Teaching Mathematics and Statistics in Sciences, IPA HU-SRB/0901/221/088 - 2011

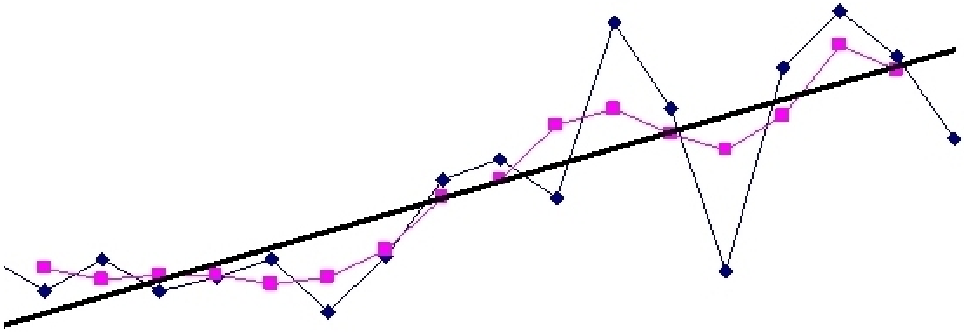
Mathematical and Statistical Modelling in Medicine

Author: **Tibor Nyári** PhD

University of Szeged
Department of Medical Physics and Informatics

www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

Types of Data



Examples to types of data

		The values (range)	
Type of data		continuous	discrete (categorical)
Qualitative	Nominal	not possible	sex, country, place of birth, profession, bloodgroup
	Ordinal	subjective statements about intensity of different things (brightness, voice)	very good-good -acceptable - wrong - very wrong, low - normal - high, etc..
Quantitative		temperature, concentration	number of hospitals, children, other counts

- A data set contains information on a number of individuals.
- Individuals are objects described by a set of data, they may be people, animals or things. For each individual, the data give values for one or more variables.
- A variable describes some characteristic of an individual, such as person's age, height, gender or salary.

The data-table

- Data of one experimental unit (“person”) must be in one record (row)
- Data of the answers to the same question (variables) must be in the same field of the record (column)
- The variables (fields) are generally named by an 8 characters long identifier (e.g.:SPSS)

Number	SEX	AGE	....
1	1	20	....
2	2	17	....
.	.	.	...

Statistical programsystems

- SPSS
- STATGRAPHICS
- SAS
- STATA
- SIGMASTAT
- BMDP
- SOLO
- CSS/STATISTICA
- STATXACT
- (EXCEL)

Distribution

- The distribution of a categorical variable describes what values it takes and how often it takes these values.
- The distribution of a continuous variable describes what values it takes and how often these values fall into an interval.
- Describing distributions with graphs:
    - Categorical variables: bar chart, pie chart
    - Continuous variable: histogram

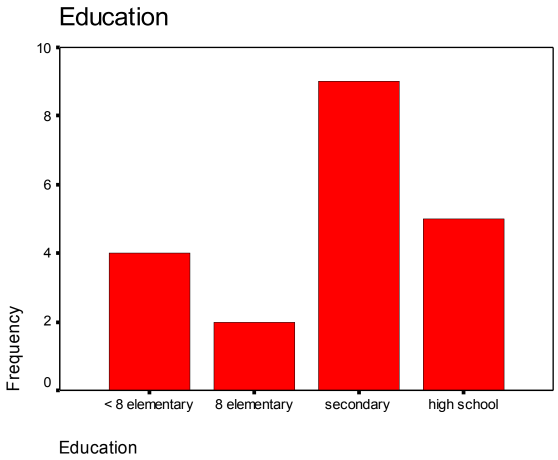
The distribution of a categorical variable, example

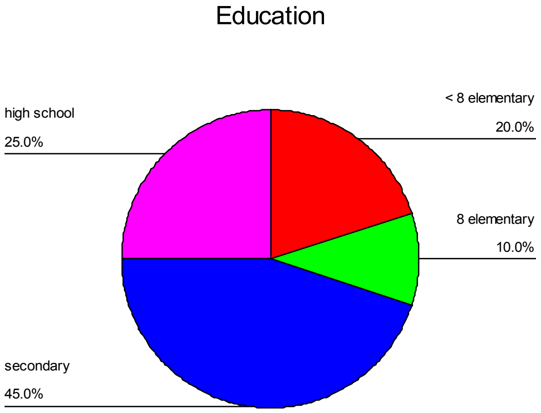
Education categories:

1: <8 elementary
2: 8 elementary
3: secondary school
4: high school or university

Frequencies:

	Frequency	Percent
< 8 elementary	4	20.0
8 elementary	2	10.0
secondary	9	45.0
high school	5	25.0
Total	20	100.0





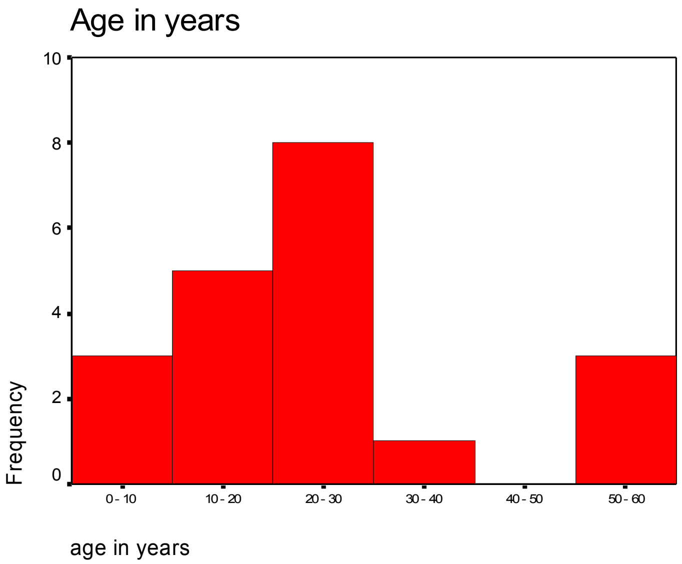
The distribution of a continuous variable, example

Values:

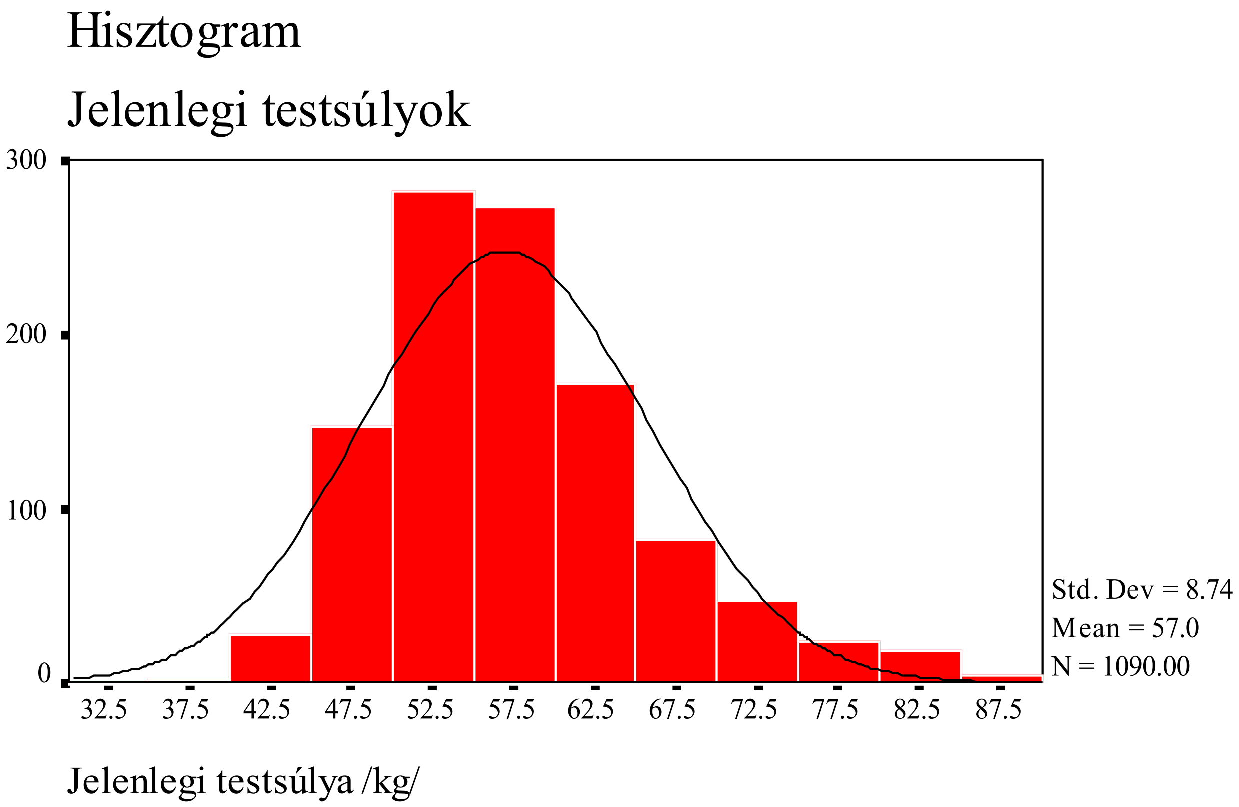
20.00
17.00
22.00
28.00
9.00
5.00
26.00
60.00
35.00
51.00
17.00
50.00
9.00
10.00
19.00
22.00
25.00
29.00
27.00
19.00

Categories:

0-10
10-20
20-30
30-40
40-50
50-60



Histogram (Body weights)

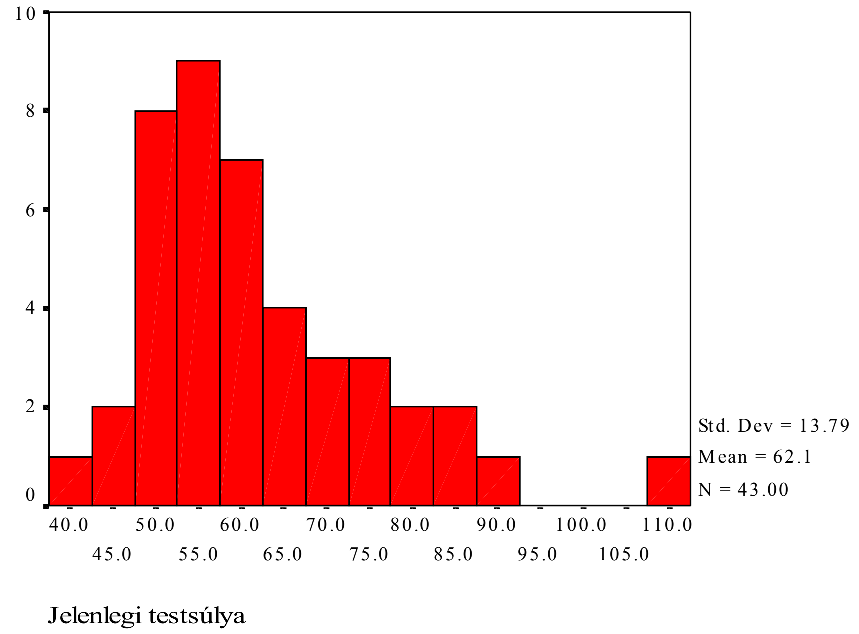


The overall pattern of a distribution:

- The center, spread and shape describe the overall pattern of a distribution.
- Some distributions have simple shape, such as symmetric and skewed. Not all distributions have a simple overall shape, especially when there are few observations.
- A distribution is skewed to the right if the right side of the histogram extends much further out then the left side.

Outliers

- Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them (real data, typing mistake or other).



Describing distributions with numbers

- Measures of central tendency:
  - the mean, the mode and the median are three commonly used measures of the center.
- Measures of variability :
  - the range, the quartiles, the variance, the standard deviation are the most commonly used measures of variability .
- Measures of an individual:
  - rank, z score

Measures of central tendency

- Mean: $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$
- Mode: is the most frequent number
- Median: is the value that half the members of the sample fall below and half above. In other words, it is the middle number when the sample elements are written in numerical order

Example

- The grades of a test written by 11 students were the following:
- 100 100 100 63 62 60 12 12 6 2 0.

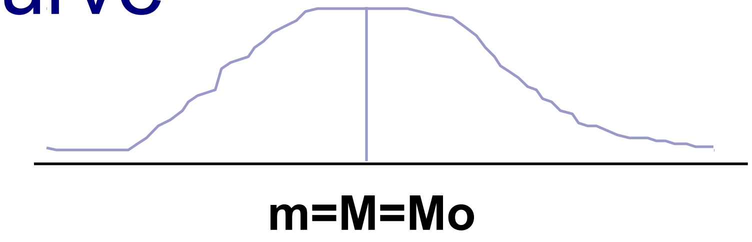
- A student indicated that the class average was 47, which he felt was rather low. The professor stated that nevertheless there were more 100s than any other grade. The department head said that the middle grade was 60, which was not unusual.

Results

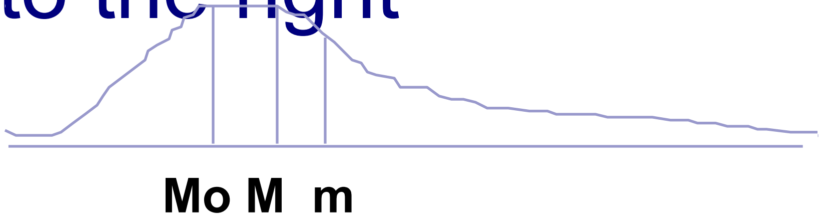
- The mean is 517/11=47,
- the mode is 100,
- the median is 60.

Relationships among the mean(m), the median(M) and the mode(Mo)

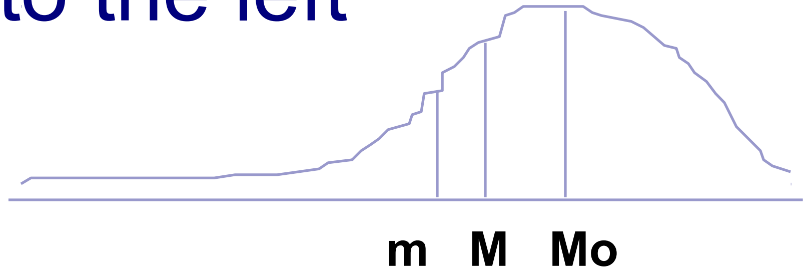
- A symmetrical curve



- A curve skewed to the right



- A curve skewed to the left



Measures of variability (dispersion)

- The range is the difference between the largest number (maximum) and the smallest number (minimum).
- The variance $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$
- The standard deviation $$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

Example

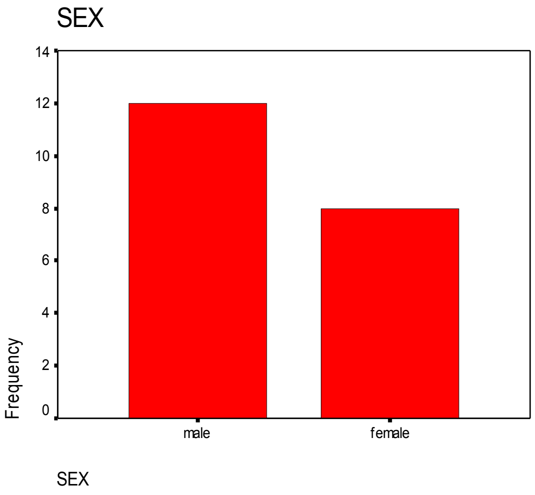
Var 1	Var 2	Var 3
2	12	20
3	13	30
4	14	40
5	15	50
8	18	80
9	19	90
9	19	90
10	20	100
40	50	400
44	54	440
44	54	440
62	72	620
mean=20	mean=30	mean=200
SD=21,0971777	SD=21,0971777	SD=210,971777

Displaying data

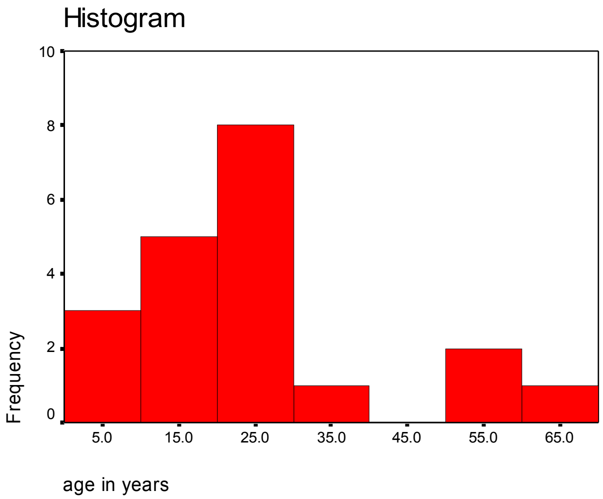
- Categorical data
  - barchart
  - piechart
- Continuous data
  - dot plot
  - histogram
  - box-whisker plot
  - mean-standard deviation plot
  - scatterplot

Bar chart and histogram

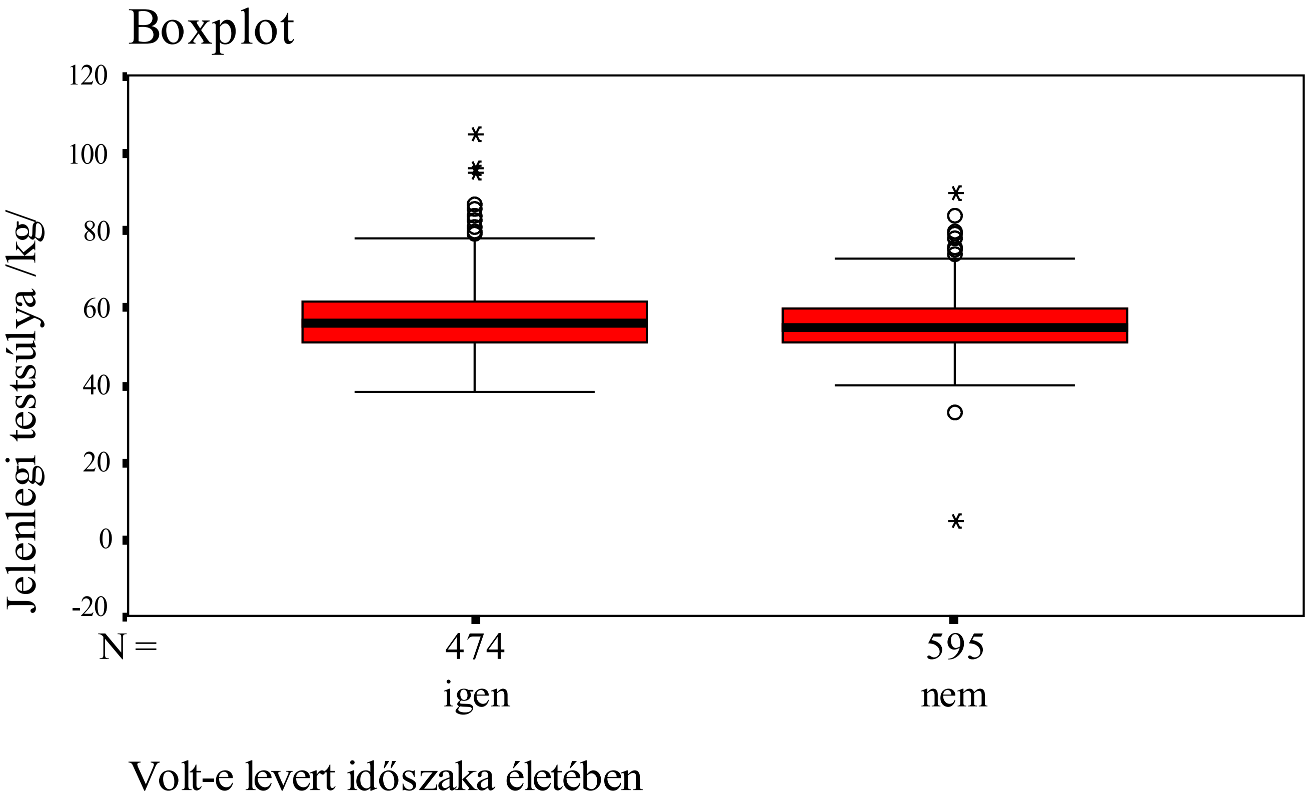
Discrete:



Continuous:

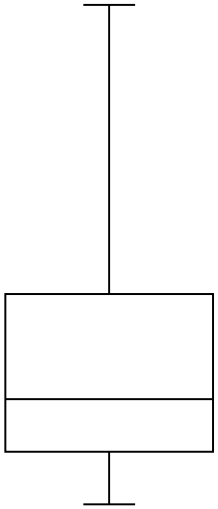


Box-plot

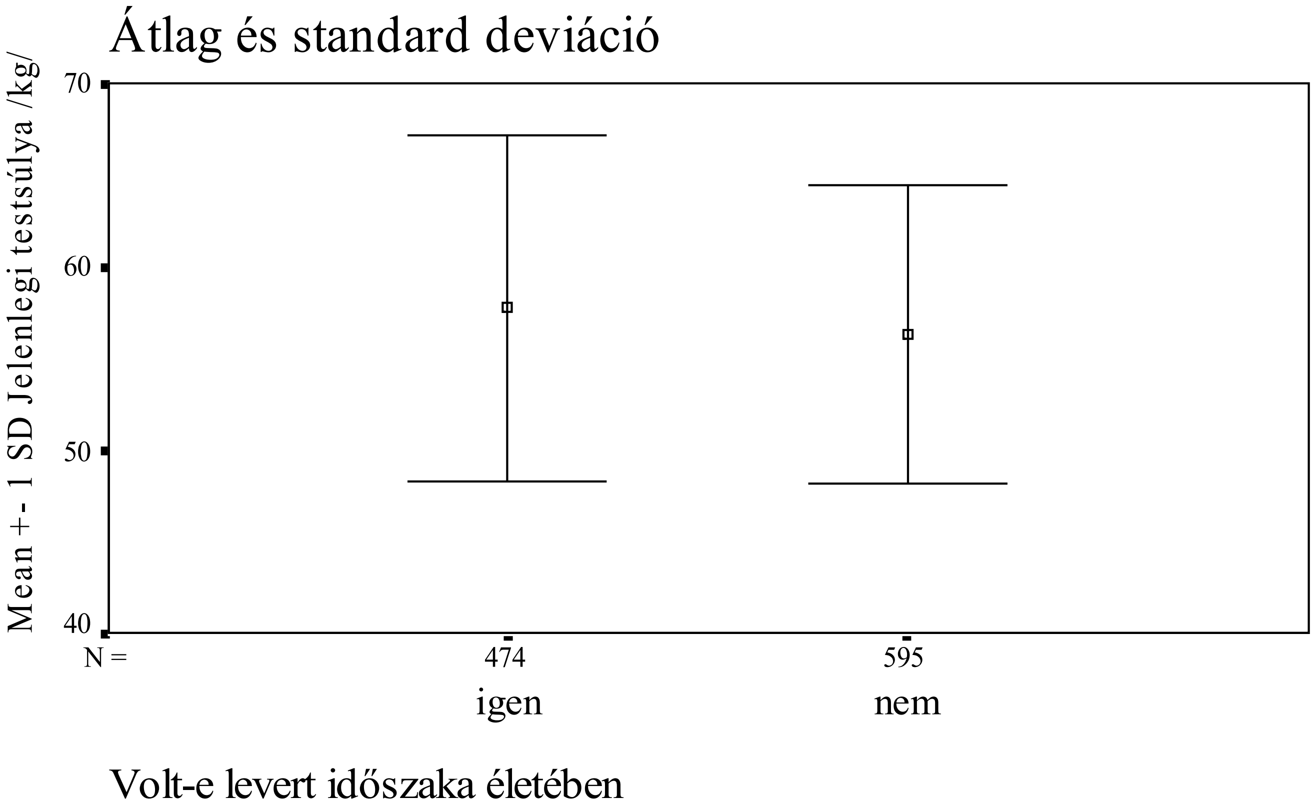

Boxplot
120
100
80
60
40
20
0
-20
Jelenlegi testsúlya /kg/
N =
474
595
igen
nem
Volt-e levert időszaka életében

How create a box-plot

- We need
  - Median($P_{50\%}$), $P_{25\%}$ and $P_{75\%}$
- Calculalate the differences of
  - $d_1=P_{50\%}$ - $P_{25\%}$ and
  - $d_2=P_{75\%}$ - $P_{50\%}$
- Then calculate 1.5 x d_1 and 1.5 x d_2.
- And plot



Mean and standard deviation



Box-plot vs Mean±SD plot

- Box-plot
  - Give information about the symmetry
- Mean and standard deviation plot
  - Could be used for normal distributed data