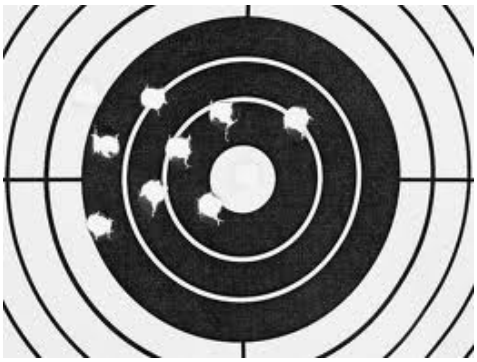
Statistical laws

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.

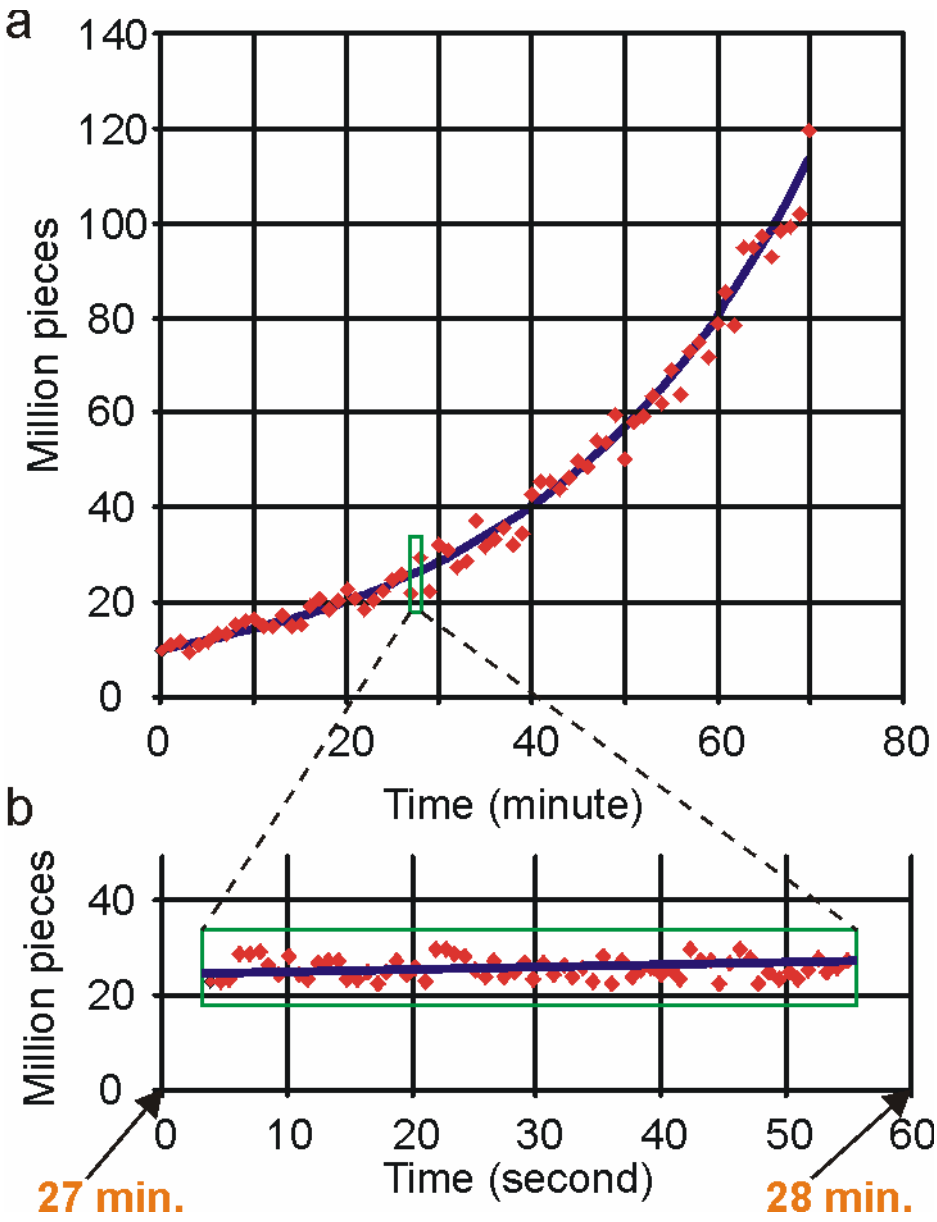
	examples			
Phenomenon	medical examination	toss of a coin (1)	waiting for a tram	toss of a coin (2)
Observation	color of skin	falling time of the coin	how many passengers	which side
Event	yellow	between 0.5 s and 1.5 s	10 passengers	head

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

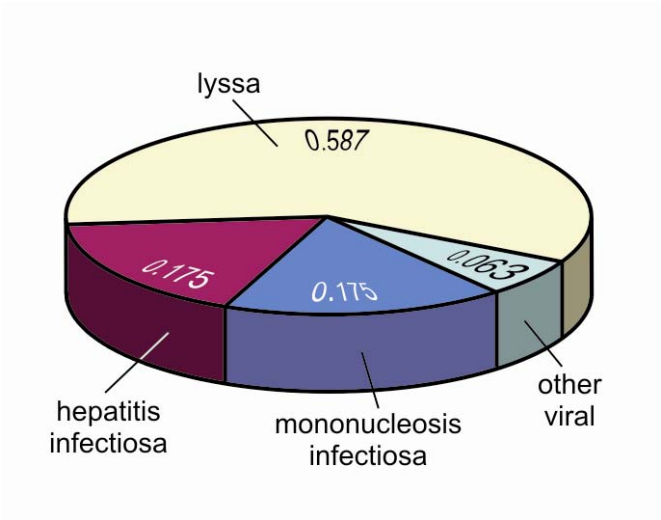
INFECTION	DISEASE	Absolute frequency		Relative frequency		Conditional relative frequency	
bacterial	Salmonellosis (Food poisoning by Salmonella)	94	208	0.280	0.619	0.452	1.000
	Scarlatina (Scarlet fever)	102		0.304		0.490	
	Other bacterial	12		0.036		0.058	
viral	Hepatitis infectiosa (Hepatitis)	22	126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

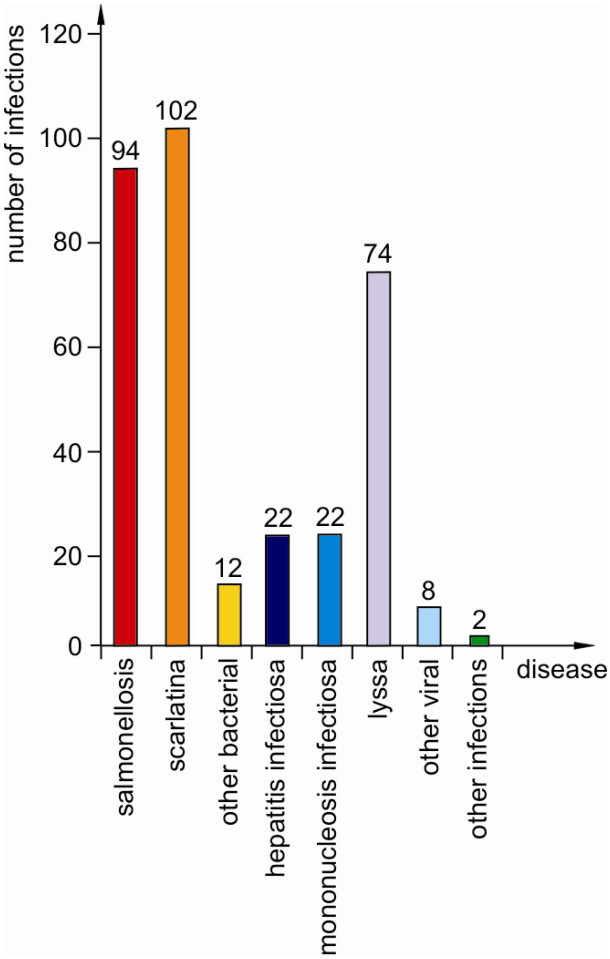
relative frequency: absolute frequency divided by the total number of elements in the set in question
It is a ratio, therefore, when speaking about relative frequency, both the category and the set we relate it to must be specified.

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

2/2. **Diagrams**

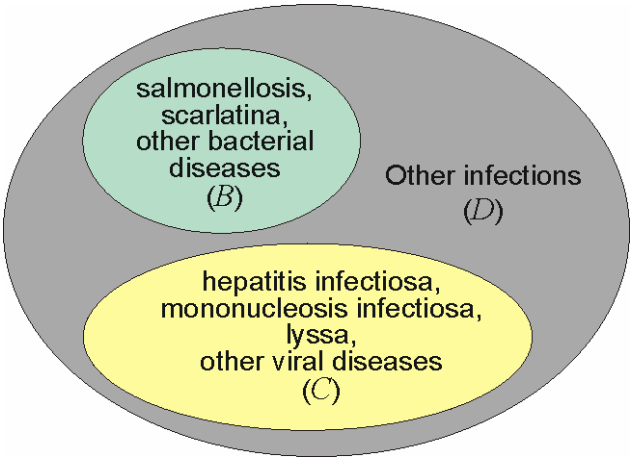


pie diagram



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.

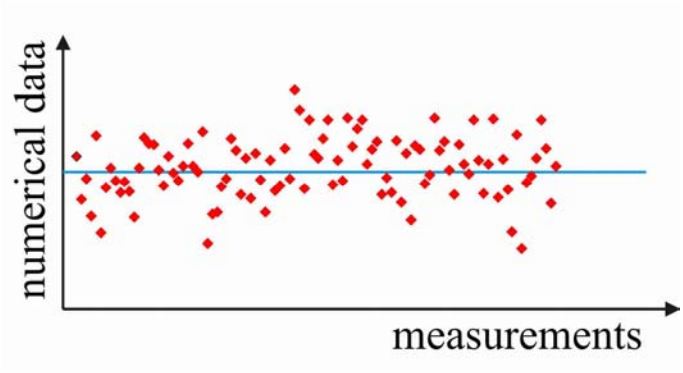
Relative frequencies are only additive within the given set.

(II)

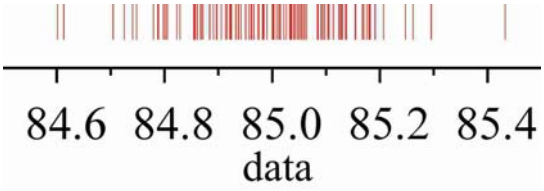
Conditional relative frequency and relative frequency (without condition) are multiplicative according to the method shown above.

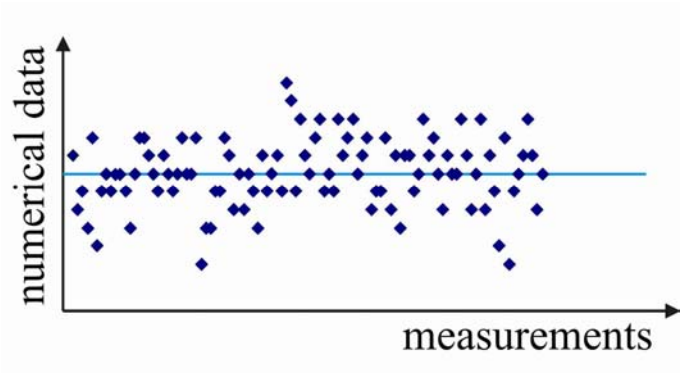
Characteristics of quantitative data

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

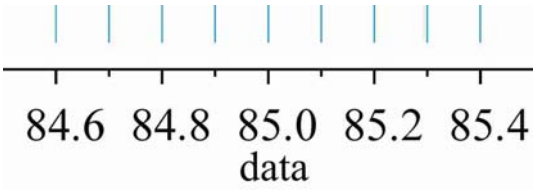


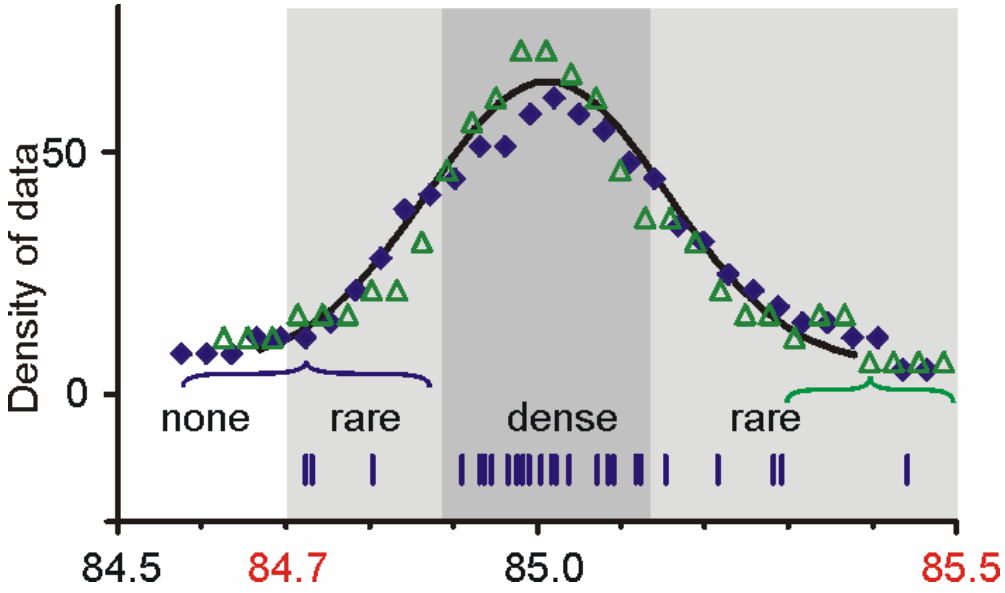
„continuous" case
"never" two identical





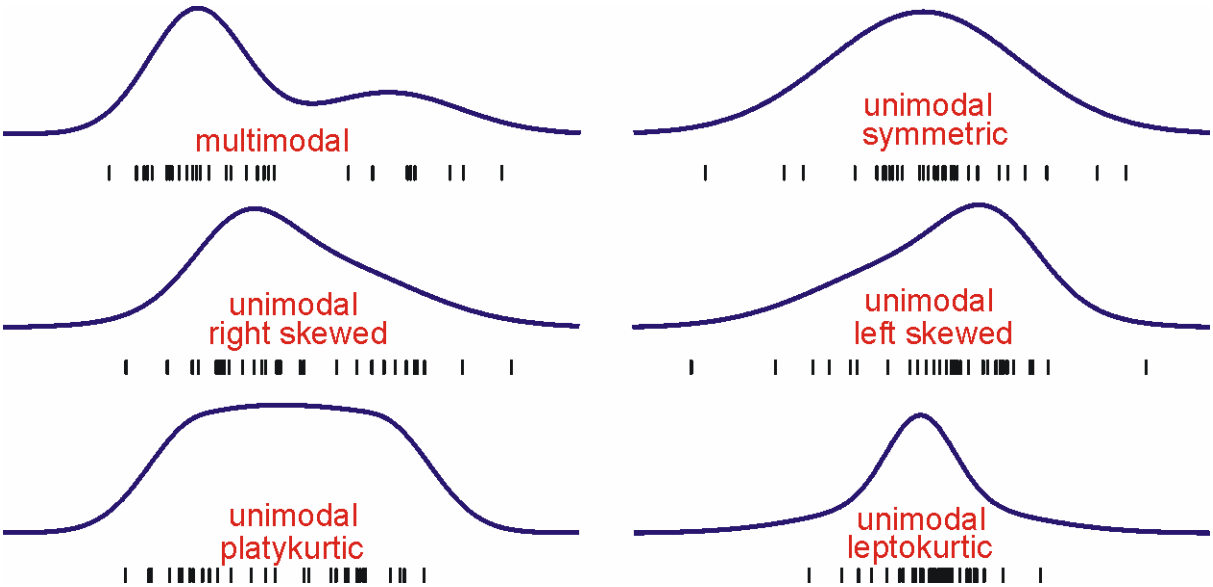
discrete case
We have to give the frequencies.





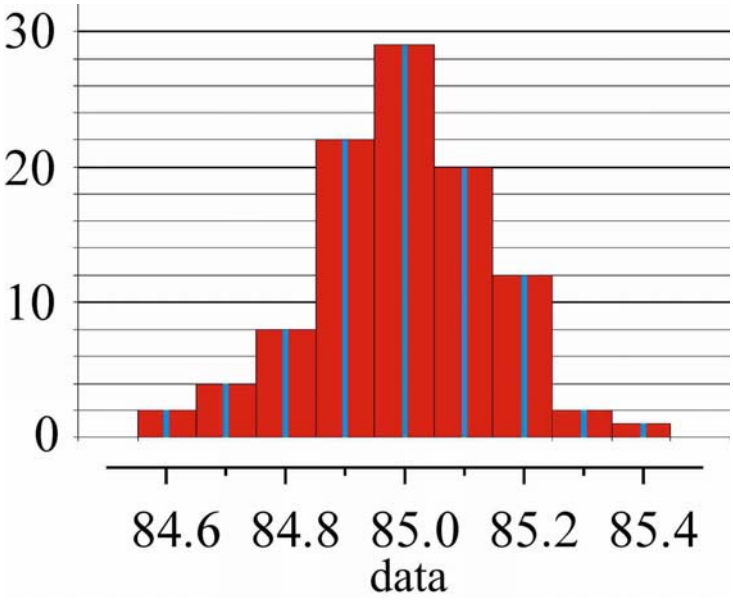
How can we characterize density of data?

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).



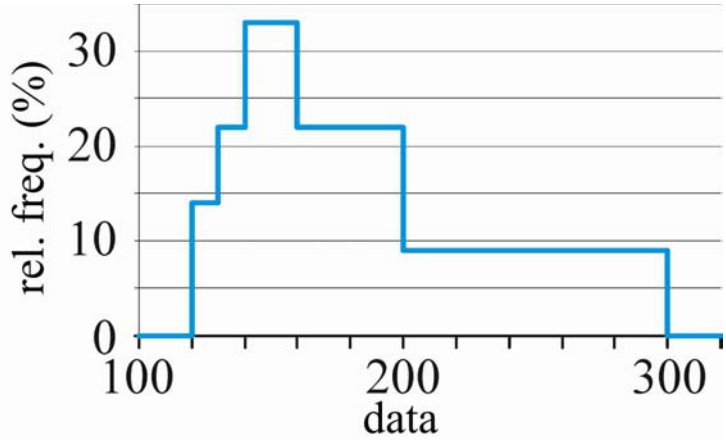
For better comparison between data sets with different bins usually the relative frequencies are given.

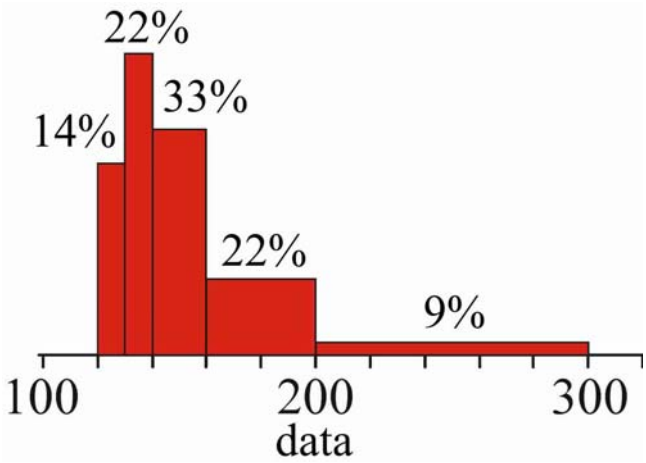
How can we characterize it if we do not know all data?

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.

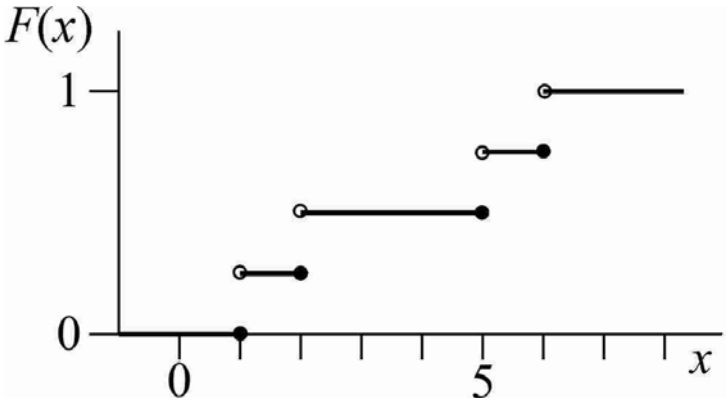
If the width of the columns is the same (equal classes) the two representations are identical.

Distribution function ($F(x)$)

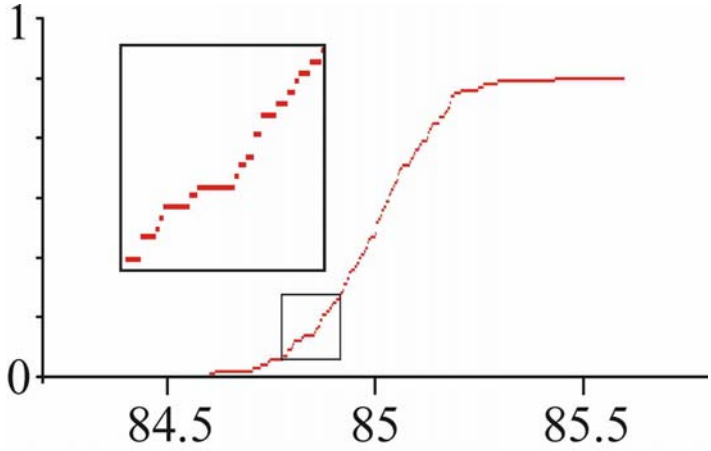
For a data set $[x_1, x_2, x_3, \ldots, x_n]$:

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

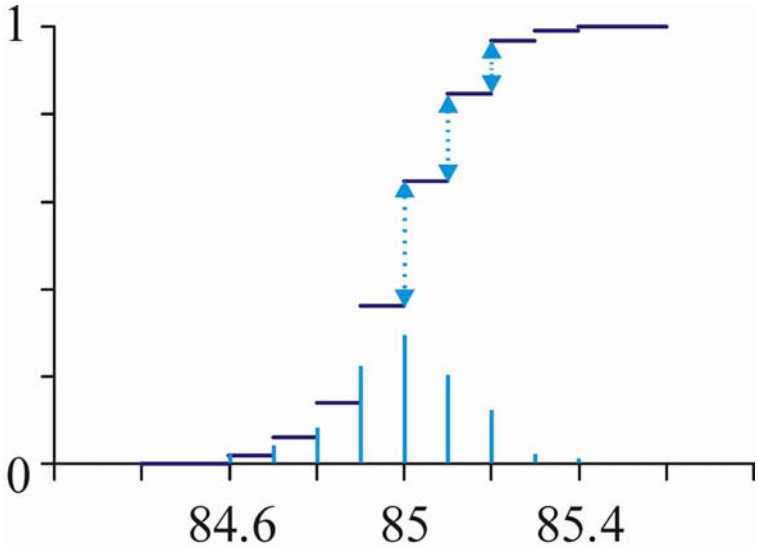
E.g. (1) [1, 2, 5, 6]



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.