Filippo Biscarini

PhD, Statistical geneticist, Bioinformatics department PTP - Lodi

Basic statistics using R

A bit of history

- public administration (surveys) → **descriptive statistics**
- chance games → **probability theory**

A.D. 1855

Statistical inference

- *Parametric methods and frequentist/"a posteriori" statistics* (R. Fisher, F. Galton, C. Peirce, K. Pearson …)
- *Bayesian/"a priori" statistics* (since 80s with Monte Carlo numerical methods)
- *Non parametric methods* (computer power, resampling, numerical/empirical statistics)

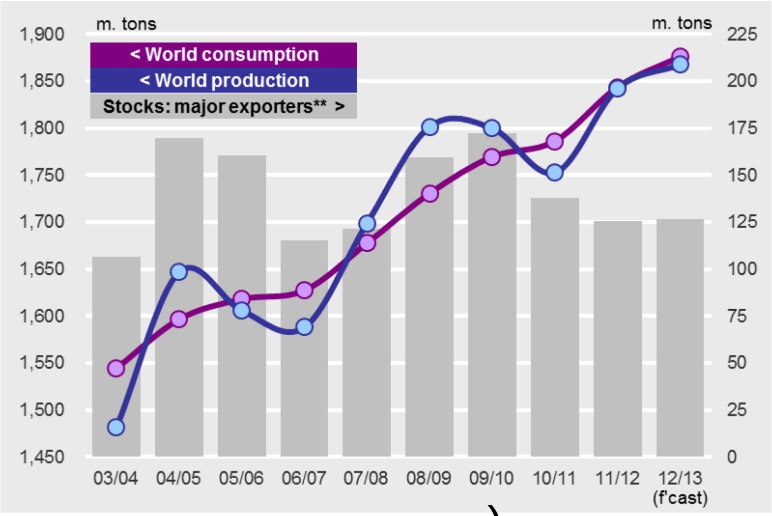
Introduction

Gathering, preparation, analysis, presentation and interpratation of **data**

- **Numeric** (e.g. milk kg) vs **categorical** data (e.g. sex, breed)
- **Nominal** data (e.g. sex), **ordinal** data (e.g. type score – "beauty contest", calving ease), **intervals** (e.g. dates), **ratios** (e.g. debt/GDP)
- **Continuous** (e.g. SCC) vs **discrete** data (e.g. SCS)
- Population → sample
- Bias (precision/accuracy)

Descriptive statistics

- “**Plots**”: very useful to summarise data or highlight interesting features *(but they are not the only nor the most important statistical tool!)*
- Bar plots (or Pareto diagrams), histograms, scatter plots, diagrams, dendrograms, pie charts etc …



Cereals (except rice)

- ↑ consumption, ↔ production, ↓ riserve
- bad weather (e.g. drought USA summer 2012)
- importance of stocks and international trade
- risks of protectionism and “km 0” (> volatility/shortage)
- especially in case of “climate change”

Descriptive statistics: measures of location

- Arithmetic mean: $$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$
  - can be computed for any set of numerical data
  - for any data set its value is unique
  - $(x_1 - \mu) + (x_2 - \mu) + \ldots + (x_n - \mu) = 0$
  - $\mu \cdot n = \sum x$
  - $$\mu_{tot} = \frac{n_1\mu_1 + n_2\mu_2 + \ldots + n_t\mu_t}{n_1 + n_2 + \ldots + n_t}$$
  - the arithmetic mean is affected by extreme values (outliers), not much by sampling fluctuations

"Problemchen": the sample mean is 40. After adding the values 50 and 64, the mean is 42. What is the size of the initial sample?

Descriptive statistics: measures of location

- Weighted average: $$\mu_w = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum w \cdot x}{\sum w}$$

- Grand mean (combined data): special case of the weighted average

- Geometric mean: **nth root of the product of the n values**

$$\mu_g = \sqrt[n]{\prod x}$$

- logarithmic identities (products → sums, power → products)

$$\mu_g = \exp\left[\frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i)\right] \qquad \left[e^{\ln(x)} = x\right]$$

Descriptive statistics: measures of location

- The geometric mean can be computed for positive values only
- The geometric mean of two numbers –a and b- is equal to the lenght of the side of a square whose area is that of a rectangle with sides of length a and b.
- The geometric mean is applied mainly when data are naturally multiplied and not summed (e.g. geometric progressions, ratios, interest rates, inflation rates …)
- Compared with the arithmetic mean, the geometric mean is more affected by small rather than large values. Specifically, one null value is sufficient to make it null.
- Somatic cell count (< 400000/ml) and germ count (< 100000/ml) in milk samples

Fibonacci integer sequence

- Leonardo Fibonacci (Italian mathematician, XIII sec.)
- Growth of a rabbit population
- $F_n := F_{n-1} + F_{n-2}$ with n>1
- Botany: flowers usually have 3, 5, 8, 13, 21, 34, 55 or 89: lilies 3 or 5, buttercups 5, larkspurs 8, marigolds 13, asters 21, daisies 34 or 55 or 89, etc …

  The leaves on the branches are placed so not to cover each other, thus allowing every leaf to receive sun light. The number of leaves between two perfectly aligned ones along a branch, is a Fibonacci number.
- Art (golden ratio), economics, informatics, music, geometry (fractals, Fibonacci's spiral), chemistry, etc …)

Measures of location: the median

- Value-ordered data (ascending or descending)
- Value of the central observation

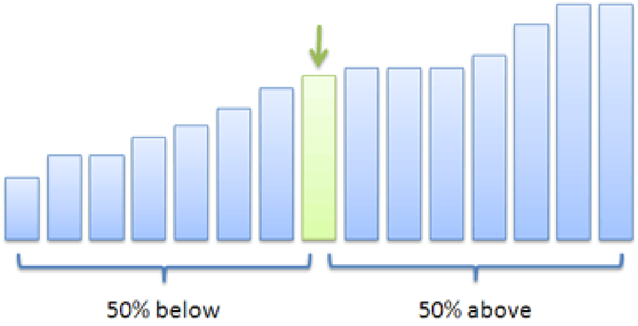
348 338 351 346 342 351 339 355 351 345 351 344 340

338 339 340 342 344 345 **346** 348 351 351 351 351 355

- Odd sequence: value of the observation with position $\frac{n+1}{2}$

- Even sequence: mean value of the two observations with positions [n/2] and [(n/2)+1]

Median



Measures of location: median and other quantiles

- The median is not affected by extreme values (the mean is)
- The median is a special **quantile**: it is the quantile that splits in two halves the ordered distribution of data
- Other quantiles are quartiles, deciles, percentiles
- **Boxplot**: based on **5** synthetic indexes (median, Q1, Q3, max, min)
- Box (interquartile range), whiskers (1.5 x box length), outliers

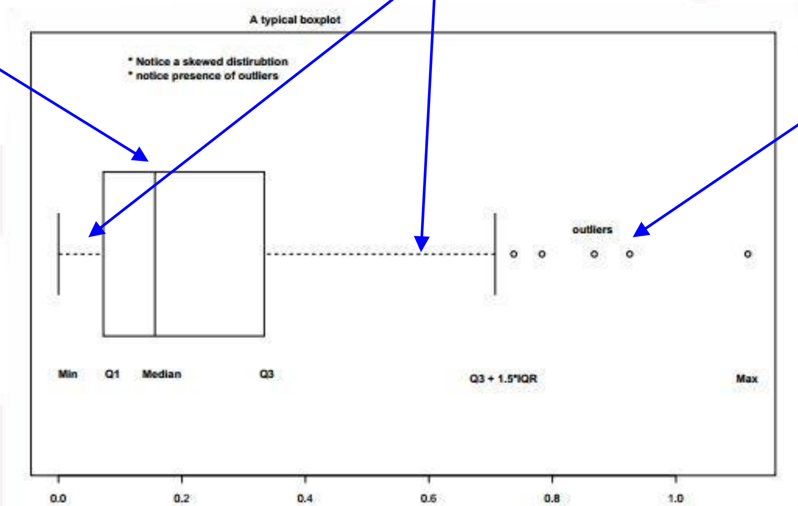


Figure 5: A typical boxplot

Measures of location: the mode

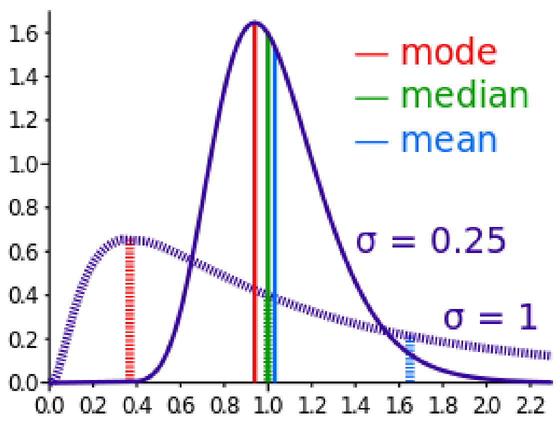
Parco Tecnologico Padano

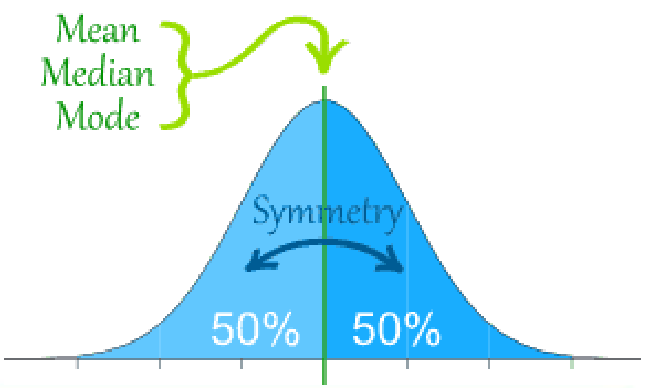
Competere per l'eccellenza

- **mode**: it is the most frequent value in the distribution

<u>BRUC</u>, TBC, TBC, BLUET, <u>BRUC</u>, <u>BRUC</u>, BLUET, TBC, <u>BRUC</u>, BLUET, <u>BRUC</u>

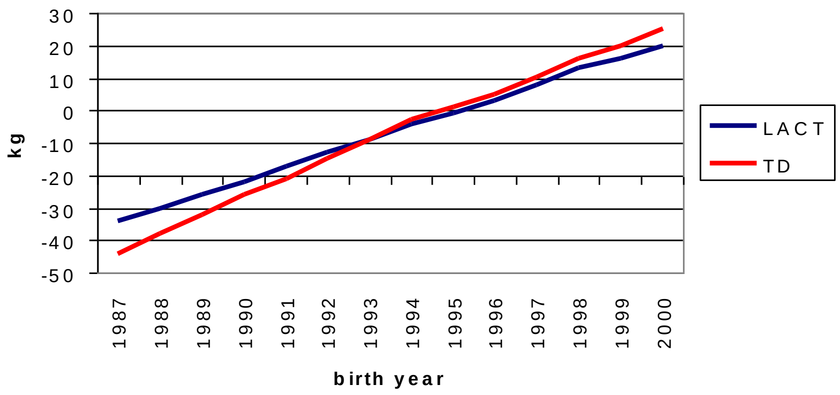
Mode (Most Popular)

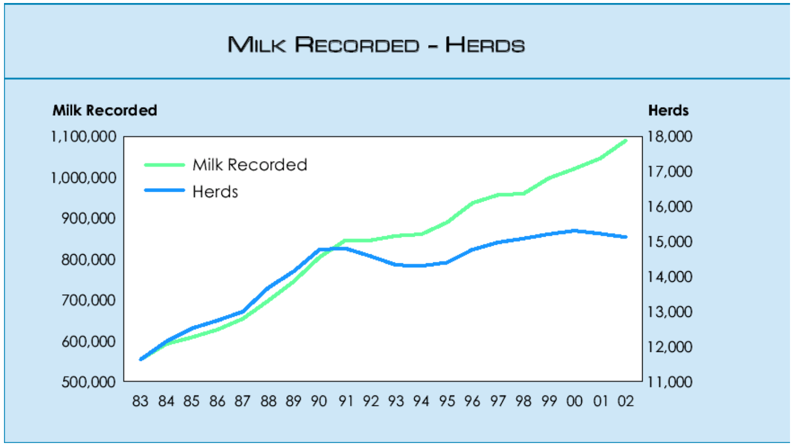




Descriptive statistics – more basics

- average over time → trend;
- "representativity" (sampling)





Descriptive statistics: imeasures of variation

Data distributions show variability (e.g. genotyping is not always successful – call rate)

1. distribution with the same mean → range (max – min)
2. distribution with the same range → different data dispersion

Variance

$$\sigma^2 \approx s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Standard deviation

$$\sigma \approx s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Why? –d.f.

Measures of variation: the standard deviation

- small σ → data close around the mean
- large σ → data scattered away from the mean

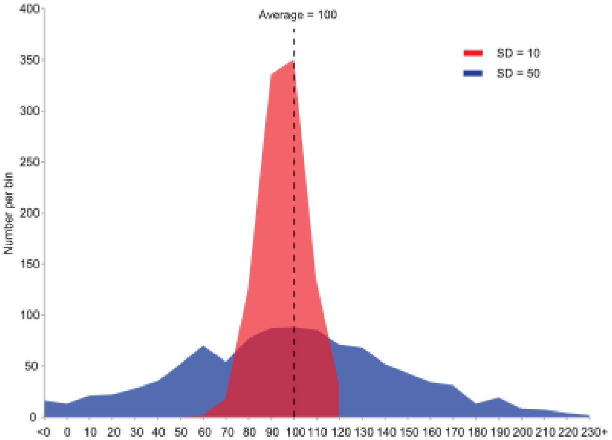
Chebyshev's theorem: for <u>any data distribution</u> and any constant $k > 1$, at least $1-1/k^2$ observations will lie within k standard deviations around the mean

$$1 - \frac{1}{2^2} = \frac{3}{4} = 0.75$$

→ at least 75% of the observations lies within $\mu \pm 2\sigma$

$$1 - \frac{1}{3^2} = \frac{8}{9} = 0.89$$

→ at least 89% of the observations lies within $\mu \pm 3\sigma$



<u>Normal distributions</u>

- 68% of the observations within $\mu \pm 1\sigma$
- 95% of the observations within $\mu \pm 2\sigma$
- 99.7% of the observations within $\mu \pm 3\sigma$

Pafnuty Chebyshev (1821-1894)

Descriptive statistics: measures of variation

- Compare two distributions (e.g. marks in different courses)
- **Data standardization** (standard units, z)

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

- **Scale change**

$$ScaledData = \frac{(newMax - newMin) \cdot (x - oldMin)}{(oldMax - oldMin)} + newMin$$

  - estimates on the same scale are easier to compare
  - changing scale may help in the interpretation of data/results
  - > numerical stability when variables are on the same scale

- Much or little variability?
- Coefficient of variation (e.g. chickens or cows live weights)

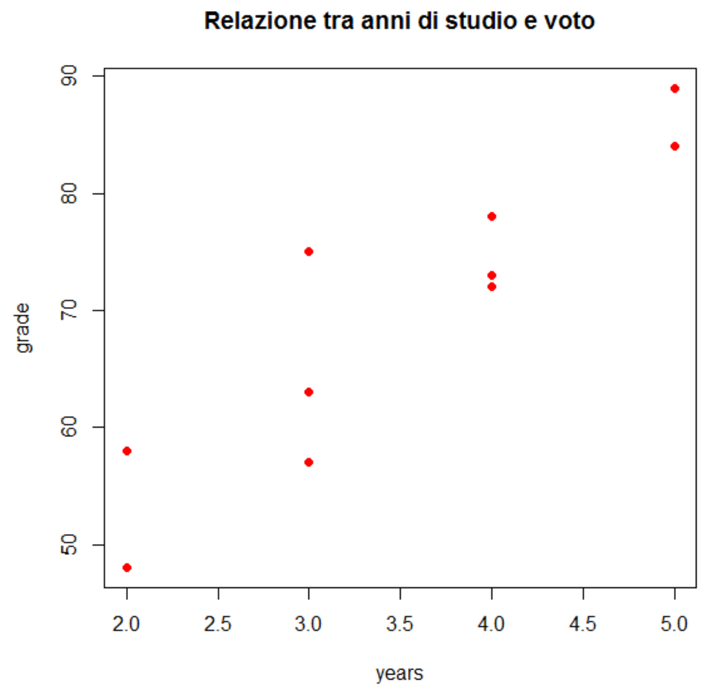
$$CV = \frac{\sigma}{\mu} \cdot 100$$

Covariance and correlation

- *How much of the total variability of* ***y*** *(marks) is due to chance, and how much to its relationship with* ***x*** *(years of study)?*

$$Cov_{xy} = \sigma(x, y) = \frac{\sum (x - \bar{x}) \cdot (y - \bar{y})}{n - 1}$$

- if large values of **x** match large values of **y**, and the same holds for small values –i.e. the two variables behave alike- then the covariance is positive



- in the opposite case the covariance is negative –the two variables behave opposite
- the sign of covariance gives the trend of the linear relationship between the variables
- the size fo covariance is not easy to interpret → **correlation**

Covariance and correlation

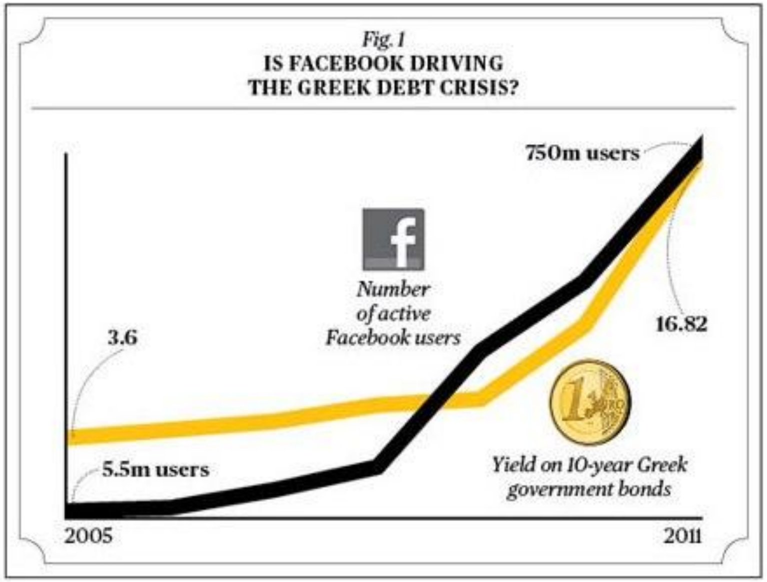
Parco Tecnologico Padano

Competere p...

- Normalizing the covariance we obtain the coefficient of correaltion

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

- Correlation vs causality
  - E.g. lighters/lung cancer
  - Police/crime
  - Facebok/Greek debt
- Analyse variables “ceteris paribus”
- Experimental method
- Etc … (which variable changes the first etc …)
  - Beware: e.g. family cars/children birth



Chance

- XVII century: chance games (limitations of Galileo-Newton's deterministic method): Chevalier de Méré, Pascal, Fermat
- Gas kinetic theory

- Arthur Stanley Eddington: typewriting monkeys
- Einleitung: 1694 characters; 26 letters, + 3 (umlaut), + punctuation= ~ 45 symbols (not considering upper/lower case) → 45^{1694} (10^{80})

Vortragstagung der DGfZ und GfT am 6./7. September 2011 in Weihenstephan

Einfluss der genetischen Architektur auf die empirische Genauigkeit der genomischen Zuchtwertschätzung

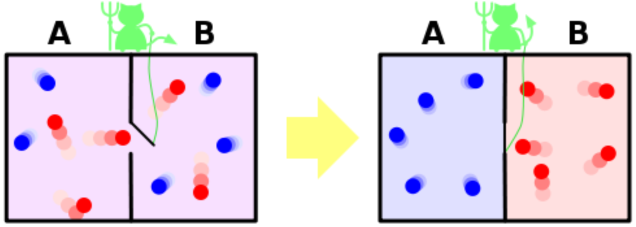
M. Kramer[1], F. Biscarini[1], B. Bapst[2], C. Stricker[3], H. Simianer[1]

[1]Abteilung Tierzucht und Haustiergenetik, Department für Nutztierwissenschaften, Georg-August-Universität Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen
[2]QUALITAS AG, Chamerstrasse 56, Zug 6300, Schweiz
[3]agn Genetics GmbH, Börtjistraße 8b, 7260, Davos, Schweiz

1. Einleitung

Wie von INTERBULL (2011) angegeben haben die GBLUP Methode (MEUWISSEN ET AL., 2001) sowie verschiedene Bayes Verfahren (GIANOLA ET AL., 2009) derzeit die größte praktische Bedeutung in der genomischen Zuchtwertschätzung. Anhand simulierter Daten haben DAETWYLER ET AL. (2010) gezeigt, dass diese Verfahren je nach Anzahl der Tiere im Trainingsset (N_p), der Effektiven Populationsgröße (N_e), der Heritabilität des betrachteten Merkmals (h^2) und der Anzahl effektiver Gene, die ein Merkmal beeinflussen (N_G) unterschiedli-

- Maxwell's demon



- Unlikely events, not impossible!

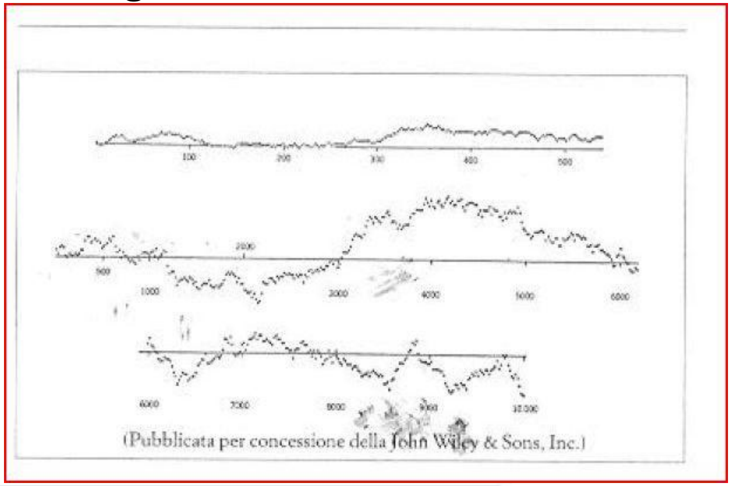
Chance

- Random process→ ordered result (e.g. typewriting monkeys)
- Deterministic process → unpredictable result (e.g. mathematical formulas to generate “pseudo-random” numbers)

- 250 times a coin is tossed:
  - 16 sequences of at least 3 consecutive heads (or tails)
  - 8 sequences of at least 4 consecutive heads
  - 4 sequences of at least 5 consecutive heads
  - 2 sequences of at least 6 consecutive heads
  - 1 sequence of at least 7 or more consecutive heads
- A pseudo-random sequence created by a person usually contains too few long sequences
- What about nucelotide sequences?

Chance

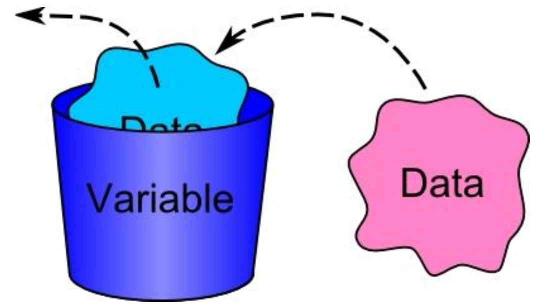
- In the long run, the **ratio** between red and black outcomes of the roulette will tend towards unity (but their **absolute number** will not necessarily tend to get closer)
  - This is why it is wrong to bet on a “late” colour -or number (at each round the probability is the same)
- Unlikely event, but high potential loss → great risk [expected value = probability x sum at stake]
  - This is why it is wrong to always double the sum at stake, hoping to make up for the loss (high probability of winning a little, small probability of losing a lot)
  - A bit like risk and hazard in biology

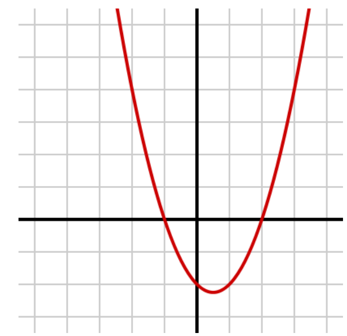


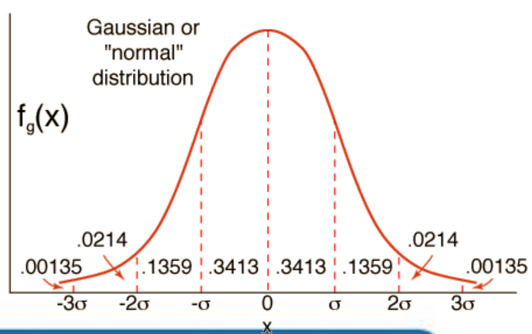
(Pubblicata per concessione della John Wiley & Sons, Inc.)

Random variables

- Variable (mathematics/informatics): $y = f(x)$
- e.g. parabole: $y = ax^2 + bx + c$
- Random (or stochastic) variable: can take different values, each with a given probability → **sample space** (Ω) and **probability distribution**
- e.g. results of scientific experiments, chance games, stochastic events (gas kinetic theory, stock market values, meteorology etc …)
- functions of random variables (they're random variables themselves)
- es. x^2 → χ^2
- several samples from a normal distribution, each squared

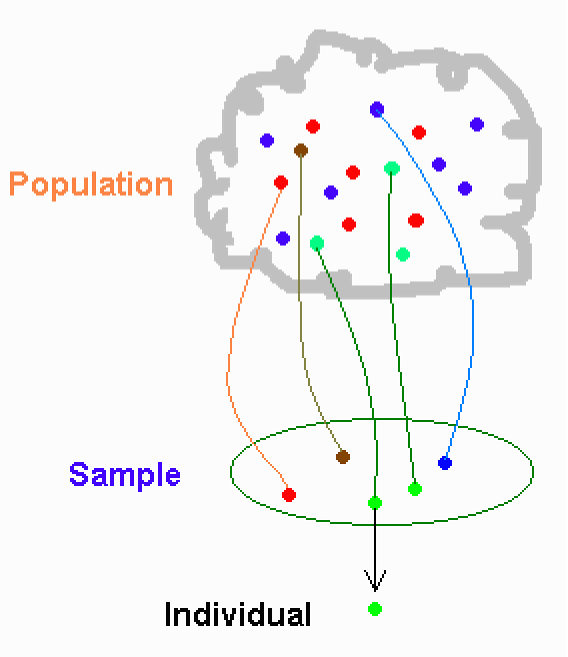






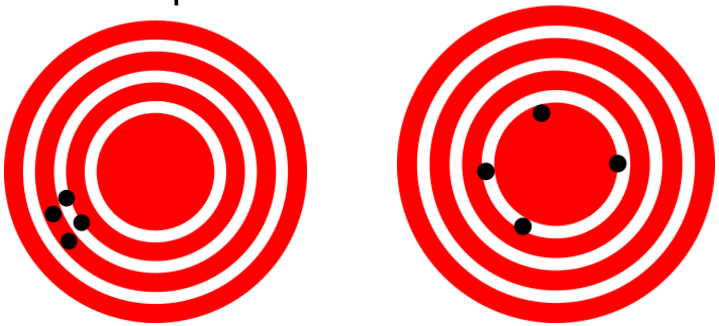
Sampling

- Choice of a population subset to describe or estimate the population characteristics
  - Define the population
  - Sampling method
  - Sample size
  - Sampling plan
  - Actual sampling and data collection
- Sampling method (representativity):
  - **random** (< bias, beware of substructure –e.g. M/F)
  - **sistematic** (every k elements: beware of periodicity e.g. house numbers)
  - **stratified** (classification factors that are relevant for the variable to be studied)
- Sampling with or without replacement
- Sample size → statistical power



Sampling errors – an example

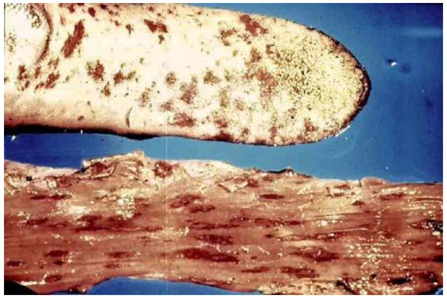
- 1936, USA election poll:
- > 2.4 million observations (from telephone directories, magazine sbscriptions)
- predicted outcome: Landon 57%, Roosvelt 43%
- results: Landon 38%, Roosvelt 62%
- very large sample (high statistical power)
- but not representative!
    - only middle/upper class had telephones or magazine subscriptions
    - low response rate (2.4 out of 10 million) → non-response bias
- → severe bias! (high precision around a biased result!)



Select animals for genotyping

<u>Case study</u>

- Cattle population from BVD outbreak (pestivirus, *Flaviviridae*)
- Infected – healthy animals → antibody titre
- Animals of different age, in different herds and husbandry groups
- Select animals to be genotyped (cases and controls)



How should they be sampled?

1. Verify whether the identified classification factors are relevant for the object of the analysis (antibody titre) → linear model/analysis of variance
2. Stratified sampling (along identified dimensions), otherwise random sampling

Analysis of variance

Partition **total variability** into components in due to **different causes**

- E.g. observed differences in the prevalence of a disease among sheep populations are due to breed, husbandry, diet, geographical region, demographic structure etc …

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})^2 \longrightarrow$$

Total sum of squares (total deviance): k factors, n samples. $\bar{x}_{..}$ Is the total (grand) mean

$$SST/(kn-1) \longrightarrow$$

Total variance

Mean of each k-group

Between-groups sum of squares

$$SST = n \cdot \sum_{i=1}^{k}(\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})^2$$

Within-groups sum of squares

Analisis of variance

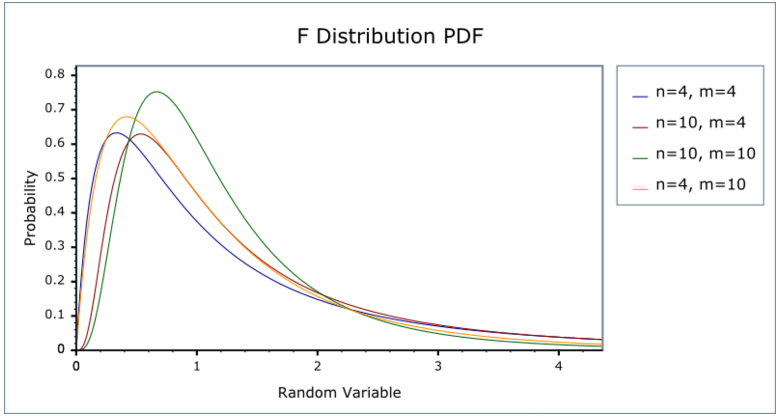
- SST = BG + WG = SS(Tr) + SSE ← basic partition of total variability

$$MS(Tr) = \frac{SS(Tr) = BG}{k-1}$$ → Mean square of treatments

$$MSE = \frac{SSE = WG}{k(n-1)}$$ → Mean square fo the error

$$F = \frac{\text{between - group variation}}{\text{within - group variation}} = \frac{MS(Tr)}{MSE}$$

Fisher-Snedecor random variable (F – distribution) with (k-1) and k(n-1) d.f.



Anova table

Source of variability	d.f.	Deviance	Mean square	F
Treatments	*k-1*	SS(Tr)	MS(Tr)	MS(Tr)/MSE
Error	*k(n-1)*	SSE	MSE	
Total	*kn-1*	SST		

Source of variability	d.f.	Deviance	Mean square	F
Fertilizer	3-1=2	456	228	15.8
Error	3(4-1)=9	130	14.44	
Total	3*4-1=11	586		

$$15.8 > 8.02, \ \text{F value for } \alpha = 0.01 \longrightarrow$$

- Differences between fertilizers are too big to be attributed to mere chance
- Fertilizers do not have all the same effect (at least one is different)

Analisis of variance – the model

$$y_{ik} = \mu + FERT_k + e_{ik} \equiv \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

observations (dependent variable)

mean (intercept)

factor (independant variable)

residuals

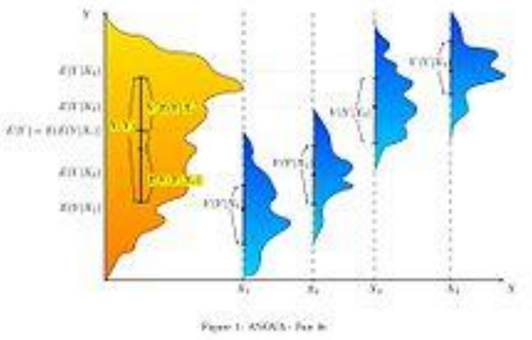
- systematic part– random part
- assumptions (base hypotheses):
  - linearity (in the parameters: intercept and coefficients, ß)
  - independence (of residuals)
  - homoscedasticity (homogeneity of variance)
  - normality

$$e_{ij} \text{ are i.i.d. } \sim N(0, \sigma^2)$$

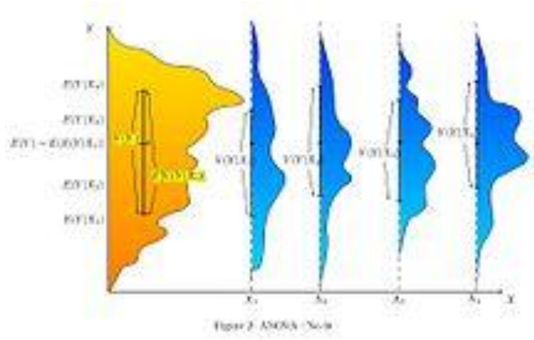
Analisis of variance

$$R^2 = 1 - \frac{SSE}{SST}$$

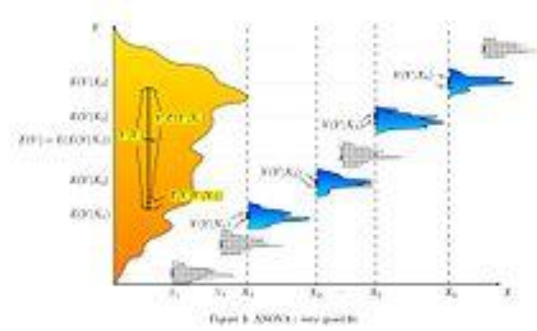
<u>Coefficient of determination</u> (fraction of the variability that is explained by the model)



FAIR FIT



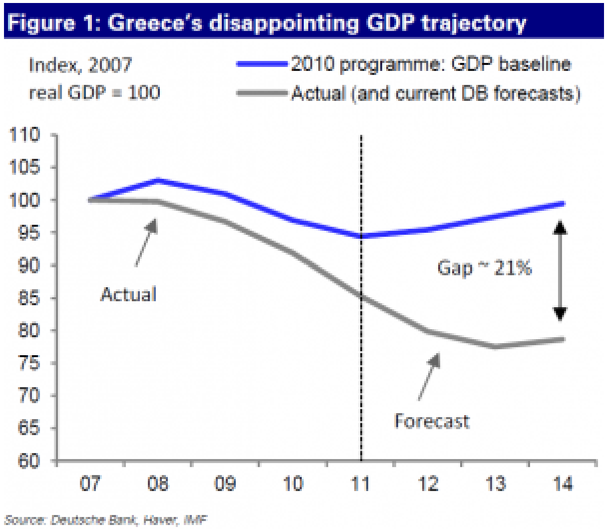
NO FIT



VERY GOOD FIT

Linear regression

- predict one variable (unknown or future observation) as function of another variable (explanatory variable)
    - family budget on holidays as function of family income
    - sheep wool quality as function of the diet
- average predictions (expected –not punctual- values)
    - e.g. average production (expected) of a variety of wheat as function of spring rainfall
- **linear regression** (predict the expected value of a variable as function of another variable)
- relationship between variables: strong, fair, weak
    - → **correlation**



Regression curve

$$V = \frac{p}{k}$$ → Boyle's law

$$N_t = k \cdot 2^t$$ → bacterial growth (exponential phase)

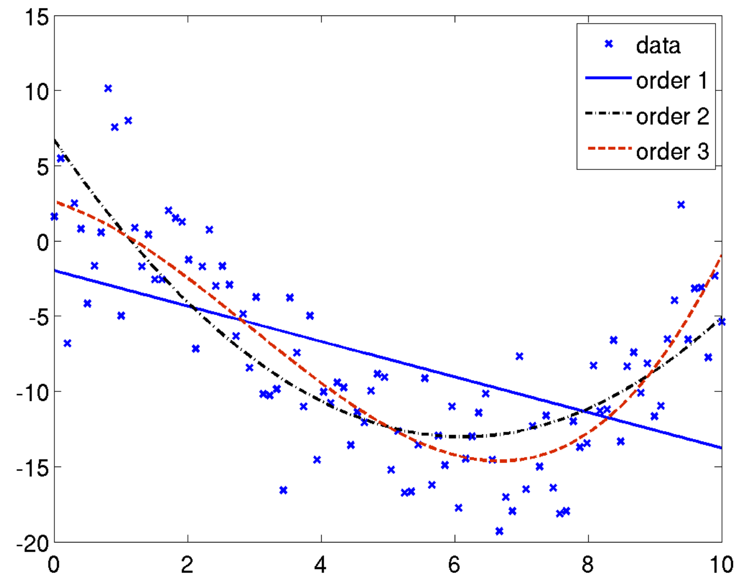
- Interpolating points on a plane (*curve-fitting*)

intercept

slope

$$y = a + bx$$

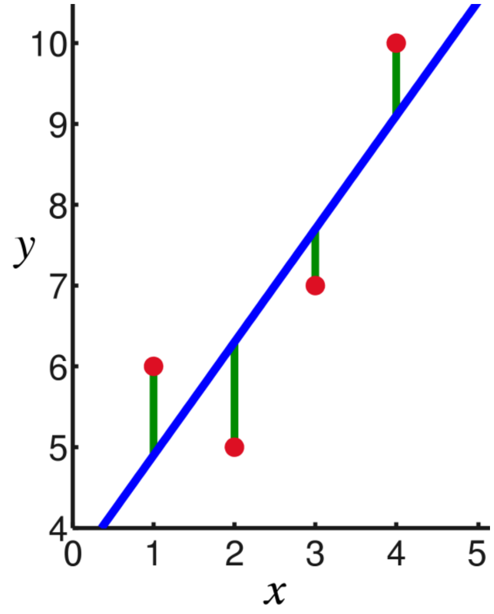
- Regression curve (linear equation)
- Most relationships are linear
- Good approximation to non-linear relationships



Least squares

- How can data be interpolated (several methods)
- Least squares method
  - Adrien-Marie Legendre (1806: published!), Carl Friedrich Gauss (1795, 18 years old)

$$\hat{y} = a + b \cdot x$$



- Many possible regression curves
- Find the one with the smallest difference between observed and predicted values
- a and b that minimize the sum of the errors
  - This involves differentiation of matrixes and vectors

$$\sum (y - \hat{y})^2 = \sum [y - (a + b \cdot x)]^2 = (\mathbf{y} - \mathbf{Xb})^{\mathbf{T}} (\mathbf{y} - \mathbf{Xb})$$

Normal equations

$$\sum y = na + b \cdot \left(\sum x \right)$$

$$\sum xy = a \cdot \left(\sum x \right) + b \cdot \left(\sum x^2 \right)$$

$$4055 = 35 \cdot a + 32275 \cdot b$$

$$4617544 = 32275 \cdot a + 37987805 \cdot b$$

- Solve the system: e.g. elimination method
    1. l.c.m. (least common multiple)
    2. multiply both equations by l.c.m./(n o Σx)
    3. subtract corresponding elements
    4. solve for b
    5. substitute b in the first equation

Defining functions in R


```
function_name <- function(function_arguments) {
            function_body
            function_return_value
}
```

keyword "function"

name (beware not to use existing names!)

```
std <- function(x) sqrt(var(x))
```

<u>Invoking the function:</u>

```
data <- c(1,2,3,4,5,6,1,1,1,2);
std(data);
```

The function arguments

Parco Tecnologico Padano

Competere per l'eccellenza

- No arguments

```
hello.world <- function() print("hello world")
```

- An argument

```
hello.someone <- function(name) print(paste("hello ",
name))
```

- A default argument

```
hello.world <- function(name="world")
print(paste("hello ",name))
```

- Multiple arguments

```
sim.t <- function(n,mu=10,sigma=5) {
    X <- rnorm(n,mu,sigma);
    (mean(X) - mu)/(sd(X)/n)
}

sim.t(40,5,10)                  #positional arguments
sim.t(sigma=0,n=10,mu=1)        #named arguments
```