Lecture 11: foundations of statistics

Content:

- statistics - basic terms
- distributions, histograms
- normal (Gaussian) distribution
- distributions - measures of central tendency
- distributions - measures of dispersion
- correlation and regression analysis
  (Least SQuares method – LSQ)

Fundaments of statistics:

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data (data-sets). The main difference to math. analysis is that it works with discrete data.

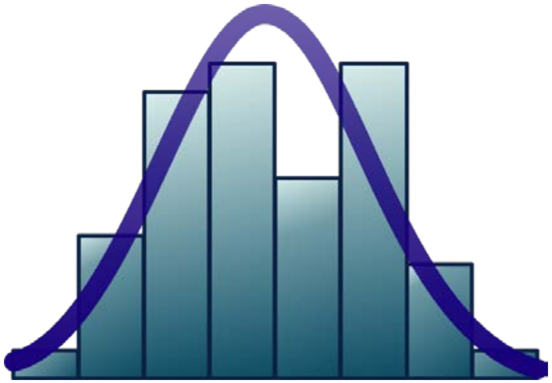
We are often interested into basic features of our data-sets: Are their values focused around some (central) value? Are they distributed around it in some typical character? What are their limits? ...

Statistics also deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments.

In other words:
Statistics is a numerical measure that describes a characteristic of a sample. You can meet also stochastics, which describes random events that are unpredictable due to the influence of a random variable.

Fundaments of statistics:

Two main statistical methodologies are used in data analysis: descriptive statistics, which summarizes data from a sample using various parameters (indexes) and inferential statistics, which draws conclusions from data that are subject to random variation (under the framework of so called probability theory).



Statistics is used for the description of scientific, industrial, or societal problems. In many branches, statistics is intensively connected with so called QC (quality check).

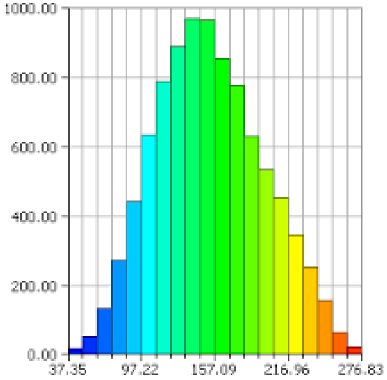
Fundaments of statistics – basic terms:

Distribution:

There are many understandings and definitions of distribution, in statistics we understand it as frequency distribution, which consists of a count of the number of occurrences of each value. The best way of frequency distribution visualization is a histogram construction.

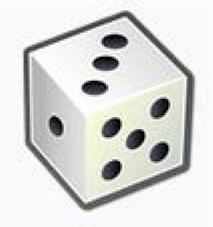
There are several parameters, which describe a frequency distribution - central tendency (or location) seeks to characterize the distribution's central or typical value, while dispersion (or variability) characterizes the extent to which members of the distribution depart from its center and each other. Both these two important parameters have their measures.

Beside this also the character (shape) of the distribution is important.



histogram

Fundaments of statistics – basic terms:



Probability:

is the measure of the likelihood that an event will occur.
Probability is quantified as a number between 0 and 1
(where 0 indicates impossibility] and 1 indicates certainty).
The higher the probability of an event, the more certain that the event will occur.

A simple example is the tossing of a fair (unbiased) coin. Since the coin is unbiased, the two outcomes ("head" and "tail") are both equally probable; the probability of "head" equals the probability of "tail.“ Since no other outcomes are possible, the probability is 1/2 (or 50%), of either "head" or "tail".

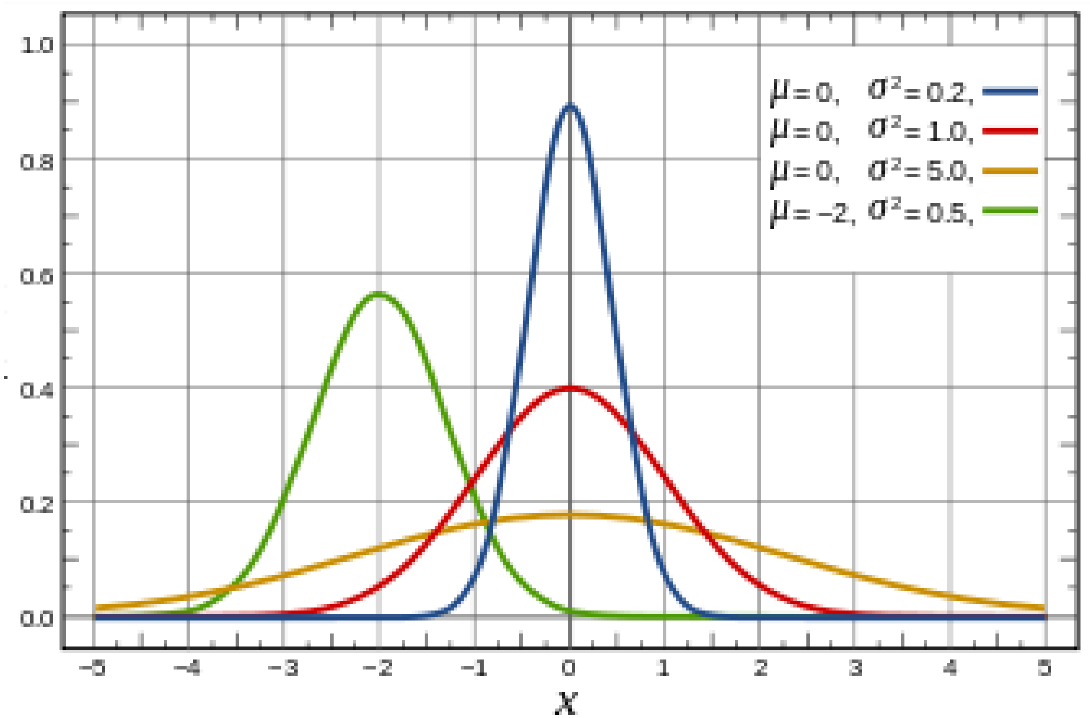
These concepts have been given an axiomatic mathematical formalization in probability theory, which is used widely in such areas of study as mathematics, statistics, finance, gambling, science (in particular physics), artificial intelligence/machine learning, computer science, game theory, and philosophy.

Fundaments of statistics – basic terms:

Distribution:

In the description of distribution, different mathematical models are used. Among them so called probability density functions (PDF) play an important role (also in probability theory). There exist also cumulative distribution functions (CDF).

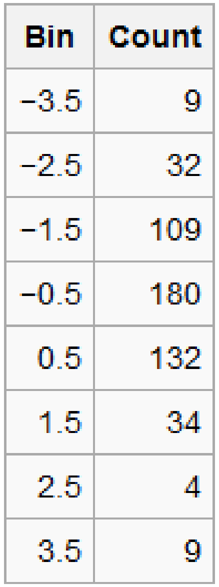
We will come to this point later on, when we will speak about the so called normal (Gaussian) distribution.



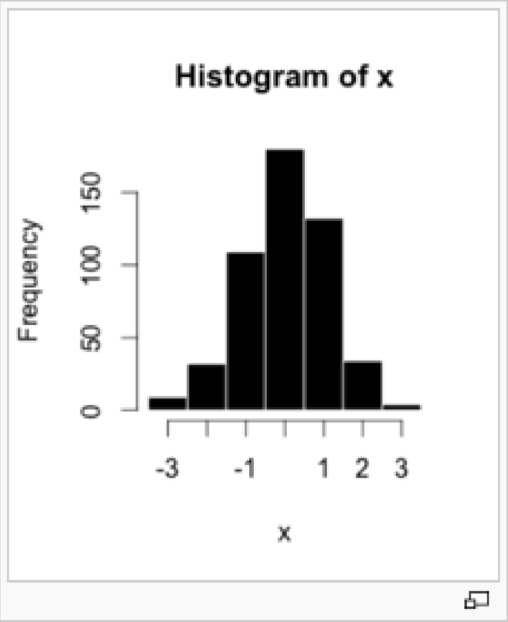
Histogram (frequency histogram)

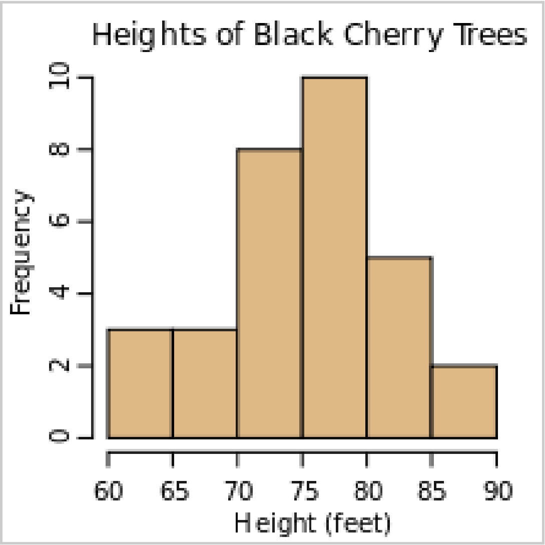
A histogram is a graphical representation of the distribution of numerical data.
To construct a histogram, the first step is to "bin" the range of values - that is, divide the entire range of values into a series of intervals - and then count how many values fall into each interval (bin). The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are usually equal size.

This is a toy example:

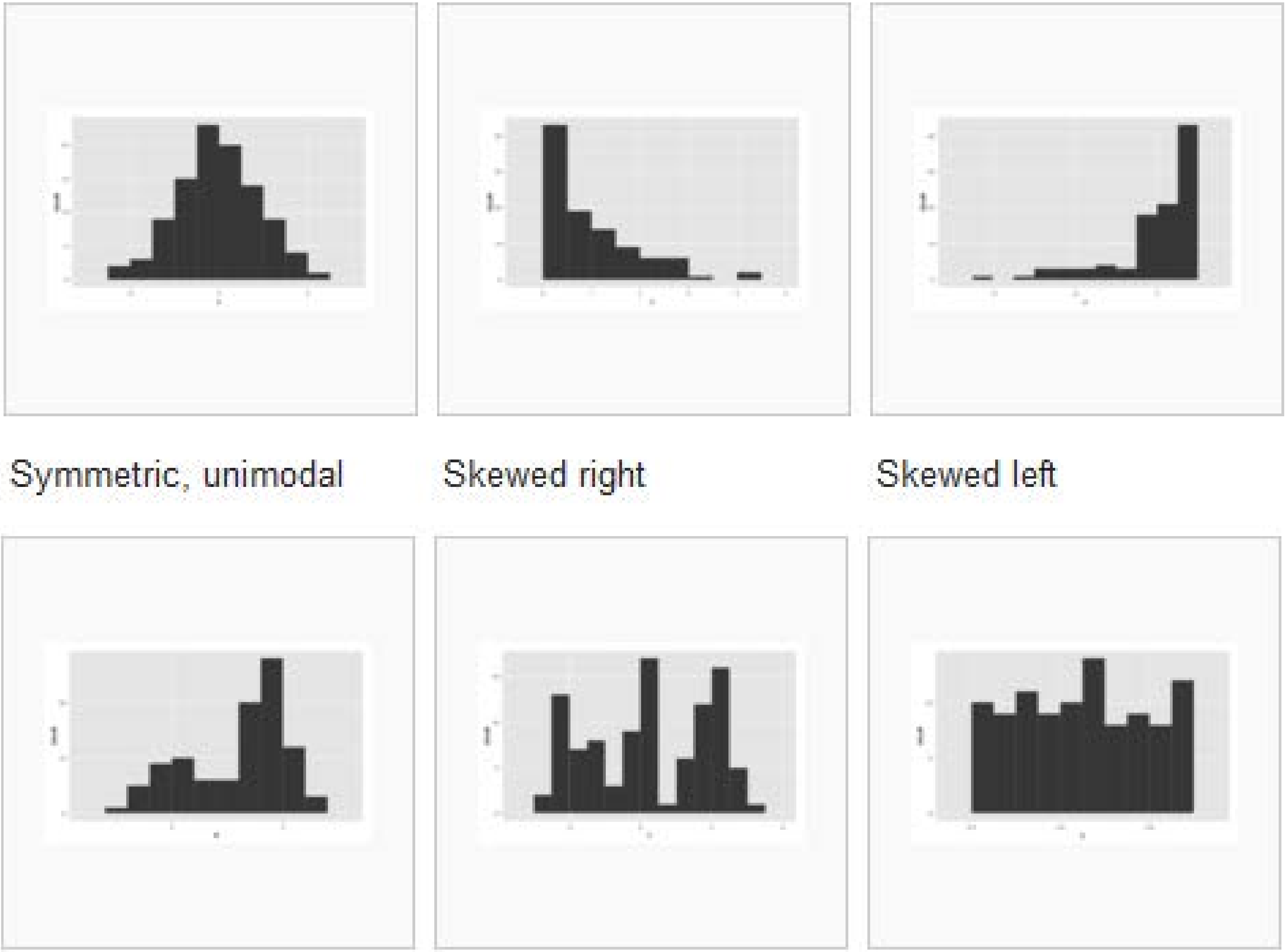


Bin	Count
−3.5	9
−2.5	32
−1.5	109
−0.5	180
0.5	132
1.5	34
2.5	4
3.5	9





An example histogram of the heights of 31 Black Cherry trees.



Symmetric, unimodal

Skewed right

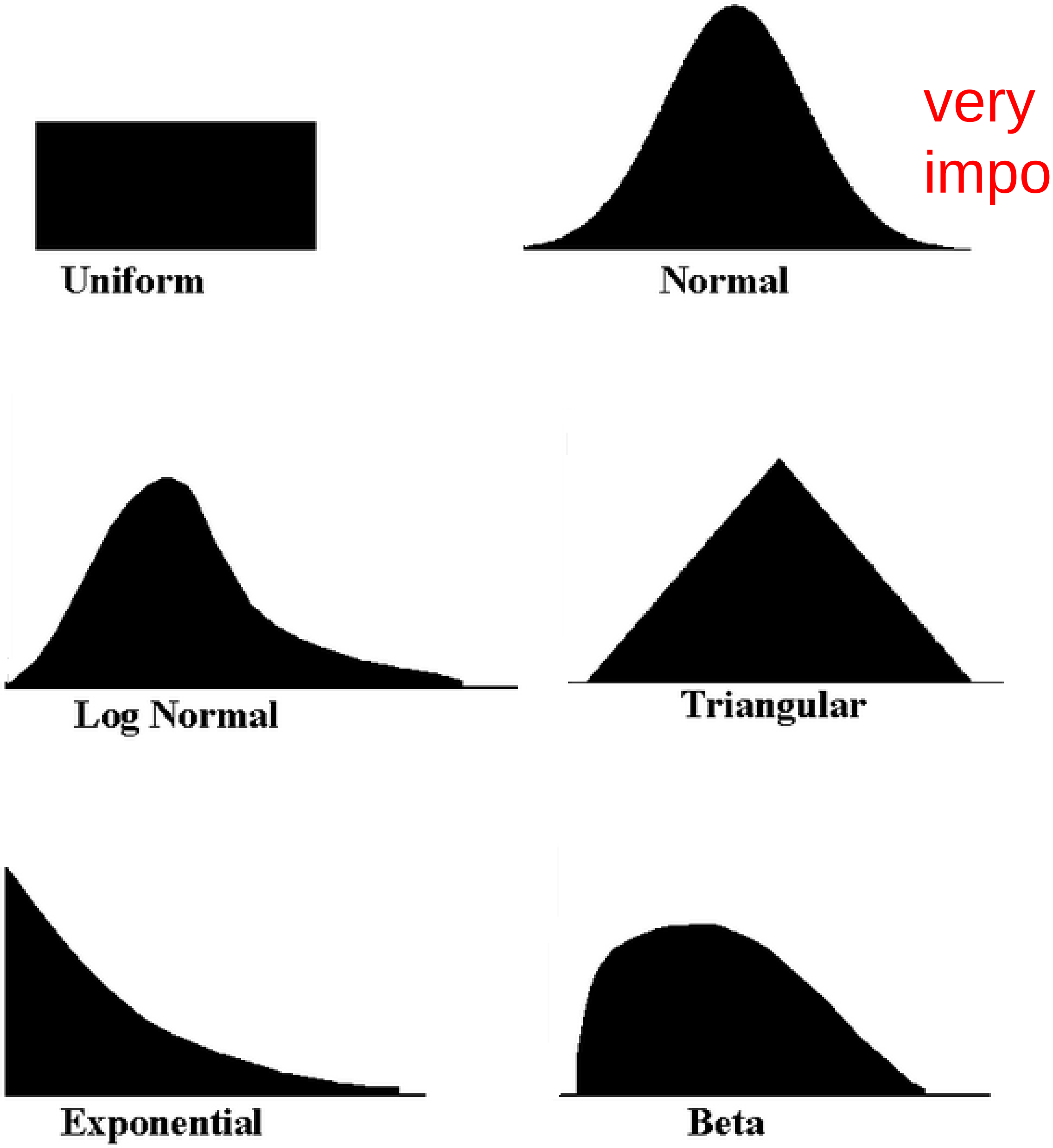
Skewed left

Bimodal

Multimodal

Symmetric

characteristics of frequency distributions



very important

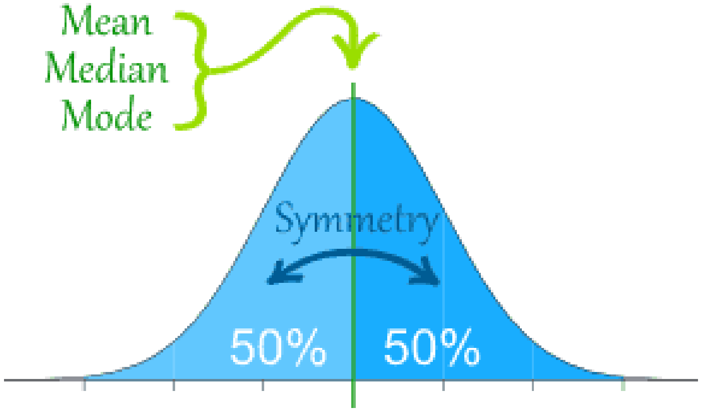
characteristics of frequency distributions (model situations)

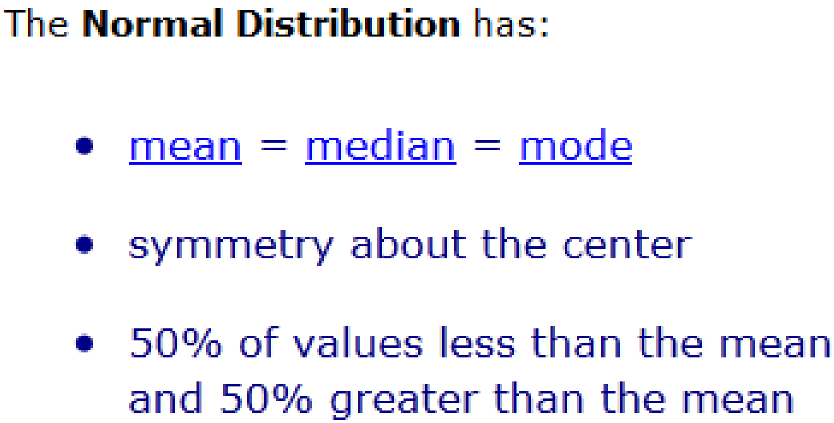
Normal (Gaussian) distribution

Among distributions, which have symmetric histogram (bell curve shaped) the most important is the so called normal (Gaussian) distribution.

Many events and phenomena are close to the normal distribution:

- heights of people,
- size of things produced by machines,
- errors in measurements,
- blood pressure,
- marks on a test,
- etc.

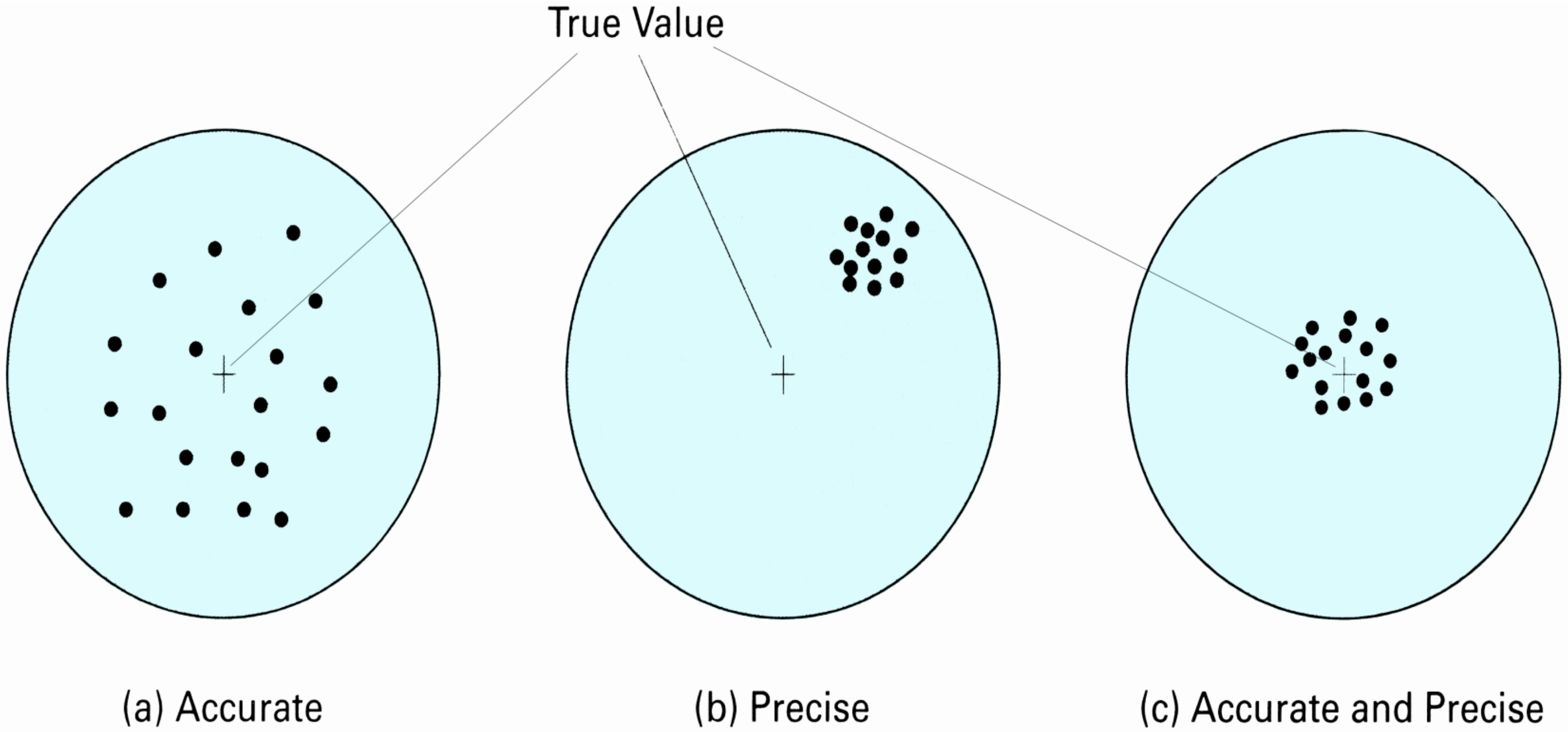




The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

Important terms: accuracy and precision

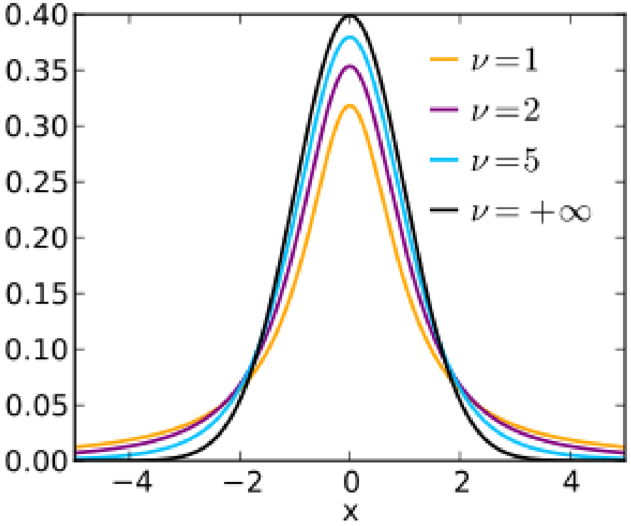


difference between accuracy and precision

Other bell curve shaped models of distributions

There exist beside the well-known normal (Gaussian) distribution other models of distribution, which have a bell curve shape:

- Cauchy (Cauchy–Lorentz) distribution,
- logistic distribution,
- Student *t*-distribution,
- etc.



<u>Interesting thing from the history:</u>
Student's distribution was introduced by William Gosset (1876-1937) who needed a distribution for small samples. He was a Irish Guiness Brewery employee and was not allowed to publish research results in scientific journal. For that reason he published his *t*-distribution under the pseudonym Student (Student, 1908, journal *Biometrika*).

Statistician Gosset, known as "Student".

distributions - measures of central tendency

statistics - measures of central tendency

The three measures in common use are the:

- □ average (mean)
- □ median
- □ mode (modus)

Average:

In mathematics, the three classical Pythagorean means are the arithmetic mean (AM), the geometric mean (GM), and the harmonic mean (HM). They are defined by:

- $AM(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n)$
- $GM(x_1, \ldots, x_n) = \sqrt[n]{x_1 \cdots x_n}$
- $HM(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$

$$AM = \bar{x} = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

were x_1... x_n are the samples of the data-set (*n* – their total number)

statistics - measures of central tendency

The three measures in common use are the:

❑ average (mean)

❑ median

❑ mode (modus)

Average:

In mathematics, the three classical Pythagorean means are the arithmetic mean (AM), the geometric mean (GM), and the harmonic mean (HM). They are defined by:

There is an ordering to these means (if all of the x_i are positive):

$$\min \leq HM \leq GM \leq AM \leq \max$$

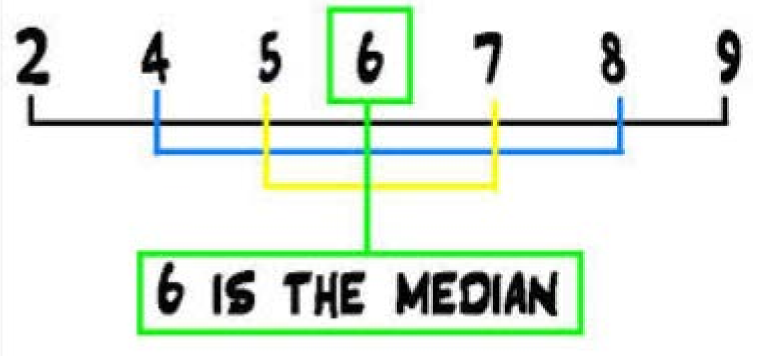
statistics - measures of central tendency

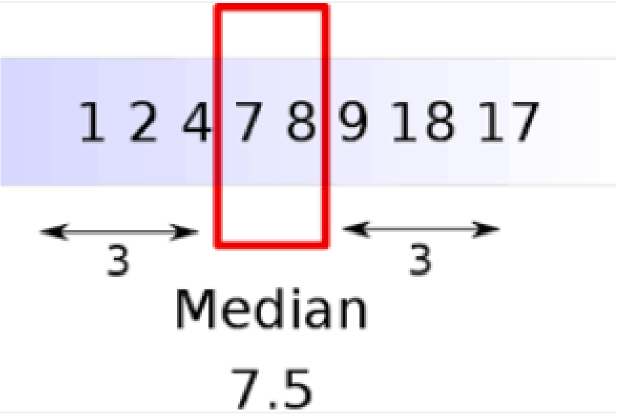
The three measures in common use are the:

- ❑ average (mean)
- ❑ median
- ❑ mode (modus)

Median is the number separating the higher half of a data sample from the lower half.

The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one .





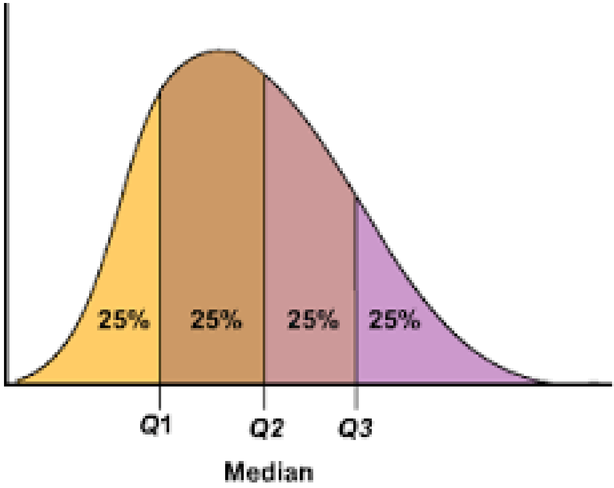
statistics - measures of central tendency

The three measures in common use are the:

- ❑ average (mean)
- ❑ median
- ❑ mode (modus)

Beside median we know also quartiles – dividing the data set into quarters ($Q_1 = Q_{0.25}$, $Q_2 = Q_{0.5}$ = median, $Q_3 = Q_{0.75}$), quintiles (dividing it into fifths), decils, … etc.

All these important values are called as quantiles.

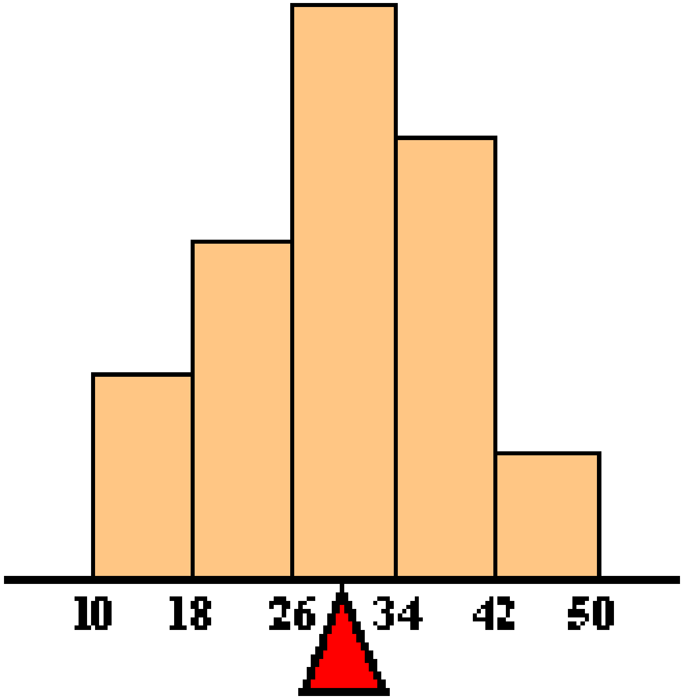


statistics - measures of central tendency

The three measures in common use are the:

- ❑ average (mean)
- ❑ median
- ❑ mode (modus)

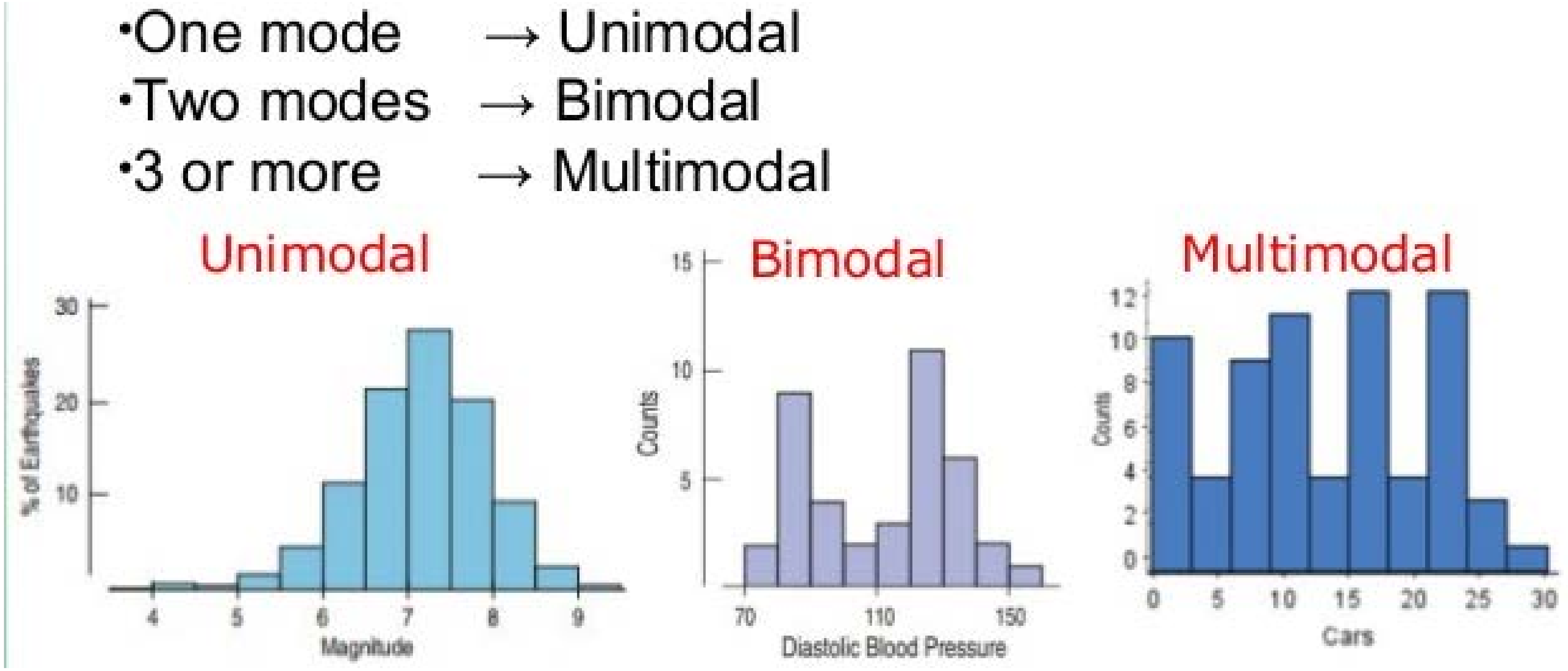
The mode is the value that occurs with the greatest frequency (it is possible to have no modes in a series or numbers or to have more than one mode).



statistics - measures of central tendency

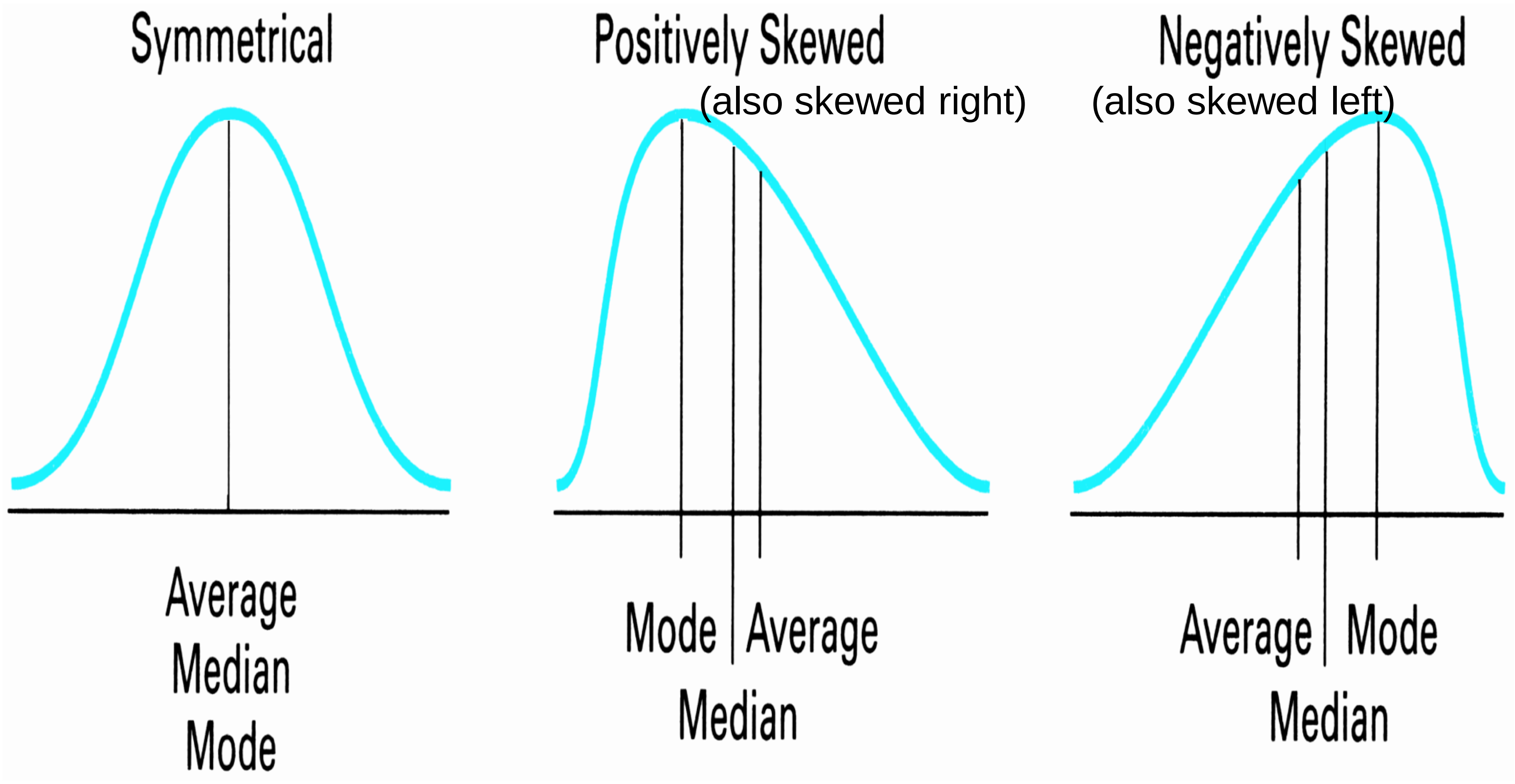
The three measures in common use are the:

- ❑ average (mean)
- ❑ median
- ❑ mode (modus)



- •One mode → Unimodal
- •Two modes → Bimodal
- •3 or more → Multimodal

statistics - measures of central tendency



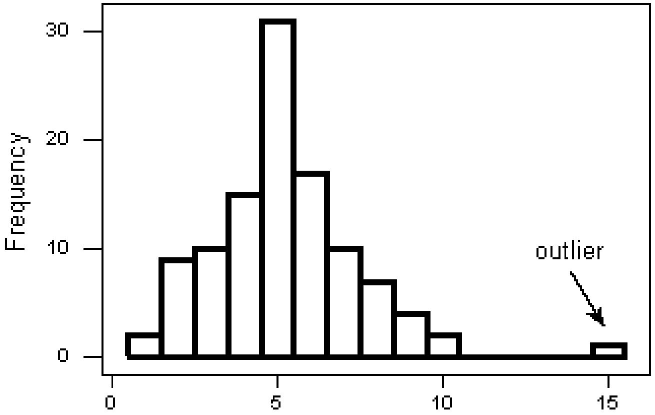
(also skewed right)

(also skewed left)

relationship among average, median and mode

statistics - measures of central tendency

Important:
Mode and median are less sensitive to outliers in comparison with average (mean).



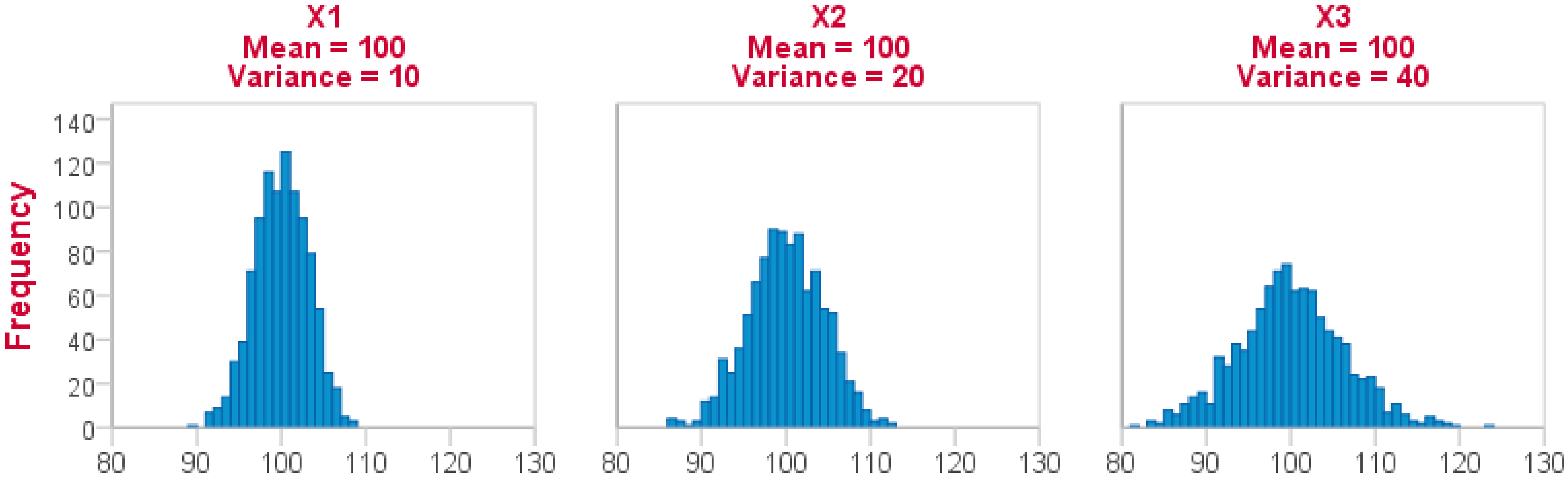
Example:
We have a set of integer numbers: {10, 18, 26, 34, 42, 50, 101}, Where the last value 101 is understood as an outlier (erroneous value). Without this value the set has mean = median = 30, but together with this outlier: mean = 40.142 and median = 34.

Comment:
In many statistical analysis the maximum and minimum values are excluded from the data-set
(so that they do not influence e.g. the mean value).

distributions - measures of dispersion

statistics - measures of dispersion

Beside the central tendency, information also is important the information about the data dispersion around this central value.

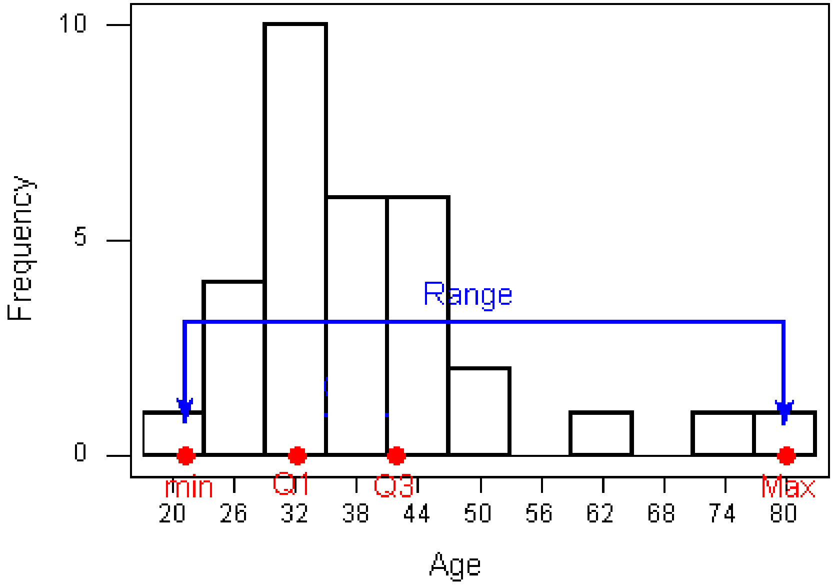


We know in statistics three important measures of this dispersion: range, standard deviation and variance.

statistics - measures of dispersion

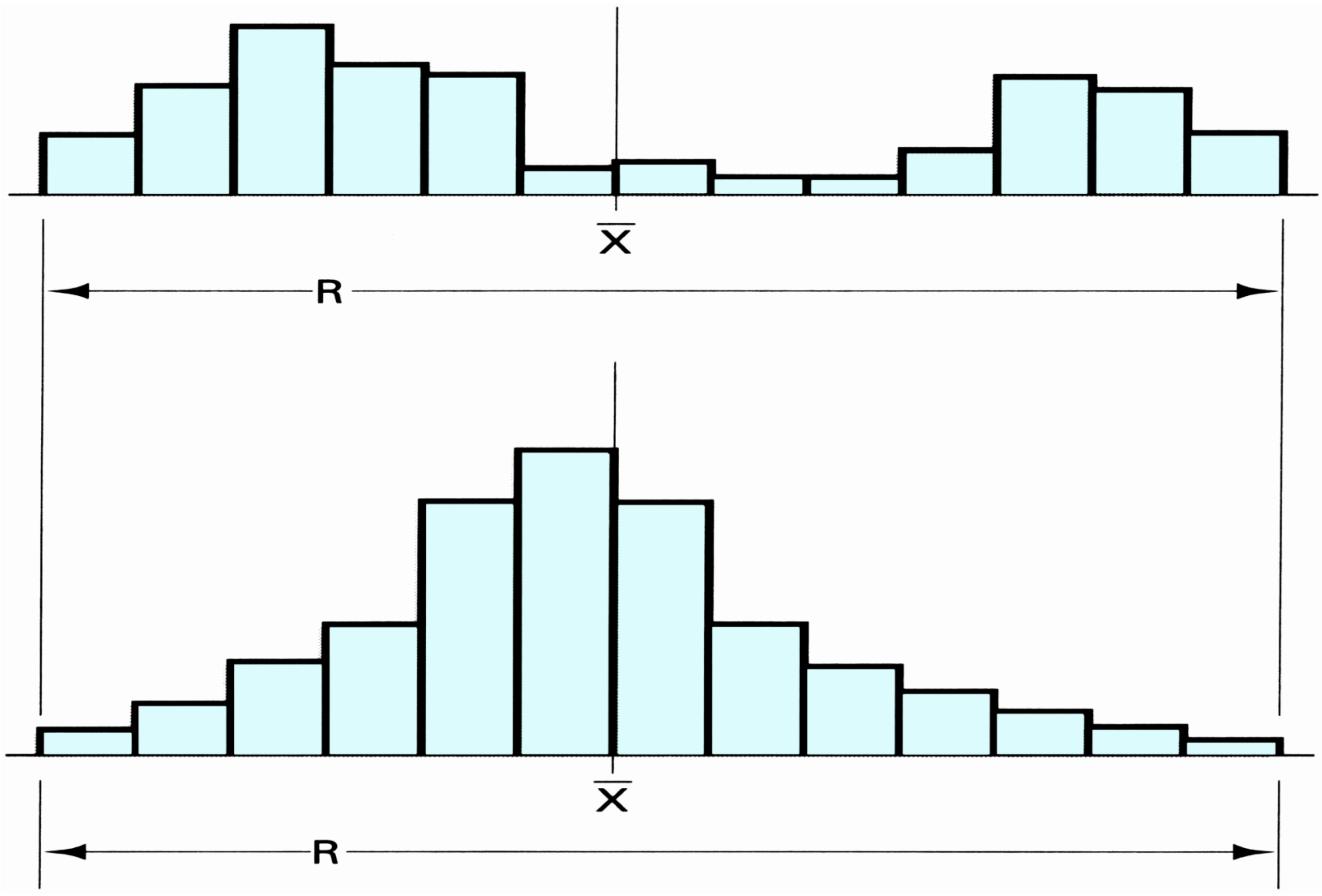
The three measures in common use are the:

- ❑ range
- ❑ standard deviation
- ❑ variance



Range is equal to the difference between the largest and smallest value in data set.
It tells how great is the interval of the data-set values, but does not tell about the character of data dispersion.

statistics - measures of dispersion (range)



comparison of two distributions with equal average and range

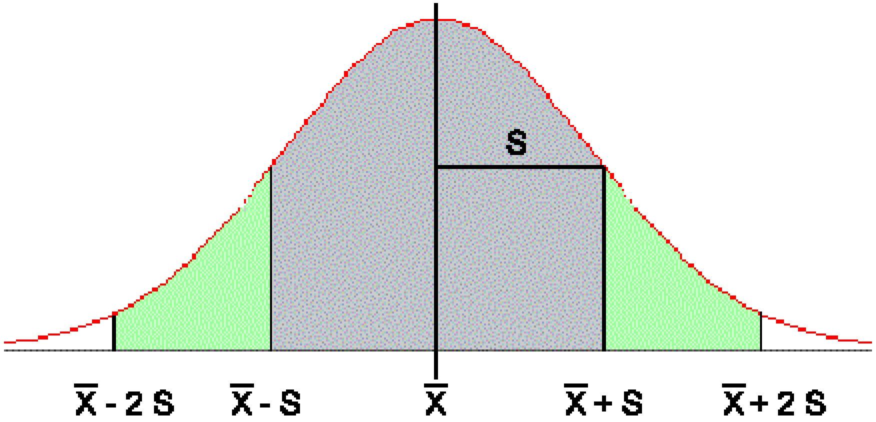
statistics - measures of dispersion

The three measures in common use are the:

❑ range

❑ standard deviation

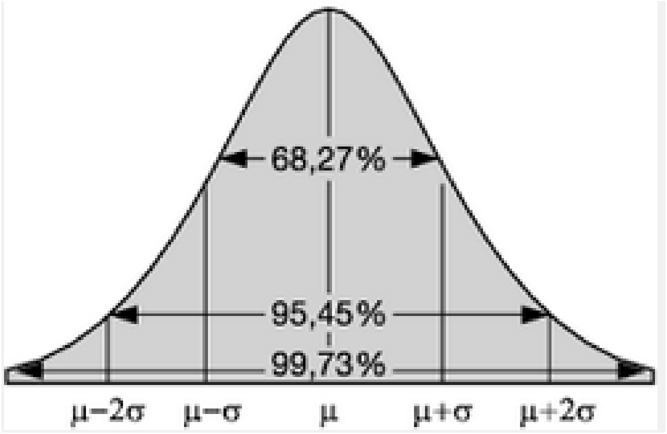
❑ variance



Shows „a kind of average distance“ from arithmetic mean – it is evaluated by following formula:

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

In the case of Gaussian normal distribution



or so called corrected or unbiased version:

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

statistics - measures of dispersion

The three measures in common use are the:

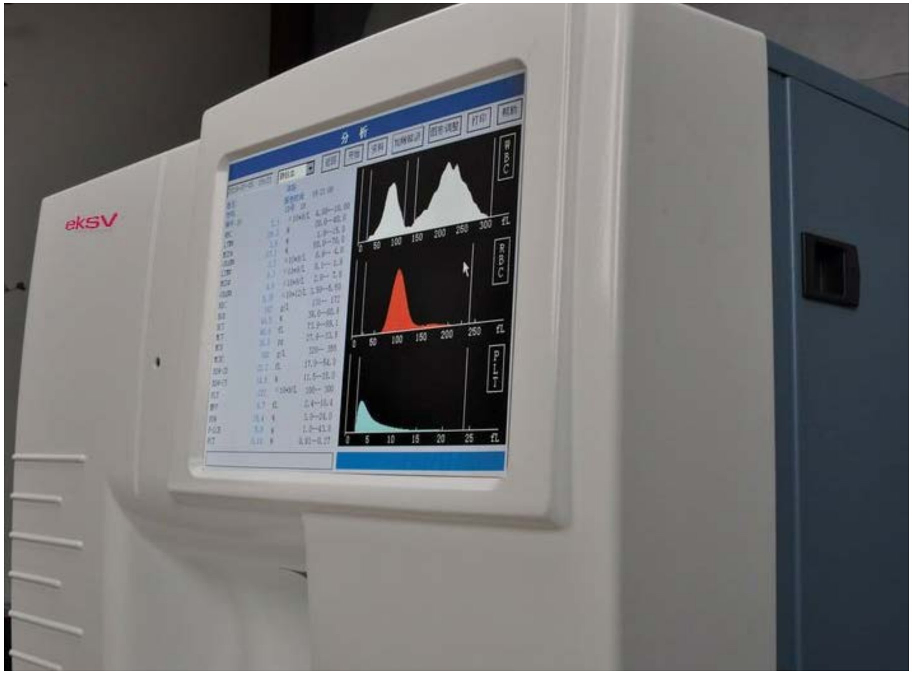
- ❑ range
- ❑ standard deviation
- ❑ variance

Shows „a kind of average distance“ from arithmetic mean – it is the square of standard deviation:

$$s^2 = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

or so called corrected or unbiased version:

$$s^2 = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

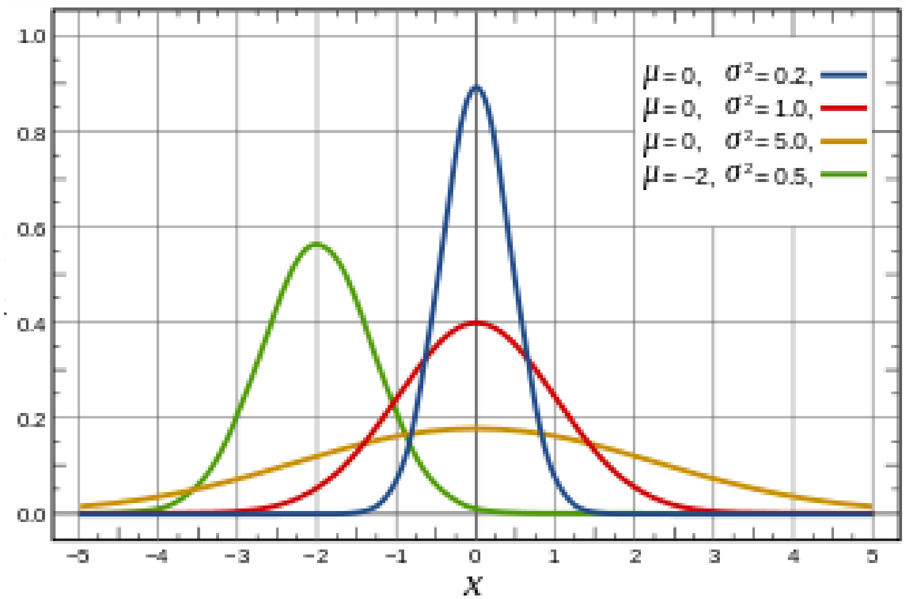


Comment to the normal (Gaussian) distribution:

Among distributions, which have symmetric histogram (bell curve shaped) the most important is the so called normal (Gaussian) distribution.

Probability density function (PDF), describing the normal (Gaussian) distribution has the form:

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



where μ is the mean of the distribution (also its median and mode), σ is its standard deviation (variance is then σ^2).

Performing simple statistics with Excel:

Using the command “=”, defining the function name and the starting and final cell numbers, e.g.:

=MIN(E2:E253)

=MAX(E2:E253)

=SUM(E2:E253)

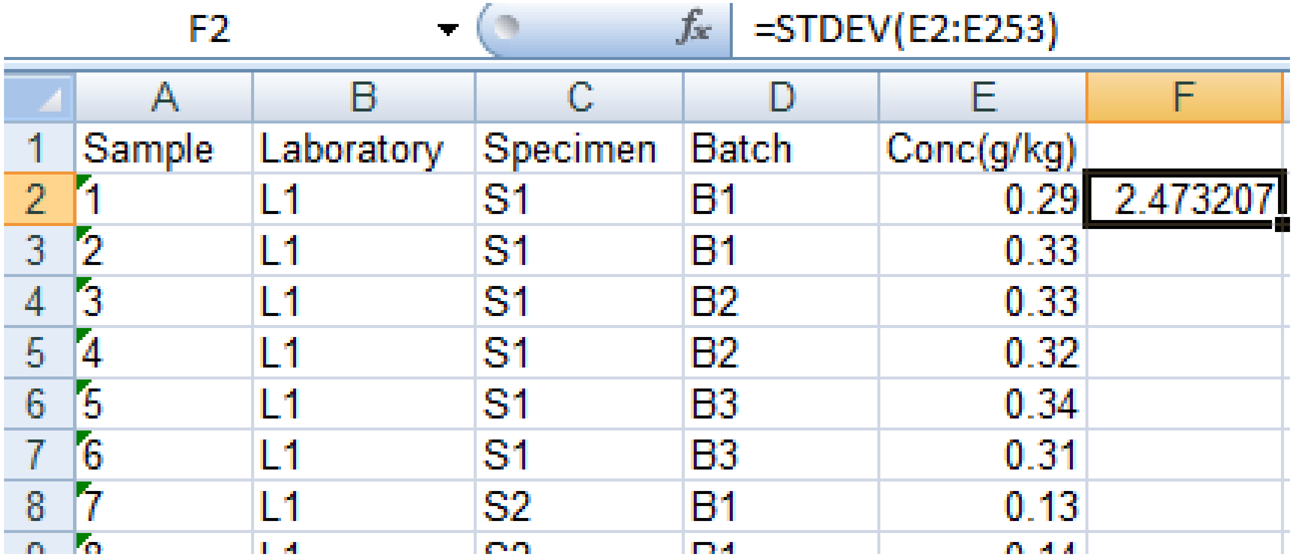
=AVERAGE(E2:E253)

=MEDIAN(E2:E253)

=MODE(E2:E253)

=STDEV(E2:E253)

… etc. (other can be found in Excel help)



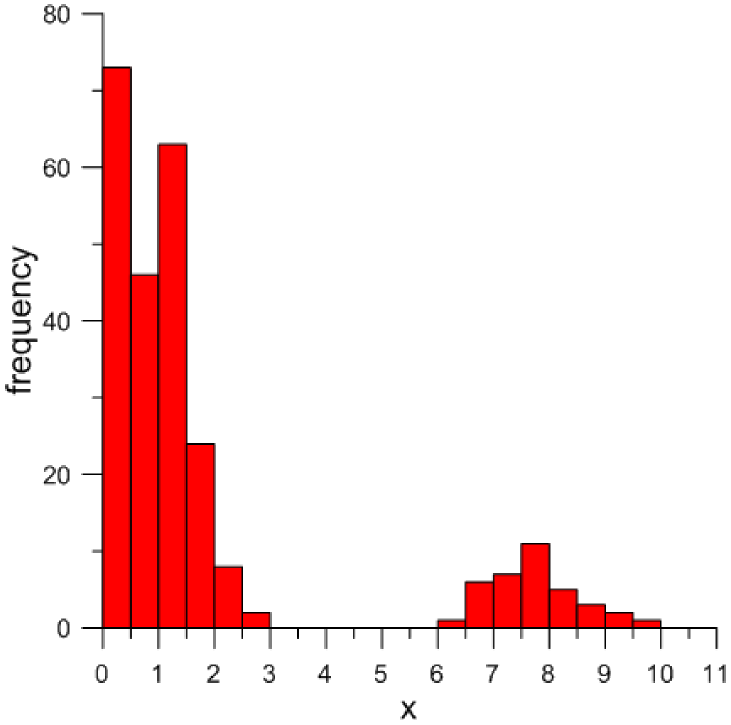
F2 =STDEV(E2:E253)

	A	B	C	D	E	F
1	Sample	Laboratory	Specimen	Batch	Conc(g/kg)	
2	1	L1	S1	B1	0.29	2.473207
3	2	L1	S1	B1	0.33	
4	3	L1	S1	B2	0.33	
5	4	L1	S1	B2	0.32	
6	5	L1	S1	B3	0.34	
7	6	L1	S1	B3	0.31	
8	7	L1	S2	B1	0.13	
9	8	L1	S2	B1	0.14	

Performing simple statistics with Excel:

Example:

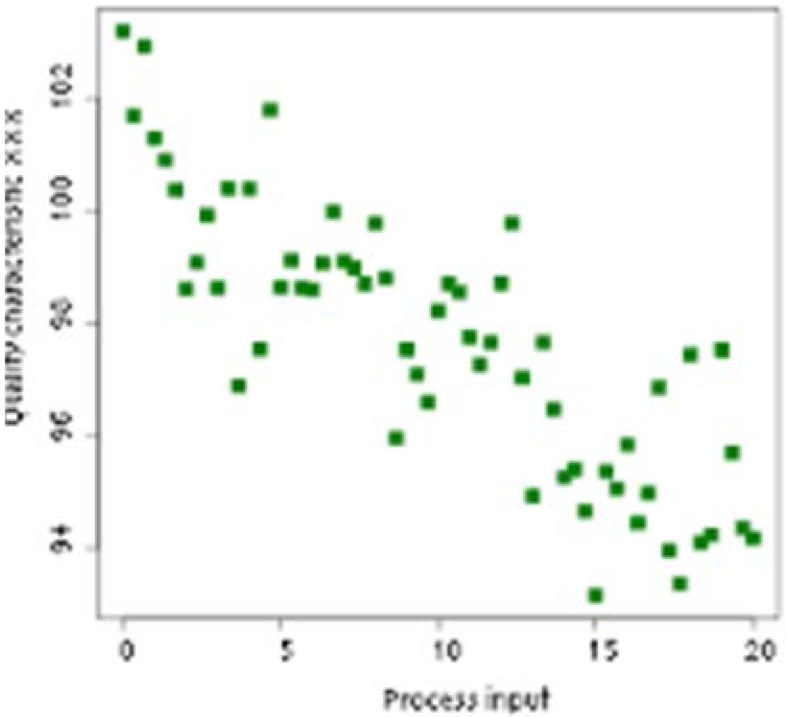
taken from a scientific paper in the journal *Analyst* in year 1987:
Seven specimens were sent to 6 laboratories in 3 separate batches and each analyzed for special analyte (results are in [g/kg]).
Each analysis was duplicated.
Results are summarized in the file "cooperation - dataset.xls".
More info, you can find in:
https://vincentarelbundock.github.io/Rdatasets/doc/MASS/coop.html



correlation and regression analysis

Cross-plots analysis:

Cross-plots are scatter plots used to describe a specialized chart that compares multiple measurements made at a single time or location along two or more axes.
The axes of the plot are commonly linear, but may also be logarithmic.
We often try to find some functional relation between these two plotted parameters (very often a linear function).



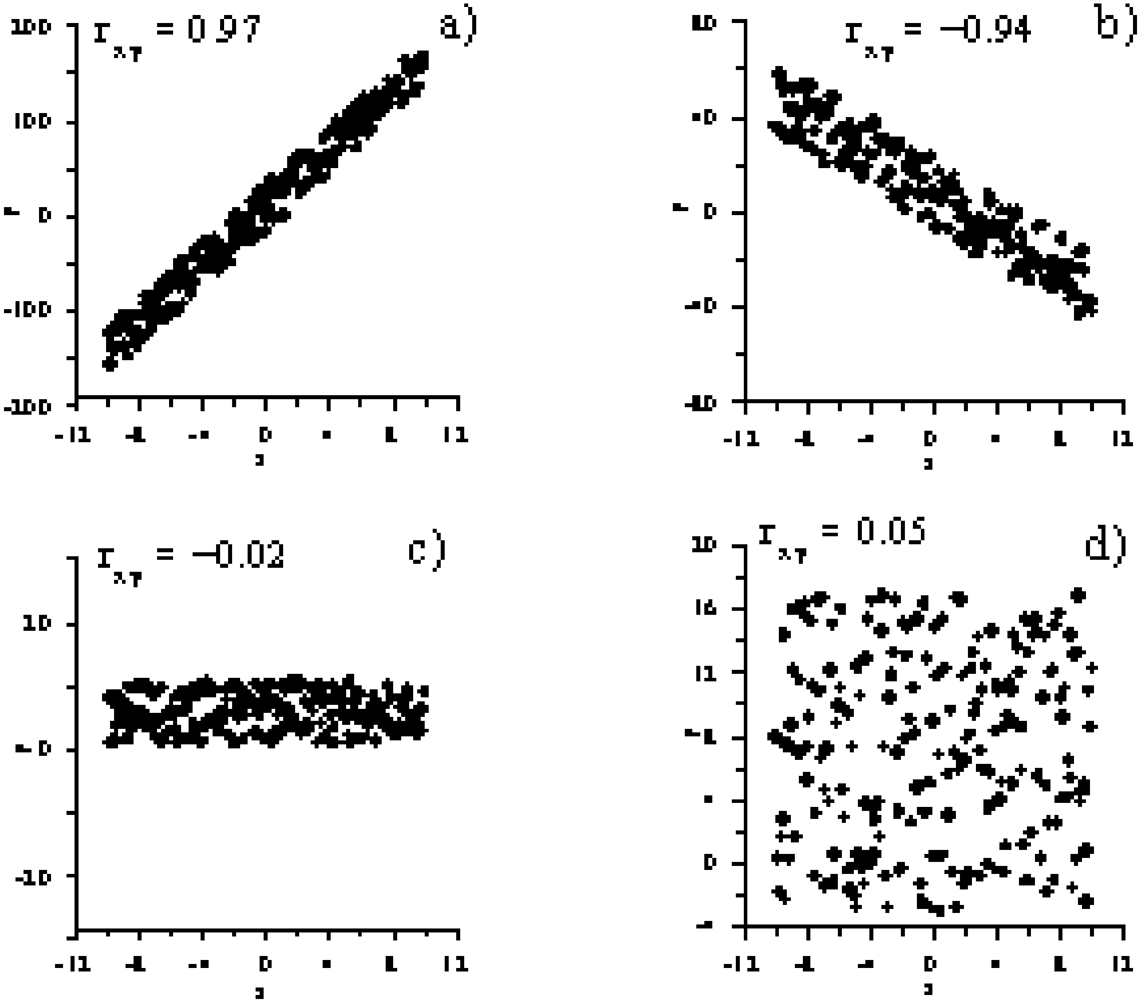
In this task two important tools are used:

- correlation analysis,
- regression analysis
  (LSQ-method: Least SQuares).

Cross-plots analysis:

Correlation analysis: evaluation of correlation coefficient r_{xy}

Its value is between -1 and 1 and it tells if the scattered data are focused around a linear trend and if this trend is increasing or decreasing.



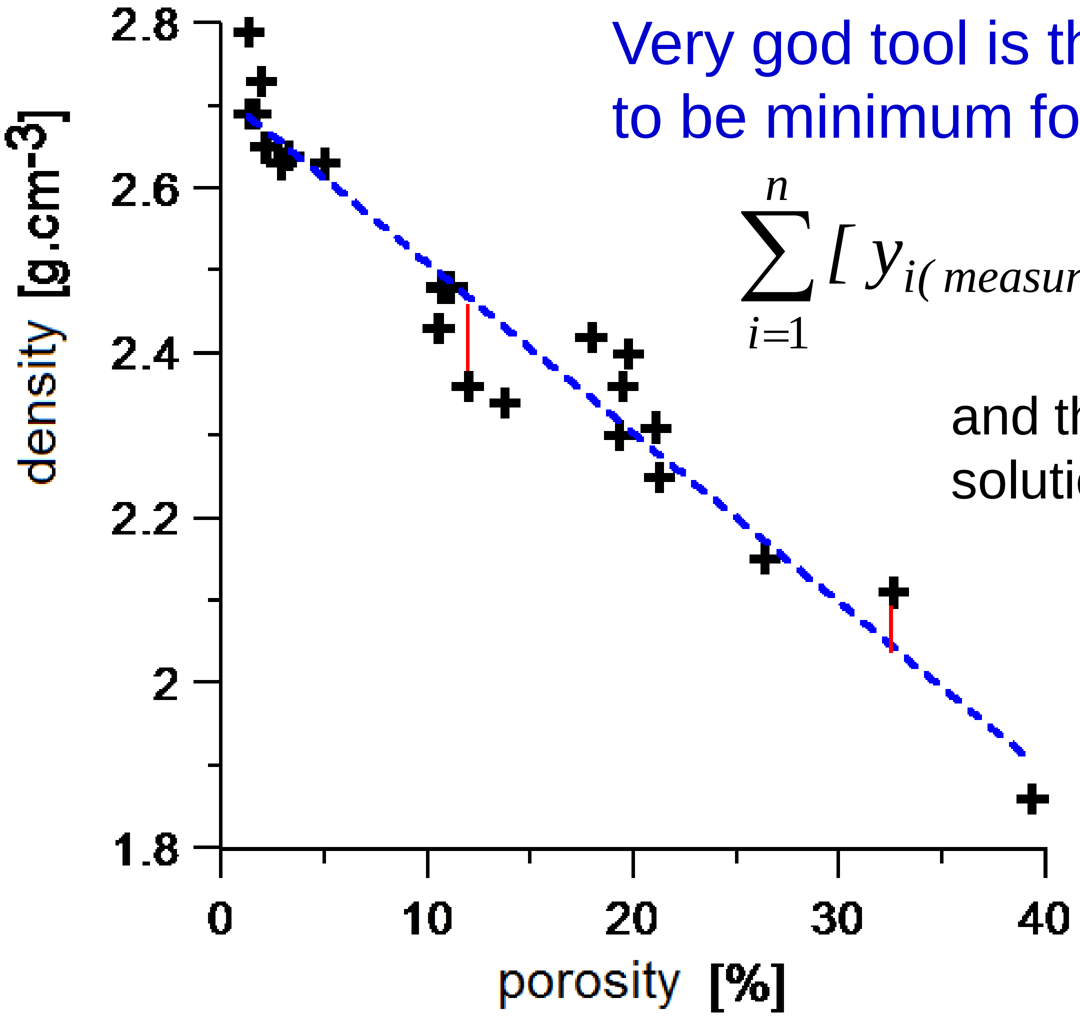
$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

here (x_i, y_i), are the samples of analysed parameters.

Cross-plots analysis:

Regression analysis:

Calculation of linear function (regression line) parameters.



Very god tool is the Least Squares (LSQ) – to be minimum for the best solution.

$$\sum_{i=1}^{n} [\, y_{i(measured)} - y(x_i)_{(teor)} \,]^2 = min.$$

and the linear function (theoretical solution) can be expressed as:

$$y(x_i)_{(teor)} = ax_i + b$$

where *a* is the slope and *b* the vertical shift of the straight line.

Regression analysis (LSQ-method):

(1/2)

We have to find such parameters *a* and *b*, for which the following simple equation is valid

$$\sum_{i=1}^{n}\left[y_i-\left(ax_i+b\right)\right]^2=\min.$$

We can understand the left-hand side of this equation as a function, for which we have to find a minimum (by means of partial derivatives evaluation and setting equal to zero):

$$\partial\left\{\sum_{i=1}^{n}\left[y_i-\left(ax_i+b\right)\right]^2\right\}\Big/\partial a=0\ ,\quad \partial\left\{\sum_{i=1}^{n}\left[y_i-\left(ax_i+b\right)\right]^2\right\}\Big/\partial b=0$$

Here we use the chain rule:

$$2\sum_{i=1}^{n}\left[y_i-\left(ax_i+b\right)\right]\left(-x_i\right)=0\ ,\quad 2\sum_{i=1}^{n}\left[y_i-\left(ax_i+b\right)\right]\left(-1\right)=0$$

and we get two equations with two unknowns (*a* and *b*):

Regression analysis (LSQ-method):

(2/2)

$$2\sum_{i=1}^{n}\left[y_i-(ax_i+b)\right](-x_i)=0\ ,\quad 2\sum_{i=1}^{n}\left[y_i-(ax_i+b)\right](-1)=0$$

and we get two equations with two unknowns (*a* and *b*):

$$\sum y_i x_i - a\sum x_i^2 - b\sum x_i = 0$$

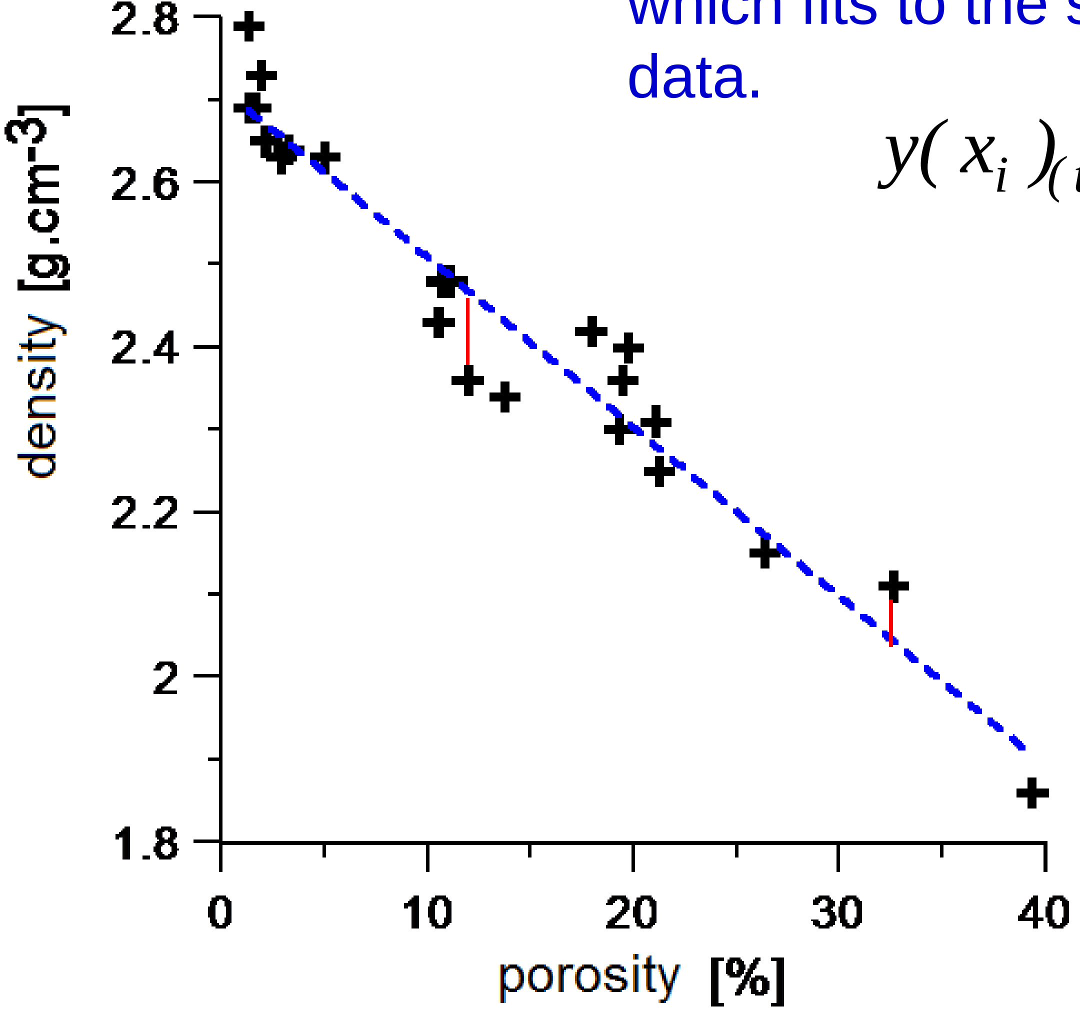
$$\sum y_i - a\sum x_i - bn = 0$$

where the summation sign is valid for *i* = 1 to *n*.

Finally we get:

$$\boxed{a=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{n\sum x_i^{2}-\left(\sum x_i\right)^{2}}\ ,\quad b=\frac{\sum y_i-a\sum x_i}{n}\ .}$$

Regression analysis (LSQ-method):



So, we have found parameters *a* and *b* for the theoretical function (a straight line), which fits to the scattered plot of measured data.

$$y(x_i)_{(teor)} = ax_i + b$$

LSQ method can be used in fitting of various functions (not only a linear function).