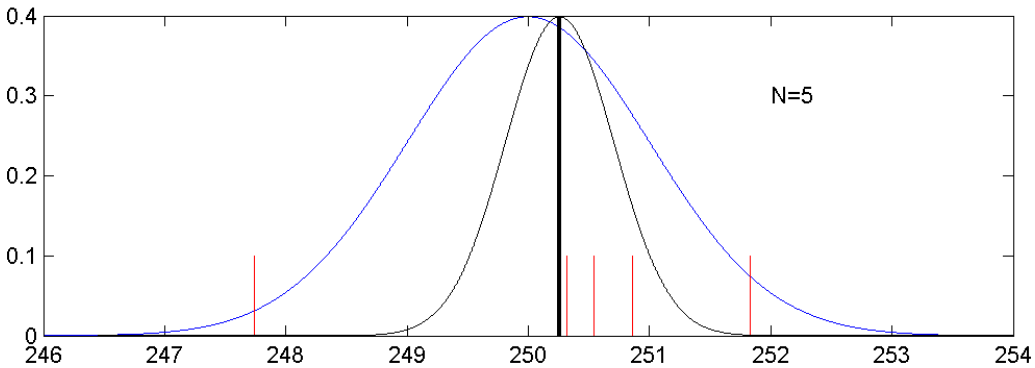




Basics of statistical inference

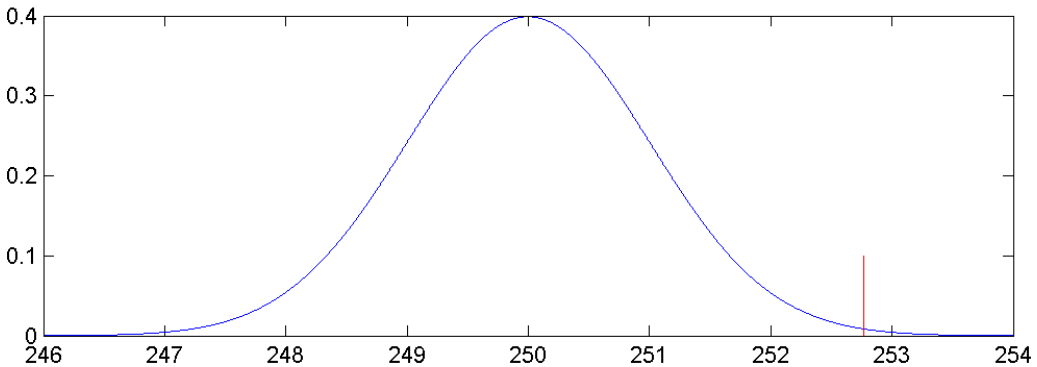
$$\hat{\mu} = \frac{1}{5}\sum_{i=1}^{5} x_i \Rightarrow \mathrm{E}\{\hat{\mu}\} = \mu, \mathrm{Var}\{\hat{\mu}\} = \frac{\sigma^2}{5}$$

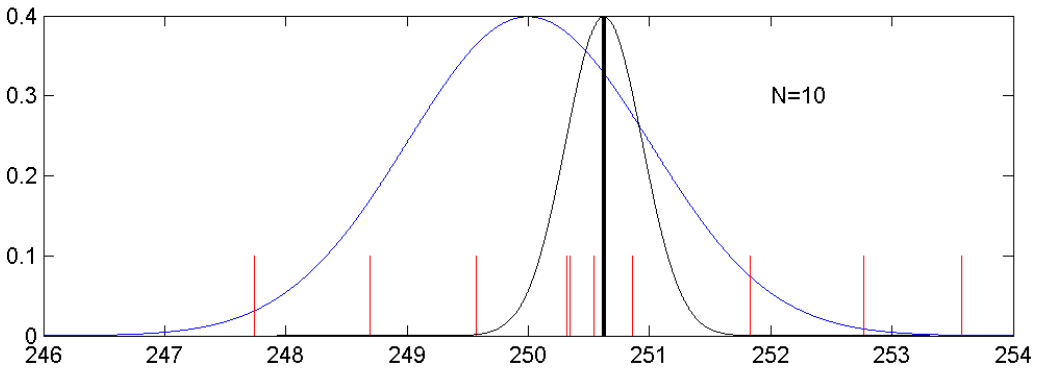




Basics of statistical inference

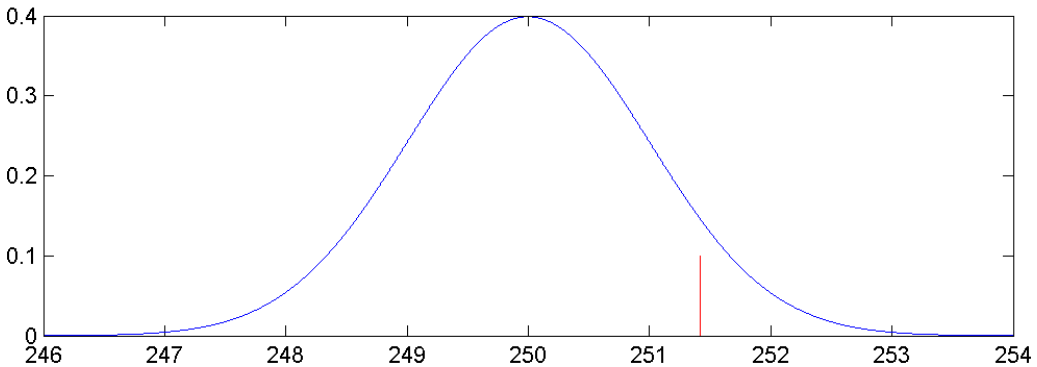
$$\hat{\mu} = \frac{1}{5}\sum_{i=1}^{10} x_i \Rightarrow \mathrm{E}\{\hat{\mu}\} = \mu, \mathrm{Var}\{\hat{\mu}\} = \frac{\sigma^2}{10}$$

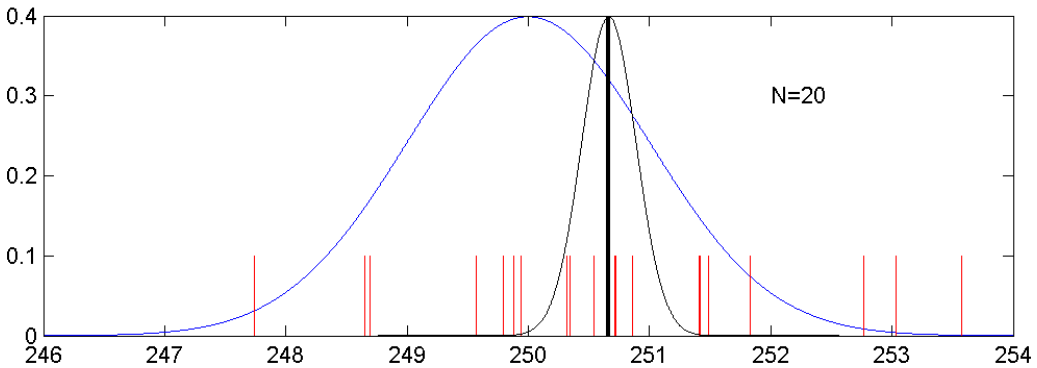




Basics of statistical inference

$$\hat{\mu} = \frac{1}{5}\sum_{i=1}^{20} x_i \Rightarrow \mathrm{E}\{\hat{\mu}\} = \mu, \mathrm{Var}\{\hat{\mu}\} = \frac{\sigma^2}{20}$$





Basics of statistical inference

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

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

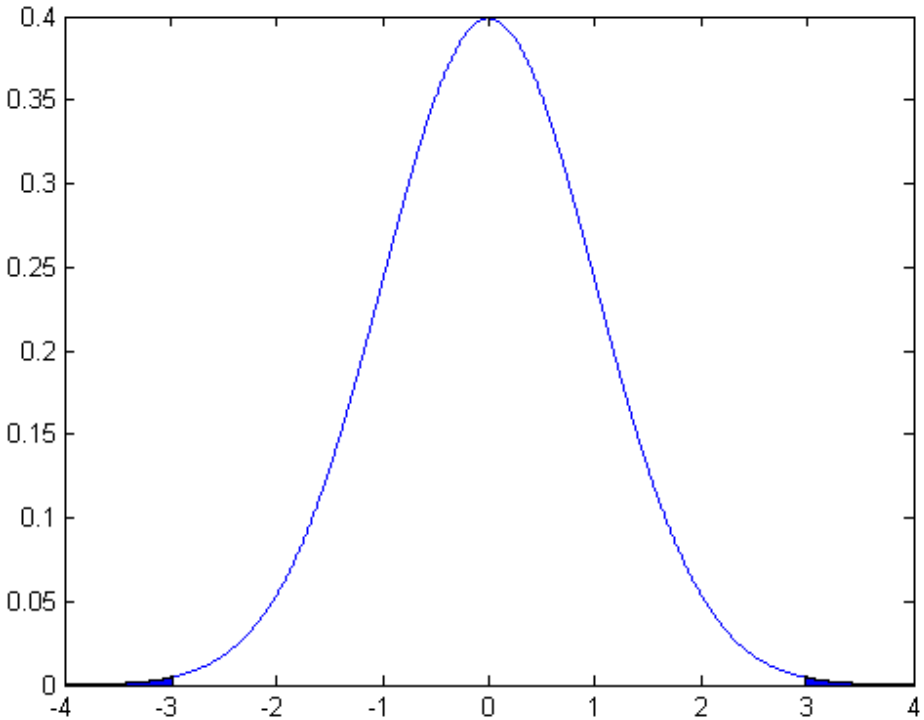
- Step 3: Plug-in the observed data if H_0 is true

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

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.

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



Basics of statistical inference

- Step 5: Reject or not reject H_0

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

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

Types of tests

- Significance tests:

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

- Equivalence tests:

$$H_0 : \mu \neq \mu_0$$
$$H_A : \mu = \mu_0$$

- Superiority tests:

$$H_0 : \mu \leq \mu_0$$
$$H_A : \mu > \mu_0$$

- Non-inferiority tests:

$$H_0 : \mu < \mu_0$$
$$H_A : \mu \geq \mu_0$$

Outline

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

$$0.3\% < 0.5\% \Rightarrow \text{Reject}$$

1. Step 1: Define the minimum meanigful 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

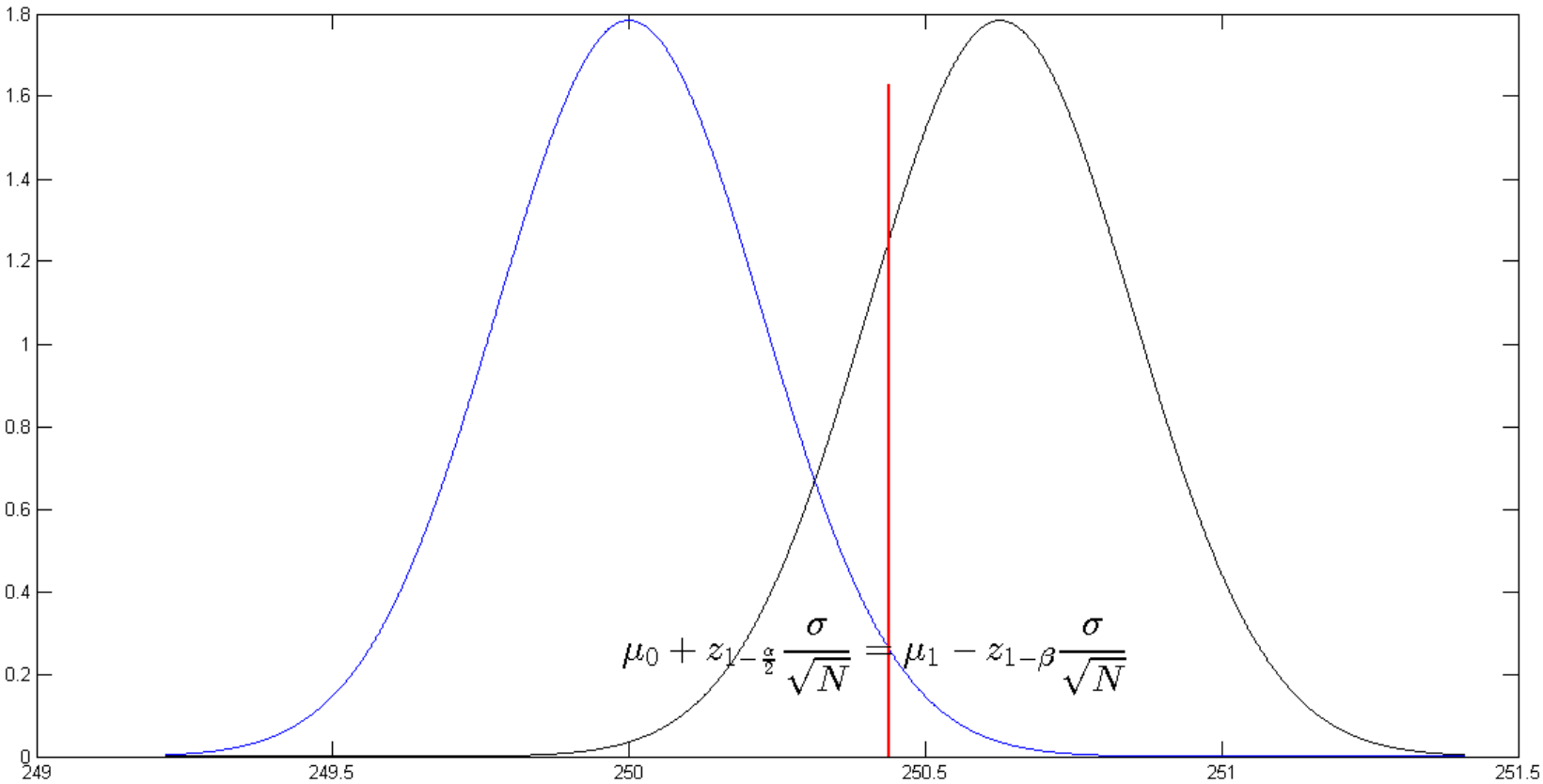
Factors that affect sample size:

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

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 diference Δ.
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.

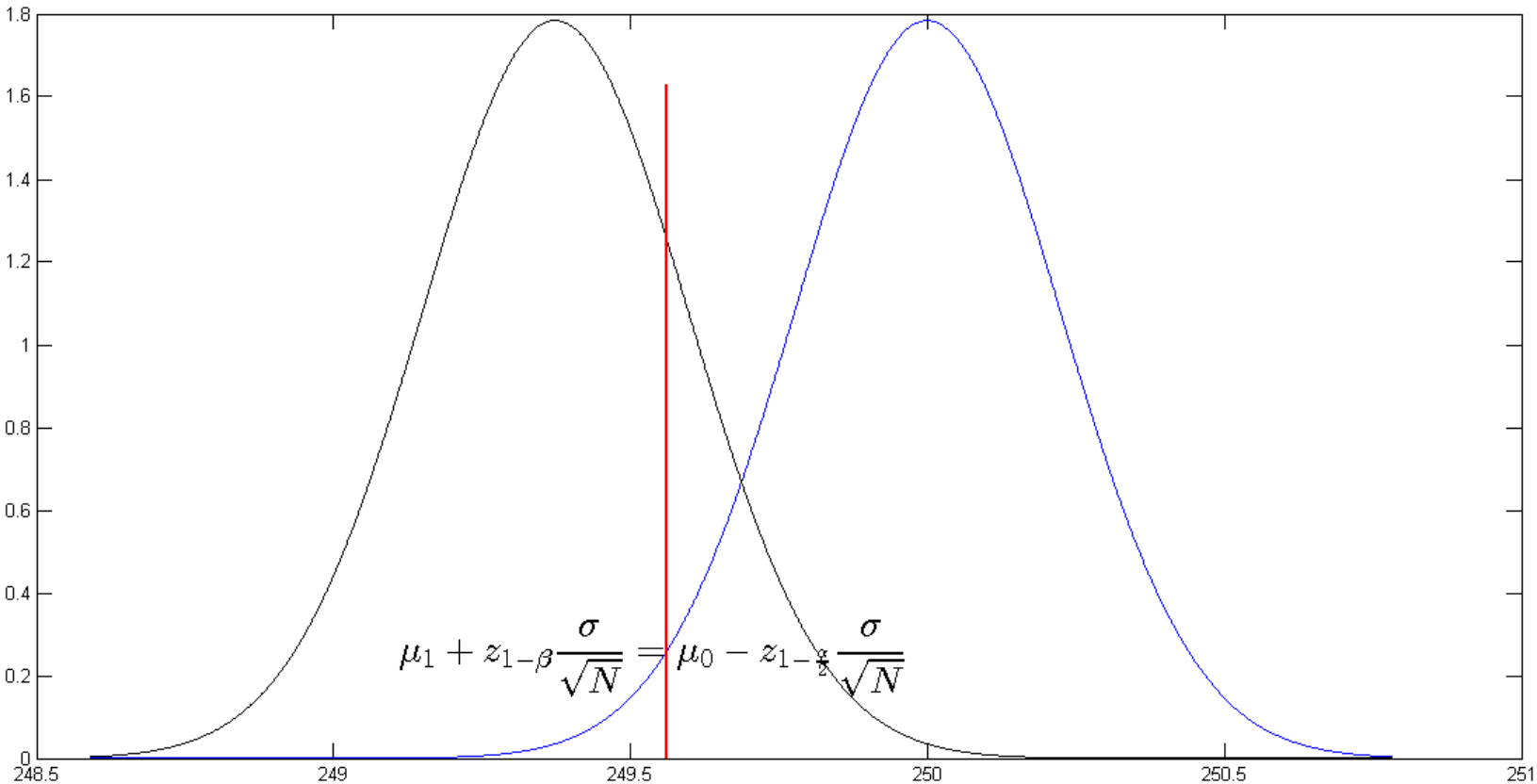
Sample size determination: Test power (right)

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



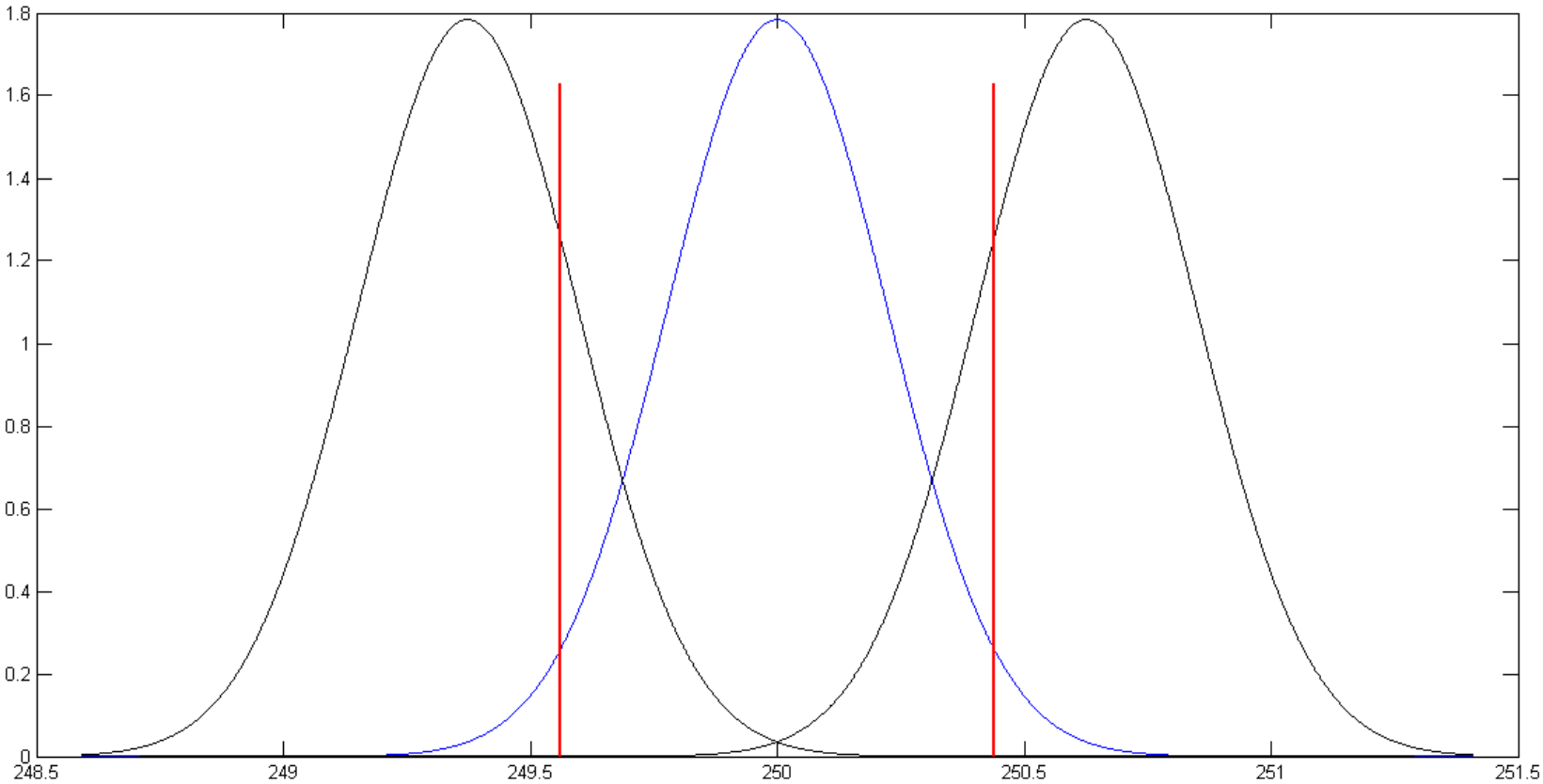
Sample size determination: Test power (left)

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



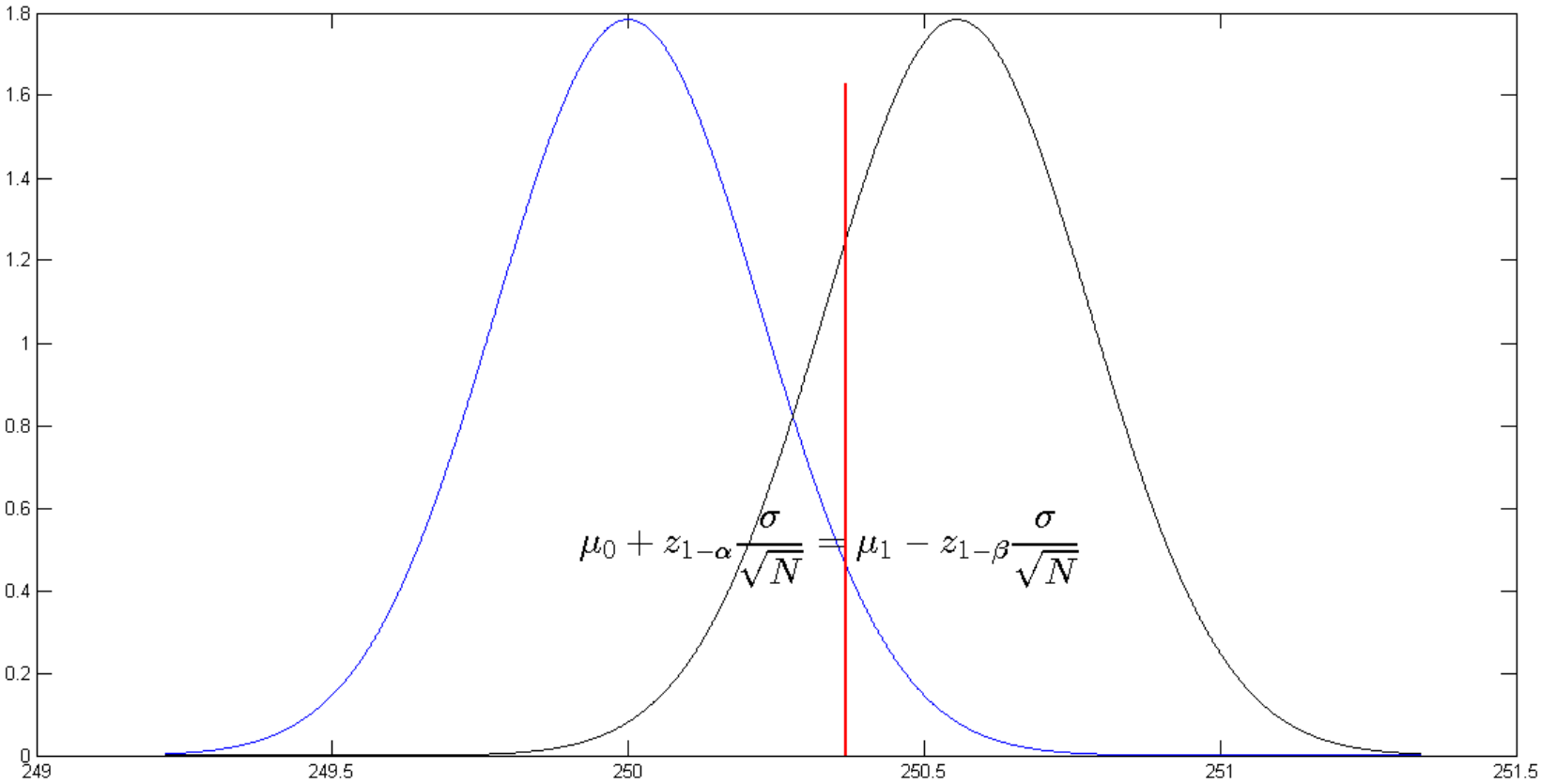
Sample size determination: Test power (two-sided)

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



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:

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

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 diference Δ.
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.

Test power calculation: Confidence level+Sample size

If we fix the sample size, then

$$z_{1-\beta} = \frac{\Delta}{\frac{\sigma}{\sqrt{N}}} - z_{1-\frac{\alpha}{2}} \Rightarrow \boxed{\text{Power} \triangleq \pi = \Pr\{z > -z_{1-\beta}\} = 1-\beta} \quad (4)$$

1. Confidence level: $1-\alpha \uparrow \Rightarrow z_{1-\frac{\alpha}{2}} \uparrow \Rightarrow z_{1-\beta} \downarrow \Rightarrow \pi \downarrow$
   More statistical confidence implies less statistical power.
2. Population variance: $\sigma^2 \uparrow \Rightarrow z_{1-\beta} \downarrow \Rightarrow \pi \downarrow$
   If the population variance increases, the statistical power decreases.
3. Effect size: $\Delta \downarrow \Rightarrow z_{1-\beta} \downarrow \Rightarrow \pi \downarrow$
   If we want to detect more subtle differences, the statistical power decreases.
4. One- or Two-sided test: Two-sided $\Rightarrow z_{1-\frac{\alpha}{2}} > z_{1-\alpha} \Rightarrow z_{1-\beta} \downarrow \Rightarrow \pi \downarrow$
   Two-sided tests have less power than one-sided tests.
5. Sample size: $N \downarrow \Rightarrow z_{1-\beta} \downarrow \Rightarrow \pi \downarrow$
   If we use fewer samples, the statistical power decreases.

Outline

Population variance $\neq$ sample variance

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

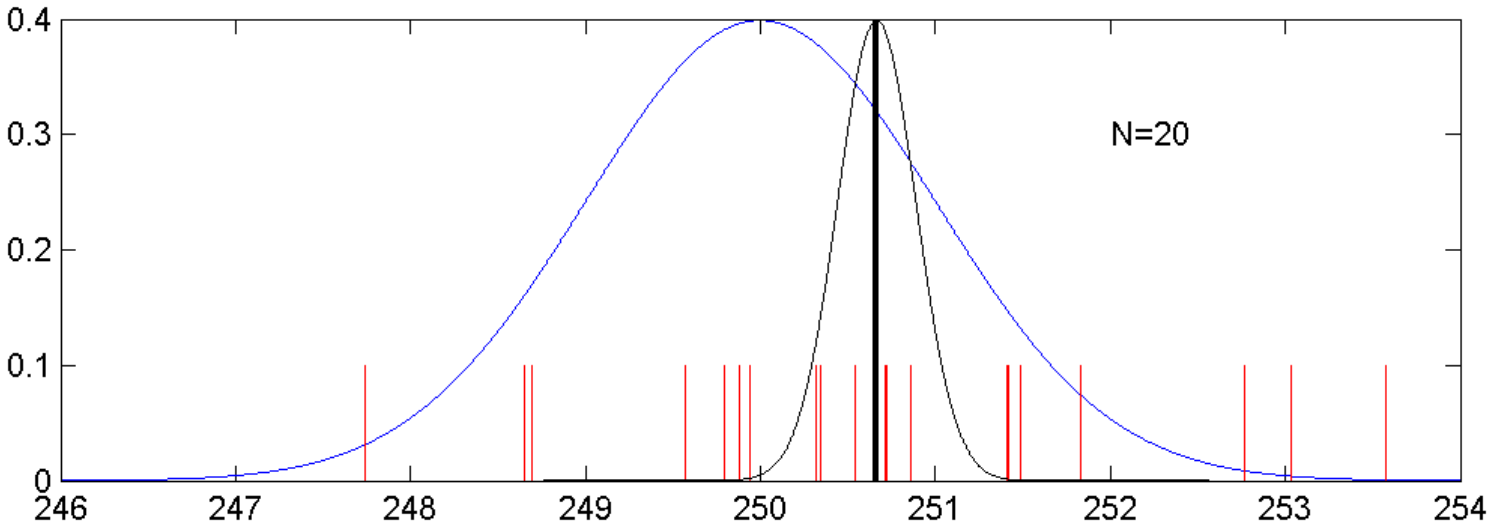
$$\begin{aligned}
\text{Population mean} &\triangleq \mu \\
\text{Sample mean} &\triangleq \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \\
\text{Population variance} &\triangleq \sigma^2 \\
\text{Sample variance} &\triangleq \hat{\sigma}^2 = s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\end{aligned}$$

$$\mu \neq \hat{\mu}$$
$$\sigma^2 \neq s^2$$

Population variance $\neq$ mean variance

$$\begin{aligned} \text{Population variance} &\triangleq \sigma^2 = \mathrm{Var}\{x_i\} \\ \text{Mean variance} &\triangleq \sigma^2_{\hat{\mu}} = \mathrm{Var}\{\hat{\mu}\} = \frac{\sigma^2}{N} \\ \text{Sample mean variance} &\triangleq s^2_{\hat{\mu}} \end{aligned}$$

$$\sigma^2 \neq \sigma^2_{\hat{\mu}} \neq s^2_{\hat{\mu}}$$



Components of the population variance: repeated measures

$$x_{im} = bp_i + \epsilon_{im}$$

$$\sigma^2 = \sigma^2_{BP} + \sigma^2_{\epsilon}$$

If we repeat the measurement process M times and average the results

$$\begin{aligned} x_i &= \frac{1}{M}\sum_{m=1}^{M} x_{im} = \frac{1}{M}\sum_{m=1}^{M}(bp_i + \epsilon_{im}) \\ &= bp_i + \frac{1}{M}\sum_{m=1}^{M}\epsilon_{im} \end{aligned}$$

$$\sigma^2 = \sigma^2_{BP} + \frac{1}{M}\sigma^2_{\epsilon} \tag{5}$$

Finite population effects

Very large population:

- Blood pressure in Spain:

$$\mu, \sigma^2$$

- We study a group of $N = 30$ people:

$$\hat{\mu}, \sigma^2_{\hat{\mu}} = \frac{\sigma^2}{N}$$

Small population:

- Blood pressure in this class:

$$\mu, \sigma^2$$

- We study a group of $N = 30$ people (out of 32!!):

$$\hat{\mu}, \quad \sigma^2_{\hat{\mu}} = \frac{\sigma^2}{N}\left(1 - \frac{N}{N_{population}}\right) \tag{6}$$

This formula is only valid for the mean. In this example,

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{30}\left(1 - \frac{30}{32}\right) = \frac{\sigma^2}{30}\frac{1}{16}$$

Repeated measures $\neq$ Replication

Repeated measures:

$\sigma^2 = \sigma^2_{BP} + \frac{1}{M}\sigma^2_{\epsilon}$

Replication: Allows estimating $\sigma^2_{\hat{\mu}}$

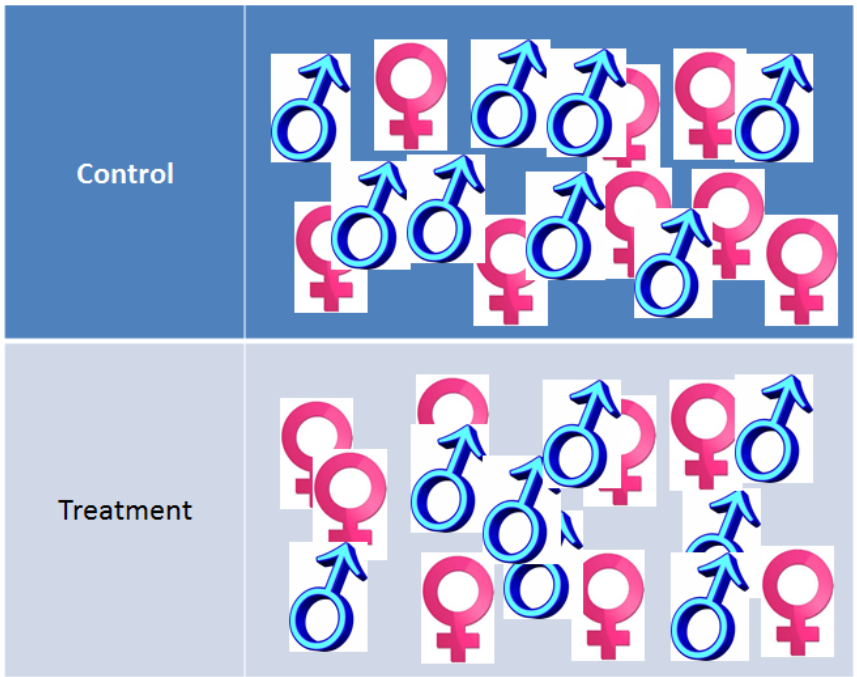
Components of the population variance: blocking variables

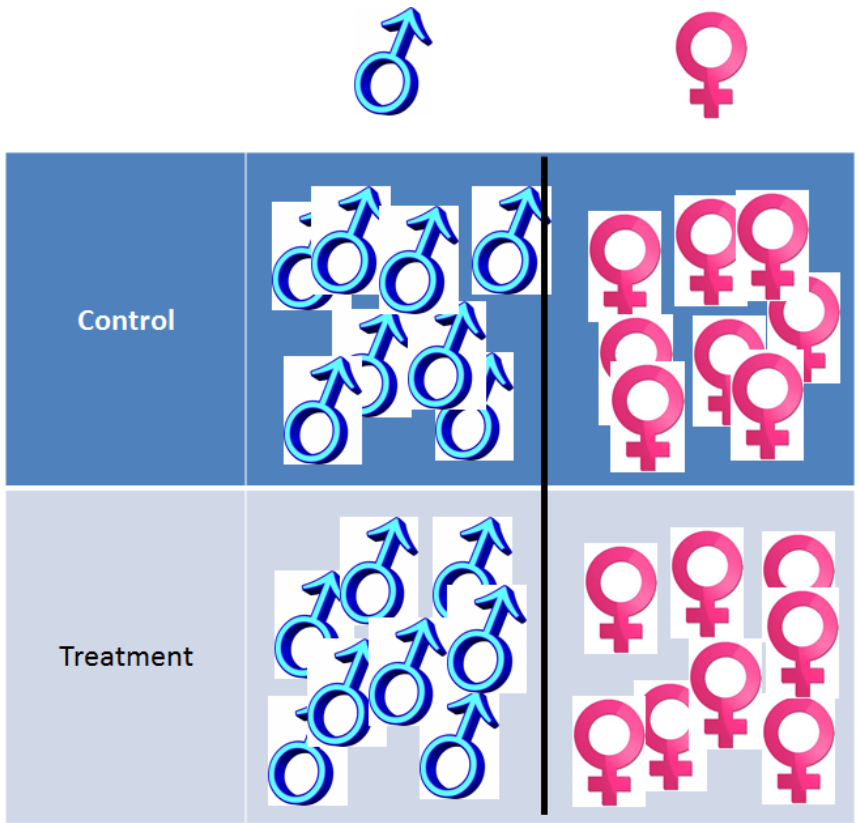
No blocking:

$$\sigma^2 = \sigma^2_{gender} + \sigma^2_{BP} + \sigma^2_{treatment} + \sigma^2_{\epsilon}$$

Blocking:

$$\sigma^2 = \cancel{\sigma^2_{gender}} + \sigma^2_{BP} + \sigma^2_{treatment} + \sigma^2_{\epsilon}$$





Replication $\neq$ Replicates

Replication: Allows estimating $\sigma^2_{\hat{\mu}}$

$$x_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Replicates: Allows estimating $\sigma^2_{\alpha_j}$

$$x_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Outline

Sample size determination fails if ...

1. The estimate of the sample variance is wrong.
2. The distribution of the samples does not follow the hypothesis.
3. Two populations are assumed to have the same variance, when they do not.
4. Two populations are assumed to have the same distribution, when they do not.
5. Variable transformation does not fully solve a distribution problem.
6. Large-sample approximation does not apply.

If the sample size is too large ...

1. Improve the measurement repeatability and reproducibility.
2. Introduce a blocking variable to reduce the precision error.
3. Use a variable with a better precision.
4. Replace a categorical response (yes/no; pass/fail; ...) by an ordinal or continuous variable.
5. Identify covariates that can help to reduce the uncertainty in the model.
6. Reduce the confidence level.
7. Reduce the test power.
8. Repeat measurements on the same subject as a way to reduce measurement variance.
9. Use paired-sample methods instead of two-independent-sample methods.

Non-parametric tests

In many ocasions 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":

$$N_{non-parametric} = \frac{N_{parametric}}{ARE} \tag{7}$$

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.

Bad practices I

1. There is one magic sample size (say, $N = 10$) for all situations.
2. Use $N = 30$ because Student's t distribution is approximately normal for that size.
3. Use $N = \sqrt{N_{total} + 1}$ in a single sampling from a population with N_{total} individuals.
4. Use a table based on Cohen's $d = \frac{\Delta}{s}$. Reason: it assumes normality in the data.
5. Sample size and power calculations are exact. Reason: they are calculated in a conext with high uncertainty (the experiment has not been performed yet).
6. The sample variance is unknown. Reason: look for previously published results or perform a pilot study.

[4]	Cohen's d		
Power	0.2	0.5	0.8
0.25	84	14	6
0.50	193	32	13
0.60	246	40	16
0.70	310	50	20
0.80	393	64	26
0.90	526	85	34
0.95	651	105	42
0.99	920	148	58

Bad practices II

7. Zero acceptance number sampling plans are superior to other samplings. Reason: they are often poorly understood, and they maybe appropriate or not.
8. Postexperiment power is a useful indicator of the value of an experiment. Reason: postexperiment power calculation assumes that the parameter estimates are the true value of the parameters. Post-experiment high power is a necessary condition for supporting the goal of the experiment, but it is not a sufficient condition.
9. Special software is required to calculate sample size and test power. Reason: accurate values are better calculated by software, but approximate values can be calculated on paper.
10. We do not need to know how the collected data will be analyzed. Reason: Every sample size calculation is matched to an analysis method and decision criterion.
11. An experiment may be practically significant but not statistically significant. Reason: to be practically significant, the experiment must be first statistically significant. The converse is not true: an experiment may be statistically significant, but not practically significant.

Good practices

1. Calculate the sample size, power and/or effect size before collecting data. Reason: Make sure you will have enough (and not too much) data to meet the goal of the experiment.
2. If necessary, perform a pilot study to estimate variance.
3. Increase the sample size to compensate for anticipated losses at random.
4. Before collecting your data, make sure that the experiment design meet the goals. You may even simulate the data collection and analysis.
5. Use power calculation to design the experiment and confidence intervals to report the results.
6. Use at least two methods (two softwares, or manual and software) to make sure there is no mistake.
7. Write up a summary describing all the steps taken to design the experiment. It will be useful for future designs and to review the outcome of the experiment if it fails.

Outline

Further reading

Introductory tests:

- G. van Belle. Statistical rules of thumb: Chapter 2. Wiley, 2008
- A. Gelman, J. Hill. Data analysis using regression and multilevel/hierarchical models: Chapter 20. Cambridge Univ. Press, 2007
- Fox N., Hunn A., and Mathers N. Sampling and sample size calculation The NIHR RDS for the East Midlands / Yorkshire & the Humber, 2007

Statistical software:

- PASS manual
- G* Power manual
- Stata power and sample size manual
- MLPowSim manual

Further reading

Web calculators:

- http://powerandsamplesize.com
- http://cran.r-project.org/web/views/ClinicalTrials.html

Outline

Summary

1. Sample size calculations are particularized to the way the data will be analyzed.
2. The goal of hypothesis testing is to prove that the null hypothesis is false. Our research hypothesis should be in the alternative hypothesis.
3. There are five variables strongly related: sample size, population variance, confidence level, test power, effect size (relevance)
4. We can measure the variance of many different, but related, variables (population, sample, sample mean, ...)

Summary

I receive quite a few questions that start with something like this:
"I'm not much of a stats person, but I tried [details...] -- am I doing it right?"

Please compare this with:
"I don't know much about heart surgery, but my wife is suffering from ... and I plan to operate ... can you advise me?"

Folks, just because you can plug numbers into a program doesn't change the fact that if you don't know what you're doing, you're almost guaranteed to get meaningless results -- if not dangerously misleading ones. Statistics really is like rocket science; it isn't easy, even to us who have studied it for a long time. Anybody who think it's easy surely lacks a deep enough knowledge to understand why it isn't! If your scientific integrity matters, and statistics is a mystery to you, then you need expert help. Find a statistician in your company or at a nearby university, and talk to her face-to-face if possible. It may well cost money. It's worth it.

Russell V. Lenth
Professor Emeritus
Department of Statistics and Actuarial Science
The University of Iowa

Outline

Chapter 1. Sample size for the mean

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

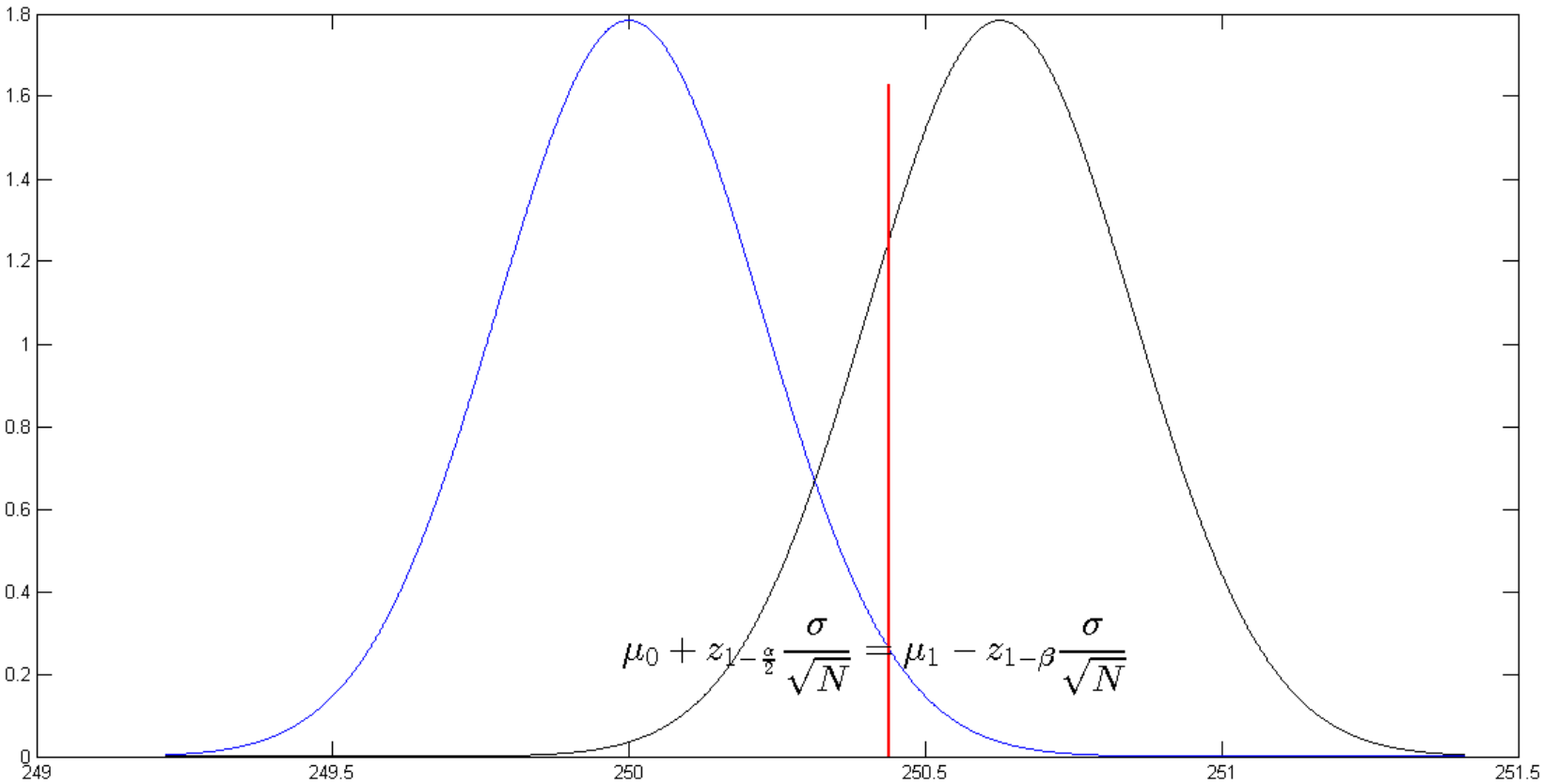
January 28, 2017

Outline

Outline

A single sample with known variance

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



A single sample with known variance

Let us focus on the limit $\hat{\mu}$ value, the one that separates the values within the confidence interval and values outside:

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

$$\hat{\mu} = \mu_1 - z_{1-\beta} \frac{\sigma}{\sqrt{N}} \Rightarrow z_{1-\beta} = \frac{\mu_1-\hat{\mu}}{\frac{\sigma}{\sqrt{N}}}$$

If we add now both equations, we reach the same result as in the previous lecture

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

where $\tilde{\Delta} = \frac{\Delta}{\sigma}$ is the normalized effect size. However, note that the first ratio

$$z = \frac{\hat{\mu} - \mu_0}{\frac{\sigma}{\sqrt{N}}}$$

is distributed as a $N(0,1)$ and that $z_{1-\frac{\alpha}{2}}$ is the z value of this distribution below which there is a probability $1 - \frac{\alpha}{2}$. The same applies to the second ratio.

A single sample with known variance

Example 1

Let us assume that we are manufacturing syrup with 3 mg/mL of a drug. The standard deviation of the manufacturing process is 0.1 mg/mL, and the deviations from the target amount follows a Gaussian distribution. How many samples do we have to screen if we want to detect a deviation from target of $\Delta = 0.03$ mg/mL, with a power of 90% and a confidence level of 95%?

Solution:

$$\text{Power} \triangleq 1-\beta = 0.9 \Rightarrow \beta = 0.1 \Rightarrow z_{1-\beta} = z_{0.9} = 1.2816$$
$$\text{Significance} \triangleq 1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.9600$$
$$\text{Effect size} \triangleq \Delta = 0.03$$
$$\text{Population variance} \triangleq \sigma^2 = 0.1^2 = 0.01$$
$$N > \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta/\sigma}\right)^2 = \left(\frac{1.9600+1.2816}{0.03/0.1}\right)^2 = 116.75 \Rightarrow \boxed{N = 117}$$

A single sample with known variance

Example

Let us assume that the measurement process (determination of the concentration of drug in each sample) has a coefficient of variation of 15%. How does this measurement error increase the variance of the samples?

Solution:

$$CV = \frac{\sigma_\epsilon}{\mu_0} \Rightarrow \sigma_\epsilon^2 = (CV\mu_0)^2 = (0.15 \cdot 3)^2 = 0.2025$$

The variance of the measurements is given by the variance of the manufacturing and the variance of the measurement process

$$\sigma^2 = \sigma^2_{manufacturing} + \sigma^2_\epsilon = 0.01 + 0.2025$$

$$\sigma = \sqrt{\sigma^2} = 0.4610$$

A single sample with known variance

Example 2

What is the sample size now if we want to be as precise as before detecting the malfunctioning of the manufacturing process?

Solution:

$$N > \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta/\sigma} \right)^2 = \left(\frac{1.9600 + 1.2816}{0.03/0.4610} \right)^2 = 2481.2$$

$$\boxed{N = 2482}$$

A single sample with known variance

Example 3

We wonder if we can reduce the cost of the experiment by dividing each syrup sample in 4 aliquotes, and determining the concentration of the sample by averaging the estimation of the concetration in the 4 aliquotes?

Solution:
The variance of the samples now reduces to

$$\sigma^2 = \sigma^2_{manufacturing} + \frac{\sigma^2_\epsilon}{4} = 0.01 + \frac{0.2025}{4} = 0.0606 \Rightarrow \sigma = 0.2462$$

$$N > \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\Delta/\sigma} \right)^2 = \left(\frac{1.9600 + 1.2816}{0.03/0.2462} \right)^2 = 707.67 \Rightarrow \boxed{N = 708}$$

However, the total number of concentration determinations is $4N = 4 \cdot 708 = 2832 > 2482$. That is, it is cheaper to perform independent concentration determinations than 4 concentration determinations from the same sample.

Outline

A single sample with unknown variance

Example 4

1-month old babies awake by night every 3 hours with a standard deviation of 0.5 h (the sleeping period is supposed to be normally distributed). We hypothesize that babies in a orphanage adapt already at this age and sleep longer (the mean shifts at least 1 hour). We do not know the standard deviation of the sleeping time since this may have also changed with respect to the general population, but it cannot be too far from 0.5. We plan to estimate the standard deviation of the sleeping time of babies in an orphanage from the data itself. How many children do I have to examine in order to prove my hypothesis?

Solution:
We cannot apply the calculations above because we do not know the population variance, but make a new theoretical development.

A single sample with unknown variance

We may substitute σ^2 (the true population variance) by s^2 (the sample variance) in the design equations above

$$z_{1-\frac{\alpha}{2}} = \frac{\hat{\mu}-\mu_0}{\frac{\sigma}{\sqrt{N}}} \rightarrow t_{1-\frac{\alpha}{2},0,N-1} = \frac{\hat{\mu}-\mu_0}{\frac{s}{\sqrt{N}}}$$

$$z_{1-\beta} = \frac{\mu_1-\hat{\mu}}{\frac{\sigma}{\sqrt{N}}} \rightarrow t_{1-\beta,\frac{\Delta}{\frac{s}{\sqrt{N}}},N-1} = \frac{\mu_1-\hat{\mu}}{\frac{s}{\sqrt{N}}}$$

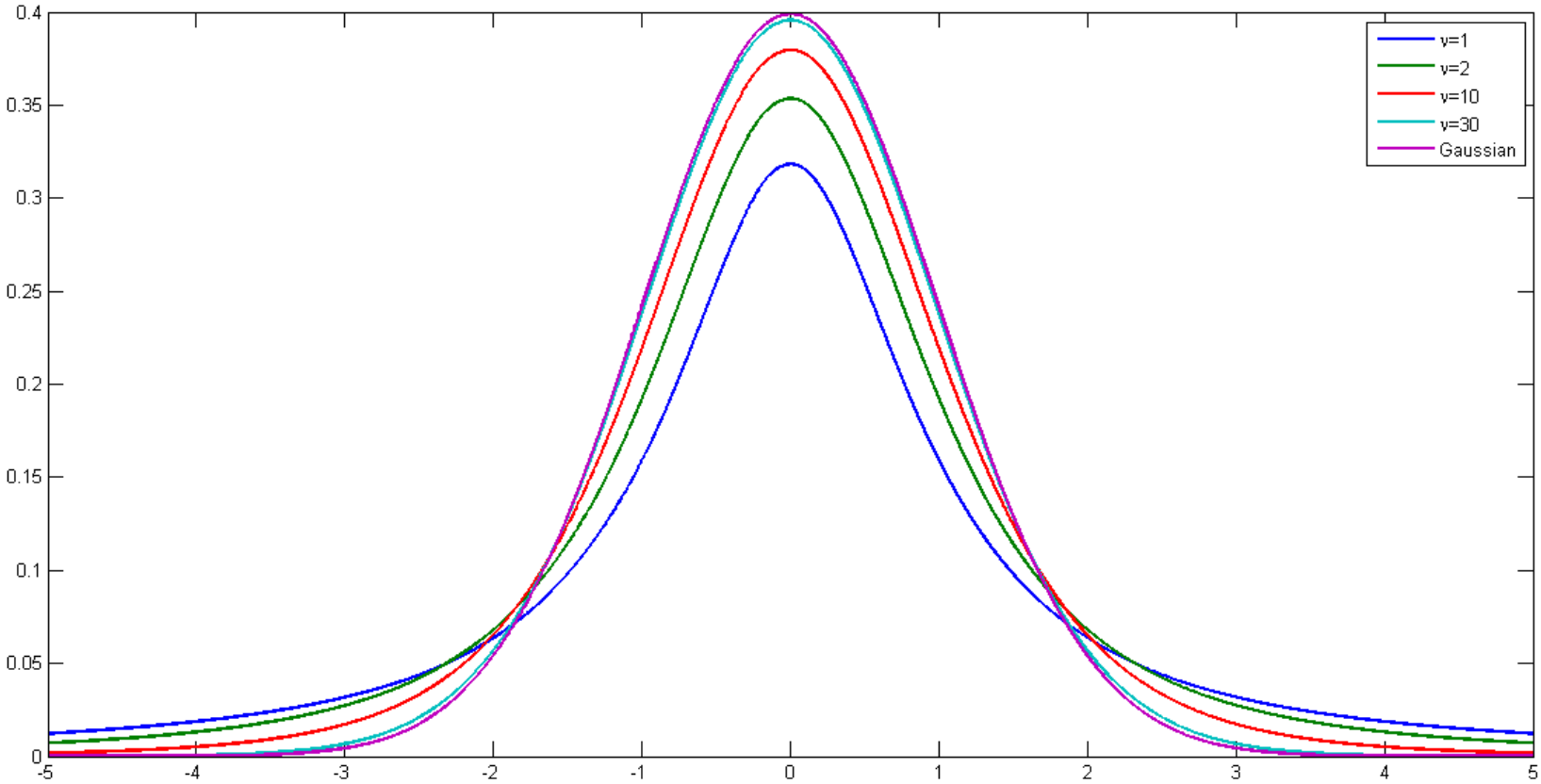
If we add now both equations

$$t_{1-\frac{\alpha}{2},0,N-1}+t_{1-\beta,\frac{\Delta}{\frac{s}{\sqrt{N}}},N-1} = \frac{\mu_1-\mu_0}{\frac{s}{\sqrt{N}}} \Rightarrow \boxed{N = \left(\frac{t_{1-\frac{\alpha}{2},0,N-1} + t_{1-\beta,\frac{\Delta}{\frac{s}{\sqrt{N}}},N-1}}{\tilde{\Delta}}\right)^2} \quad (2)$$

where the normalized effect size is $\tilde{\Delta} = \frac{\Delta}{s}$.

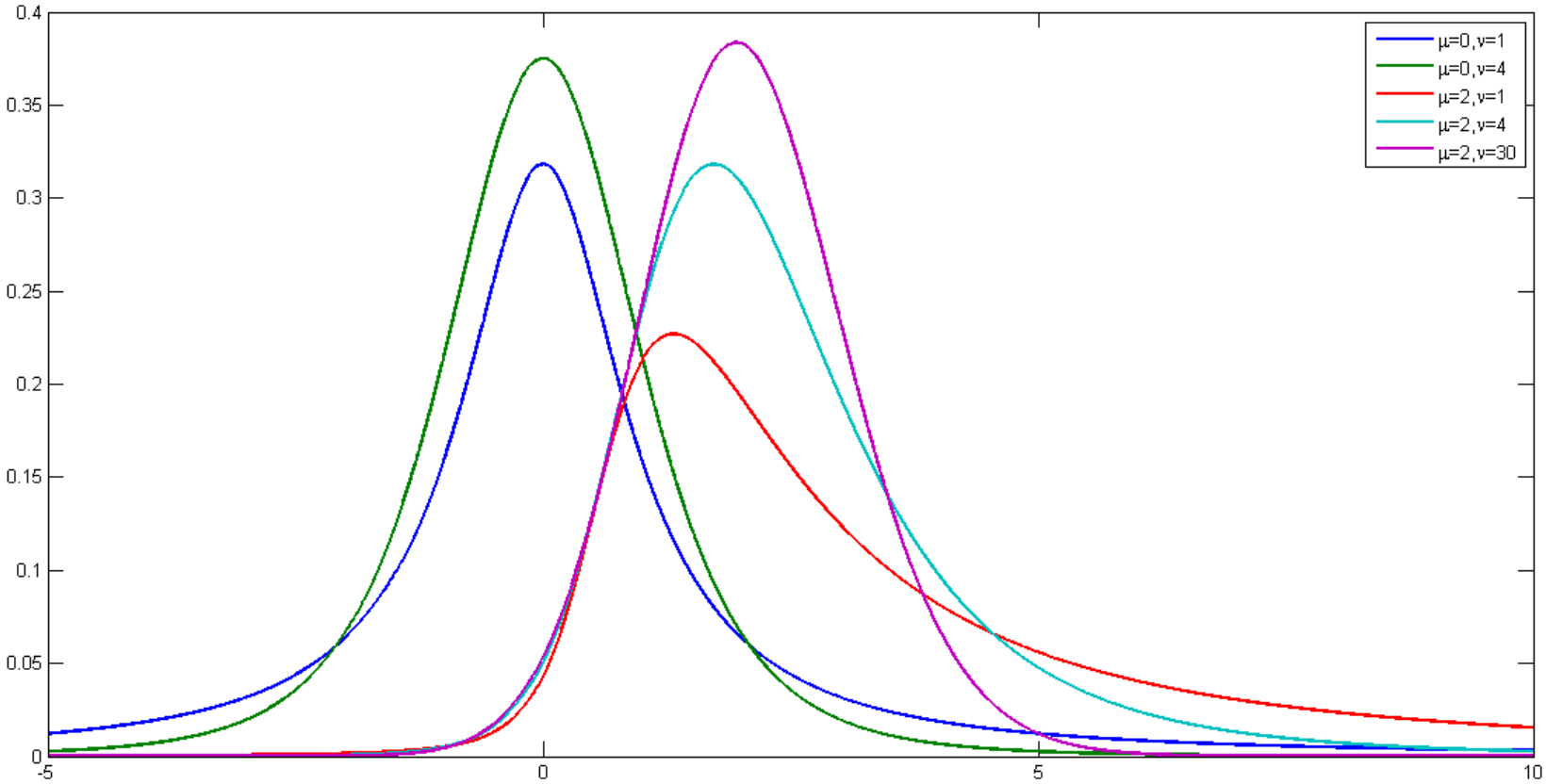
Central Student's t distributions

t_ν : ν degrees of freedom



Noncentral Student's t distributions

$t_{\mu,\nu}$: ν degrees of freedom, μ non-centrality parameter



Example (continued)

Solution:
Our hypothesis test is one-sided:

$$H_0 : \mu \leq 3$$
$$H_A : \mu > 3$$

We need to solve the equation

$$N = \left(\frac{t_{1-\alpha,0,N-1} + t_{1-\beta,\frac{\Delta}{\frac{s}{\sqrt{N}}},N-1}}{\tilde{\Delta}} \right)^2$$

We will use $\alpha = 0.05$, $\beta = 0.1$, and $\Delta = 1$. We do not know s yet, we will use $s_{guess} = \sigma = 0.5$ instead. This gives

$$\tilde{\Delta} = \frac{\Delta}{s_{guess}} = \frac{\Delta}{\sigma} = \frac{1}{0.5} = 2$$

A single sample with unknown variance

Example (continued)

Substituting

$$N = \left(\frac{t_{0.95,0,N-1} + t_{0.9,\frac{2}{\sqrt{N}},N-1}}{2} \right)^2$$

This is a non-linear equation with no analytical form. We can solve it iteratively. For the first iteration we will use the Gaussian parameters:

Iter.	$t_{0.95,0,N-1}$	$t_{0.9,\frac{2}{\sqrt{N}},N-1}$	N
1	$\approx z_{0.95} = 1.65$	$\approx z_{0.9} = 1.28$	2.14
2	5.14	9.05	50.36
3	1.68	1.59	2.66
4	3.37	5.39	19.17
...	...	...	...
30	1.98	2.69	$5.63 \rightarrow \boxed{N = 6}$

A single sample with unknown variance

An approximate formula for this case is given by

$$N = \left(\frac{t_{1-\frac{\alpha}{2},0,N-1} + t_{1-\beta,0,N-1}}{\tilde{\Delta}} \right)^2 \tag{3}$$

where the non-centrality parameter is disregarded.

Example (continued)

The iterative procedure for this approximate method converges to 3.99, which would make $\boxed{N = 4}$. As seen in the exact formula (with non-centrality parameter), $N = 4$ would result in an insufficient level of confidence and/or test power.

Test power improvement

Example (continued)

Due to the increase from $N = 5.63$ to $N = 6$, we increase a little bit the test power (if we keep fixed the confidence level in our hypothesis test). To calculate the new test power we need to find $1 - \beta$ such that the following equation is satisfied

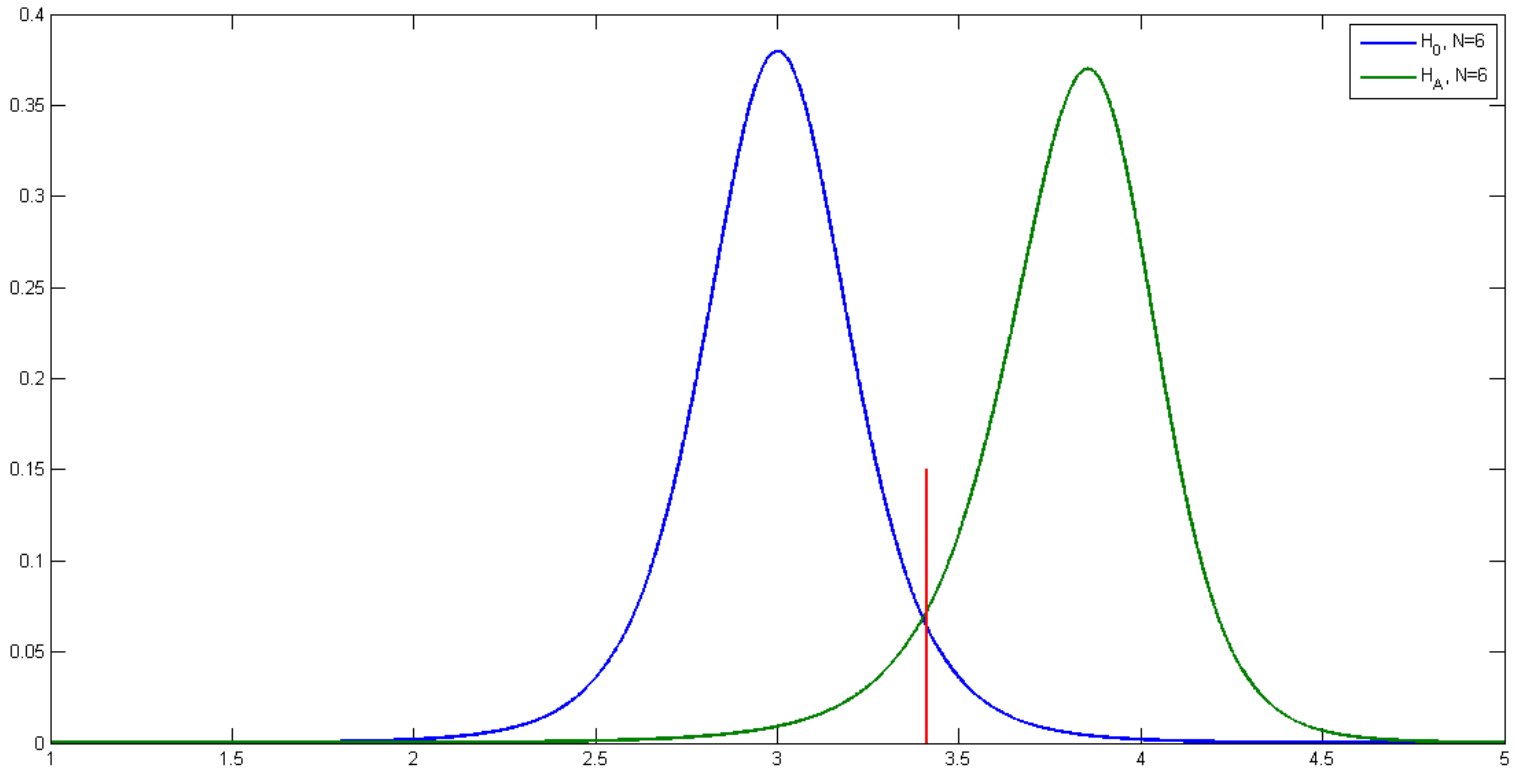
$$N = \left(\frac{t_{1-\alpha,0,N-1} + t_{1-\beta,\frac{\Delta}{\frac{s}{\sqrt{N}}},N-1}}{\tilde{\Delta}} \right)^2$$

$$6 = \left(\frac{t_{0.95,0,5} + t_{1-\beta,\frac{2}{\sqrt{6}},5}}{2} \right)^2$$

We find $\boxed{1 - \beta = 0.9282}$, that is the power has slightly increased from 90% to 92.82%.

Test power improvement

Example (continued)



Outline

Paired samples

Example 5

We are developing a cough mixture and we would like to know its effectiveness. For doing so, we will take N different people with cough. Measure the frequency of coughing (coughs/min) before taking the mixture and 1 hour after taking the mixture. For a severe allergic response, this value is about 5 coughs/min. We would like to detect a reduction to at most 3.5 coughs/min. The standard deviation in the population of people with severe allergic response is 0.4 coughs/min. Let us assume that due to the limited time of observation, we may have a standard deviation due to measurements of 0.1 coughs/min. Confidence level=95%, Power=90%.

Solution:
We will see that with a small transformation of the data, we can use the case of one sample with known or unknown means.

Paired samples

Let us denote as x_{i1} and x_{i2} the two measurements for the i-th individual. Each measurement alone has a variance

$$\text{Var}\{X_1\} = \text{Var}\{X_2\} = \sigma^2_{total} = \sigma^2_{population} + \sigma^2_{measurement}$$

Let us define now the difference between the two measurements

$$\Delta x_i = x_{i1} - x_{i2}$$

If the two measurements are independent from each other, then the variance of the difference is given by

$$\boxed{\text{Var}\{\Delta x\} = 2\sigma^2_{total}} \Rightarrow \sigma_{\Delta x} = \sqrt{2}\sigma_{total} \tag{4}$$

If there is no difference in the treatment, then

$$\mu_{\Delta x} = 0$$

Paired samples

Example (continued)

The hypotheses in our case are

$$H_0 : \Delta_x \leq 0$$
$$H_A : \Delta_x > 0$$

For simplicity, let us assume a design with known variance. The design values are $\Delta = 1.5$, $\sigma^2_{total} = 2 \cdot (0.4^2 + 0.1^2) = 0.34$, $\tilde{\Delta} = \frac{\Delta}{\sigma} = 2.57$, $\alpha = 0.05$ ($z_{1-\alpha} = 1.65$), $\beta = 0.1$ ($z_{1-\beta} = 1.28$). With these values we have

$$N = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\tilde{\Delta}} \right)^2 = 1.30 \rightarrow \boxed{N = 2}$$

Outline

Two-samples with known variance

Example 6

We are interested in knowing if two competitive drugs taken by pregnant women has an effect on the birthweight of their babies. We presume that the standard deviation of the birthweight is $\sigma = 800$ g, and the mean of the whole population $\mu = 3$ kg. We are interested in detecting differences larger than 300 g. Confidence level=95%, Power=90%.

For this experiment we plan to observe a group of N_1 women taking drug 1 and N_2 women taking drug 2. Then, we will calculate the mean of each group and check if their difference is significant or not.

Solution:
We need to derive a new theoretical framework, although it is very similar the one-sample case.

Two-samples with known variance

Measurements in group 1: $x_{1,1}, x_{1,2}, ..., x_{1,N_1} \rightarrow \hat{\mu}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} x_{1,i}$

Measurements in group 2: $x_{2,1}, x_{2,2}, ..., x_{2,N_2} \rightarrow \hat{\mu}_2 = \frac{1}{N_2} \sum_{i=1}^{N_2} x_{1,i}$

$$\begin{aligned} \widehat{\Delta\mu} &= \hat{\mu}_1 - \hat{\mu}_2 \\ \mathrm{Var}\{\widehat{\Delta\mu}\} &= \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} \end{aligned} \tag{5}$$

Optimal sampling:

$$\underset{N_1, N_2}{\mathrm{argmin}} \mathrm{Var}\{\widehat{\Delta\mu}\} \text{ s.t. } N_1 + N_2 = ct \Rightarrow \quad N_2 = N_1 \frac{\sigma_2}{\sigma_1} \tag{6}$$

Hypothesis test:

$$H_0 : \Delta\mu = 0$$
$$H_A : \Delta\mu \neq 0$$

Two-samples with known variance

We can solve this problem as we did for the one-sample case through the use of statistics with known distributions

$$z_{1-\frac{\alpha}{2}} = \frac{\widehat{\Delta\mu}}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}$$
$$z_{1-\beta} = \frac{\Delta-\widehat{\Delta\mu}}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}$$

Adding both equations we have

$$z_{1-\frac{\alpha}{2}} + z_{1-\beta} = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}$$

Using the relationship given by the optimal sampling we reach to

$$N_1 = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\Delta}{\sqrt{\sigma_1(\sigma_1+\sigma_2)}}}\right)^2 \qquad N_2 = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\Delta}{\sqrt{\sigma_2(\sigma_1+\sigma_2)}}}\right)^2 \tag{7}$$

Two-samples with known variance

Example (continued)

We note that $\sigma_1 = \sigma_2$, so $N_1 = N_2 = N$. The data provided yields

$$\alpha = 0.05 \Rightarrow z_{1-\frac{\alpha}{2}} = 1.96$$
$$\beta = 0.1 \Rightarrow z_{1-\beta} = 1.28$$
$$\Delta = 300$$
$$\sigma_1 = \sigma_2 = 800$$
$$\tilde{\Delta} = \frac{\Delta}{\sqrt{\sigma_2(\sigma_1+\sigma_2)}} = 0.2652$$
$$N = \left(\frac{1.96+1.28}{0.2652}\right)^2 = 149.4 \rightarrow \boxed{N = 150}$$

Outline

Two-samples with unknown variance: $\sigma_1 = \sigma_2$

Since the variance of both groups is presumed to be the same, we can estimate the variance from both at the same time as

$$s_{12}^2 = \frac{s_1^2 + s_2^2}{2}$$

The sample variance of the difference sought is

$$s_{\widehat{\Delta\mu}}^2 = \frac{s_{12}^2}{N_1} + \frac{s_{12}^2}{N_2} = s_{12}^2 \frac{2}{N}$$

The statistics defined as

$$t_{1-\frac{\alpha}{2},0,df} = \frac{\widehat{\Delta\mu}}{s_{12}\sqrt{\frac{2}{N}}} \qquad t_{1-\beta,\frac{\Delta}{s_{\widehat{\Delta\mu}}},df} = \frac{\Delta - \widehat{\Delta\mu}}{s_{12}\sqrt{\frac{2}{N}}}$$

follow Student's t distributions (central and noncentral, respectively) with $df = 2(N-1)$ degrees of freedom.

Two-samples with unknown variance: $\sigma_1 = \sigma_2$

Adding both equations and solving for N we get

$$N = 2\left(\frac{t_{1-\frac{\alpha}{2},0,df} + t_{1-\beta,\frac{\Delta}{s_{\widehat{\Delta\mu}}},df}}{\tilde{\Delta}}\right)^2 \tag{8}$$

where the normalized effect size is $\tilde{\Delta} = \frac{\Delta}{s_{12}}$. Remind that for sample size calculation s_{12} is normally unknown (yet) and it is substituted by s_{guess}, the standard deviation of the two populations (assumed to be the same in both).

Two-samples with unknown variance: $\sigma_1 \neq \sigma_2$

The sample variance of the difference sought is

$$s^2_{\widehat{\Delta\mu}} = \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}$$

The statistics defined as

$$t_{1-\frac{\alpha}{2},0,df} = \frac{\widehat{\Delta\mu}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \qquad t_{1-\beta,\frac{\Delta}{s_{\widehat{\Delta\mu}}},df} = \frac{\Delta - \widehat{\Delta\mu}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

follow Student's t distributions (central and noncentral, respectively) with (Welch–Satterthwaite)

$$df = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{1}{N_1-1}\left(\frac{s_1^2}{N_1}\right)^2 + \frac{1}{N_2-1}\left(\frac{s_2^2}{N_2}\right)^2}$$

Two-samples with unknown variance: $\sigma_1 \neq \sigma_2$

We exploit now the optimal allocation for unequal variances, which states that

$$\frac{\sigma_2}{\sigma_1} = \frac{N_2}{N_1}$$

to obtain

$$N_1 = \left(\frac{t_{1-\frac{\alpha}{2},0,df} + t_{1-\beta,\frac{\Delta}{\widehat{s}_{\Delta\mu}},df}}{\tilde{\Delta}_1} \right)^2 \qquad N_2 = \left(\frac{t_{1-\frac{\alpha}{2},0,df} + t_{1-\beta,\frac{\Delta}{\widehat{s}_{\Delta\mu}},df}}{\tilde{\Delta}_2} \right)^2 \tag{9}$$

where $\tilde{\Delta}_1 = \frac{\Delta}{\sqrt{s_1(s_1+s_2)}}$ and $\tilde{\Delta}_2 = \frac{\Delta}{\sqrt{s_2(s_1+s_2)}}$.

Outline

Equivalence test for one mean

Example 7

We are manufacturing tablets with a target amount of drug of 250 mg. We have introduced a new manufacturing process, and we want to show that the new system is equivalent to the old system. How many tablets we need to analyze in order to show that both system perform equally? Presume that the variance of the new manufacturing process has to be estimated from the data itself. We want to detect departures from the mean of at least 2.5mg ($\delta = 2.5mg$), and the standard deviation of the old manufacturing process is 3mg. We want to have a power of 90% if the difference is larger than $\Delta = 5$ mg. Confidence level=95%.

Solution:

This is an equivalence test with hypotheses

$$H_0 : \mu \neq 250$$
$$H_A : \mu = 250$$

Equivalence test for one mean

Actually, the hypotheses tested are

$$H_0 : |\mu - \mu_0| > \delta$$
$$H_A : |\mu - \mu_0| \leq \delta$$

where δ is a deviation from μ_0. This is equivalent to

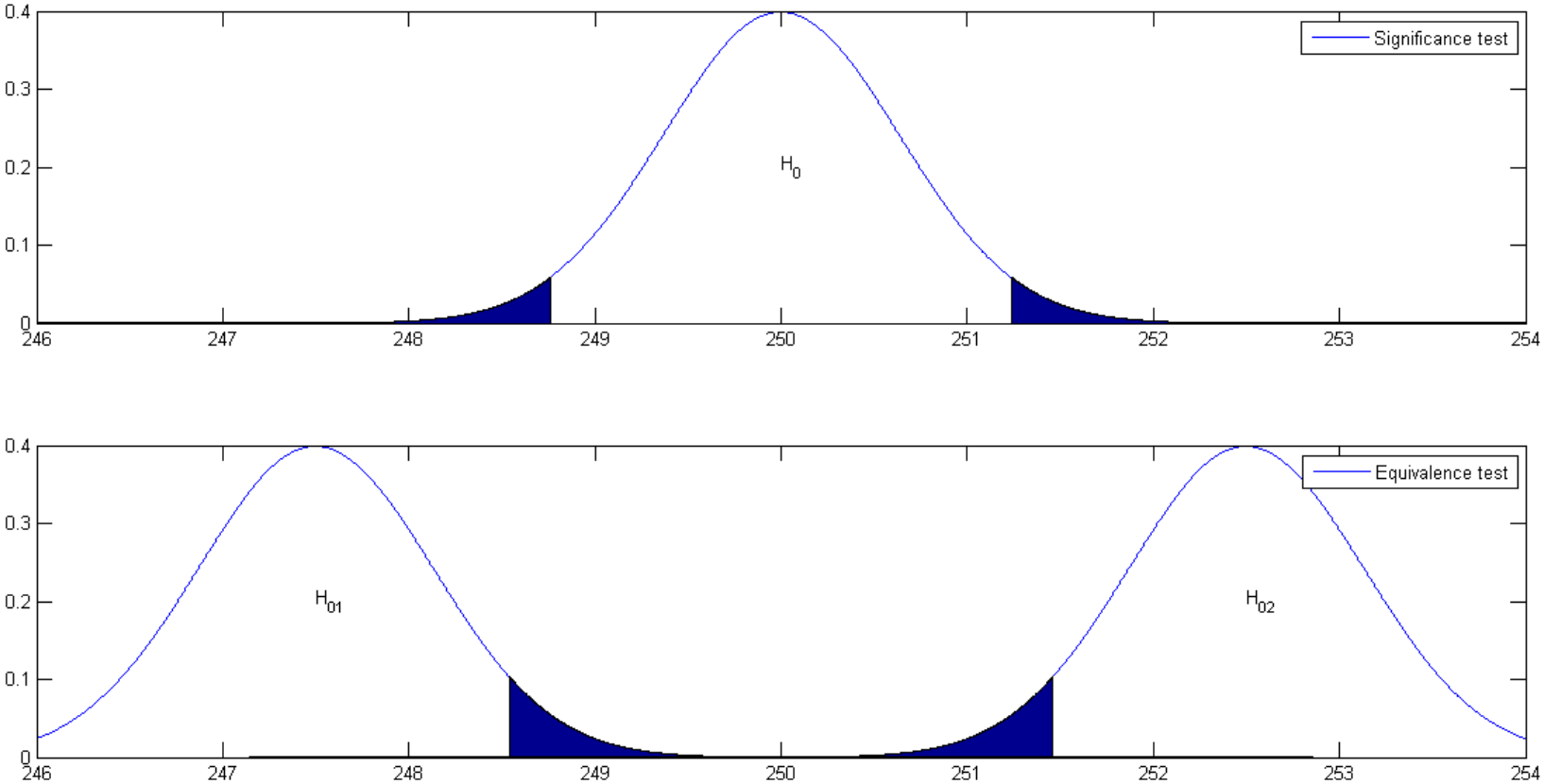
$$\begin{array}{lllll} H_0 : & \mu - \mu_0 < -\delta & \text{or} & \mu - \mu_0 > \delta \\ H_A : & \mu - \mu_0 \geq -\delta & \text{and} & \mu - \mu_0 \leq \delta \end{array}$$

Applying the Two One-Sided Tests (TOST) methodology, we decompose the hypotheses above into two subproblems

$$H_{01} : \mu - \mu_0 < -\delta$$
$$H_{A1} : \mu - \mu_0 \geq -\delta$$

$$H_{02} : \mu - \mu_0 > \delta$$
$$H_{A2} : \mu - \mu_0 \leq \delta$$

Equivalence test for one mean



Equivalence test for one mean

Known variance:

$$N = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\tilde{\Delta}} \right)^2 \tag{10}$$

Unknown variance:

$$N = \left(\frac{t_{1-\alpha,0,N-1} + t_{1-\beta,\frac{\Delta+\delta}{\frac{s}{\sqrt{N}}},N-1}}{\tilde{\Delta}} \right)^2 \tag{11}$$

where $\tilde{\Delta} = \frac{\Delta-\delta}{\sigma}$ is the normalized effect size.

Equivalence test for one mean

Example (continued)

Substituting in our example: $s_{guess} = 3$,
$\tilde{\Delta} = \frac{5-2.5}{3} = 0.83$

$$N = \left(\frac{t_{0.95,0,N-1} + t_{0.9,\frac{0.83}{\sqrt{N}},N-1}}{0.83} \right)^2 \Rightarrow \boxed{N = 16}$$

Outline

Equivalence test for two means

Example 8

We want to show the bioequivalence of two different drugs, i.e., the mean effect of the two drugs are similar. The diastolic blood pressure taking a reference drug is about 96 mmHg, and with an experimental drug, it is presumed to be similar. The population variation is $\sigma = 18$ mmHg. The two drugs are supposed to be similar if their difference is smaller than $\Delta = 19.2 = 20\% \cdot 96$. $\alpha = 0.05$, $\beta = 0.1$. The minimum detectable difference should be $\delta = 10$ mmHg.

Solution:

This is an equivalence test with hypotheses

$$H_0 : \mu_1 - \mu_2 \neq 0$$
$$H_A : \mu_1 - \mu_2 = 0$$

Equivalence test for two means

Sample size formulas are the same as in the case of the significance tests (Eq. 7, Eq. 8, and Eq. 9) but substituting Δ by $|\Delta - \delta|$ and $\frac{\alpha}{2}$ by α.

Example (continued)

$\alpha = 0.05$

$\beta = 0.1$

$s_{guess} = 18$

$\tilde{\Delta} = \frac{|\Delta - \delta|}{s_{guess}} = \frac{|19.2 - 10|}{18} = 0.51$

$s_{\widehat{\Delta\mu}} = s_{guess}\sqrt{\frac{2}{N}}$

$df = 2(N-1)$.

Finally we need to solve for N the equation

$$N = 2\left(\frac{t_{1-\alpha,0,df} + t_{1-\beta,\frac{\Delta}{s_{\widehat{\Delta\mu}}},df}}{\tilde{\Delta}}\right)^2 \Rightarrow \boxed{N = 83}$$

Outline

1-way ANOVA contrasts

Example 9

We are studying the effect of two different drugs on the blood pressure of patients. We have three study groups: placebo (drug 0), drug 1 and drug 2. We wonder which is the sample size for each group if we want to test:

1. There is no difference between placebo and the other two treatments.
2. There is no difference between the two treatments

Let us assume that the variance of the population for each of the treatments is $\sigma = 18$ mmHg, and that we want to be able to detect effect sizes of $\Delta = 10$ mmHg.

Solution:

The two problems above can be addressed through an ANOVA contrast given by

1. $\mu_0 - \frac{\mu_1+\mu_2}{2} = 0$
2. $\mu_1 - \mu_2 = 0$

1-way ANOVA contrasts

The following table shows an example of the expected measurements observed for the balanced (all cells have the same size) experiment.

Placebo	**Drug 1**	**Drug 2**	
$x_{01}, x_{02}, \ldots, x_{0N}$	$x_{11}, x_{12}, \ldots, x_{1N}$	$x_{21}, x_{22}, \ldots, x_{2N}$	
$\hat{\mu}_{0\cdot} = \frac{1}{N}\sum_{j=1}^{N} x_{0j}$	$\hat{\mu}_{1\cdot} = \frac{1}{N}\sum_{j=1}^{N} x_{1j}$	$\hat{\mu}_{2\cdot} = \frac{1}{N}\sum_{j=1}^{N} x_{2j}$	$\hat{\mu}_{\cdot\cdot} = \frac{1}{TN}\sum_{i=1}^{T}\sum_{j=1}^{N} x_{ij}$

Each observation is modelled as

$$x_{ij} = \hat{\mu}_{\cdot\cdot} + \alpha_i + \epsilon_{ij}$$

The α's are the effect of each of the treatments and it is calculated as

$$\alpha_i = \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}$$

Note that the sum of α_i's is 0 and that

$$\hat{\mu}_{i\cdot} = \hat{\mu}_{\cdot\cdot} + \alpha_i$$

1-way ANOVA contrasts

The ANOVA table accounts for the contributions of the different sources of variation to the total variation. The total variation is measured by the variance of all observations. This variance is decomposed as the sum of different sources

$$SS_{tot} = SS_{\alpha} + SS_{\epsilon}$$

$$\sum_{i=1}^{T}\sum_{j=1}^{N}(x_{ij} - \hat{\mu}_{..})^2 = \sum_{i=1}^{T}\sum_{j=1}^{N}(\hat{\mu}_{i.} - \hat{\mu}_{..})^2 + \sum_{i=1}^{T}\sum_{j=1}^{N}(x_{ij} - \hat{\mu}_{i.})^2$$

Source	SS	MS	F	df
Total	SS_{tot}			$df_{tot} = NT - 1$
ϵ	SS_{ϵ}	$MS_{\epsilon} = \frac{SS_{\epsilon}}{df_{\epsilon}}$		$df_{\epsilon} = T(N-1)$
α	SS_{α}	$MS_{\alpha} = \frac{SS_{\alpha}}{df_{\alpha}}$	$F = \frac{MS_{\alpha}}{MS_{\epsilon}}$	$df_{\alpha} = T - 1$

Example

$45 = 40 + 5 \Rightarrow$ Treatments are significant

$45 = 5 + 40 \Rightarrow$ Treatments are not significant

1-way ANOVA contrasts

The following formulas show the true underlying contrast and how it can be estimated

$$\mu_c = \sum_{i=1}^{T} c_i \mu_i \qquad\qquad \hat{\mu}_c = \sum_{i=1}^{T} c_i \hat{\mu}_i.$$

The variance of the estimate is given by

$$\mathrm{Var}\{\hat{\mu}_c\} = \sum_{i=1}^{T} c_i^2 \frac{\sigma_\epsilon^2}{N_i}$$

If all cells have the same number of observations $N_i = N$

$$\mathrm{Var}\{\hat{\mu}_c\} = \frac{\sigma_\epsilon^2}{N} \sum_{i=1}^{T} c_i^2$$

1-way ANOVA contrasts

We may design the sample size through a one-sample mean design with the hypotheses:

$$H_0 : \mu_c = 0$$
$$H_A : \mu_c \neq 0$$

This is equivalent to the derivation of Eqs. 1 and 2 to give

$$N = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\tilde{\Delta}} \right)^2 \tag{12}$$

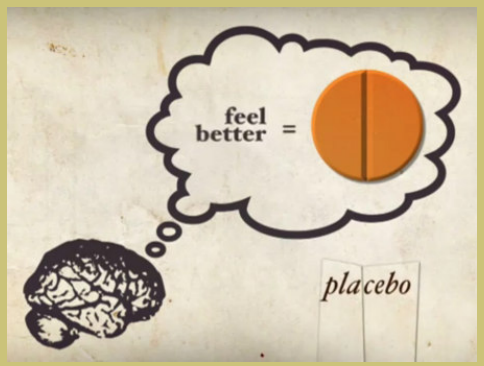
or

$$N = \left(\frac{t_{1-\frac{\alpha}{2},0,df_\epsilon} + t_{1-\beta,\frac{\Delta}{\frac{s}{\sqrt{N}}},df_\epsilon}}{\tilde{\Delta}} \right)^2 \tag{13}$$

with $\tilde{\Delta} = \frac{\Delta}{\sigma_\epsilon \sqrt{\sum_{i=1}^{T} c_i^2}}$ or $\tilde{\Delta} = \frac{\Delta}{s_\epsilon \sqrt{\sum_{i=1}^{T} c_i^2}}$.

1-way ANOVA contrasts

Example (continued)



Design for $\mu_0 - \frac{\mu_1+\mu_2}{2} = 0$

$$\tilde{\Delta} = \frac{10}{18\sqrt{1^2+\left(-\frac{1}{2}\right)^2+\left(-\frac{1}{2}\right)^2}} = 0.45$$

$$N = \left(\frac{z_{0.975}+z_{0.9}}{0.45}\right)^2 = 51.07 \rightarrow \boxed{N = 52}$$

Design for $\mu_1 - \mu_2 = 0$

$$\tilde{\Delta} = \frac{10}{18\sqrt{0^2+1^2+(-1)^2}} = 0.39$$

$$N = \left(\frac{z_{0.975}+z_{0.9}}{0.39}\right)^2 = 68.08 \rightarrow \boxed{N = 69}$$

The most limiting comparison is the second one, so we use $\boxed{N = 69}$.

Outline

Multiple testing correction

Example 10

Let us suppose that we are screening 1000 compounds vs a control, and that we have a confidence level of 95% ($\alpha = 0.05$) in each test. Let us assume that none of the compounds is effective for our disease. However, on average, we will reject the null hypothesis for 50 compounds, and we will incorrectly assume that these 50 compounds make a difference with respect to the control.

This problem is known as multiple testing.

For K comparisons:

Bonferroni correction:

$$\alpha = \frac{\alpha_{family}}{K} \tag{14}$$

Dunn-Sidak correction:

$$\alpha = 1 - (1 - \alpha_{family})^{(1/K)} \tag{15}$$

Multiple testing correction

When analyzing ANOVA data:

1. If we are to compare all effects vs all others: $K = \binom{T}{2} = \frac{T(T-1)}{2}$
2. If we are to compare all effects vs control: $K = T - 1$

ANOVA actually tests

$$H_0 : \mu_1 = \mu_2 ... = \mu_T$$
$$H_A : \text{ At least one } \mu_i \text{ is different from the rest}$$

Once the ANOVA test fails (not all treatments are the same), post-hoc comparisons are used. Remind to use a method that compensates for the multiple comparisons

- All vs all: Tukey's Honestly Significant Difference
- All vs control: Dunnett's test
- All vs best: Hsu's test
- Unplanned constrasts: Scheffé's test

All vs control

Suppose we have T treatments (Groups 1, 2, ..., T) that will be compared to a control group (Group 0). These are contrasts of the form

$$\begin{aligned}
\mu_0 - \mu_1 &= 0 \\
\mu_0 - \mu_2 &= 0 \\
&\dots \\
\mu_0 - \mu_T &= 0
\end{aligned}$$

The variance of the contrast estimate is

$$\mathrm{Var}\{\hat{\mu}_c\} = \sigma_\epsilon^2 \left(\frac{1}{N_0} + \frac{1}{N_i} \right)$$

All vs control

This variance is minimized (subject to $N_0 + TN_i = ct$) for

$$N_0 = N_i\sqrt{T} \tag{16}$$

and N_i given by the design formulas for ANOVA contrasts (Eqs. 12 and 13) with $\tilde{\Delta} = \frac{\Delta}{\sigma_\epsilon \sqrt{\frac{1}{N_i}\left(1+\frac{1}{\sqrt{T}}\right)}}$ or $\tilde{\Delta} = \frac{\Delta}{s_\epsilon \sqrt{\frac{1}{N_i}\left(1+\frac{1}{\sqrt{T}}\right)}}$.

Remind to correct α in some way (Bonferroni, Sidak, ...) to account for the multiple (T) comparisons, and that the number of degrees of freedom for ϵ of this design is

$$df_\epsilon = (N_0 - 1) + T(N_i - 1)$$

Outline

Summary

$$N = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\tilde{\Delta}} \right)^2 \qquad \tilde{\Delta} = \frac{\Delta}{\sigma} \tag{17}$$

1. One-sided or two-sided alternative test?
2. 1 or 2 independent samples?
3. 2 samples that can be reduced to 1 independent sample? (paired, ANOVA contrasts, ...)
4. Known or unknown variance?
5. Equivalence tests $\neq$ significance tests
6. Increase group size if more variance $N_2 = N_1 \frac{\sigma_2}{\sigma_1}$.
7. Decrease α for multiple comparisons.
8. Increase control size $N_0 = N_i \sqrt{T}$.

Outline

Chapter 1. Basic designs

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

October 14, 2016

Outline

Outline

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}
...	...
$y_{1,500}$	$y_{2,500}$

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

Completely Randomized Design

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

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

- μ 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.
- ϵ_{jk} is the part of the observed measurement that cannot be explained by the average and the treatment.

Completely Randomized Design

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

- $y_{..}$: average of all observations

$$y_{..} = \frac{1}{n} \sum_{jk} y_{jk} \approx \mu$$

- $y_{j.}$: average of observations in treatment j

$$y_{j.} = \frac{1}{n_j} \sum_{k} y_{jk} \approx \mu + t_j$$

Completely Randomized Design

The total variation of the data is

$$
\begin{aligned}
SS &= \sum_{jk}(y_{jk} - y_{..})^2 = \sum_{jk}\left(y_{jk}^2 + y_{..}^2 - 2y_{jk}y_{..}\right) \\
&= \sum_{jk} y_{jk}^2 + \sum_{jk} y_{..}^2 - \sum_{jk} 2y_{jk}y_{..} = \sum_{jk} y_{jk}^2 + ny_{..}^2 - 2y_{..}\sum_{jk} y_{jk} \\
&= \sum_{jk} y_{jk}^2 + ny_{..}^2 - 2ny_{..}^2 = \sum_{jk} y_{jk}^2 - ny_{..}^2 \\
&= \sum_{jk} y_{jk}^2 - n\left(\frac{1}{n}\sum_{jk} y_{jk}\right)^2 = \sum_{jk} y_{jk}^2 - \frac{\left(\sum_{jk} y_{jk}\right)^2}{n} \boxed{= \sum_{jk} y_{jk}^2 - \frac{Y_{..}^2}{n}}
\end{aligned}
$$

Completely Randomized Design

The treatment effect is estimated as

$$\hat{t}_j = y_{j\cdot} - y_{\cdot\cdot} \approx (\mu + t_j) - \mu = t_j$$

and its associated variance

$$\boxed{SS_T = \sum_{jk} \hat{t}_j^2 = \left(\sum_j \frac{Y_{j\cdot}^2}{n_j} \right) - \frac{Y_{\cdot\cdot}^2}{n}}$$

Similarly, for the residuals

$$\hat{\epsilon}_{jk} = y_{jk} - y_{j\cdot} \approx (\mu + t_j + \epsilon_{jk}) - (\mu + t_j) = \epsilon_{jk}$$

the sum of squares of the residuals (within the treatments)

$$\boxed{SS_\epsilon = \sum_{jk} \hat{\epsilon}_{jk}^2 = \sum_{jk} y_{jk}^2 - \sum_j \frac{Y_{j\cdot}^2}{n_j}}$$

Completely Randomized Design

The sum of squares of all measurements can be decomposed into a sum of different components

$$\begin{array}{rcl} SS & = & SS_T + SS_\epsilon \\ \sum_{jk}(y_{jk} - y_{..})^2 & = & \sum_{jk}(y_{j.} - y_{..})^2 + \sum_{jk}(y_{jk} - y_{j.})^2 \end{array}$$

and similarly for the degrees of freedom

$$n - 1 = \sum_j (n_j - 1) + (t - 1)$$

Remind in our example, $n = 1000$ (=total population), $t = 2$ (two treatments: drug and placebo), and $n_1 = n_2 = 500$ (500 individuals in each treatment).

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	$SS_T = \sum_{jk}(y_{j\cdot} - y_{\cdot\cdot})^2$	$t-1$	$MS_T = \frac{SS_T}{df_t}$
Residuals	$SS_\epsilon = \sum_{jk}(y_{jk} - y_{j\cdot})^2$	$\sum_j(n_j - 1) = n - t$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	$SS = \sum_{jk}(y_{jk} - y_{\cdot\cdot})^2$	$n-1$	

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

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

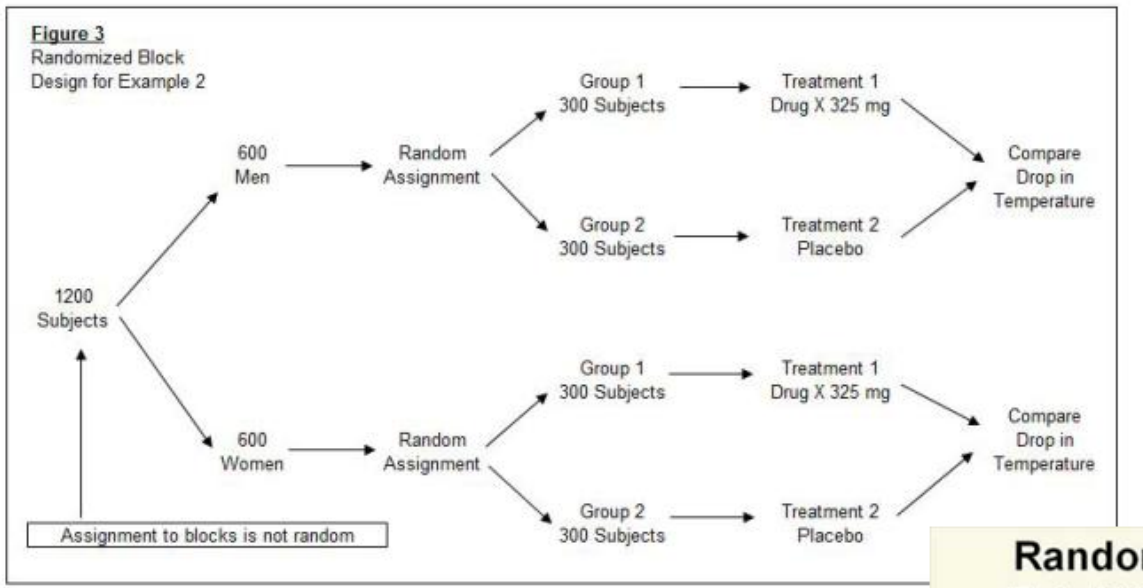
In this case

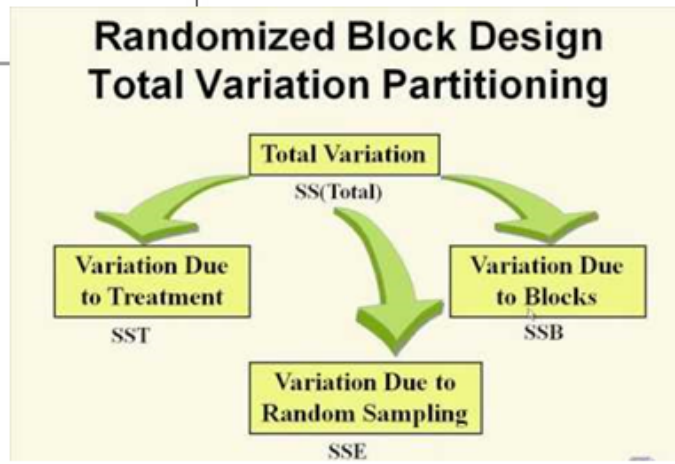
$$F = \frac{256.88}{13.61} = 18.87 \gg 3.85 = F_{0.95,1,998}$$

Outline

Reducing variance

Randomized block designs





Randomized Complete Block Design

Blocks are groups of experimental units that are formed to be as homogeneous as possible with respect to the block characteristics. The term block comes from the agricultural heritage of experimental design where a large block of land was selected for the various treatments, that had uniform soil, drainage, sunlight, and other important physical characteristics. Homogeneous clusters improve the comparison of treatments by randomly allocating levels of the treatments within each block. (SAS)

2 X 2

4 207	6 208	9 209	8 407	3 408	6 409
7 204	2 205	3 206	1 404	4 405	9 406
5 201	1 202	8 203	7 401	5 402	2 403
2 107	4 108	5 109	9 307	4 308	1 309
1 104	8 105	7 106	6 304	3 305	2 306
9 101	3 102	6 103	8 301	5 302	7 303

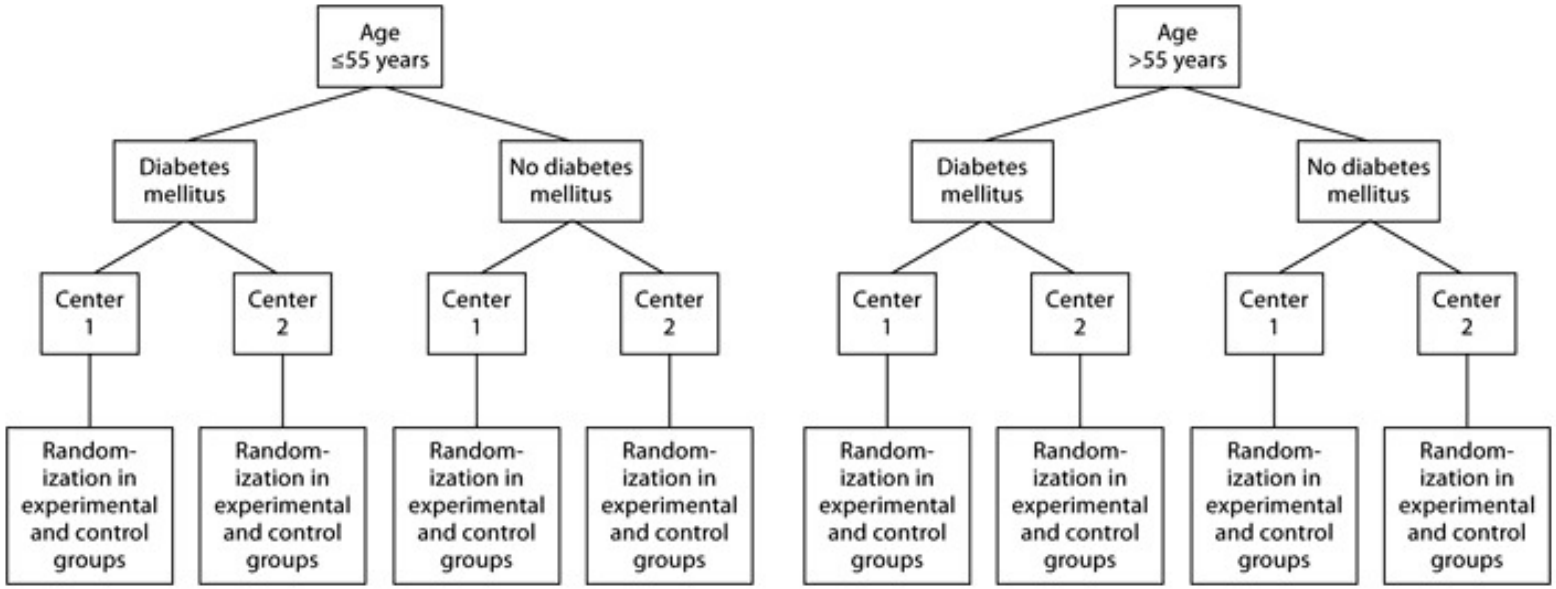
4 x 1

8 407	3 408	6 409
1 404	4 405	9 406
7 401	5 402	2 403
9 307	4 308	1 309
6 304	3 305	2 306
8 301	5 302	7 303
4 207	6 208	9 209
7 204	2 205	3 206
5 201	1 202	8 203
2 107	4 108	5 109
1 104	8 105	7 106
9 101	3 102	6 103

1 x 4

2 107	4 108	5 109	4 207	6 208	9 209	9 307	4 308	1 309	8 407	3 408	6 409
1 104	8 105	7 106	7 204	2 205	3 206	6 304	3 305	2 306	1 404	4 405	9 406
9 101	3 102	6 103	5 201	1 202	8 203	8 301	5 302	7 303	7 401	5 402	2 403

Randomized Complete Block Design



Within each block, experimental units must be randomly assigned to treatments. When several variables must be blocked, each combination (e.g. >55, Diabetes, Center 1) can be treated as a block. Alternatively, each block may be treated independently (we will see how later).

Randomized Complete Block Design

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

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

- μ 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.
- ϵ_{ijk} is the part of the observed measurement that cannot be explained by the average, block and treatment.

Randomized Complete Block Design

We now have the relationships

$$
\begin{aligned}
\hat{\mu} &= y_{...} \\
\hat{b}_i &= y_{i..} - y_{...} \approx (\mu + b_i) - \mu = b_i \\
\hat{t}_j &= y_{.j.} - y_{...} \approx (\mu + t_j) - \mu = t_j \\
\hat{\epsilon}_{ijk} &= y_{ijk} - y_{i..} - y_{.j.} + y_{...} = y_{ijk} - (\hat{\mu} + \hat{b}_i + \hat{t}_j) \\
&\approx (\mu + b_i + t_j + \epsilon_{ijk}) - (\mu + b_i) - (\mu + t_j) + \mu = \epsilon_{ijk} \\
SS &= \sum_{ijk}(y_{ijk} - y_{...})^2 = \sum_{ijk} y_{ijk}^2 - \frac{Y_{...}^2}{n} \\
SS_B &= \sum_{ijk} \hat{b}_i^2 \\
SS_T &= \sum_{ijk} \hat{t}_j^2 \\
SS_\epsilon &= \sum_{ijk} \hat{\epsilon}_{ijk}^2
\end{aligned}
$$

$$
\boxed{
\begin{aligned}
SS &= SS_B + SS_T + SS_\epsilon \\
n-1 &= (b-1) + (t-1) + (n-b-t+1)
\end{aligned}
}
$$

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_ϵ	$n-b-t+1$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
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

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

Randomized Complete Block Design

Example 3

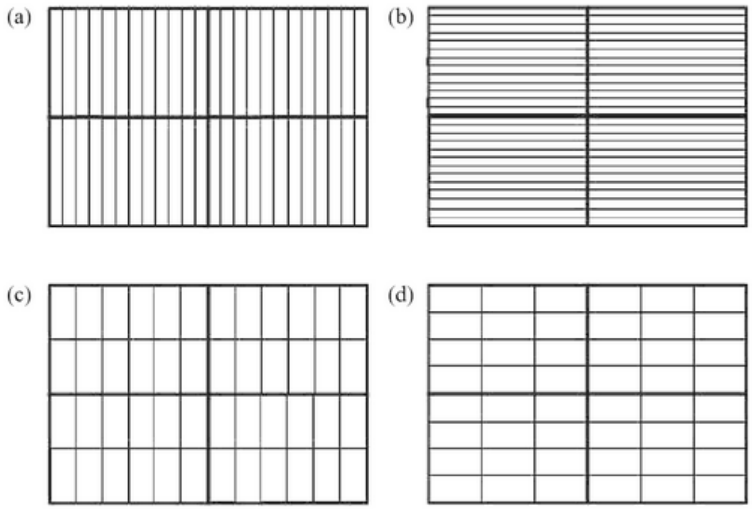
We want to analyze the optimal spacing (in terms of yield measured in kilos) between plants (10 treatments: 30×30, 30×24, 30×20, 30×15, 24×24, 24×20, 24×15, 20×20, 20×15, 15×15). To avoid possible land effects, we divide the land in 4 blocks, and within each block we randomly apply the 10 treatments.

We may compute the difference between many pairs of treatments, creating a problem of Type I error inflation by multiple testing. Instead, we may analyze the data converting the treatments to a numerical variable (area per plant, e.g. $30 \times 30 = 900$) and performing a regression analysis of yield versus area and making the hypothesis testing only on a single parameter, the slope.

Randomized Complete Block Design

- If there are clear variables to block, they should be blocked. Litters are normally chosen as blocks (and birth weight as covariate).

- If there are no obvious blocking variables, but we may create blocks, we may do as an "insurance" against possible patterns not yet identified.

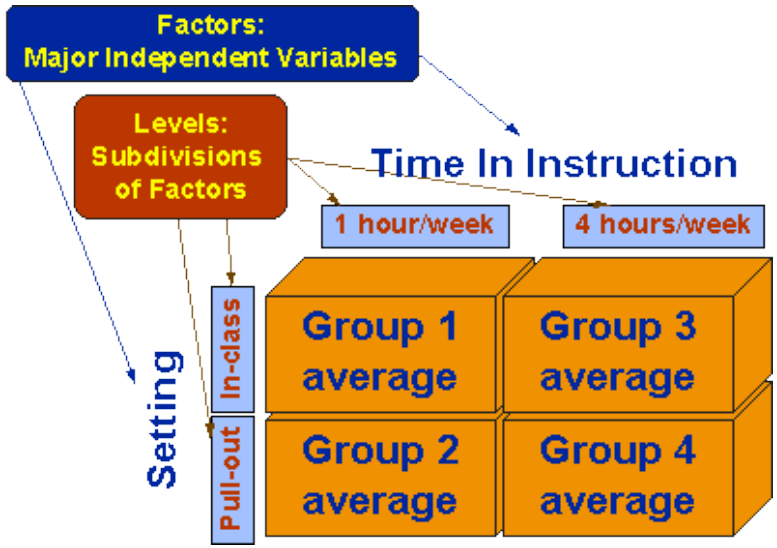


(e.g. 4 block, 12 treatments)

Outline

Factorial Design

Let's imagine a design where we have an educational program where we would like to look at a variety of program variations to see which works best. For instance, we would like to vary the amount of time the children receive instruction with one group getting 1 hour of instruction per week and another getting 4 hours per week. And, we'd like to vary the setting with one group getting the instruction in-class (probably pulled off into a corner of the classroom) and the other group being pulled-out of the classroom for instruction in another room.

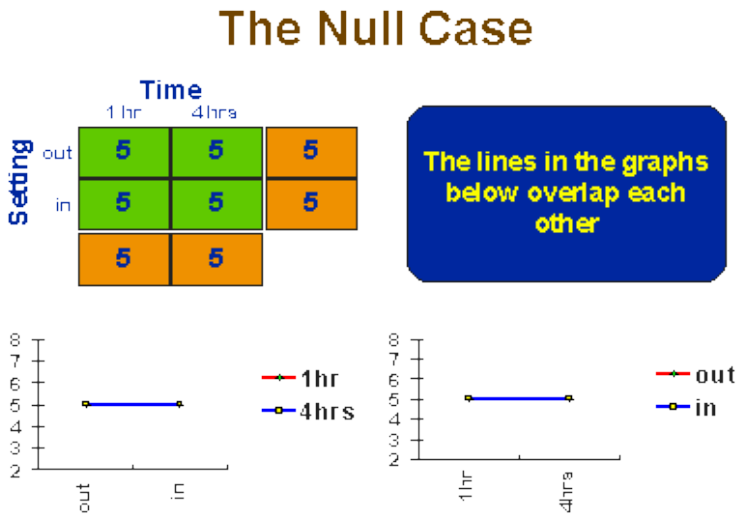


Factorial Design

The data is supposed to be generated as

$$y_{ijk} = \mu + p_i + q_j + \epsilon_{ijk}$$

Treatment variables are P(=amount of time) and Q (=setting).

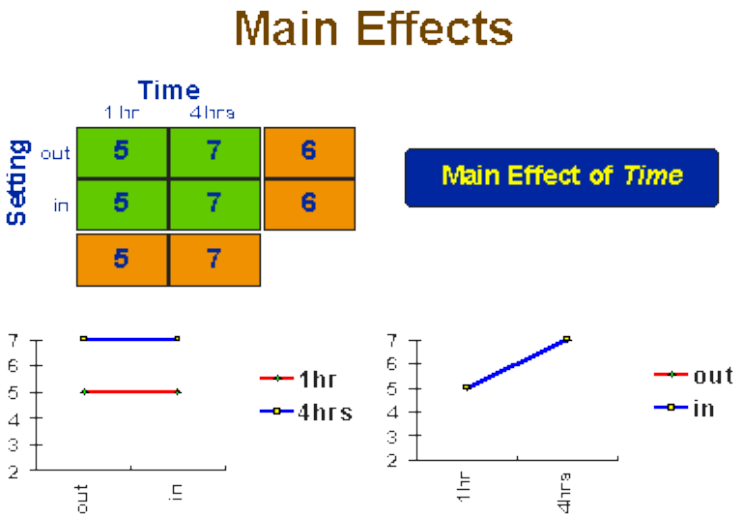


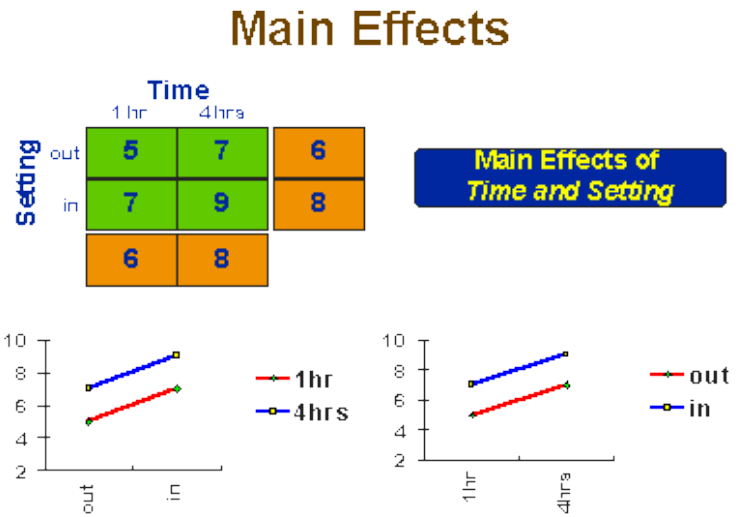
In case that there is no effect of any of the variables, we should not observe differences amongst the groups.

5+0+0	5+0+0	$q_1 = 0$
5+0+0	5+0+0	$q_2 = 0$
$p_1 = 0$	$p_2 = 0$	$\mu = 5$

Factorial Design

Main effects are the consistent differences observed for the levels of each one of the factors.





Outcome example if the amount of time has an effect but the setting does not.

6-1+0	6+1+0	$q_1 = 0$
6-1+0	6+1+0	$q_2 = 0$
$p_1 = -1$	$p_2 = 1$	$\mu = 6$

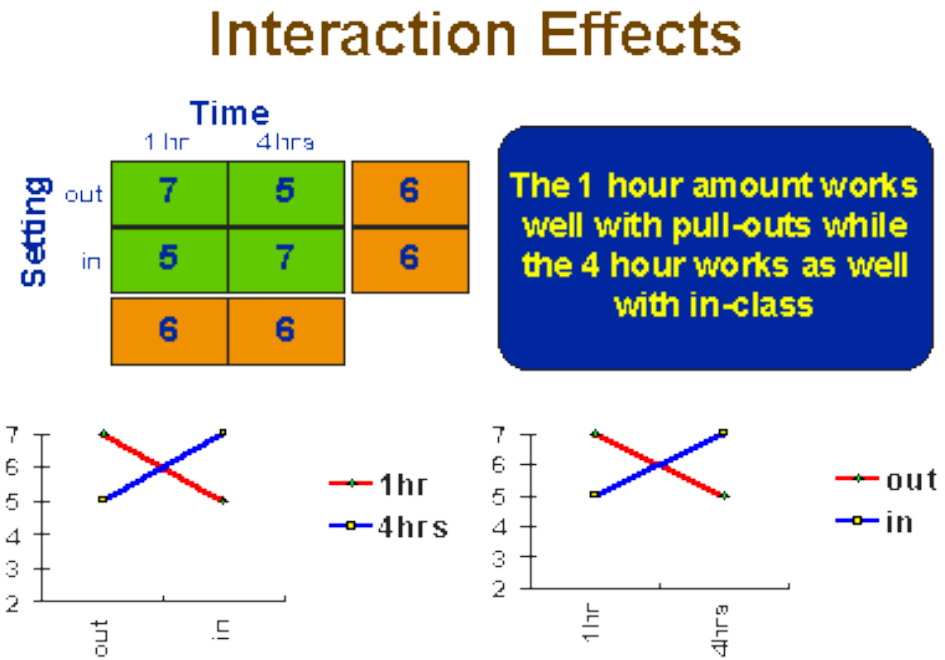
Outcome example if the amount of time and the setting have an effect.

7-1-1	7+1-1	$q_1 = -1$
7-1+1	7+1+1	$q_2 = 1$
$p_1 = -1$	$p_2 = 1$	$\mu = 7$

Factorial Design

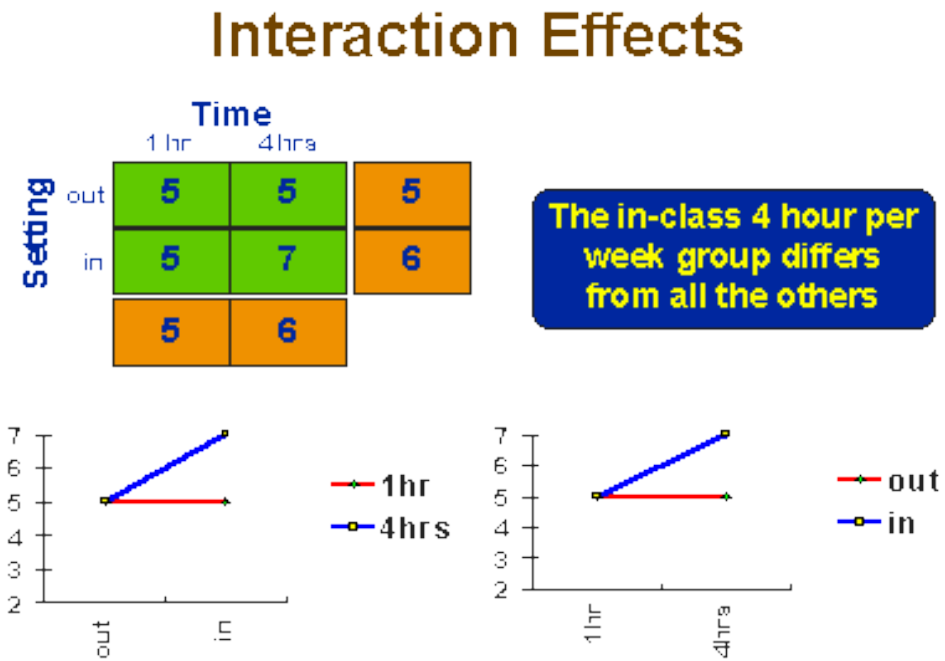
Interaction effects exist when differences on one factor depend on the level you are on another factor. The interactions are between factors and not between levels.

$$y_{ijk} = \mu + p_i + q_j + (pq)_{ij} + \epsilon_{ijk}$$



6+0+0+1$(= (pq)_{11})$	6+0+0-1$(= (pq)_{12})$	$q_1 = 0$
6+0+0-1$(= (pq)_{21})$	6+0+0+1$(= (pq)_{22})$	$q_2 = 0$
$p_1 = 0$	$p_2 = 0$	$\mu = 6$

Factorial Design



5.5-0.5-0.5+0.5$(= (pq)_{11})$	5.5+0.5-0.5-0.5$(= (pq)_{12})$	$q_1 = -0.5$
5.5-0.5+0.5-0.5$(= (pq)_{21})$	5.5+0.5+0.5+0.5$(= (pq)_{22})$	$q_2 = 0.5$
$p_1 = -0.5$	$p_2 = 0.5$	$\mu = 5.5$

Factorial Design

Given the linear model

$$y_{ijk} = \mu + p_i + q_j + (pq)_{ij} + \epsilon_{ijk}$$

The model constraints are

$$\sum_i p_i = \sum_j q_j = \sum_i (pq)_{ij} = \sum_j (pq)_{ij} = 0$$

and we may estimate each one of the components as

$$\hat{\mu} = y_{...} \qquad SS = \sum_{ijk}(y_{ijk} - \hat{\mu})^2 \qquad df = n - 1$$

$$\hat{p}_i = y_{i..} - y_{...} \qquad SS_P = \sum_{ijk} \hat{p}_i^2 \qquad df_P = p - 1$$

$$\hat{q}_j = y_{.j.} - y_{...} \qquad SS_Q = \sum_{ijk} \hat{q}_j^2 \qquad df_Q = q - 1$$

$$\widehat{(pq)}_{ij} = y_{ij.} - y_{i..} - y_{.j.} + y_{...} \qquad SS_{PQ} = \sum_{ijk} \widehat{(pq)}_{ij}^2 \qquad df_{PQ} = (p-1)(q-1)$$

$$\hat{\epsilon}_{ijk} = y_{ijk} - y_{ij.} \qquad SS_\epsilon = \sum_{ijk} \hat{\epsilon}_{ijk}^2 \qquad df_\epsilon = n - pq$$

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_ϵ	$n-pq$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	SS	$n-1$	

Factorial Design

Example 4

We are testing water uptake by amphibia. Frogs and toads (species factor S) are kept in most 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
Species	515.06	1	
Moisture	471.33	1	
Hormone	218.01	1	
SM	39.50	1	
SH	165.12	1	
MH	57.73	1	
SMH	43.43	1	
Error	276.05	8	$s^2 = 34.51$
Total	1786.33	15	

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

Factorial Design

Factors and blocks: Factors and blocks may be combined, the difference between a block and a factor is that it makes no sense to study the interaction of blocks

$$y_{ijkl} = \mu + b_i + p_j + q_k + (pq)_{jk} + \epsilon_{ijkl}$$

The model constraints are

$$\sum_i b_i = \sum_j p_j = \sum_k q_k = \sum_j (pq)_{jk} = \sum_k (pq)_{jk} = 0$$

and we may estimate each one of the components as

$\hat{\mu} = y....$	$SS = \sum_{ijkl}(y_{ijkl} - \hat{\mu})^2$	$df = n - 1$
$\hat{b}_i = y_{i...} - y....$	$SS_B = \sum_{ijkl} \hat{b}_i^2$	$df_B = b - 1$
$\hat{p}_j = y_{.j..} - y....$	$SS_P = \sum_{ijkl} \hat{p}_j^2$	$df_P = p - 1$
$\hat{q}_k = y_{..k.} - y....$	$SS_Q = \sum_{ijkl} \hat{q}_k^2$	$df_Q = q - 1$
$\widehat{(pq)}_{jk} = y_{.jk.} - y_{.j..} - y_{..k.} + y....$	$SS_{PQ} = \sum_{ijkl} \widehat{(pq)}_{jk}^2$	$df_{PQ} = (p-1)(q-1)$
$\hat{\epsilon}_{ijkl} = y_{ijkl} - y_{i...} - y_{.jk.} + y....$	$SS_\epsilon = \sum \hat{\epsilon}_{ijkl}^2$	$df_\epsilon = n - pq - b - 1$

Factorial Design

Advantages of factorial design:

- Interactions between factors can be estimated and their significance tested.
- Wider validity of main effects: they have been tested in many different cases (e.g. the effect of moisture have been tested with frogs and toads, and with and without hormone)
- Several experiments are done simultaneously: the variance of pairwise comparisons is minimal, as shown in the following experiment

Factorial Design

Example 5

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 6(continued)

We now want to test if there is a difference induced by the hormone injection, for which we construct the statistic

$$\Delta h = h_0 - h_1$$

Its variance in the three experiments are

1. $\sigma^2_{\Delta h} = 2\frac{\sigma^2_\epsilon}{4}$
2. $\sigma^2_{\Delta h} = 2\frac{\sigma^2_\epsilon}{6}$
3. $\sigma^2_{\Delta h} = 2\frac{\sigma^2_\epsilon}{12}$

The factorial design yields the smallest variance for the comparison of any of its components.

Factorial design: Hold all factors constant except ~~the one~~ those whose effects we are investigating.

Outline

Non-orthogonal Designs

Example 6

We are testing 2 spray treatments (t_k) using 2 different concentrations of a chemical growth regulator. We also include a control spray without the chemical. We have 9 plots (3×3) for the experiment and we allow for row (r_i) and column (c_j) differences

$$y_{ijkl} = \mu + r_i + c_j + t_k + \epsilon_{ijkl}$$

Results are	A 3.72	B 3.39	C 2.95
	C 3.50	A 3.08	B 1.72
	B 4.18	C 4.36	A 0.81

This is a latin square and the analysis techniques are not the same as in the randomized complete block design (the reason is that in block designs, for each block (in our case row and column) we assume that we have all treatments, and this is not the case.

Example 6(continued)

The solution comes through Least Squares fitting

$$
\begin{aligned}
3.72 &= \mu + r_1 + c_1 + t_A \\
3.39 &= \mu + r_1 + c_2 + t_B \\
2.95 &= \mu + r_1 + c_3 + t_C \\
3.50 &= \mu + r_2 + c_1 + t_C \\
3.08 &= \mu + r_2 + c_2 + t_A \\
1.72 &= \mu + r_2 + c_3 + t_B \\
4.18 &= \mu + r_3 + c_1 + t_B \\
4.36 &= \mu + r_3 + c_2 + t_C \\
0.81 &= \mu + r_3 + c_3 + t_A
\end{aligned}
$$

Non-orthogonal Designs

Example 6(continued)

$$\mathbf{y} = A\theta$$

$$\begin{pmatrix} 3.72 \\ 3.39 \\ 2.95 \\ 3.50 \\ 3.08 \\ 1.72 \\ 4.18 \\ 4.36 \\ 0.81 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ r_1 \\ r_2 \\ r_3 \\ c_1 \\ c_2 \\ c_3 \\ t_A \\ t_B \\ t_C \end{pmatrix}$$

However we have not introduced yet the constraints

$$r_3 = -r_1 - r_2, c_3 = -c_1 - c_2, t_C = -t_A - t_B$$

Example 6(continued)

With the constraints, the LS problem becomes

$$\begin{pmatrix} 3.72 \\ 3.39 \\ 2.95 \\ 3.50 \\ 3.08 \\ 1.72 \\ 4.18 \\ 4.36 \\ 0.81 \end{pmatrix} = \begin{bmatrix} \mu & r_1 & r_2 & c_1 & c_2 & t_A & t_B \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & 0 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 0 \end{bmatrix} \begin{pmatrix} \mu \\ r_1 \\ r_2 \\ c_1 \\ c_2 \\ t_A \\ t_B \end{pmatrix}$$

Note that for any pair of factor, their corresponding columns in the design matrix are orthogonal

$$\langle \mu, r_i \rangle = \langle \mu, c_j \rangle = \langle \mu, t_k \rangle = \langle r_i, c_j \rangle = \langle r_i, t_k \rangle = \langle c_j, t_k \rangle = 0$$

Non-orthogonal Designs

Example 7

We are now given 3 extra plots (another row), which we employ to replicate the treatments and have better estimates.

Results are now

A 3.72	B 3.39	C 2.95
C 3.50	A 3.08	B 1.72
B 4.18	C 4.36	A 0.81
C 5.45	B 5.26	A 4.85

Non-orthogonal Designs

Example 7(continued)

$$\begin{pmatrix} 3.72 \\ 3.39 \\ 2.95 \\ 3.50 \\ 3.08 \\ 1.72 \\ 4.18 \\ 4.36 \\ 0.81 \\ 5.45 \\ 5.26 \\ 4.85 \end{pmatrix} = \begin{bmatrix} \mu & r_1 & r_2 & r_3 & c_1 & c_2 & t_A & t_B \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & 0 & 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & -1 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & -1 & -1 \\ 1 & 0 & 0 & 1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & -1 & 1 & 0 & -1 & -1 \\ 1 & -1 & -1 & -1 & 0 & 1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 1 & 0 \end{bmatrix} \begin{pmatrix} \mu \\ r_1 \\ r_2 \\ r_3 \\ c_1 \\ c_2 \\ t_A \\ t_B \end{pmatrix}$$

Factor columns in the design matrix are no longer orthogonal (in particular $\langle c_j, t_k \rangle \neq 0$).

Non-orthogonal Designs

- Orthogonal designs are insensitive to the order in which the parameters are fitted. We may fit all of them at the same time (as shown), or
  1. fit first μ, produce a new experiment dataset removing the part we have already fitted (μ)
  2. fit then r_i and c_j, produce a new experiment dataset removing the part we have already fitted (μ, r_i, c_j)
  3. fit finally the treatments (t_k)
- Non-orthogonal designs depend on the order in which parameters are fitted (nothing terrible, but something to keep in mind).

Outline

Covariates

Researchers cannot control covariates, but can measure them and use them to increase the predictive power of the Linear Model.

Example 8

We suspect that the effect of the growth chemical depends on the ambient temperature, we extend the model with this covariate

$$y_{ijkl} = \mu + r_i + c_j + t_k + \beta T_{ijkl} + \epsilon_{ijkl}$$

T_{ijkl} is the ambient temperature measured when the spray was applied.

A 3.72 (T=28)	B 3.39 (T=22)	C 2.95 (T=23)
C 3.50 (T=24)	A 3.08 (T=25)	B 1.72 (T=26)
B 4.18 (T=20)	C 4.36 (T=22)	A 0.81 (T=26)

Covariates

Example 8(continued)

$$\mathbf{y} = A\boldsymbol{\theta}$$

$$\begin{pmatrix} 3.72 \\ 3.39 \\ 2.95 \\ 3.50 \\ 3.08 \\ 1.72 \\ 4.18 \\ 4.36 \\ 0.81 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 28 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 22 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 23 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 24 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 25 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 26 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 20 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 22 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 26 \end{pmatrix} \begin{pmatrix} \mu \\ r_1 \\ r_2 \\ r_3 \\ c_1 \\ c_2 \\ c_3 \\ t_A \\ t_B \\ t_C \\ \beta \end{pmatrix}$$

Contrasts

Example 9

Remind that our simplified parameter vector is

$$\boldsymbol{\theta} = (\mu, r_1, r_2, c_1, c_2, t_A, t_B)^T$$

We want to know whether there is a difference in the spray treatment

$$t_A - t_B = 0 = (0, 0, 0, 0, 0, 1, -1)^T \boldsymbol{\theta}$$

or if there are differences in the rows

$$\begin{aligned} r_1 - r_2 = 0 &= (0, 1, -1, 0, 0, 0, 0)^T \boldsymbol{\theta} \\ r_2 - r_3 = 0 &= r_2 - (-r_1 - r_2) = 2r_2 + r_1 \\ &= (0, 1, 2, 0, 0, 0, 0)^T \boldsymbol{\theta} \end{aligned}$$

In general, many interesting tests are of the form $\mathbf{c}^T \boldsymbol{\theta} = 0$.
If $\mathbf{1}^T \mathbf{c} = 0$, $\mathbf{c}$ is called a contrast.

Outline

Least squares

The linear model is of the form

$$\mathbf{y} = A\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

and it assumes

$$\begin{aligned} E\{\boldsymbol{\epsilon}\} &= \mathbf{0} \\ \Sigma_{\boldsymbol{\epsilon}} &= \sigma_{\epsilon}^2 I \end{aligned}$$

Consequently

$$E\{\mathbf{y}\} = A\boldsymbol{\theta}$$

And the deviations from the expected value is the sum of squares

$$SS = (\mathbf{y} - A\boldsymbol{\theta})^T(\mathbf{y} - A\boldsymbol{\theta})$$

The minimizer of this Sum of Squares is

$$\hat{\boldsymbol{\theta}} = (A^T A)^{-1} A^T \mathbf{y}$$

Least squares

The covariance matrix of the fitting parameters (assuming that ϵ is a multivariate normal) is

$$Cov\{\hat{\boldsymbol{\theta}}\} = \sigma_\epsilon^2 (A^T A)^{-1}$$

If we diagonalize $A^T A$, then after some suitable rotation P

$$Cov\{P\hat{\boldsymbol{\theta}}\} = \begin{pmatrix} \frac{\sigma_\epsilon^2}{\lambda_1^2} & 0 & \dots & 0 \\ 0 & \frac{\sigma_\epsilon^2}{\lambda_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\sigma_\epsilon^2}{\lambda_M^2} \end{pmatrix}$$

being λ_1, λ_2, ... λ_M the Singular Values of the matrix A

The goal of the Experimental Design is to construct a matrix A such that: 1) $A^T A$ has a determinant as small as possible; or 2) the variance of a specific parameter is as small as possible. We would also like the matrix A to be well-conditioned (otherwise some parameter will be too variable).

Least squares

If in our experiment the most important test is of the form

$$c = \mathbf{c}^T \boldsymbol{\theta} = 0$$

we may design our experiment such that the variance of c is minimized

$$\mathrm{Var}\{c\} = \sigma_\epsilon^2 \mathbf{c}^T (A^T A)^{-1} \mathbf{c}$$

The goal of the Experimental Design is to construct a matrix A such that: ... or 3) the variance of a specific statistic is as small as possible.

Particular structures (Factorial Design, Completely Randomized Design, Randomized Complete Block Design) are "precooked" A constructions, which additionally allow very easy Least Squares fitting.

Outline