



CEU

*Universidad  
San Pablo*

# Practical Statistical Questions

Session 0: Course outline

Carlos Óscar Sánchez Sorzano, Ph.D.  
Madrid, July 7th 2008

# Motivation for this course

http://lib.stat.cmu.edu/datasets/Plasma\_Retinol - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Reload Print

Address http://lib.stat.cmu.edu/datasets/Plasma\_Retinol

Google G multivariate dataset Go

Determinants of Plasma Retinol and Beta-Carotene Levels

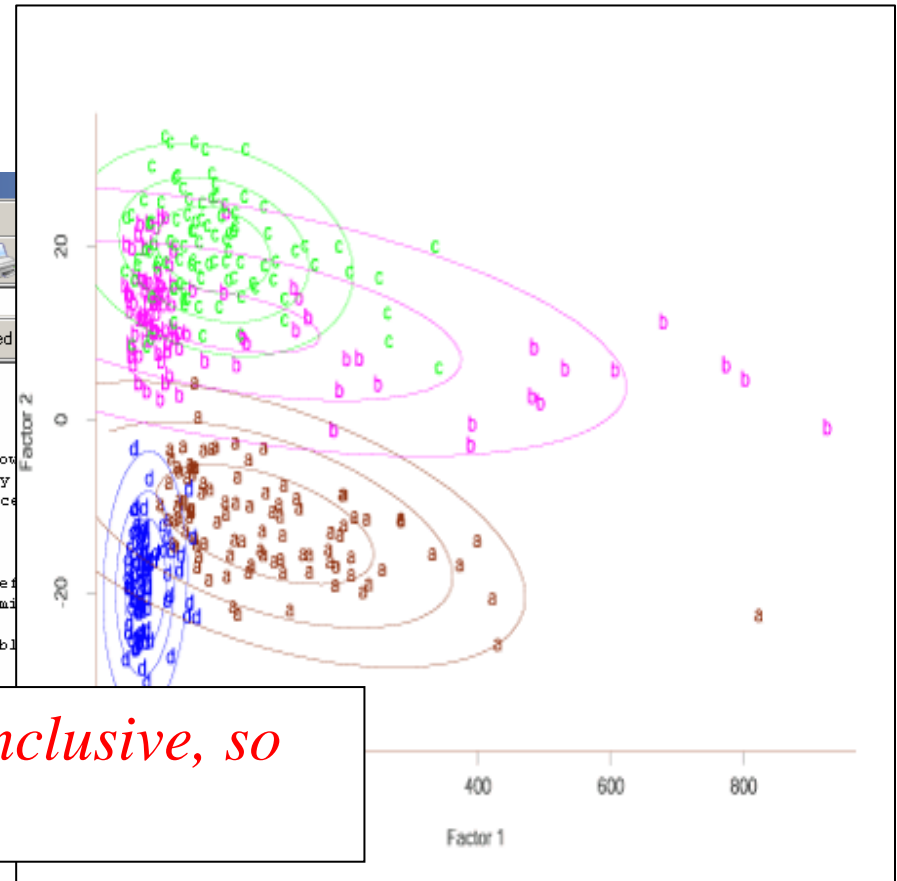
Summary:  
Observational studies have suggested that low dietary intake or low plasma concentrations of the micronutrients varied widely. We conclude that there is wide variability in plasma concentrations.

Authorization: Contact Authors

Reference: These data have not been published yet but a related reference is Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of Plasma Retinol and Beta-Carotene Levels. JAMA. 1993;269:1000-1006.

Description: This datafile contains 315 observations on 14 variables. The variables are:

Variable Names in order from left to right:



*The result of the experiment was inconclusive, so we had to use statistics!!*

ALCOHOL: Number of alcoholic drinks consumed per week.													
CHOLESTEROL: Cholesterol consumed (mg per day).													
BETADIET: Dietary beta-carotene consumed (mcg per day).													
RETDIET: Dietary retinol consumed (mcg per day)													
BETAPLASMA: Plasma beta-carotene (ng/ml)													
RETPLASMA: Plasma Retinol (ng/ml)													
64	2	2	21.4838	1	1298.8	57	6.3	0	170.3	1945	890	200	915
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108	2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562

## Motivation for this course



June 2008

Statistics  
506.000.000

Rodríguez Zapatero  
4.660.000



June 2008

Statistics  
>5000

Quantitative Research Analyst- Statistics Expert	eka finance New York	May 17
<a href="#">+ Expand</a>	<a href="#">Save</a>   <a href="#">More like this</a>	
Statistics Research Analyst 137933	Walt Disney Parks & Resorts Orlando	May 17
<a href="#">+ Expand</a>	<a href="#">Save</a>   <a href="#">More like this</a>	
Manager of Clinical Development Statistics	Trans Tech Pharma High Point, NC 27265	May 16
<a href="#">+ Expand</a>	<a href="#">View Map</a>   <a href="#">Save</a>   <a href="#">More like this</a>	
Visual C++/MFC Developer with a statistics background	Robert Half Technology PARAMUS	May 15
<a href="#">+ Expand</a>	<a href="#">Save</a>   <a href="#">More like this</a>	

## Course outline



# Course outline

1. I would like to know the intuitive definition and use of ...: The basics
  1. Descriptive vs inferential statistics
  2. Statistic vs parameter. What is a sampling distribution?
  3. Types of variables
  4. Parametric vs non-parametric statistics
  5. What to measure? Central tendency, differences, variability, skewness and kurtosis, association
  6. Use and abuse of the normal distribution
  7. Is my data really independent?

## Course outline

2. How do I collect the data? Experimental design
  1. Methodology
  2. Design types
  3. Basics of experimental design
  4. Some designs: Randomized Complete Blocks, Balanced Incomplete Blocks, Latin squares, Graeco-latin squares, Full  $2^k$  factorial, Fractional  $2^{k-p}$  factorial
  5. What is a covariate?
3. Now I have data, how do I extract information? Parameter estimation
  1. How to estimate a parameter of a distribution?
  2. How to report on a parameter of a distribution? What are confidence intervals?
  3. What if my data is “contaminated”? Robust statistics



## Course outline

4. Can I see any interesting association between two variables, two populations, ...?
  1. What are the different measures available?
  2. Use and abuse of the correlation coefficient
  3. How can I use models and regression to improve my measure of association?

## Course outline

5. How can I know if what I see is “true”? Hypothesis testing
  1. The basics: What is a hypothesis test? What is the statistical power? What is a p-value? How to use it? What is the relationship between sample size, sampling error, effect size and power? What are bootstraps and permutation tests?
  2. What are the assumptions of hypothesis testing?
  3. How to select the appropriate statistical test
    - i. Tests about a population central tendency
    - ii. Tests about a population variability
    - iii. Tests about a population distributions
    - iv. Tests about differences randomness
    - v. Tests about correlation/association measures
  4. Multiple testing
  5. Words of caution



## Course outline

6. How many samples do I need for my test?: Sample size
  1. Basic formulas for different distributions
  2. Formulas for samples with different costs
  3. What if I cannot get more samples? Resampling: Bootstrapping, jackknife

# Course outline

7. Can I deduce a model for my data?
  1. What kind of models are available?
  2. How to select the appropriate model?
  3. Analysis of Variance as a model
    1. What is ANOVA really?
    2. What is ANCOVA?
    3. How do I use them with pretest-posttest designs?
    4. What are planned and post-hoc contrasts?
    5. What are fixed-effects and random-effects?
    6. When should I use Multivariate ANOVA (MANOVA)?
  4. Regression as a model
    1. What are the assumptions of regression
    2. Are there other kind of regressions?
    3. How reliable are the coefficients? Confidence intervals
    4. How reliable are the coefficients? Validation

## Suggested readings: Overviews

It is suggested to read:

- Basics of probability
- Basics of design of experiments
- Basics of Hypothesis Testing
- Basics of ANOVA
- Basics of regression

## Bibliography

- D. J. Sheskin. Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC (2007)
- G. van Belle. Statistical Rules of Thumb. Wiley-Interscience (2002)
- R. R. Newton, K. E. Rudestam. Your Statistical Consultant: Answers to Your Data Analysis Questions. Sage Publications, Inc (1999)
- P. I. Good, J. W. Hardin. Common Errors in Statistics (and How to Avoid Them). Wiley-Interscience (2006)
- D. C. Montgomery. Design and analysis of experiments. John Wiley & Sons (2001)
- S. B. Vardeman. Statistics for engineering problem solving. IEEE Press (1994)
- G. K. Kanji. 100 Statistical tests. Sage publications (2006)



CEU

*Universidad  
San Pablo*

# Practical Statistical Questions

Session 1

Carlos Óscar Sánchez Sorzano, Ph.D.  
Madrid, July 7th 2008

# Course outline

1. I would like to know the intuitive definition and use of ...: The basics
  1. Descriptive vs inferential statistics
  2. Statistic vs parameter. What is a sampling distribution?
  3. Types of variables
  4. Parametric vs non-parametric statistics
  5. What to measure? Central tendency, differences, variability, skewness and kurtosis, association
  6. Use and abuse of the normal distribution
  7. Is my data really independent?

# 1.1 Descriptive vs Inferential Statistics

Statistics  
(=“state  
arithmetic”)

## **Descriptive: describe data**

- How rich are our citizens on average? → Central Tendency
- Are there many differences between rich and poor? → Variability
- Are more intelligent people richer? → Association
- How many people earn this money? → Probability distribution
- Tools: tables (all kinds of summaries), graphs (all kind of plots), distributions (joint, conditional, marginal, ...), statistics (mean, variance, correlation coefficient, histogram, ...)

## **Inferential: derive conclusions and make predictions**

- Is my country so rich as my neighbors? → Inference
- To measure richness, do I have to consider EVERYONE? → Sampling
- If I don't consider everyone, how reliable is my estimate? → Confidence
- Is our economy in recession? → Prediction
- What will be the impact of an expensive oil? → Modelling
- Tools: Hypothesis testing, Confidence intervals, Parameter estimation, Experiment design, Sampling, Time models, Statistical models (ANOVA, Generalized Linear Models, ...)





## 1.1 Descriptive vs Inferential Statistics

Of 350 randomly selected people in the town of Luserna, Italy, 280 people had the last name Nicolussi.

Which of the following sentences is descriptive and which is inferential:

1. 80% of THESE people of Luserna has Nicolussi as last name.
2. 80% of THE people of ITALY has Nicolussi as last name.

On the last 3 Sundays, Henry D. Carsalesman sold 2, 1, and 0 new cars respectively.

Which of the following sentences is descriptive and which is inferential:

1. Henry averaged 1 new car sold of the last 3 sundays.
2. Henry never sells more than 2 cars on a Sunday

What is the problem with the following sentence:

3. Henry sold no car last Sunday because he fell asleep inside one of the cars.

Source: [http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Data\\_Descr\\_Infer.htm](http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Data_Descr_Infer.htm)



## 1.1 Descriptive vs Inferential Statistics

The last four semesters an instructor taught Intermediate Algebra, the following numbers of people passed the class: 17, 19, 4, 20

Which of the following conclusions can be obtained from purely descriptive measures and which can be obtained by inferential methods?

- a) The last four semesters the instructor taught Intermediate Algebra, an average of 15 people passed the classs
- b) The next time the instructor teaches Intermediate Algebra, we can expect approximately 15 people to pass the class.
- c) This instructor will never pass more than 20 people in an Intermediate Algebra class.
- d) The last four semesters the instructor taught Intermediate Algebra, no more than 20 people passed the class.
- e) Only 5 people passed one semester because the instructor was in a bad mood the entire semester.
- f) The instructor passed 20 people the last time he taught the class to keep the administration off of his back for poor results.
- g) The instructor passes so few people in his Intermediate Algebra classes because he doesn't like teaching that class.

Source: [http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Data\\_Descr\\_Infer.htm](http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Data_Descr_Infer.htm)

## 1.2 Statistic vs. Parameter. What is a sampling distribution?

**Statistic:** characteristic of a sample

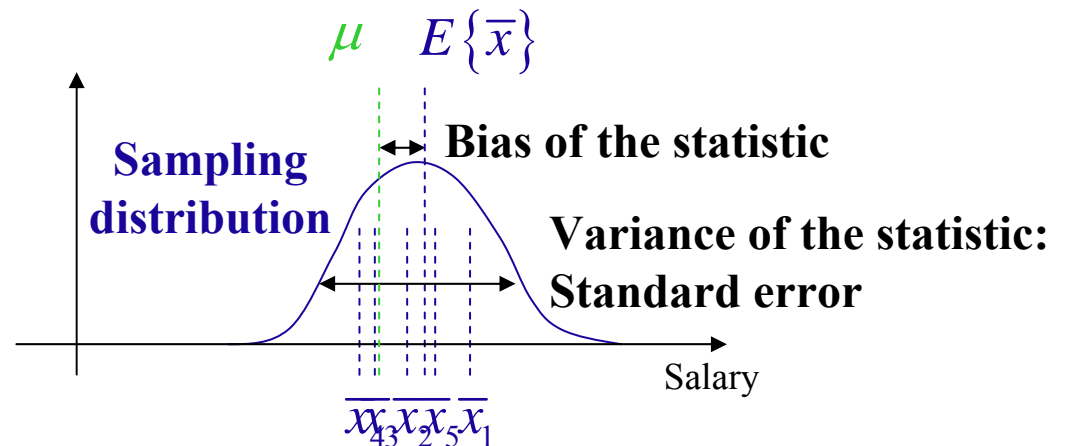
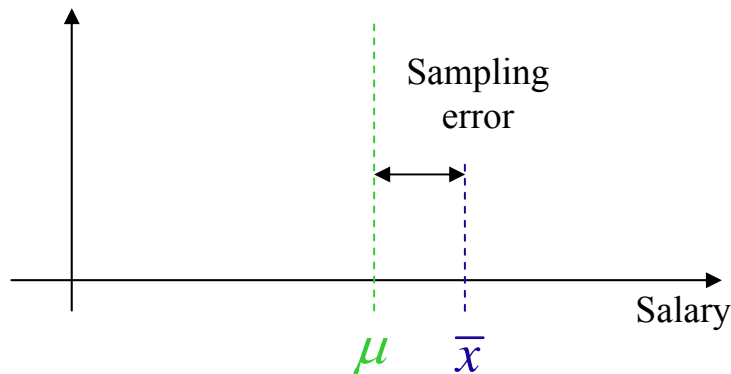
What is the average salary of 2000 people randomly sampled in Spain?

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

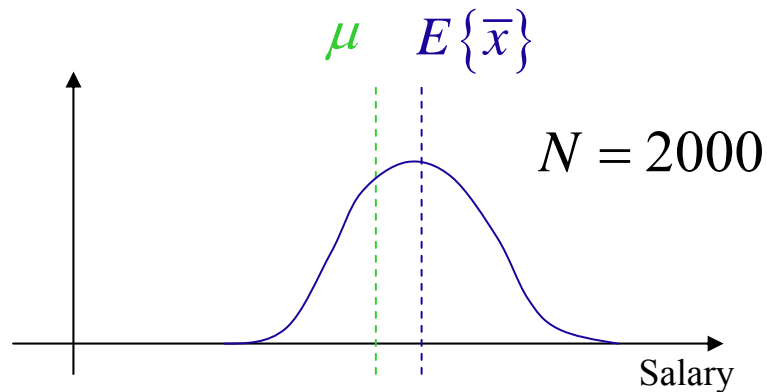
**Parameter:** characteristic of a population

What is the average salary of all Spaniards?

$\mu$



## 1.2 Statistic vs. Parameter. What is a sampling distribution?

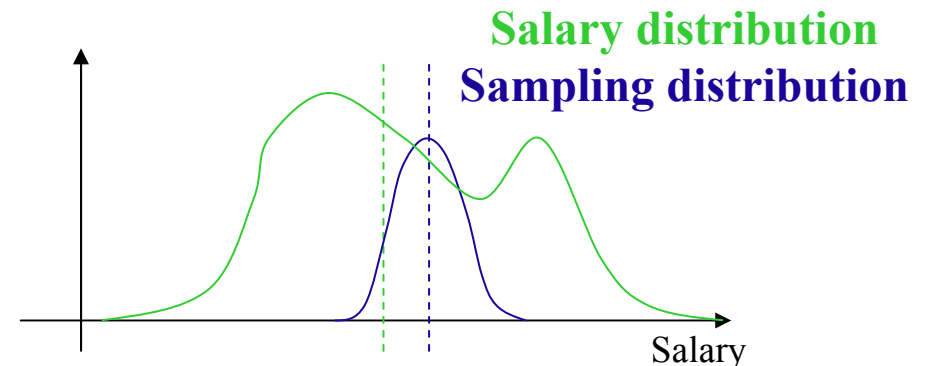
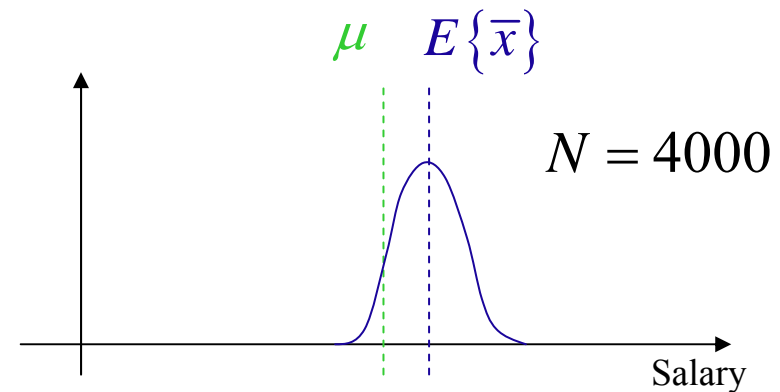


Unbiased

$$\mu - E\{\bar{x}\} = 0$$

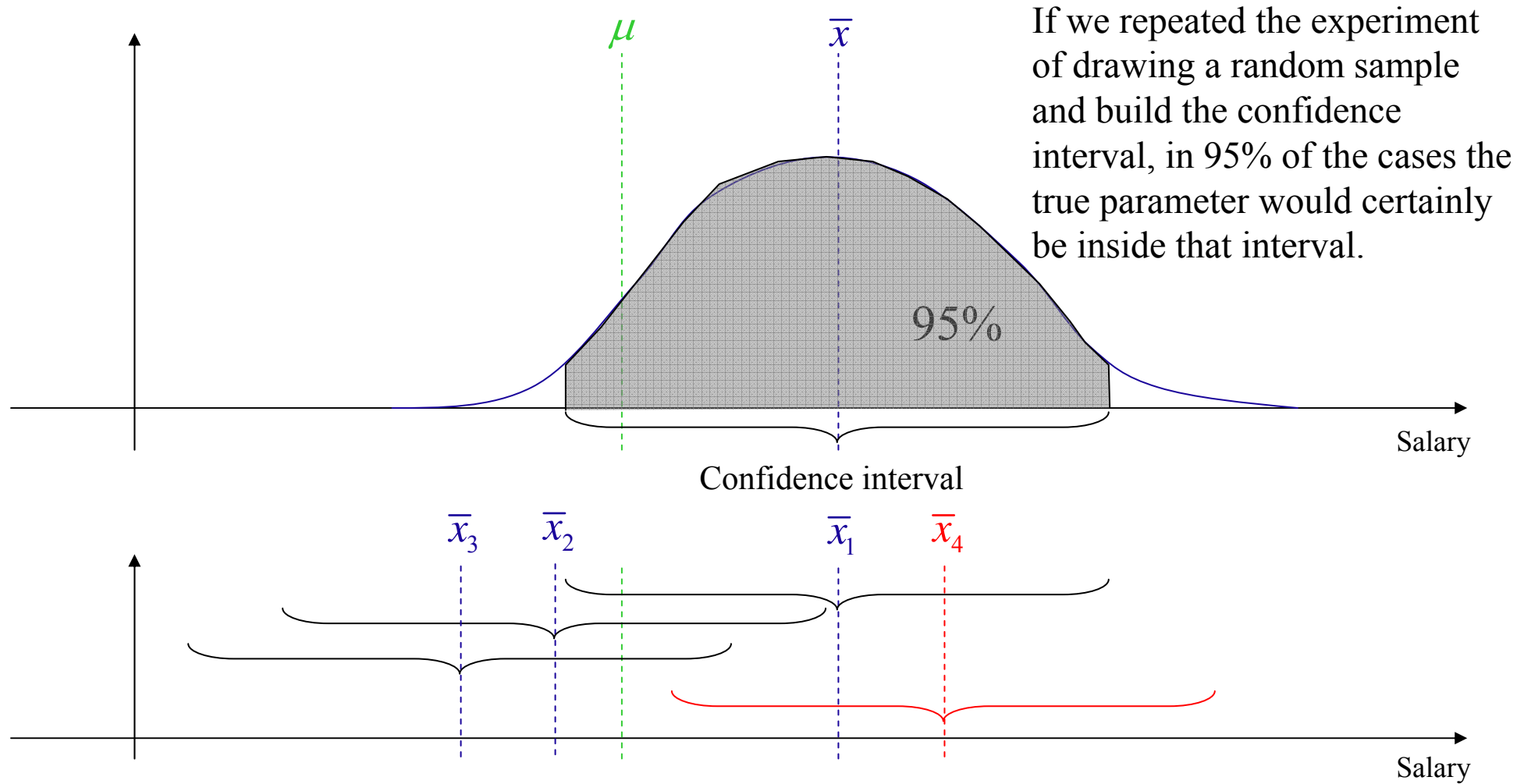
Asymptotically unbiased

$$\lim_{N \rightarrow \infty} \mu - E\{\bar{x}\} = 0$$



Sampling distribution: distribution of the statistic if all possible samples of size  $N$  were drawn from a given population

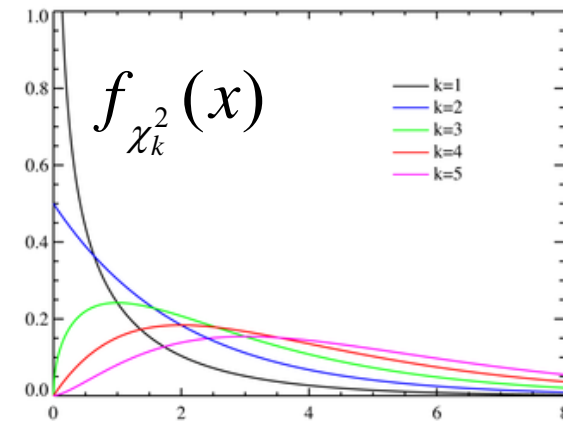
## 1.2 Statistic vs. Parameter. What is a sampling distribution?



## 1.2 Statistic vs. Parameter. What is a sampling distribution?

Sometimes the distribution of the statistic is known

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \frac{1}{N} \sum_{i=1}^N X_i \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$
$$\sum_{i=1}^N \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_N^2$$



Sometimes the distribution of the statistic is NOT known, but still the mean is well behaved

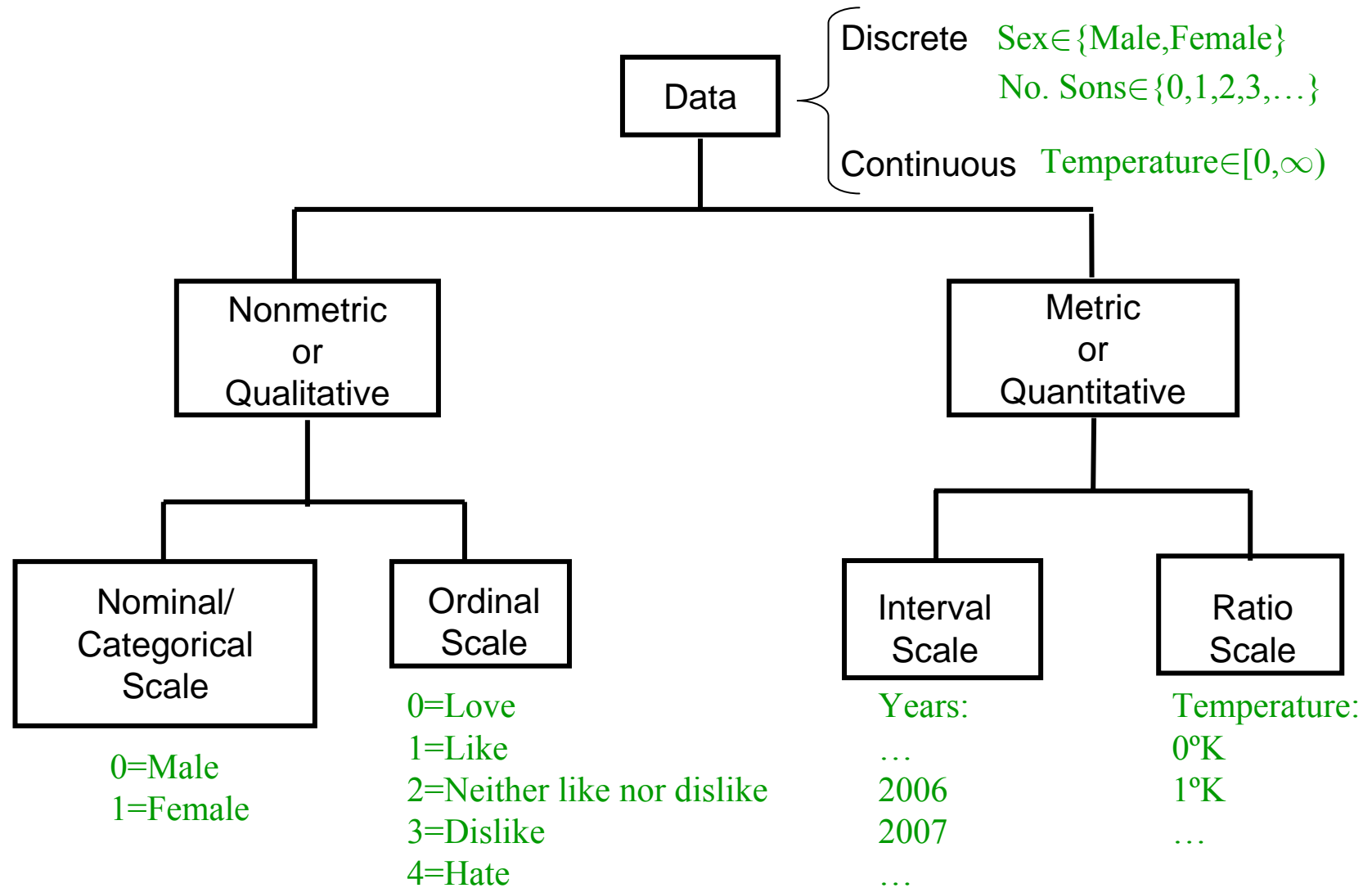
$$E\{X_i\} = \mu$$
$$Var\{X_i\} = \sigma^2$$
$$\Rightarrow \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N X_i \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

Central limit theorem!!

But:

- The sample must be truly random
- Averages based on samples whose size is more than 30 are reasonably Gaussian

## 1.3 Introduction: Types of variables





## 1.3 Introduction: Types of variables

### Coding of categorical variables

Hair Colour  
{Brown, Blond, Black, Red}  $\xrightarrow{\text{No order}}$   $(x_{\text{Brown}}, x_{\text{Blond}}, x_{\text{Black}}, x_{\text{Red}}) \in \{0,1\}^4$

---

Peter: Black

Peter: {0,0,1,0}

Molly: Blond

Molly: {0,1,0,0}

Charles: Brown

Charles: {1,0,0,0}

Company size  
{Small, Medium, Big}  $\xrightarrow{\text{Implicit order}}$   $x_{\text{size}} \in \{0,1,2\}$

---

Company A: Big

Company A: 2

Company B: Small

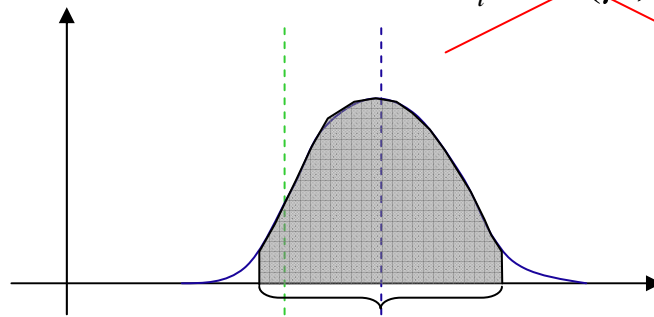
Company B: 0

Company C: Medium

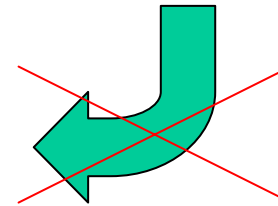
Company C: 1

## 1.4 Parametric vs. Non-parametric Statistics

### Parameter estimation



$$\cancel{X_i \sim N(\mu, \sigma^2)} \Rightarrow \frac{1}{N} \sum_{i=1}^N X_i \sim \cancel{N\left(\mu, \frac{\sigma^2}{N}\right)}$$



Solution: Resampling (bootstrap, jackknife, ...)

### Hypothesis testing

Cannot use statistical tests based on any assumption about the distribution of the underlying variable (t-test, F-tests,  $\chi^2$ -tests, ...)

Solution:

- discretize the data and use a test for categorical/ordinal data (non-parametric tests)
- use randomized tests

## 1.5 What to measure? Central tendency

During the last 6 months the rentability of your account has been:  
5%, 5%, 5%, -5%, -5%, -5%. Which is the average rentability of your account?

### Arithmetic mean

- (-) Very sensitive to large outliers, not too meaningful for certain distributions
- (+) Unique, **unbiased estimate of the population mean**,  
better suited for symmetric distributions

$$x_{AM}^* = \frac{1}{N} \sum_{i=1}^N x_i$$

$$x_{AM}^* = \frac{1}{6} (5 + 5 + 5 - 5 - 5 - 5) = 0\%$$

Property  $E\{x_{AM}^*\} = \mu$


$$x_{AM}^* = \frac{1}{6} (1.05 + 1.05 + 1.05 + 0.95 + 0.95 + 0.95) = 1 = 0\%$$

### Geometric mean

- (-) Very sensitive to outliers
- (+) Unique, used for the mean of **ratios and percent changes**,  
less sensitive to asymmetric distributions

$$x_{GM}^* = \sqrt[N]{\prod_{i=1}^N x_i} \Rightarrow \log x_{GM}^* = \frac{1}{N} \sum_{i=1}^N \log x_i$$

$$x_{GM}^* = \sqrt[6]{1.05 \cdot 1.05 \cdot 1.05 \cdot 0.95 \cdot 0.95 \cdot 0.95} = 0.9987 = -0.13\%$$

Which is right?   $\rightarrow 1050 \rightarrow 1102.5 \rightarrow 1157.6 \rightarrow 1099.7 \rightarrow 1044.8 \rightarrow 992.5$

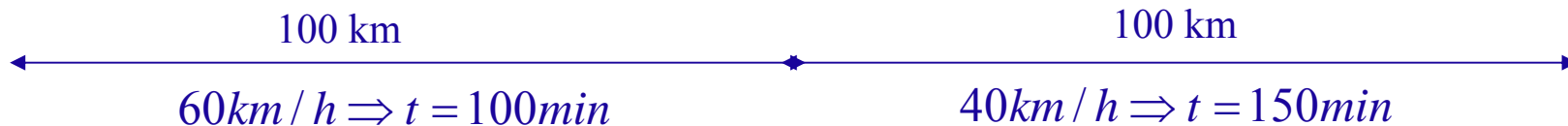
## 1.5 What to measure? Central tendency

### Harmonic mean

- (-) Very sensitive to small outliers
- (+) Usually used for the average of **rates**,  
less sensitive to large outliers

$$x_{HM}^* = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}} \Rightarrow \frac{1}{x_{HM}^*} = \frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}$$

A car travels 200km. The first 100 km at a speed of 60km/h, and the second 100 km at a speed of 40 km/h.



$$x_{HM}^* = \frac{1}{\frac{1}{2} \left( \frac{1}{60} + \frac{1}{40} \right)} = 48 \text{ km/h}$$

$$x_{AM}^* = \frac{1}{2} (60 + 40) = 50 \text{ km/h}$$



Which is the right average speed?

## 1.5 What to measure? Central tendency

Property: For positive numbers

$$x_{HM}^* \leq x_{GM}^* \leq x_{AM}^*$$

↑ Less affected by extreme values

↑ More affected by extreme small values

↑ More affected by extreme large values  
↑ Less affected by extreme small values

Generalization: Generalized mean  $x^* = \left( \frac{1}{N} \sum_{i=1}^N x_i^p \right)^{\frac{1}{p}}$

Minimum  $p = -\infty$

Harmonic mean  $p = -1$

Geometric mean  $p = 0$

Arithmetic mean  $p = 1$

Quadratic mean  $p = 2$

Maximum  $p = \infty$

## 1.5 What to measure? Robust central tendency

During the last 6 months the rentability of your account has been:

5%, 3%, 7%, -15%, 6%, 30%. Which is the average rentability of your account?

$$\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_{(3)} & x_{(2)} & x_{(5)} & x_{(1)} & x_{(4)} & x_{(6)} \end{array}$$

Trimmed mean, truncated mean, Windsor mean:

Remove p% of the extreme values on each side

$$x^* = \frac{1}{4} (x_{(2)} + x_{(3)} + x_{(4)} + x_{(5)}) = \frac{1}{4} (3 + 5 + 6 + 7) = 5.25\%$$

Median

Which is the central sorted value? (50% of the distribution is below that value) It is not unique

Any value between  $x_{(3)} = 5\%$  and  $x_{(4)} = 6\%$

Winsorized mean:

Substitute p% of the extreme values on each side

$$x^* = \frac{1}{6} (x_{(2)} + x_{(2)} + x_{(3)} + x_{(4)} + x_{(5)} + x_{(5)}) = \frac{1}{6} (3 + 3 + 5 + 6 + 7 + 7) = 5.1\hat{6}\%$$

## 1.5 What to measure? Robust central tendency

### M-estimators

Give different weight to different values

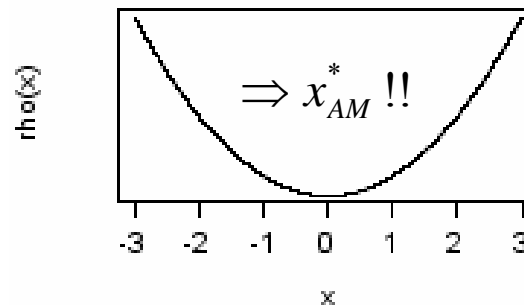
$$x^* = \arg \min_x \frac{1}{N} \sum_{i=1}^N \rho(x_i - x)$$

### R and L-estimators

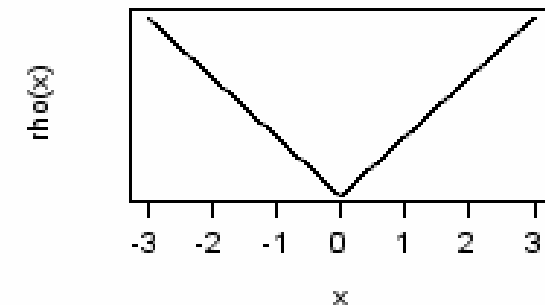
Now in disuse

The distribution of robust statistics is usually unknown and has to be estimated experimentally (e.g., bootstrap resampling)

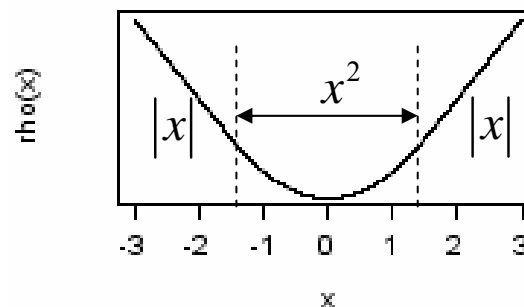
**Squared errors**



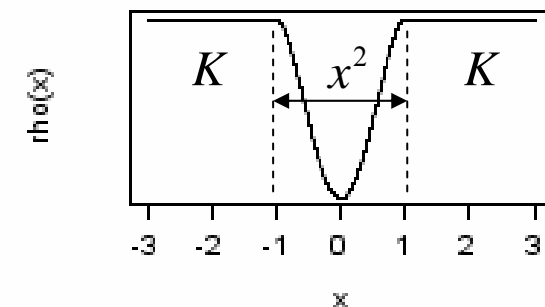
**Absolute errors**



**Winsorizing at 1.5**



**Biweight**





## 1.5 What to measure? Central tendency

### Mode:

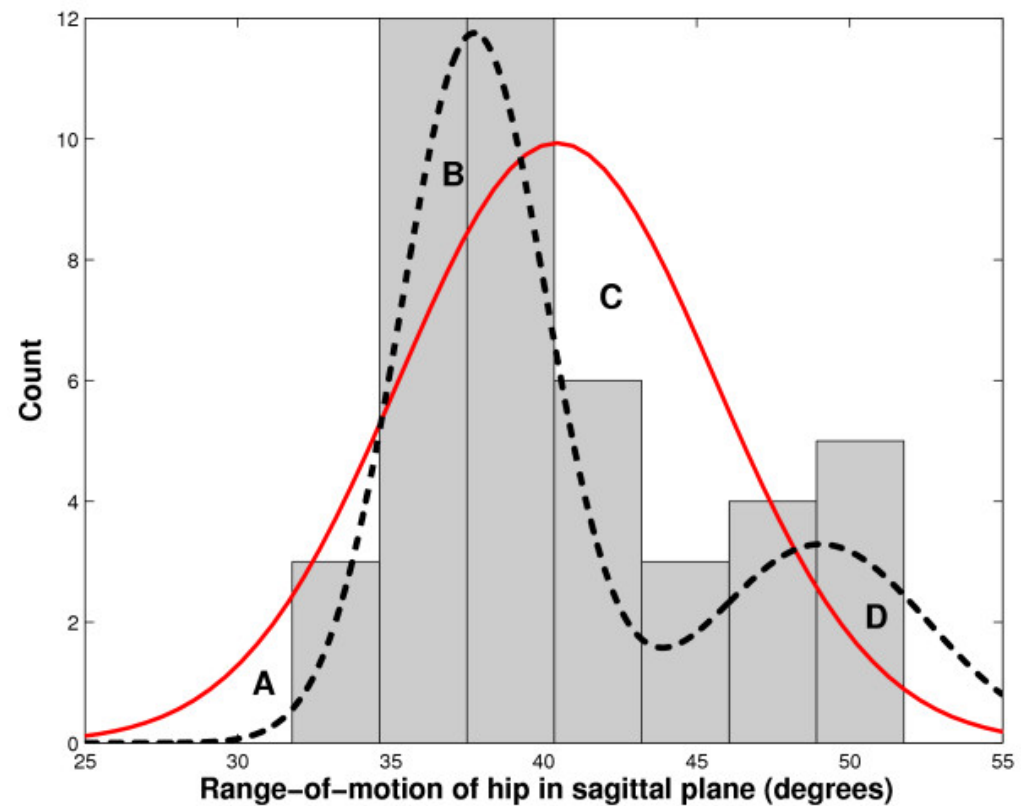
Most frequently occurring

(-) Not unique (multimodal)

(+) representative of the most “typical” result

If a variable is multimodal,  
most central measures fail!

$$x^* = \arg \max f_X(x)$$



## 1.5 What to measure? Central tendency



- What is the geometric mean of  $\{-2, -2, -2, -2\}$ ? Why is it so wrong?
- The arithmetic mean of  $\{2, 5, 15, 20, 30\}$  is 14.4, the geometric mean is 9.8, the harmonic mean is 5.9, the median is 15. Which is the right central value?

## 1.5 What to measure? Differences

An engineer tries to determine if a certain modification makes his motor to waste less power. He makes measurements of the power consumed with and without modifications (the motors tested are different in each set). The nominal consumption of the motors is 750W, but they have from factory an unknown standard deviation around 20W. He obtains the following data:

Unmodified motor (Watts): 741, 716, 753, 756, 727       $\bar{x} = 738.6$

Modified motor (Watts): 764, 764, 739, 747, 743       $\bar{y} = 751.4$

Not robust measure of unpaired differences       $d^* = \bar{y} - \bar{x}$

Robust measure of unpaired differences       $d^* = \text{median}\{y_i - x_j\}$

If the measures are paired (for instance, the motors are first measured, the modified and remeasured), then we should first compute the difference.

Difference: 23, 48, -14, -9, 16       $d^* = \bar{d}$

## 1.5 What to measure? Variability

During the last 6 months the rentability of an investment product has been:

-5%, 10%, 20%, -15%, 0%, 30% (geometric mean=5.59%)

The rentability of another one has been: 4%, 4%, 4%, 4%, 4%, 4%

Which investment is preferable for a month?

### Variance

(-) In squared units

(+) Very useful in analytical expressions

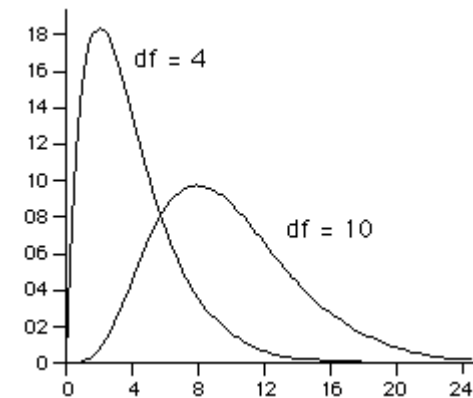
$$\sigma^2 = E\{(X - \mu)^2\}$$

$$s_N^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

$$E\{s_N^2\} = \frac{N-1}{N} \sigma^2$$

Substitution  
of the variance

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad E\{s^2\} = \sigma^2$$



$$X_i \sim N(\mu, \sigma^2) \Rightarrow (N-1) \frac{s^2}{\sigma^2} \sim \chi_{N-1}^2$$

$$s^2 \{0.95, 1.10, 1.20, 0.85, 1.00, 1.30\} = 0.0232$$

$$\text{Rentability} = 5.59 \pm 2.32\% \quad ?$$

# 1.5 What to measure? Variability

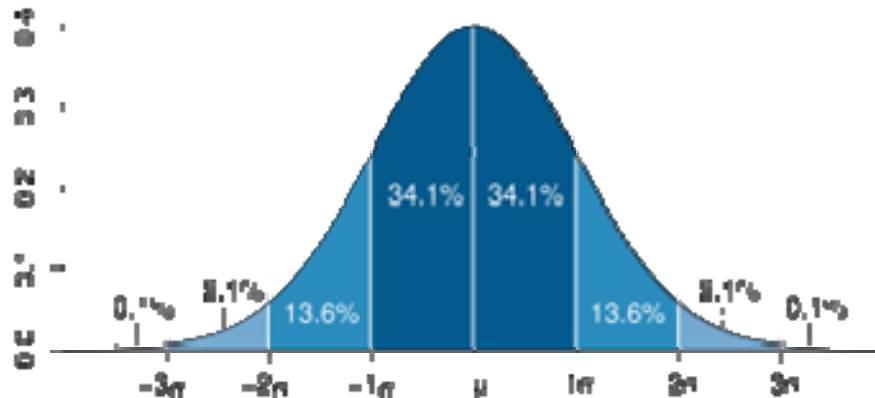
## Standard deviation

- (+) In natural units,
  - provides intuitive information about variability
  - Natural estimator of measurement precision
  - Natural estimator of range excursions

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

$$\text{Rentability} = 5.59 \pm \sqrt{0.0232} = 5.59 \pm 15.23\%$$

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \sqrt{N-1} \frac{s}{\sigma} \sim \chi_{N-1}$$



## Tchebychev's Inequality

$$\Pr\{\mu - K\sigma \leq X \leq \mu + K\sigma\} = 1 - \frac{1}{K^2}$$

- At least 50% of the values are within  $\sqrt{2}$  standard deviations from the mean.
- At least 75% of the values are within 2 standard deviations from the mean.
- At least 89% of the values are within 3 standard deviations from the mean.
- At least 94% of the values are within 4 standard deviations from the mean.
- At least 96% of the values are within 5 standard deviations from the mean.
- At least 97% of the values are within 6 standard deviations from the mean.
- At least 98% of the values are within 7 standard deviations from the mean.

For any distribution!!!

## 1.5 What to measure? Variability

### Percentiles

- (-) Difficult to handle in equations
- (+) Intuitive definition and meaning
- (+) Robust measure of variability

$$\Pr\{X \leq x^*\} = q$$

Someone has an IQ score of 115. Is he clever, very clever, or not clever at all?



### Deciles

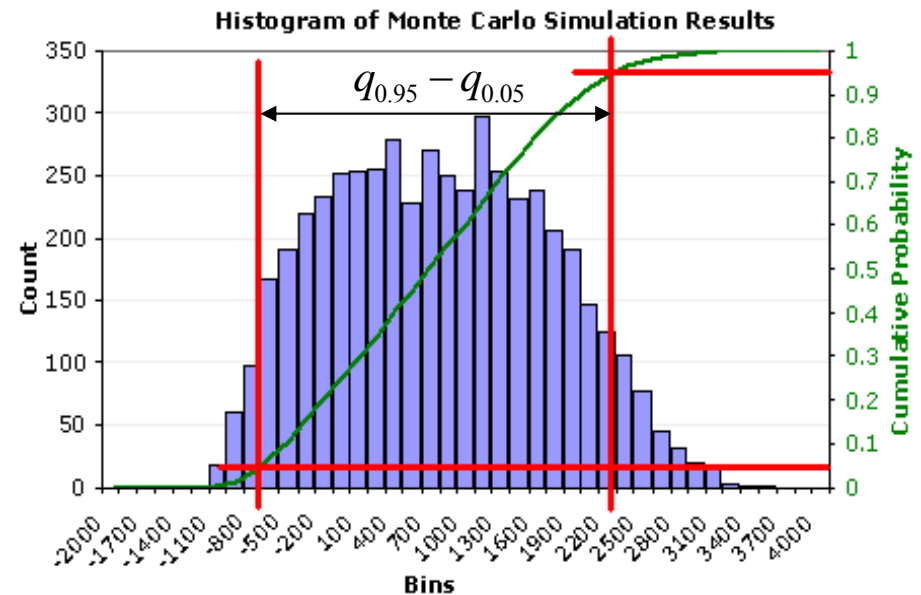
$q_{0.10}, q_{0.20}, q_{0.30}, q_{0.40}, q_{0.50}$   
 $q_{0.60}, q_{0.70}, q_{0.80}, q_{0.90}$

$$q_{0.90} - q_{0.10}$$

### Quartiles

$q_{0.25}, q_{0.50}, q_{0.75}$

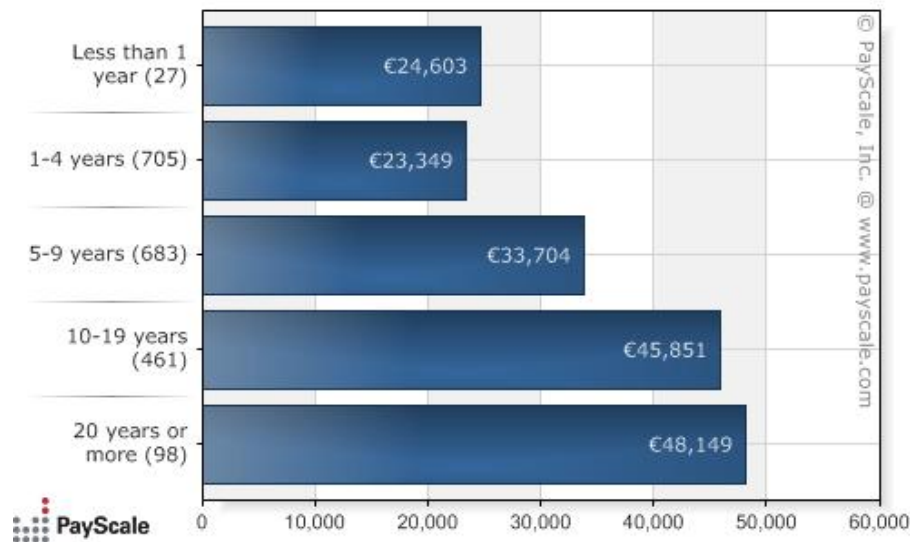
$$q_{0.75} - q_{0.25}$$



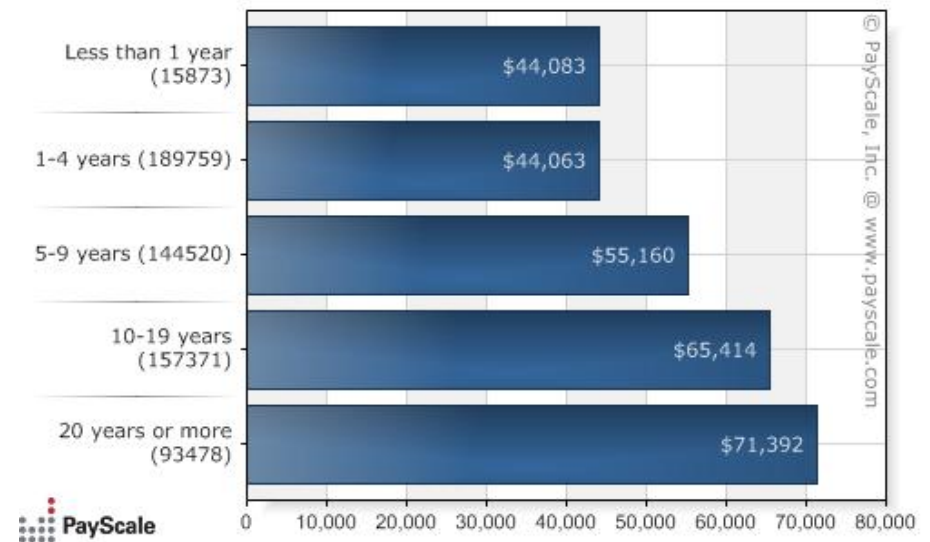
## 1.5 What to measure? Variability

### Coefficient of variation

Median salary in Spain by years of experience



Median salary in US by years of experience



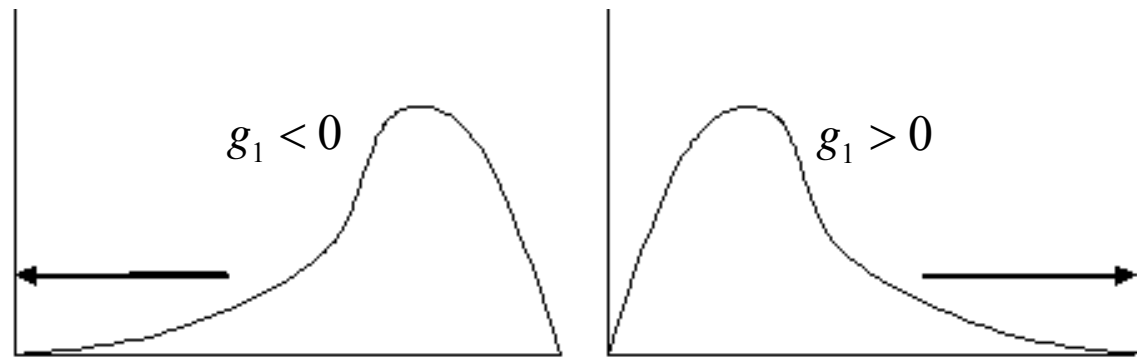
In which country you can have more progress along your career?





## 1.5 What to measure? Skewness

Skewness: Measure of the assymetry of a distribution



Negative Skew

Elongated tail at the **left**  
More data in the left tail than would be expected in a normal distribution

$$\mu < Med < Mode$$

Positive Skew

Elongated tail at the **right**  
More data in the right tail than would be expected in a normal distribution

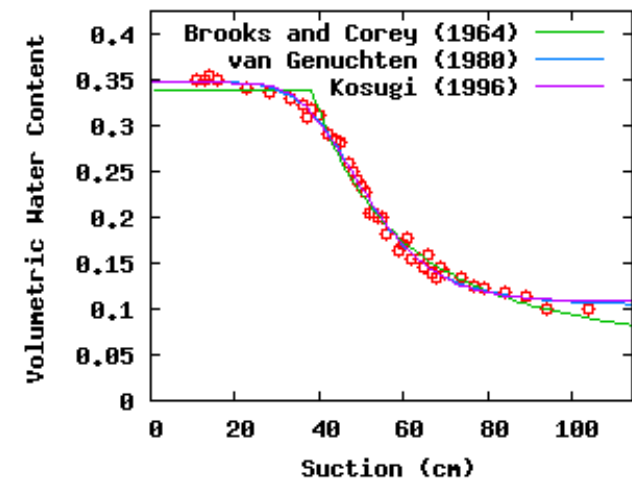
$$Mode > Med > \mu$$

$$\gamma_1 = \frac{E\{(X - \mu)^3\}}{\sigma^3}$$

Unbiased estimator

$$g_1 = \frac{m_3}{s^3}$$

$$m_3 = \frac{N \sum_{i=1}^N (x_i - \bar{x})^3}{(N-1)(N-2)}$$



The residuals of a fitting should not be skew! Otherwise, it would mean that positive errors are more likely than negative or viceversa. This is the rationale behind some goodness-of-fit tests.

## 1.5 What to measure? Correlation/Association

Is there any relationship between education, free-time and salary?

Person	Education (0-10)	Education	Free-time (hours/week)	Salary \$	Salary
A	10	High	10	70K	High
B	8	High	15	75K	High
C	5	Medium	27	40K	Medium
D	3	Low	30	20K	Low

Pearson's correlation coefficient

$$\rho = \frac{E\{(X - \mu_X)(Y - \mu_Y)\}}{\sigma_X \sigma_Y} \in [-1, 1]$$

$$r = \frac{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$$

*Salary* ↑ ⇒ *FreeTime* ↓

*Education* ↑ ⇒ *Salary* ↑

Correlation	Negative	Positive
Small	-0.3 to -0.1	0.1 to 0.3
Medium	-0.5 to -0.3	0.3 to 0.5
Large	-1.0 to -0.5	0.5 to 1.0

## 1.5 What to measure? Correlation/Association

Correlation between two ordinal variables? Kendall's tau

Is there any relationship between education and salary?

Person	Education	Salary \$
A	10	70K
B	8	75K
C	5	40K
D	3	20K



Person	Education	Salary \$
A	1st	2nd
B	2nd	1st
C	3rd	3rd
D	4th	4th

P=Concordant pairs

Person A

Education: (A>B) (A>C) (A>D) }  
 Salary: (A>C) (A>D) } 2

Person B

Education: (B>C) (B>D) }  
 Salary: (B>A) (B>C) (B>D) } 2

Person C

Education: (C>D) }  
 Salary: (C>D) } 1

Person D

Education: }  
 Salary: } 0

$$\tau = \frac{P}{\frac{N(N-1)}{2}}$$

$$\tau = \frac{2+2+1+0}{\frac{4(4-1)}{2}} = \frac{5}{6} = 0.83$$

## 1.5 What to measure? Correlation/Association

Correlation between two ordinal variables? Spearman's rho

Is there any relationship between education and salary?

Person	Education	Salary \$
A	10	70K
B	8	75K
C	5	40K
D	3	20K



Person	Education	Salary \$	di
A	1st	2nd	-1
B	2nd	1st	1
C	3rd	3rd	0
D	4th	4th	0

$$\rho = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}$$

$$\rho = 1 - \frac{6((-1)^2 + 1^2 + 0^2 + 0^2)}{4(4^2 - 1)} = 0.81$$

## 1.5 What to measure? Correlation/Association

Other correlation flavours:

- Correlation coefficient: How much of Y can I explain given X?
- 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  $X_2$ ?

## 1.6 Use and abuse of the normal distribution

Univariate  $X \sim N(\mu, \sigma^2) \Rightarrow f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

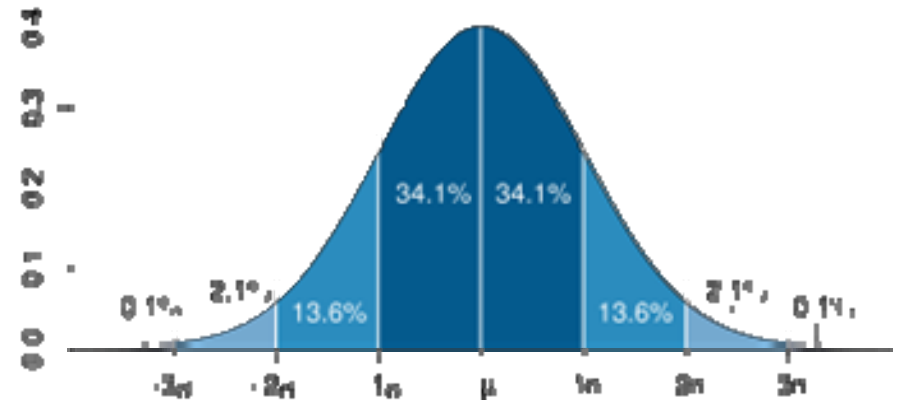
Multivariate  $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \Rightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$

↑  
Covariance matrix

Use: Normalization

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X - \mu}{\sigma} \sim N(0, 1)$$

↑  
Z-score



Compute the z-score of the IQ ( $\mu = 100, \sigma = 15$ ) of:  
 Napoleon Bonaparte (emperor): 145  
 Gary Kasparov (chess): 190

$$z_{\text{Napoleon}} = \frac{145 - 100}{15} = 3$$

$$z_{\text{Kasparov}} = \frac{190 - 100}{15} = 6$$

## 1.6 Use and abuse of the normal distribution

Use: Computation of probabilities **IF** the underlying variable is normally distributed

$$X \sim N(\mu, \sigma^2)$$

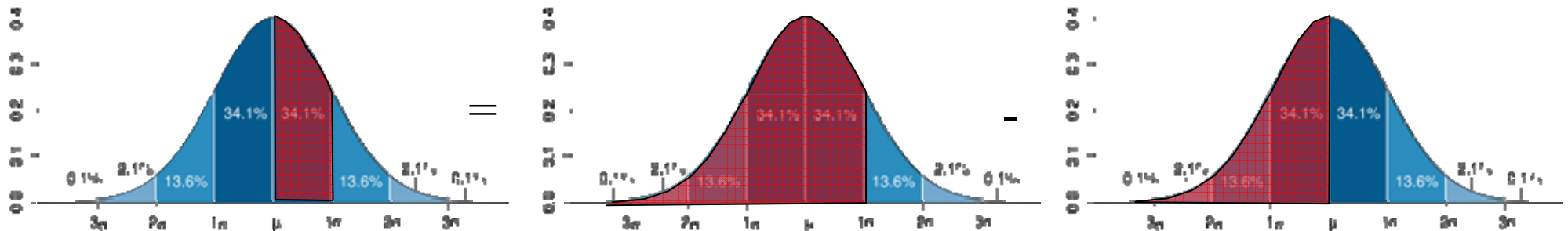
What is the probability of having an IQ between 100 and 115?

$$\Pr\{100 \leq IQ \leq 115\} = \int_{100}^{115} \frac{1}{\sqrt{2\pi}15^2} e^{-\frac{1}{2}\left(\frac{x-100}{15}\right)^2} dx = \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx - \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.341$$

Normalization



Use of tabulated values



## 1.6 Use and abuse of the normal distribution

Use: Computation of probabilities **IF** the underlying variable is normally distributed

$$X \sim N(\mu, \sigma^2)$$

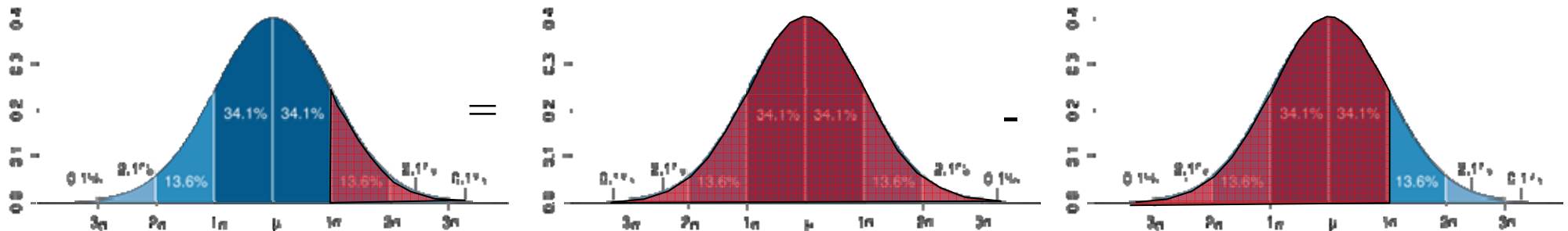
What is the probability of having an IQ larger than 115?

$$\Pr\{100 \leq IQ \leq 115\} = \int_{115}^{\infty} \frac{1}{\sqrt{2\pi}15^2} e^{-\frac{1}{2}\left(\frac{x-100}{15}\right)^2} dx = \int_1^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1 - \int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.159$$

Normalization



Use of tabulated values





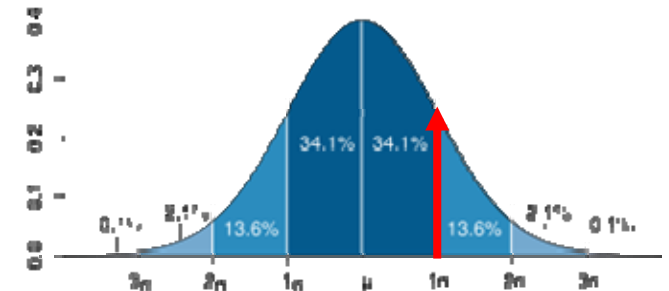
## 1.6 Use and abuse of the normal distribution

### Abuse: Computation of probabilities of a single point

What is the probability of having an IQ exactly equal to 115?

$$\Pr\{IQ = 115\} = 0$$

$$\text{Likelihood}\{IQ = 115\} = \text{Likelihood}\{z_{IQ} = 1\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}$$

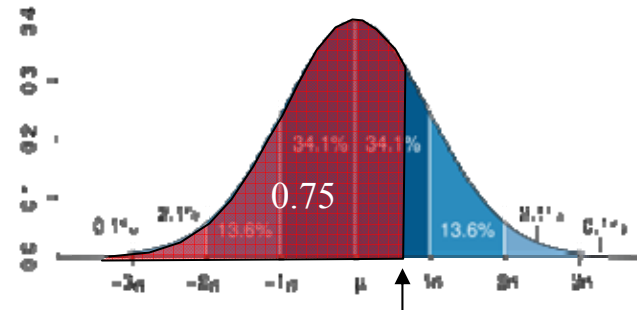


### Use: Computation of percentiles

Which is the IQ percentile of 75%?

$$\int_{-\infty}^{q_{0.75}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.75 \Rightarrow q_{0.75} = 0.6745$$

$$IQ_{0.75} = \mu_{IQ} + q_{0.75}\sigma_{IQ} = 100 + 0.6745 \cdot 15 = 110.1$$



0.6745

## 1.6 Use and abuse of the normal distribution

### Abuse: Assumption of normality

Many natural phenomena are normally distributed (thanks to the central limit theorem):

Error in measurements

Light intensity

Counting problems when the count number is very high (persons in the metro at peak hour)

Length of hair

The logarithm of weight, height, skin surface, ... of a person

But many others are not

The number of people entering a train station in a given minute is not normal, but the number of people entering all the train stations in the world at a given minute is normal.

Many distributions of mathematical operations are normal

$$X_i \sim N \longrightarrow aX_1 + bX_2; a + bX_1 \sim N$$

But many others are not

$$X_i \sim N \longrightarrow \frac{X_1}{X_2} \sim \text{Cauchy}; e^X \sim \text{LogNormal}; \frac{\sum X_i^2}{\sum X_j^2} \sim F - \text{Snedecor}$$

$$\sum X_i^2 \sim \chi^2; \sqrt{\sum X_i^2} \sim \chi; \sqrt{X_1^2 + X_2^2} \sim \text{Rayleigh}$$

Some distributions can be safely approximated by the normal distribution

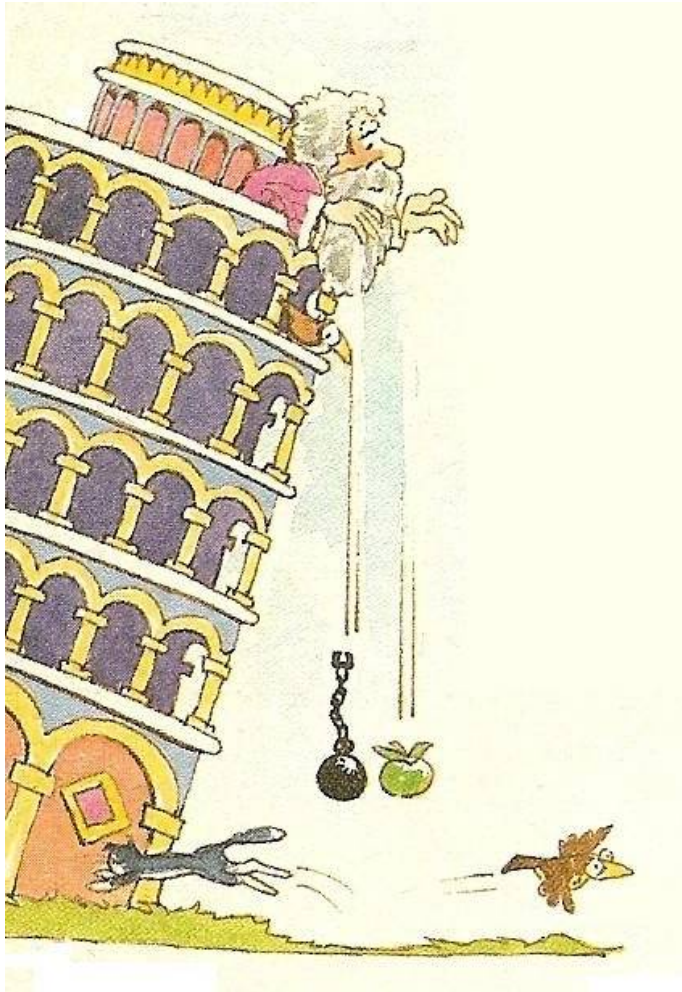
Binomial  $np > 10$  and  $np(1-p) > 10$ , Poisson  $\lambda > 1000$

## 1.6 Use and abuse of the normal distribution

Abuse: banknotes



## 1.6 Use and abuse of the normal distribution



$$t(\text{sec}) \sim N(t_0, \sigma^2)$$

$$t(\text{msec}) \sim N$$

$$h = \frac{1}{2}gt^2 \sim N$$



## 1.7 Is my data really independent?

### Independence is different from mutual exclusion

In general,

Mutual exclusion is when two results are impossible to happen at the same time.

Independence is when the probability of an event does not depend on the results that we have had previously.

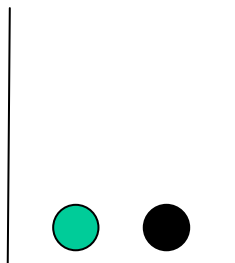
$$p(A \cap B) = p(A)p(B | A)$$

$$p(B | A) = 0$$

$$p(A \cap B) = 0$$

$$p(A \cap B) = p(A)p(B)$$

Knowing A does not give any information about the next event



### Example: Sampling with and without replacement

What is the probability of taking a black ball as second draw, if the first draw is green?



## 1.7 Is my data really independent?

### Sampling without replacement

In general samples are not independent except if the population is so large that it does not matter.

### Sampling with replacement

Samples may be independent. However, they may not be independent (see Example 1)

Examples: tossing a coin, rolling a dice

Random sample: all samples of the same size have equal probability of being selected

Example 1: Study about child removal after abuse, 30% of the members were related to each other because when a child is removed from a family, normally, the rest of his/her siblings are also removed. Answers for all the siblings are correlated.

Example 2: Study about watching violent scenes at the University. If someone encourages his roommate to take part in this study about violence, and the roommate accepts, he is already biased in his answers even if he is acting as control watching non-violent scenes.

Consequence: The sampling distributions are not what they are expected to be, and all the confidence intervals and hypothesis testing may be seriously compromised.

## 1.7 Is my data really independent?



- A newspaper makes a survey to see how many of its readers like playing videogames. The survey is announced in the paper version of the newspaper but it has to be filled on the web. After processing they publish that 66% of the newspaper readers like videogames. Is there anything wrong with this conclusion?

## 1.7 Is my data really independent?



“A blond woman with a ponytail snatched a purse from another woman. The thief fled in a yellow car driven by a black man with a beard and moustache”.

A woman matching this description was found. The prosecution assigned the following probabilities: blond hair ( $1/3$ ), ponytail ( $1/10$ ), yellow car ( $1/10$ ), black man with beard ( $1/10$ ), moustache ( $1/4$ ), interracial couple in car ( $1/1000$ ). The multiplication of all these probabilities was  $1/12M$  and the California Supreme Court convicted the woman in 1964.

Is there anything wrong with the reasoning?



# Course outline

2. How do I collect the data? Experimental design
  1. Methodology
  2. Design types
  3. Basics of experimental design
  4. Some designs: Randomized Complete Blocks, Balanced Incomplete Blocks, Latin squares, Graeco-latin squares, Full  $2^k$  factorial, Fractional  $2^{k-p}$  factorial
  5. What is a covariate?

## 2.1 Methodology

- Case-study method (or clinical method):
  - Observes a phenomenon in the real-world.
    - Example: Annotation of habits of cancer patients and tumor size
    - Advantages: Relevance
    - Disadvantages: There is no control, quantification is difficult, statistical procedures are not easy to apply, lost of precision.
- Experimental method (or scientific method):
  - Conduction of controlled experiments.
    - Example: Dosis of a certain drug administered and tumor size
    - Advantages: Precision, sets of independent (controlled dosis) and dependent (resulting tumor size) variables, statistical procedures are well suited
    - Disadvantages: Lost of relevance, artificial setting
- Correlational method (or survey method):
  - Conduction of surveys on randomly chosen individuals
    - Example: Survey on the habits of a random sampling among cancer patients
    - Advantages: cheap and easy, trade-off between the previous two approaches.
    - Disadvantages: lost of control, the sample fairness is crucial, poor survey questions, participants may lie to look better, or have mistaken memories.

## 2.2 Design types

- Control of variables: A design must control as many variables as possible otherwise external variables may invalidate the study.
- Control groups: a control group must be included so that we can know the changes not due to the treatment.
- Pre-experimental designs: Usually fail on both controls
- Quasi-experimental designs: Usually fail on the control of variables
- True experimental designs: Succeeds in both controls

## 2.2 Design types: Control of variables

- 20 clinically depressed patients are given a pretest to assess their level of depression.
- During 6 months they are given an antidepressant (treatment)
- After 6 months, a posttest show a significant decrease of their level of depression



- Can we conclude that the treatment was effective?

## 2.2 Design types: Control Groups

- The blood pressure of 20 people is taken as a pretest.
- Then they are exposed to 30 minutes of hard rock (treatment)
- Finally, as a posttest we measure the blood pressure again, and we find that there is an increase of blood pressure.



- Can we conclude that hard rock increases blood pressure?

## 2.2 Design types: Random selection

- 20 depressed patients from Hospital A are selected as study group, and 20 depressed patients from Hospital B will be selected as control group. They are assumed to be equally depressed and a pretest is not considered necessary.
- The study group is given an antidepressant for 6 months, and the control group is given a placebo (treatment).
- Finally, as a posttest we measure level of depression of patients from Hospital A with respect to that of patients from Hospital B, finding that the depression level in Hospital A is lower than in Hospital B.
- Can we conclude that the antidepressant was effective?

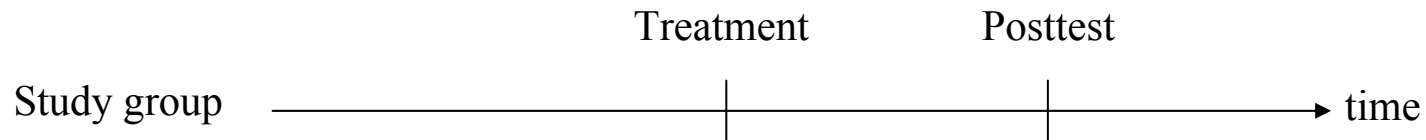


## 2.2 Design types: Group stability

- 20 depressed patients from Hospital A and Hospital B are randomly split in a control and a study group. They are given a pretest to assess their level of depression.
- The study group is given an antidepressant for 6 months, and the control group is given a placebo (treatment). Unfortunately, some of the patients dropped the study from the treatment group.
- Finally, as a posttest we measure level of depression obtaining that patients of the study group is lower than the depression of the control group.
- Can we conclude that the antidepressant was effective?



## 2.2 Design types: Pre-experimental design



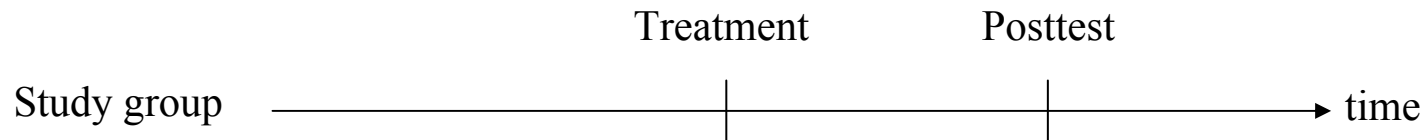
- 20 arthritis patients undergo a novel surgical technique (treatment)
- After a year, a posttest shows that they have minimal symptoms
- Can we conclude that the improvement is due to the new surgery?





## 2.2 Design types: Pre-experimental design

### One-shot case study

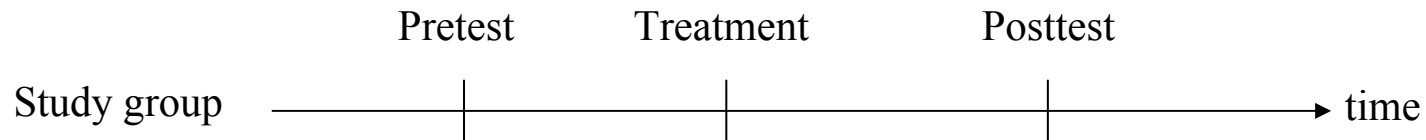


- 20 arthritis patients undergo a novel surgical technique (treatment)
- After a year, a posttest shows that they have minimal symptoms
- Can we conclude that the minimal symptoms are due to the new surgery?



## 2.2 Design types: Pre-experimental design

### One-group pretest-posttest case study

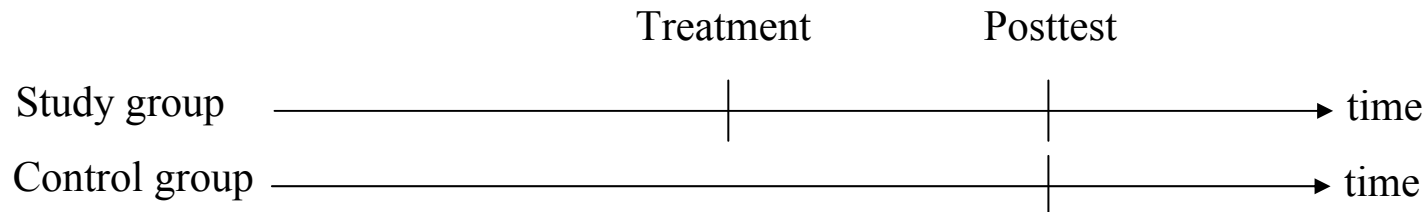


- 20 arthritis patients are given a pretest to evaluate their status
- Then, they undergo a novel surgical technique (treatment)
- After a year, a posttest shows that they have minimal symptoms
- Can we conclude that the improvement is due to the new surgery?



## 2.2 Design types: Pre-experimental design

### Nonequivalent posttest-only design

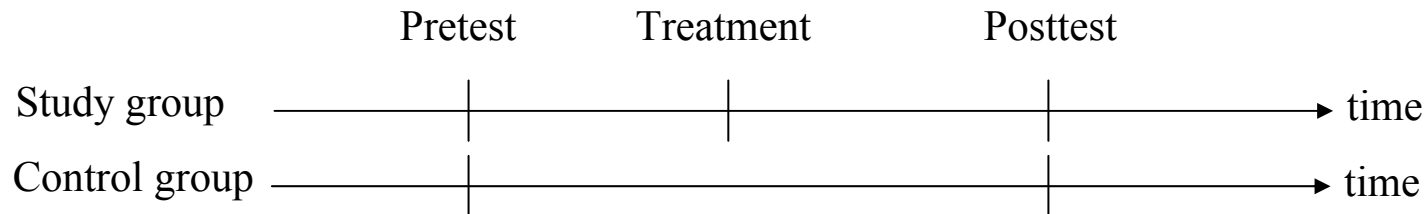


- 20 arthritis patients of Dr.A undergo a novel surgical technique (treatment)
- After a year, a posttest shows that they have less symptoms than those of Dr. B
- Can we conclude that the difference is due to the new surgery?



## 2.2 Design types: Quasi-experimental design

### Nonequivalent control group design

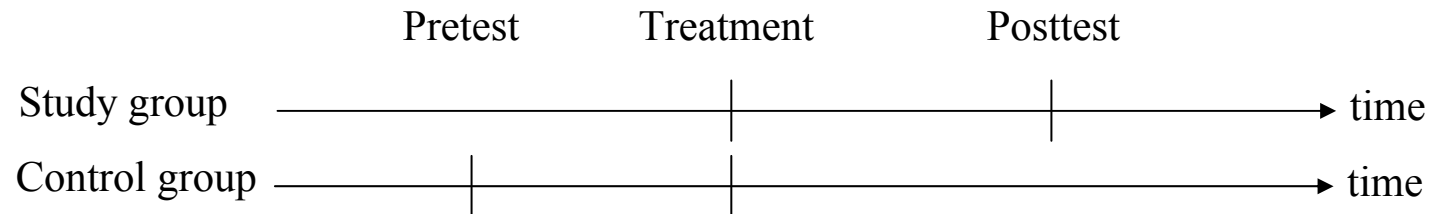


- 20 arthritis patients of Dr.A and 20 arthritis patients of Dr.B are subjected to a pretest.
- The patients of Dr.A undergo a novel surgical technique (treatment)
- After a year, a posttest shows that they have less symptoms than those of Dr. B
- Can we conclude that the difference is due to the new surgery?



## 2.2 Design types: Quasi-experimental design

### Separate sample pretest-posttest design

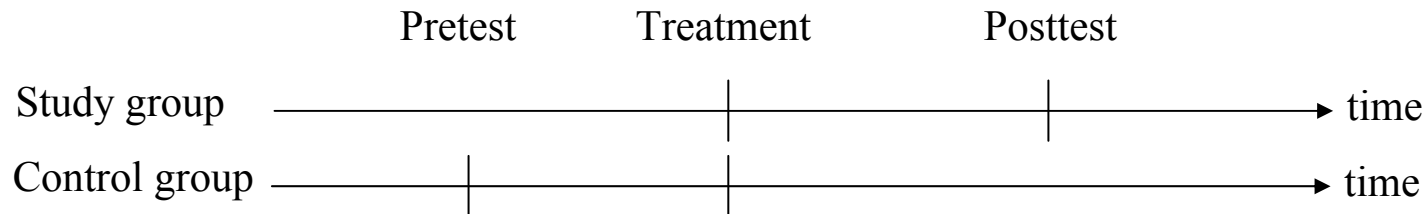


- A control group undergoes a pretest to know the audience attitude towards a TV program
- Then a publicity campaign is performed (treatment affecting to both groups).
- Finally a posttest on a study group reveals that the attitude towards the program has improved.
- Can we conclude that the difference is due to the campaign?



## 2.2 Design types: Quasi-experimental design

### Separate sample pretest-posttest design

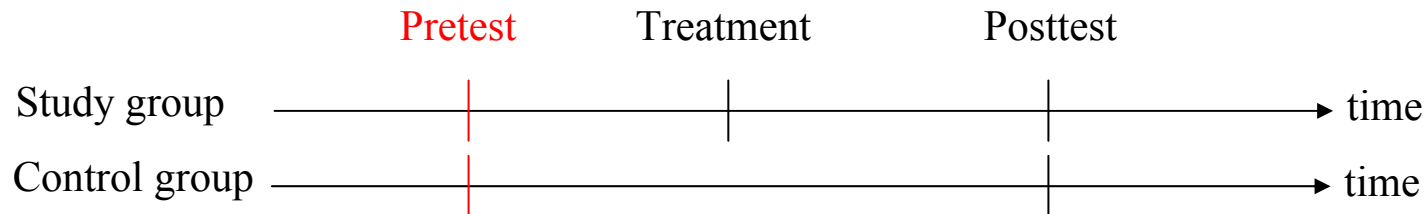


- A control group undergoes a pretest to know the audience attitude towards a TV program
- Then a publicity campaign is performed (treatment affecting to both groups).
- Finally a posttest on a study group reveals that the attitude towards the program has improved.
- Can we conclude that the difference is due to the campaign?



## 2.2 Design types: True-experimental design

### Dependent samples design (randomized-blocks)

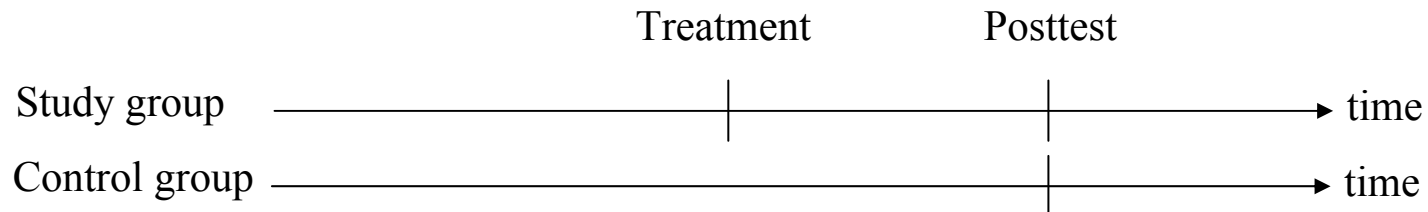


- 50 arthritis patients randomly chosen from Dr.A and Dr. B undergo a novel surgical technique (treatment)
- After a year, a posttest shows that they have less symptoms than those of the control group (also randomly chosen from Dr. A and Dr. B?)
- Can we conclude that the difference is due to the new surgery?



## 2.2 Design types: True-experimental design

### Dependent samples design (randomized-blocks)



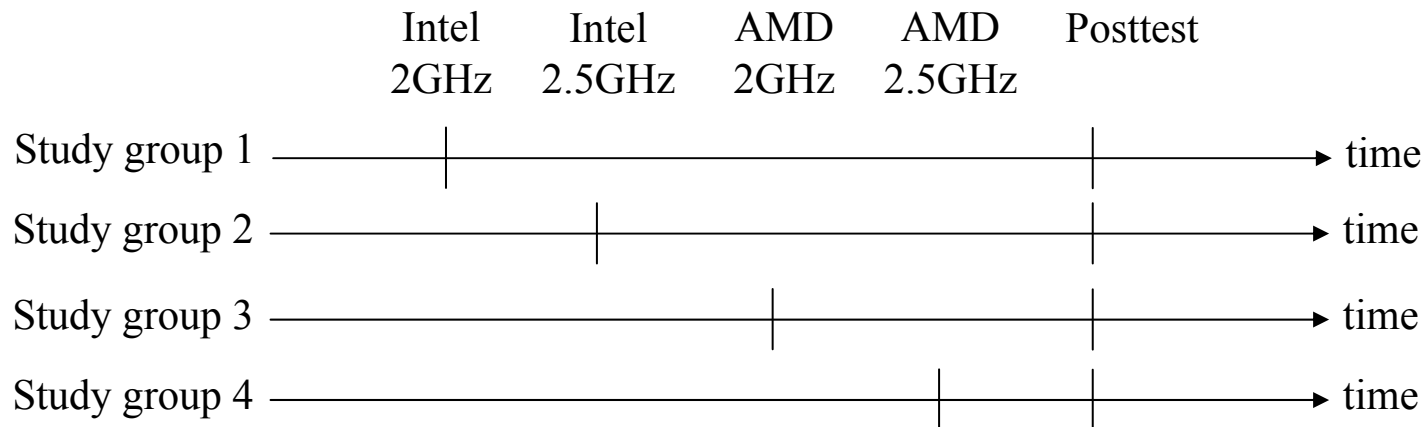
- 50 arthritis patients are given a placebo for 6 months after which they are evaluated.
- The same patients are given a drug against arthritis for another 6 months after which they are reevaluated.
- Can we conclude that the difference is due to the new drug?





## 2.2 Design types: True-experimental design

Factorial design: Effects of two or more levels on two or more variables

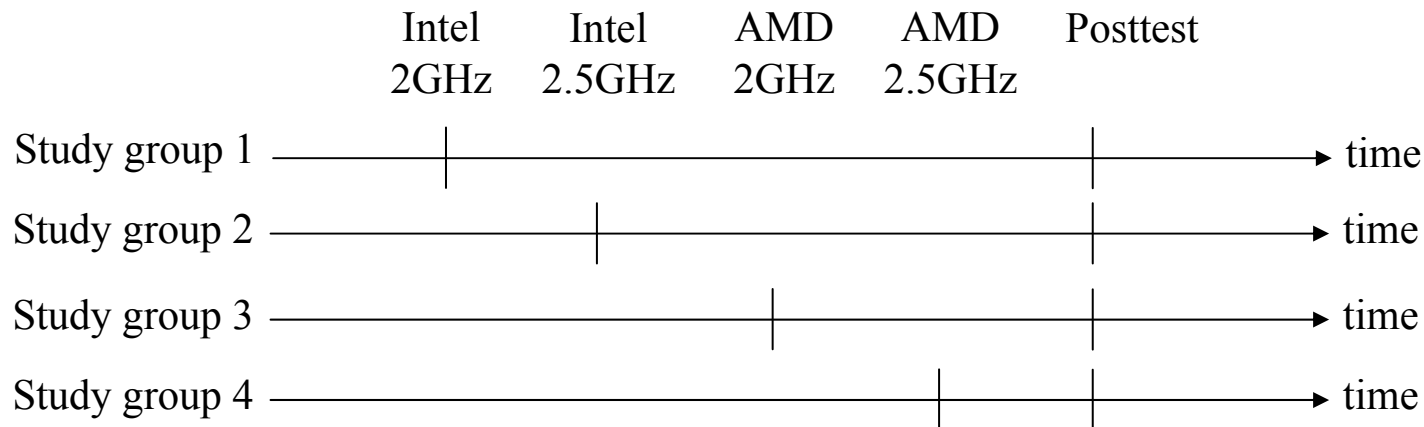


- What is the consumption of an Intel and AMD microprocessor running at 2 and 2.5GHz?



## 2.2 Design types: True-experimental design

Factorial design: Effects of two or more levels on two or more variables



- What is the consumption of an Intel and AMD microprocessor running at 2 and 2.5GHz?

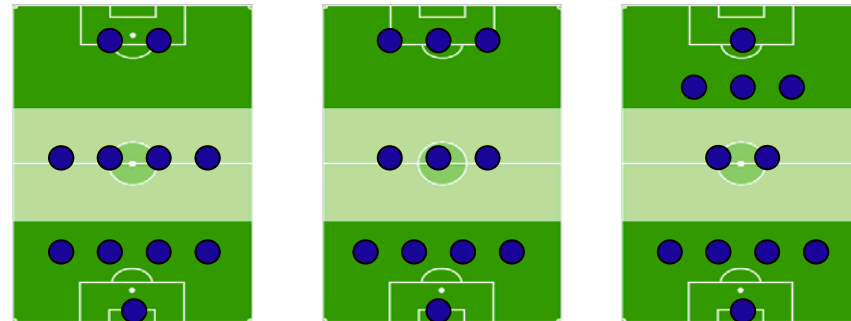


## 2.3 Basics of experimental design



We are the coaches of Real Madrid wanting to maximize our numbers of goals. We think that the following variables are important:

- Playing scheme: 4-4-2, 4-3-3, 4-2-3-1



3

- Does Raúl play or not?



2

- How many days do we train along the week? 3, 4, 5 or 6
- How many sessions do we have per day? 1 or 2
- Do we have gym sessions? Yes or no.
- Number of trials per combination: 3

$$\begin{array}{r}
 4 \\
 2 \\
 2 \\
 \times 3 \\
 \hline
 288
 \end{array}$$

## 2.3 Basics of experimental design



Best-guess approach: start with one configuration, change “randomly” any of the variables and see what happens (there can be a lot of thinking and *a priori* knowledge in this change)

Configuration	Scores in 3 matches	Avg
(4-4-2, Raúl does not play, 3 training days, 1 session, no gym session) →	1, 0, 1	2/3
(4-4-2, Raúl does not play, 5 training days, 1 session, no gym session) →	3, 0, 0	3/3
(4-3-3, Raúl does not play, 5 training days, 1 session, no gym session) →	2, 0, 3	5/3
(4-3-3, Raúl does not play, 5 training days, 2 sessions, no gym session) →	2, 1, 3	6/3
(4-3-3, Raúl plays, 5 training days, 2 sessions, no gym session) →	4, 3, 5	12/3
...		

### Drawbacks:

- The coach has to “guess” which variable to change
- We may stop in a reasonable solution, but not the best solution

## 2.3 Basics of experimental design



One-factor-at-a-time approach: start with one baseline configuration, change systematically all variables one at a time and keep for each one the best level.

Configuration	Scores in 3 matches	Avg
(4-4-2, Raúl does not play, 3 training days, 1 session, no gym session) →	1, 0, 1	2/3
(4-3-3, Raúl does not play, 3 training days, 1 session, no gym session) →	2, 0, 2	4/3
(4-2-3-1, Raúl does not play, 3 training days, 1 session, no gym session) →	1, 2, 0	3/3
(4-3-3, Raúl plays, 3 training days, 1 session, no gym session) →	3, 2, 2	7/3
(4-3-3, Raúl plays, 4 training days, 1 session, no gym session) →	3, 2, 4	9/3

...

### Drawbacks:

- Interactions are lost:
  - What if the combination of 4-4-2 with 4 training days have an explosive synergy? (2 factors)  
I will never try this combination with this strategy
  - What if the combination of 4-2-3-1, without Raúl, and 6 training days is even more explosive? (3 factors)

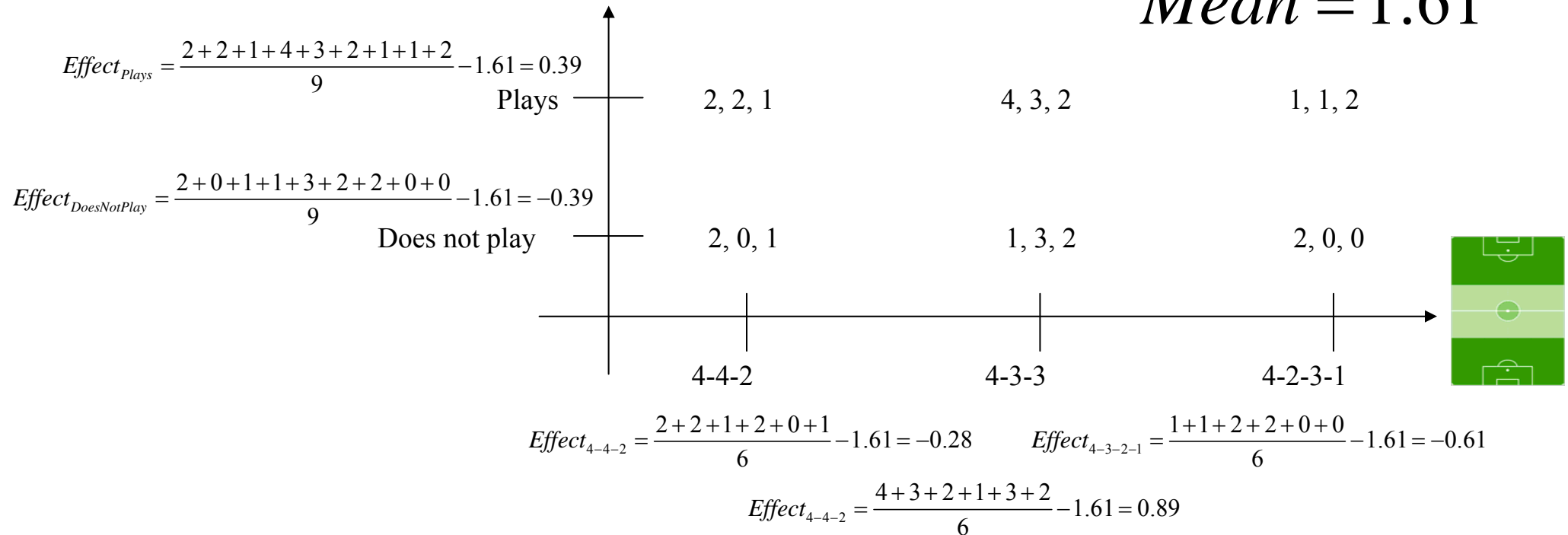
## 2.3 Basics of experimental design



Factorial approach: factors are varied together



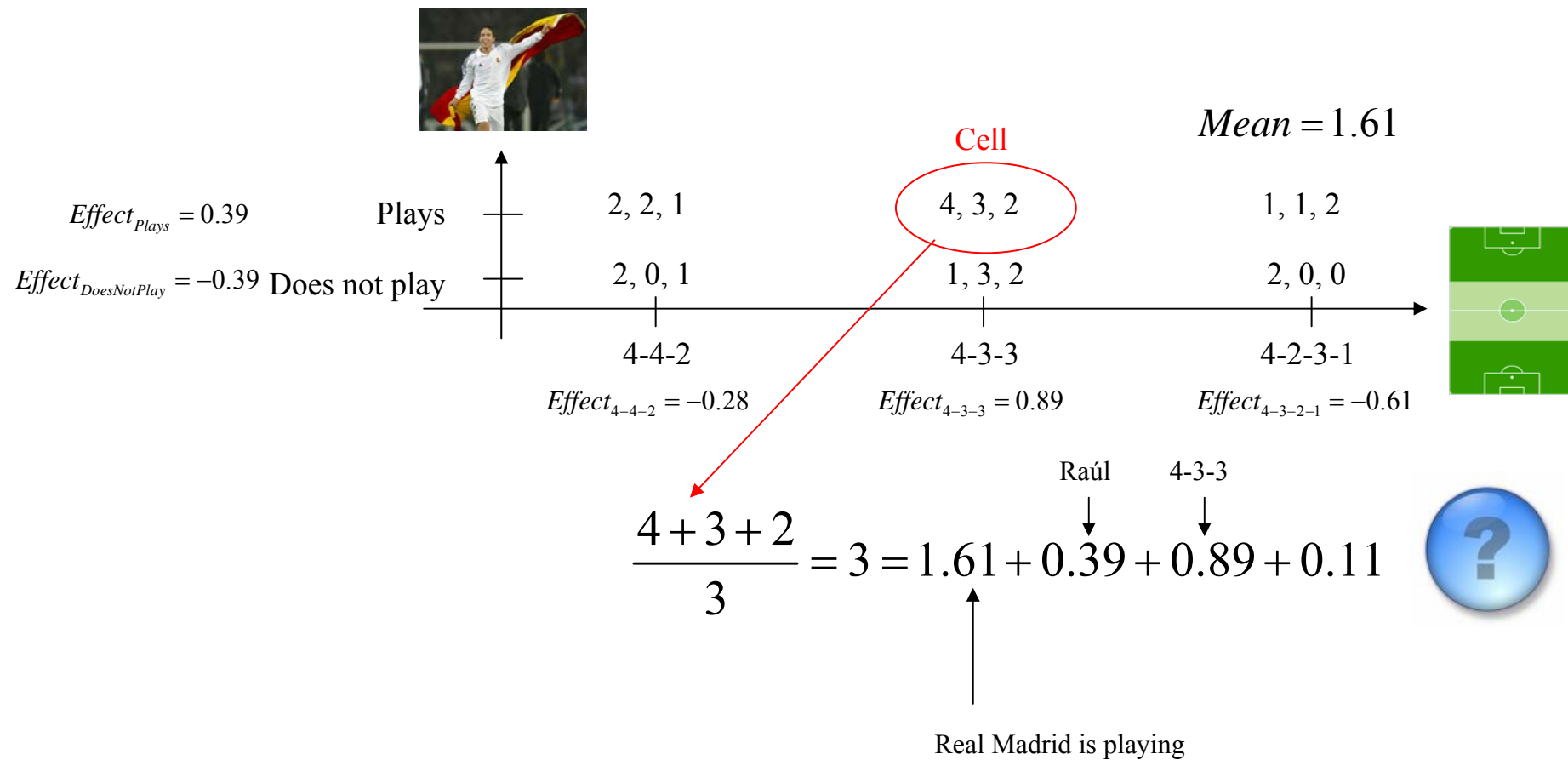
*Mean* = 1.61



## 2.3 Basics of experimental design



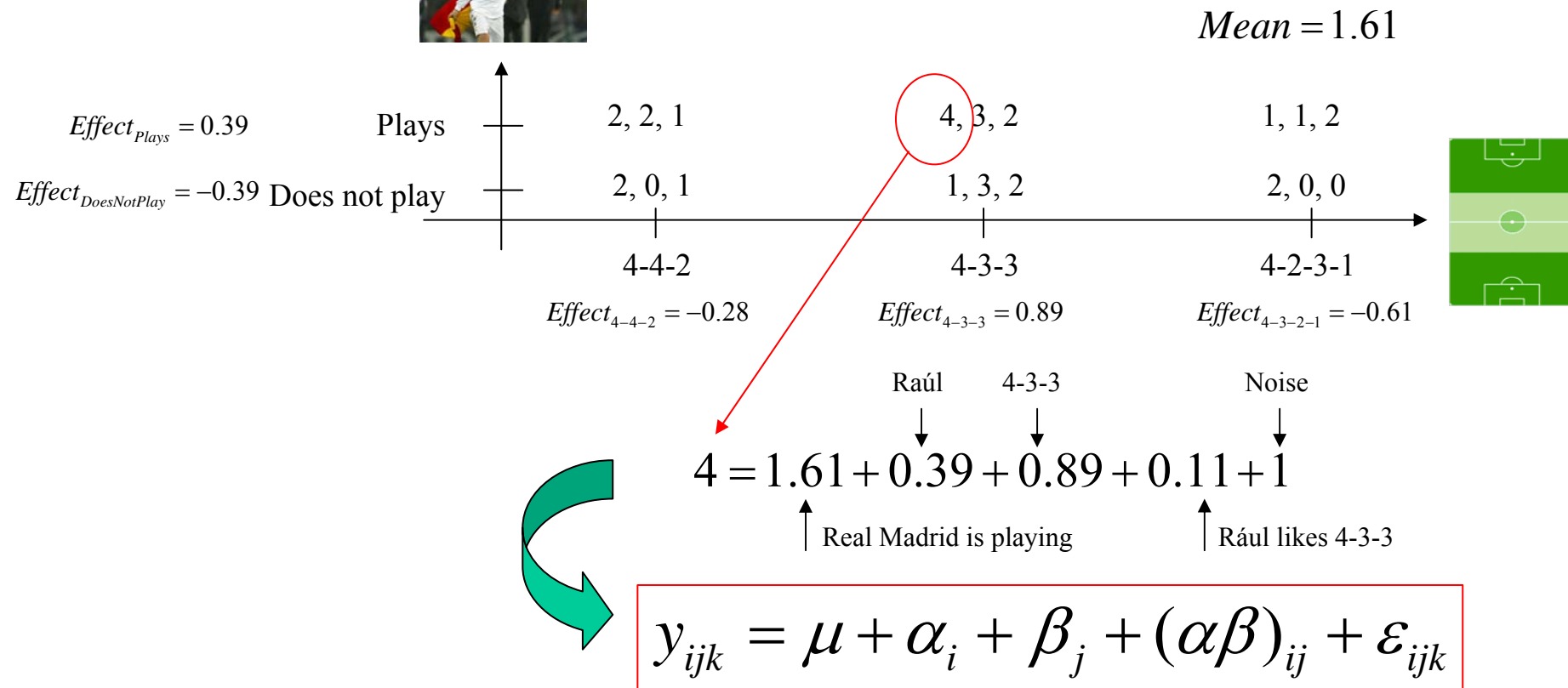
Factorial approach: factors are varied together



## 2.3 Basics of experimental design



### Analysis of Variance: ANOVA





## 2.3 Basics of experimental design



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
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

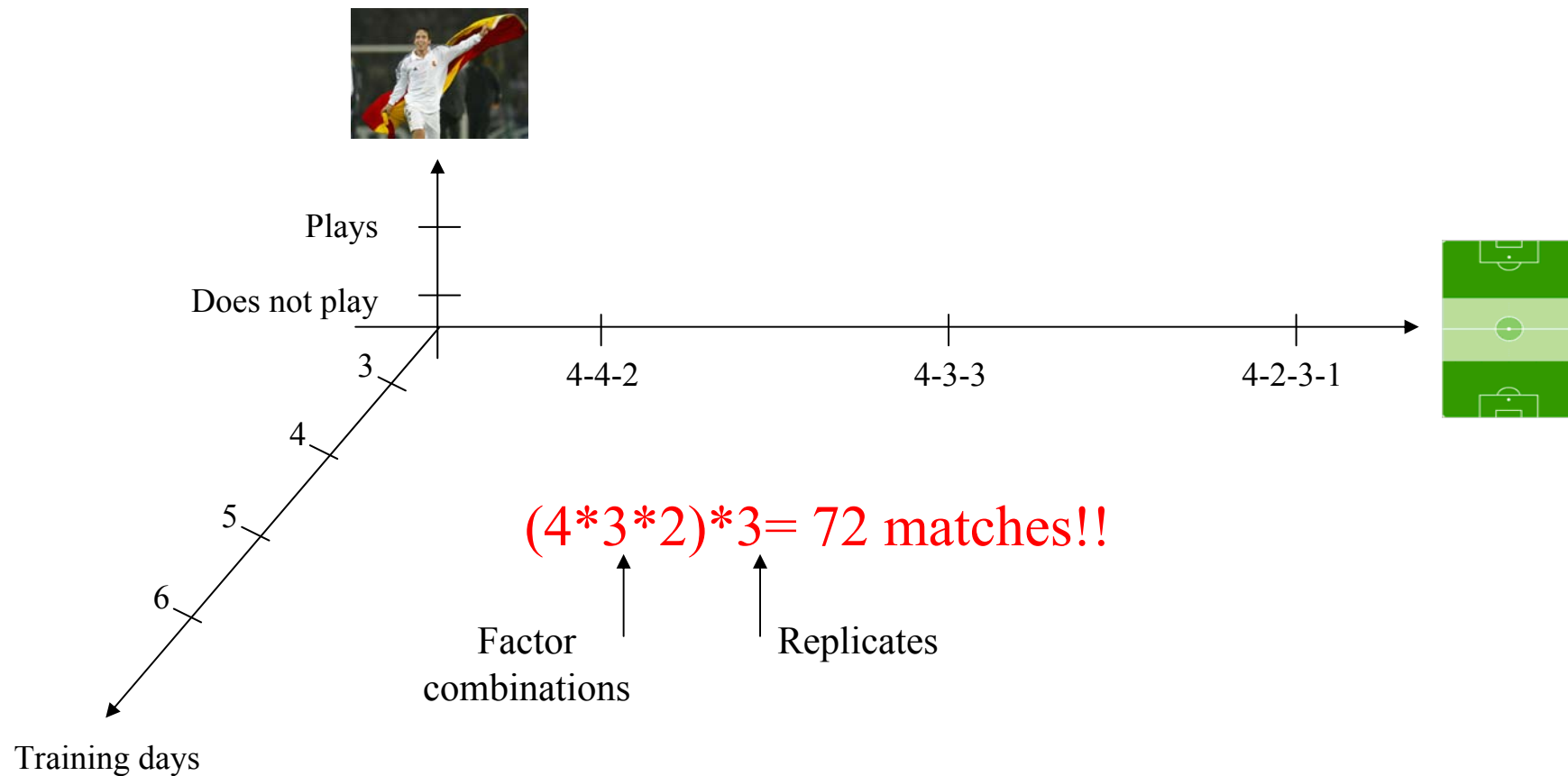
a=number of different levels in treatment A

b=number of different levels in treatment B

## 2.3 Basics of experimental design



Factorial approach: factors are varied together



## 2.3 Basics of experimental design: Principles of experimental design

Replication: Repetition of the basic experiment (3 matches for each combination)

- o It permits to estimate the experimental error
- o The experimental error allows us to assess whether an effect is significant
- o More replicates allow a better determination of the effect size (sampling distrib.)
- o Replication is different from repeated measurements (measuring several times the height of the same person)

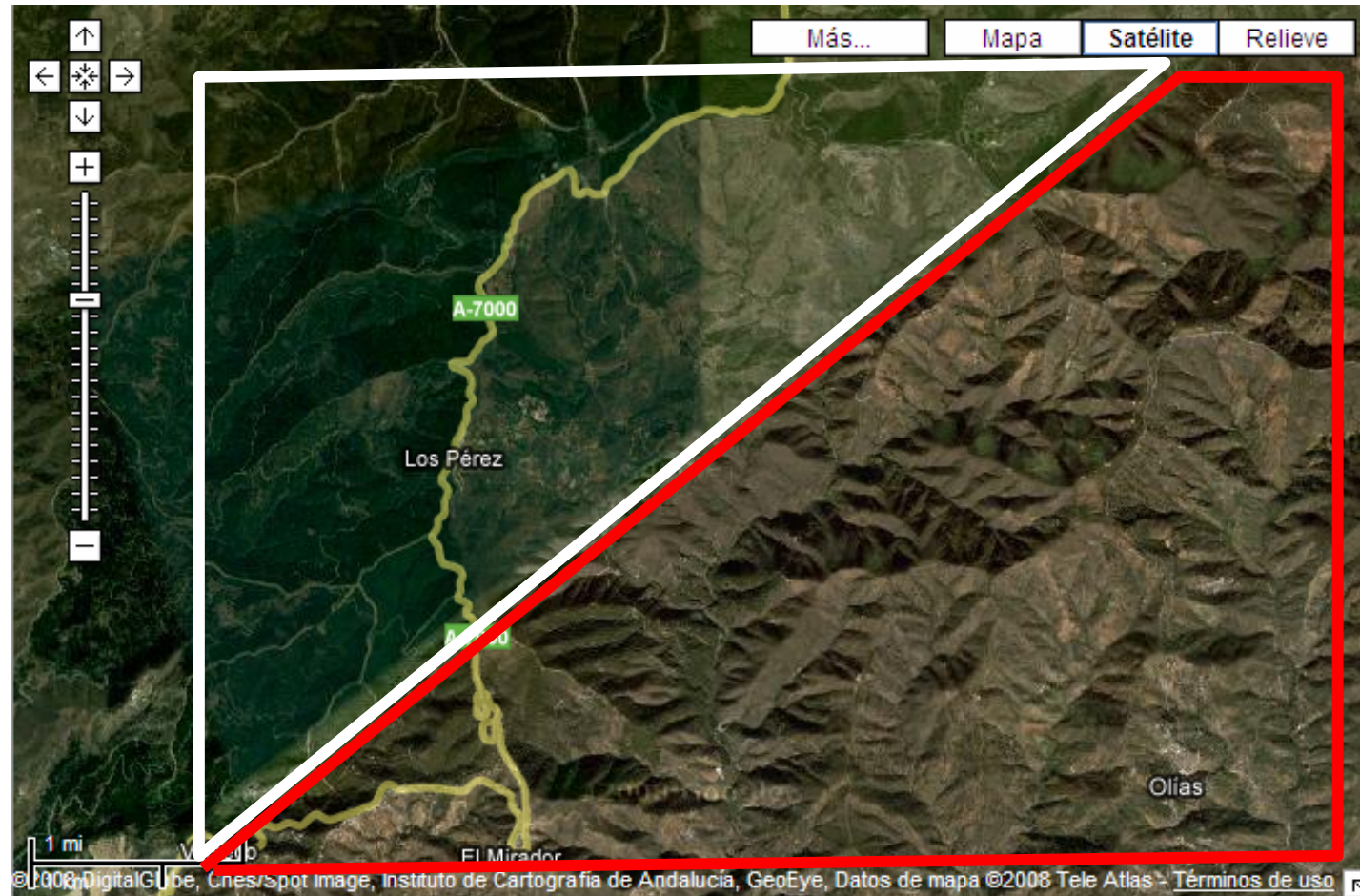
• Randomization: experiments must really be independent of each other, no bias should be introduced

• Blocking: removal of variability due to factors in which we are not interested in (*nuisance factors*)

- o For instance, in a chemical experiment we may need two batches of raw material. The two batches may come from two different suppliers and may differ. However, we are not interested in the differences due to the supplier.
- o Nuisance factor is unknown and uncontrollable → Randomization
- o Nuisance factor is known but uncontrollable → Analysis of covariance
- o Nuisance factor is known and controllable → Blocked designs

## 2.4 Some designs: Randomized complete blocks

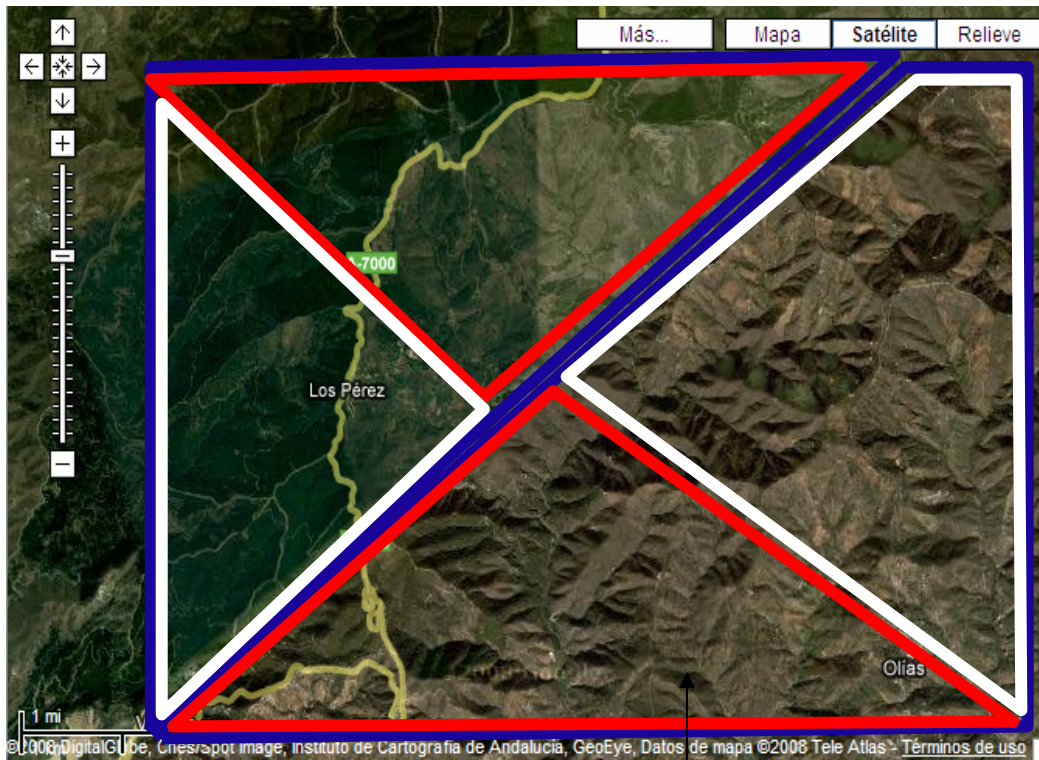
What do you think of the following land division for trying two different fertilizers?





## 2.4 Some designs: Randomized complete blocks

Randomized complete blocks: each block (of nuisance factors) is tested with all treatments and only once



$$N=2*2=tb$$

	DOF
Treatments	$1=t-1$
Blocks	$1=b-1$
Residual	$1=N-1-(t-1)-(b-1)$ $=(t-1)(b-1)$
Total	$3=N-1$

### Example with 4 fertilizers

Block I	A B C D	Note that we used only 4 out of the 4! possible arrangements
Block II	D A B C	
Block III	B D C A	
Block IV	C A B D	

Random assignment of which fertilizer goes in which subplot!!

## 2.4 Some designs: Balanced incomplete blocks

Randomized incomplete blocks: What if I cannot test all treatments in each block?  
Trying 4 fertilizers in four blocks, but in each block I have only space for trying 3 fertilizers

### Example with 4 fertilizers

Block	I	A	B	C	D
Block	II	D	A	B	C
Block	III	C	D	A	B
Block	IV	B	C	D	A

Balanced because every pair of treatments appearing in each block appears the same number of times as any other pair

	A	B	C	D
A	-	2	2	2
B	-	-	2	2
C	-	-	-	2
D	-	-	-	-

$$N=4 \times 3$$

	DOF
Treatments	$3=t-1$
Blocks	$3=b-1$
Residual	$5=N-1-(t-1)-(b-1)$
Total	$11=N-1$

## 2.4 Some designs: Latin squares



We want to measure the effect of 5 different car lubricants. To perform a statistical study we try with 5 different drivers and 5 different cars.

- Do we need to perform the  $125=5^3$  combinations?
- How many nuisance factors do we have?



## 2.4 Some designs: Latin squares

Latin squares: each treatment is tried only once in each nuisance level (sudoku!)

This design is used to remove two nuisance factors.

	Driver 1	Driver 2	Driver 3	Driver 4	Driver 5
Car 1	Oil 1	Oil 2	Oil 3	Oil 4	Oil 5
Car 2	Oil 5	Oil 1	Oil 2	Oil 3	Oil 4
Car 3	Oil 4	Oil 5	Oil 1	Oil 2	Oil 3
Car 4	Oil 3	Oil 4	Oil 5	Oil 1	Oil 2
Car 5	Oil 2	Oil 3	Oil 4	Oil 5	Oil 1

$$N=5*5=p^2$$

	DOF
Treatments	$4=p-1$
Car blocks	$4=p-1$
Driver blocks	$4=p-1$
Residual	$12=N-1-3(p-1)$
Total	$24=N-1$

A latin square is a square matrix in which each element occurs only once in each column and row. There are 161280 latin squares of size 5x5, we have used one of them.



## 2.4 Some designs: Graeco-Latin squares

Graeco-Latin squares: Two orthogonal (if when superimposed, each combination appears only once) latin squares superimposed. This design is used to remove three nuisance factors.

We want to measure the effect of 3 different car lubricants. To perform a statistical study we try with 3 different drivers, 3 different cars, and 3 different driving situations (city, highway, mixture)

	Driver 1	Driver 2	Driver 3
Car 1	Oil 1 ← City	Oil 2 Highway	Oil 3 → Mixture
Car 2	Oil 3 Highway	Oil 1 Mixture	Oil 2 → City
Car 3	Oil 2 ← Mixture	Oil 3 City	Oil 1 → Highway

$$N=3*3=p^2$$

	DOF
Treatments	$2=p-1$
Car blocks	$2=p-1$
Driver blocks	$2=p-1$
Situation blocks	$2=p-1$
Residual	$0=N-1-4(p-1)$
Total	$8=N-1$

## 2.4 Some designs: Replicated Graeco-Latin squares

We want to measure the effect of 3 different car lubricants. To perform a statistical study we try with 3 different drivers, 3 different cars, and 3 different driving situations (city, highway, mixture). We will perform 2 replicates per combination.

$$N=3*3*2=p^2r$$

	Driver 1	Driver 2	Driver 3
Car 1	Oil 1 City	Oil 2 Highway	Oil 3 Mixture
Car 2	Oil 3 Highway	Oil 1 Mixture	Oil 2 City
Car 3	Oil 2 Mixture	Oil 3 City	Oil 1 Highway

	DOF
Treatments	$2=p-1$
Car blocks	$2=p-1$
Driver blocks	$2=p-1$
Situation blocks	$2=p-1$
Residuals	$9=N-1-4(p-1)$
Total	$17=N-1$

## 2.4 Some designs: Full $2^k$ Factorial designs

We want to measure the effect of  $k$  factors, each one with two levels (yes, no; high, low; present, absent; ...)

Factor <sub>1</sub>	Factor <sub>2</sub>	Factor <sub>3</sub>	Factor <sub>4</sub>	Replicate <sub>1</sub>	Replicate <sub>2</sub>	Replicate <sub>3</sub>	N=2 <sup>kr</sup>	
No	No	No	No	10	12	15		DOF
No	No	No	Yes	8	9	11	Factor 1	1
No	No	Yes	No	...			...	1
No	No	Yes	Yes	...			Factor k	1
No	Yes	No	No	...			Interaction 1,2	1
No	Yes	No	Yes	...			...	1
No	Yes	Yes	Yes	...			Interaction k-1,k	1
Yes	No	No	No	...			Interaction 1,2,3	1
Yes	No	No	Yes	...			...	1
Yes	No	Yes	No	...			Interaction 1,2,...,k-1,k	1
Yes	No	Yes	Yes	...			Residuals	N-(2 <sup>k</sup> -1)
Yes	Yes	No	No	...			Total	N-1
Yes	Yes	No	Yes	...				
Yes	Yes	Yes	No	...				
Yes	Yes	Yes	Yes	...				

## 2.4 Some designs: Fractional $2^{k-p}$ Factorial designs

$N=2^k r$				$N=128r$	
	DOF				
Factor 1	1	}	$k$	Main effects	7
...	1				
Factor k	1				
Interaction 1,2	1	}	$\binom{k}{2} = \frac{k(k-1)}{2}$	Second order interactions	21
...	1				
Interaction k-1,k	1				
Interaction 1,2,3	1	}	$\sum_{i=3}^k \binom{k}{i}$	Higher order interactions	99
...	1				
Interaction 1,2,...,k-1,k	1				
Residuals	$N-(2^k-1)$				
Total	$N-1$				

If they can be disregarded (as in screening experiments) we can save a lot of experiments

## 2.4 Some designs: Fractional $2^{k-p}$ Factorial designs

Example: Fractional  $2^{k-1}$  factorial  $\rightarrow$   $\frac{1}{2}$  Experiments

Factor<sub>1</sub> Factor<sub>2</sub> Factor<sub>3</sub> Factor<sub>4</sub> Replicate<sub>1</sub> Replicate<sub>2</sub> Replicate<sub>3</sub>

No	No	No	No	10	12	15
Yes	No	No	Yes	15	16	15
No	Yes	No	Yes	...		
No	No	Yes	Yes	...		
Yes	Yes	No	No	...		
No	No	Yes	Yes	...		
Yes	No	Yes	No	...		
No	Yes	Yes	No	...		
Yes	Yes	Yes	Yes	...		

$$N=2^{k-1}r$$

	DOF
Factor 1+Interaction 234	1
Factor 2+Interaction 134	1
Factor 3+Interaction 124	1
Factor 4+Interaction 123	1
Interaction 12+Interaction 34	1
Interaction 13+Interaction 24	1
Interaction 14+Interaction 23	1
Residuals	$N-(2^{k-1}-1)$
Total	$N-1$

aliasing

Didn't we expect  
them to be  
negligible?

Cannot be  
cleanly estimated

This is not "a kind of magic",  
there is science behind!

Normally a fractional design is used to identify important factors in a exploratory stage, and then a full factorial analysis is performed only with the important factors.

## 2.4 Some designs

An experiment was performed to investigate the effectiveness of five insulating materials. Four samples of each material were tested at an elevated voltage level to accelerate the time to failure. The failure time in minutes are shown below:

Material	Failure time (minutes)
1	110, 157, 194, 178
2	1, 2, 4, 18
3	880, 1256, 5276, 4355
4	495, 7040, 5307, 10050
5	5, 7, 29, 2

- How many factors do we have?
- Which is this kind of design?
- Are there blocks?
- Write the DOF table



## 2.4 Some designs

An industrial engineer is investigating the effect of four assembly methods (A, B, C and D) on the assembly time for a color television component. Four operators are selected for the study. Furthermore, the engineer knows that each assembly produces such fatigue that the time required for the last assembly may be greater than the time required for the first, regardless of the method.

- How many factors do we have?
- What kind of design would you use?
- Are there blocks?
- How many experiments do we need to perform?



## 2.4 Some designs

An industrial engineer is conducting an experiment on eye focus time. He is interested in the effect of the distance of the object from the eye on the focus time. Four different distances are of interest. He has five subjects available for the experiment.

- How many factors do we have?
- What kind of design would you use?
- Are there blocks?
- How many experiments do we need to perform?





## 2.5 What is a covariate?

A covariate is variable that affects the result of the dependent variable, can be measured but cannot be controlled by the experimenter.

We want to measure the effect of 3 different car lubricants. To perform a statistical study we try with 3 different drivers, 3 different cars, and 3 different driving situations (city, highway, mixture). All these are variables that can be controlled. However, the atmospheric temperature also affects the car consumption, it can be measured but cannot be controlled.

Covariates are important in order to build models, but not for designing experiments.

# Course outline

3. Now I have data, how do I extract information? Parameter estimation
  1. How to estimate a parameter of a distribution?
  2. How to report on a parameter of a distribution? What are confidence intervals?
  3. What if my data is “contaminated”? Robust statistics

### 3. Now I have data, how do I extract information? Parameter estimation

In a class of 20 statisticians, 4 of them smoke.  
What is the proportion of smokers among statisticians?

The height of the statisticians in this class is:  
1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76,  
1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71  
What is the average height of statisticians?

The height of 4 Spanish statisticians is:  
1.73, 1.79, 1.76, 1.76  
What is the average height of Spanish statisticians  
knowing that the average should be around 1.70  
because that is the average height of Spaniards?



### 3. Now I have data, how do I extract information? Parameter estimation

**Statistic:** characteristic of a sample

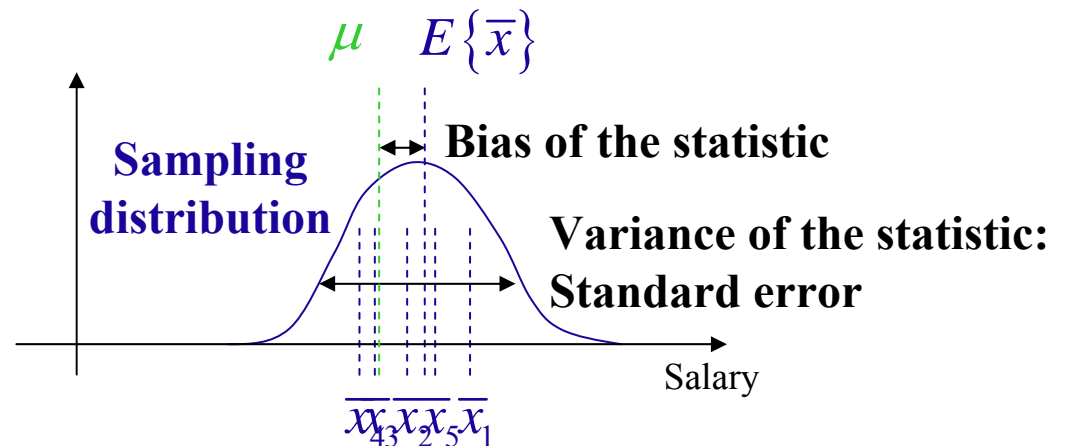
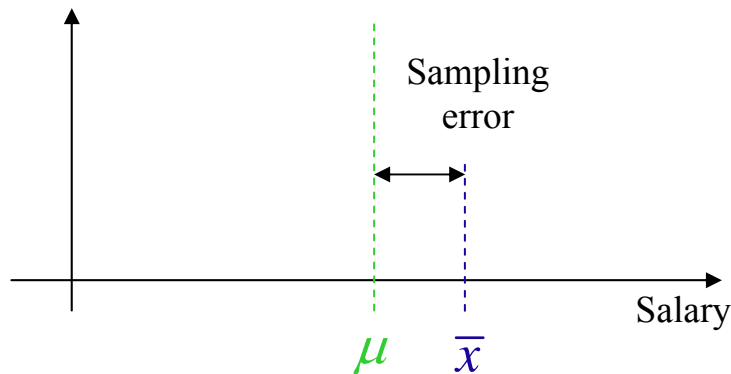
What is the average salary of 2000 people randomly sampled in Spain?

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

**Parameter:** characteristic of a population

What is the average salary of all Spaniards?

$\mu$



# 3.1 How to estimate the parameter of a distribution?

## Maximum probability

What is the proportion of smokers among statisticians?

In a class of 20 statisticians, 4 of them smoke.

$$\hat{p} = \frac{n}{N} = \frac{4}{20} = 20\%$$

Why? Couldn't it be any other number like 19%, 25%, 50%, 2%?

!!

$$Smokers \sim Binomial(N, p) \Rightarrow \Pr\{Smokers = n\} = \binom{N}{n} p^n (1-p)^{N-n}$$

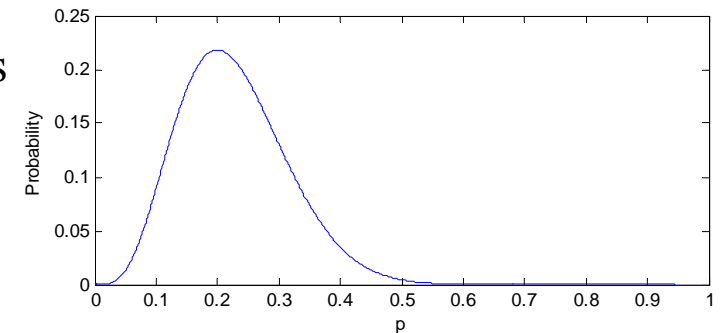
$$E\{Smokers\} = Np \longrightarrow p = \frac{E\{Smokers\}}{N}$$

$p$	$\Pr\{Smokers = 4\}$
0.20	0.218
0.19	0.217
0.25	0.190
0.50	0.005
0.02	0.001

0.2 is the parameter that maximizes the probability of our observation

$$\hat{\theta} = \arg \max_{\theta} \Pr\{X | \theta\}$$

Our data  $X \equiv Smokers = 4$



# 3.1 How to estimate the parameter of a distribution? Maximum likelihood

What is the average height of statisticians?

1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76,  
1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71

$$\bar{x} = 1.76$$

Why? Couldn't it be any other number like 1.75, 1.60, 1.78?

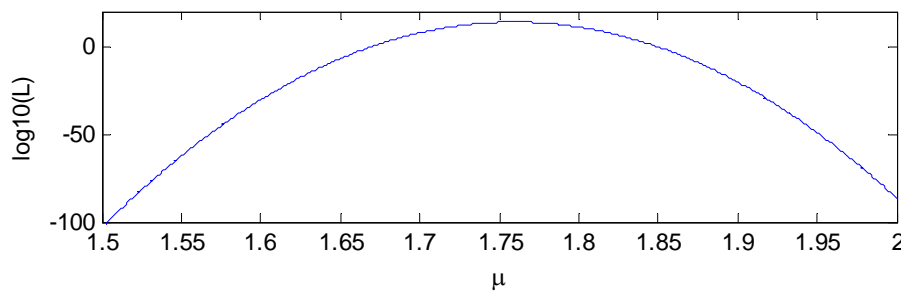
$$\hat{\theta} = \arg \max_{\theta} \Pr \{ X | \theta \}$$

$$Height \sim N(\mu, 0.05^2) \Rightarrow \Pr \{ X | \mu \} = \Pr \{ X_1 = 1.73 | \mu \} \Pr \{ X_2 = 1.67 | \mu \} \dots \Pr \{ X_{20} = 1.71 | \mu \} = 0!!$$

$$L \{ X | \mu \} = f_{N(\mu, 0.05^2)}(1.73) f_{N(\mu, 0.05^2)}(1.67) \dots f_{N(\mu, 0.05^2)}(1.71) \approx 9e13$$

$$f_{N(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\mu = \bar{x}$$



$$\hat{\theta}_{ML} = \arg \max_{\theta} \log L(X | \theta) \rightarrow \frac{\partial L \{ X | \mu \}}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \bar{x}$$

## 3.1 How to estimate the parameter of a distribution?

### Bayesian approach

What is the average height of Spanish statisticians knowing that the average should be around 1.70 because that is the average height of Spaniards?

1.73, 1.79, 1.76, 1.76

Now,  $\hat{\mu} = \bar{x} = 1.76$  is rather strange. Maybe we were unlucky in our sample

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \log L(X | \theta) L(\theta)$$

Now, our parameter is itself a random variable with an *a priori* known distribution

$$\left. \begin{array}{l} \text{Height} \sim N(\mu, 0.05^2) \\ \mu \sim N(1.70, 0.05^2) \end{array} \right\} \Rightarrow \hat{\mu} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \longrightarrow \hat{\mu} = 1.75$$

$\uparrow$        $\uparrow$   
 $\mu_0$      $\sigma_0^2$

Bayesian parameter estimation is one of the most powerful estimates **IF** you have the right a priori distribution

### 3.1 How to estimate the parameter of a distribution?

#### Other criteria

Minimum Mean Squared Error	$\hat{\theta}_{MMSE} = \arg \min_{\hat{\theta}} E \left\{ (\theta - \hat{\theta})^2 \right\} = \arg \min_{\hat{\theta}} Var \left\{ \hat{\theta} \right\} + \left( Bias(\hat{\theta}, \theta) \right)^2$ <p style="text-align: center;"> <math>\uparrow</math> Depends on something I don't know. Solution: <math>\hat{\theta}_{SURE}</math> Stein's unbiased risk estimator         </p>
Minimum risk	$\hat{\theta}_{risk} = \arg \min_{\hat{\theta}} E \left\{ Cost(\hat{\theta}, \theta) \right\} \longrightarrow Cost(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2 \Rightarrow \theta_{risk} = E \{ \theta   x \}$ $Cost(\hat{\theta}, \theta) =  \theta - \hat{\theta}  \Rightarrow \theta_{risk} = Median \{ \theta   x \}$ $Cost(\hat{\theta}, \theta) = \begin{cases} 0 &  x  < \Delta \\ \Delta &  x  \geq \Delta \end{cases} \Rightarrow \theta_{risk} = Mode \{ \theta   x \}$
Minimum Variance Unbiased Estimator	$\hat{\theta}_{MVUE} = \arg \min_{\hat{\theta}} Var \left\{ \hat{\theta} \right\}$
Best Linear Unbiased Estimator	$\hat{\theta}_{BLUE} = \arg \min_{\hat{\theta}} Var \left\{ \hat{\theta} \right\} \quad s.t. \quad \hat{\theta}_{BLUE} = \sum_{i=1}^N \alpha_i x_i$
Cramer-Rao Lower Bound	$Var \left\{ \hat{\theta} \right\} \geq \frac{1}{I(\theta)} \longleftarrow \text{Fisher's information}$

In all of them you need to know the posterior distribution of  $\theta$  given the data  $x$



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

What is the proportion of smokers among statisticians?

In a class of 20 statisticians, 4 of them smoke.

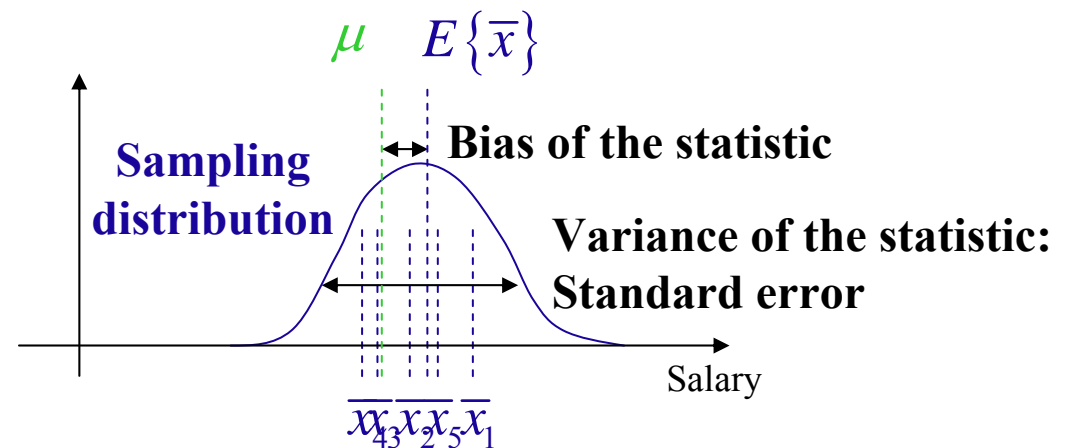
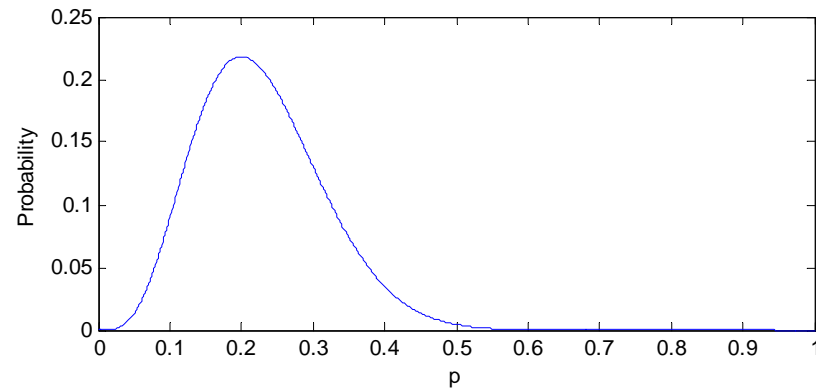
$$\hat{p} = \frac{n}{N} = \frac{4}{20} = 20\%$$

$p$	$\Pr\{\text{Smokers} = 4\}$
0.20	0.218
0.19	0.217
0.25	0.190
0.50	0.005
0.02	0.001

It would be safer to give a range:

$p \in [0, 100]\%$  confidence=100%

$p \in [18, 22]\%$  confidence=??



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

$$\text{Smokers} \sim \text{Binomial}(N, \hat{p}) \Rightarrow \Pr\{\text{Smokers} = n\} = \binom{N}{n} \hat{p}^n (1 - \hat{p})^{N-n}$$

$\hat{p} = 0.2$		$\hat{p} = 0.07135$		$\hat{p} = 0.401$	
$n$	$\Pr\{\text{Smokers} = n\}$	$n$	$\Pr\{\text{Smokers} = n\}$	$n$	$\Pr\{\text{Smokers} = n\}$
0	0.0115	0	0.2275	0	0.0000
1	0.0576	→ 1	0.3496	1	0.0005
2	0.1369	2	0.2552	2	0.0030
3	0.2054	3	0.1176	3	0.0121
→ 4	0.2182	4	0.0384	4	0.0344
5	0.1746	5	0.0094	5	0.0737
6	0.1091	6	0.0018	6	0.1234
7	0.0545	7	0.0003	7	0.1652
8	0.0222	8	0.0000	→ 8	0.1797
9	0.0074	9	0.0000	9	0.1604
10	0.0020	10	0.0000	10	0.1181
11	0.0005	11	0.0000	11	0.0719
12	0.0001	12	0.0000	12	0.0361

$\frac{\alpha}{2} = 0.05$

$\frac{\alpha}{2} = 0.05$

$p \in [0.07135, 0.401]$   
 $1 - \alpha = 0.90$   
 $\alpha = 0.10$

## 3.2 How to report on a parameter of a distribution?

### What are confidence intervals?

#### Meaning

The confidence of 90% means that our method of producing intervals produces an interval that 90% of the times contains the distribution parameter. Another way of viewing this is that if we routinely build intervals like this, on average 10 times of every 100 experiments, we will be wrong about our the intervals where the true distribution parameter is.

A wrong interpretation is that the confidence interval contains the parameter with probability 0.95.

#### More confidence

We can increase our confidence in the interval if our method builds larger intervals (decreasing the 0.05 used in the previous example).

$$p \in [0.07135, 0.401] \quad 1 - \alpha = 0.90$$

$$p \in [0.0573, 0.4366] \quad 1 - \alpha = 0.95$$

#### More accuracy

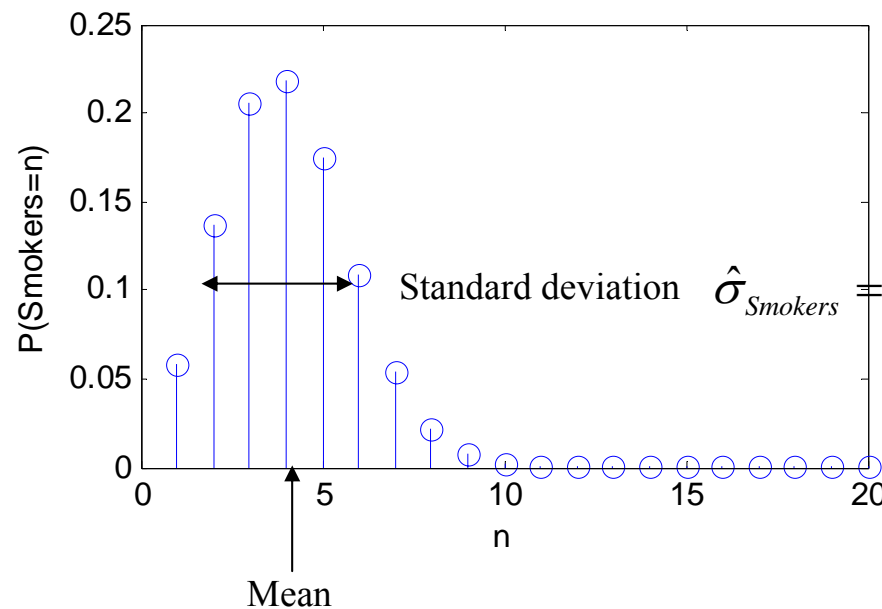
If we want smaller intervals with a high confidence, we have to increase the number of samples.

#### Centered intervals

In this example our intervals were centered “in probability”, i.e., the probability of error used to compute the limits was the same on both sides (0.05). We can build assymetric intervals, or even one-sided intervals.

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Shortcut calculation of the variability of an estimated proportion



We need to know  
the distribution of  
the statistic!!!

$$\hat{\sigma}_{\hat{p}} = \frac{1}{N} \hat{\sigma}_{Smokers} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = 0.09$$

$$p \in "0.20 \pm 2 \cdot 0.09" \longleftrightarrow p \in [0.07135, 0.401]$$

$\alpha = 0.90$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

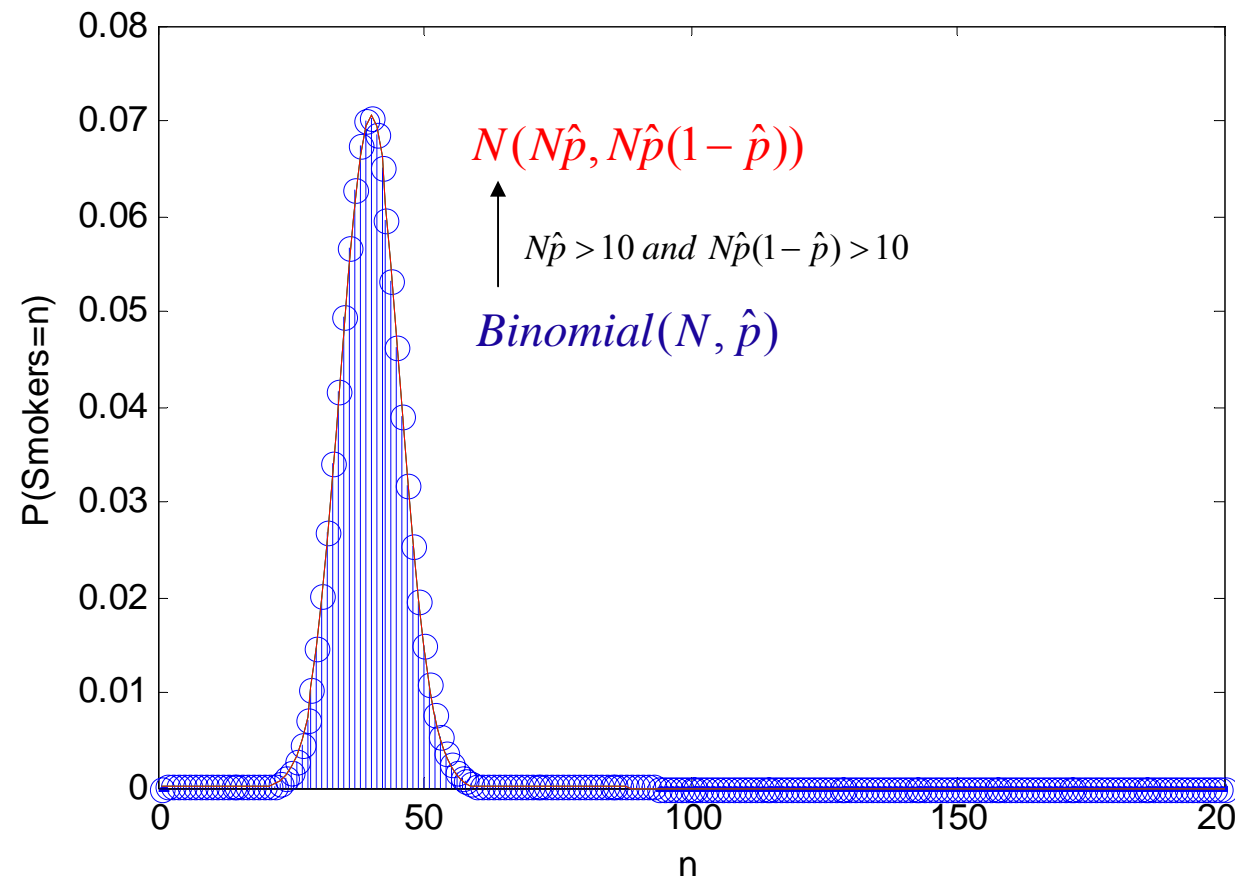
What is the proportion of smokers among statisticians?  
In a class of 200 statisticians, 40 of them smoke.

$$\hat{p} = 20\%$$

$$p \in [0.15455, 0.25255]$$

$$\alpha = 0.90$$

$$p \in "0.20 \pm 2 \cdot 0.03"$$



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

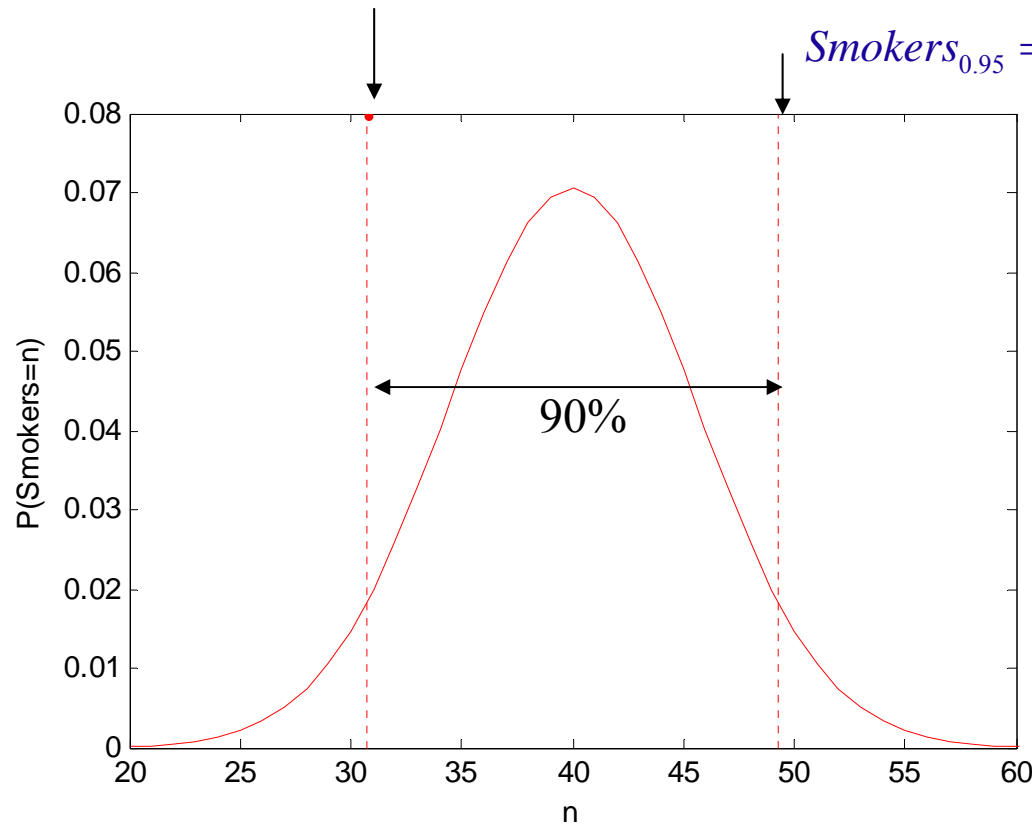
What is the proportion of smokers among statisticians?

In a class of 200 statisticians, 40 of them smoke.

$$\hat{p} = 20\%$$

$$Smokers_{0.05} = \mu_{Smokers} + q_{0.05} \sigma_{Smokers} = 40 - 1.6449 \cdot 5.66 = 30.7$$

$$Smokers_{0.95} = \mu_{Smokers} + q_{0.95} \sigma_{Smokers} = 40 + 1.6449 \cdot 5.66 = 49.3$$



$$Smokers \in [30.7, 49.3]$$

$$p \in [0.1535, 0.2465] \text{ vs. } p \in [0.15455, 0.25255]$$

$$p \in \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$$p \in \frac{1}{1 + \frac{z_{\frac{\alpha}{2}}^2}{N}} \left( \left( \hat{p} + \frac{z_{\frac{\alpha}{2}}^2}{2N} \right) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z_{\frac{\alpha}{2}}^2}{4N^2}} \right)$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

Shortcut through the Poisson distribution for rare events

The binomial can be approximated by a Poisson of parameter  $\lambda = np$  if  $\begin{cases} n \geq 20 & \text{and} & p \leq 0.05 \\ \text{or} \\ n \geq 100 & \text{and} & np \leq 10 \end{cases}$

$$p \in \left[ 0, \frac{1}{2N} \chi^2_{1-\alpha, 2(N\hat{p}+1)} \right]$$

I have to go under a dangerous operation and my doctor told me that they have carried out 20 operations, and no one died from it. Does it mean that there is a 0 probability of fatal result?

$$p \in \left[ 0, \frac{1}{2 \cdot 20} \chi^2_{1-0.05, 2(20 \cdot 0 + 1)} \right] \approx \left[ 0, \frac{3}{20} \right] = [0, 15\%]$$

The probability of finding a certain bacteria in a liter of water is 0.1%. How many liters do I need to take to be able to find 1 organism with a confidence of 95%.

$$\frac{1}{2 \cdot N} \chi^2_{1-\alpha, 2(Np+1)} = p \Rightarrow N \approx \frac{3}{p} = \frac{3}{0.001} = 3000$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

$$X_i \sim N(\mu_X, \sigma_X^2) \Rightarrow \bar{x} \sim N\left(\mu_X, \frac{\sigma_X^2}{N_X}\right) \Rightarrow \frac{\bar{x} - \mu_X}{\frac{\sigma_X}{\sqrt{N_X}}} \sim N(0,1)$$

$$\Rightarrow \frac{\bar{x} - \mu_X}{\frac{s_X}{\sqrt{N_X}}} \sim t_{N_X-1}$$

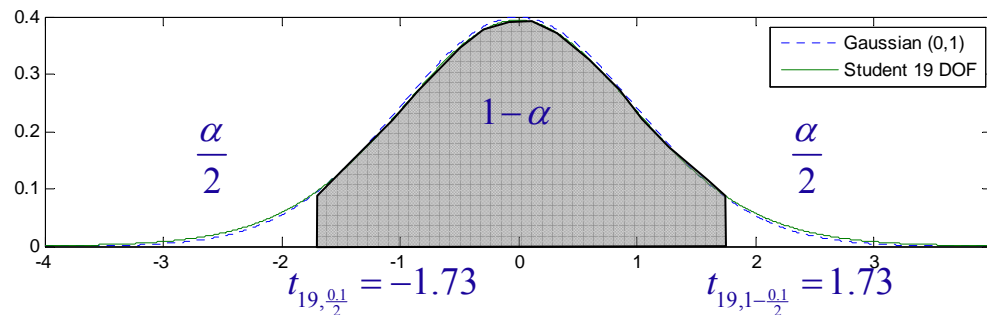
Sample mean distribution  
when the variance is known

Sample mean distribution when  
the variance is **unknown**

What is the average height of statisticians?

1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76,  
1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71

$$\begin{aligned} \bar{x} &= 1.76 \\ s &= 0.04 \end{aligned} \longrightarrow \frac{1.76 - \mu_X}{\frac{0.04}{\sqrt{20}}} \sim t_{19}$$



$$\Pr\left\{\left|\frac{1.76 - \mu_X}{\frac{0.04}{\sqrt{20}}}\right| < t_{19, 1-\frac{\alpha}{2}}\right\} = 1 - \alpha$$

$$\left|\frac{1.76 - \mu_X}{\frac{0.04}{\sqrt{20}}}\right| < 1.73$$

$$1.76 - 1.73 \frac{0.04}{\sqrt{20}} < \mu_X < 1.76 + 1.73 \frac{0.04}{\sqrt{20}} \Rightarrow \mu_X \in [1.74, 1.78]$$



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

Sample mean distribution when the variance is **unknown** and the data is normal.

$$\frac{\bar{x} - \mu_X}{\frac{s_X}{\sqrt{N_X}}} \sim t_{N_X-1} \longrightarrow \mu_X \in \bar{x} \pm t_{N_X-1, \frac{\alpha}{2}} \frac{s_X}{\sqrt{N_X}}$$

$$\mu_X \in [1.74, 1.78]$$

Realistic

$\alpha$	$z_{1-\frac{\alpha}{2}}$
0.001	3.2905
0.005	2.8075
0.01	2.5758
0.05	1.9600
0.1	1.6449

$$N_X \geq 30$$

$$\mu_X \in \bar{x} \pm z_{\frac{\alpha}{2}} \frac{s_X}{\sqrt{N_X}}$$

$$\mu_X \in [1.75, 1.77]$$

Optimistic

Sample mean distribution when the variance is **unknown** and the data distribution is **unknown**.

$$\Pr \left\{ \left| \frac{X - \mu}{\sigma} \right| \leq K \right\} = 1 - \frac{1}{K^2}$$

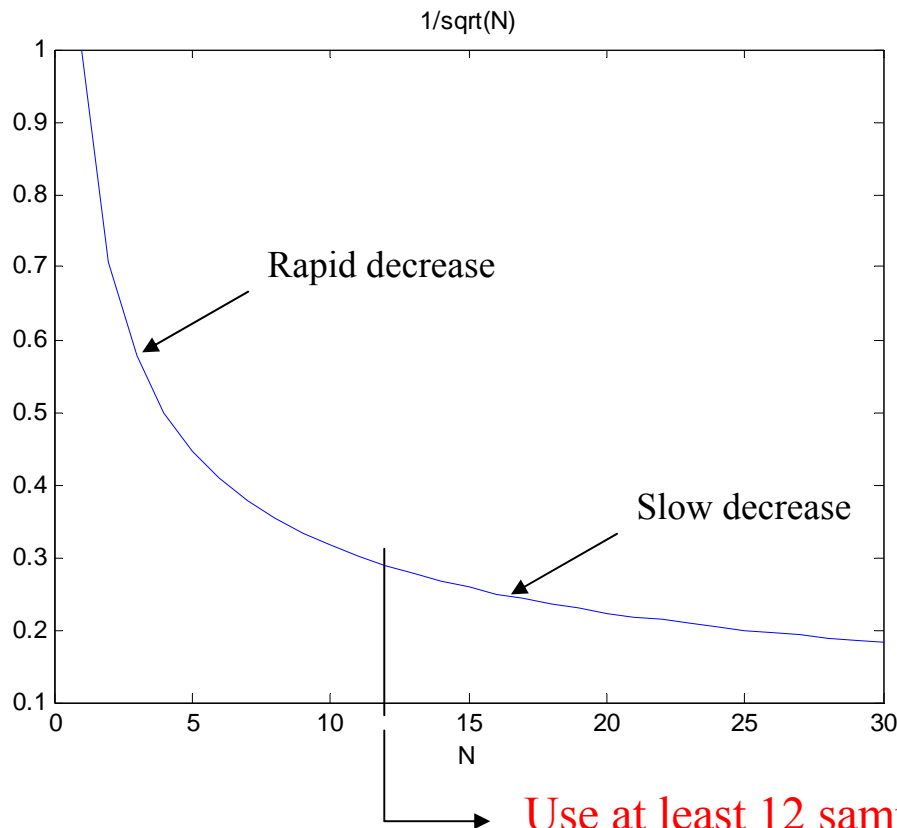
$$\mu_X \in [1.73, 1.79]$$

Pesimistic

$$\mu_X \in \bar{x} \pm \frac{1}{\sqrt{\alpha}} \frac{s_X}{\sqrt{N_X}}$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

Note that the variance is multiplied by the inverse of the square root of N



$$\mu_X \in \bar{x} \pm t_{N_X-1, \frac{\alpha}{2}} \frac{s_X}{\sqrt{N_X}}$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Unpaired differences

Sample difference distribution  
when the variances are known

$$\begin{aligned} X_i &\sim N(\mu_X, \sigma_X^2) \\ Y_i &\sim N(\mu_Y, \sigma_Y^2) \Rightarrow \bar{d} \sim N\left(\mu_d, \frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}\right) \\ &\quad \uparrow \\ &\quad \bar{d} = \bar{x} - \bar{y} \end{aligned}$$

Sample difference distribution when  
the variances are **unknown** and  $N_X = N_Y$

$$\frac{\bar{d} - \mu_d}{\sqrt{\frac{s_X^2}{N} + \frac{s_Y^2}{N}}} \sim t_{2N-2}$$

Sample difference distribution when  
the variances are **unknown but the same** and  $N_X \neq N_Y$

$$\frac{\bar{d} - \mu_d}{\sqrt{\frac{(N_X - 1)s_X^2 + (N_Y - 1)s_Y^2}{N_X + N_Y - 2} \left( \frac{1}{N_X} + \frac{1}{N_Y} \right)}} \sim t_{N_X + N_Y - 2}$$

Sample difference distribution when  
the variances are unknown and  
different and  $N_X \neq N_Y$

$$\frac{\bar{d} - \mu_d}{\sqrt{\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}}} \sim t \quad \frac{\left( \frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y} \right)^2}{\frac{s_X^2}{N_X^2(N_X - 1)} + \frac{s_Y^2}{N_Y^2(N_Y - 1)}}$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

Sample difference distribution when  
the variances are **unknown** and  $N_X = N_Y$

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{N} + \frac{s_Y^2}{N}}} \sim t_{2N-2}$$

An engineer tries to determine if a certain modification makes his motor to waste less power. He makes measurements of the power consumed with and without modifications (the motors tested are different in each set). The nominal consumption of the motors is 750W, but they have from factory an unknown standard deviation around 20W. He obtains the following data:

Unmodified motor (Watts): 741, 716, 753, 756, 727

$$\bar{x} = 738.6 \quad s_X = 17.04$$

Modified motor (Watts): 764, 764, 739, 747, 743

$$\bar{y} = 751.4 \quad s_Y = 11.84$$

$$\bar{d} = 751.4 - 738.6 = 12.8$$

$$\mu_d \in 12.8 \pm t_{8, \frac{0.05}{2}} \sqrt{\frac{17.04^2}{5} + \frac{11.84^2}{5}} \in [-8.6, 34.2]$$



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

Our engineer is convinced that he did it right, and keeps on trying:

Unmodified motor (Watts): 752 771 751 748 733 756 723 764 782  
736 767 775 718 721 761 742 764 766 764 776 763 774 726 750 747  
718 755 729 778 734

Modified motor (Watts): 751 744 722 697 739 720 752 750 774 752  
727 748 720 740 739 740 734 762 703 749 758 755 752 741 754 751  
735 732 734 710

$$\bar{x} = 751.6 \quad s_x = 19.52$$

$$\bar{y} = 739.4 \quad s_y = 17.51$$

$$\bar{d} = 739.4 - 751.6 = -12.19$$

$$\mu_d \in -12.19 \pm t_{58, \frac{0.05}{2}} \sqrt{\frac{19.52^2}{30} + \frac{17.51^2}{30}} \in [-21.78, -2.61]$$



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Paired differences

Sample difference distribution  
when the variance of the  
difference is known

$$\begin{aligned} X_i &\sim N(\mu_X, \sigma_X^2) \\ Y_i &\sim N(\mu_Y, \sigma_Y^2) \Rightarrow \bar{d} \sim N\left(\mu_d, \frac{\sigma_d^2}{N}\right) \end{aligned}$$

Sample difference distribution when  
the variance of the difference is  
**unknown**

$$\frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{N}}} \sim t_{N-1}$$

$$\mu_d \in \bar{d} \pm t_{N-1, \frac{\alpha}{2}} \frac{s_d}{\sqrt{N}}$$

Our engineer learnt the lesson and now will test the same motor before and after  
modification

Unmodified motor (Watts): 755 750 730 731 743

Modified motor (Watts): 742 738 723 721 730

Difference: -13 -12 -7 -10 -13

$$\bar{d} = -11 \quad s_d = 2.56$$

$$\mu_d \in -11 \pm t_{4, \frac{0.05}{2}} \frac{2.56}{\sqrt{5}} \in [-14.17, -7.83]$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Variance and standard deviation

Sample variance distribution  $X_i \sim N(\mu_X, \sigma_X^2) \Rightarrow \frac{s^2}{\frac{\sigma^2}{N-1}} \sim \chi_{N-1}^2$

$$\Pr \left\{ \chi_{N-1, \frac{\alpha}{2}}^2 < \frac{s^2}{\frac{\sigma^2}{N-1}} < \chi_{N-1, 1-\frac{\alpha}{2}}^2 \right\} = 1 - \alpha \longrightarrow \frac{(N-1)s^2}{\chi_{N-1, 1-\frac{\alpha}{2}}^2} < \sigma^2 < \frac{(N-1)s^2}{\chi_{N-1, \frac{\alpha}{2}}^2}$$

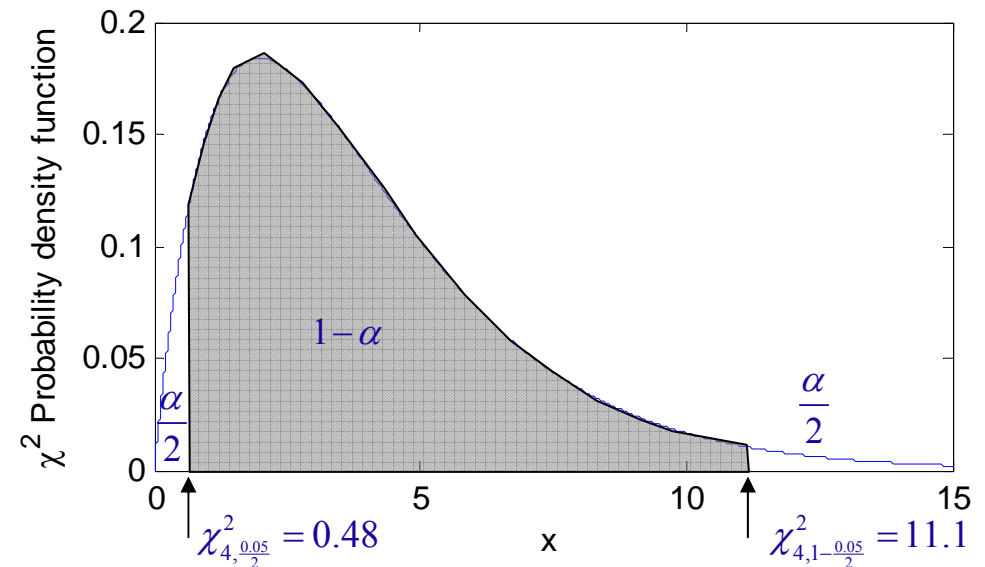
Our engineer wants to know which is the variability of the unmodified motors

Unmodified motor (Watts):

755 750 730 731 743

$$\bar{x} = 741.8 \quad s = 11.17$$

$$\sigma^2 \in \left[ \frac{4 \cdot 11.17^2}{11.1}, \frac{4 \cdot 11.17^2}{0.48} \right] = [6.7^2, 32.2^2]$$



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Variance difference

Sample variance distribution

$$X_i \sim N(\mu_X, \sigma_X^2) \Rightarrow \frac{\frac{s_X^2}{s_Y^2}}{\frac{\sigma_X^2}{\sigma_Y^2}} \sim F_{N_X-1, N_Y-1}$$

$$Y_i \sim N(\mu_Y, \sigma_Y^2)$$

$$\Pr \left\{ F_{N_X-1, N_Y-1, \frac{\alpha}{2}} < \frac{\frac{s_X^2}{s_Y^2}}{\frac{\sigma_X^2}{\sigma_Y^2}} < F_{N_X-1, N_Y-1, 1-\frac{\alpha}{2}} \right\} = 1 - \alpha \longrightarrow \boxed{\frac{\frac{s_X^2}{s_Y^2}}{F_{N_X-1, N_Y-1, 1-\frac{\alpha}{2}}} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{\frac{s_X^2}{s_Y^2}}{F_{N_X-1, N_Y-1, \frac{\alpha}{2}}}}$$

Our engineer wants to know if the  
modification introduces an extra variance  
in the motors

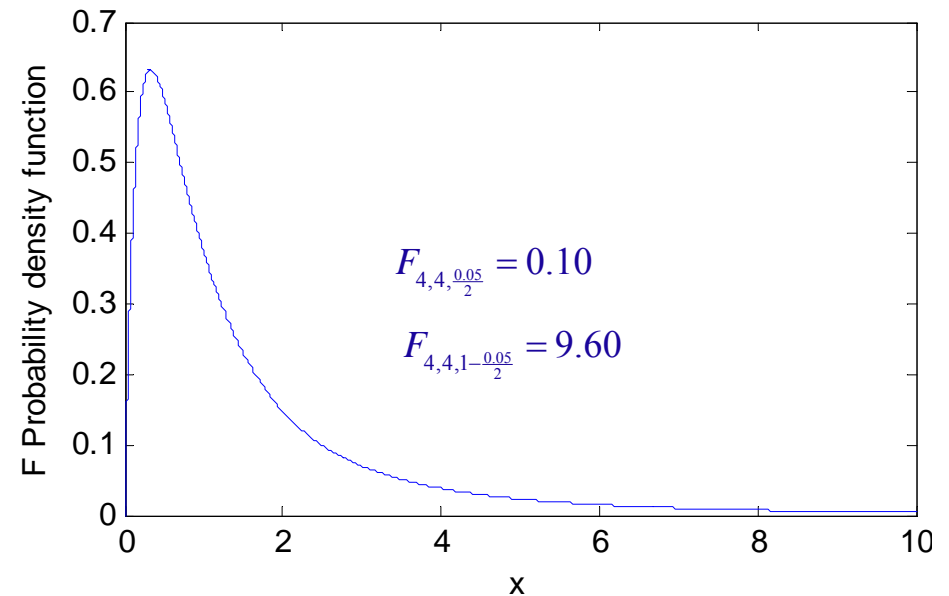
Unmodified motor (Watts):  $s_X = 11.17$

755 750 730 731 743  $s_Y = 9.28$

Modified motor (Watts):

742 738 723 721 730

$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left[ \frac{\frac{11.17^2}{9.28^2}}{9.6}, \frac{\frac{11.17^2}{9.28^2}}{0.1} \right] = [0.39^2, 3.8^2]$$





## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Correlation

Sample correlation distribution

$$\begin{aligned} X_i &\sim N(\mu_X, \sigma_X^2) \Rightarrow Z = \frac{1}{2} \log \frac{1+r}{1-r} \sim N\left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{N-3}\right) \\ Y_i &\sim N(\mu_Y, \sigma_Y^2) \end{aligned}$$

$$\rho \in \left[ \tanh\left(r - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{N-3}}\right), \tanh\left(r + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{N-3}}\right) \right]$$

Our engineer wants to know if there is a relationship between the power consumption and room temperature. He collected the following data:

Room temperature: 23.26 24.31 20.66 22.08 22.48

Unmodified motor (Watts): 755 750 730 731 743

$$r = 0.84$$

$$\rho \in \left[ \tanh\left(0.84 - 1.96 \frac{1}{\sqrt{2}}\right), \tanh\left(0.84 + 1.96 \frac{1}{\sqrt{2}}\right) \right] \in [-0.497, 0.977]$$

## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

### Finite population correction

Correction for a simple random  
sampling without replacement

$$s_{\bar{x}, corrected}^2 = s_{\bar{x}}^2 \left( 1 - \frac{N_{sample}}{N_{population}} \right)$$

It should be applied if the sample  
represents more than 5% of the  
total population

What is the average height of statisticians?

$$\begin{array}{l} 1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76, \\ 1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71 \end{array} \longrightarrow \begin{array}{l} \bar{x} = 1.76 \\ s = 0.04 \end{array} \longrightarrow \frac{1.76 - \mu_x}{\frac{0.04}{\sqrt{20}}} \sim t_{19}$$

$$\mu_x \in [1.74, 1.78] \quad 1 - \alpha = 0.90$$

What is the average height of statisticians in this class of 30 students?

$$\begin{array}{l} 1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76, \\ 1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71 \end{array} \longrightarrow \begin{array}{l} \bar{x} = 1.76 \\ s = 0.04 \sqrt{1 - \frac{20}{30}} \end{array}$$

$$\frac{1.76 - \mu_x}{\frac{0.04}{\sqrt{20}} \sqrt{1 - \frac{20}{30}}} \sim t_{19} \quad \mu_x \in [1.75, 1.77] \quad 1 - \alpha = 0.90$$

### 3.3 How to report on a parameter of a distribution? What are confidence intervals?

We are interested in comparing two supposedly identical coffee machines.  
We performed 9 measures on both machines obtaining

$$\bar{x}_1 = 150ml \quad s_1 = 3ml$$

$$\bar{x}_2 = 152ml \quad s_2 = 4ml$$

Which are the confidence intervals for the mean filling of each machine?

Can we conclude that one of them is dispensing more coffee?



## 3.2 How to report on a parameter of a distribution? What are confidence intervals?

Facing the national elections there is a poll trying to guess which political party will win the elections. Of a population of 30M people who can vote, we take a random sample of 5k people. These are the results

Party A: 38.9%

Party B: 36.5%

Rest of parties: 24.6%

Can we guarantee that Party A will win the elections?



## 3.3 What if my data is “contaminated”?

### Robust statistics

A “contamination” is anything that keeps your data away from the assumptions of the methods to compute confidence intervals (mainly “normality”):

- The data is not normal
  - Use non-parametric estimators of confidence intervals (e.g., bootstrap)
  - Fit a general probability density function (e.g. Gaussian mixture)
- The data is normal but there are outliers
  - Remove outliers if possible
  - Use robust estimators of the mean, variance, ...

**Bootstrap** estimate of the confidence interval for the mean of  $N(>100)$  samples:

1. Take a random subsample of the original sample of size  $N$  (with replacement)
2. Estimate the mean of your subsample.
3. Repeat steps 1 and 2 at least 1000 times.

This gives you the empirical distribution of the mean from which the confidence interval can be computed. This empirical distribution can be computed with any statistic (median, mode, regression coefficient, ...)



### 3.3 What if my data is “contaminated”? Robust statistics

The average length of *Chelonia Mydas* (Green Turtle) is about 1 meter. However, there is a group of green turtles in Australia that may be larger.



We monitor the glucose level of a patient to track its evolution over time. The glucose sensor is a needle under the skin. Sometimes, due to the patient movement our measurements go to zero (which is clearly not a valid glucose value).



Discuss about the normality of the dataset and the way of computing confidence intervals.

## Course outline

4. Can I see any interesting association between two variables, two populations, ...?
  1. What are the different measures available?
  2. Use and abuse of the correlation coefficient
  3. How can I use models and regression to improve my measure of association?

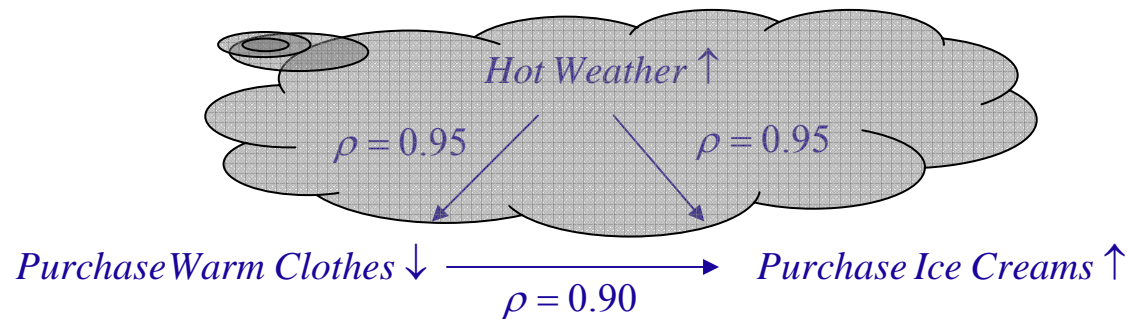
## 4.1 What are the different measures of association available?

- 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  $X_2$ ?

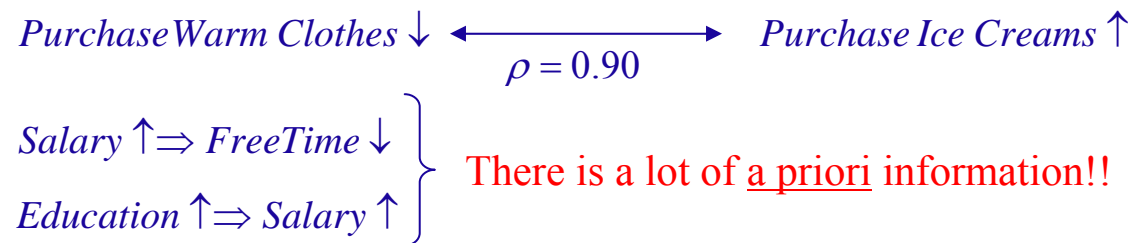


## 4.2 Use and abuse of the correlation coefficient

Pitfall: Correlation means causation

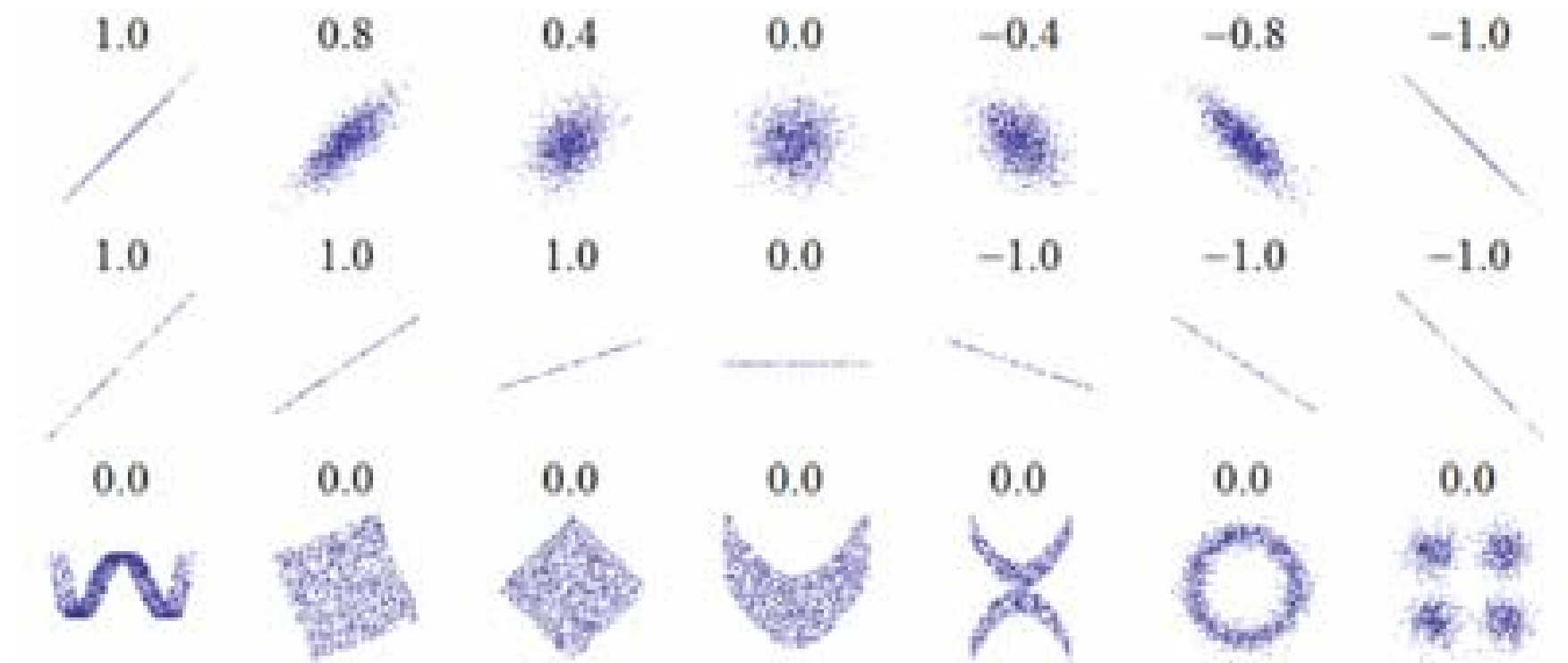


Correct: Correlation means linear covariation



## 4.2 Use and abuse of the correlation coefficient

**Pitfall:** Correlation measures all possible associations

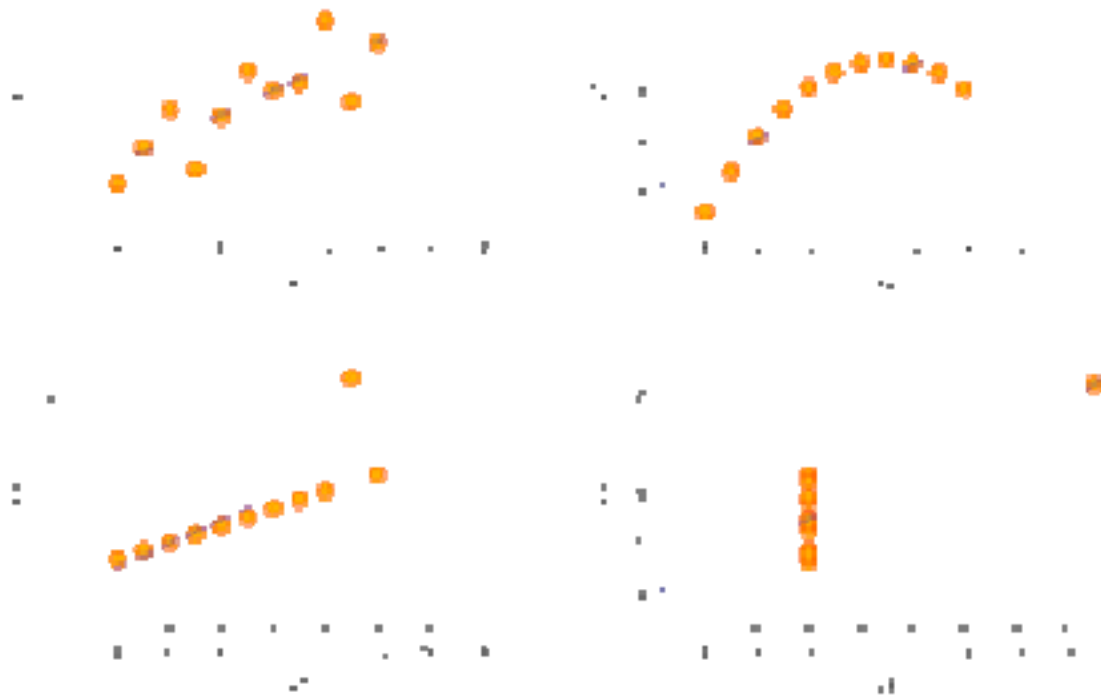


**Correct:** Correlation measures only linear associations

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

## 4.2 Use and abuse of the correlation coefficient

Pitfall: Correlation summarizes well the relationship between two variables



$$\bar{y} = 7.5$$

$$s_Y = 4.12$$

$$y = 3 + 0.5x$$

$$r = 0.81$$

Correct: Visual inspection of the data structure is always needed

## 4.2 Use and abuse of the correlation coefficient

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

## 4.2 Use and abuse of the correlation coefficient

Pitfall: Correlation between combinations with common variables



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

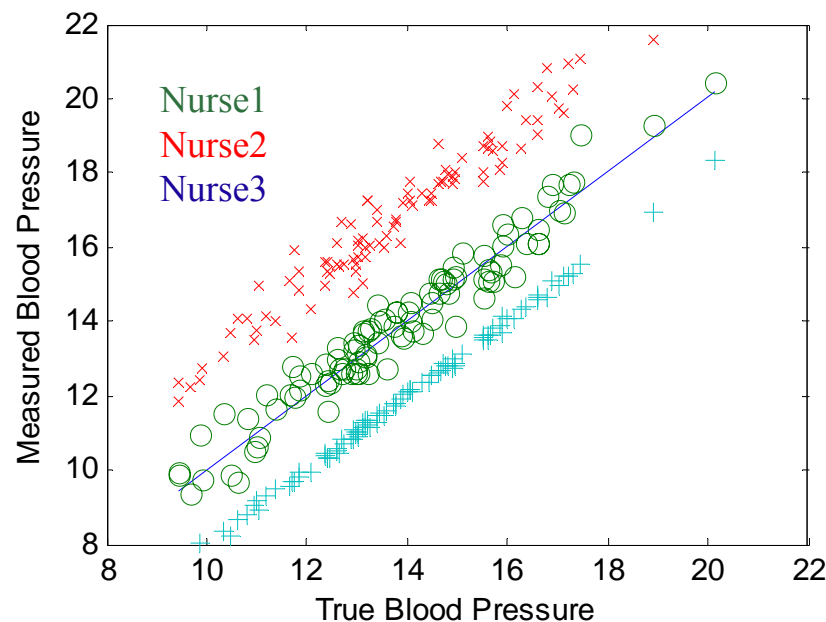
$r_{BabiesPerWoman, StorkPerWoman} = 0.63!! \quad (p < 0.00001)$

## 4.2 Use and abuse of the correlation coefficient

**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.



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

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

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

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

## 4.2 Use and abuse of the correlation coefficient

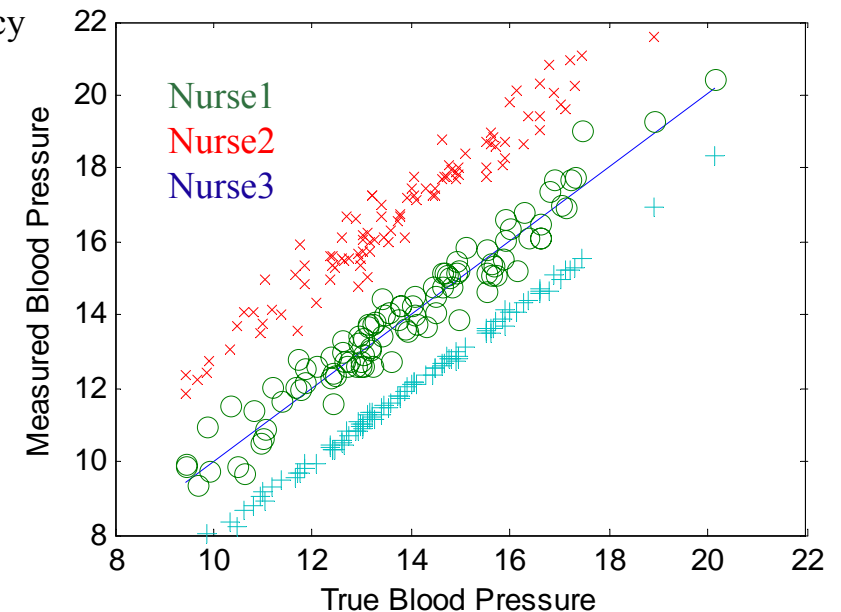
**Solution:** Assess agreement through bias, scale difference and accuracy

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

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

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.



## 4.2 Use and abuse of the correlation coefficient

**Pitfall:** Summarize a regression experiment through correlation

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

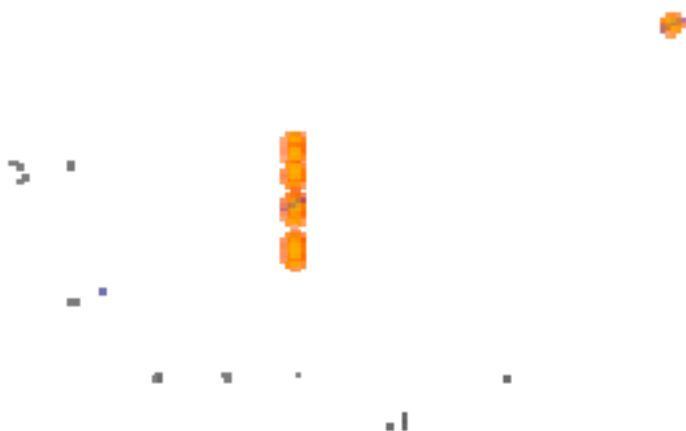
Which is the relationship between height and weight in a random sample?

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.



$$r^2 = \frac{1}{1 + K^2 \frac{s_{Y|X}^2}{s_X^2}}$$

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

$s_X^2$  The experimenter controls the width of values tested.

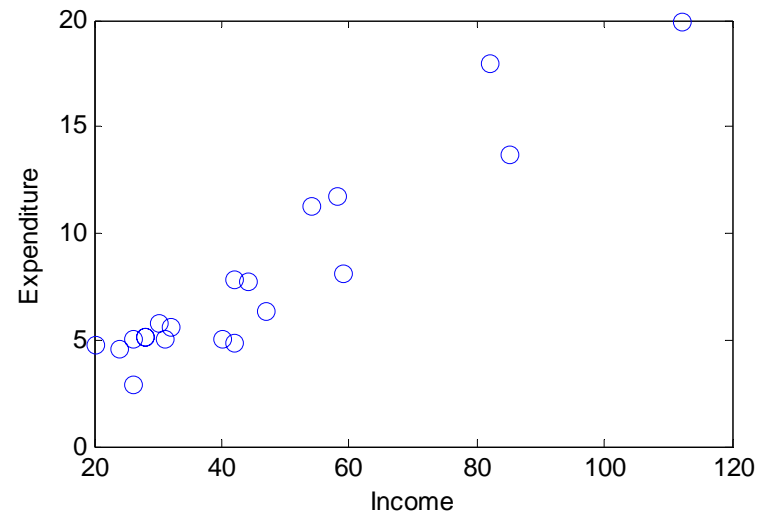
By making the range of X large, we can have a correlation as closed to 1 as desired.



## 4.3 How can I use models and regression to improve my measure of association?

Models not only provide an “index of association” but also an “explanation”.

Modelled Expenditure (k\$)	Annual Food Expenditure (k\$)	Annual Income (k\$)
$\hat{y}_i$ 4.7	$y_i$ 5.2	28
4.4	5.1	26
5.4	5.6	32
4.0	4.6	24
9.5	11.3	54
10.5	8.1	59
7.7	7.8	44
5.1	5.8	30
7.0	5.1	40
14.7	18.0	82
7.3	4.9	42
10.3	11.8	58
4.7	5.2	28
3.3	4.8	20
7.3	7.9	42
8.2	6.4	47
20.2	20.0	112
15.2	13.7	85
5.3	5.1	31
4.4	2.9	26
$\bar{\hat{y}} = 7.97$	$\bar{y} = 7.97$	



$$r_{\text{Expenditure, Income}} = 0.95$$

$$\text{Expenditure} = 18\% \text{Income} - 412\$$$

$$413.35 = 369.57 + 43.78$$

$$R^2 = 1 - \frac{43.78}{413.35} = 0.89$$

$$= r_{\text{Expenditure, Income}}^2$$

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

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

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

The coefficient of determination represents the percentage of “unexplained” variance

## 4.3 How can I use models and regression to improve my measure of association?

Models not only provide an “index of association” but also an “explanation”.

Modelled Expenditure (k\$)	Annual Food Expenditure (k\$)	Annual Income (k\$)	Family members
$\hat{y}_i$ 5.4	$y_i$ 5.2	28	3
5.1	5.1	26	3
5.2	5.6	32	2
3.2	4.6	24	1
10.1	11.3	54	4
9.2	8.1	59	2
7.8	7.8	44	3
4.9	5.8	30	2
5.6	5.1	40	1
15.8	18.0	82	6
7.5	4.9	42	3
10.7	11.8	58	4
3.8	5.2	28	1
5.8	4.8	20	5
7.5	7.9	42	3
6.6	6.4	47	1
20.2	20.0	112	6
15.4	13.7	85	5
5.1	5.1	31	2
4.3	2.9	26	2
$\bar{\hat{y}} = 7.97$	$\bar{y} = 7.97$		

$$\text{Expenditure} = 15\% \text{Income} + 790\$ \cdot \text{Members} - 1120\$$$

$$413.35 = 386.31 + 27.04$$

$$R^2 = 1 - \frac{27.04}{413.35} = 0.93$$

## 4.3 How can I use models and regression to improve my measure of association?

25 hypertense patients from the same hospital were randomly assigned to 5 groups. In each group a different treatment was given to test the efficacy of a new drug.

No Treatment	No salt diet	Salt diet	Dosis 1mg	Dosis 2mg	$y_i$	Sistolic pressure
180	172	163	158	147		
173	158	170	146	152		
175	167	158	160	143		
182	160	162	171	155		
181	175	170	155	160		
					$\bar{y} = 163.7$	

	178.2	166.4	164.6	158.0	151.4
$\alpha_{treatment}$	14.5	2.7	0.9	-5.7	-12.3

No Treatment	No salt diet	Salt diet	Dosis 1mg	Dosis 2mg	$\hat{y}_i$	
178.2	166.4	164.6	158.0	151.4		
178.2	166.4	164.6	158.0	151.4		
178.2	166.4	164.6	158.0	151.4		
178.2	166.4	164.6	158.0	151.4		
178.2	166.4	164.6	158.0	151.4		
					$\bar{\hat{y}} = 163.7$	

$$SistolicPressure = 163.7 + \alpha_{treatment}$$

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

$$2905 = 2011 + 894$$

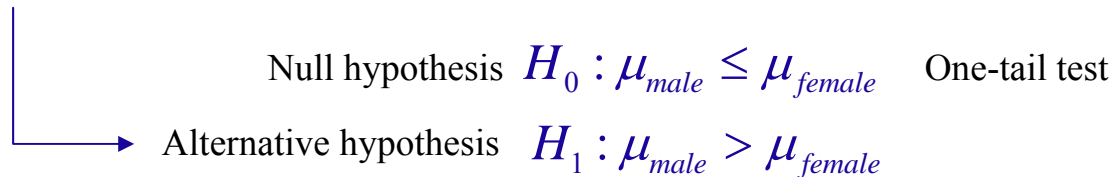
$$R^2 = 1 - \frac{894}{2905} = 0.69$$

## Course outline

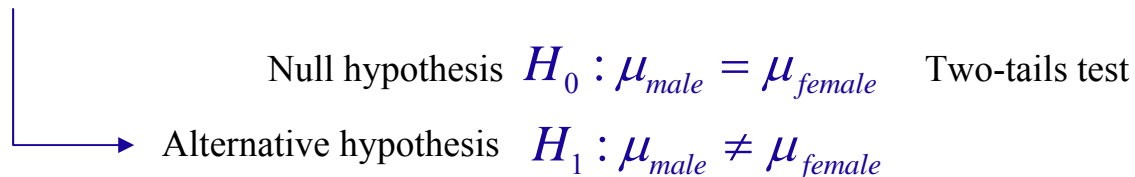
5. How can I know if what I see is “true”? Hypothesis testing
  1. The basics: What is a hypothesis test? What is the statistical power? What is a p-value? How to use it? What is the relationship between sample size, sampling error, effect size and power? What are bootstraps and permutation tests?
  2. What are the assumptions of hypothesis testing?
  3. How to select the appropriate statistical test
    - i. Tests about a population central tendency
    - ii. Tests about a population variability
    - iii. Tests about a population distributions
    - iv. Tests about differences randomness
    - v. Tests about correlation/association measures
  4. Multiple testing
  5. Words of caution

## 5.1 Basics: What is a hypothesis test?

Research hypothesis: Among couples married for 5 years or more, males report higher levels of intimacy than females.



Research hypothesis: Among couples married for 5 years or more, males report a different level of intimacy than females.



Research hypothesis: Our motor engineer wants to know if his modified motors really consume less than the unmodified ones.



## 5.1 Basics: What is a hypothesis test?

We can have more complex hypothesis

25 hypertense patients from the same hospital were randomly assigned to 5 groups. In each group a different treatment was given to test the efficacy of a new drug.

Research hypothesis: Not all treatments are the same.

Null hypothesis  $H_0 : \mu_{NoTreatment} = \mu_{NoSaltDiet} = \mu_{SaltDiet} = \mu_{Dosis1mg} = \mu_{Dosis2mg}$

Alternative hypothesis  $H_1 : \exists i, j \mid \mu_i \neq \mu_j$

Research hypothesis: The effect (positive or negative) of a dose of 1mg is larger than 2.

Null hypothesis  $H_0 : \{ \mu_{Dosis1mg} \geq -2 \} \cap \{ \mu_{Dosis1mg} \leq 2 \}$

Alternative hypothesis  $H_1 : \{ \mu_{Dosis1mg} < -2 \} \cup \{ \mu_{Dosis1mg} > 2 \}$

Research hypothesis: The effect of a dose of 1mg is larger than 2, and the effect of a dose of 2mg is non-negative.

Null hypothesis  $H_0 : \{ \mu_{Dosis1mg} \leq 2 \} \cup \{ \mu_{Dosis2mg} < 0 \}$

Alternative hypothesis  $H_1 : \{ \mu_{Dosis1mg} > 2 \} \cap \{ \mu_{Dosis2mg} \geq 0 \}$

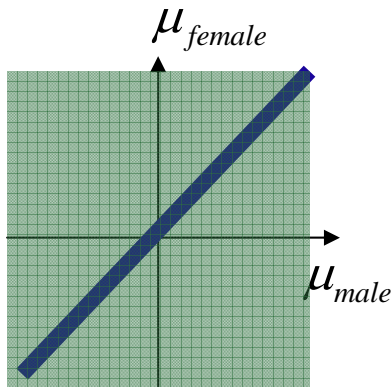
Always complementary regions!!

## 5.1 Basics: What is a hypothesis test?

Hypothesis are ALWAYS about parameter regions

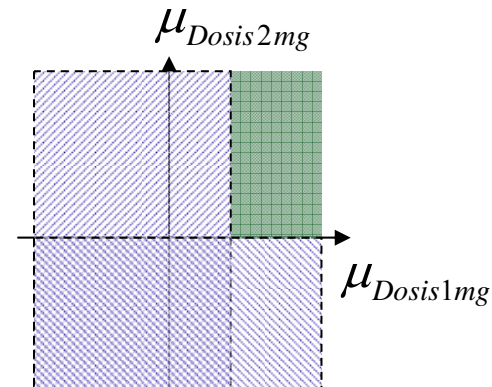
$$H_0 : \mu_{male} = \mu_{female}$$

$$H_1 : \mu_{male} \neq \mu_{female}$$



$$H_0 : \{ \mu_{Dosis1mg} \leq 2 \} \cup \{ \mu_{Dosis2mg} < 0 \}$$

$$H_1 : \{ \mu_{Dosis1mg} > 2 \} \cap \{ \mu_{Dosis2mg} \geq 0 \}$$



Hypothesis are NEVER about specific realizations of the random variable

Joe is a hypertense patient. Research hypothesis: Given our previous study, the effect of 1mg of the drug will have a positive effect larger than 2 on Joe.

Joe and Mary have been married for then 5 years. Research hypothesis: Joe will report higher intimacy level than Mary.

## 5.1 Basics: What is a hypothesis test?

Hypothesis NEVER use “Not all”, “Some”, “None”, “All”

Research hypothesis: All hypertense patients benefit from a new drug.

Research hypothesis: None hypertense patients benefit 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.

Research hypothesis: Some hypertense patients benefit from a new drug.

Problem: Too imprecise, being true does not provide much information



## 5.1 Basics: What can I do with hypothesis testing?

- 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 couples married for more than 5 years).

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)

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).

## 5.1 Basics: An example

An engineer works for Coca Cola. He knows that the filling machine has a variance of 6 cl. (the filling process can be approximated by a Gaussian). Knowing this, he sets the machine to a target fill of 348 cl ( $=330+3*6$ ). In a routine check with 25 cans, he measures an average of 345 cl. Is it possible that the machine is malfunctioning?

Step 1. Define your hypothesis

$$H_0 : \mu = 348$$

$$H_1 : \mu \neq 348$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad Z = \frac{\bar{x} - 348}{\frac{6}{\sqrt{25}}} \sim N(0,1)$$



## 5.1 Basics: An example

### Step 3. Plug-in observed data

$$z = \frac{345 - 348}{\frac{6}{\sqrt{25}}} = -2.5$$

### Step 4. Compute probabilities

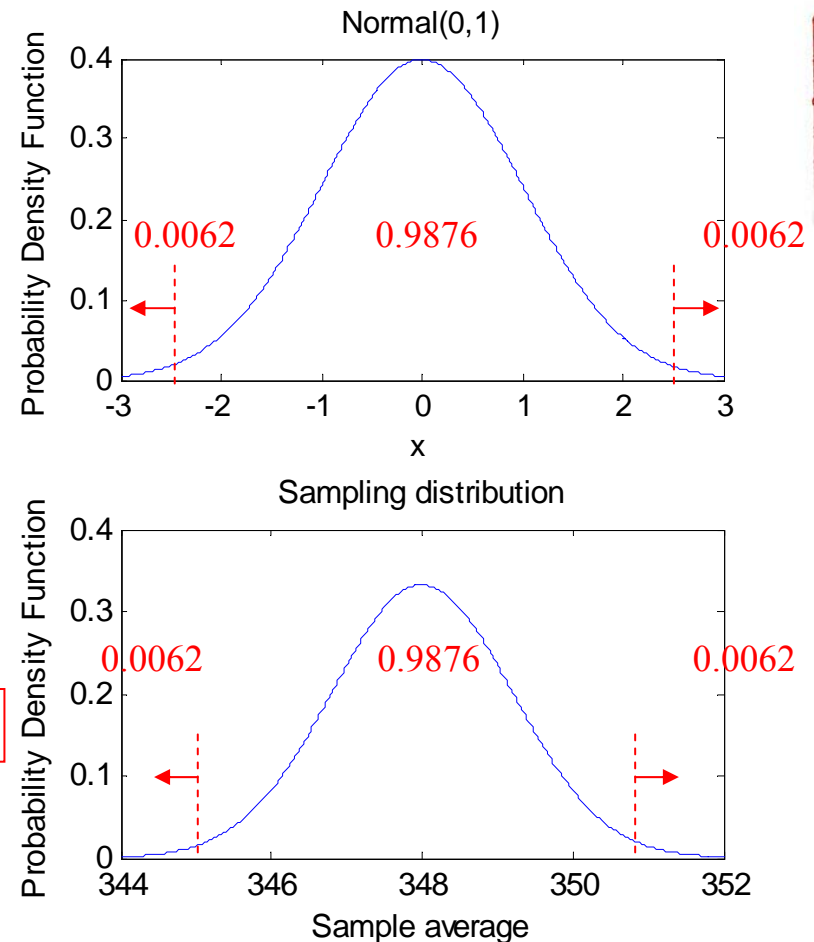
The probability of observing a sample average AT LEAST AS EXTREME AS THE OBSERVED is

$$P\{Z < -2.5\} + P\{Z > 2.5\} = 0.0062 \cdot 2 = 0.0124$$

### Step 5. Reject or not $H_0$

Extremes happen: normally in 1.24% of the cases I will observe a sample average that is at least as far from 348 as 345.

$p$ -value



$0.0124 < 0.05$  I will take the risk of not going for machine maintenance when I should have gone in 5% of the cases

$0.0124 > 0.01$  I will take the risk of not going for machine maintenance when I should have gone in 1% of the cases

## 5.1 Basics: Another example

Is it possible that the machine is filling less than programmed?



Step 1. Define your hypothesis

$$H_0 : \mu \geq 348$$

$$H_1 : \mu < 348$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad Z = \frac{\bar{x} - 348}{\frac{6}{\sqrt{25}}} \sim N(0,1)$$

Step 3. Plug-in observed data

$$z = \frac{345 - 348}{\frac{6}{\sqrt{25}}} = -2.5$$

Step 4. Compute probabilities

The probability of observing a sample average AT LEAST AS EXTREME AS THE OBSERVED is

$$P\{Z < -2.5\} = 0.0062$$

Step 5. Reject or not  $H_0$

Extremes happen: normally in 1.24% of the cases I will observe a sample average that is at least as far from 348 as 345.

$$0.0062 < 0.01$$

I will take the risk of going for machine maintenance when I should have not in 1% of the cases

## 5.1 Basics: Even another example

A country that has 4 ethnics (Balzacs 40%, Crosacs 25%, Murads 30%, Isads 5%) imposes by law that medical schools accept students proportionally to the ethnics. The distribution of 1000 students admitted this year in a medical school has been 300, 220, 400 and 80. Is the school respecting the law?

Step 1. Define your hypothesis

$$H_0 : O_{Balzacs} = E_{Balzacs} \cap O_{Crosacs} = E_{Crosacs} \cap O_{Murads} = E_{Murads} \cap O_{Isads} = E_{Isads}$$

$$H_1 : O_{Balzacs} \neq E_{Balzacs} \cup O_{Crosacs} \neq E_{Crosacs} \cup O_{Murads} \neq E_{Murads} \cup O_{Isads} \neq E_{Isads}$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad X = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-1}^2$$

Step 3. Plug-in observed data

$$\begin{aligned} x &= \frac{(300 - 1000 \cdot 40\%)^2}{1000 \cdot 40\%} + \frac{(220 - 1000 \cdot 25\%)^2}{1000 \cdot 25\%} + \frac{(400 - 1000 \cdot 30\%)^2}{1000 \cdot 30\%} + \frac{(80 - 1000 \cdot 5\%)^2}{1000 \cdot 5\%} = \\ &= 25 + 3.6 + 33.3 + 18 = 79.9 \end{aligned}$$

Step 4. Compute acceptance region

$$P\{X > 79.9\} \approx 0$$

Step 5. Reject or not  $H_0$

Reject!

## 5.1 Basics: Even another example

We want to know if there is any relationship between political affiliation and personality. We study 200 individuals obtaining the following data

	Democrat	Republican	Sum
Introvert	20	80	110 (55%)
Extrovert	50	40	90 (45%)
Sum	70 (35%)	120 (65%)	

Step 1. Define your hypothesis

$$H_0 : O_{Introvert, Democrat} = E_{Introvert, Democrat} \cap \dots \cap O_{Extrovert, Republican} = E_{Extrovert, Republican}$$

$$H_1 : O_{Introvert, Democrat} \neq E_{Introvert, Democrat} \cup \dots \cup O_{Extrovert, Republican} \neq E_{Extrovert, Republican}$$

Step 3. Plug-in observed data

$$\begin{aligned}
 x &= \frac{(30 - 200 \cdot 35\% \cdot 55\%)^2}{200 \cdot 35\% \cdot 55\%} + \frac{(80 - 200 \cdot 65\% \cdot 55\%)^2}{200 \cdot 65\% \cdot 55\%} + \frac{(50 - 200 \cdot 35\% \cdot 45\%)^2}{200 \cdot 35\% \cdot 45\%} + \frac{(40 - 200 \cdot 65\% \cdot 45\%)^2}{200 \cdot 65\% \cdot 45\%} = \\
 &= 1.88 + 1.01 + 10.87 + 5.85 = 19.61
 \end{aligned}$$

Step 4. Compute acceptance region  $P\{X > 19.61\} = 0.0002$

## 5.1 Basics: Even another example

We want to know if there is any relationship between political affiliation and personality. We study 200 individuals obtaining the following data

Step 1. Define your hypothesis

$$H_0 : O_{Balzacs} = E_{Balzacs} \cap O_{Crosacs} = E_{Crosacs} \cap O_{Murads} = E_{Murads} \cap O_{Isads} = E_{Isads}$$

$$H_1 : O_{Balzacs} \neq E_{Balzacs} \cup O_{Crosacs} \neq E_{Crosacs} \cup O_{Murads} \neq E_{Murads} \cup O_{Isads} \neq E_{Isads}$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad X = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

Step 3. Plug-in observed data

$$\begin{aligned} x &= \frac{(300 - 1000 \cdot 40\%)^2}{1000 \cdot 40\%} + \frac{(220 - 1000 \cdot 25\%)^2}{1000 \cdot 25\%} + \frac{(400 - 1000 \cdot 30\%)^2}{1000 \cdot 30\%} + \frac{(80 - 1000 \cdot 5\%)^2}{1000 \cdot 5\%} = \\ &= 25 + 3.6 + 33.3 + 18 = 79.9 \end{aligned}$$

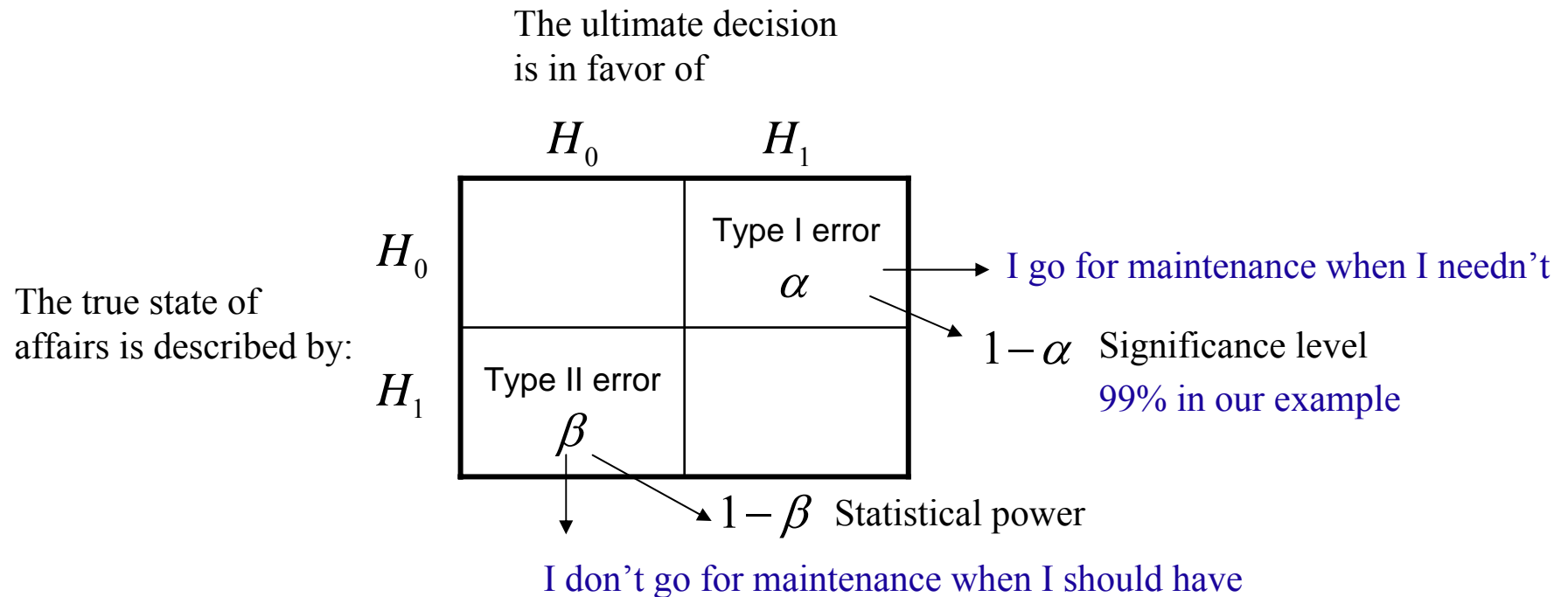
Step 4. Compute acceptance region

$$P\{X > 79.9\} \approx 0$$

Step 5. Reject or not  $H_0$

Reject!

## 5.1 Basics: Decision making



Decision Rule: If  $p - value < \alpha$ , then reject  $H_0$   
Otherwise, you cannot reject  $H_0$

Drawback: it is exclusively driven by Type I errors



## 5.1 Basics: Type I errors and Statistical significance

We can think of  $\alpha$  in different ways:

1. The probability of committing the error of rejecting the null hypothesis when it is true. The smaller  $\alpha$ , the better. Normal values are 0.05, 0.01, 0.005, 0.001.
2. The ability of the sample to be generalized to the population. This is measured by  $1 - \alpha$  (statistical significance). Again, the smaller the  $\alpha$ , the better. Normal significances are 95%, 99%, 99.5%, 99.9%.
3. Statistical significance is not substantial significance (a statistically significant difference in salary of 1\$ is not substantially different, for instance, it does not affect your annual expenditure in food).

But, it IS NOT:

1. A hint that our finding is very important, it simply means that it would be rare to observe if the null hypothesis is true.
2. The statistical significance is not the probability of being able to repeat the study.

## 5.1 Basics: Type II errors and Statistical power

The true state of affairs is  $H_1 : \mu < 348$  but we don't have enough evidence to reject  $H_0 : \mu \geq 348$

It is impossible to compute  $\beta$ , but we can understand its behaviour

With  $\alpha = 0.01$  we cannot reject  $H_0$  if  $z > z_{0.01} = -2.33$

Let us suppose that  $\mu = 347$

$$\begin{aligned}\beta &= \Pr\{z > -2.33\} = 1 - \Pr\{z \leq -2.33\} = 1 - \Pr\left\{\frac{\bar{x} - 348}{\frac{6}{\sqrt{25}}} \leq -2.33\right\} = 1 - \Pr\left\{\bar{x} \leq 348 - 2.33 \frac{6}{\sqrt{25}}\right\} \\ &= 1 - \Pr\{\bar{x} \leq 345.2\} = 1 - \Pr\left\{\zeta \leq \frac{345.2 - 347}{\frac{6}{\sqrt{25}}}\right\} = 1 - \Pr\{\zeta \leq -1.493\} = 0.9323\end{aligned}$$

$\bar{x} \sim N\left(347, \frac{6^2}{25}\right); \zeta \sim N(0,1)$

## 5.1 Basics: Type II errors and Statistical power

The true state of affairs is  $H_1 : \mu < 348$  but we don't have enough evidence to reject  $H_0 : \mu \geq 348$

$N = 25$

$$\mu = 347.5 \Rightarrow \beta = 0.9719$$

$$\mu = 347 \Rightarrow \beta = 0.9323$$

$$\mu = 345 \Rightarrow \beta = 0.4323$$

$N = 100$

$$\mu = 347.5 \Rightarrow \beta = 0.9323$$

$$\mu = 347 \Rightarrow \beta = 0.7453$$

$$\mu = 345 \Rightarrow \beta = 0.0038$$

The further the underlying parameter is from our supposed value, the easier for us to prove that  $H_0$  is false. Or in other words, we have less probability of not being able to reject  $H_0$ .

The more samples we have, the easier to prove that  $H_0$  is false. We say in this case that the test with 100 samples is more powerful than the test with 25 samples.

## 5.1 Basics: Type II errors and Statistical power

We can think of  $\beta$  in different ways:

1. The probability of committing the error of not rejecting the null hypothesis when it is false. The smaller  $\beta$ , the better. It cannot be easily computed.
2. Since the hypothesis testing is more focused in Type I errors, Type II errors are very common.
3. That is, research hypothesis are not accepted unless there is much evidence against the null hypothesis (hypothesis testing is rather conservative).
4.  $1 - \beta$  is the statistical power and it measures the ability of the test to detect relationships between the variables, changes in parameters, etc. (for instance, detecting a filling mean of 347.5 instead 348). A statistical power of 80% means that 80% of the times the experimenter will be able to find a relationship that really exist.
5. Small changes and small correlations are difficult to detect with small sample sizes.

## 5.1 Basics: What is the relationship between effect size, sample size, significance and power.

As the effect size increases:

1. Statistical power will increase for any given sample size.
2. We need less samples to achieve a certain statistical significance.
3. The probability of Type I errors (wrong rejection) decreases.

As the sample size increases:

1. Statistical power will increase.
2. The sampling error will decrease.
3. The probability of Type II errors (not being able to reject) decreases.

As the sampling error decreases:

1. Statistical power will increase for any given effect size.
2. We need less samples to achieve the same statistical significance.
3. The probability of Type II errors (not being able to reject) decreases.

## 5.1 Basics: Relationship to confidence intervals

Our engineer wants to know if the modification introduces an extra variance in the motors

Unmodified motor (Watts): 755 750 730 731 743  $s_X = 11.17$

Modified motor (Watts): 742 738 723 721 730  $s_Y = 9.28$

$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left[ \frac{\frac{11.17^2}{9.28^2}}{9.6}, \frac{\frac{11.17^2}{9.28^2}}{0.1} \right] = \left[ 0.39^2, 3.8^2 \right]_{\alpha = 0.05}$$

$$\begin{array}{ccc} H_0 : \sigma_X^2 = \sigma_Y^2 & \longrightarrow & H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 & & H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1 \end{array}$$

If the confidence interval includes the null hypothesis, then we cannot reject  $H_0$ .  
If the confidence interval does not include the null hypothesis, then we reject  $H_0$ .

## 5.1 Basics: Relationship to confidence intervals

Do not misinterpret confidence intervals, overlapping confidence intervals do not mean that we cannot reject the null hypothesis.

Our engineer wants to know if the modification introduces an extra variance in the motors

Unmodified motor (Watts):	755	750	740	741	743	$\bar{x} = 745.8$	$s_x = 6.45$
Modified motor (Watts):	732	738	723	721	730	$\bar{y} = 728.8$	$s_y = 6.91$
						$\bar{d} = -17$	

$$\mu_X \in \bar{x} \pm t_{N_X-1, \frac{\alpha}{2}} \frac{s_X}{\sqrt{N_X}} \longrightarrow \mu_X \in [727.9, 763.7]$$

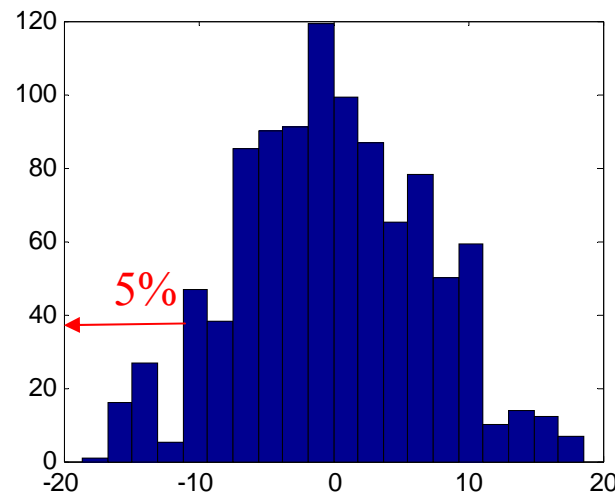
$$\mu_Y \in [709.6, 748.0]$$

$$\frac{\bar{d} - \mu_d}{\sqrt{\frac{s_X^2}{N} + \frac{s_Y^2}{N}}} \sim t_{2N-2} \longrightarrow \mu_d \in [-26.8, -7.2] \longrightarrow \text{Reject } H_0 \text{ because } \mu_d = 0 \text{ does not belong to the confidence interval}$$

## 5.1 Basics: Relationship to confidence intervals (Permutation/Randomization tests)

Our engineer wants to know if the modification reduces the consumption

	Unmodified					Modified					Diff
Actual measurement	755	750	730	731	743 (741.8)	742	738	723	721	730 (730.8)	-11
Permutation 1	742	731	721	755	730 (735.8)	750	723	730	743	738 (736.8)	1
Permutation 2	742	721	750	730	755 (739.6)	743	730	731	723	738 (733.0)	-6.6
Permutation 3	750	755	742	730	743 (744.0)	721	723	731	730	738 (728.6)	-15.4
...											



$$\Pr\{\hat{\mu} < -11\} = 0.05$$

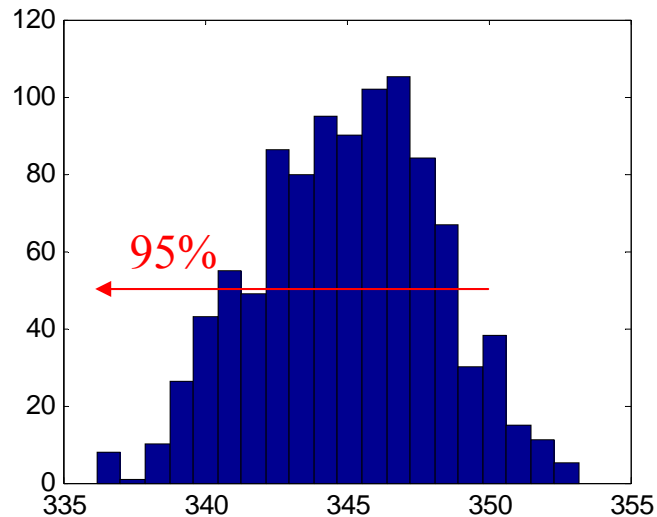


## 5.1 Basics: Relationship to confidence intervals (Bootstrap tests)

Is it possible that the machine is filling less than programmed?

Data:

341 335 354 345 350 → 354 350 350 354 350 → 351.6  
341 335 345 341 335 → 339.4  
335 335 354 335 335 → 338.8  
341 345 354 350 354 → 348.8  
...



$$H_0 : \mu \geq 348 \longrightarrow \Pr\{\hat{\mu} \geq 348\} = 0.166$$
$$H_1 : \mu < 348$$

Reject  $H_0$  if the confidence interval at confidence level  $1-\alpha$  does not intersect the null region.

## 5.1 Basics: Bayesian hypothesis tests

### Classical Hypothesis Testing

25 hypertense patients were given 1mg during 1 month of a new drug.

Research hypothesis: The treatment is effective against hypertension

$H_0$  : *Opposite of  $H_1$*

$H_1$  : *Some condition on model parameters*

### Bayesian Hypothesis Testing

25 hypertense patients were given 1mg during 1 month of a new drug.

Research hypothesis: The treatment is effective against hypertension knowing that 50% of the drugs tested are usually ineffective.

$H_0$  : *Data generated by model<sub>0</sub>*

$H_1$  : *Data generated by model<sub>1</sub>*

Also known as likelihood ratio tests.

## 5.1 Basics: Examples



Research hypothesis: the majority of people overinflate car tires

Research hypothesis: the more people go to cinema, the happier they are

Research hypothesis: There is a difference in the performance in mathematics among students among the different countries in the EU.

Research hypothesis: There is NO difference in the performance in mathematics among students among the different countries in the EU.

## 5.2 What are the assumptions of hypothesis tests?

- Ideally the violation of the test assumptions invalidate the test result.
- However, there are tests that are more robust than others to violations.
- And the same test may be more robust to violations in one assumption than another.

That's why it is better to know the assumptions about:

- The population distribution
- Residuals
- The nature of the sample
- Measurement errors
- The model being tested.

## 5.2 What are the assumptions about the population?

### ➤ Parametric tests:

- usually make an assumption about the population (normally, normality).
  - o Tests: Shapiro-Wilks (1D), Kolmogorov-Smirnov(1D), Smith-Jain (nD)
- Samples are independent and identically distributed
- Homoscedasticity is also usually assumed :
  - o Regression: the variance of the dependent variables is the same across the range of predictor variables
  - o ANOVA-like: the variance of the variable studied is the same among groups
  - o Tests: Levene, Breusch-Pagan, White
- From these assumptions the statistic distribution is derived.
- Normally applied to ratio/interval variables.

### ➤ Nonparameteric tests:

- Don't make any "parametric assumption"
- Samples are independent and identically distributed
- Normally applied to ordinal/categorical variables
- Permutation tests assume that labels can be permuted

## 5.2 What are the assumptions about the population?

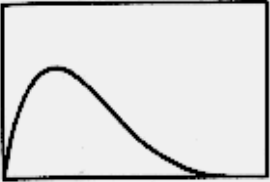
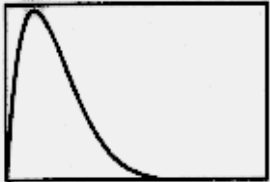

### Normality

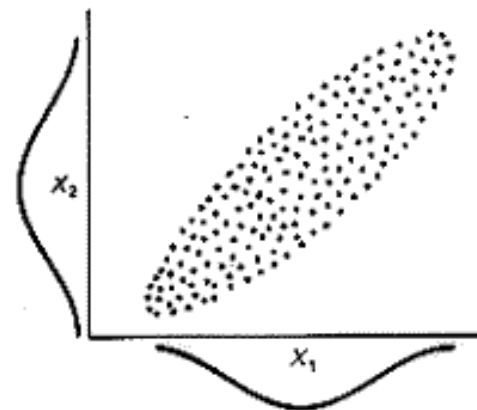
The solution to most assumption violations is provided by data transformations.

### Homoscedasticity

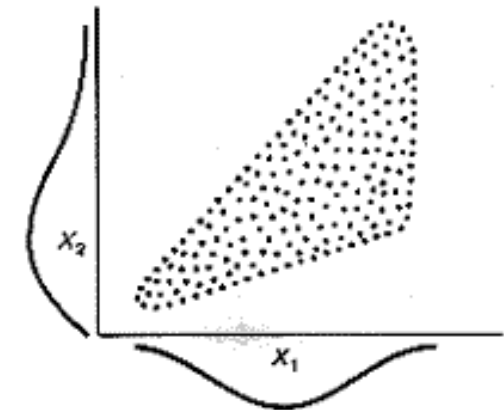
Sometimes can be detected by bare eye.

Table of sample pdfs and suggested transformation

Form	Transformation
	Square Root $Y = \sqrt{X}$
	Logarithm $Y = \log X$
	Inverse $Y = \frac{1}{X}$



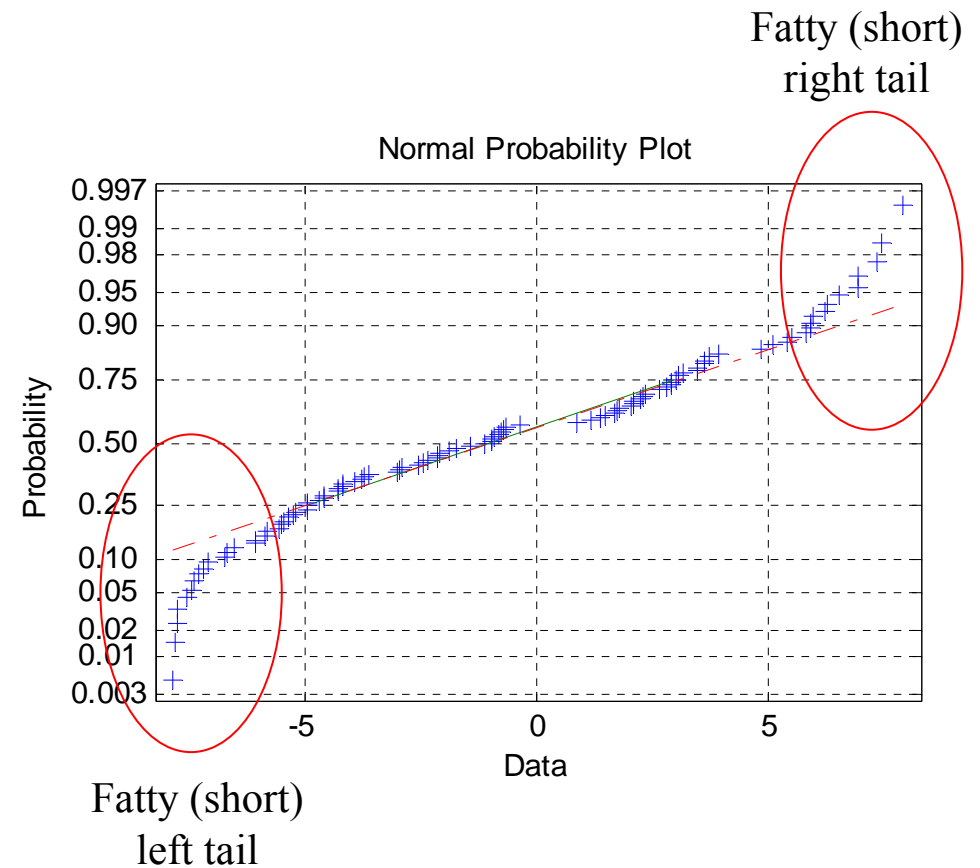
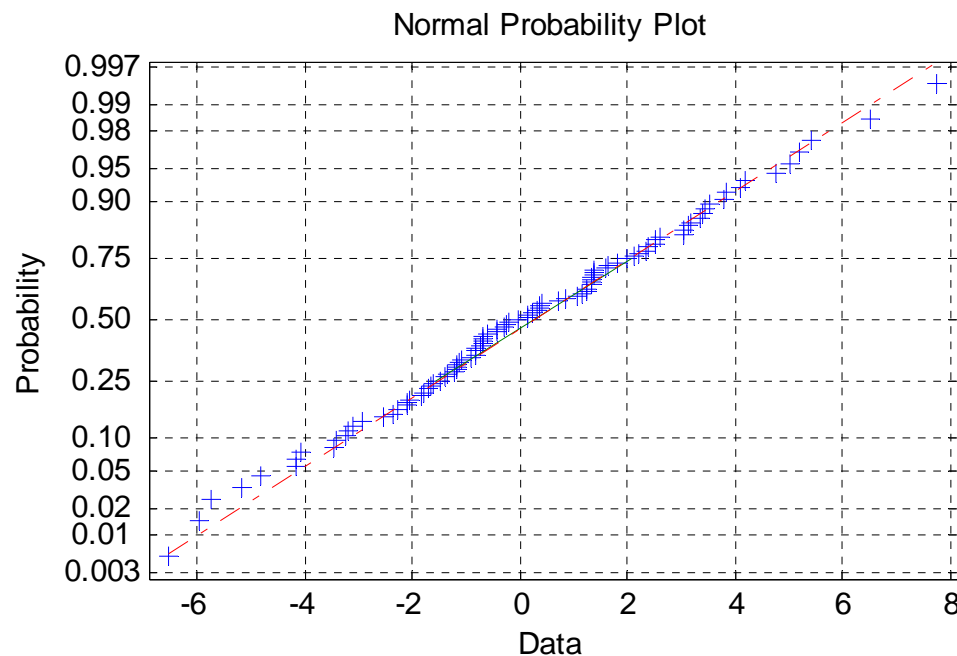
Homoscedasticity with both variables normally distributed



Heteroscedasticity with skewness on one variable

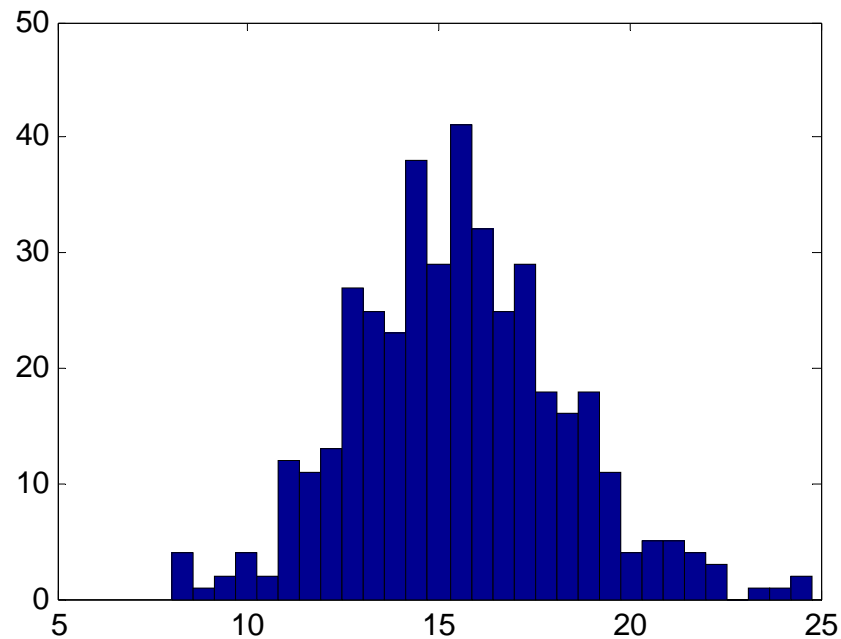
## 5.2 What are the assumptions about the population?

### Normality: Probability plots



## 5.2 What are the assumptions about the population?

### Normality: Probability plots





## 5.2 What are the assumptions about the population?

Data: Number of people entering a shop within 15 minutes

Data: Electrical power consumption in a city (consider blackouts from time to time)

Data: Yes-No answers of a questionnaire

Data: Number of people entering the underground at peak hour

Data: The number of children as a function of age



## 5.2 What are the assumptions about the residuals?

$$\text{SystolicPressure} = 163.7 + \alpha_{\text{treatment}} + \varepsilon$$

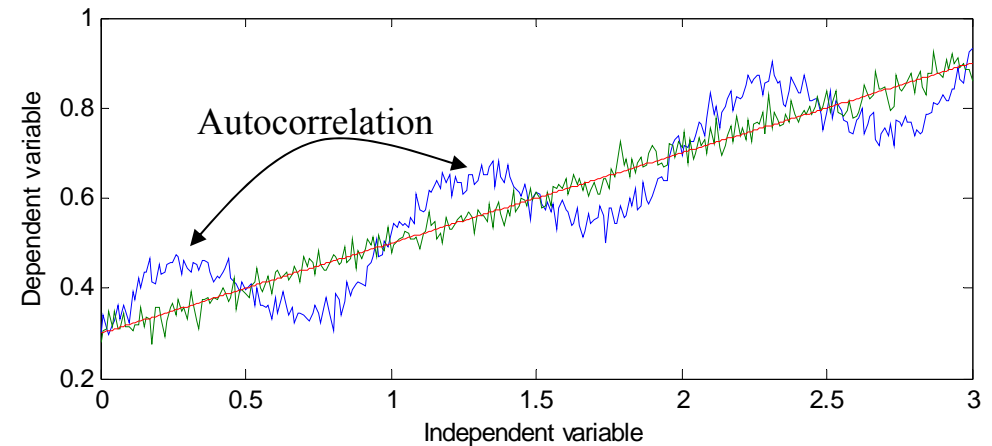
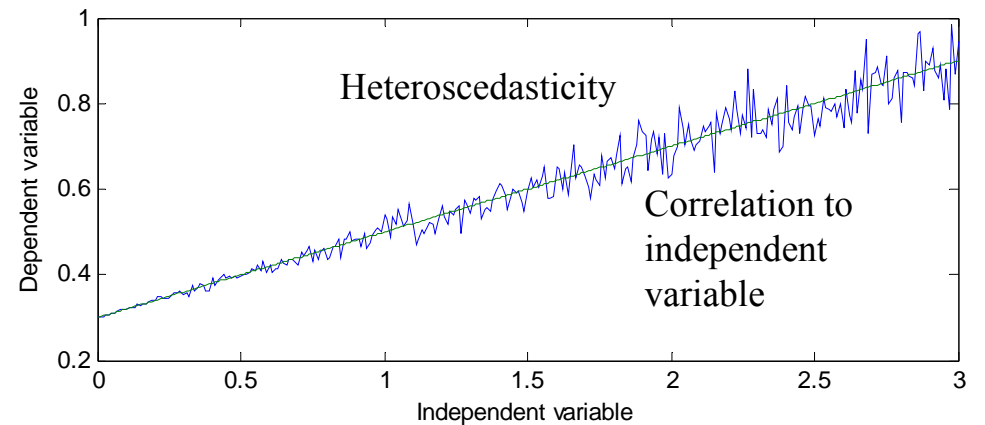
$$\text{Score} = \mu_{\text{RealMadrid}} + \alpha_{\text{strategy}} + \alpha_{\text{Raúl}} + \alpha_{\text{Raúl, strategy}} + \varepsilon$$

### ➤ Residuals

- Zero mean
- Homoscedasticity
- No autocorrelation
- Not correlated to the independent variable

### ➤ Solution:

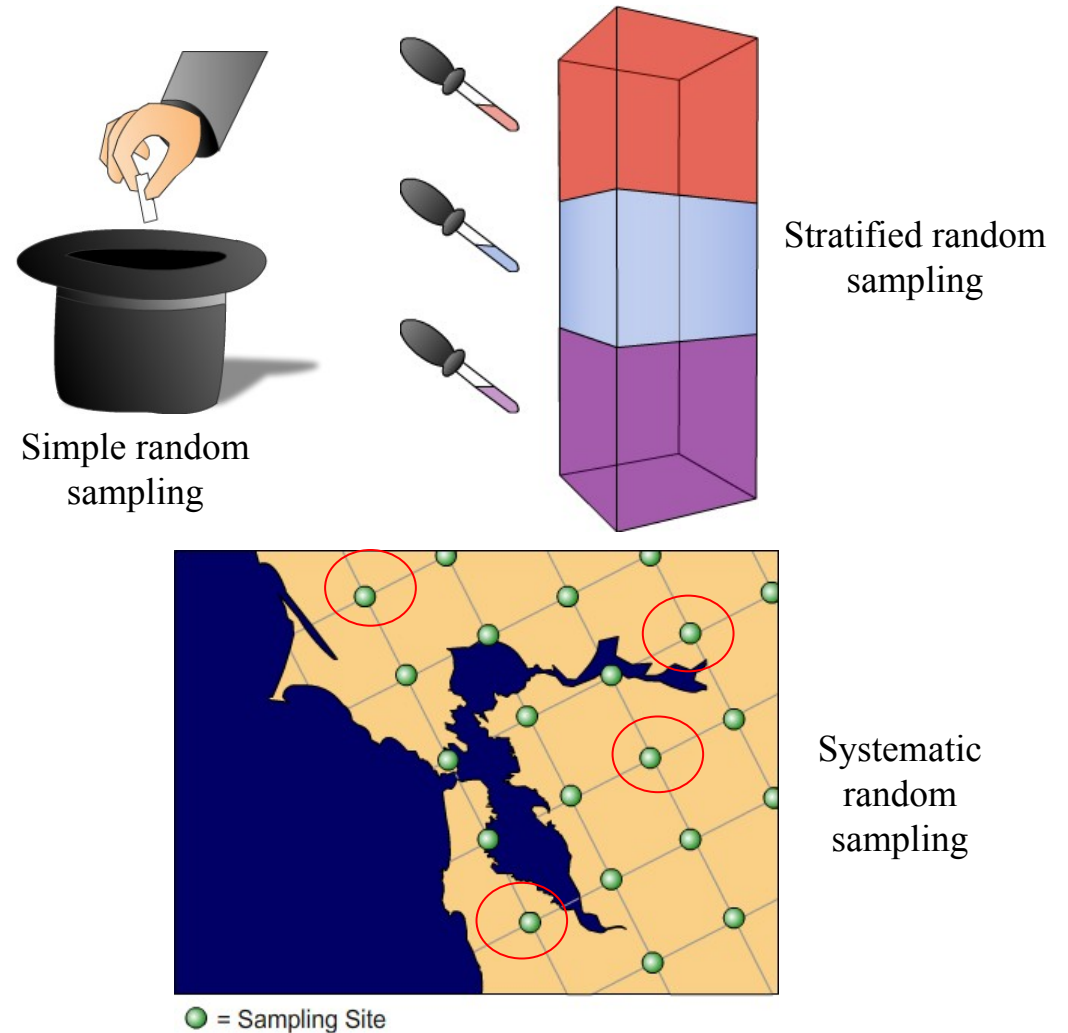
- Change your model
- Use non parametric test



## 5.2 What are the assumptions about the nature of the sample?

- Sample
  - Simple random sampling
  - Samples are independent
- Solution:
  - Random sampling but not simple → readjust variances
  - Sampling not random → inference cannot be used! You cannot generalize to the whole population

Sampling: A study carried on volunteers  
**Violates random sampling**



## 5.2 What are the assumptions about variables, measurements, and models?

### ➤ Nature of variables

- Categorical: Sex
- Ordinal: Degree of preference
- Interval: Temperature
- Ratio: Length

### ➤ Measurements

- Regression assumes that the independent variable (X) is measured without error, but the dependent variable is measured with error. If there is error in X, then association measures are underestimated.
- Check if your measurements are reliable.

### ➤ Models

- Many models assume linearity (for instance, the correlation coefficient)
- We assume that our model fully explain the phenomenon

*Purchase Warm Clothes* ↓ —————→ *Purchase Ice Creams* ↑

## 5.3 How to select the appropriate test? Central tendency

### ➤ Interval/ratio variables

- ❖ One independent variable
  - ❑ Single sample: Hypothesis about population mean
    - Z-test and t-test
  - ❑ Two samples: Hypothesis about differences between 2 population means
    - Independent samples: z-test and t-test, 1-factor between-subjects ANOVA and ANCOVA
    - Dependent samples: z-test, t-test, Sandler's A test, 1-factor within-subjects ANOVA
  - ❑ Two or more samples
    - Independent samples: 1-factor between-subjects ANOVA and ANCOVA
    - Dependent samples: 1-factor within-subjects ANOVA
- ❖ Two independent variables: Hypothesis about differences between 2 population means
  - Independent samples: Factorial between-subjects ANOVA
  - Dependent samples: Factorial within-subjects ANOVA

## 5.3 How to select the appropriate test? Central tendency

### ➤ Ordinal/rank order variables

- ❑ Single sample: Hypothesis about population median
  - Wilcoxon signed-rank test, binomial sign test
- ❑ Two samples: Hypothesis about differences between 2 population medians
  - Independent samples: Mann-Whitney U test; permutation, bootstrap, jackknife tests; median test (chi-square)
- ❑ Two or more samples: are the median different in two or more samples?
  - Independent samples: Kruskal-Wallis 1-way ANOVA by ranks, Jonckheere-Terpstra test for ordered alternatives
  - Dependent samples: Friedman 2-way ANOVA by ranks, Page test for ordered alternatives

## 5.3 How to select the appropriate test? Variability

### ➤ Interval/ratio variables

#### ❖ One independent variable

- ❑ Single sample: Hypothesis about population variance
  - Chi-square
- ❑ Two samples: Hypothesis about differences between 2 population variances
  - Independent samples: Snedecor's F test, Hartley's  $F_{\max}$
  - Dependent samples: t test
- ❑ Two or more samples
  - Independent samples: Sphericity test in ANOVA

### ➤ Ordinal/rank order variables

#### ❖ One independent variable

- ❑ Two samples: Hypothesis about differences between 2 population variances
  - Independent samples: Siegel-Tukey test, Moses test

## 5.3 How to select the appropriate test? Distributions

### ➤ Interval/ratio variables

#### ❖ One independent variable

##### ☐ Single sample: Hypothesis about the data distribution

- General: Kolmogorov-Smirnov goodness-of-fit
- Normality: Lilliefors, D'Agostino-Pearson

##### ☐ Two samples: Hypothesis about differences between 2 population distributions

- Independent samples: Kolmogorov-Smirnov, Wilcoxon U test

##### ☐ Two or more samples

- Independent samples: van der Waerden normal scores

### ➤ Ordinal/Rank order variables

#### ❖ One independent variable

##### ☐ Single sample: Hypothesis about the data distribution

- Kolmogorov-Smirnov goodness-of-fit

##### ☐ Two samples: Hypothesis about differences in the data distribution

- Dependent samples: Wilcoxon matched-pairs signed ranks, binomial sign test



## 5.3 How to select the appropriate test? Distributions

### ➤ Categorical variables

- ❖ One independent variable
  - ❑ Single sample: Hypothesis about the data distribution
    - General: Chi-square goodness-of-fit
    - Binomial: Binomial sign test, z-test for population proportion
  - ❑ Two samples: Hypothesis about differences between 2 population distributions
    - Independent samples: Chi-square test for homogeneity, Fisher's exact test, z-test for two independent proportions
    - Dependent samples: McNemar test
- ❖ Two independent variables: Are two variables really independent?
  - ❑ Single sample:
    - Chi-square test for independence
  - ❑ Two samples: differences between 2 population distributions
    - Independent samples: Chi-square test for homogeneity, Fisher's exact test
    - Dependent samples: McNemar test, Gart test for order effects, Bowker test of internal symmetry, Stuart-Maxwell test of marginal homogeneity
  - ❑ Two or more samples: differences between several population distributions
    - Independent samples: Chi-square test for homogeneity
    - Dependent samples: Cochran's Q test

## 5.3 How to select the appropriate test? Randomness

ABABABABAB,AAAAABBBBBB are not random

### ➤ Interval/ratio variables

#### ❖ One independent variable

##### ❑ Single sample series:

- Single-sample runs (only for the sign)), mean square successive difference test, autocorrelation tests (correlation test, Durbin-Watson test, Cox-Stuart test for trend)

##### ❑ Two samples: Hypothesis about differences between 2 population distributions

- Independent samples:
- Dependent samples:

##### ❑ Two or more samples

- Independent samples:
- Dependent samples:

### ➤ Categorical variables

#### ❖ One independent variable

##### ❑ Single sample series: Is a series of symbols random?

- Single-sample runs, frequency test, gap test, poker test, maximum test, coupon's collector test

## 5.3 How to select the appropriate test? Association

- Interval/ratio variables
  - ❖ Two independent variables: Pearson correlation test, t-test
  - ❖ More than two variables: multiple correlation coefficient, partial correlation coefficient, semipartial correlation coefficient
- Ordinal/rank order variables
  - ❖ Two variables/Sets of ranks: Spearman's rho, Kendall's tau, Goodman and Kruskal's gamma
  - ❖ More than two variables/Sets of ranks: Kendall's coefficient of concordance
- Categorical variables
  - ❖ Two binary variables: contingency coefficient, phi coefficient, Yule's Q test, odds ratio, Cohen's kappa, binomial effect size display
  - ❖ Two non-binary variables: contingency coefficient, Cramér's phi, odds ratio, Cohen's kappa.
- Interval/ratio variable and categorical variable
  - ❖ Non-binary categorical: Omega squared, eta squared, Cohen's f index
  - ❖ Binary categorical: Cohen's d index, point-biserial correlation

## 5.3 How to select the appropriate test?

Situation: A librarian wants to know if it is equally likely that a person takes a book any of the 6 days (Monday to Saturday) that the library is open. She records the following frequencies: Monday, 20; Tuesday, 14; Wednesday, 18; Thursday, 17; Friday, 22; and Saturday, 29.



## 5.3 How to select the appropriate test?

Situation: The instruction manual of a hearing aid device claims that the average life of the battery is 7 hours with a standard deviation of 2.23 hours. A customer thinks that this standard deviation is too low. In order to test this hypothesis he purchases 10 batteries and records the following average lives: 5, 6, 4, 3, 11, 12, 9, 13, 6, 8



## 5.3 How to select the appropriate test?

Situation: A physician states that the median number of times he sees each of his patients during the year is 5. He randomly selects 10 of his patients and determines the number of times each one of them visited him (9, 10, 8, 4, 8, 3, 0, 10, 15, 9). Do the data support his statement?



## 5.3 How to select the appropriate test?

Situation: 10 women are asked to judge which of two perfume brands has better fragrance. 8 women selected brand A, while 2 selected brand B. Is there a significant difference with respect to the preference for the perfumes?



## 5.3 How to select the appropriate test?

Situation: 10 clinically depressed patients are preevaluated and assessed as equally depressed. They are randomly assigned to two groups (5 in each group). One of the groups are given a new antidepressant drug, while the other group is given a placebo. After 6 months of treatments, a psychiatrist reevaluates the patients (the doctor does not know who received the placebo and who received the drug). The evaluation of each group is (11, 1, 0, 2, 0) and (11, 11, 5, 8, 4). Is the new drug effective?





## 5.3 How to select the appropriate test?

Situation: 10 clinically depressed patients are preevaluated and assessed as equally depressed. They are randomly assigned to two groups (5 in each group). One of the groups are given a new antidepressant drug, while the other group is given a placebo. After 6 months of treatments, a psychiatrist reevaluates the patients (the doctor does not know who received the placebo and who received the drug). The evaluation of each group is (11, 1, 0, 2, 0) and (11, 11, 5, 8, 4). Is the new drug effective?



## 5.3 How to select the appropriate test?

Situation: 10 clinically depressed patients are preevaluated and assessed as equally depressed. They are randomly assigned to two groups (5 in each group). One of the groups are given a new antidepressant drug, while the other group is given a placebo. After 6 months of treatments, a psychiatrist reevaluates the patients (the doctor does not know who received the placebo and who received the drug). The evaluation of each group is (11, 1, 0, 2, 0) and (11, 11, 5, 8, 4). Is the new drug effective?



## 5.3 How to select the appropriate test?

Situation: 10 clinically depressed patients are preevaluated and assessed as equally depressed. They are randomly assigned to two groups (5 in each group). One of the groups are given a new antidepressant drug, while the other group is given a placebo. After 6 months of treatments, a psychiatrist reevaluates the patients (the doctor does not know who received the placebo and who received the drug). The evaluation of each group is (10, 10, 9, 1, 0) and (6, 6, 5, 5, 4, 4). With this data we cannot discard that the two means are the same. Is it possible that the new drug increases the depression in some patients and decreases it in some others?



## 5.4 Multiple testing

$$H_0 : p_{\text{Head}} = 0.5$$



...



The ultimate decision  
is in favor of

$H_0$

$H_1$

The true state of  
affairs is described by:

$H_0$

$H_1$

	$H_0$	$H_1$
$H_0$		Type I error $\alpha = 0.05$
$H_1$	Type II error	

$p(\text{Type I error})$	$p(\text{Correct})$
--------------------------	---------------------

$\alpha$

$1 - \alpha$

$\alpha$

$1 - \alpha$

Type I error of a single test

$\alpha$

$1 - \alpha$

$1 - (1 - \alpha)^N$

$(1 - \alpha)^N$

0.994

0.006

Type I error of the whole family

I decide the coin is biased when it is not.  
Expected: 5 (=5%\*100) coins

$$H_0 : \mu_{\text{drug}} = \mu_{\text{placebo}} !!$$

## 5.4 Multiple testing

$$p(\text{Type I family error}) = 1 - (1 - \alpha)^N$$

- Choose a smaller  $\alpha$  or equivalently, recompute the p-values of each individual test
- Fix the False Discovery Rate and choose a smaller  $\alpha$
- Compute empirical p-values via permutation tests or bootstrapping

Bonferroni: Choose a smaller  $\alpha$ .

$$\alpha_{used} = \frac{\alpha_{desired}}{N} \quad \alpha_{used} = \frac{0.05}{100} = 0.0005 \quad 1 - (1 - \alpha_{used})^N = 0.0488 \quad (1 - \alpha_{used})^N = 0.9512$$

Problems:  $\alpha_{used} \downarrow \Rightarrow SampleSize \uparrow$

Too conservative when tests are not independent (e.g., genes from the same person)

Acceptance of a single tests depends on the total number of tests!

The underlying assumption is that all null hypothesis are true (likely not to be true)

Low family power (not rejecting the family null hypothesis when it is false)

## 5.4 Multiple testing

### Recompute p-values

$$\begin{aligned}
 p_{\text{Bonferroni}} &= \min(Np_{\text{value}}, 1) \\
 p_{\text{Bonferroni-Holm}}^{(i)} &= \min((N - (i - 1))p_{\text{value}}^{(i)}, 1) \\
 p_{\text{FDR}}^{(i)} &= \min\left(\frac{N}{i} p_{\text{value}}^{(i)}, 1\right)
 \end{aligned}
 \left. \vphantom{\begin{aligned} p_{\text{Bonferroni}} \\ p_{\text{Bonferroni-Holm}}^{(i)} \\ p_{\text{FDR}}^{(i)} \end{aligned}} \right\} \text{Single-step procedures: each p-value is corrected individually.}$$

	Sequence	p-value	Bonferroni	Bonferroni-Holm	FDR
Sorted in ascending p-value ↓	1 0 1 1 1 1 1 1 1 1	0.0098	0.9800	0.0098*100=0.9800	0.0098*100/ 1=0.9800
	1 0 0 0 0 0 0 0 0 0	0.0098	0.9800	0.0098* 99=0.9668	0.0098*100/ 2=0.4883
	1 1 1 1 1 1 0 1 1 1	0.0098	0.9800	0.0098* 98=0.9570	0.0098*100/ 3=0.3255
	1 1 1 1 1 1 0 1 1 1	0.0098	0.9800	0.0098* 97=0.9473	0.0098*100/ 4=0.2441
	1 1 1 1 1 1 0 0 1 1	0.0439	1.0000	0.0439* 96=1.0000	0.0439*100/ 5=0.8789
	0 1 1 1 1 1 1 1 1 0	0.0439	1.0000	0.0439* 95=1.0000	0.0439*100/ 6=0.7324
	1 1 0 1 1 1 1 1 0 1	0.0439	1.0000	0.0439* 94=1.0000	0.0439*100/ 7=0.6278
	0 1 0 1 1 1 1 1 1 1	0.0439	1.0000	0.0439* 93=1.0000	0.0439*100/ 8=0.5493
	1 0 0 1 0 0 0 0 0 0	0.0439	1.0000	0.0439* 92=1.0000	0.0439*100/ 9=0.4883
	1 0 1 1 1 1 1 1 1 0	0.0439	1.0000	0.0439* 91=1.0000	0.0439*100/10=0.4395
	1 0 1 1 1 1 0 1 1 1	0.0439	1.0000	0.0439* 90=1.0000	0.0439*100/11=0.3995
	1 1 1 1 1 0 1 1 1 0	0.0439	1.0000	0.0439* 89=1.0000	0.0439*100/12=0.3662
	1 1 1 0 1 1 1 1 0 1	0.0439	1.0000	0.0439* 88=1.0000	0.0439*100/13=0.3380
	1 1 0 0 1 1 1 1 1 1	0.0439	1.0000	0.0439* 87=1.0000	0.0439*100/14=0.3139
	1 0 1 0 0 1 1 1 1 1	0.1172	1.0000	0.1172* 86=1.0000	0.1172*100/15=0.7813
	...				

## 5.4 Multiple testing

False Discovery Rate: Choose a smaller  $\alpha$ .

$$P(H_0 | p \leq \alpha) = \frac{P(p \leq \alpha | H_0)P(H_0)}{P(p \leq \alpha | H_0)P(H_0) + P(p \leq \alpha | H_1)P(H_1)} = \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

$$\alpha_{used} = \underbrace{\frac{P(H_0 | p \leq \alpha)}{1 - P(H_0 | p \leq \alpha)}}_{\text{FDR}} \frac{1 - \pi_0}{\pi_0} (1 - \beta)$$

Proportion of tests that follows the null hypothesis

Proportion of false positives

Power to detect tests following the alternative hypothesis

Proportion of tests following the alternative hypothesis

Example: FDR=0.05,  $\pi_0 = 0.9$ ,  $\beta = 0.3$

$$\alpha_{used} = 0.05 \frac{1-0.9}{0.9} (1-0.3) = 0.0039$$

## 5.4 Multiple testing

- What about prefiltering experiments (according to intensity, variance etc.) to reduce the proportion of false positives - e.g. tests with consistently low intensity may not be considered interesting?
- Can be useful, but:
  - The criteria for filtering have to be chosen before the analysis, i.e. not dependent on the results of the analysis.
  - The criteria have to be independent of the distribution of the test statistic under the null hypothesis - otherwise no control of the type I error.



## 5.5 Words of caution

- Use two-sided p-values if you expect deviations from the null hypothesis in any direction (different from).
- Use one-sided p-values if you expect deviations from the null hypothesis in only one direction (better than, worse than).
- Wrong: Design a test after collecting data (probably you will pick the hypothesis that this data support).
- Right: Design a test before collecting data (the data will tell you if your hypothesis is confirmed or not).
- Wrong: Decide whether to take a two-sided p-value or one-sided p-value after collecting data.
- Right: Decide whether to take a two-sided p-value or one-sided p-value before collecting data.
- Wrong: Run several tests and pick the one supporting your beliefs.
- Right: Decide beforehand which will be the test to run.
- Right: Perform preliminary analysis (two-sided). Then, design a new experiment with the corresponding one-sided hypothesis if necessary.
- Use control groups: You find that among 10 patients of pancreas cancer, 9 of them drink coffee. Drinking coffee causes pancreas?



## 5.5 Words of caution

Unequal variances: What can I do if I'm comparing the mean of a variable in two groups and the two groups have different variances?

Problem: Permutation and classical test will not be accurate.

Solutions:

- Apply a transformation (e.g. square root, arcsine) so that the variance are more equal
- Compare medians (which are less affected by the variance)
- Determine why the variances are different.

Possible reason: systematic bias → use proper experiment design (randomization)

## 5.5 Words of caution

Dependent observations: What can I do if my observations are dependent?

Problem: Tests will not be accurate.

Solutions:

- Use multivariate techniques that explicitly consider dependency between variables.
- Use clustering techniques for comparing means and variances
- Use orthogonal projections (like Principal Component Analysis) to make the data less dependent
- If the dependency is introduced through time, use time series models.
- Use Generalized Linear Mixed Models (GLMM) or Generalized Estimating Equations (GEE)

## Course outline

6. How many samples do I need for my test?: Sample size
  1. Basic formulas for different distributions
  2. Formulas for samples with different costs
  3. What if I cannot get more samples? Resampling: Bootstrapping, jackknife

## 6.1 How many samples do I need for my test?

An engineer works for Coca Cola. He knows that the filling machine has a variance of 6 cl. (the filling process can be approximated by a Gaussian). How many cans does he need to ascertain with a 95% confidence that the machine is malfunctioning?



### Step 1. Define your hypothesis

$$H_0 : \mu = 348$$

$$H_1 : \mu \neq 348$$

### Step 2. Define a test statistic

$$\text{Under } H_0 \quad Z = \frac{\bar{x} - 348}{\frac{6}{\sqrt{N}}} \sim N(0,1)$$

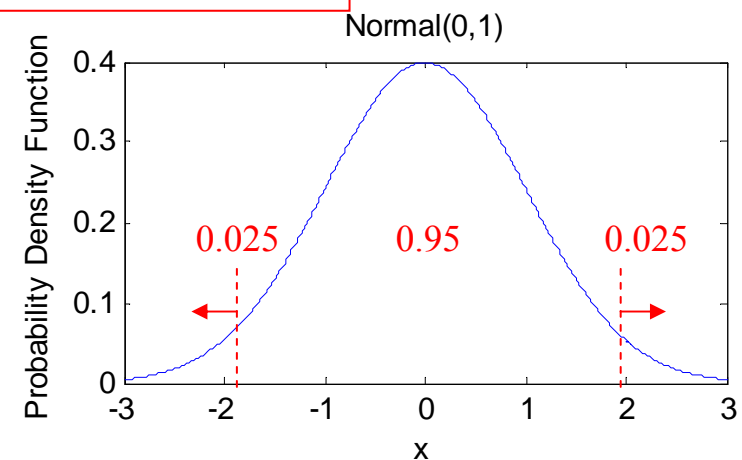
### Step 3. Define minimum detectable deviation and confidence level

We want to detect a minimum deviation of 0.5ml  
Confidence 95%,  $z_{1-\frac{0.05}{2}} = 1.96$

### Step 4. Compute non-rejection region and sample size

$$P\{|Z| < 1.96\} = 95\% \Rightarrow \frac{0.5}{\frac{6}{\sqrt{N}}} < 1.96 \Rightarrow N > \left(\frac{1.96 \cdot 6}{0.5}\right)^2 = 553.2$$

$$P\{|Z| < z_{1-\frac{\alpha}{2}}\} \Rightarrow N > \left(z_{1-\frac{\alpha}{2}} \frac{\sigma}{d}\right)^2 = \frac{z_{1-\frac{\alpha}{2}}^2}{\Delta^2} \quad \Delta = \frac{d}{\sigma}$$

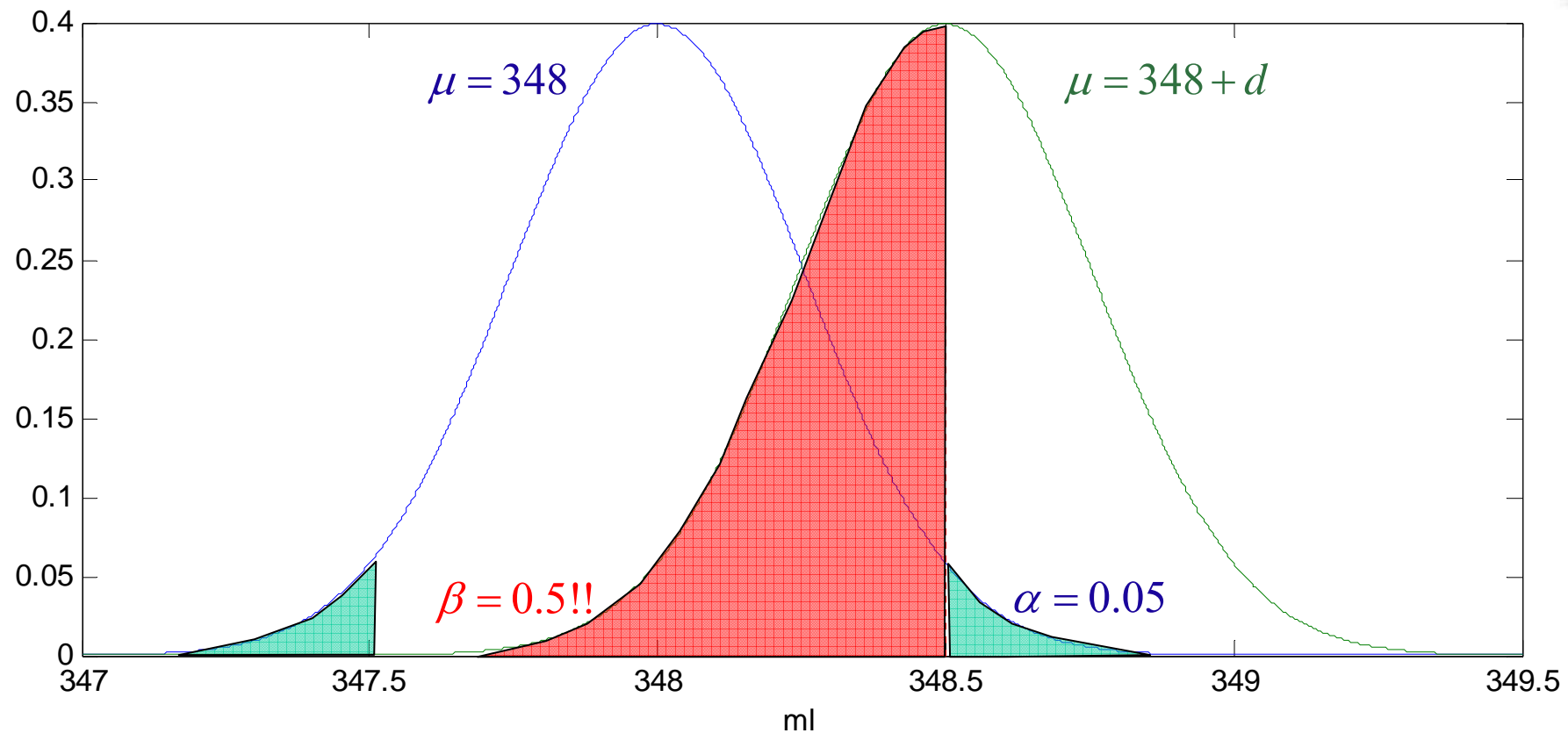


## 6.1 How many samples do I need for my test?



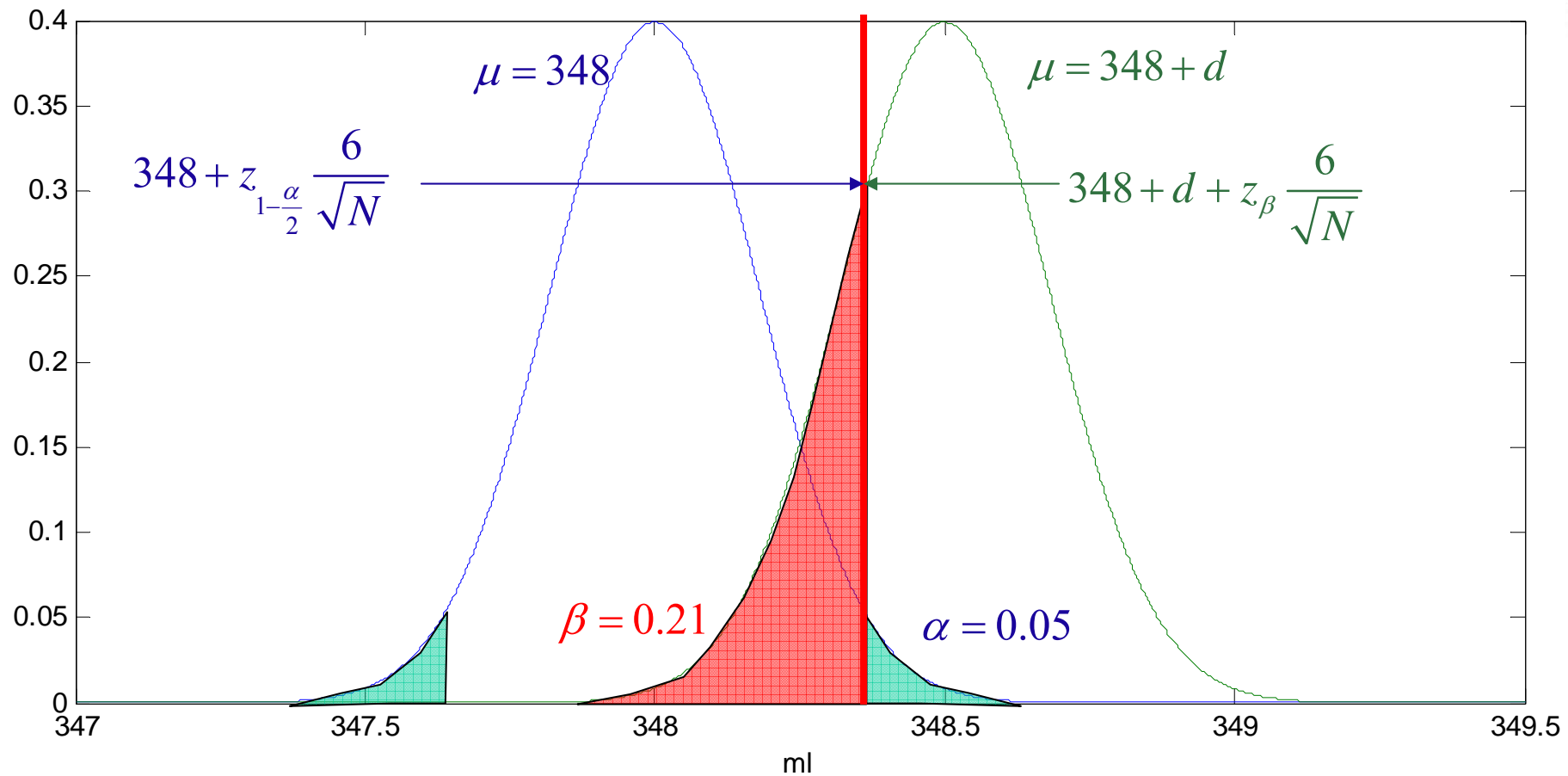
I know now how to control the Type I error (I don't want to stop the plant unnecessarily)  
But, what about the Type II error? What if I should stop the plant and I don't do it?

$N = 554$



## 6.1 How many samples do I need for my test?

$N = 1108$



## 6.1 How many samples do I need for my test?



$$348 + z_{1-\frac{\alpha}{2}} \frac{6}{\sqrt{N}} = 348 + d + z_{\beta} \frac{6}{\sqrt{N}}$$

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

more power, more confidence  
or smaller differences require  
more samples.

Step 1. Define your hypothesis

$$H_0 : \mu = 348$$

$$H_1 : \mu \neq 348$$

Step 3. Define minimum detectable deviation,  
confidence level and statistical power

$$d = 0.5 \quad z_{1-\frac{0.05}{2}} = 1.96 \quad z_{0.05} = -1.64$$

Step 2. Define a test statistic

$$\text{Under } H_0 \quad Z = \frac{\bar{x} - 348}{\frac{6}{\sqrt{N}}} \sim N(0,1)$$

Step 4. Compute sample size

$$N = \frac{(1.96 - (-1.64))^2}{\left(\frac{0.5}{6}\right)^2} = 1866.2$$



## 6.1 How many samples do I need for my test?

### Finite samples

$$\sigma_{avg}^2 = \frac{\sigma_X^2}{N_{sample}} \left( 1 - \frac{N_{sample}}{N_{population}} \right) \longrightarrow \text{The equation is not as easy as before but still is solvable}$$

### Percentages

- What kind of treatment effect are you anticipating?
- Oh!, I'm looking for a 20% change in the mean
- Mmm, and how much variability is there in your observations?
- About 30%

$$N = \left( z_{1-\frac{\alpha}{2}} - z_{\beta} \right)^2 \frac{CV^2}{PC^2} \quad CV = \frac{\sigma}{\mu} \quad PC = \frac{d}{\mu}$$

$$z_{1-\frac{0.05}{2}} = 1.96$$

$$z_{0.05} = -1.64$$

$$N = (1.96 - (-1.64))^2 \frac{0.3^2}{0.2^2} = 29.16$$

## 6.1 Basic formulas for different distributions: Gaussians

Means of Gaussians  $\xi \in [\hat{\xi} - \varepsilon, \hat{\xi} + \varepsilon]$

I want to determine a parameter with a given sampling error, for instance, the mean power consumption of the unmodified motor within  $\pm 2$  with a confidence of 95%

$$\varepsilon = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \Rightarrow N \geq \frac{z_{1-\frac{\alpha}{2}}^2}{\left(\frac{\varepsilon}{\sigma}\right)^2} \quad \varepsilon = t_{1-\frac{\alpha}{2}, N-1} \frac{s}{\sqrt{N}} \Rightarrow N \geq \frac{t_{1-\frac{\alpha}{2}, N-1}^2}{\left(\frac{\varepsilon}{s}\right)^2}$$

Detection of differences in Gaussian means

I want to determine if the filling volume is 348 with a smallest detectable difference of  $d=0.5\text{ml}$ , with a confidence of 95% and a statistical power of 95%.

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

## 6.1 Basic formulas for different distributions: Bernoulli

### Proportions

Binomial approximated by Gaussian

$np > 10$  and  $np(1-p) > 10$

$$\xi \in \left[ \hat{\xi} - \varepsilon, \hat{\xi} + \varepsilon \right]$$

I want to determine what is the proportion of people living alone in Oslo with a precision of  $\pm 1\%$  with a confidence of 95%

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{N}} \Rightarrow N_{\text{sample}} \geq \frac{z_{1-\frac{\alpha}{2}}^2}{\left( \frac{\varepsilon}{\sqrt{p(1-p)}} \right)^2}$$

← Worse case  $p=0.5$

### Finite sample correction

$$N_{\text{correctedSample}} = \frac{N_{\text{sample}}}{1 + \frac{N_{\text{sample}} - 1}{N_{\text{population}}}}$$

### Detection of difference in proportions

Binomial approximated by Gaussian

I want to determine if the proportion of people living alone in Oslo is 70% with a minimum detectable change of 1% with a confidence of 95%, and statistical power of 95%

$$p_0 + z_{1-\frac{\alpha}{2}} \frac{\sqrt{p_0(1-p_0)}}{\sqrt{N}} = p_1 + z_{\beta} \frac{\sqrt{p_1(1-p_1)}}{\sqrt{N}} \Rightarrow N \geq \frac{\left( z_{1-\frac{\alpha}{2}} \sqrt{p_0(1-p_0)} - z_{\beta} \sqrt{p_1(1-p_1)} \right)^2}{(p_0 - p_1)^2}$$

## 6.2 Formulas for samples with different costs

In an epidemiologic study it is easy to have more healthy patients than ill patients. If the illness is very rare, having ill patients may be extremely costly. Let's say we need 30 samples in both ill and control groups to be able to detect a certain difference. However, in the ill group we can only have 20 patients, how many healthy person do I need in the control group to have the same accuracy?

$$\left. \begin{array}{l} \bar{x} \sim N\left(\mu_X, \frac{\sigma^2}{N_X}\right) \\ \bar{y} \sim N\left(\mu_Y, \frac{\sigma^2}{N_Y}\right) \end{array} \right\} \bar{d} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{N_X} + \frac{\sigma^2}{N_Y}\right)$$

Valid if  $\frac{N_{Sample}}{2} < N_X < N_{Sample}$

$$\frac{\sigma^2}{N_X} + \frac{\sigma^2}{N_Y} \leq \frac{\sigma^2}{N_{Sample}} + \frac{\sigma^2}{N_{Sample}} \Rightarrow N_Y \geq \frac{N_{Sample} N_X}{2N_X - N_{Sample}}$$

$$N_{control} \geq \frac{30 \cdot 20}{2 \cdot 20 - 30} = 60$$

## 6.2 Formulas for samples with different costs

Assume that the cost of evaluating an ill person is 4 times that of evaluating a healthy person. How should I distribute the samples in the two groups if I have a certain budget for the experiment?

$$\begin{array}{l} \min \frac{\sigma^2}{N_X} + \frac{\sigma^2}{N_Y} \\ \text{s.t. } N_X c_X + N_Y c_Y = C \end{array} \xrightarrow{\text{Solution}} \left\{ \begin{array}{l} N_X = \frac{C}{\sqrt{c_X} (\sqrt{c_X} + \sqrt{c_Y})} \\ N_Y = \frac{C}{\sqrt{c_Y} (\sqrt{c_X} + \sqrt{c_Y})} \end{array} \right\} \quad \left\{ \begin{array}{l} N_Y \sqrt{c_Y} = N_X \sqrt{c_X} \end{array} \right.$$

$$N_Y \sqrt{c_X} = N_X \sqrt{4c_X} \Rightarrow N_Y = 2N_X$$

## 6.3 What if I cannot get more samples? Resampling: Bootstrapping and Jackknife

### Jackknife

#### **Jackknife** resampling

1. Take a subsample with all samples except 1
2. Estimate the mean of your subsample.
3. Repeat steps 1 and 2 leaving out once all samples

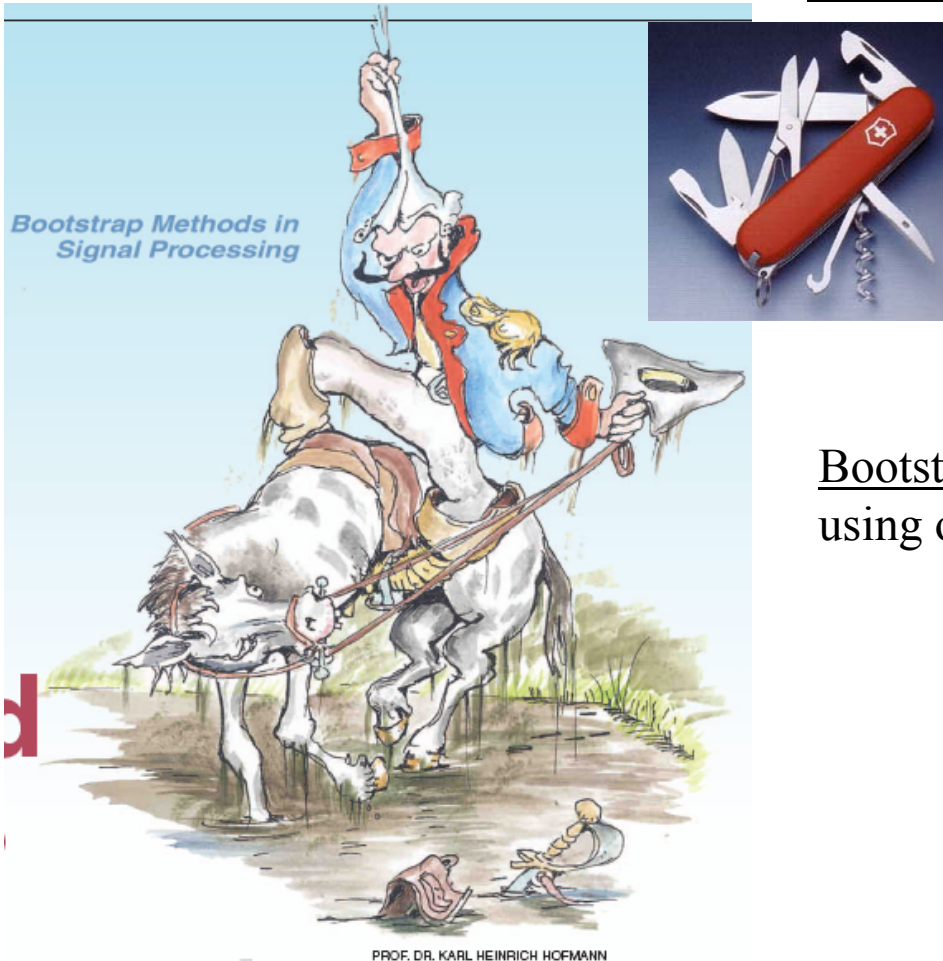
This gives you the empirical bias and variance of the mean.

Bootstrapping: To do something seemingly impossible using only the available resources.

#### **Bootstrap** resampling

1. Take a random subsample of the original sample of size  $N$  (with replacement)
2. Estimate the mean of your subsample.
3. Repeat steps 1 and 2 at least 1000 times.

This gives you the empirical distribution of the mean.



# Course outline

7. Can I deduce a model for my data?
  1. What kind of models are available?
  2. How to select the appropriate model?
  3. Analysis of Variance as a model
    1. What is ANOVA really?
    2. What is ANCOVA?
    3. How do I use them with pretest-posttest designs?
    4. What are planned and post-hoc contrasts?
    5. What are fixed-effects and random-effects?
    6. When should I use Multivariate ANOVA (MANOVA)?
  4. Regression as a model
    1. What are the assumptions of regression
    2. Are there other kind of regressions?
    3. How reliable are the coefficients? Confidence intervals
    4. How reliable are the coefficients? Validation

## 7.1 What kind of models are available?

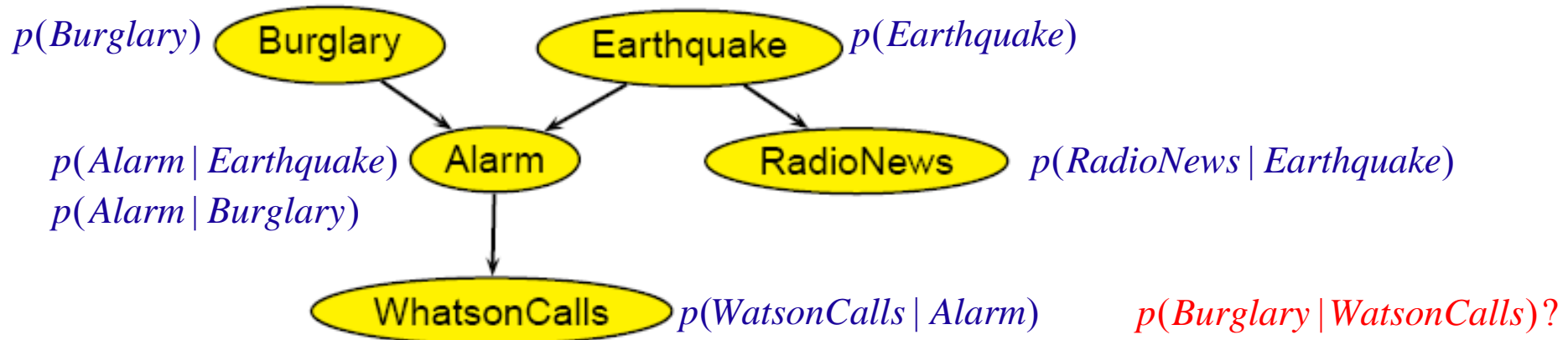
### Probability distributions:

Part of the modelling process is determining the distribution parameters, providing confidence intervals and testing that the model fits well the data

- The number of people entering a shop follows a Poisson distribution
- The joint probability of being democrat/republican and introvert/extrovert is not the product of the corresponding two marginals (i.e., the two variables are not independent)

### Graphical probability models (Bayesian networks):

Bayesian networks are a nice way of representing complex conditional probabilities.

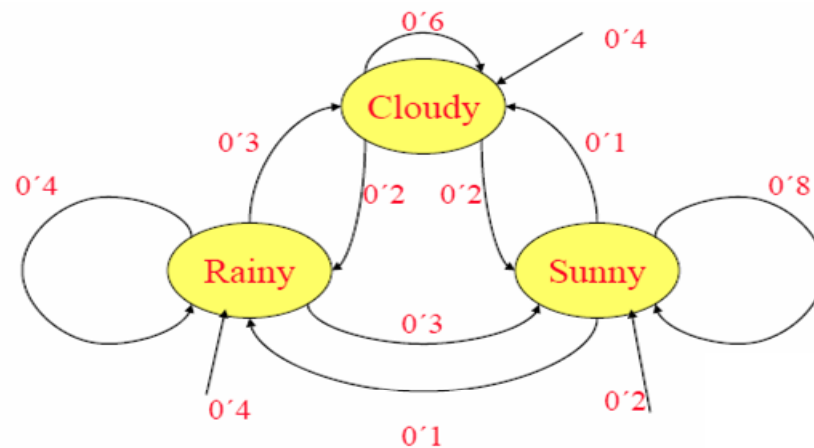




## 7.1 What kind of models are available?

### Markov models and Hidden Markov Models:

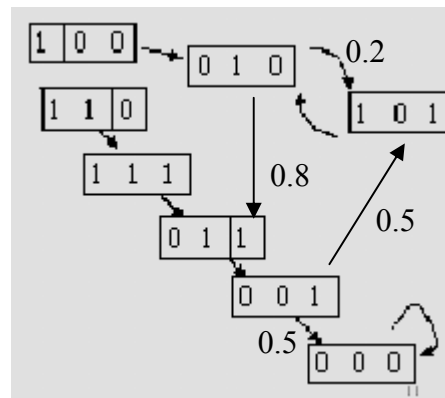
Markov models are probabilistic models to represent time dependent events.



$$p(State_{i+1} | State_i)$$

### Probabilistic Boolean Networks

All variables have a given state (ON/OFF). Given a state it changes to the next state following some probability law.



0	1	1	0	0	0	0	1	1	0
0	0	0	0	0	0	0	0	0	0
1	0	0	1	1	0	1	0	1	1
0	0	0	1	0	1	1	1	0	1
1	1	1	1	1	1	0	0	0	1
0	1	1	0	0	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1
1	1	1	0	1	0	1	0	1	1
1	1	1	1	1	0	0	1	0	0
1	0	0	0	1	0	1	0	0	1

## 7.1 What kind of models are available?

### Classifiers (Supervised classification)

Given a set of features determine which is the class of the object

### Clustering (Unsupervised classification)

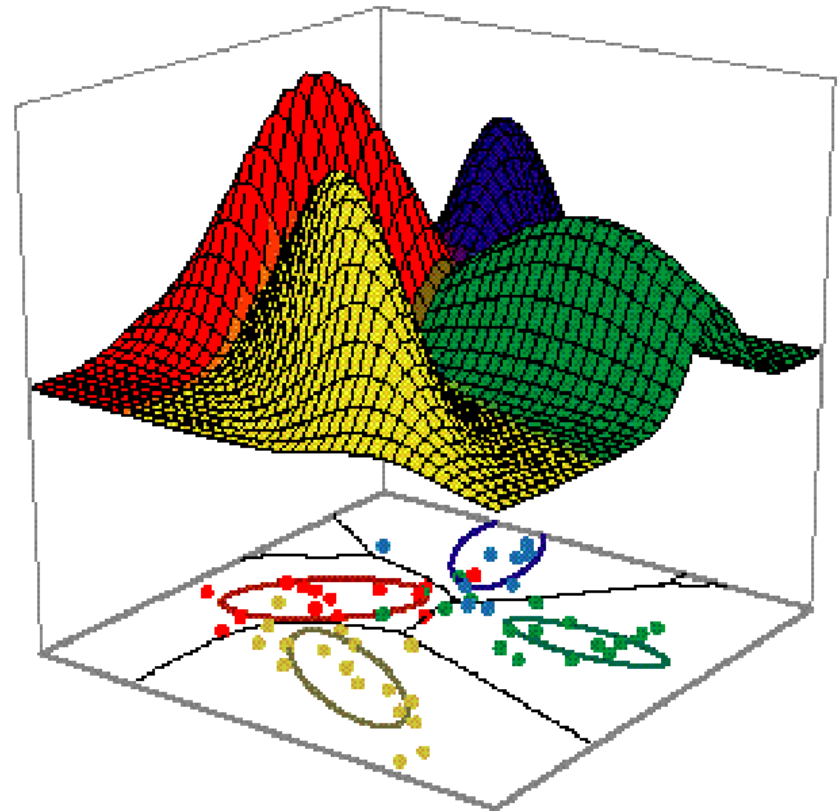
If you don't know the class, at least, tell me which objects go together.

### Association rules

Describe highly populated regions of the PDF in terms of combinations of the object variables.

Continuous PDFs:  $x_1 \geq 3 \Rightarrow x_2 < -2 \quad p = 0.3$

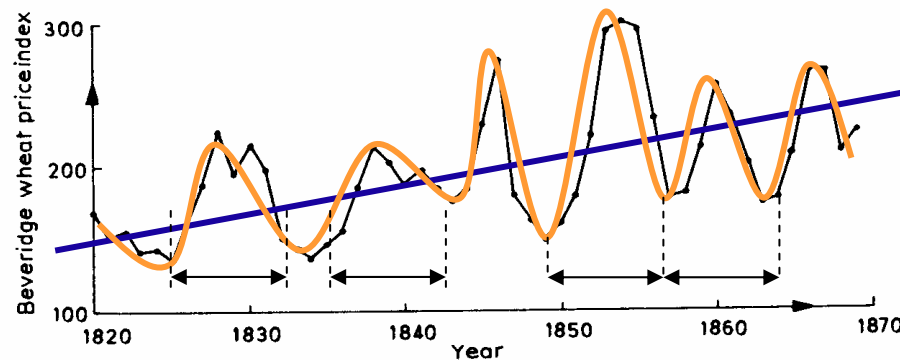
Discrete PDFs:  $A \cap B \Rightarrow C \cap D \quad p = 0.7$



## 7.1 What kind of models are available?

### Time series

Explain the underlying behaviour of a time series (AR, MA, ARMA, ARIMA, SARIMA, ...)



**Figure 1.1** Part of the Beveridge wheat price index series.

$$x[n] = \text{trend}[n] + \text{periodic}[n] + \text{random}[n]$$

$$\rightarrow p(x[n] | x[n-1], x[n-2], \dots)$$

## 7.1 What kind of models are available?

### Regression

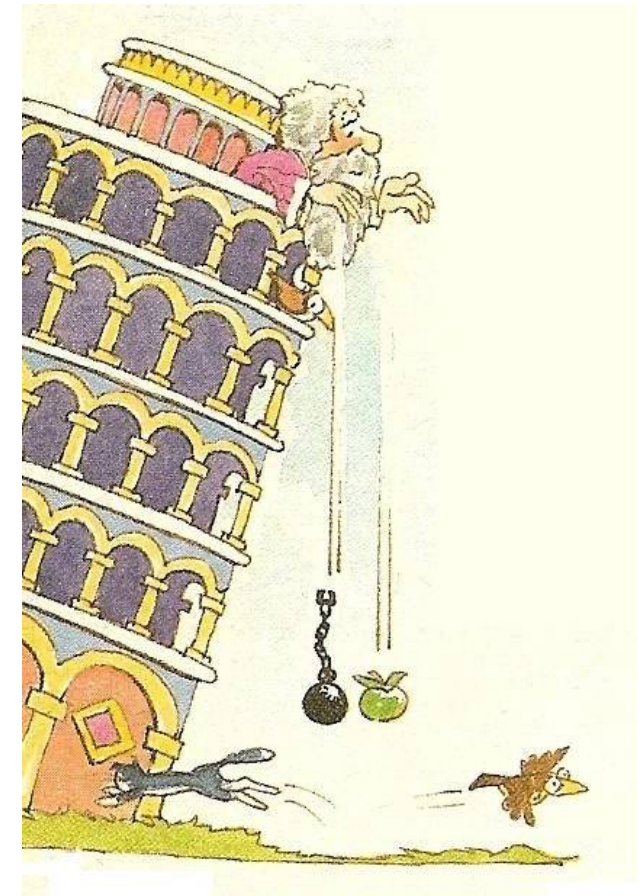
Bivariate: Find a functional relationship between two variables

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

Ideally it will find  $a_0 = a_1 = 0; a_2 = \frac{1}{2} g$

$$h(t) = \frac{\log(a_0 + a_1 t + a_2 t^2)}{\sin \sqrt{(t\varepsilon)^{a_3}}}$$

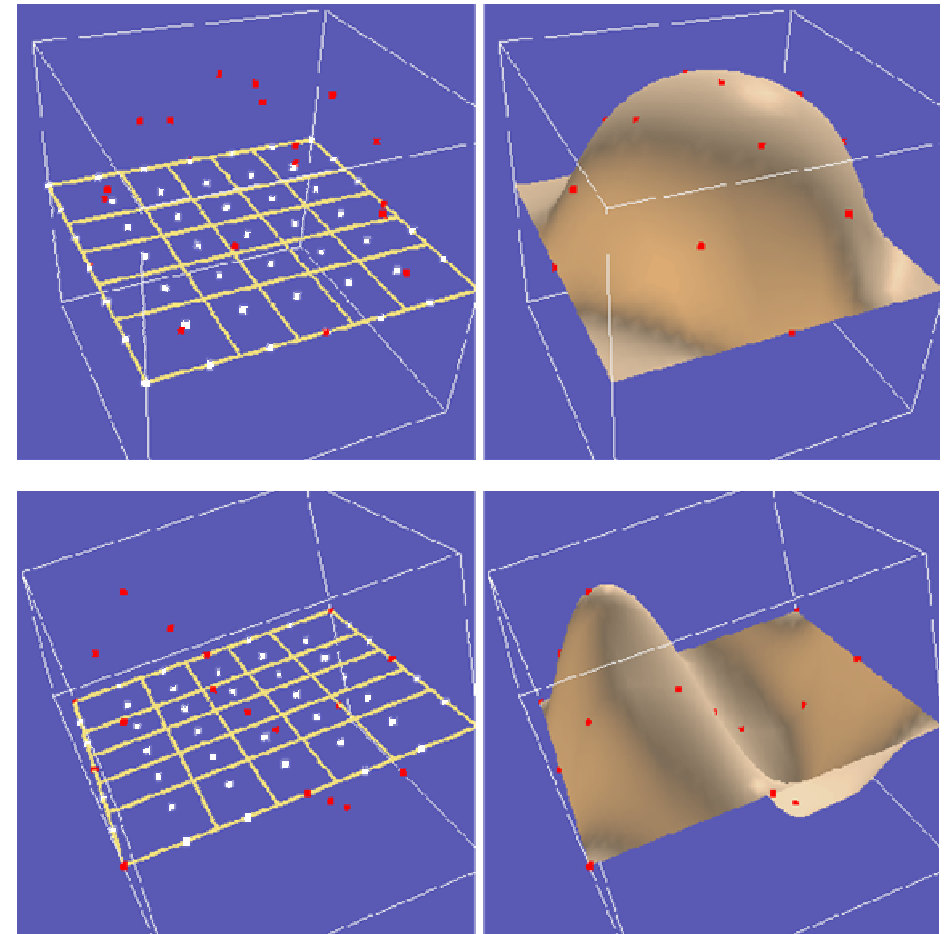
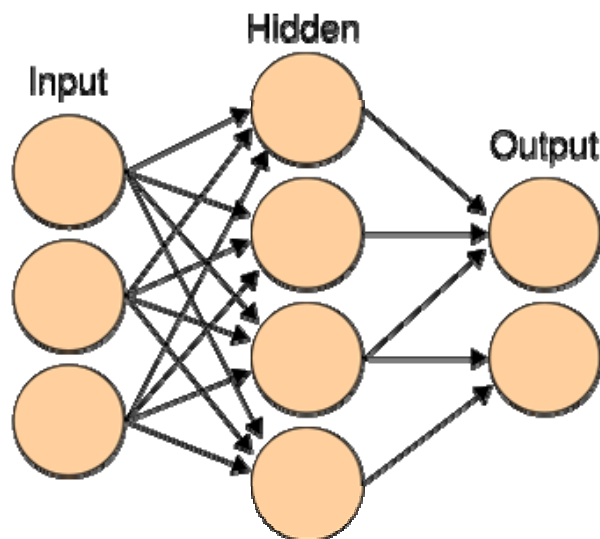
We propose the model which may be nonlinear (and nonsense!!)



## 7.1 What kind of models are available?

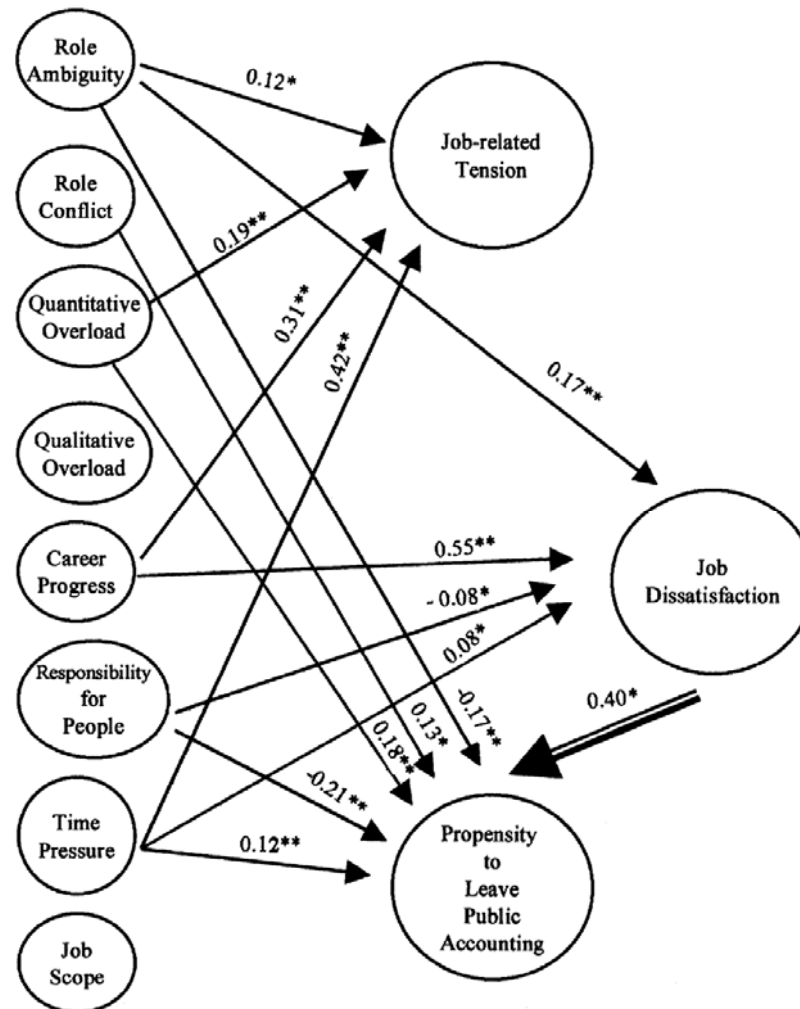
### Neural networks:

They are no more than complex non-linear regressions from  $N$  input variables to  $M$  outputs.



## 7.1 What kind of models are available?

Structural Equation Modelling:  
Generalized regression in which predictor variables can be of any type, there can be latent variables. It is a confirmatory technique.



\* 5% significance level

\*\* 1% significance level

## 7.1 What kind of models are available?

Linear models (ANOVA, MANOVA, Linear filters, Kalman filters):

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

$$Score_{ijk} = RealMadrid + Raúl_i + Strategy_j + (Raúl.Strategy)_{ij} + \varepsilon_{ijk}$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

↑  
Supposed to be gaussian

Generalized Linear Models (GLM, ANCOVA, MANCOVA, Extended Kalman, Particle filters):

$$\mathbf{y} = g^{-1}(X\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

↑  
Can have any distribution

Generalized Linear Mixed Models (GLMM)

$$\mathbf{y} = g^{-1}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1) + \boldsymbol{\varepsilon}_2$$

↑  
Can have any distribution

↑  
Can have any distribution  
depending even on  $X\boldsymbol{\beta}$

Generalized Estimating Equations (GEE):

$$\mathbf{y} = g^{-1}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1) + \boldsymbol{\varepsilon}_2$$

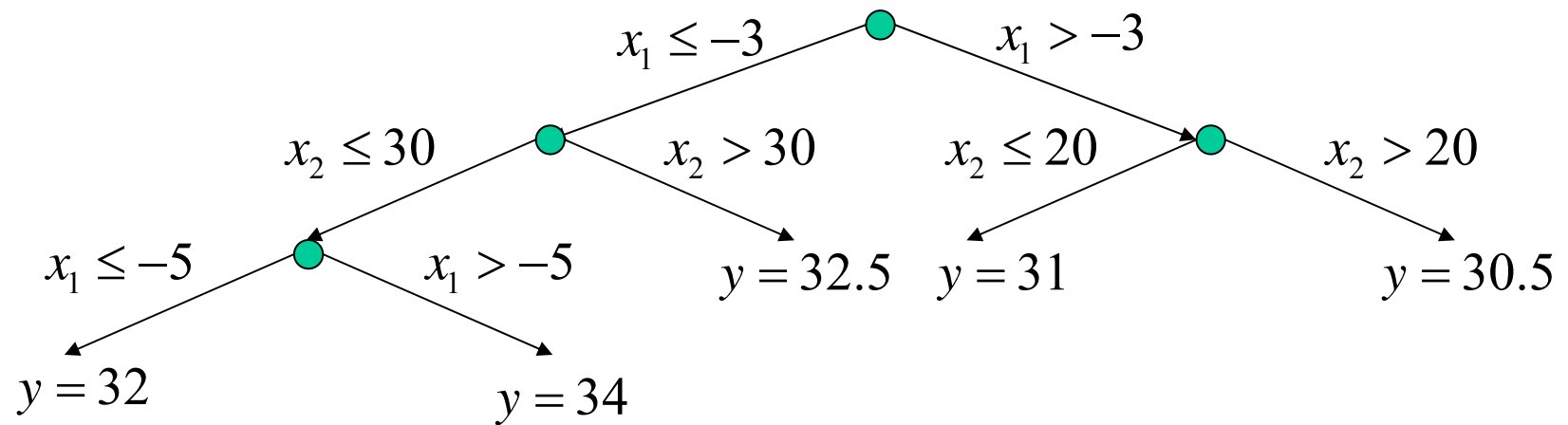
↑  
Can be auto-correlated

↑  
Can have any distribution

↑  
Can have any distribution  
depending even on  $X\boldsymbol{\beta}$

## 7.1 What kind of models are available?

### Classification and regression tree (CART)

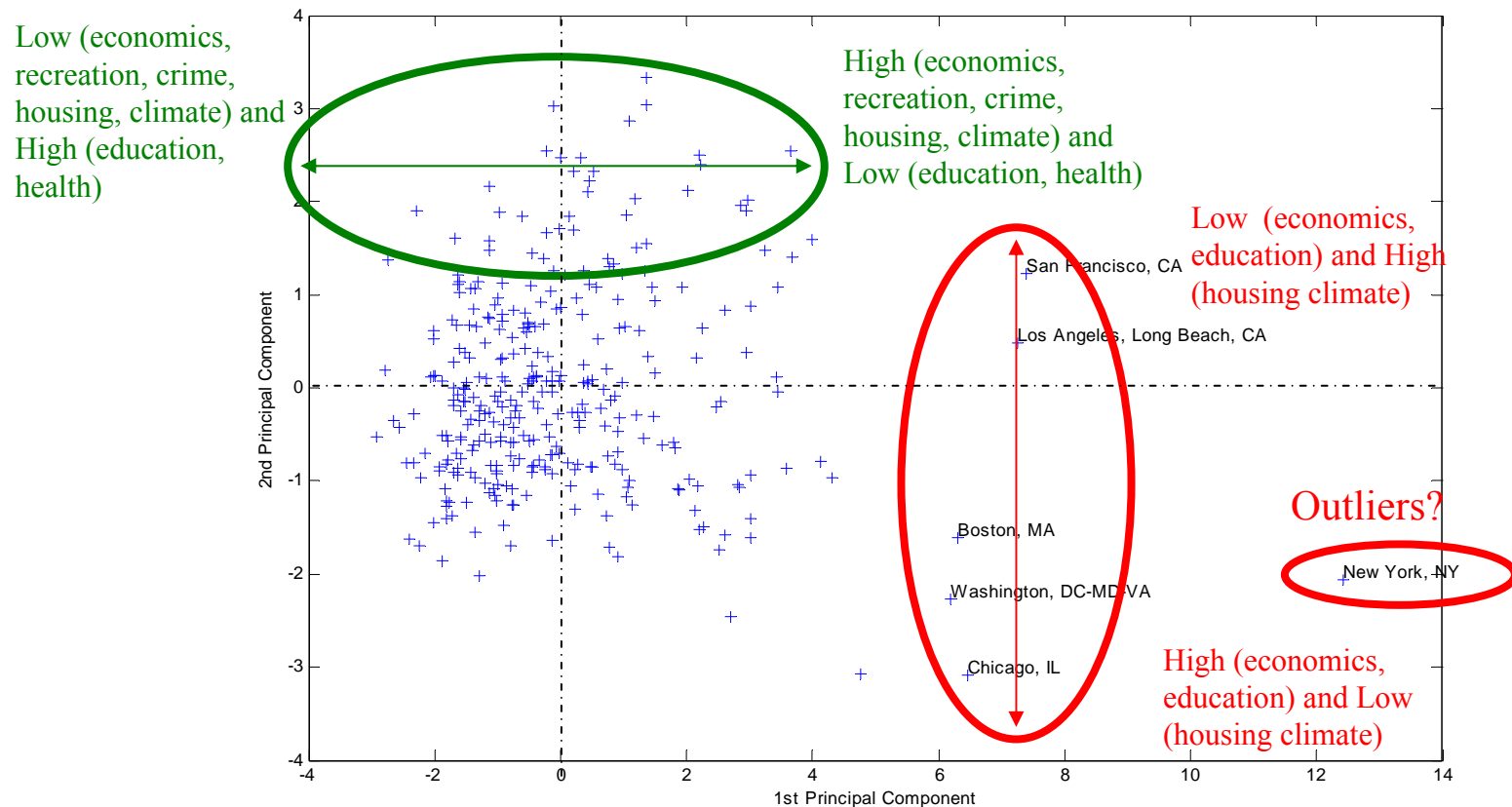




## 7.1 What kind of models are available?

### Multivariate models

Find explanatory factors for multivariate data (PCA, Factor Analysis, Multidimensional Scaling, Correspondence Analysis, MANOVA, Canonical Correlation Analysis, )



## 7.1 What kind of models are available?

Ockham's razor:

“Everything should be made as simple as possible, but not simpler” (Albert Einstein)

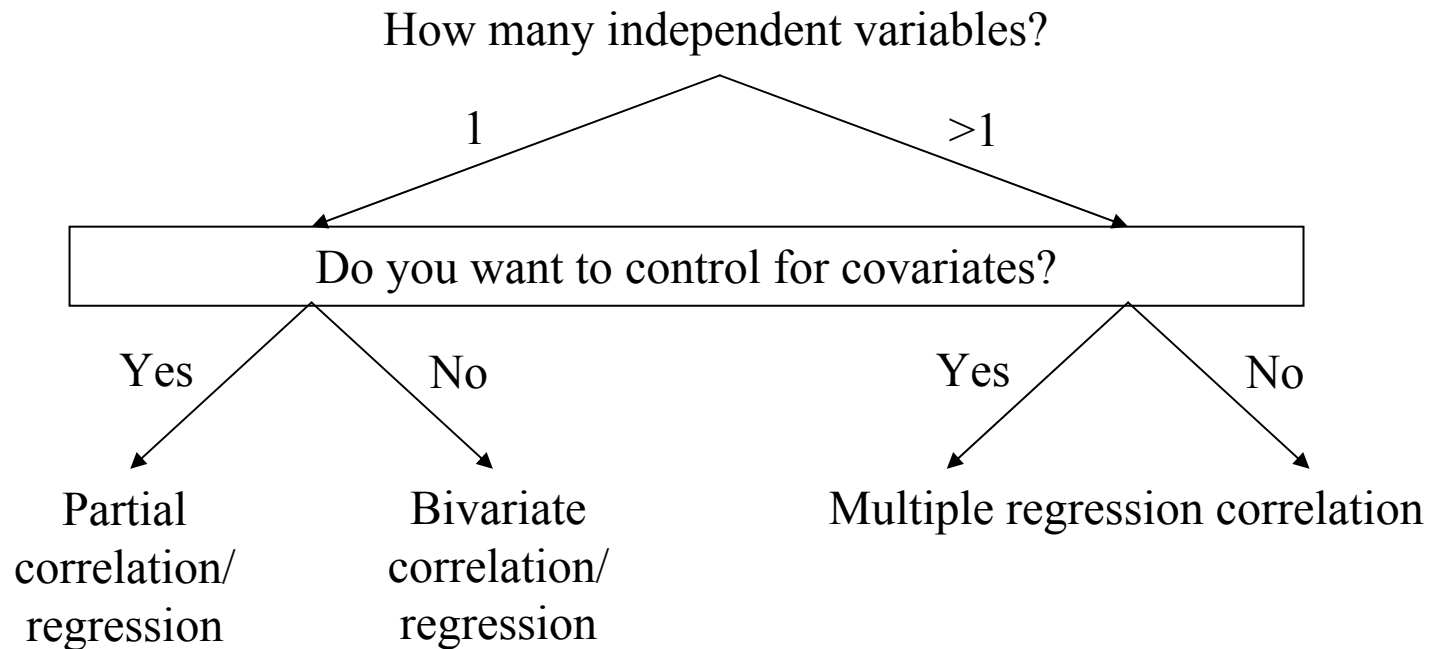
1. Use your own knowledge of the data to propose a model.
2. Check that coefficients in the model are significantly different from 0. If they are not, remove that term from the model and reestimate the model.

## 7.2 How to select the appropriate model?

Independent variables (X): interval/ratio

Dependent variable (Y): interval/ratio

Typical question: What is the degree of relationship between X and Y?

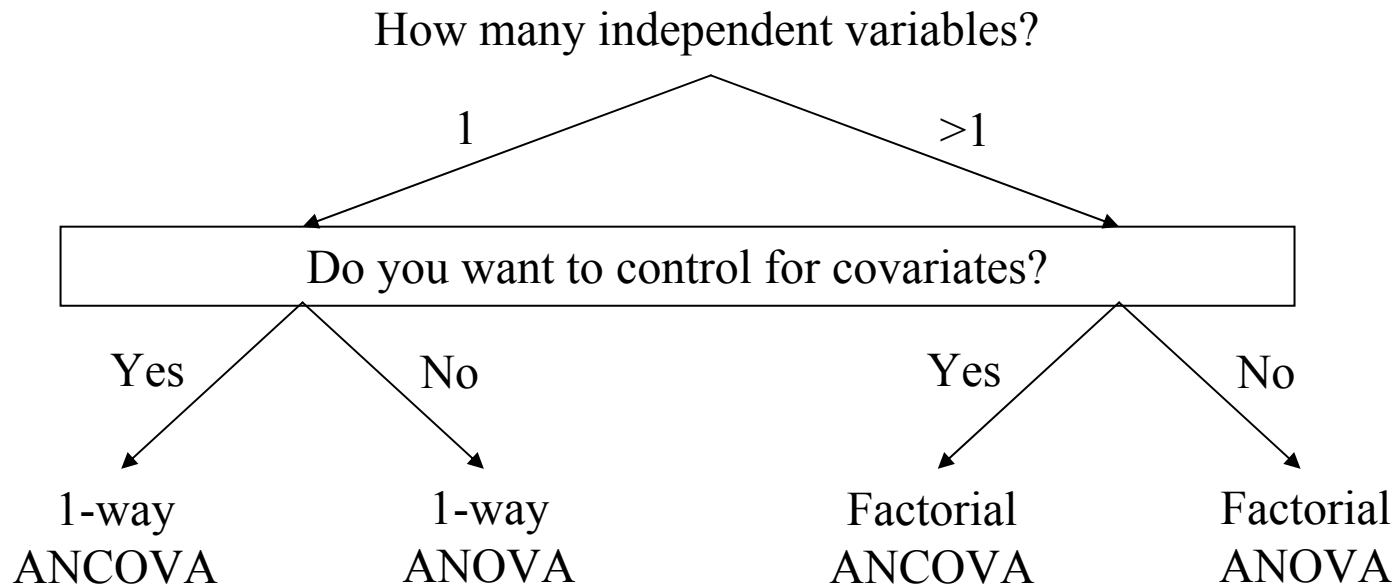


## 7.2 How to select the appropriate model?

Independent variables (X): ordinal/categorical

Dependent variable (Y): interval/ratio

Typical question: Are there significant group differences in Y between groups defined by X combinations?

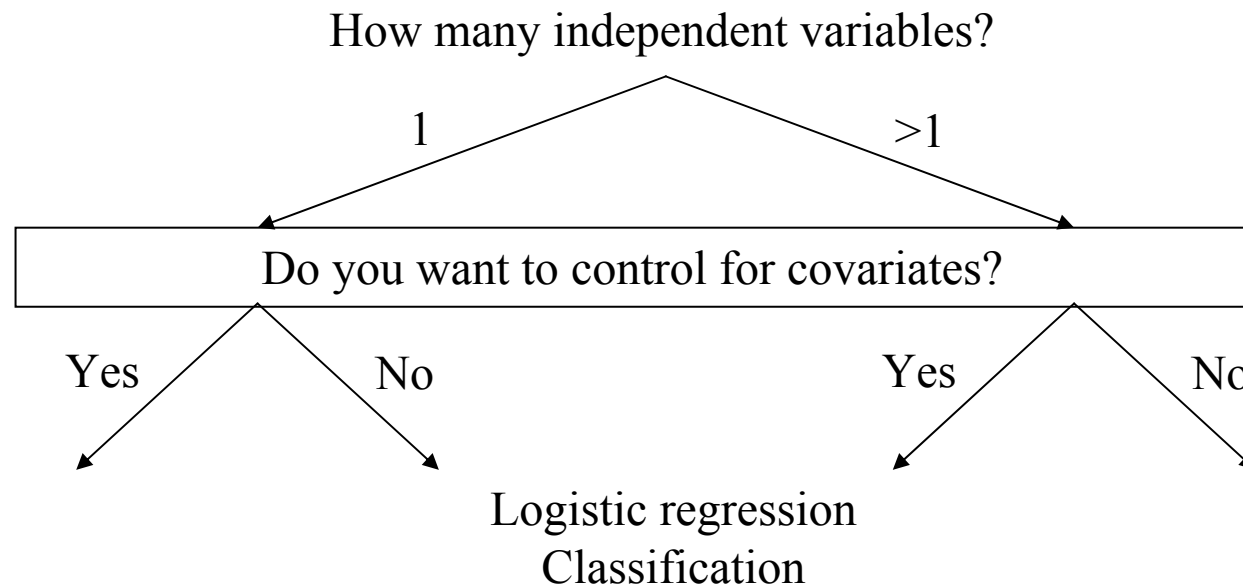


## 7.2 How to select the appropriate model?

Independent variables: interval/ratio

Dependent variable: ordinal/categorical

Typical question: Are there significant differences in the frequency of occurrence of Y related to differences in X?

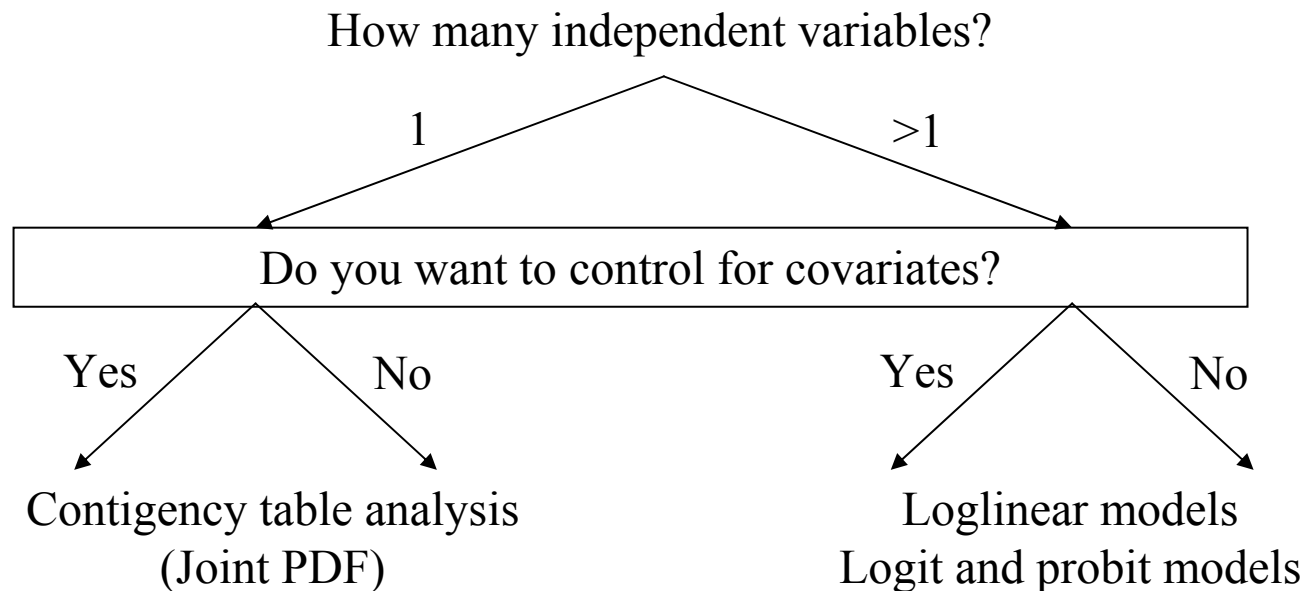


## 7.2 How to select the appropriate model?

Independent variables: ordinal/categorical

Dependent variable: ordinal/categorical

Typical question: Are there significant differences in Y for groups defined by X combinations?



## 7.2 How to select the appropriate model?

**Poll:** Let's fill the following table and give an example of each technique

	Used	Known		Used	Known
Probability Distrib.			Regression		
Bayesian Networks			Neural Networks		
Markov chains			Structural Eq. Modelling		
Probability Bool. Netw.			Linear models		
Classification			GLM		
Clustering			GLMM		
Association rules			GEE		
Time series			Multivariate		

## 7.2 How to select the appropriate model?

Situation: In our previous poll we have collected the following data about technique usage:

Person A: Regression, probability, time series

Person B: Clustering, classification, neural networks, bayesian networks

Person C: Regression, inference, time series, GLM

...

We would like to know which techniques go together, how do we do it?





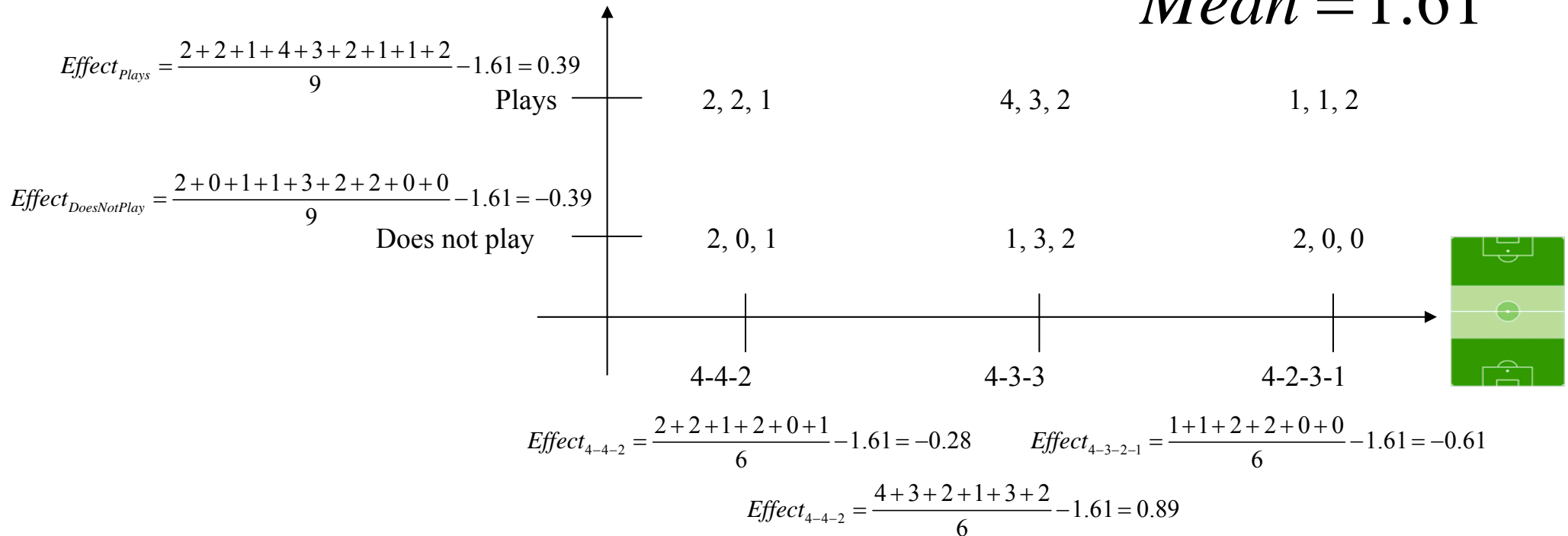
## 7.3 ANOVA as a model



Factorial approach: factors are varied together



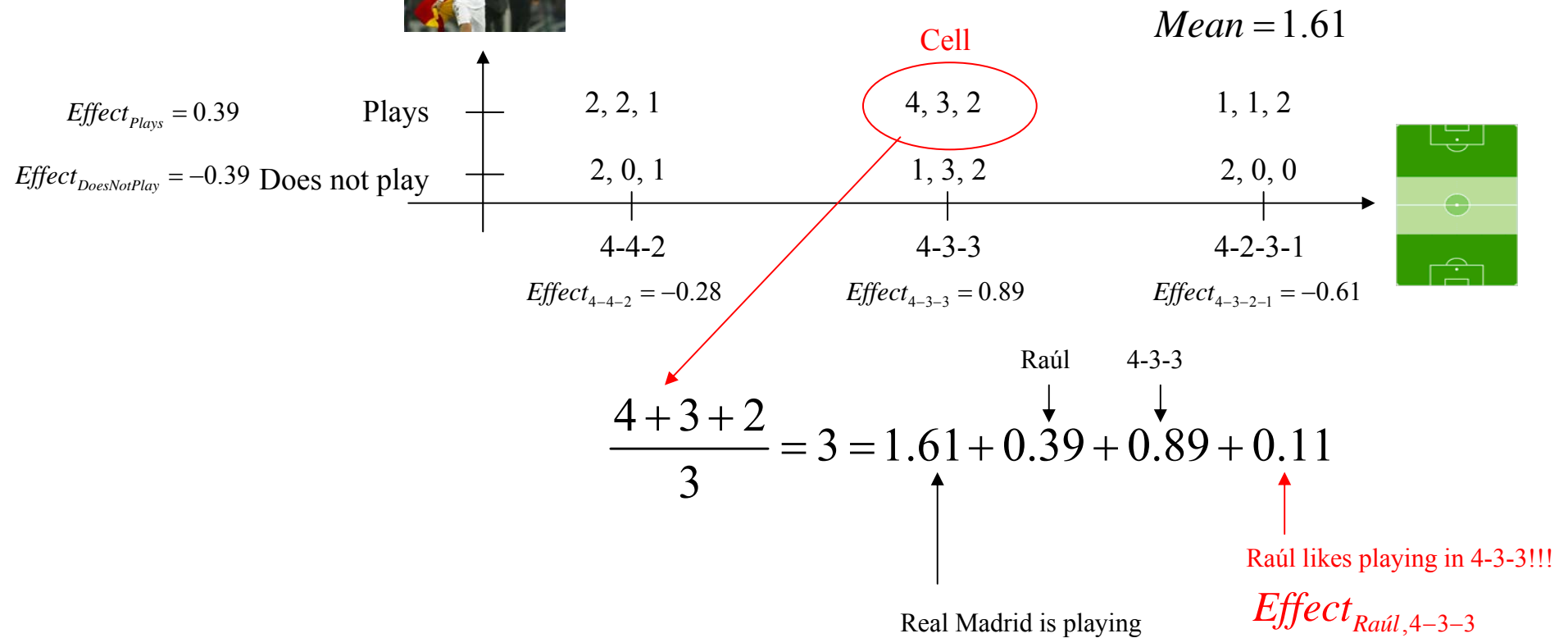
*Mean* = 1.61



## 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

↑ Raúl 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
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

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

$\uparrow$  Individual score       $\uparrow$  Drug effect

### 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

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

$\uparrow$  Individual effect

## 7.3.1 What is ANOVA really?

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

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

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_K$$

$$H_1 : \exists i, j \mid \alpha_i \neq \alpha_j$$

There is no effect of the treatment

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

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_K$$

$$H_1 : \exists i, j \mid \alpha_i \neq \alpha_j$$

There is no effect of the first treatment

---

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{K'}$$

$$H_1 : \exists i, j \mid \beta_i \neq \beta_j$$

There is no effect of the second treatment

---

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots = (\alpha\beta)_{KK'}$$

$$H_1 : \exists i, j, k, l \mid (\alpha\beta)_{ij} \neq (\alpha\beta)_{kl}$$

There is no interaction

## 7.3.1 What is ANOVA really?

How are the tests performed? F-tests

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

$$\mu = 1.61$$

$$\alpha_{Raul} = 0.39$$

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

$$\beta_{4-4-2} = -0.28$$

$$\beta_{4-3-3} = 0.89$$

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

$$(\alpha\beta)_{Raul.4-4-2} = -0.05$$

$$(\alpha\beta)_{Raul.4-3-3} = 0.11$$

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

$$(\alpha\beta)_{NoRaul.4-4-2} = 0.05$$

$$(\alpha\beta)_{NoRaul.4-3-3} = -0.11$$

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

Source	SumSquare	df	MeanSquare	F	Prob>F
Raúl	2.7222	1	2.7222	3.2667	0.0958
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			

Also called  
within-  
groups

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

$$MS_i = \frac{SS_i}{df_i}$$

$$F_i = \frac{MS_i}{MS_{Error}}$$

$F_{1,12}$   
 $F_{2,12}$   
 $F_{2,12}$

Significant at a  
confidence  
level of 95%

## 7.3.1 What is ANOVA really?

How are the different terms computed?

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

Every term is the average of all input data with the indexes being determined fixed.

Overall mean  $\hat{\mu} = \bar{x}_{..} = \frac{1}{IJK} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K X_{ijk}$

Effects of variable  $\alpha$   $\hat{\alpha}_i = \bar{x}_{i.} - \hat{\mu} = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K (X_{ijk} - \underbrace{\hat{\mu}}_{\text{What we have explained so far}})$

Effects of variable  $\beta$   $\hat{\beta}_j = \bar{x}_{.j} - \hat{\mu} = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (X_{ijk} - \underbrace{\hat{\mu}}_{\text{What we have explained so far}})$

Effects of interaction  $\alpha\beta$   $\hat{\alpha}\hat{\beta}_{ij} = \bar{x}_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) = \frac{1}{K} \sum_{k=1}^K (X_{ijk} - \underbrace{(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)}_{\text{What we have explained so far}})$

Residuals  $\varepsilon_{ijk} = X_{ijk} - \underbrace{(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha}\hat{\beta}_{ij})}_{\text{What we have explained so far}}$



## 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  $\alpha$

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

SS explained by variable  $\beta$

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

SS explained by interaction  $\alpha\beta$

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

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”

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

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			

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

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

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

## 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	$F_{2,27}$
Error	50.00	27	1.85			
Total	100.00	29				

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	$F_{1,26}$
Vocabulary	50.00	2	25.00	21.7	2.8e-6	$F_{2,26}$
Error	30.00	26	1.15			
Total	100.00	29				

## 7.3.2 What is ANCOVA?

How are the different terms computed?

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

Every term is the average of all input data with the indexes being determined fixed.

Overall mean  $\hat{\mu} = \bar{x}_{..} = \frac{1}{IJK} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K X_{ijk}$

Age regressor  $\hat{a} = \frac{1}{IJK} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \frac{X_{ijk} - \hat{\mu}}{age_{ijk}}$

Effects of variable  $\alpha$   $\hat{\alpha}_i = \bar{x}_{i.} - (\hat{\mu} + \hat{a} \cdot age_{i.}) = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K (X_{ijk} - (\hat{\mu} + \hat{a} \cdot age_{ijk}))$

Effects of variable  $\beta$   $\hat{\beta}_j = \bar{x}_{.j} - (\hat{\mu} + \hat{a} \cdot age_{.j}) = \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K (X_{ijk} - (\hat{\mu} + \hat{a} \cdot age_{ijk}))$

Effects of interaction  $\alpha\beta$   $\hat{\alpha}\hat{\beta}_{ij} = \bar{x}_{ij} - (\hat{\mu} + \hat{a} \cdot age_{ij} + \hat{\alpha}_i + \hat{\beta}_j) = \frac{1}{K} \sum_{k=1}^K (X_{ijk} - (\hat{\mu} + \hat{a} \cdot age_{ijk} + \hat{\alpha}_i + \hat{\beta}_j))$

Residuals  $\varepsilon_{ijk} = X_{ijk} - (\hat{\mu} + \hat{a} \cdot age_{ijk} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha}\hat{\beta}_{ij})$

## 7.3.2 What is ANCOVA?

How are the different terms computed?

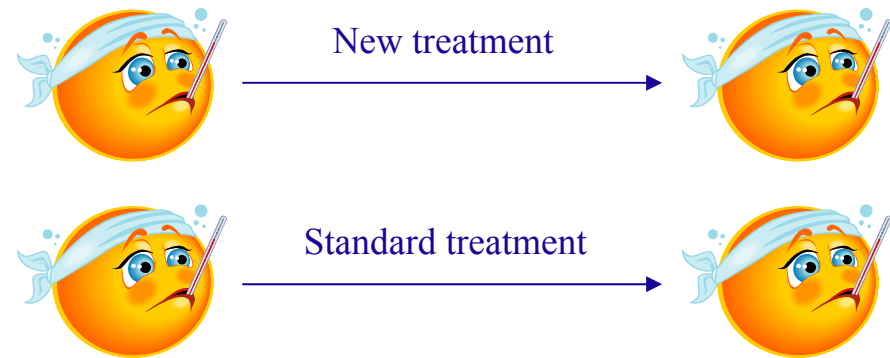
$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

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 Regression	$SS_{age} = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (X_{ijk} - \hat{\mu})^2 - (X_{ijk} - (\hat{\mu} + \hat{a} \cdot age_{ijk}))^2$
SS explained by variable $\alpha$	$SS_{\alpha} = JK \sum_{i=1}^I (\bar{x}_{i.} - (\hat{\mu} + \hat{a} \cdot age_{i.}))^2$
SS explained by variable $\beta$	$SS_{\beta} = IK \sum_{j=1}^J (\bar{x}_{.j} - (\hat{\mu} + \hat{a} \cdot age_{.j}))^2$
SS explained by interaction $\alpha\beta$	$SS_{\alpha\beta} = K \sum_{i=1}^I \sum_{j=1}^J (\bar{x}_{ij} - (\hat{\mu} + \hat{a} \cdot age_{ij} + \hat{\alpha}_i + \hat{\beta}_j))^2$
SS Residuals	$SS_E = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (X_{ijk} - (\hat{\mu} + \hat{a} \cdot age_{ijk} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha}\hat{\beta}_{ij}))^2$

## 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



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!!

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

## 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:

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

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

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

$$H_0 : \frac{\alpha_{stocks} - \alpha_{housing}}{2} - \frac{\alpha_{fixedFunds} - \alpha_{variableFunds}}{2} = 0 \text{ is orthogonal to the previous contrast}$$

## 7.3.4 What are planned and post-hoc contrasts?

What is a contrast?

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

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

$$\left. \begin{array}{ll} H_0 : \alpha_{\text{drugA}} - \alpha_{\text{placebo}} = 0 & \mathbf{c}_1 = (1, 0, 0, 0, -1) \\ H_0 : \alpha_{\text{drugB}} - \alpha_{\text{placebo}} = 0 & \mathbf{c}_2 = (0, 1, 0, 0, -1) \\ H_0 : \alpha_{\text{drugC}} - \alpha_{\text{placebo}} = 0 & \mathbf{c}_3 = (0, 0, 1, 0, -1) \\ H_0 : \alpha_{\text{drugD}} - \alpha_{\text{placebo}} = 0 & \mathbf{c}_4 = (0, 0, 0, 1, -1) \end{array} \right\} \begin{array}{l} \text{They are not} \\ \text{orthogonal} \\ \text{contrasts} \end{array}$$

I will make 4 hypothesis tests (multiple tests) with the same dataset, I have to be careful with Type I error inflation.

If the contrasts were orthogonal, Bonferroni's correction would be fine, instead of pessimistic as it tends to be.



## 7.3.4 What are planned and post-hoc contrasts?

- Planned contrast: A contrast that you decided to test before collecting data, it is driven by theory, we perform a confirmatory data analysis
- Post-hoc test: A contrast you decide to test only after observing all or part of the data, it is driven by data, we perform an exploratory data analysis

This is rather philosophical, isn't it? At the end it is the same!

No!!

### Example:

Let's say that I'm testing 100 drugs at a confidence level of 95%. If before conducting the experiment I have the preconceived idea that drug A is effective, I have a Type I error (accepting a drug as effective when it is not) of 5%. If I don't have any preconceived idea, I will have an inflated Type I error probability because of the multiple testing.

Ok, I have to take care with the multiple testing effect.

## 7.3.4 What are planned and post-hoc contrasts?

Example:

Let's go back to our investment problem.

After data collection we obtain  $\hat{\alpha}_{stocks} = 0.3; \hat{\alpha}_{housing} = 0.1; \hat{\alpha}_{fixedFunds} = -0.2; \hat{\alpha}_{variableFunds} = -0.2$

My planned contrast  
has a confidence level of 95%.  
$$H_0 : \frac{\alpha_{stocks} + \alpha_{housing}}{2} - \frac{\alpha_{fixedFunds} + \alpha_{variableFunds}}{2} = 0$$

While my post-hoc contrast  
has a lower confidence level.  
$$H_0 : \frac{\alpha_{(Best)} + \alpha_{(SecondBest)}}{2} - \frac{\alpha_{(SecondWorse)} + \alpha_{(Worse)}}{2} = 0$$

**Why?**

Because I could have obtained  $\hat{\alpha}_{stocks} = 0.3; \hat{\alpha}_{housing} = -0.1; \hat{\alpha}_{fixedFunds} = 0.2; \hat{\alpha}_{variableFunds} = -0.4$

Imagine that in fact, the different investments really have no effect, and the differences are due to chance. My planned contrast would be the same and have a probability of error of 5%. However, my post-hoc contrast would change now. And in fact, I have more chances that the difference between the best two investments are different (by chance) from the worse two.

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

## 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

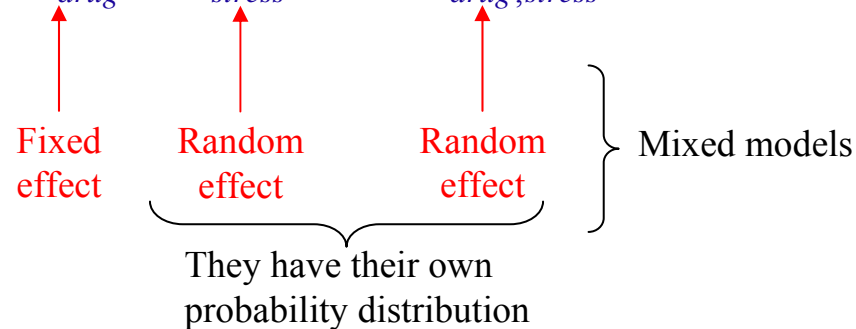
Situation: We want to compare the effect of 4 different drugs with respect to a placebo.

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

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

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



## 7.3.6 When should I use Multivariate ANOVA (MANOVA)?

What is MANOVA?

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

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

$$H_0 : \alpha_{\text{mathTextA}} = \alpha_{\text{mathTextB}}$$

$$H_0 : \beta_{\text{physicsTextA}} = \beta_{\text{physicsTextB}}$$

$$H_0 : \gamma_{\text{collegeA}} = \gamma_{\text{collegeB}}$$

	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?

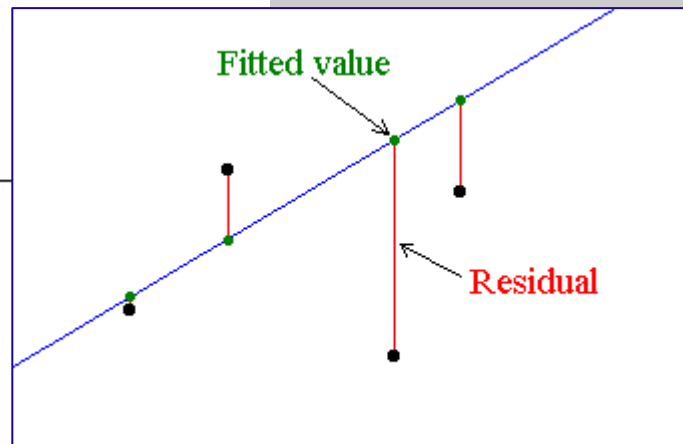
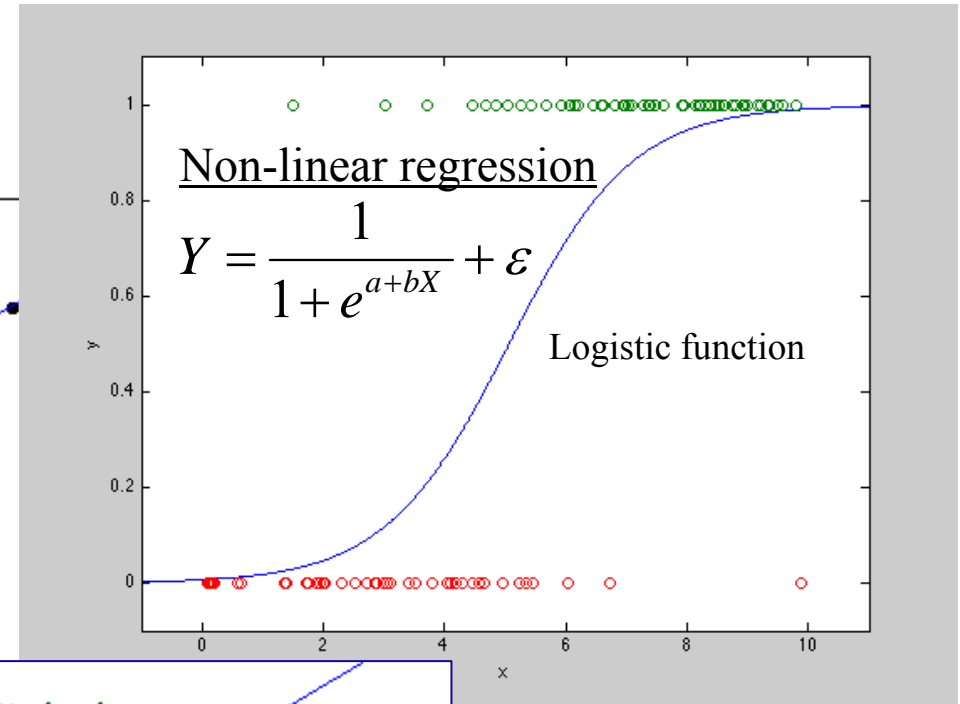
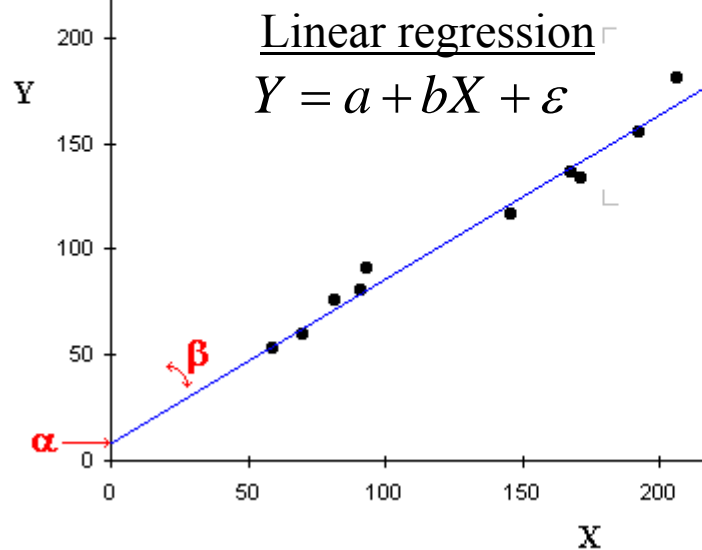
MANOVA is a very powerful technique but it is even more “omnibus” than ANOVA. If you expect correlation among the two dependent variables (ability in math, and ability in physics), then use MANOVA. If they are clearly independent, use two separate ANOVAs.

## 7.4 Regression as a model

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

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

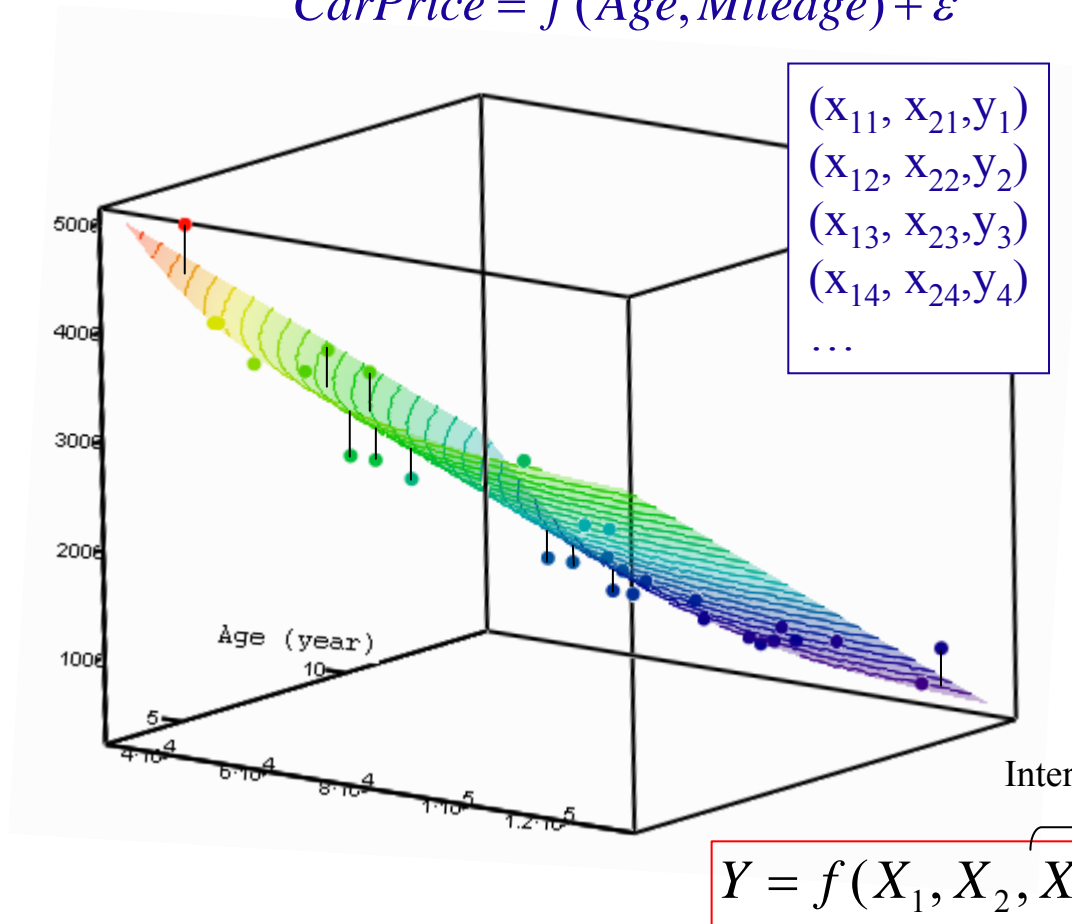
$$\text{CarPrice} = f(\text{Age}) + \varepsilon$$



## 7.4 Regression as a model

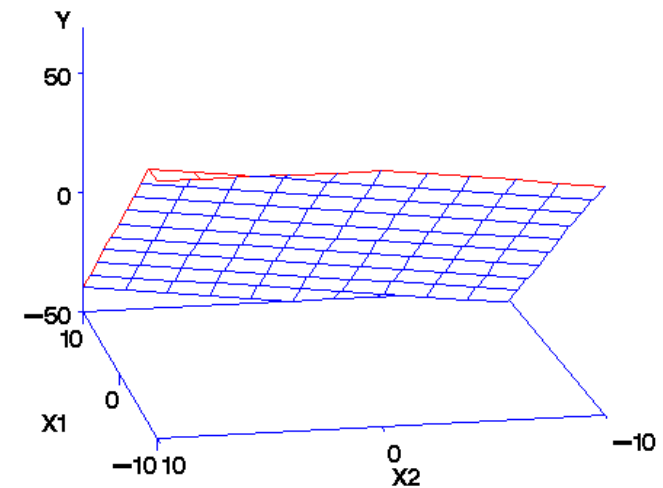
$$Y = f(X_1, X_2) + \varepsilon$$

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



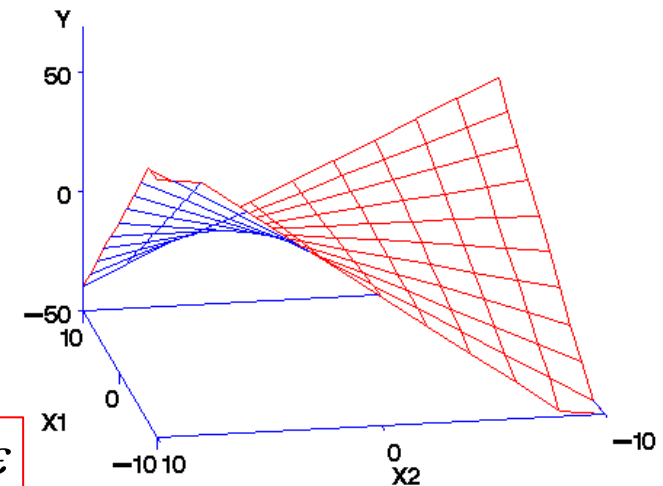
Regression surface, varying  $b_1$

$$Y = -5 \cdot X_1 + 1 \cdot X_2$$



Interaction regression surface, varying  $b_{12}$

$$Y = 0 \cdot X_1 + 1 \cdot X_2 - 0.5 \cdot X_1 \cdot X_2$$



## 7.4 Regression as a model: Goodness-of-fit

### Coefficient of determination

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

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

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

The coefficient of determination represents the percentage of “unexplained” variance

### Multiple correlation coefficient

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

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

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

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



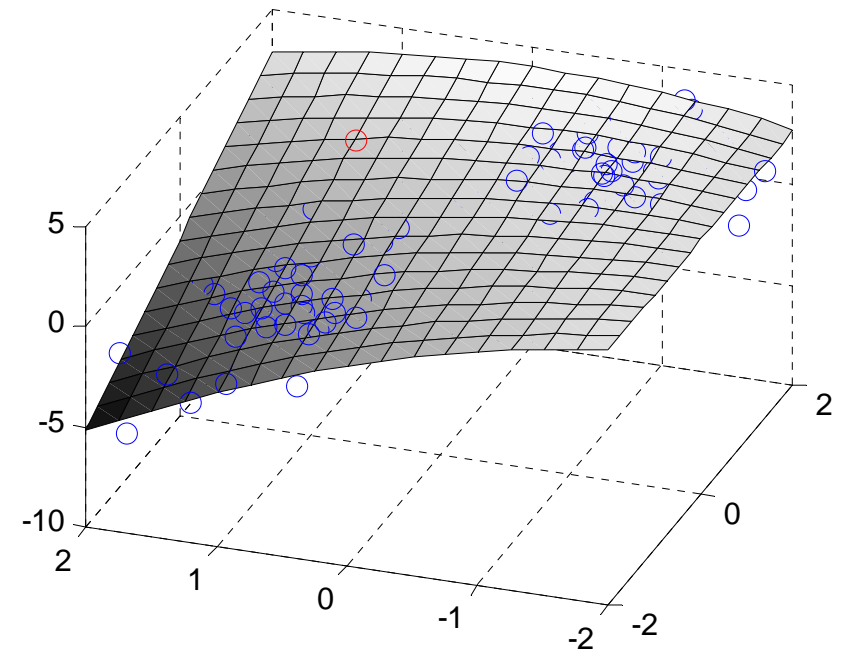
## 7.4.1 What are the assumptions of regression?

### Multiple regression

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

↑  
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.

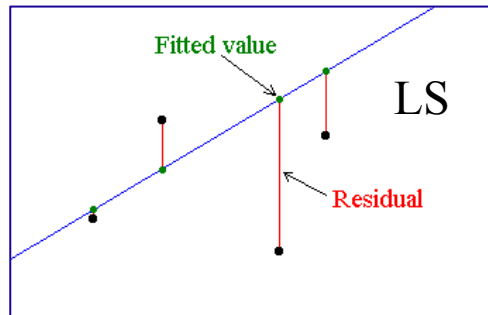
**Solution:** Make sure that your predictor vector  $(X_1, \dots, X_p)$  is not an outlier of the “training sample”.

## 7.4.1 What are the assumptions of regression?

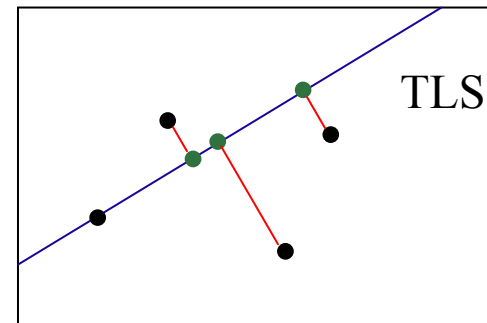
### Assumptions

- The dependent variable is noisy, but the predictors are not!!

**Solution:** If the predictors are noisy, use a scheme like Total Least Squares



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



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

Zero-mean  
Random variable

Systematic errors are not contemplated

### Least Squares

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

### Total Least Squares

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

## 7.4.1 What are the assumptions of regression?

### Assumptions

- 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

## 7.4.1 What are the assumptions of regression?

### 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).

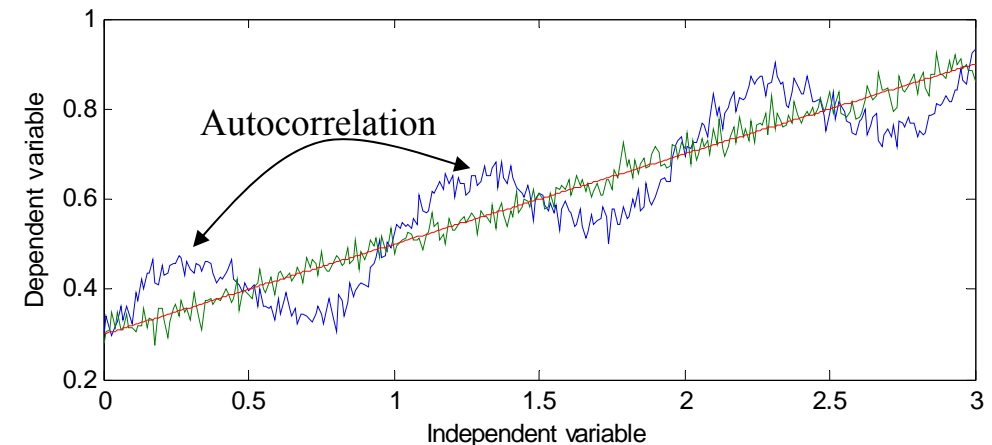
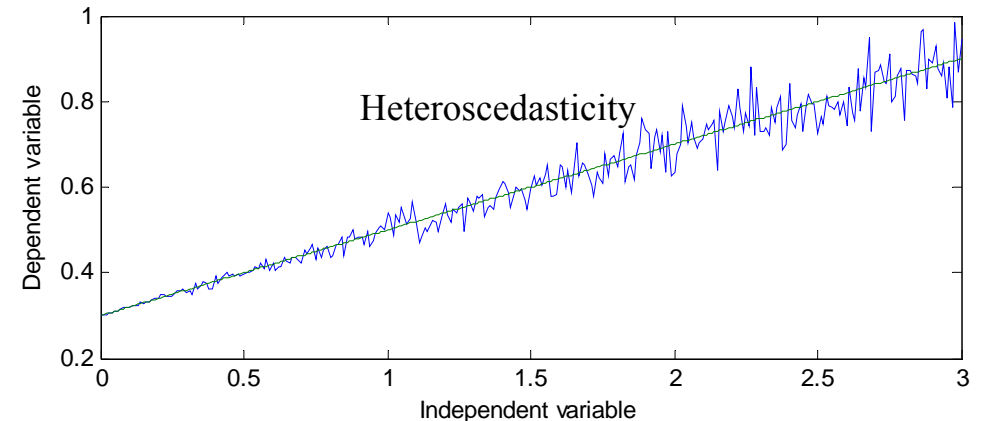
#### **Solution:**

- Use Generalized Least Squares.

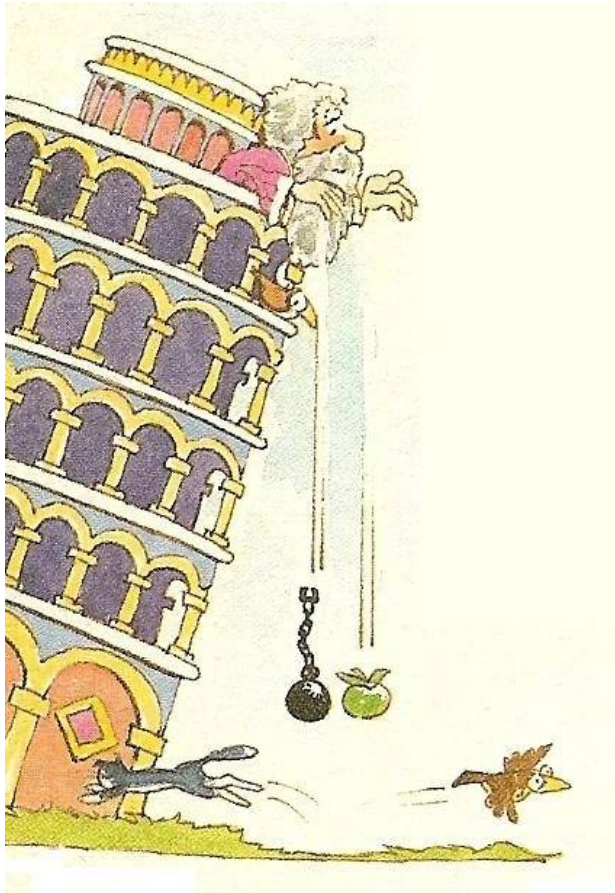
6. The errors follow a normal distribution.

#### **Solution:**

- Use Generalized Linear Models.



## 7.4.1 What are the assumptions of regression?



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?



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

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

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

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

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

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

## 7.4.2 Are there other kinds of regressions?

### Poisson regression

In this case we are fitting a parameter of a distribution, i.e, we don't try to predict Y but its distribution.

Let's assume that Y has a Poisson distribution (this is particularly valid for counts)

Which will be the number of calls arriving at a call center?

$$\log(E\{NumberOfCalls\}) = a + b \cdot PeopleAssigned$$

$$\log(E\{Y\}) = a + bX$$

Which is the number of trees per kilometer square in this region, if my data is the number of trees in a number of forests around?

$$\log\left(\frac{E\{NumberOfTrees\}}{ForestArea}\right) = a + b \cdot AciditySoil$$

$$\log\left(\frac{E\{Y\}}{Exposure}\right) = a + bX$$

## 7.4.2 Are there other kinds of regressions?

### Penalized Least Squares

This is a way of computing the regression coefficients avoiding overfitting. Normally the penalization imposes some kind of smoothness on the solution.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^N (Y_i - f(\mathbf{X}_i, \boldsymbol{\beta}))^2 + \underbrace{\lambda T(f(\mathbf{X}, \boldsymbol{\beta}))}_{\text{Penalization weight and function}}$$

### Partial Least Squares

This is a multivariate extension of linear regression

$$\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

### Bayesian regression

Incorporate a priori information about the distribution of the regression coefficients

$$\max_{\boldsymbol{\beta}} \left( \prod_{i=1}^N f_{\varepsilon}(Y_i | f(\mathbf{X}_i, \boldsymbol{\beta})) \right) f_{\boldsymbol{\beta}}(\boldsymbol{\beta})$$

### Robust regression

Use M-estimators or least absolute distance instead of least squares

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

## 7.4.2 Are there other kinds of regressions?

### Non-parametric regression

#### Kernel regression

The regression function is implicitly formed by convolution of the data with a kernel

$$\hat{Y} = E(Y | X) \Rightarrow Y = \frac{\sum_{i=1}^N K_h(\mathbf{X} - \mathbf{X}_i) Y_i}{\sum_{i=1}^N K_h(\mathbf{X} - \mathbf{X}_i)}$$

### Non-parametric regression

#### Quantile regression

Predict the quantiles of Y from the quantiles of X

$$\hat{\beta}(\tau) = \min_{\beta} \sum_{i=1}^N \rho_{\tau}(Y_i - f(\mathbf{X}_i, \beta))$$

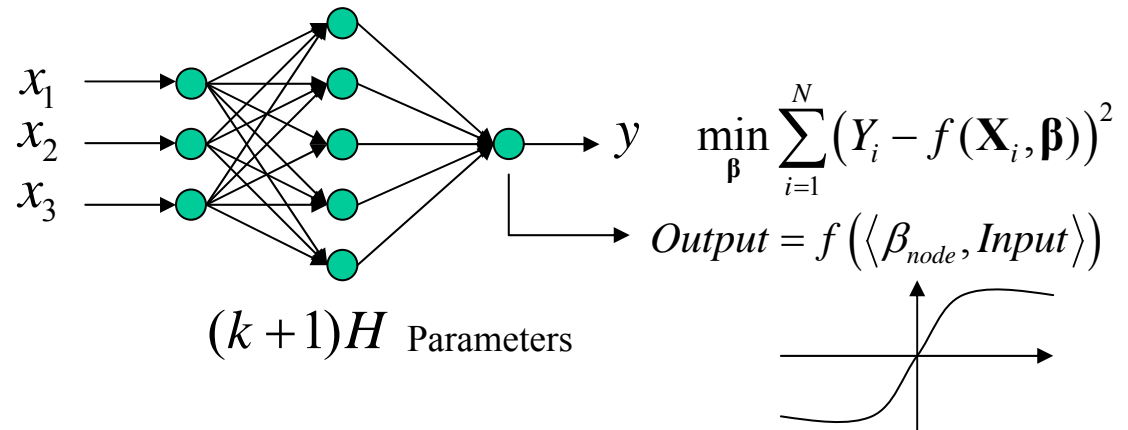
A different set of parameters is fitted for each quantile

$$\rho_{\tau}(x) = x(\tau - I(x < 0))$$

Indicator function

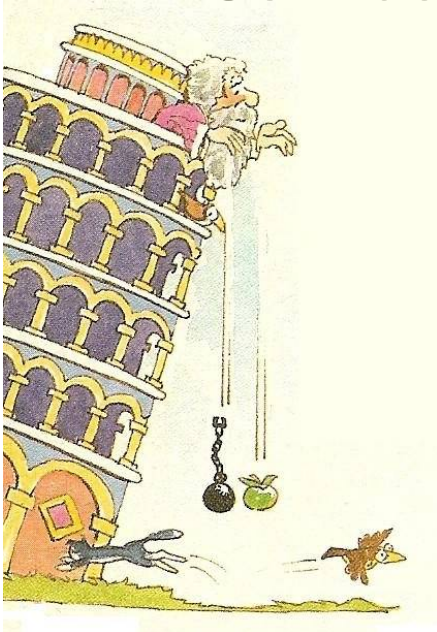
### Neural networks

Strong nonlinear regression





## 7.4.3 How reliable are the coefficients? Confidence intervals



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

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

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

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

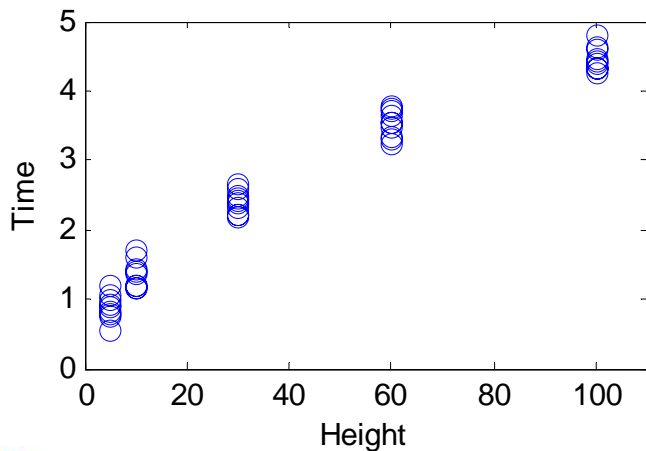
$$t(h) = -0.15 + 0.51\sqrt{h} + 0h + \varepsilon \quad R^2 = 0.9772$$

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

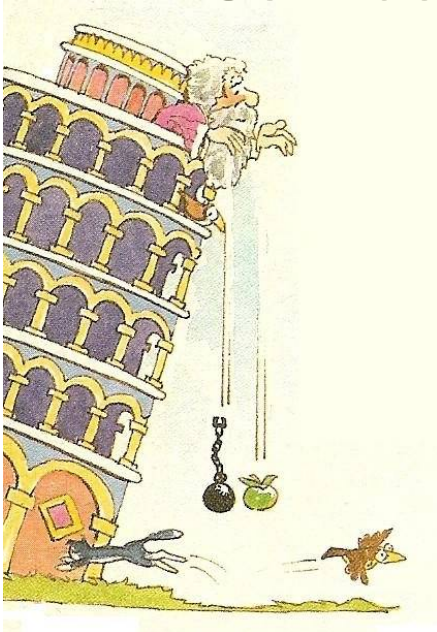
$$t(h) = 0 + 0.45\sqrt{h} + \varepsilon \quad R^2 = 0.9766$$

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

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



## 7.4.3 How reliable are the coefficients? Confidence intervals



Adjusted R: this is a way of reducing overfitting

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

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

$$R^2 = 0.9773 \quad R_{adjusted}^2 = 0.9760$$

$$t(h) = -0.15 + 0.51\sqrt{h} + 0h + \varepsilon$$

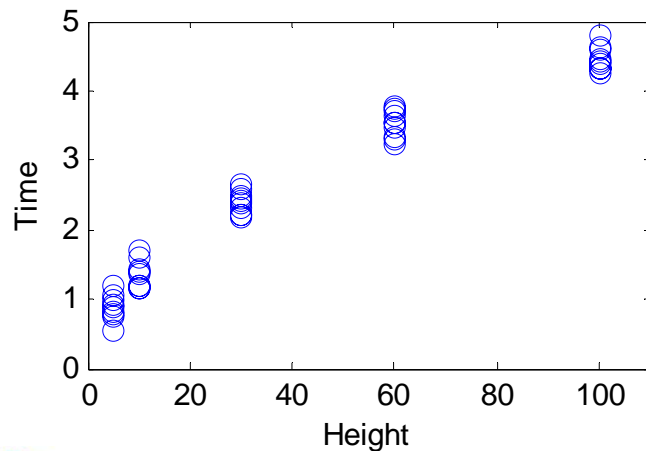
$$R^2 = 0.9772 \quad R_{adjusted}^2 = 0.9762$$

$$t(h) = 0 + 0.45\sqrt{h} + \varepsilon$$

$$R^2 = 0.9766 \quad R_{adjusted}^2 = 0.9759$$

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

$$R^2 = 0.9766 \quad R_{adjusted}^2 = 0.9762$$



## 7.4.3 How reliable are the coefficients?

### Confidence intervals

Can we distinguish between important coefficients and non important coefficients?

Linear regression

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

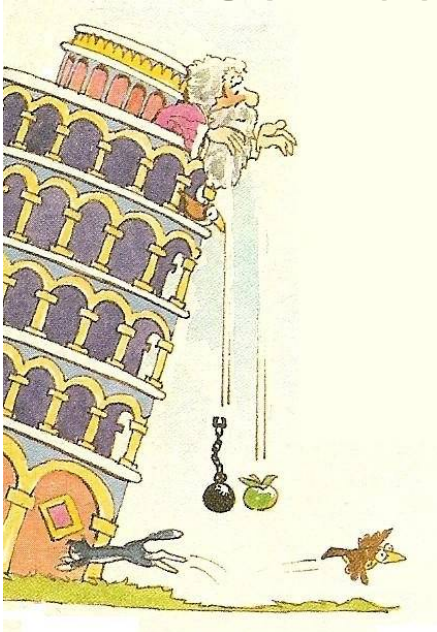
$$\boldsymbol{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \longleftarrow \text{Regression coefficients}$$

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

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

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

## 7.4.3 How reliable are the coefficients? Confidence intervals



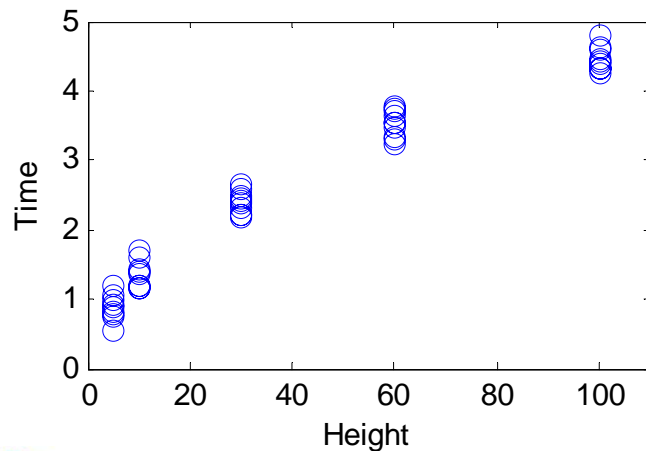
We climb to a few towers (with heights of 5, 10, 20, 30, 60 and 100 meters), let a ball fall 15 times and measure the time it takes to reach the floor. We know there is an error in our time measurement that is assumed to be normal, with zero mean, and a standard deviation of 0.2s.

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

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

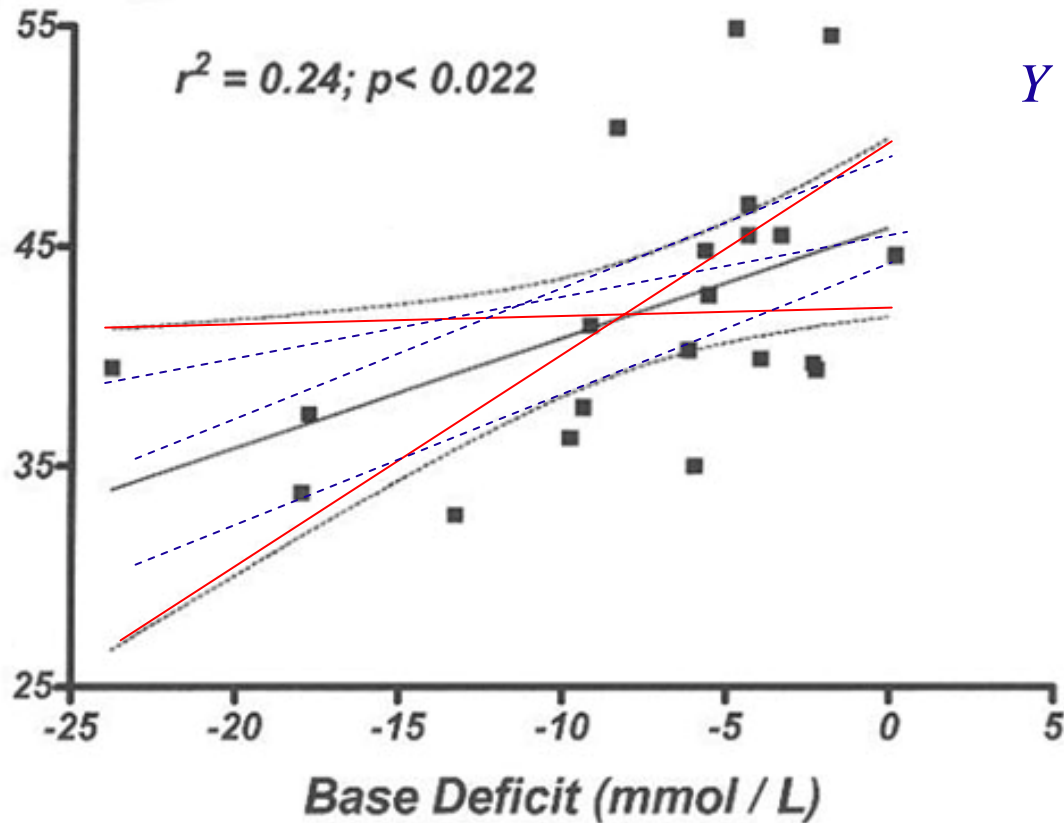
$$t(h) = [-0.90, 0.23] + [0.30, 0.93]\sqrt{h}$$

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



### 7.4.3 How reliable are the coefficients? Confidence intervals

■  $P_{\text{eCO}_2}$   
(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.

## 7.4.3 How reliable are the coefficients?

### Confidence intervals

What should the sample size be?

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

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

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

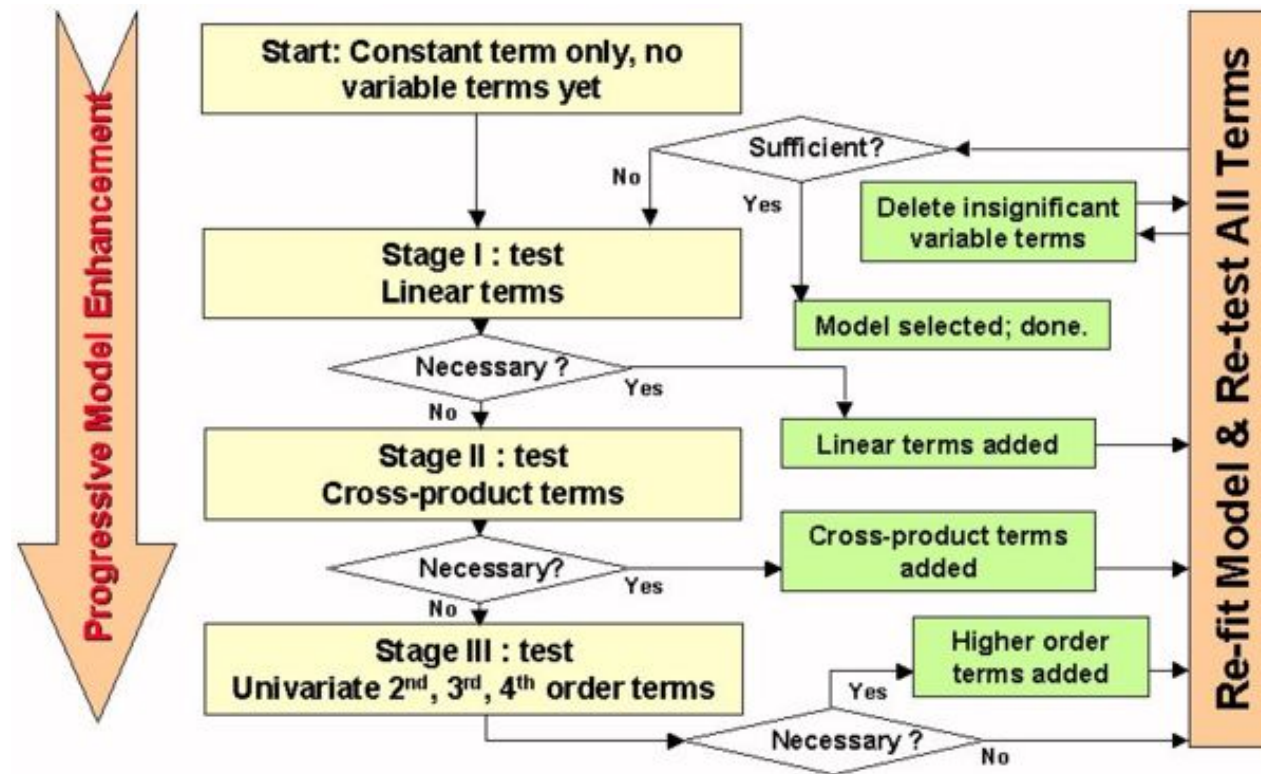
Determine minimum effect size for the regression coefficients ( $f^2$ ), the confidence level ( $1-\alpha$ ), the statistical power ( $1-\beta$ ) and the number of predictors  $k$ .

1. Compute degrees of freedom of denominator  $df = k + 1$
2. Compute  $F_{1-\alpha, k, df}$  with a central F distribution with  $k$  and  $df$  degrees of freedom
3. Compute the centrality parameter  $\lambda = f^2(k + df + 1)$
4. Compute the current power with a noncentral F (NCF)  $Power = \int_0^{F_{1-\alpha, k, df}} f_{NCF_{k, df, \lambda}}(x) dx$
5. 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>

## 7.4.3 How reliable are the coefficients? Confidence intervals

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

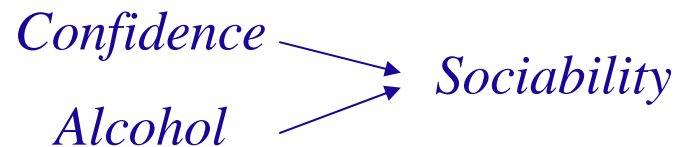


## 7.4.3 How reliable are the coefficients?

### Confidence intervals

#### 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!!!



## 7.4.4 How reliable are the coefficients? Validation

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.

# Conclusions



It is difficult to draw a single conclusion from this course. Probably the most important one is **Take Care!!!**

