

Descriptive statistics

The statistical work can be split into four steps, but there are no sharp borders between them:

1.	collecting data	descriptive statistics
2.	organizing data	
3.	analysis of data	inductive statistics
4.	conclusions	

In the first two the concept of **probability** is not essential, in the last two the basis of **probability calculus** is **essential**.

1. Collecting data (sampling: see later)

data collection is motivated by a **goal**
(identification, discrimination)

Some part of data is **known**, just we have to ask from somebody, some part can be gained by **observation** and some part is **measurable** (medical examination).



2. Organizing data

In everyday life, we often deal with a large number of data that are connected to a given problem. We need to organize and summarize our observations because **we need an overview of the data.**

2/1. Tables

INFECTION	DISEASE	Absolute frequency		Relative frequency		Conditional relative frequency	
bacterial	Salmonellosis (Food poisoning by Salmonella)	94	208	0.280	0.619	0.452	1.000
	Scarlatina (Scarlet fever)	102		0.304		0.490	
	Other bacterial	12		0.036		0.058	
viral	Hepatitis infectiosa (Hepatitis)	22	126	0.065	0.375	0.175	1.000
	Mononucleosis infectiosa (Mono)	22		0.065		0.175	
	Lyssa (Rabies)	74		0.220		0.587	
	Other viral	8		0.025		0.063	
other	Other infections	2	2	0.006	0.006	1.000	1.000
total:		336	336	1.000	1.000		

absolute frequency: number of data in a given category

relative frequency: absolute frequency divided by the **total** number of elements in the set in question

It is a **ratio**, therefore, when speaking about relative frequency, both the **category** and the **set** we relate it to **must be specified.**

conditional relative frequency: absolute frequency divided by the number of elements in a **subset** of the set in question

Rules of summation

Absolute frequencies are additive **without condition.**

Relative frequencies are only additive **within the given set.**

Rule of multiplication

$$\frac{\text{number of Scarlatina}}{\text{number of bacterial infections}} \cdot \frac{\text{number of bacterial infections}}{\text{number of all infections}} =$$

$$= \frac{\text{number of Scarlatina}}{\text{number of all infections}}$$

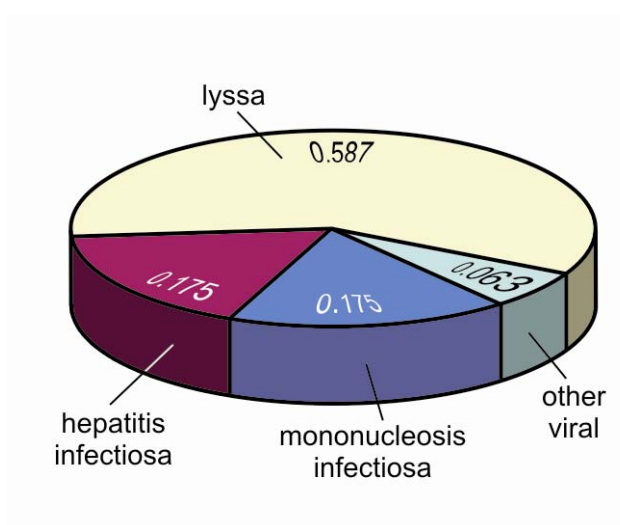
Problem:

Last year the **relative frequency** of fails at the final exam was 0.15, the **relative frequency** of excellents **among the passes** was 0.2. What was the relative frequency of excellents among all the exams?

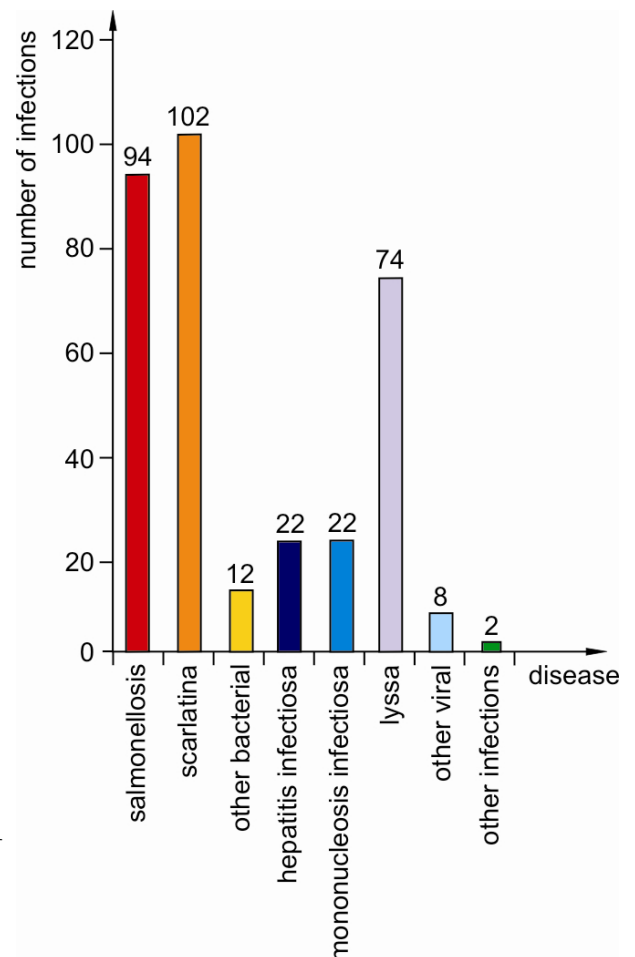
$$\frac{\text{number of fails}}{\text{number of all students}} + \frac{\text{number of passes}}{\text{number of all students}} = 1$$

$$\frac{\text{number of passes}}{\text{number of all students}} \cdot \frac{\text{number of excellents}}{\text{number of passes}} = \frac{\text{number of excellents}}{\text{number of all students}}$$

2/2. Diagrams



pie diagram

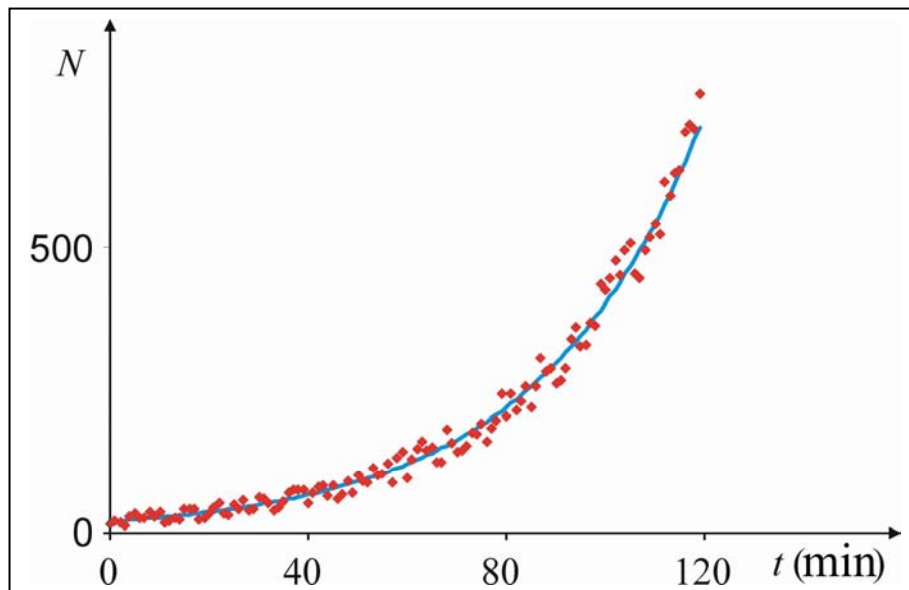


bar diagram

2/n. Organizing **numerical** data

2/n₁. Characteristics of quantitative data

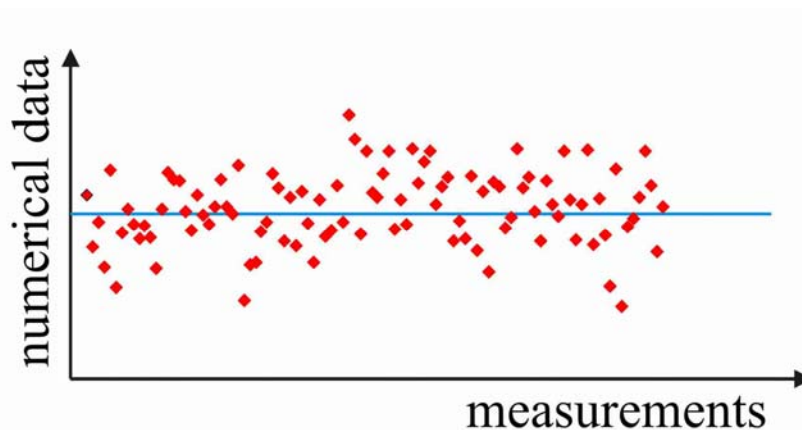
Premise: reproduction of a population of bacteria



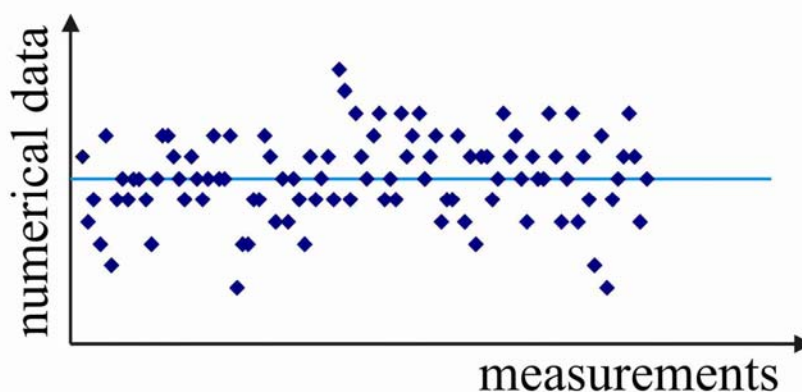
Deterministic and **stochastic** parts of the change appear **simultaneously**.

The question is whether the two parts can be separated?

At the beginning, **let us suppose** that we study data which have **no deterministic** changes.



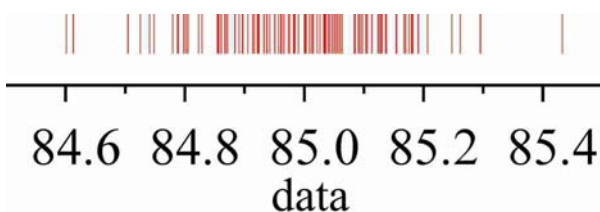
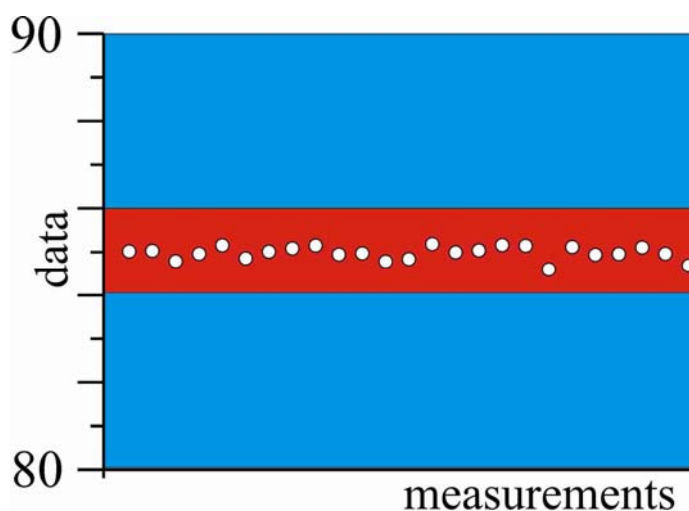
„continuous” case



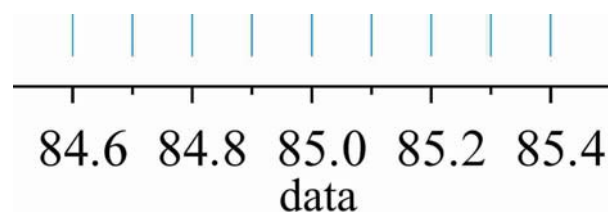
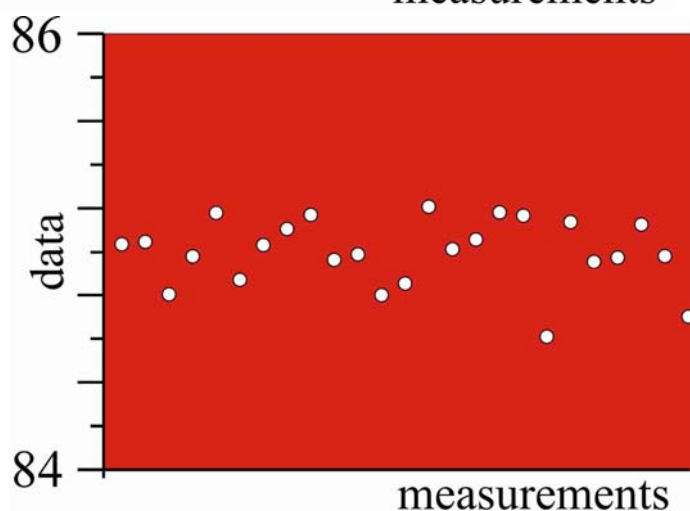
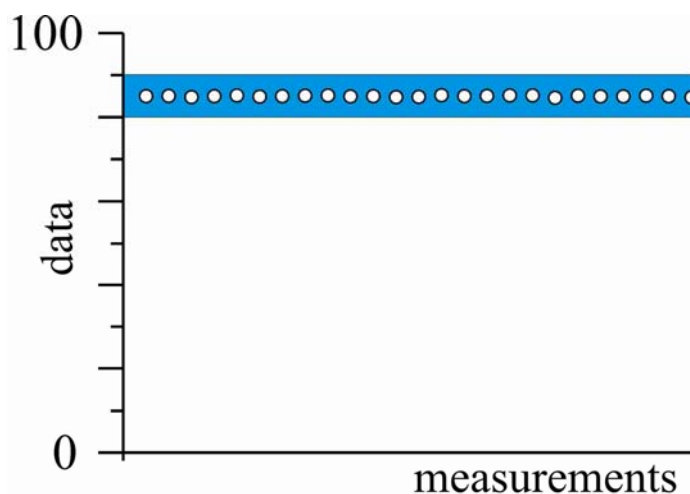
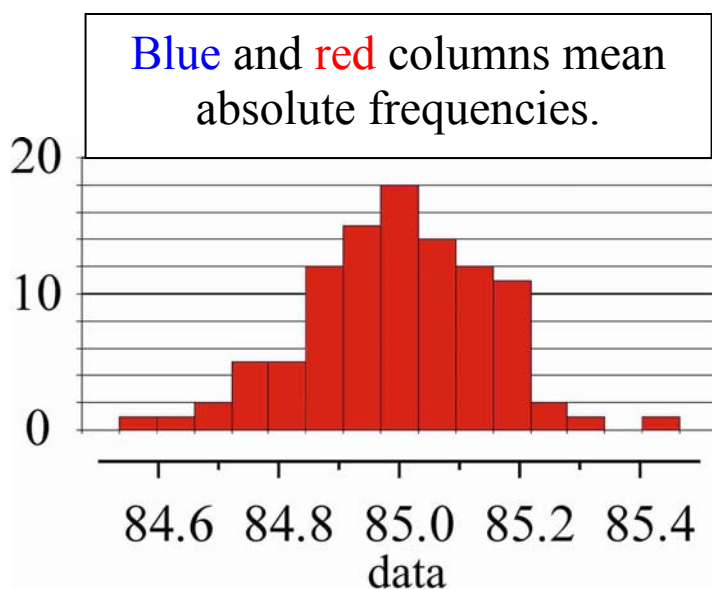
discrete case

Let us take specific numbers.

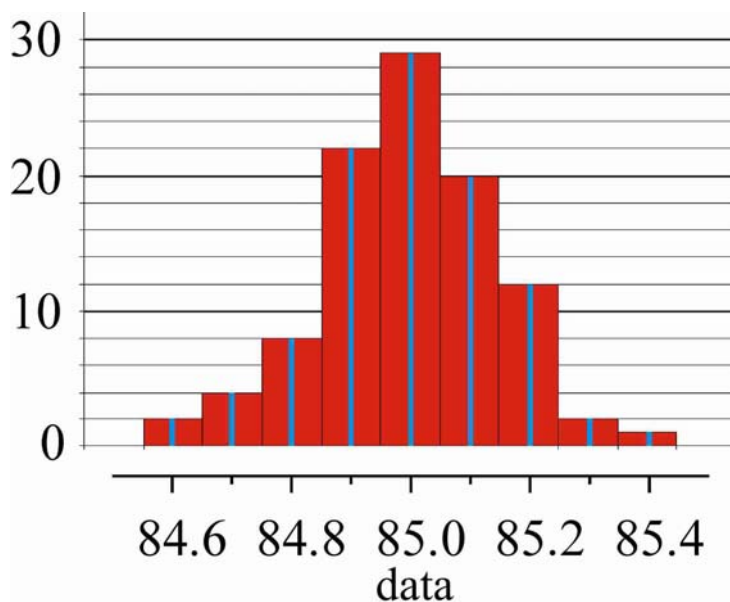
85.0361847182	85.0
85.0472573819	85.0
84.8050250975	84.8
84.9817905059	85.0
85.1795262649	85.2
+ 95 numerical data	



At **continuous case** we “never” get identical data.



At **discrete case** we have to give the **frequencies**.



Frequency distribution

In the **discrete case** it is unambiguous.

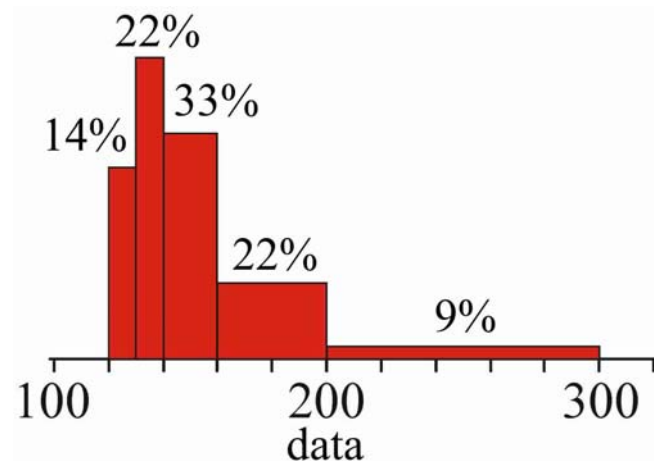
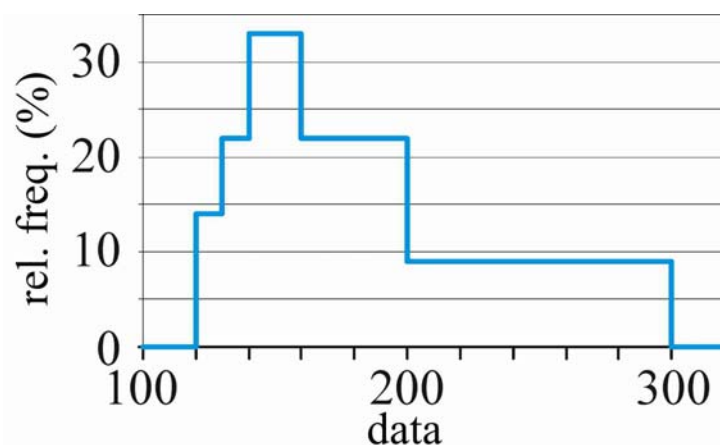
In the **continuous case** its shape depends on the width and location of intervals named **classes** or **bins** (but not so much).

For better comparison between data sets with different bins usually the **relative frequencies** are given.

E.g. Financial statement of salaries (HUF):

	abs. freq.	rel. freq.
between 120 and 130 thousand	124	14%
between 130 and 140 thousand	195	22%
between 140 and 160 thousand	293	33%
between 160 and 200 thousand	195	22%
between 200 and 300 thousand	80	9%
total	887	100%

How can we represent it?



Histogram

relative frequencies are proportional to the area of the columns. Total area is $100\% = 1$.

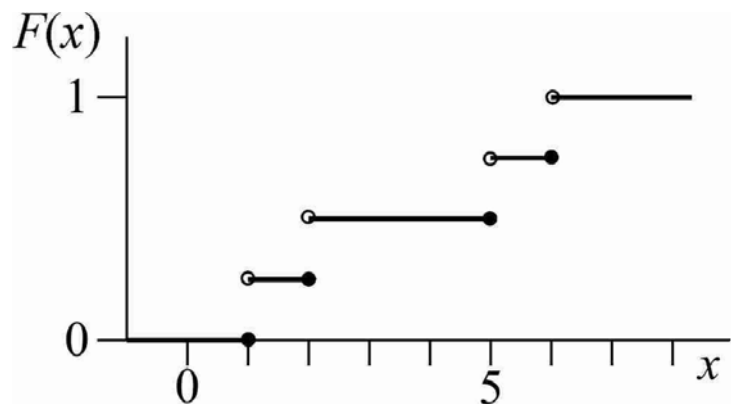
If the width of the columns is the same (**equal classes**) the two representations are **identical**.

Distribution function ($F(x)$)

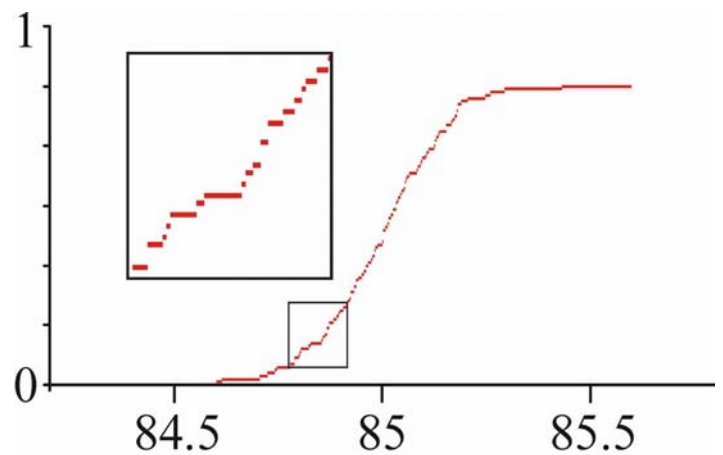
For a data set $[x_1, x_2, x_3, \dots, x_n]$:

$$F(x) = \frac{\text{number of data smaller than } x}{n}$$

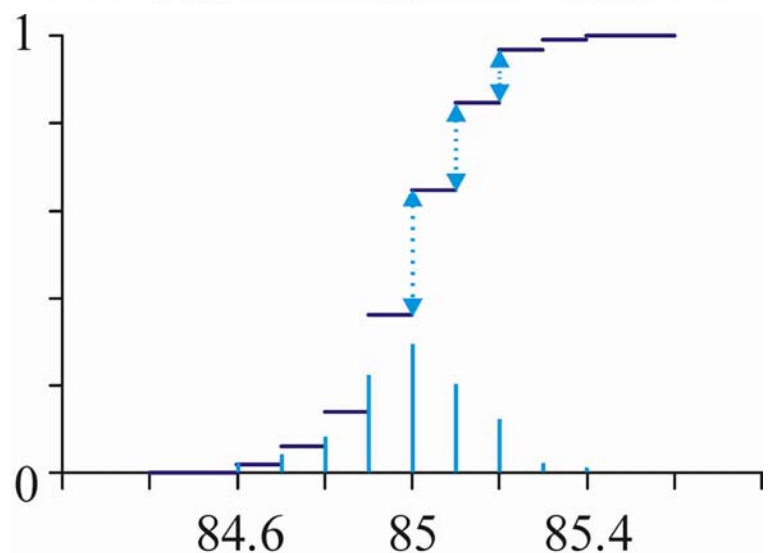
E.g. (1) $[1, 2, 5, 6]$



E.g. (2) The earlier 100 data
at continuous case.



E.g. (3) The earlier 100 data
at discrete case.



The columns show the **relative frequencies**. If we consecutively add up these columns, we get the respective values of the distribution function. The other way round, the **differences** of distribution function give us the **relative frequencies**.

unimodal			multimodal
symmetrical	right skewed	left skewed	
leptokurtic		platykurtic	

If we have only **few data** we can **not** see the **characteristics** of frequency distribution.

2/n₂. Numerical characteristics of quantitative data

(can be determined **in any cases**)

Location measures (**data set** with n elements)

1. **mean** (arithmetical average)

1. E.g. [1, 2, 2, 7]; 2. E.g. [1, 1, 3, 7]

$$x_{\text{mean}} = \bar{x} = \frac{1+2+2+7}{4} = 3 = \frac{1}{4} \sum_{i=1}^4 x_i = \frac{1}{n} \sum_{i=1}^n x_i$$

It is **sensitive** to the extreme values!

Location measures because of their position

2. **median** (x_{median})

We order the data according to their magnitudes, and look for the middle or middles.

1. E.g. [1, 2, 2, 7]

$$x_{\text{median}} = 2;$$

2. E.g. [1, 1, 3, 7]

$$1 \leq x_{\text{median}} \leq 3$$

3. **mode** (x_{mode})

If the data set has identical data, the one which has the most copy called as mode. (But more mode also may exist in the same data set.)
(„mode” → fashionable)

1. E.g. [1, 2, 2, 7]

$$x_{\text{mode}} = 2;$$

2. E.g. [1, 1, 3, 7]

$$x_{\text{mode}} = 1$$

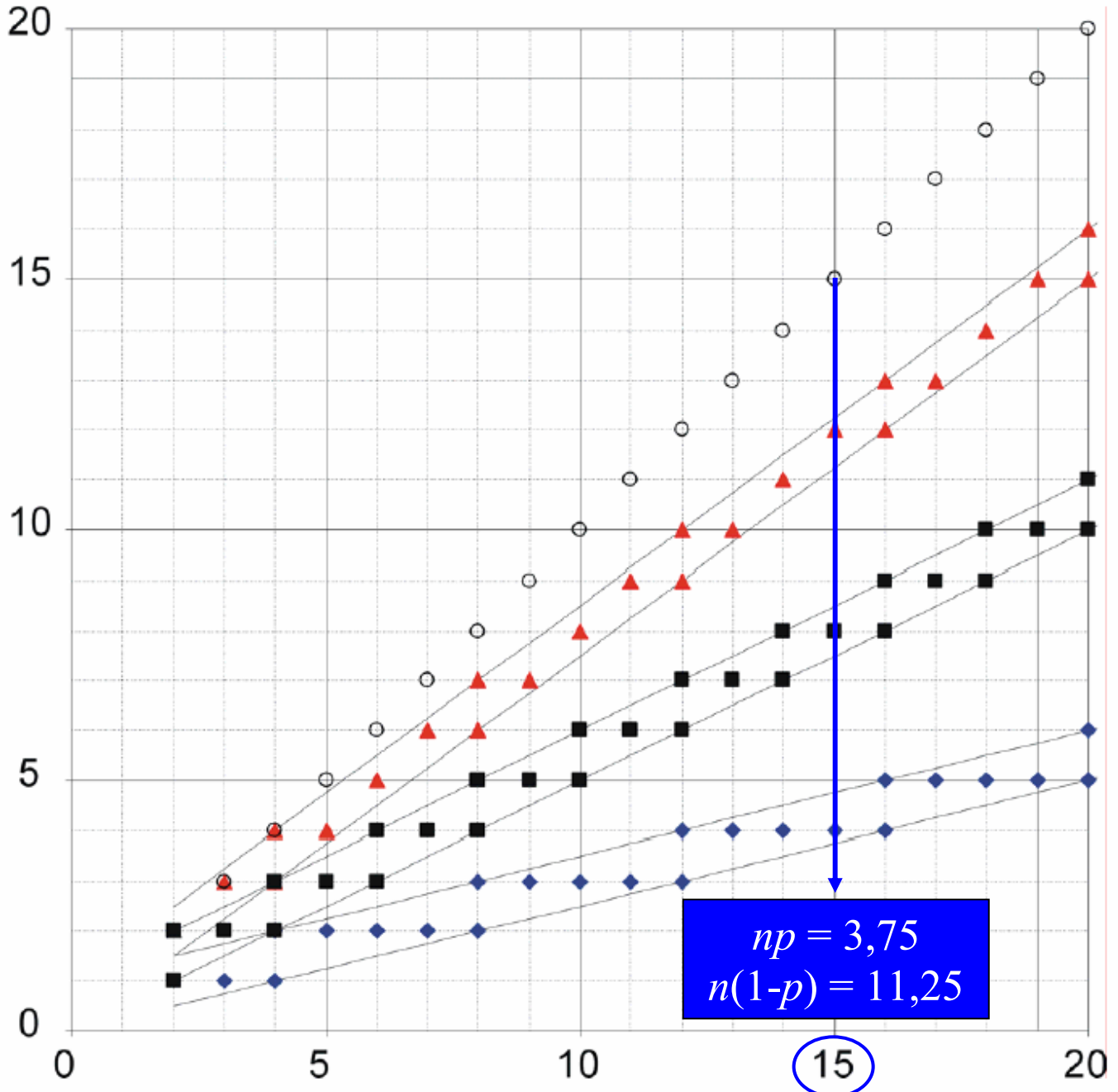
They are **not sensitive** to the extreme values!

Quantiles

What riches makes a person become a member of the “upper ten thousand”.

First we order the data by magnitude again.

E.g. lower quartile, middle quartile = median, upper quartile



Let p be a number between 0 and 1. p -quantile of a data set is the number for which the number of smaller elements is at the most np , and the number of bigger elements is at the most $n(1 - p)$.

quintiles ($p = 1/5$), deciles ($p = 1/10$), percentiles ($p = 1/100$)

Characteristics of measures of spread

1. **range**

the difference of the biggest and the smallest elements of the data set

How much is the **average deviation of the data from the mean**?

1. E.g. [1, 2, 2, 7] $(3 - 1) + (3 - 2) + (3 - 2) + (3 - 7) = 0$

2. E.g. [1, 1, 3, 7] $(3 - 1) + (3 - 1) + (3 - 3) + (3 - 7) = 0$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = 0$$

How much is the **average absolute deviation of the data from the median**?

If x^* is a fixed value, it can be proven that the expression

$$\frac{1}{n} \sum_{i=1}^n |x_i - x^*| \text{ is minimal if } x^* = x_{\text{median}}.$$

How much is the **average squared deviation of the data from the mean**?

If x^* is a fixed value, it can be proven that the expression

$$\frac{1}{n} \sum_{i=1}^n (x_i - x^*)^2 \text{ is minimal if } x^* = \bar{x}.$$

2. **variance** (s_x^2)

average of the squared deviation of the data from the mean

3. **standard deviation**

of the data set is given by the formula

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

Further characteristics can also be quantified (**skewness, kurtosis**).