The role of biostatistics and informatics in the every day medical practice

The **purpose** of medical science:

- Prevention of diseases,
- Healing of the sick

Diagnostics: **scientific** methodology of recognition of diseases.
Therapy

Auxiliary sciences: e.g. anatomy, physiology, physics, chemistry, biology; *and*

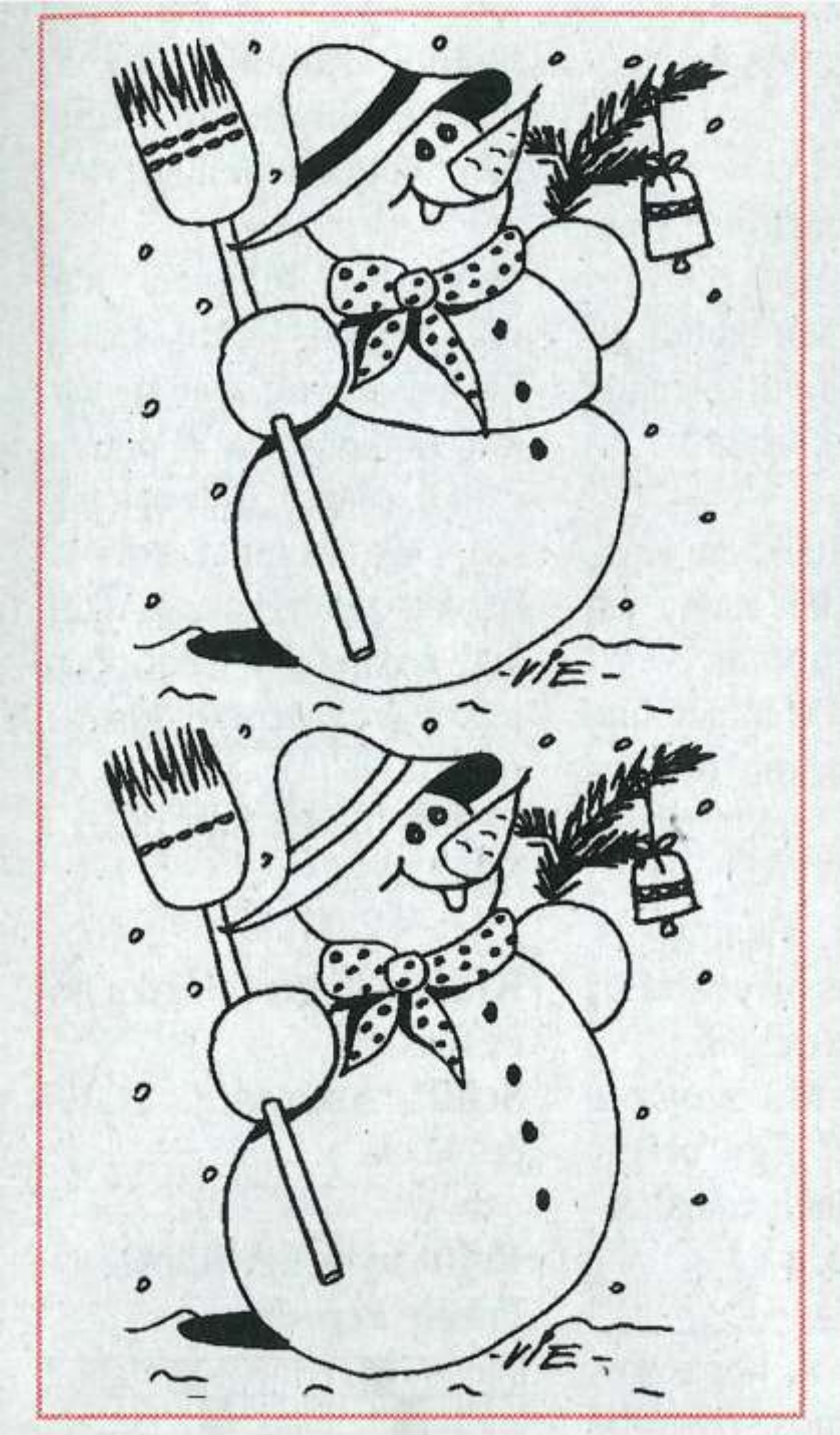
Biostatistics and informatics

Medical doctors: series of decesions
Confidence
Lots of uncertainty

General experience:
many of us make inappropriate conclusions easily, and make decisions based on them. (whisky)

Main purpose of biostatistics and informatics:

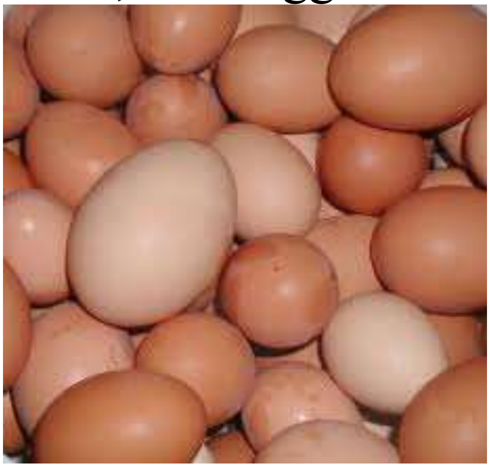
to know **quantitatively**;
two or more things are similar or different



Data: facts for the cognition, characterization of somebody or something;
qualitative and quantitative characteristics of the surrounding world.

Signals: transmitter units of data (suitable for description of data)

Identity: sometime names, “two eggs”



“theoretical ball” model
spherical
white
with 38 mm diameter
with 2.5 g mass

“practical ball” reality
measurements

Honvédelmi Minisztérium Állami Egészségügyi Központ
1134 Budapest, Róbert Károly krt. 44. Tel.: 06-1-465-1800 Fax: 06-1-340-3129
Főigazgató:
Működési engedély száma:

Központi Laboratóriumi Diagnosztikai Osztály
Osztályvezető főorvos :

Laboratóriumi eredmények

Megnevezés	Érték	M.e.	Megjegyzés	Eltérés	Referencia értékek	
Klinikai kémia						
Glukóz	3,5	mmol/l			3,1	5,6
~~Karbamid~~	6,1	mmol/l			1,7	8,3
Kreatinin meghat.	75	µmol/l			44	80

Zuglói Egészségügyi Szolgálat
1148 Budapest, Örs Vezér tér 23.
Telefon: 469-4600

LABORATÓRIUMI LELET

Szakorvosi Rendelőintézet
Laboratórium
Labor vezető:

Lelet kelte:

Páciens neve:
TAJ szám: Nem:
Született: **Napi sorszám: 749** Beut. egység: **340092019** Azon.: 012101003
Anyja neve:

Kért vizsgálatok:		Eredmény: mértékegység	Referencia érték:
VÉRKÉP XT WBC		**10,71** 10'3/u	4,0 - 13,0
RBC vvt szám		**4,22** 10'6u	3,9 - 5,6
~~KARBAMID~~	+	**9,6** mmol/l	1,7 - 8,3
KREATININ	+	**113,0** umol/l	50,0 - 110,0

Semmelweis Egyetem ÁOK Központi Laboratórium
1083 Budapest, Korányi Sándor u. 2/a.
Intézetvezető:
Tel: 06 1 2100 278/1522,1457

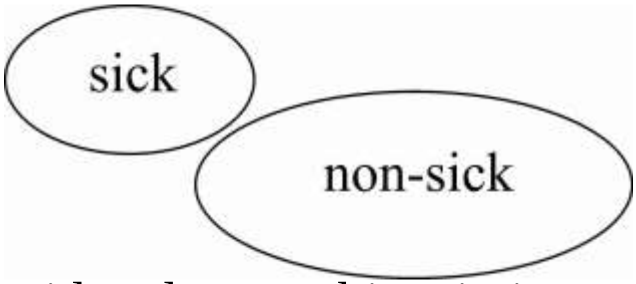
LABORATÓRIUMI EREDMÉNYKÖZLŐ LAP

Név :
Születési idő : Nem:
TAJ/azonosító : Rendelés sorszáma: 6037990

Vizsgálat	Eredmény		M.Egység	Ref.tart
VVt süllyedés	2		mm/h	1-20
~~Karbamid~~	6,7		mmol/l	2,5-8
Kreatinin	108	*	umol/l	62-106

The most important fundamental concepts and the connected problems

Who is sick and who is healthy?



Set: collection of distinct objects, considered as an object in its own right. They are characterized uniquely. Things belonging to the set are the **elements** of the set.
In general: **variable**

In which set the given element can be found?
Systematization, classification, separation

What is the similarity between this toy and the diagnosis?

Not so much!

3 very different bodies
Characteristics:

„case/variable”	**shape**	**color**	**size**
1	sphere	**yellow**	4,3 cm
2	tetrahedron	**blue**	4,5 cm
3	cube	**red**	3,8 cm

3 very different holes
Characteristics:

„case/variable”	**shape**	**color**	**size**
1	circle	**yellow**	4,3 cm
2	triangle	**blue**	4,5 cm
3	square	**red**	3,8 cm

„Can not” miss it.
There is one-to-one correspondence.

Same (exists only exceptional cases)
("We both step and do not step in the same rivers." Heraclitus)

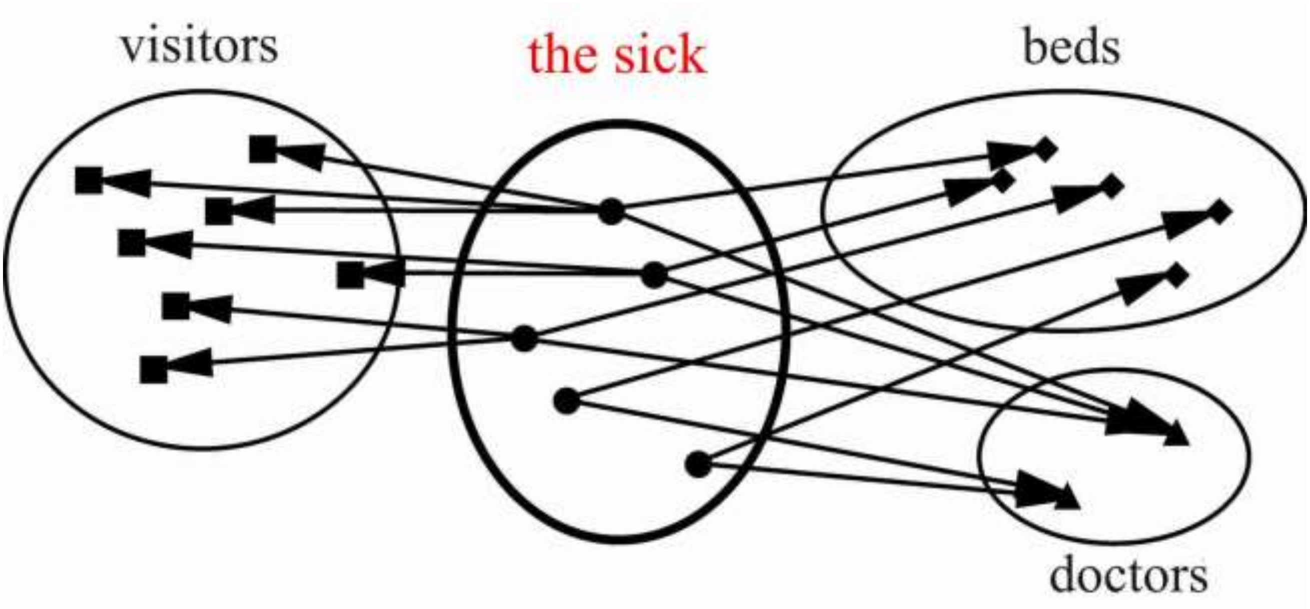
Instead, more or less **similar**

The absence of uniqueness can cause the problems.
We are not able to take into account all of the circumstances.

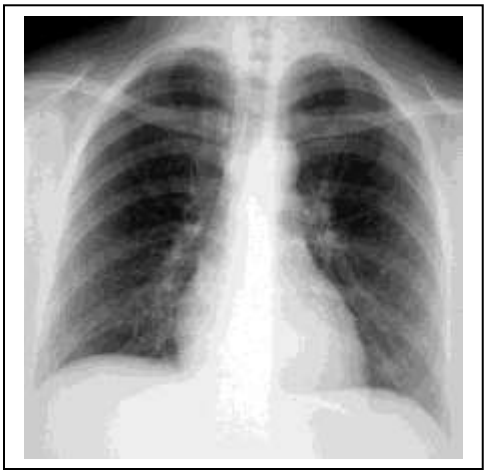
Measure of similarity: confidence

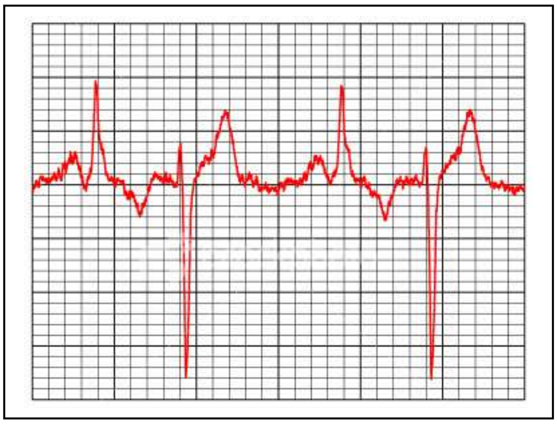
The essence of the problem:
There are no two identically sick people.
There are diseases which can produce very similar symptoms.

Function (mapping),
but there are exceptions



changes in space or time, (or in both) e.g. change of light, sound, any sensation or a measurable quantity





The role of „change” in theory and practice

The “most important” feature of a function is the **change.**

How does it change?
Increases or decrease; quickly or slowly

The simplest function is the linear one: $y = ax + b$
(in most cases we prefer it)

Some further important functions

1. Exponential function

$$y = b\,2^{ax}$$

2. Logarithmic function

$$y = a \log_2 x + b$$

3. Powerfunction

$$y = b\,x^{a}$$

Remarks:

1. $\log_2 y = \log_2 b + a\,x\,\cancel{\underbrace{\log_2 2}_{=1}}$

3. $\log_2 y = \log_2 b + a \log_2 x$

After this transformation we get a linear function in all cases.

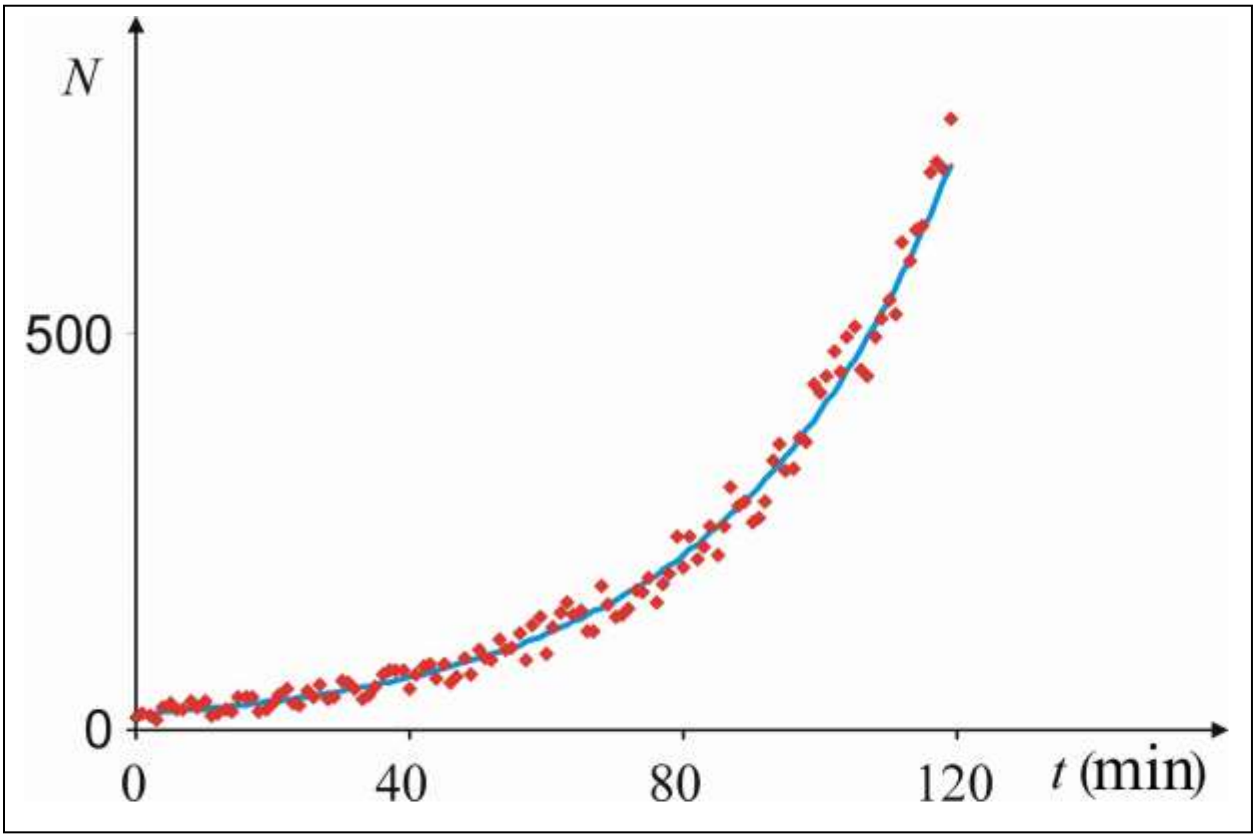
Deterministic part (determined by circumstances, which could be taken into account) and stochastic part (determined by circumstances, which could not be taken into account) of changes appear simultaneously.

E.g. reproduction of a bacteria population

Theory (model)

$$N(t) = N_0 2^{\frac{t}{T}}$$

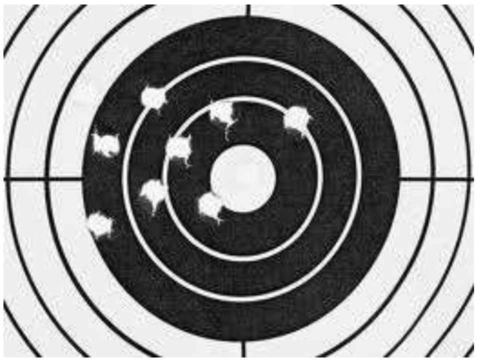
Practice (we should measure)



There are uncertainties which come from the measurement, but they can also be caused by the properties of the measured quantity.

Statistical laws

There are circumstances which we can not take into account (target)



Model: probability calculus

Fundamental concept:

Phenomenon: all the things which are repeatable in essence at identical conditions, in connection with them we can do observations we can make “experiments”.

Observation: we give what we are interested in, in connection with the phenomenon and how we can detect or measure it.

Event: a statement which comes true or not.

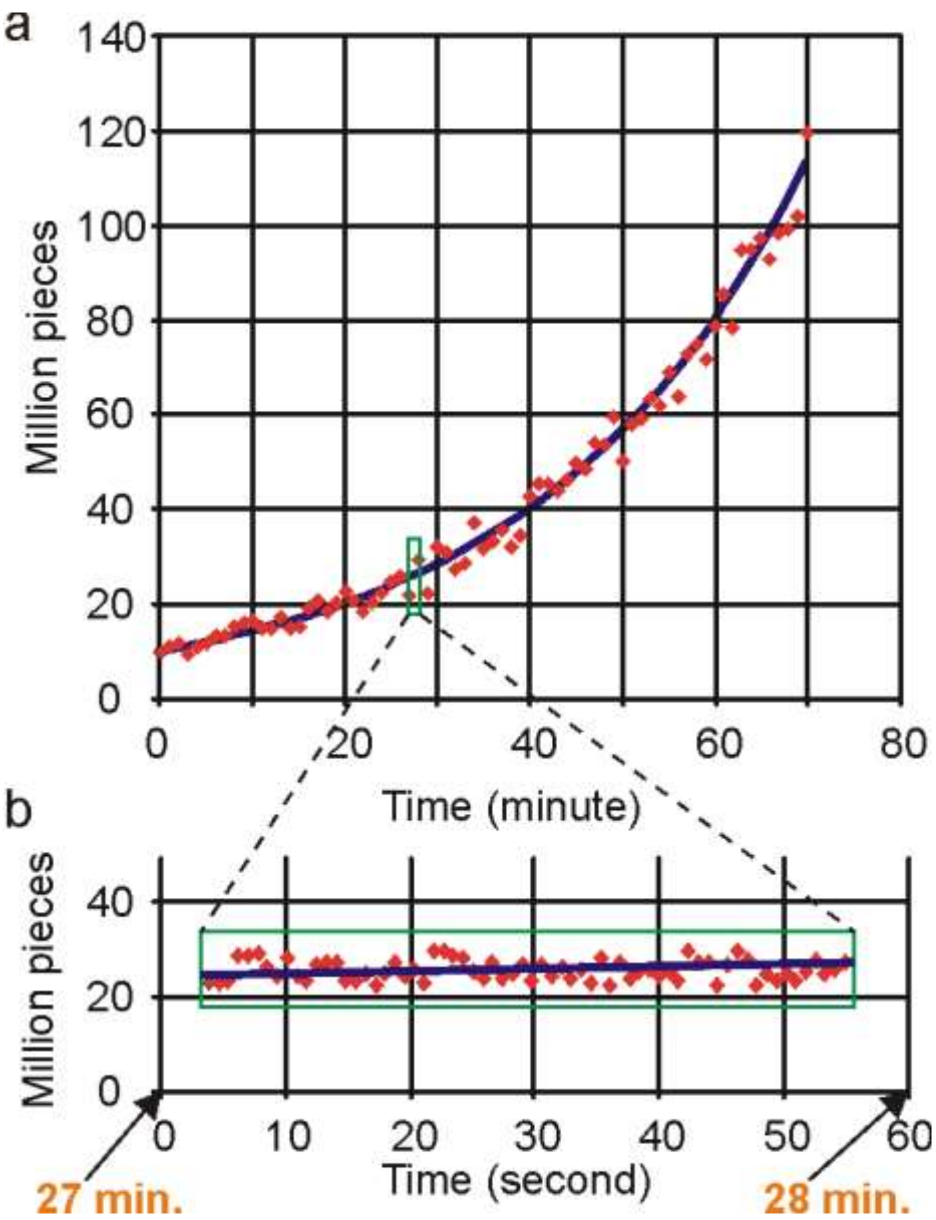
	examples			
Phenomenon	medical examination	toss of a coin (1)	waiting for a tram	toss of a coin (2)
Observation	color of skin	falling time of the coin	how many passengers	which side
Event	yellow	between 0.5 s and 1.5 s	10 passengers	head

The more frequent, the more probable

Multiplication of a bacterial colony in theory, according to the suitable deterministic mathematical model (blue curve) and in practice, based on measurements (red symbols).

Theory (model) **Practice** (have to measure)

$$N(t) = N_0 2^{\frac{t}{T}}$$



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

The question is whether the two parts can be separated?

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

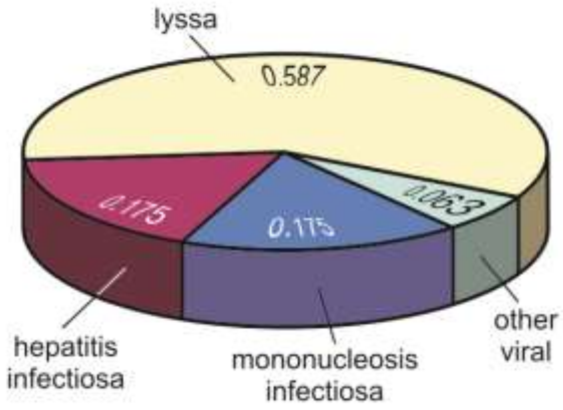
INFECTION	DISEASE	Absolute frequency		Relative frequency		Conditional relative frequency	
bacterial	Salmonellosis (Food poisoning by Salmonella)	94	208	0.280	0.619	0.452	1.000
	Scarlatina (Scarlet fever)	102		0.304		0.490	
	Other bacterial	12		0.036		0.058	
viral	Hepatitis infectiosa (Hepatitis)	22	126	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

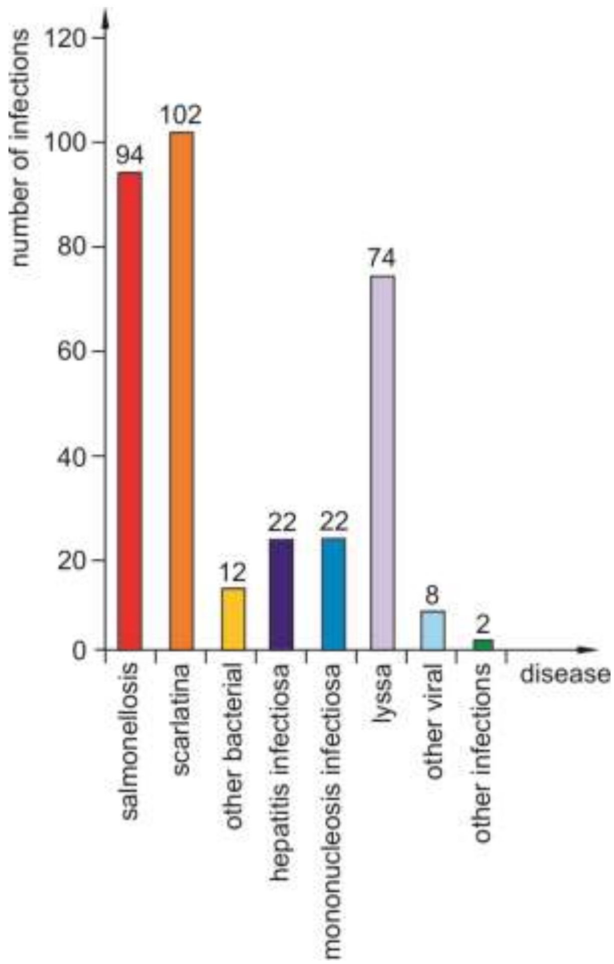
relative frequency: absolute frequency divided by the total number of elements in the set in question
It is a ratio, therefore, when speaking about relative frequency, both the category and the set we relate it to must be specified.

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

2/2. **Diagrams**

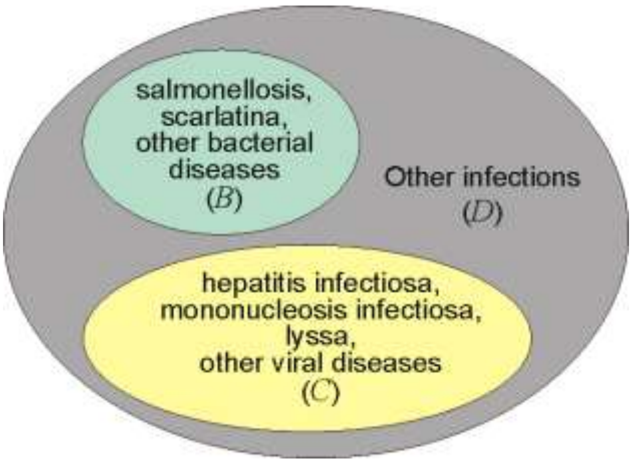


pie diagram



bar diagram

Diseases as subsets



$$B \cup C \cup D = A$$
$$B \cap C \cap D = \emptyset$$

Subsets and events correspond to each other

Viral disease as event

	example
Phenomenon	medical examination
Observation	origin of disease
Event (C)	viral

Rules of summation (I) and multiplication (II)

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?

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

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

(I)

Absolute frequencies are additive without condition.

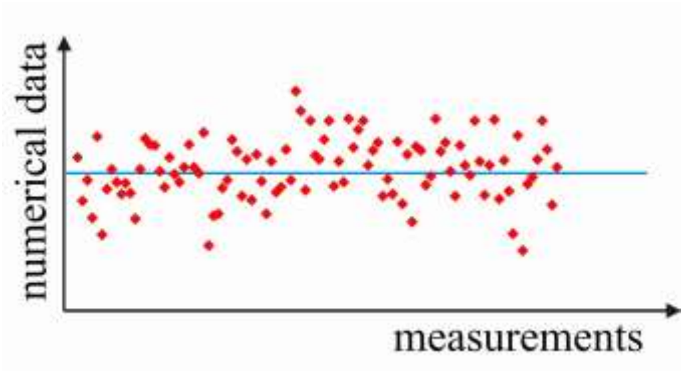
Relative frequencies are only additive within the given set.

(II)

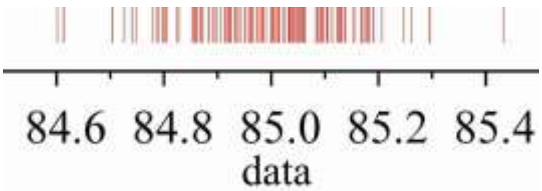
Conditional relative frequency and relative frequency (without condition) are multiplicative according to the method shown above.

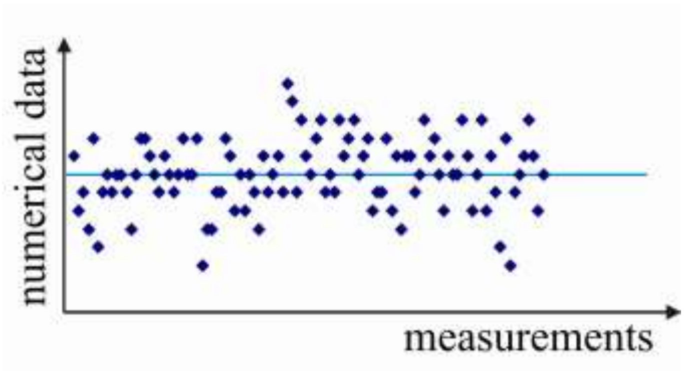
Characteristics of quantitative data

At the beginning, let us suppose that we study data which have no deterministic changes. (order is not important)

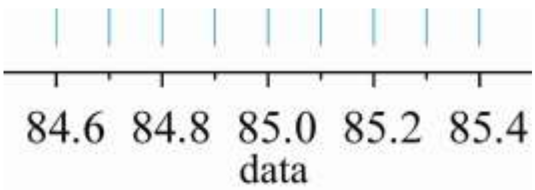


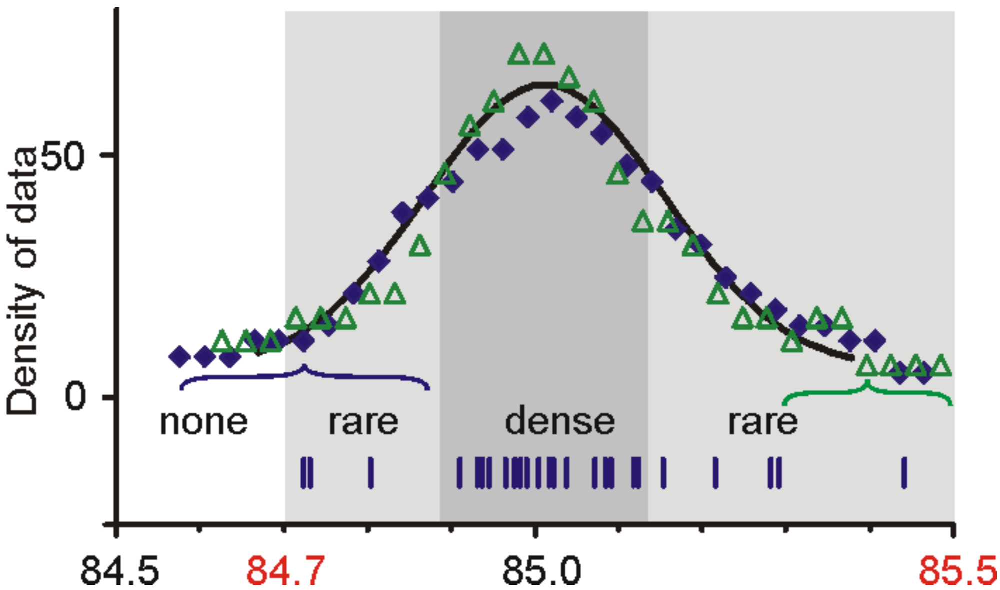
„continuous" case
"never" two identical



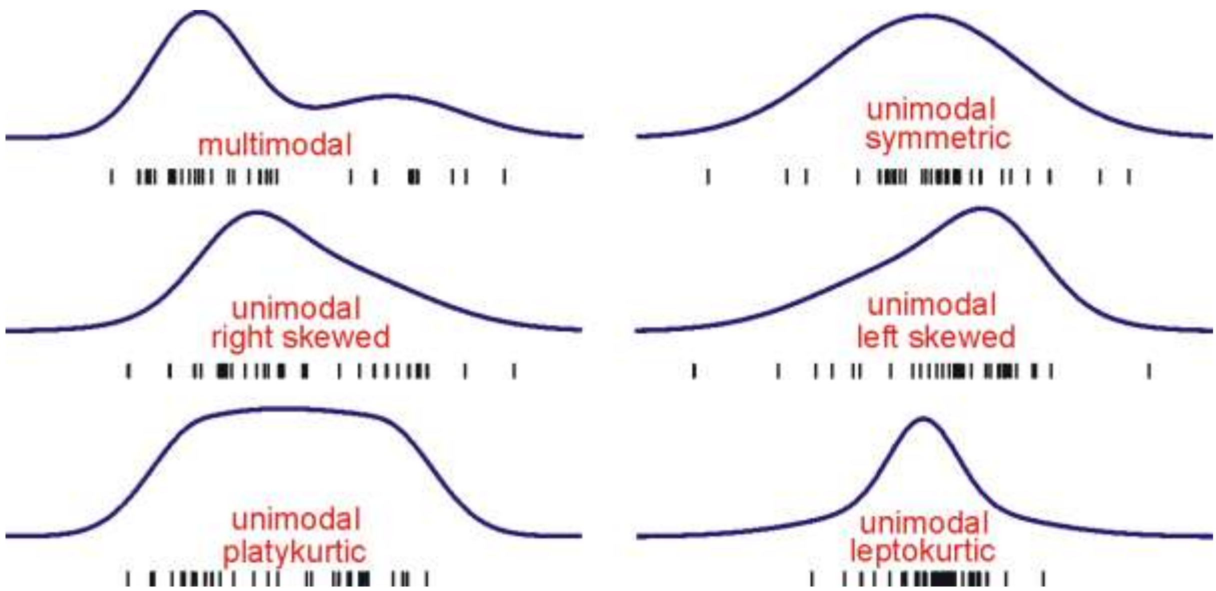


discrete case
We have to give the frequencies.



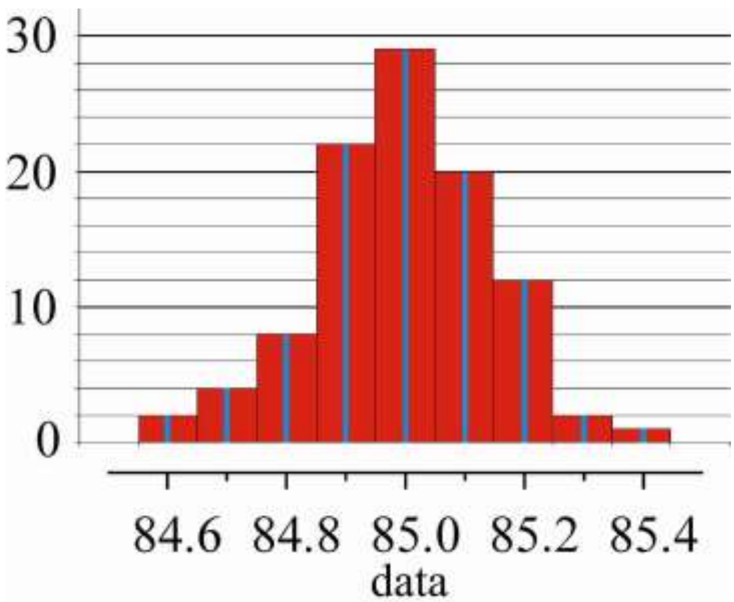


How can we characterize density of data?
Main types



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).



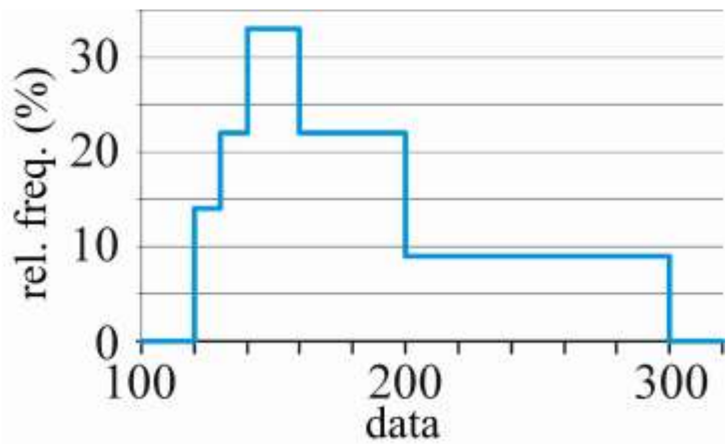
For better comparison between data sets with different bins usually the relative frequencies are given.

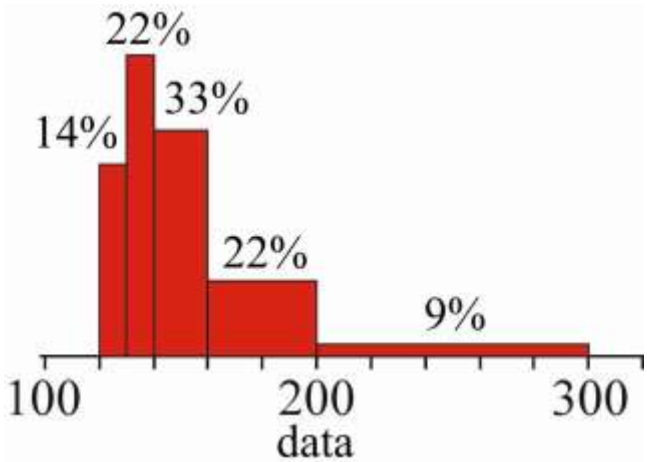
How can we characterize it if we do not know all data?

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.

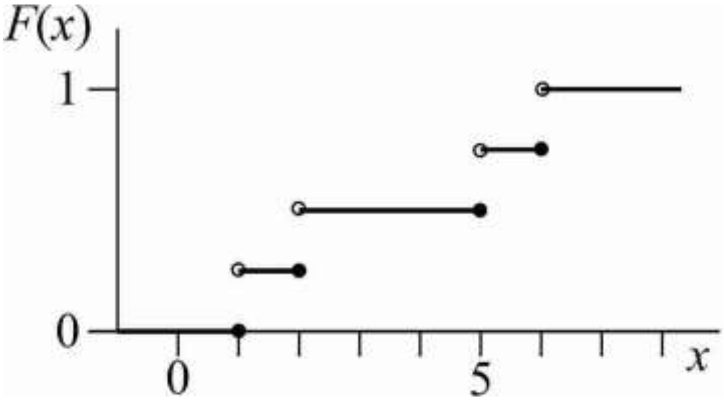
If the width of the columns is the same (equal classes) the two representations are identical.

Distribution function ($F(x)$)

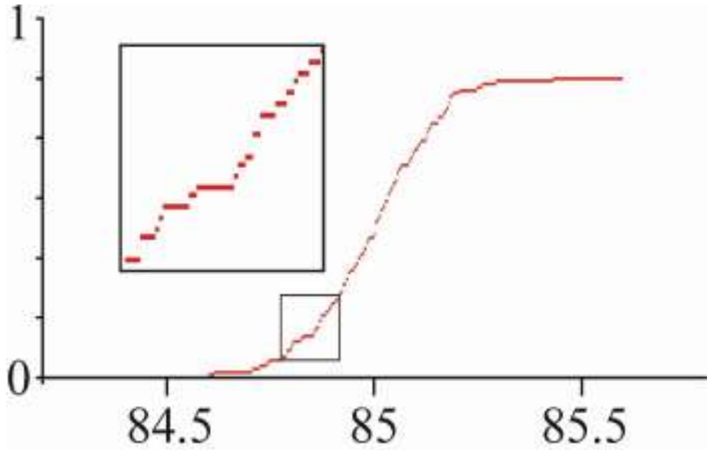
For a data set $[x_1, x_2, x_3, \ldots, x_n]$:

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

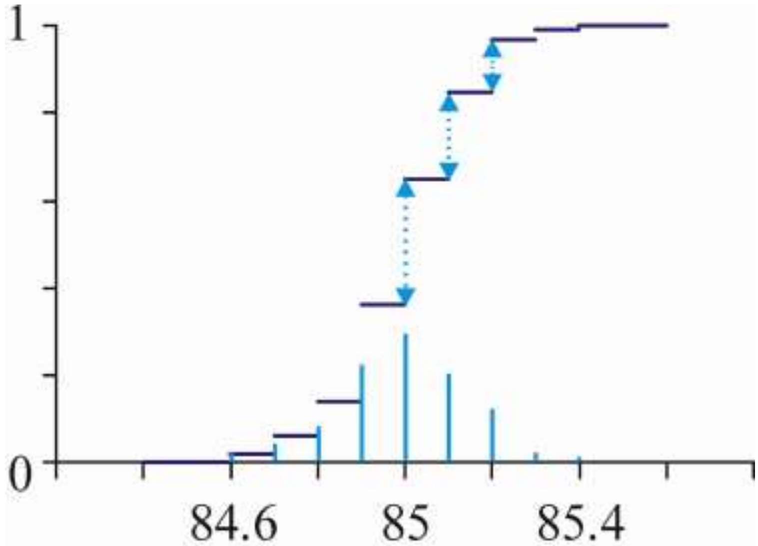
E.g. (1) [1, 2, 5, 6]



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



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

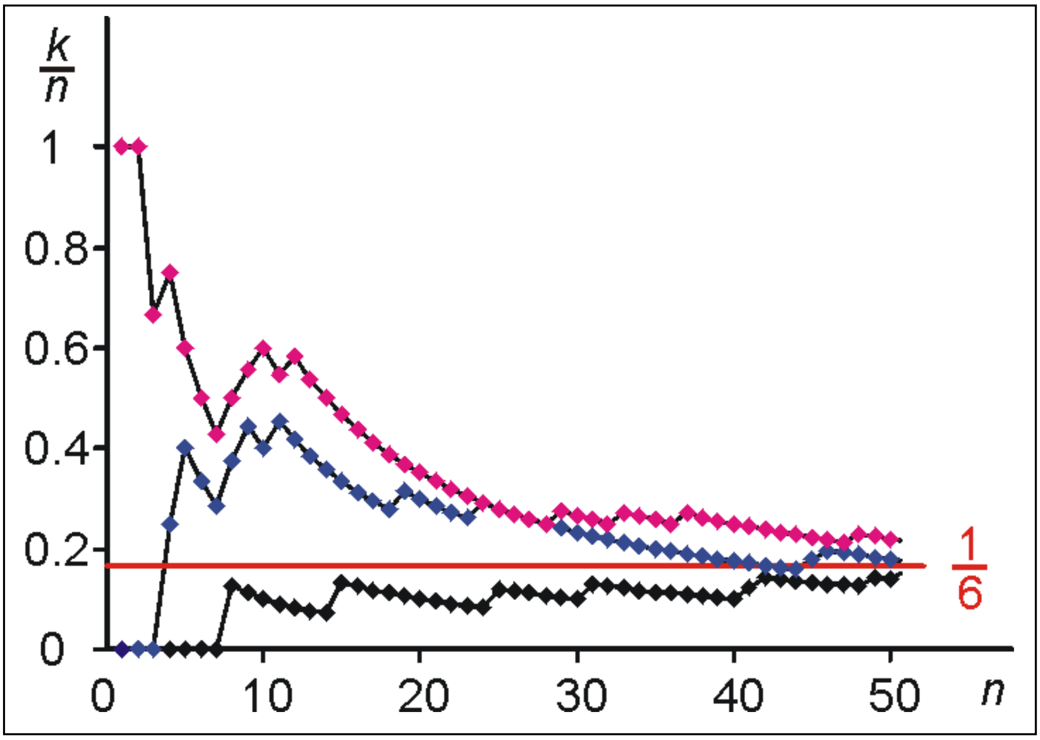


The columns show the relative frequencies. If we consecutively add up these columns, we get the respective values of the distribution function. The other way round, the differences of distribution function give us the relative frequencies.

Elements of probability calculus

Relative frequency of the event in the series of trials:
k/*n*, where *k* is the absolute frequency of the occurrence of the event and *n* is the number of experiments.

E.g. Phenomenon: a die is rolled.
Observation: what is the outcome.
Event: the result is 6.



Law of large numbers (for the relative frequencies):

As the *n* (number of die rolls) increases, the relative frequency, *k*/*n* becomes stable around a certain value. This value is independent of the actual series of trials.

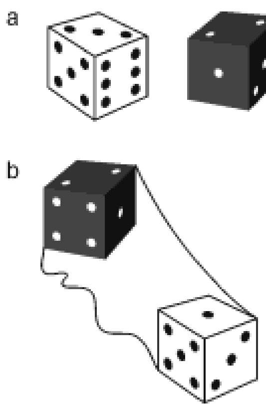
(It is an empirical fact it can not be proven by logical sequence.)
(Karl Pearson 1857-1936)
We assign a number to the event: **probability**.

Properties of probability:

1. The probability of an event [$P(A)$] is always: $0 \le P(A) \le 1$.
2. The probability of certain event is: $P(sure) = 1$.
3. The probability of the union of two mutually exclusive events (e.g. A and B, $A \cap B = \emptyset$) is:
   $P(A \cup B) \equiv P(A+B) = P(A) + P(B)$.

Independence

1000 rolls



a

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	30	25	30	29	28	25
⚁	24	27	31	27	24	27
⚂	28	30	39	32	24	29
⚃	28	28	22	26	27	33
⚄	27	24	26	21	31	27
⚅	30	25	32	30	29	25

b

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	40	41	46	12	9	21
⚁	51	38	37	13	22	15
⚂	42	49	52	8	20	17
⚃	8	10	15	36	52	44
⚄	11	16	9	45	39	35
⚅	10	17	8	43	41	28

Conditional probability

The probability that "the result of the black die is 1" (event A) if "the result of the white die is 1" (event B),
$P(A|B)$: the probability of event A is conditioned on the prior occurrence of event B.

If $P(A|B) = P(A)$ then event A is statistically **independent** of event B

If $P(A \cap B) \equiv P(AB)$ is the probability of occurrence of A and B, then

$$P(A|B)\,P(B) = P(A \cap B) \qquad \text{(rule of multiplication)}$$

Independence (equivalent equation): $P(A)P(B) = P(A \cap B)$.

Problem:

After a die roll, are the next two events independent or not? The result is smaller than 3 (event A), the result is even (event B).

Use the previous equation!

Random variable

We observe a quantitative thing in connection with a phenomenon.

1. We give what and how to "measure".
2. Random variable is characterized by its distribution or by the parameters of distribution, if they exist.

In general we do not know these parameters.

Practically all the "change" which based on any observation and we may assign numbers are of these kind. Its value depends on circumstances what we are not able to take into account, thus depends on "chance".

Characterization of discrete random variable

E.g. roll of a pair of (independent) dice with 36 possible outcomes.

Let $\xi = i + k$ be the random variable
$i = 1, 2, 3, 4, 5, 6$ and $k = 1, 2, 3, 4, 5, 6$, thus
ξ may have 11 different values:

The possible values are: $x_j = 2, 3, \ldots, 12$.

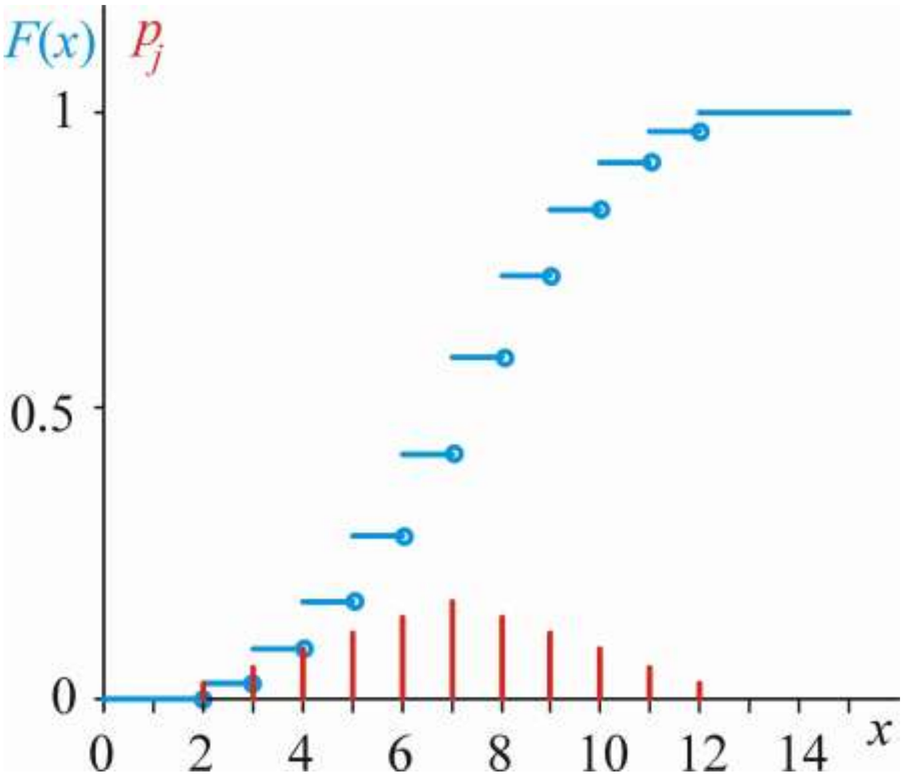
The „result" of the roll is one of the possible values.

Characterization (by):

Distribution function [$F(x)$] and **Probabilities** [p_j]

$$F(x) = p(\xi < x) = \sum_{x_j < x} p(\xi = x_j)$$

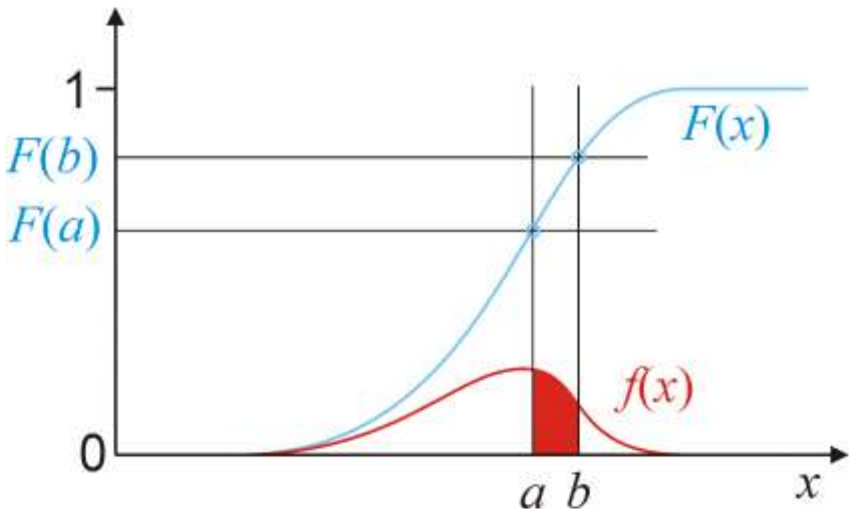
$$p_j = p(\xi = x_j)$$



x_j	p_j
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Characterization of continuous random variable

Cumulative distribution function [$F(x)$] and Probability density function [$f(x)$]



$$F(b) - F(a) =$$
$$= p(a < \xi < b) =$$
$$= \int_a^b f(x)dx =$$
$$= [\text{red area}]$$

Numerical parameters of a **random variable**
or rather its distribution.

Where is the **"middle"** of distribution?

1a. **expected value** [$M(\xi)$]

Discrete case: $$M(\xi) = \sum_i x_i p_i$$

Continuous case: $$M(\xi) = \int_{-\infty}^{\infty} x f(x) dx$$

(roll of a pair of dice)

x_i	p_i	$x_i p_i$
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	20/36
6	5/36	30/36
7	6/36	42/36
8	5/36	40/36
9	4/36	36/36
10	3/36	30/36
11	2/36	22/36
12	1/36	12/36
		252/36 = 7

demonstration: location of center of mass

If we have only a **few data** we can not see the characteristics of **data set**.
Numerical characteristics of quantitative data
(can be determined in every case)

Where is the **"middle"** of data set with *n* elements?

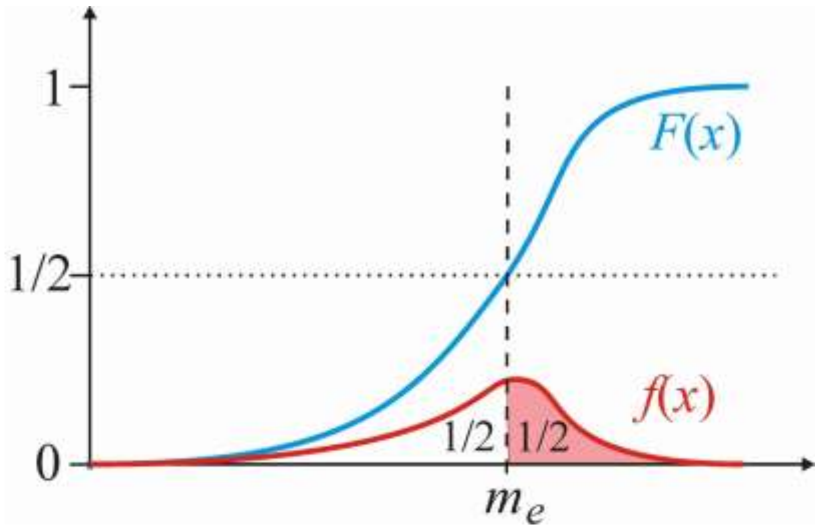
1b. **mean** (arithmetical average)

$$x_{\text{mean}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{j=1}^{m} w_j x_j}{\sum_{j=1}^{m} w_j}$$

It is sensitive to the extreme values!

2a. **median** (m_e)

$F(m_e) = 1/2$



demonstration: quantile of two uniform probability (1/2) mass (weight) or rather area.

2b. **median** (x_{median}) of data set

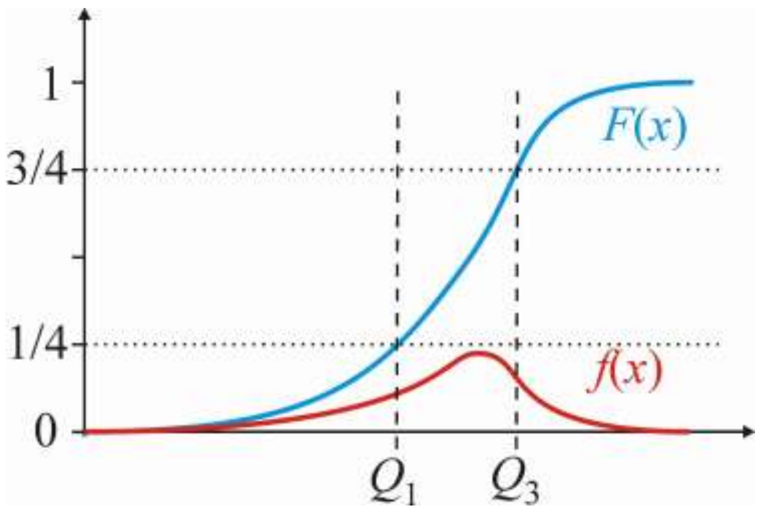
We order the data according to their magnitudes, and look for the middle or middles.

3a. **quantiles**

other ratio of probability or mass (weight), or rather ratio of area (Q_1 lower, Q_3 upper quartile)

$F(Q_1) = 1/4$

$F(Q_3) = 3/4$



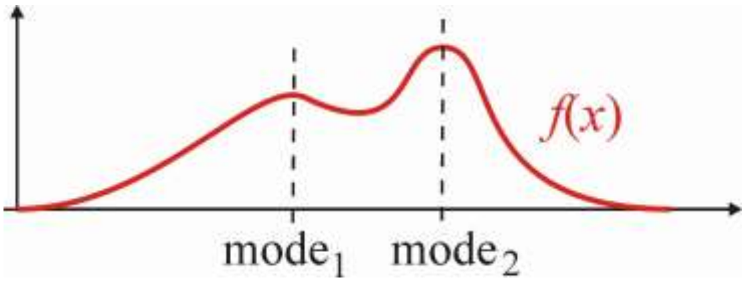
3b. for a set of data; e.g.: What income makes a person become a member of the "upper ten thousand".

First we order the data by magnitude again.

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

4a. **mode(s)**

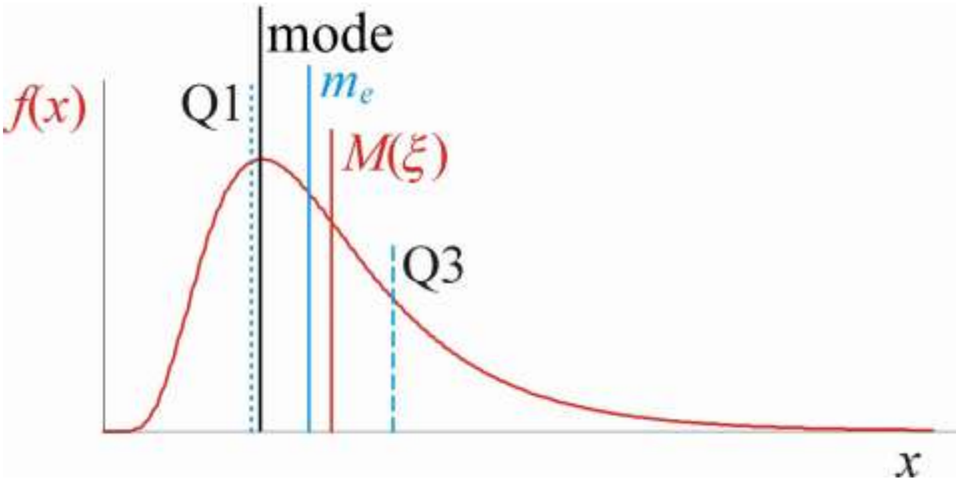
most probable value(s),
local maxima of the probability density function



4b. 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)

They are not sensitive to the extreme values!

Relation of the numerical parameters of the „middle”:



How large is the **spread** of the distribution?

1. **variance**

$$D^2(\xi) = M[(\xi - M(\xi))^2]$$

Characteristics of measures of spread of data set

0. **range**

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

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

   average of the squared deviation of the data from the mean

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

2. 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).

End of lecture

Some properties of expected value

$M(k\xi) = kM(\xi)$

$M(\xi + \eta) = M(\xi) + M(\eta)$

if ξ and η are independent random variables, than
$M(\xi\eta) = M(\xi)M(\eta)$,

Some properties of variance

$D^2(a\xi + b) = a^2D^2(\xi)$

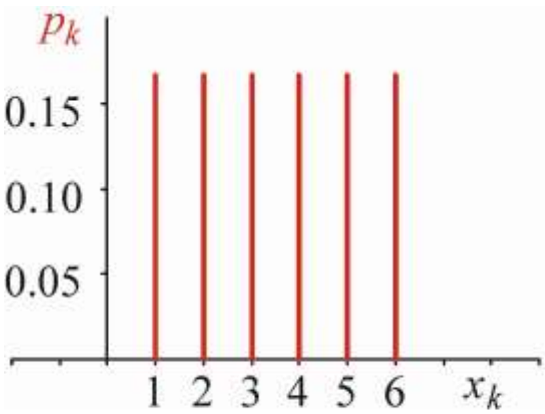
if ξ and η are independent random variables, than
$D^2(\xi + \eta) = D^2(\xi) + D^2(\eta)$

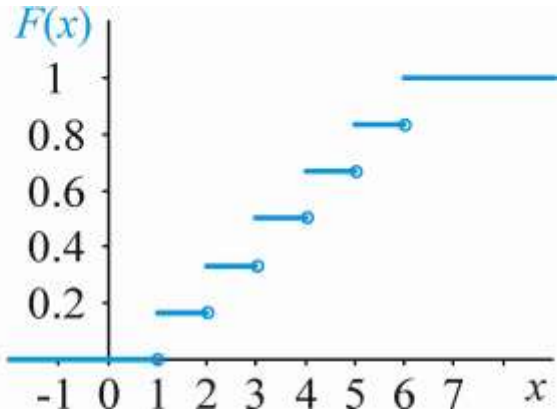
Some remarkable (model) **distributions**

1. Discrete probability distributions

Uniform distribution

In a specific case
Example: dice; probability of an outcome $p = 1/6$.
Possible values: 1, 2, 3, 4, 5, 6.





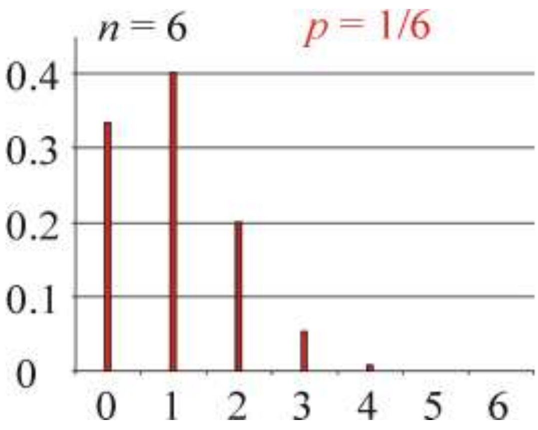
Binomial distribution (Bernoulli-distribution)

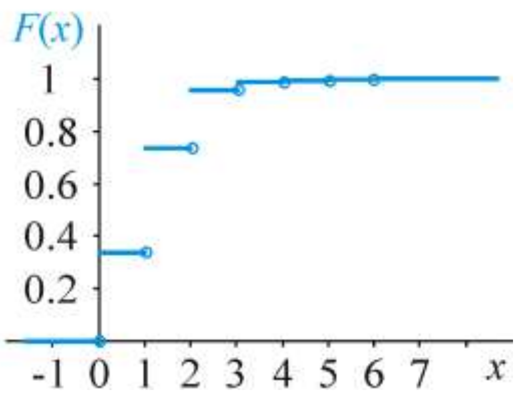
alternative p, $(1–p)$
n trials $P(\xi = k) = B(n, k)$

k	P
0	0.33
1	0.4
2	0.2
3	0.05
4	0.008
5	0.0006
6	0.00002

Example: dice, 6 rolls, $n = 6$ ($p = 1/6$)
What is the probability that we get never ($k = 0$), ones, twice (k-times) a result of 6?

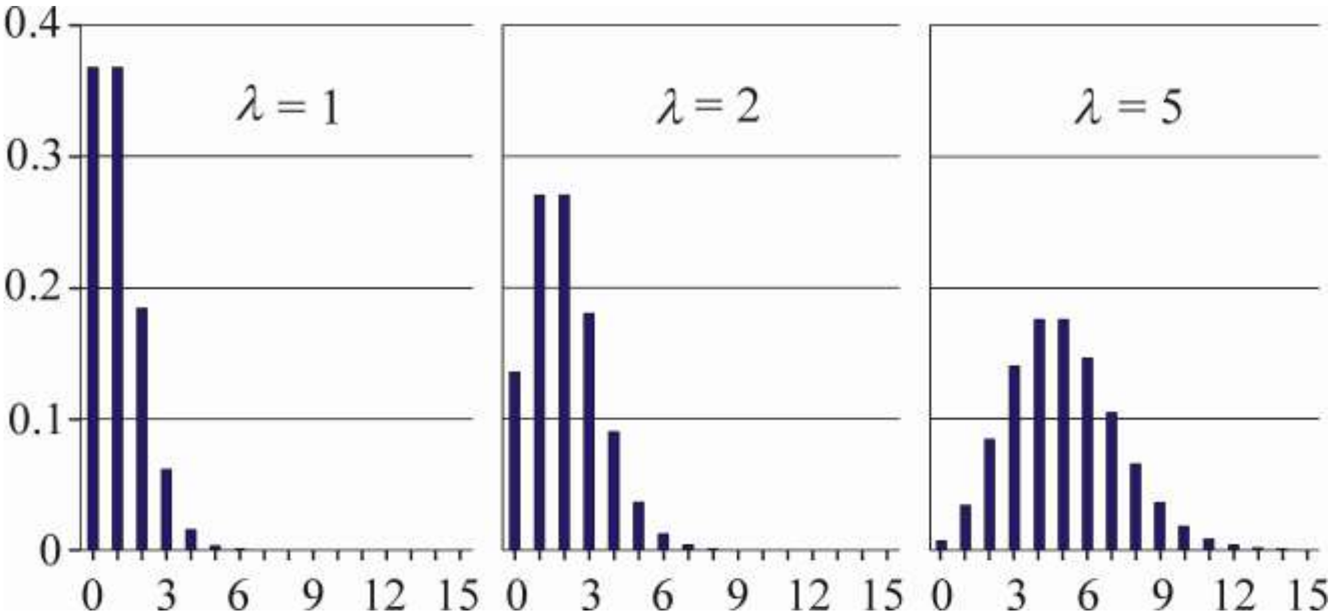
$M(\xi) = np$,
$D^2(\xi) = np(1–p)$





Poisson-distribution

$M(\xi) = \lambda$, $D^2(\xi) = \lambda$

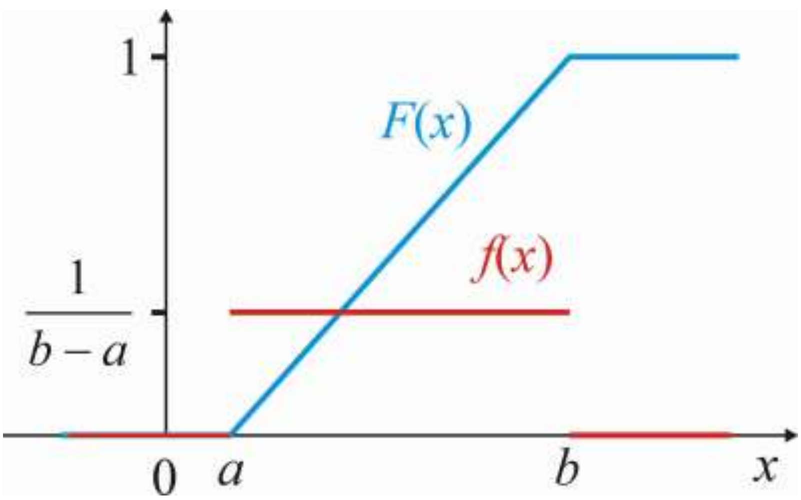


Examples: number of particles in a given volume
number of decayed atoms in a radioactive substance during a given time interval

2. Continuous probability distributions

Uniform distribution

$M(\xi) = (a + b)/2$ $D^2(\xi) = (b - a)^2/12$

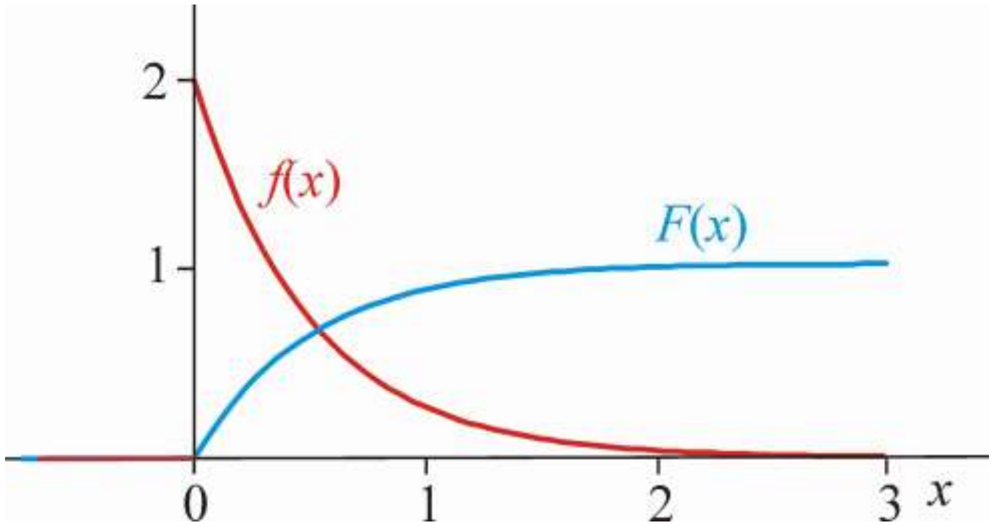


Example: the density or temperature of air in a room

Exponential distribution

$M(\xi) = 1/\lambda$, $(\lambda = 2)$

$D^2(\xi) = 1/\lambda^2$

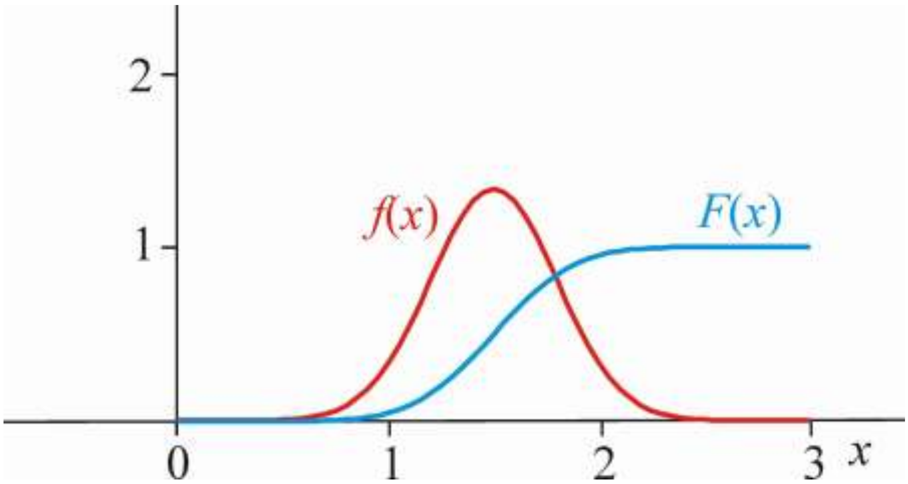


Example: lifetime of the individual atoms in the course of radioactive decay

Normal distribution (Gaussian distribution)

$M(\xi) = \mu$, $N(\mu;\sigma)$

$D^2(\xi) = \sigma^2$ $N(1.5;0.3)$



Examples:

The height of men in Hungary, given in cm: $N(171;7)$

Diastolic blood pressure of schoolboys, given in Hgmm: $N(58;8)$

Standard normal distribution

$M(\xi) = 0$
$D^2(\xi) = 1$

Transformation: x [N(μ;σ)] $\rightarrow z$ [N(0;1)] $$z = \frac{x-\mu}{\sigma}$$

Both the χ^2-distribution and the t-distribution are results of the transformations of variables having standard normal distribution (ξ_n).

Why the normal distribution is a favoured one?

Central limit theorem

If a random variable is a result of a sum of several small independent changes, than it should be a random variable having normal distribution with a good approximation.

You may try it!