Useful material for the course

Suggested textbooks:

•Mood A.M., Graybill F.A., Boes D.C., Introduction to the Theory of Statistics. McGraw-Hill, New York, 1974. [very complete]
•M.C. Whitlock, D. Schluter, Analisi statistica dei dati biologici, Zanichelli, Bologna 2010 [focussing on (population) biology problems]
•S. M. Iacus, G. Masarotto, Laboratorio di statistica con R, McGraw-Hill, 2006 [practical books (in Italian) on using R for statistics]

Software for statistics:

My advice is to use R   http://www.r-project.org/
a programmable environment suitable for statistics.
Many simple things can be done using Excel, or similar software...
[I will not teach how to use software, but will show some examples of R]

These notes and programs will be available at

http://www.science.unitn.it/~pugliese/
http://www.science.unitn.it/%7epugliese/

Statistics

Descriptive	Inferential
Aim: present useful information on the data	**Aim:** understand the mechanism that generated the data
Methods: *histograms, mean, variance* for univariate data. More complex for multivariate data	**Methods:** *point estimates, confidence intervals, hypothesis testing, analysis of variance...*

Some problems that can be tackled with inferential statistics

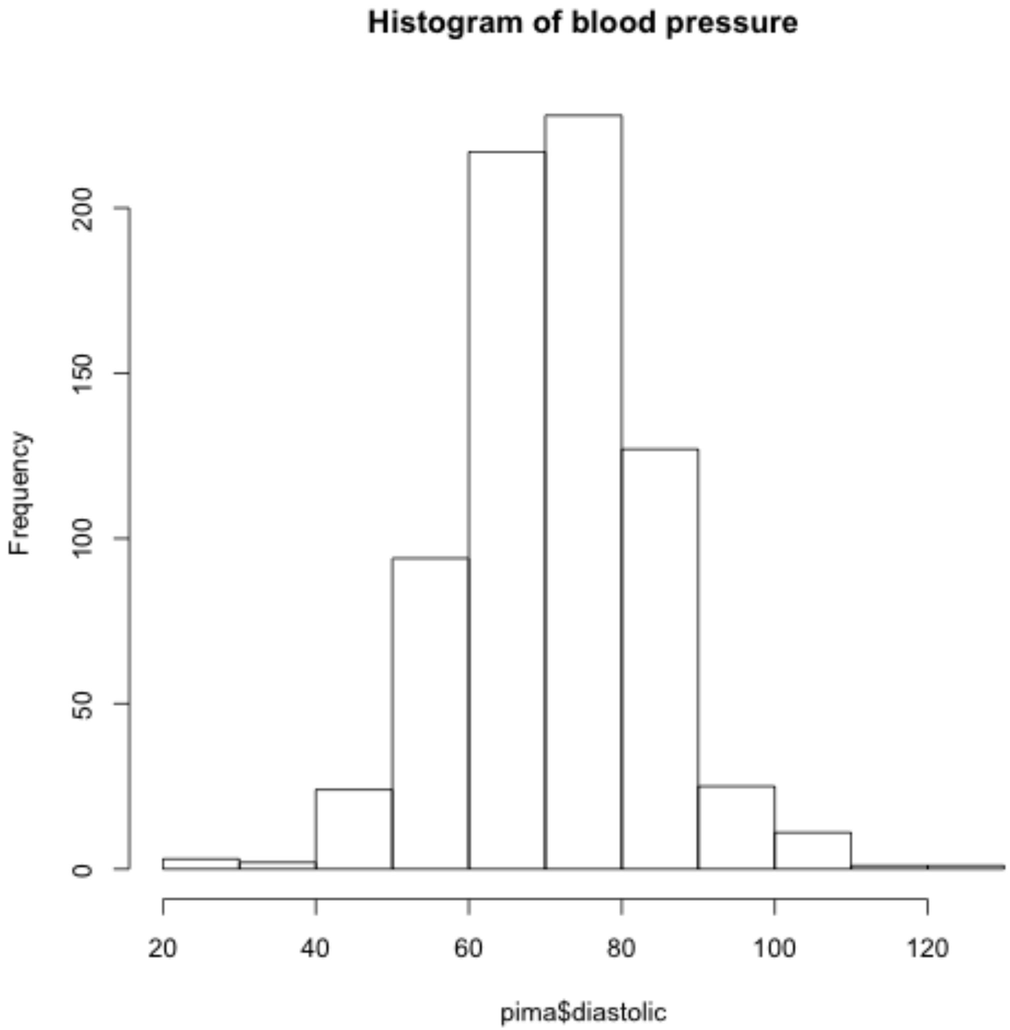
- Can I say whether the experimental group has a lower risk of heart attack than the control group? or has a lower blood pressure? and of how much?
- How large should I choose the two groups to be able to detect an effect of treatment?
- Which is the precision associated to a measurement performed?
- Is there a (linear) relationship between chlorophyll concentration and photosynthetic rate?

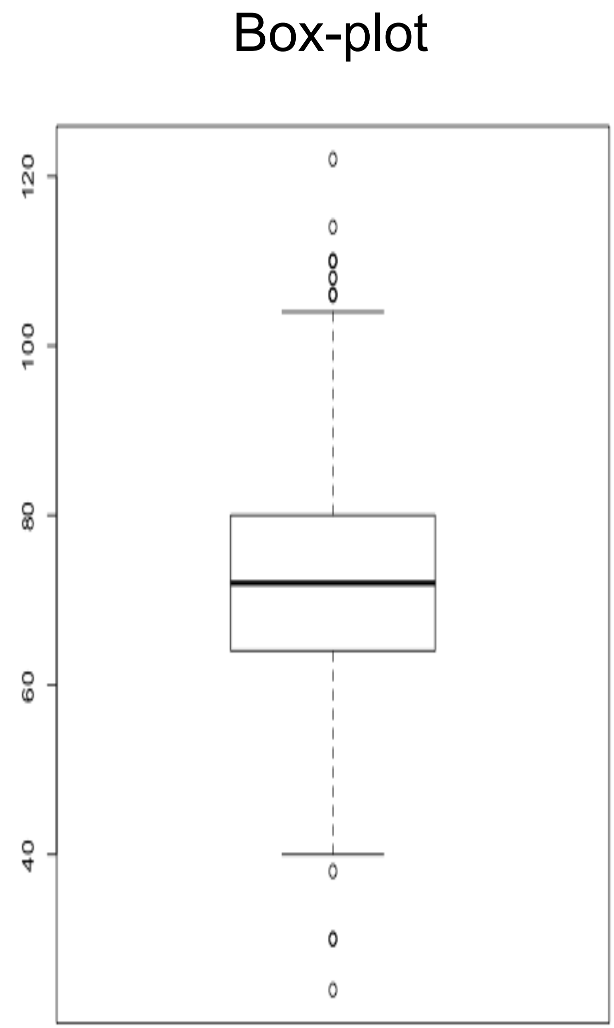
These questions involve *experimental design* and mathematics. I will (almost) only deal with the latter.

Summary statistics from data

- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- Variance: $V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (reason for dividing by $n-1$ clear with inferential statistics).
- Median: m, the value such that 50% of the data are below m, and 50% are above m (a precise computation depends on whether the number of data is odd or even...)
- Quantiles: q_α is the value that a fraction α of the data is below q_α and $1-\alpha$ is above (the median is the 50% quantile).

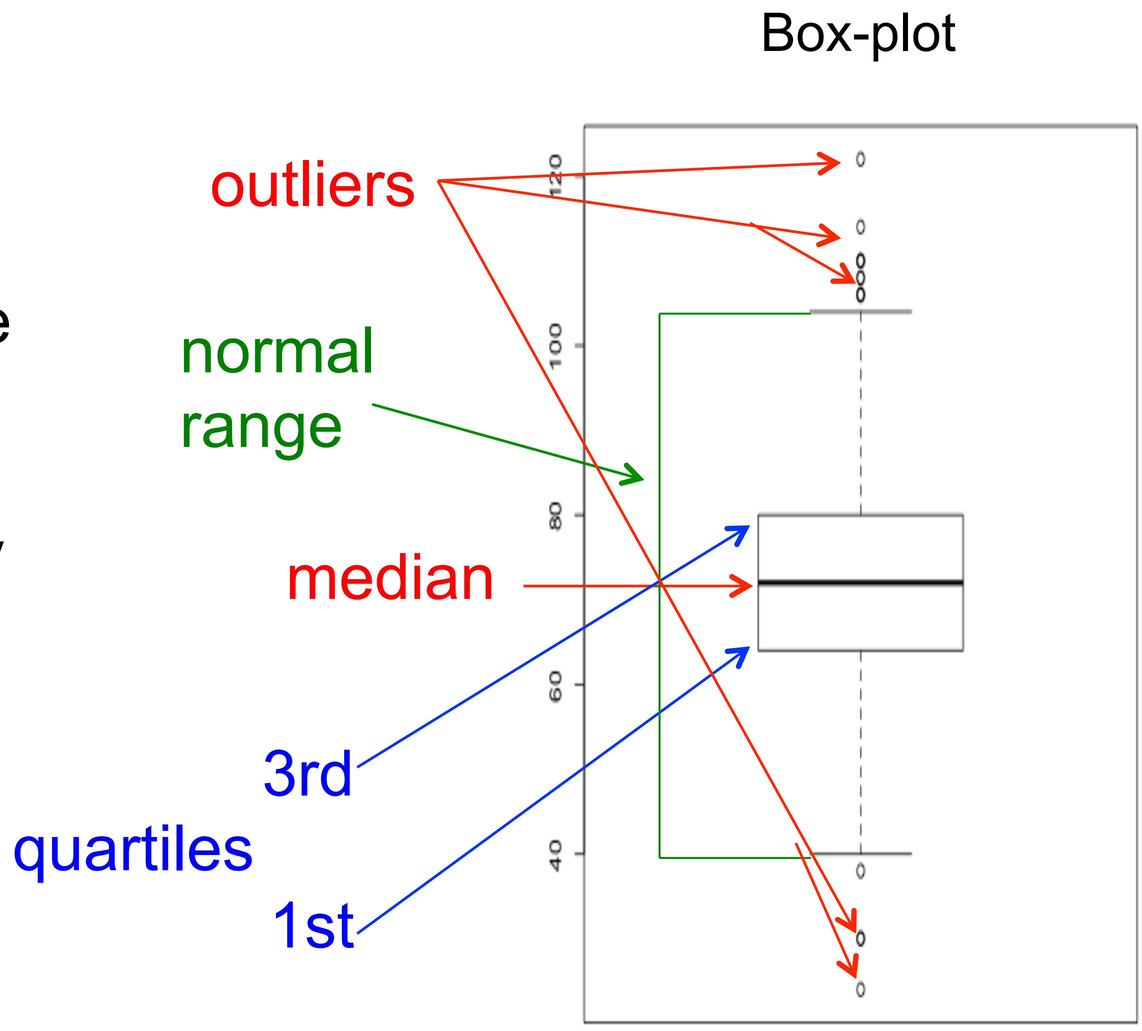
Description of continuous variables



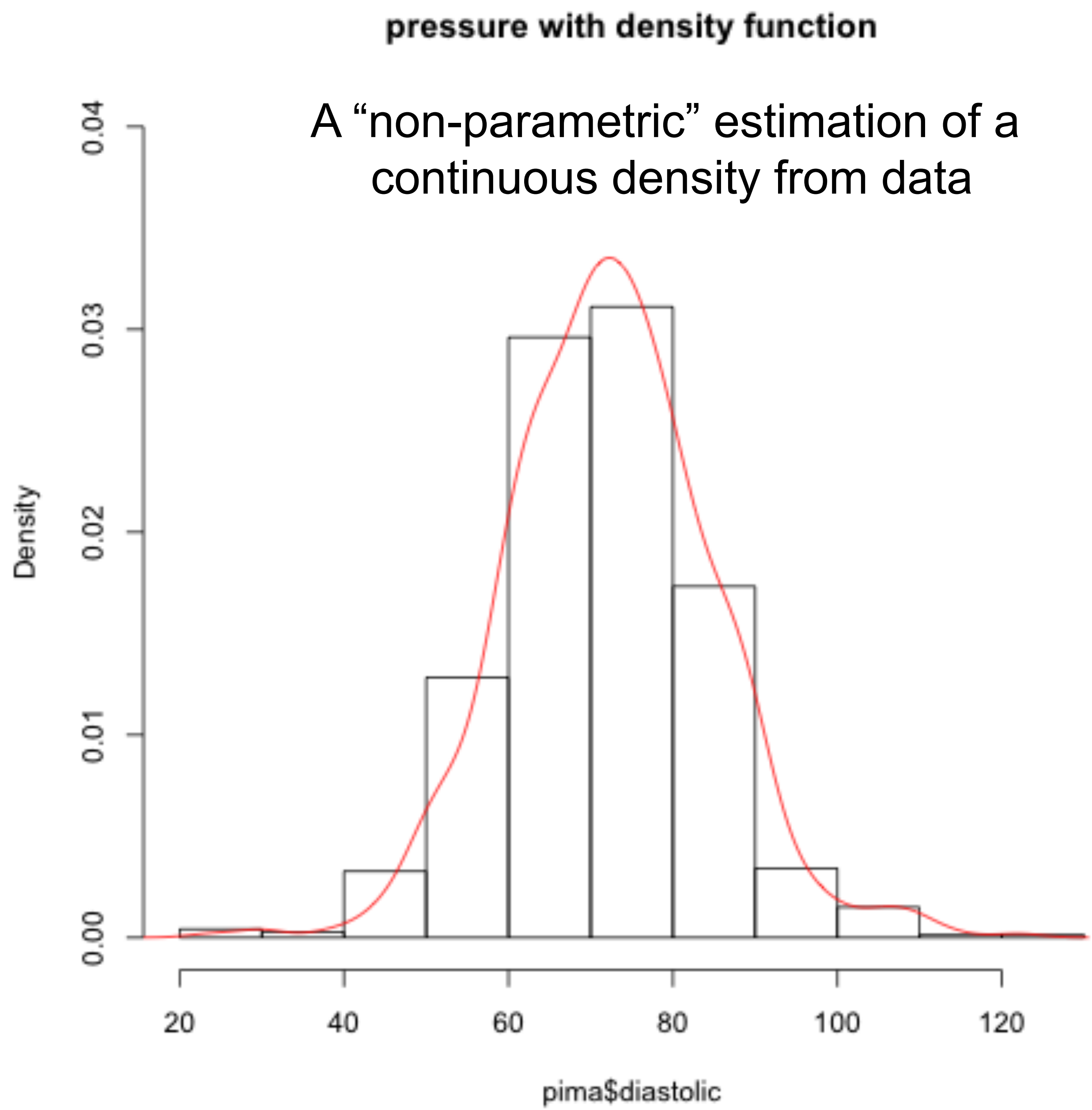


Reading box-plots

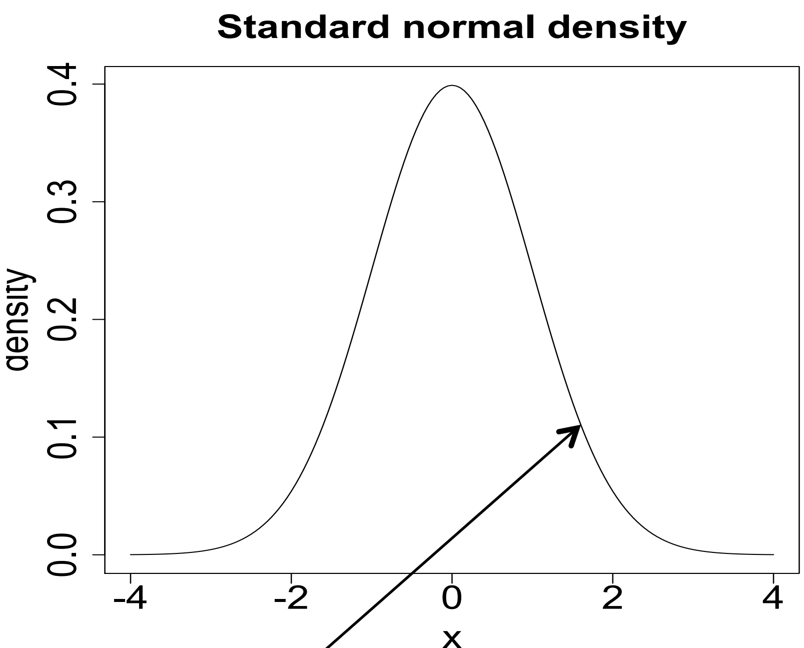
A useful tool to summarize information on the distribution of a variable
(we can put many side by side)

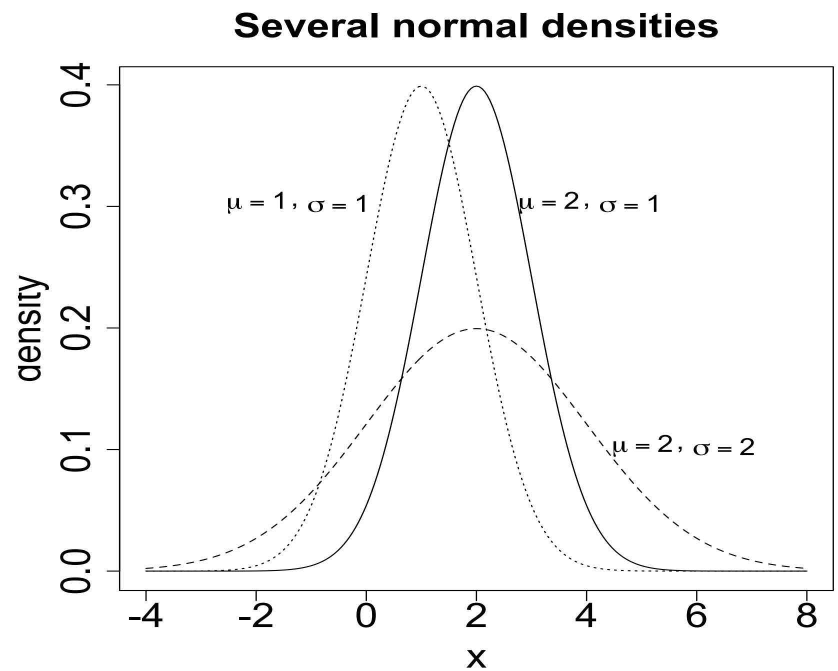


A “non-parametric” estimation of a continuous density from data



Normal (or Gaussian) distribution

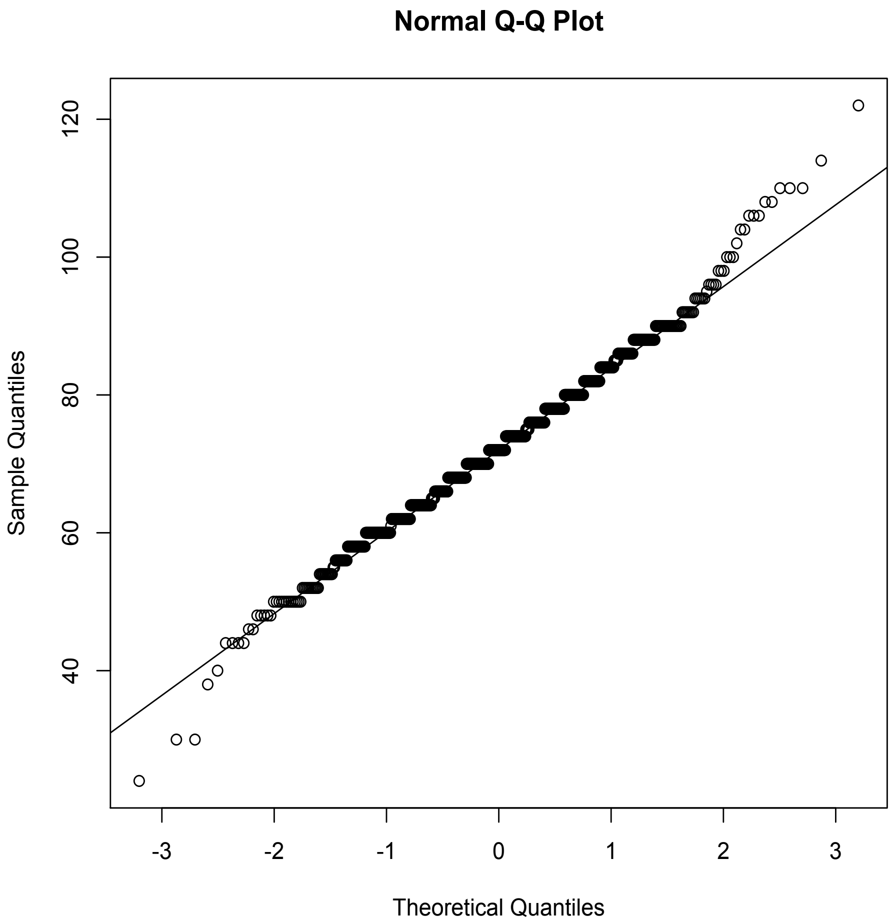




$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Visually comparing a distribution to a normal (Q-Q plots)



compares theoretical quantiles (those of a standard normal) to observed quantiles.
If data were normally distributed, points should lie on a line *(a line is added to help visual impression)*

Summary of methods in ` univariate descriptive statistics

- Mean, variance, median (summary indices)
- Quantiles …
- Histogram, box-plots
- Empirical density
- Comparison with a normal distribution
- Q-Q plot (to compare two distributions, in particular data with a normal)

- Thumb rule: approximately 2/3 of a distribution lies between *E(X)-sqrt(V(X)*) and *E(X)+sqrt(V(X))*

Basic probability

Inferential statistics is based on probability theory (we do not have certainty, but only *confidence*).

- Events: something that may or may not happen: A; $\mathbb{P}(A)=$ *probability* that A happens;

For instance $\mathbb{P}$(it rains tomorrow in Trento); $\mathbb{P}$(there is at least one son in a family with three children); $\mathbb{P}$(the ball number 90 is extracted at 'lotto').

Basic probability

Inferential statistics is based on probability theory (we do not have certainty, but only *confidence*).

- Events: something that may or may not happen: A; $\mathbb{P}(A)=$ *probability* that A happens;

For instance $\mathbb{P}$(it rains tomorrow in Trento); $\mathbb{P}$(there is at least one son in a family with three children); $\mathbb{P}$(the ball number 90 is extracted at 'lotto').
Formally, $A \subset \Omega$, the sample space (all possible occurrence).
We consider $A \cap B$ (both A and B occur), $A \cup B$ (A *or* B occurs, *or both*)...

Computing probabilities

How do we assign probabilities? We generally use models based on experience and intuition.
After seeing data, statistics helps in deciding whether the model used was correct.
Often, it is assumed that all *elementary events* are equally likely (*classical probability*).
Examples...

- Sequences of heads and tails
- Drawing balls from an urn

Random variables

Often, we are more interested in events that concern a quantitative measure:

- Random variable: something that takes an unpredictable numerical value: X
  $\mathbb{P}(X = k) =$ *probability* that X takes value k.

For instance, X is the number of 'tails' when tossing a coin 10 times.

Random variables

Often, we are more interested in events that concern a quantitative measure:

- Random variable: something that takes an unpredictable numerical value: X
  $\mathbb{P}(X = k) =$ *probability* that X takes value k.

For instance, X is the number of 'tails' when tossing a coin 10 times.
Formally,

$$X : \Omega \to \mathbb{R}, \quad \mathbb{P}(X = k) = \mathbb{P}\left(X^{-1}(\{k\})\right).$$

Binomial distribution

Assumptions:

- X represents the number of successes in n trials;
- Trials can result only in 'success' or 'failure';
- Trials are independent;
- The probability of success is the same p in all trials.

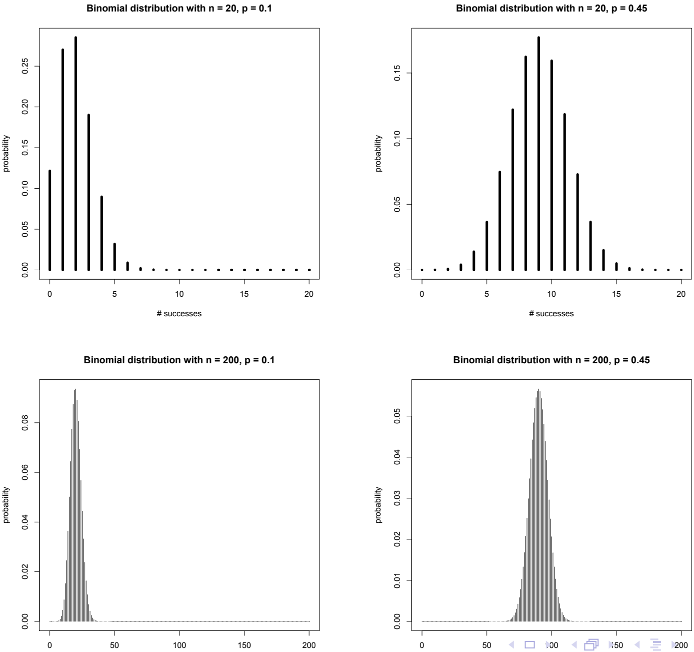
Then

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k}$ [binomial coefficient] $= \dfrac{n \cdot (n-1) \cdots (n-k+1)}{1 \cdot 2 \cdots k}$

$= \dfrac{n!}{k!(n-k)!}$ with $n!$ [n *factorial*]$= n \cdot (n-1) \cdots 2 \cdot 1$.

Graphical illustration of binomial distributions



Problem: are data consistent with the assumption of a binomial distribution?
A classical case study are the sex ratios obtained by Geissler (1889) on the sex of 6115 sibships, each of 12 children.

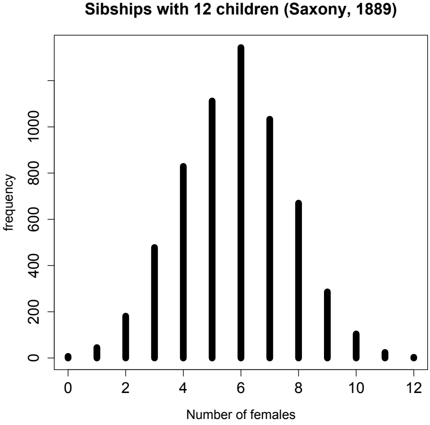
Problem: are data consistent with the assumption of a binomial distribution?
A classical case study are the sex ratios obtained by Geissler (1889) on the sex of 6115 sibships, each of 12 children.

# females	# sibships
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

Problem: are data consistent with the assumption of a binomial distribution?

A classical case study are the sex ratios obtained by Geissler (1889) on the sex of 6115 sibships, each of 12 children.

# females	# sibships
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3



Fitting a binomial

$$p = \text{frequency of female newborns} = \frac{total\#females}{total\#children} \approx 0.480785.$$

$$\mathbb{P}(\#\text{ females in a sibship}) = k) = \binom{12}{k} p^k (1-p)^{12-k}$$

Fitting a binomial

$$p = \text{frequency of female newborns} = \frac{total\#females}{total\#children} \approx 0.480785.$$

$$\mathbb{P}(\#\text{ females in a sibship}) = k) = \binom{12}{k} p^k (1-p)^{12-k}$$

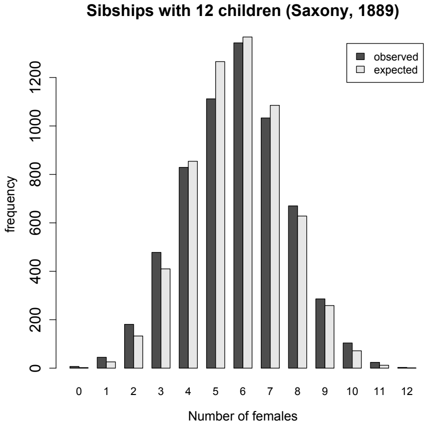
#	obs.	exp.
0	7	2.35
1	45	26.08
2	181	132.84
3	478	410.01
4	829	854.25
5	1112	1265.63
6	1343	1367.28
7	1033	1085.21
8	670	628.06
9	286	258.48
10	104	71.8
11	24	12.09
12	3	0.93

Fitting a binomial

$$p = \text{frequency of female newborns} = \frac{total\#females}{total\#children} \approx 0.480785.$$

$$\mathbb{P}(\#\text{ females in a sibship}) = k) = \binom{12}{k} p^k (1-p)^{12-k}$$

#	obs.	exp.
0	7	2.35
1	45	26.08
2	181	132.84
3	478	410.01
4	829	854.25
5	1112	1265.63
6	1343	1367.28
7	1033	1085.21
8	670	628.06
9	286	258.48
10	104	71.8
11	24	12.09
12	3	0.93

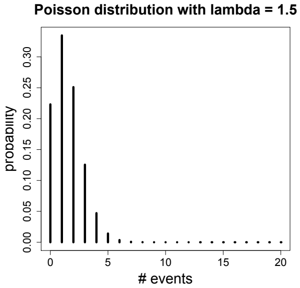


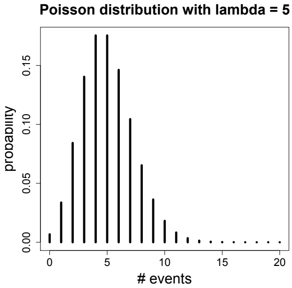
Poisson distribution

Another *discrete* distribution often used is the *Poisson* distribution, used for the occurrence of 'rare' events:

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \; k = 0, 1, 2, \ldots \qquad k! = 1 \cdot 2 \cdots k.$$

λ is the only parameter of the Poisson *[relations with binomial]*.



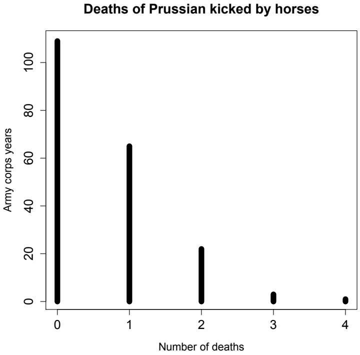


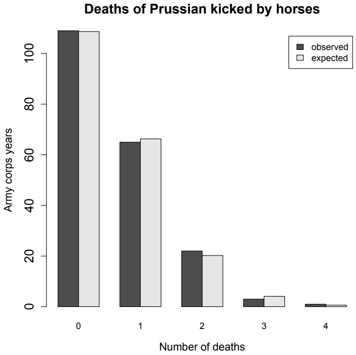
Poisson approximation

Poisson can be viewed as a limiting case of binomial (*law of small numbers*.

The figure shows how binomials with larger n and the same value for np can be approximated by a Poisson with parameter $\lambda = np$

Poisson fit of a distribution





A Poisson distribution fits a famous dataset by von Bortkiewicz (1898) on the number of soldiers killed by being kicked by a horse each year in each of 14 cavalry corps over a 20-year period.

Mean and variance of a random variable

In general, the *distribution* of a discrete random variable is given by

- the list of possible values $\{x_1, \dots, x_n\}$;
- the respective probabilities $\{p_1, \dots, p_n\}$, i.e. $p_k = \mathbb{P}(X = x_k)$

For a random variable, one can compute its *expected value* or mean:

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i \qquad \text{will be denoted also as } \mu_X .$$

To describe its spread, one uses the variance, i.e. the expected value of the squared deviations from the mean:

$$\mathbb{V}(X) = \mathbb{E}((X - \mu_X)^2) = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i = \sum_{i=1}^{n} x_i^2 p_i - \mu_X^2 .$$

Mean and variance of some distributions

If $X \sim Bin(n, p)$ [binomial of parameters n and p]

$$\mathbb{E}(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = n \cdot p. \quad [\# \textit{ trials} \cdot \textit{prob. success}]$$

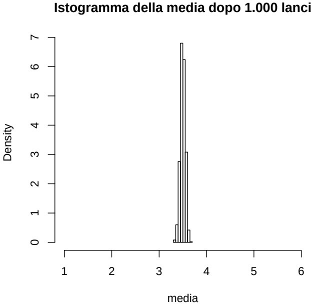
$$\mathbb{V}(X) = \sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k (1-p)^{n-k} = n \cdot p \cdot (1-p).$$

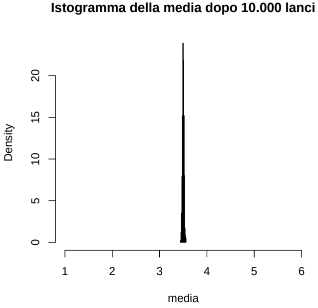
If $X \sim P(\lambda)$ [Poisson of parameters λ]

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda.$$

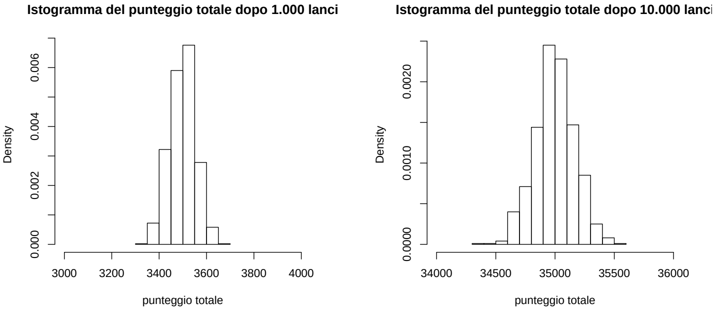
$$\mathbb{V}(X) = \sum_{k=0}^{n} (k - \lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda.$$

Limit theorems of probability. I. The law of large numbers

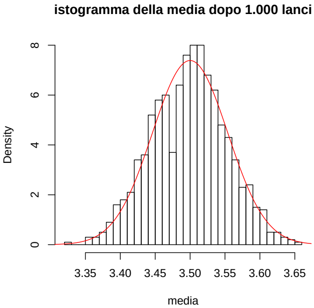


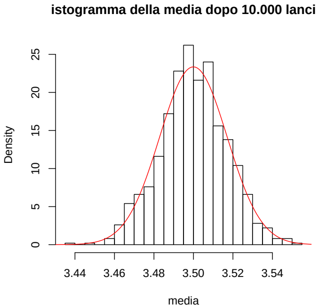


Limit theorems of probability. II. Summing variables



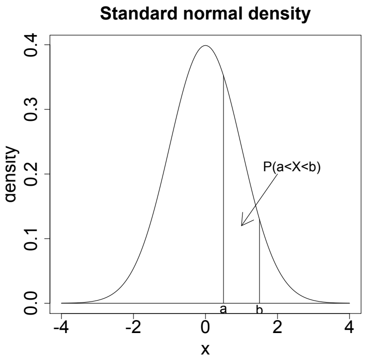
Limit theorems of probability. III. Central limit theorem





With an appropriate scaling, the deviations from the mean follow a universal distribution, the *normal* or *Gaussian*.

Normal distribution



Standard normal:

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \le X \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

More on normal distribution

Generic normal:

$$X \sim N(\mu, \sigma^2), \text{ density } p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} :$$

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\, dx$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xp(x)\, dx = \mu, \qquad \mathbb{V}(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\, dx = \sigma^2.$$

If $X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, i.e. standard normal.

Normal approximation to the binomial

If $X \sim Bin(n, p)$, # successes after n trials

$$\mathbb{E}(X) = np \qquad \mathbb{V}(X) = np(1-p)$$

for n large [say $n \geq 25$, np, $n(1-p) \geq 10$] approximate

$$X \sim N\left(np, np(1-p)\right).$$

$$i.e. \quad \mathbb{P}(a \leq Bin(n,p) \leq b) \approx \mathbb{P}(a \leq N(np, np(1-p)) \leq b).$$

Continuity approximation

True value: $\mathbb{P}(40 \leq Bin(100, 0.42) \leq 48) = 0.598$

$$\mathbb{P}(40 \leq Bin(100, 0.42) \leq 48) = \mathbb{P}(39.5 \leq Bin(100, 0.42) \leq 48.5) \approx$$

$$\approx \mathbb{P}(39.5 \leq N(42, 24.36) \leq 48.5) \approx 0.600.$$

while $\mathbb{P}(40 \leq N(42, 24.36) \leq 48) \approx 0.545$