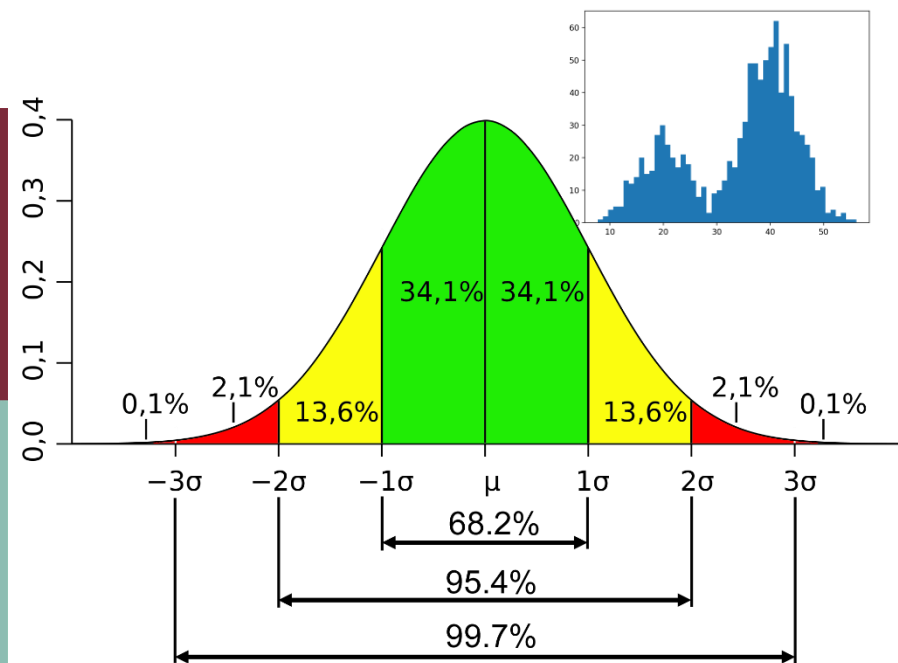
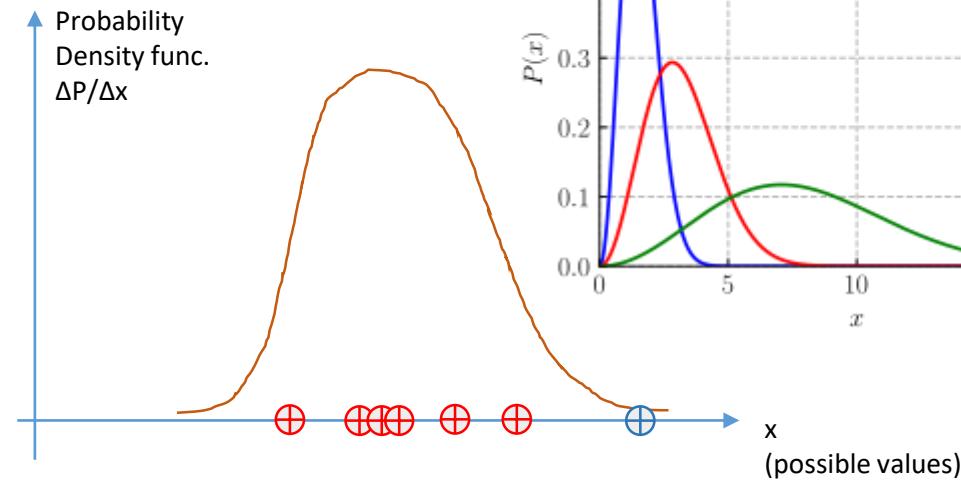
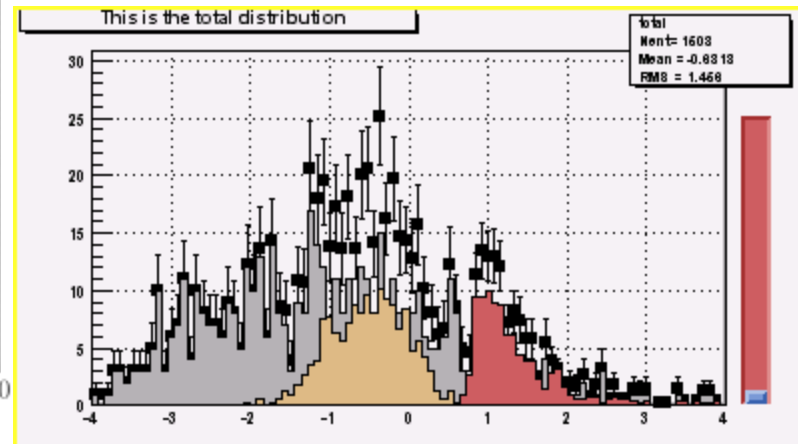
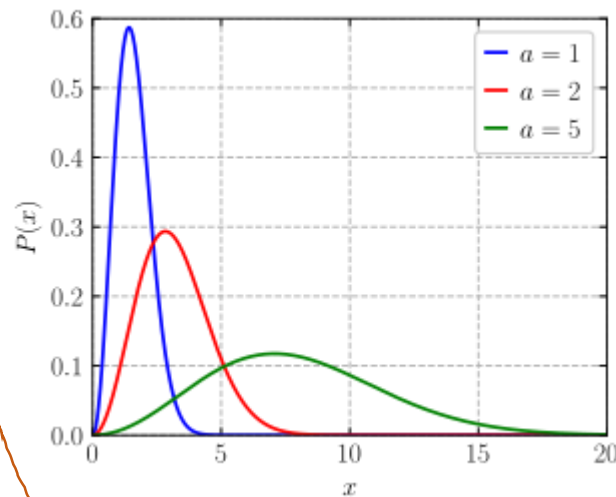
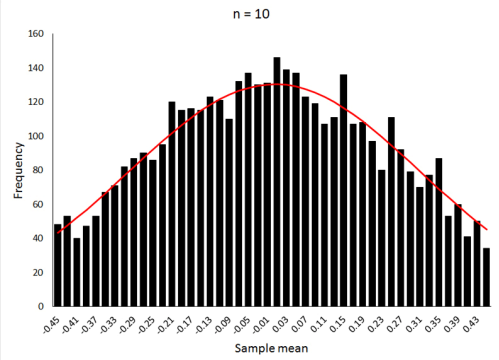
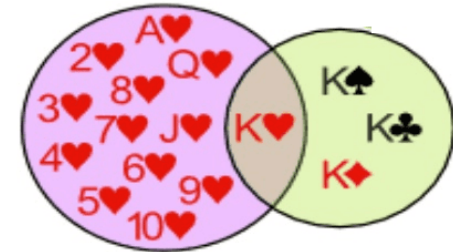


# Event, probability, distribution.

G. Schay



### Event:

In a random (real) situation or setting (the “experiment”) one particular outcome (which is of interest). The event is always *observable*.



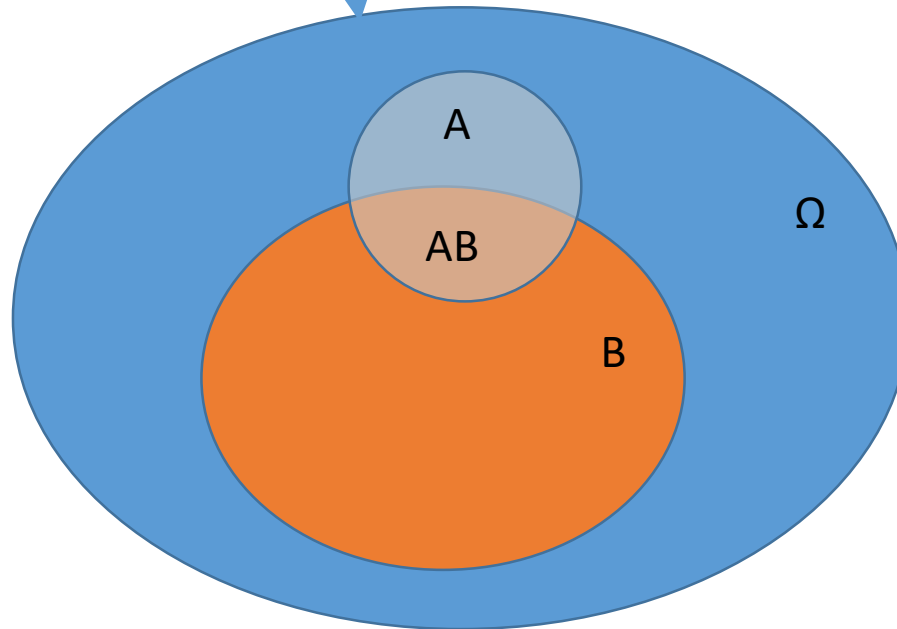
It is unambiguously possible to tell if an event **happened**, or **did not happen**.



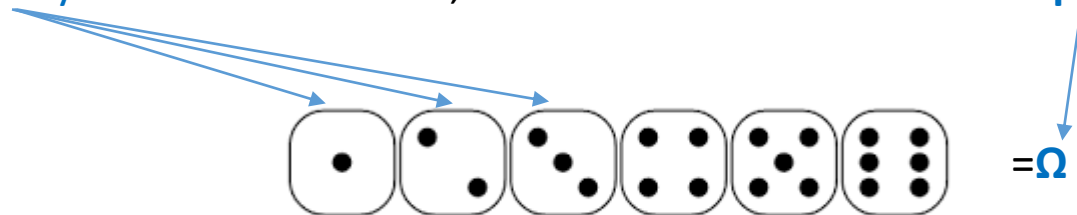
© Dreamworks: The Road to El Dorado – Zaragoza's dice

Event space and embedded events in it.

We can have elementary or composite events.



If we take every elementary event into account, then we know the full event space.



Every event has a particular **frequency**

We conduct N identical experiments, and count the number of times an event happens.  
e.g.

- > in a game we throw a dice 5 times, and count the number of “6”-s.
- > how many times we have to get up during the hospital night shift.
- > how many emergency cesarean cuts happen in one month.

*The number of times a particular event happens is termed as the (absolute) frequency of the event*

Usually it is denoted by k, but also with n, etc.

e.g.  $k_A$  is the frequency of the event A during the observation period.

### **Relative frequency**

We calculate the *ratio* of the frequency and total number of experiments

So relative frequency of A =  $k_A/N$  (it can be given in %)

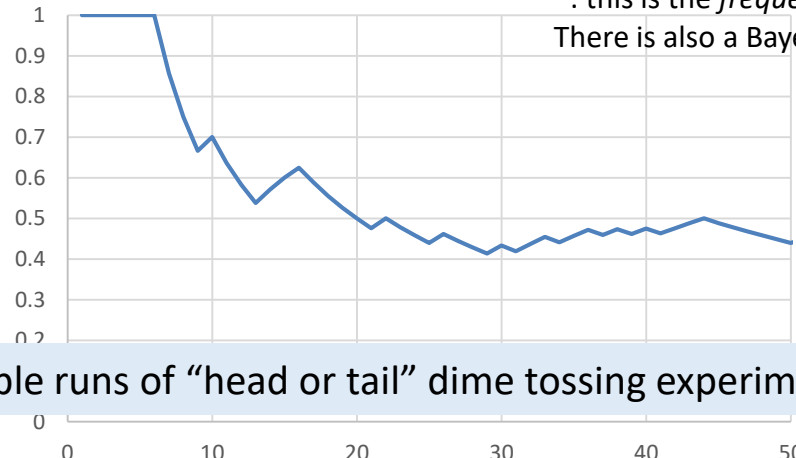
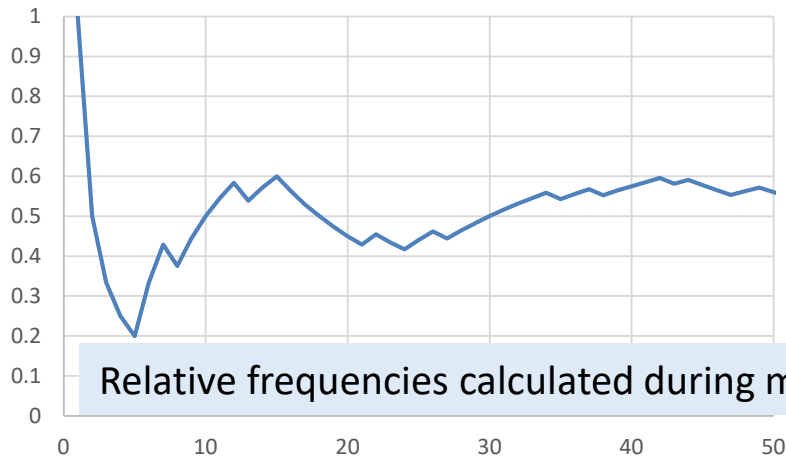
## The law of very large numbers

The relative frequency is only *approximatel the same* if we compare multiple runs of identical experiments.

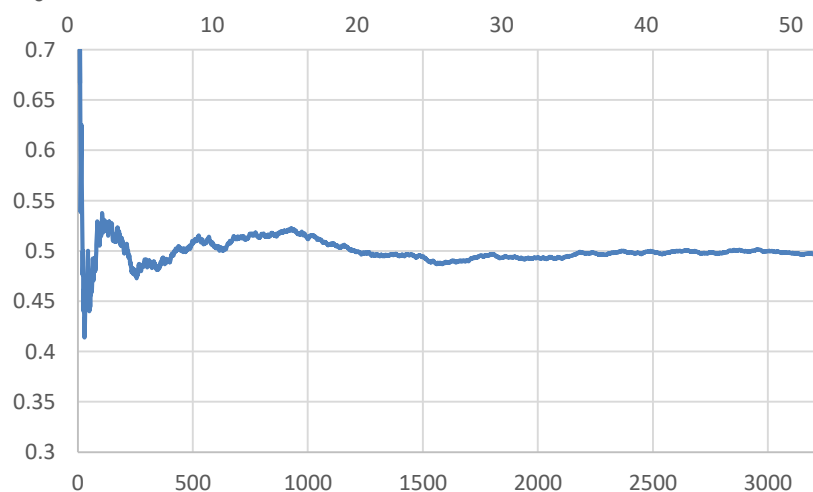
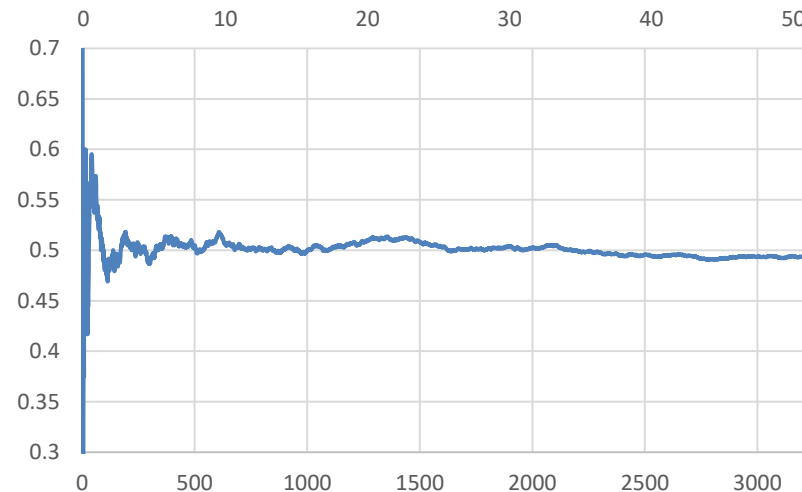
But it gets *stabilized IF we have a LOT of experiments.*

**The number around which the realitve frequency stabilizes\* is called the Probability (P)**

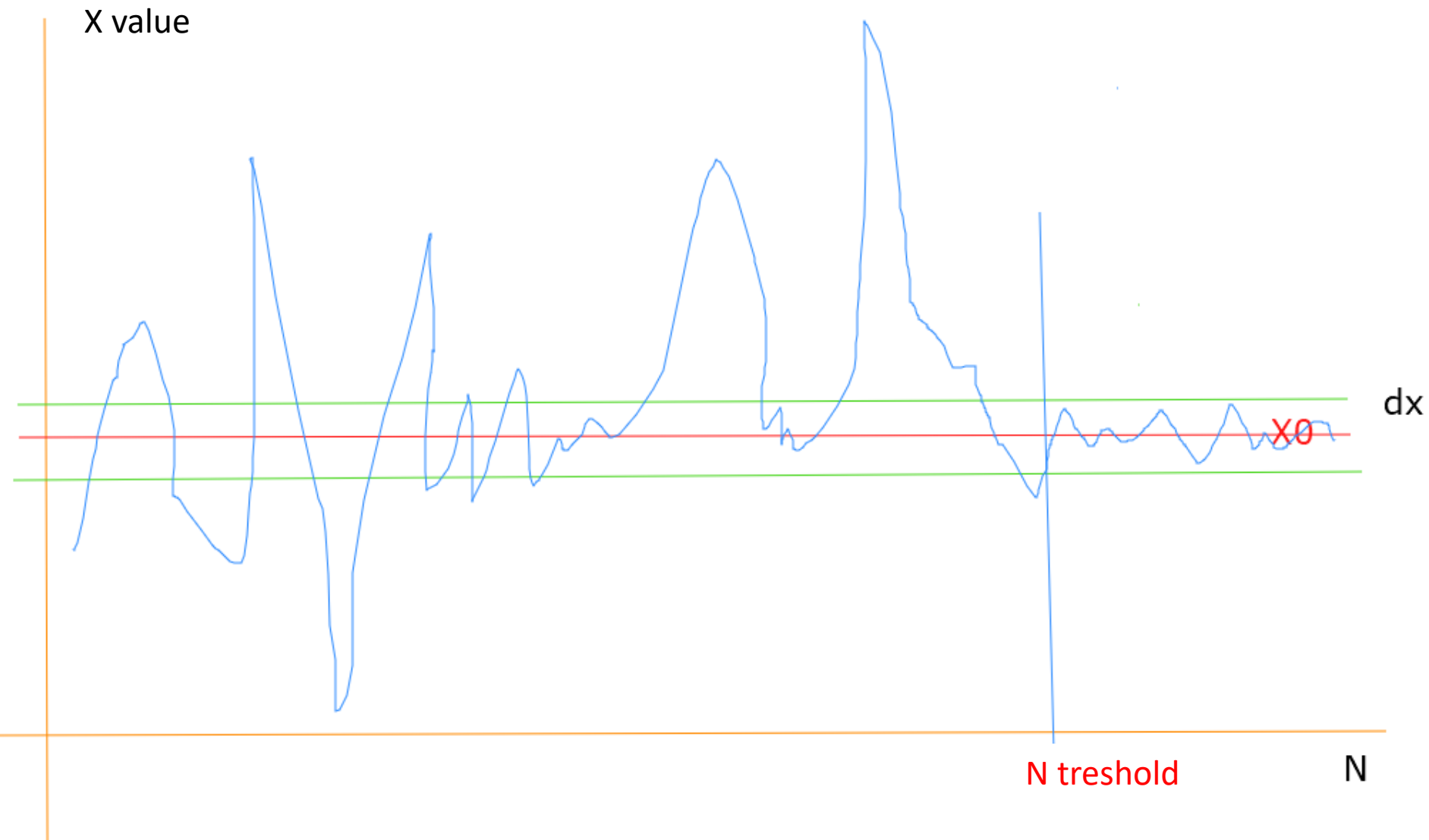
\*: this is the *frequentist* definition.  
There is also a Bayesian definition



Relative frequencies calculated during multiple runs of “head or tail” dime tossing experiments

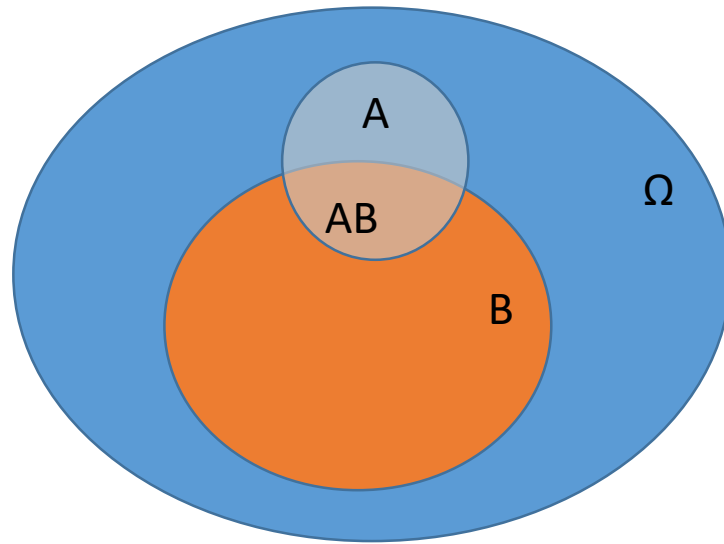


**Extension material:** There exists a so called “limit” of  $x$  which is  $X_0$ , IF and only if there exists a threshold of  $N$ , the number of experiments, such that if  $N > N_{\text{threshold}}$  then  $|X - X_0| < dx$ . With other words, for ANY small  $dx$  (but not exactly 0) it is possible to give an  $N$  threshold so that if  $N$  is larger than this then the deviation of  $X$  from  $X_0$  is  $<$  than  $dx$ .



# Relation of events

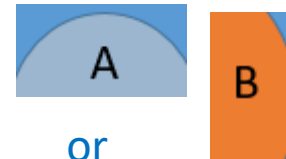
Event is something in which we are interested. (we can define it freely)



There are compound events

- Eg. we have thrown an even number on the dice
- We get some deviation on the blood report

Logical relations between events:



either one happens



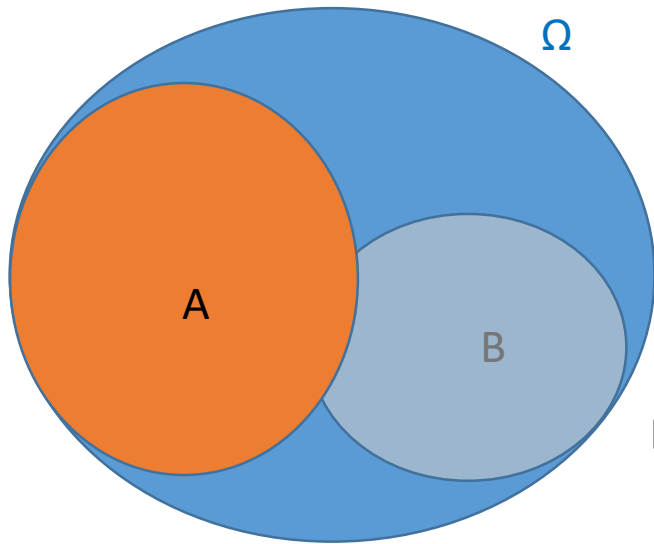
They happen together

**Mutually exclusive events:** ONLY one of the set of events can happen at once.

Eg. Even or Odd. It can't be both at once.

**Opposite event:** Not-A : any event but A has happened.

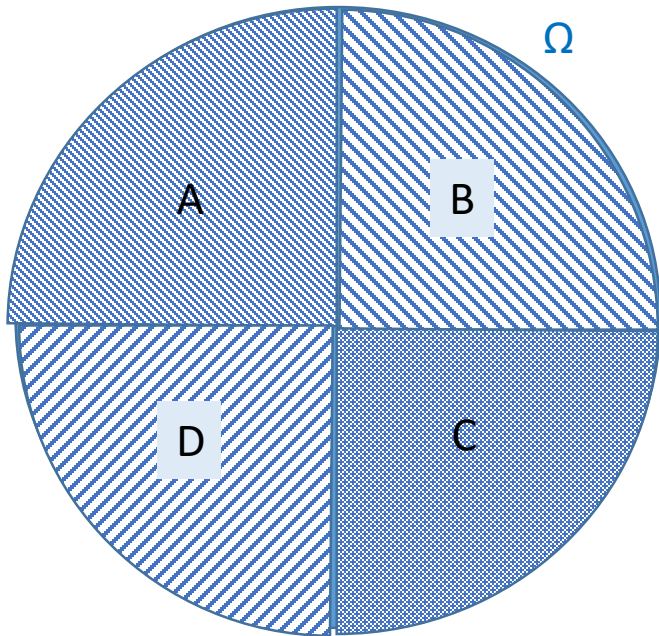
e.g. „the patient does not have fever”



## Mutually exclusive events

either-or

For these events  $P(A \text{ or } B) = P(A) + P(B)$



## Elementary events:

Such events which can not (or is not needed) to decompose into “smaller” events.

It is extremely important to cover the whole  $\Omega$  event space with elementary events!



conventional notations:

AND  $P(A \text{ és } B) = P(A * B) = P(A \cap B) = P(A \cdot B) = P(AB)$

OR  $P(A \text{ vagy } B) = P(A + B) = P(A \cup B) = P(A \vee B)$

NOT  $P(\text{nem } A) = P(!A) = P(\sim A) = P(\bar{A})$

**Kolmogorov-axioms:**

$$0 \leq P(A) \leq 1$$

$$P(A \cdot \bar{A}) = P(\emptyset) = 0 \quad \text{impossible event}$$

$$P(A + \bar{A}) = P(\Omega) = 1 \quad \text{sure event}$$

$\emptyset$  : emptiness (no event, “nothing”)

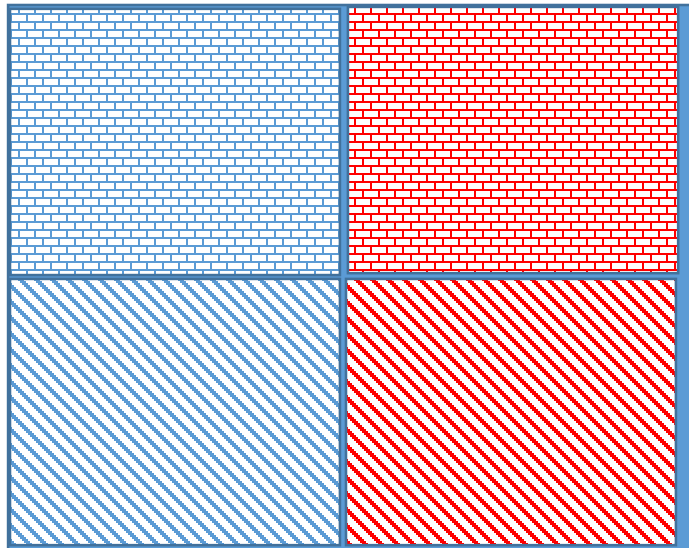
$\Omega$  : full event space or „anything”

## Stochastic independence

Two events are independent, if they mutually do not have any influence on the probabilities. (so either event B happens or not, the probability of A stays the same)

e.g. the first patient is coughing, the second has skin irritation, the third has ... etc.

Or, we throw a dice 2 times, the result of the first one does not impact the result of the next (the dice does not have a memory)



$$P(\text{grid}) = 2/4 = 0.5$$

$$P(\text{stripes}) = 2/4 = 0.5$$

$$P(\text{blue}) = 2/4 = 0.5$$

$$P(\text{red}) = 2/4 = 0.5$$

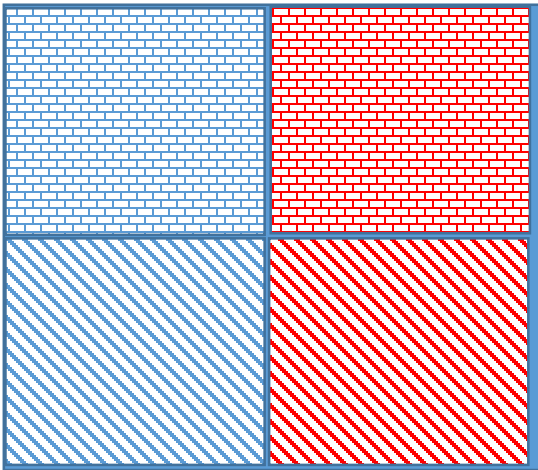
$$P(\text{stripes AND red}) = 1/4 = 1/2 * 1/2 = P(\text{red}) * P(\text{stripes})$$

**Generally, two events (A and B) are independent *only if*  $P(A \text{ AND } B) = P(A) * P(B)$  holds, and vica versa, *if* it holds *then* A and B are independent.**

## Conditional probability

$P(A \mid B)$  = the probability of A **given that** condition B has occurred / is true.

e.g. :      the patient has fever, *given that* she/he is COVID-19 infected.  
              I get the grade 5 in statistics *given that* I have attended every lecture.



*We are interested only in a subset of  $\Omega$* , and we need the relative frequency in that subset only.

$P(\text{blue} \mid \text{stripes})$  = the number of blues among the ones with stripes = 1 blue AND stripes / 2 stripes =  $\frac{1}{2}$

remarks:

1. For independent A and B we have  $P(A \mid B) = P(A)$
2. For any two events  $P(AB) = P(A \mid B) * P(B)$  Bayes-theorem, or product law

-> see more in the lecture on Bayesian methods!

examples:

- Subset of event space: assume the probability of having diabetes (B) to be 15%.  
The probability of Type II. típusú (A) let be 10%.

Then the probability of Type II. IF we know that the patient HAS diabetes is

$$P(A|B) = P(AB)/P(B) = P(A)/P(B) = 10\%/15\% = 66.7\%$$

Remark: if A is a subset of B, then in the case of A happening B is “automatically” fulfilled, thus  $P(AB)=P(A)$ .

- Independence: What is the probability of the next patient wearing glasses (A), IF the previous patient was a woman (B)?

using  $P(A|B) = P(AB)/P(B)$ , and applying the independence rule =  $P(A)*P(B) / P(B) = P(A)$

Remark: we assume that if A and B are independent, then B can not influence the probability of A, thus it should not even appear in the final result. This is indeed what we get 😊

## Extension material

From the probability it is possible to calculate further measures:

**Odds:**  $O = \frac{P}{1-P}$

This tells how many times more is the propensity of happening than not happening

It is useful if the event space is dichotomic, e.g.: the COVID-19 test comes out positive **or not**.

**Risk:** we have two – assumed to be related - events: one is called the *risk factor (R)*.

We relate the two odds of the event: one with and one without the R happening.

B      R

$P(\text{Disease} | \text{Risk}) =$  „a betegség kockázata ha van rizikófaktor jelen” =  $P(D^*R)/P(R)$

$P(\text{Disease} | \text{no Risk}) = P(D^*\bar{R})/P(\bar{R})$

$$\text{Relative Risk}_{(RR)} = \frac{P(D | R) = P(D^*R)/P(R)}{P(D | \bar{R}) = P(D^*\bar{R})/P(\bar{R})}$$

**Odds Ratio:** Az esélyek hányadosa rizikófaktor mellett és anélkül.

(OR)

$$OR = \frac{O_{D|R}}{O_{D|\bar{R}}}$$

This needs the knowledge of 4 conditional probabilities

## Important, useful distributions

Distribution: if we have multiple possible events, then the distribution gives the probability (mass function) or probability density, i.e. the probability for every event.

For numerous cases the  $P(x)$  function can be defined with a closed formula. For continuous numeric variables we can give the integral or cumulative function  $P(\xi < x)$ -et (here the greek  $\xi$  denotes the random variable, and  $x$  is a particular value)

These known functions can be used for many common problems, or we can transform the original variables to follow one of these well-known distributions.

The most important ones in medicine: Do not memorize the list!

- normal or Gaussian distribution

- Student's t-distribution

- even distribution

- exponential

- binomial

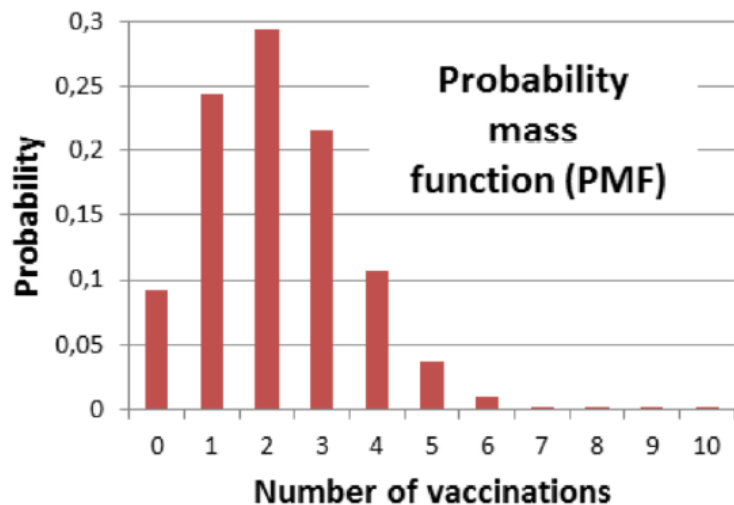
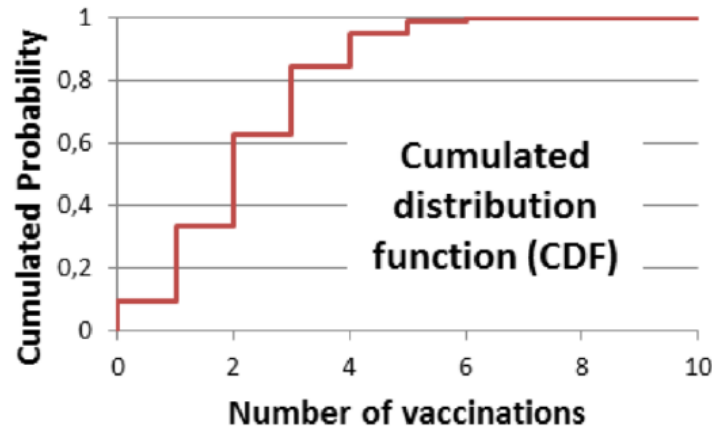
- $\chi^2$

- geometric

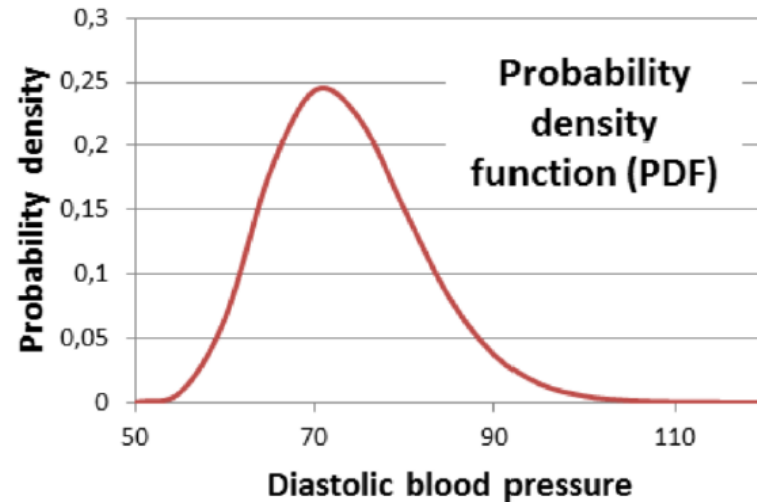
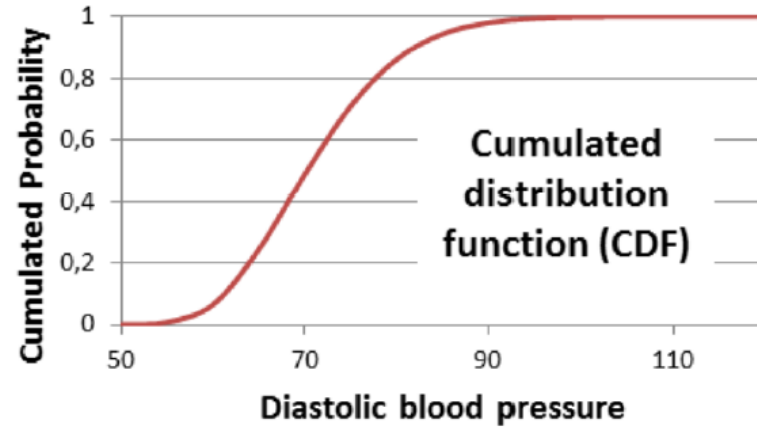
- log-normal

## Ways to graph a distribution

### Discrete



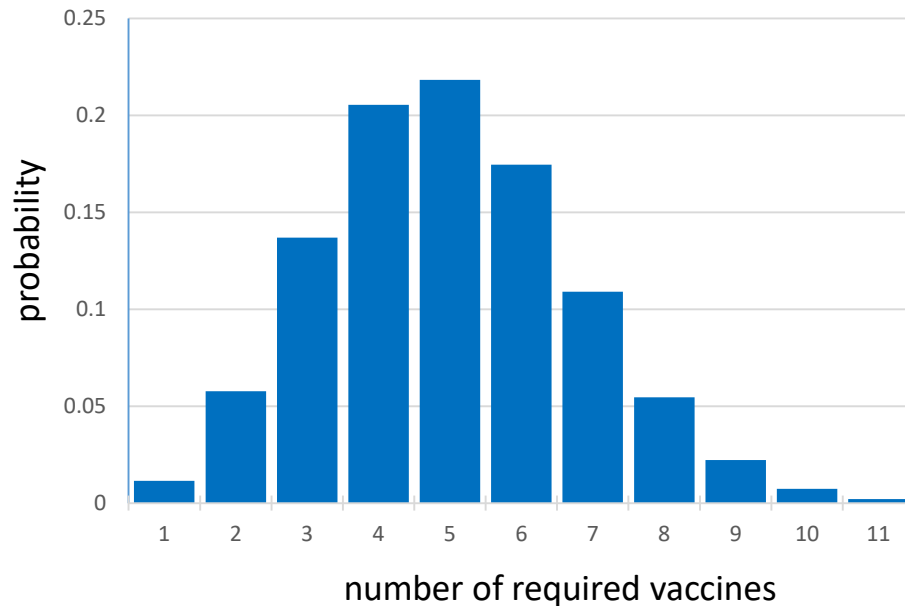
### Continuous



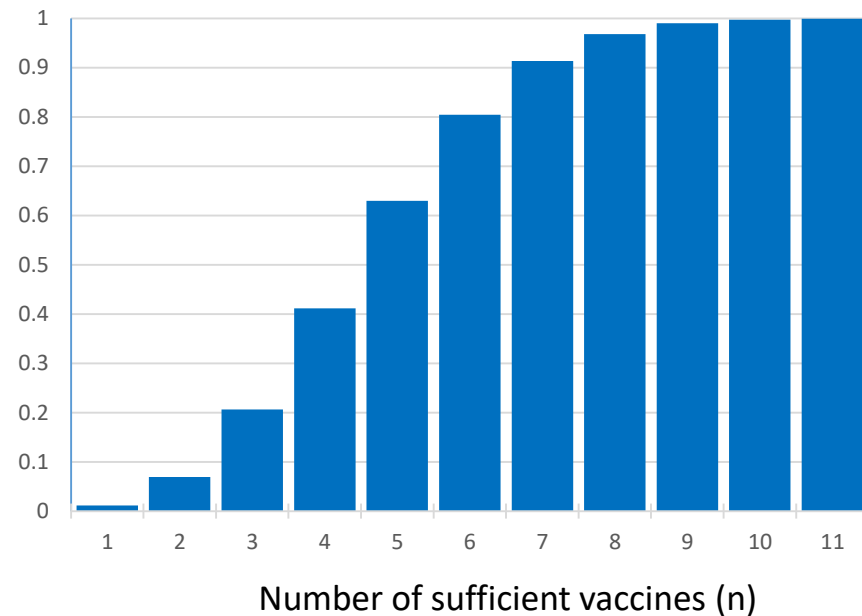
We can choose based on the kind of variable we have (discrete: only certain values are possible, or continuous)

## Binomial (Bernoulli) distribution

It gives that out of  $N$  experiments with what probability does  $A$  happen  $k$ -times.  
Thus we get  $P(A,k,N)$  for every  $k \leq N$  if  $P(A)$  is known.



$P(x)$   
(mass function)

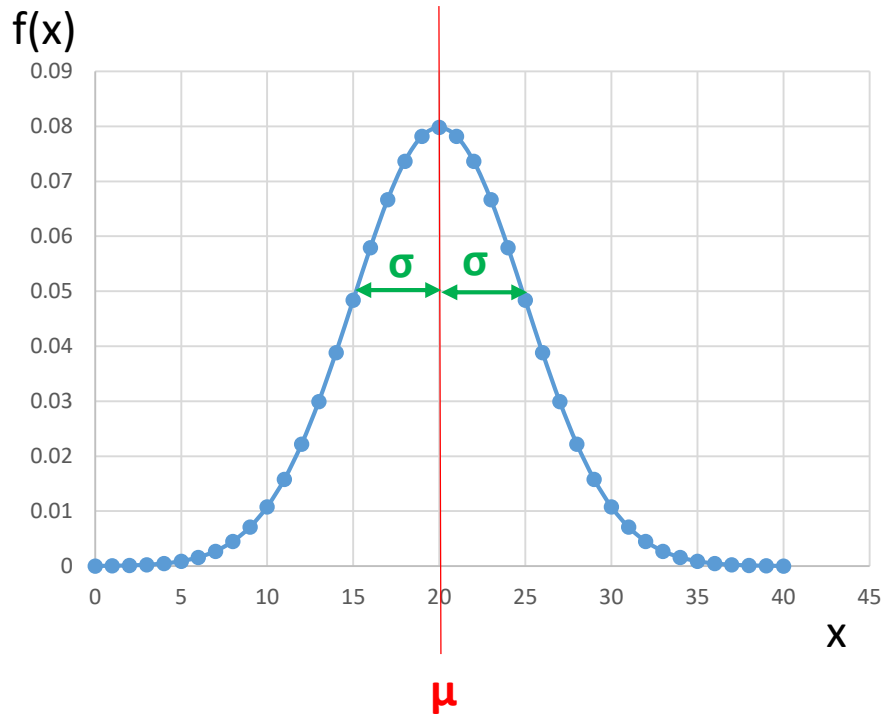


$P(x \leq n)$  function  
(cumulative)

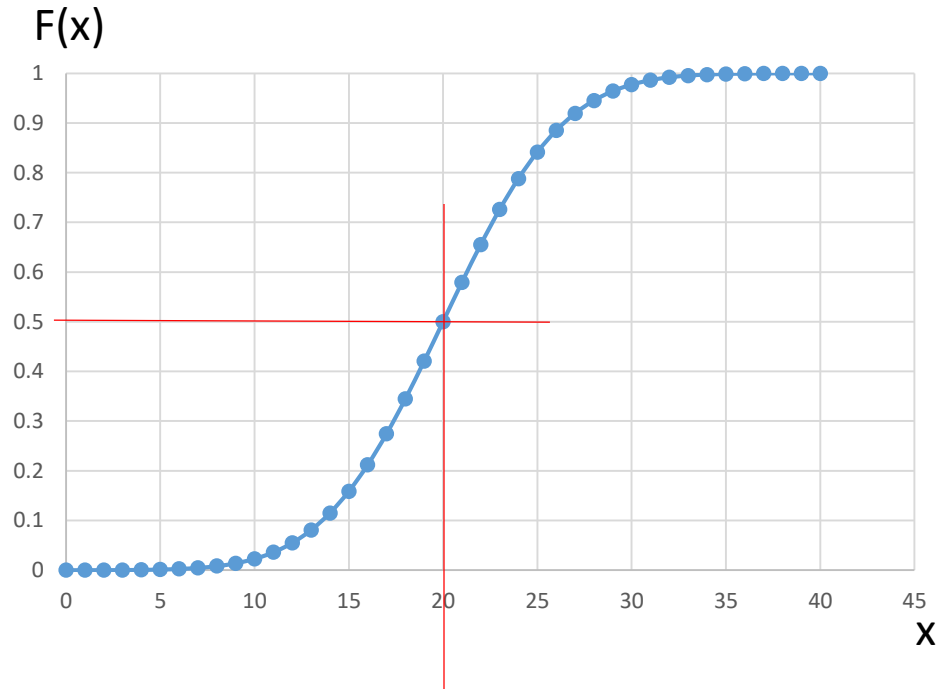
*example:* If during the last month out of 3000 cases we had 12 requiring acute operations, and now I have the night shift with an estimated number of 20 patients, then what can I expect in terms of acute operations?



## Normal (Gaussian)



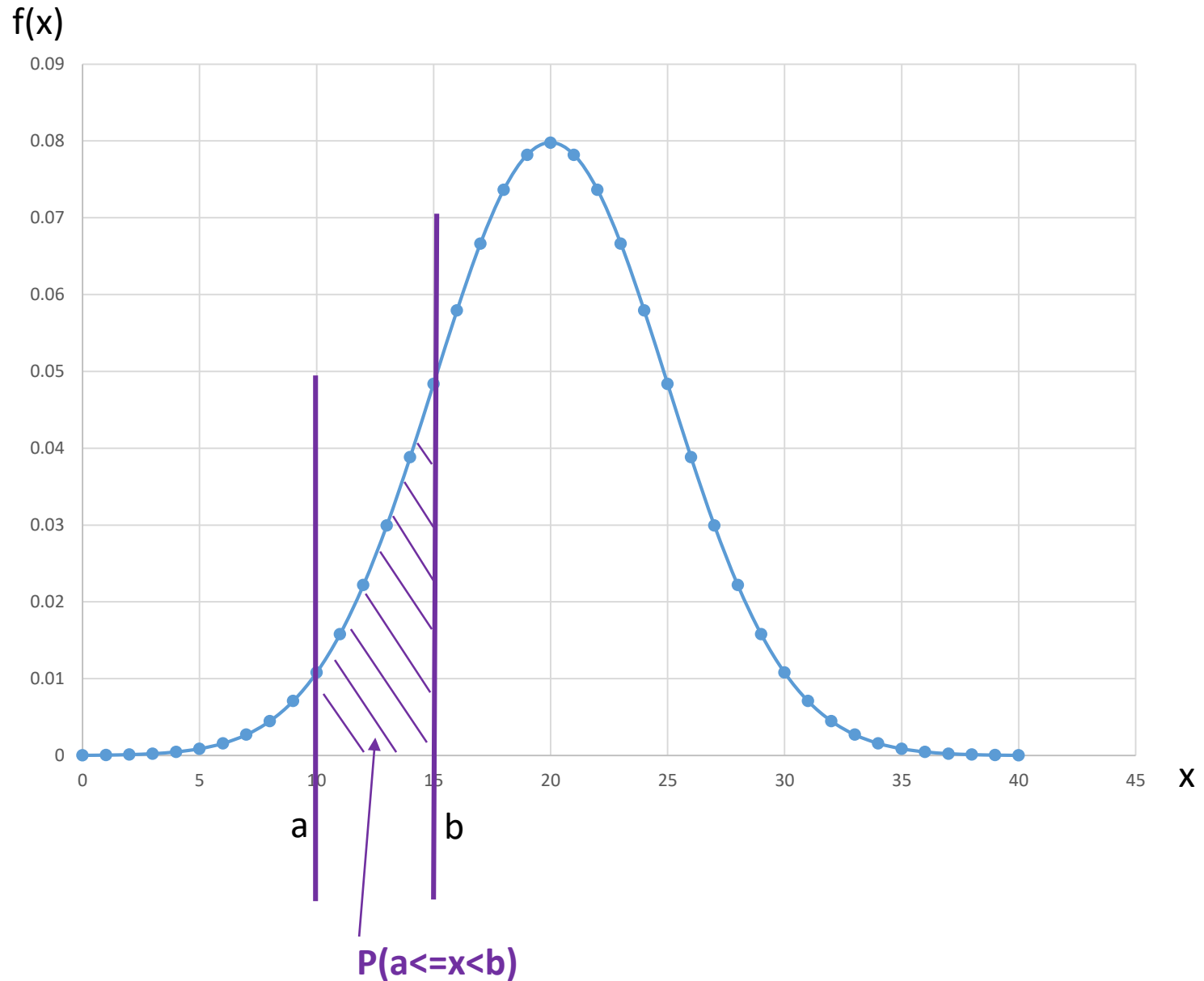
density function



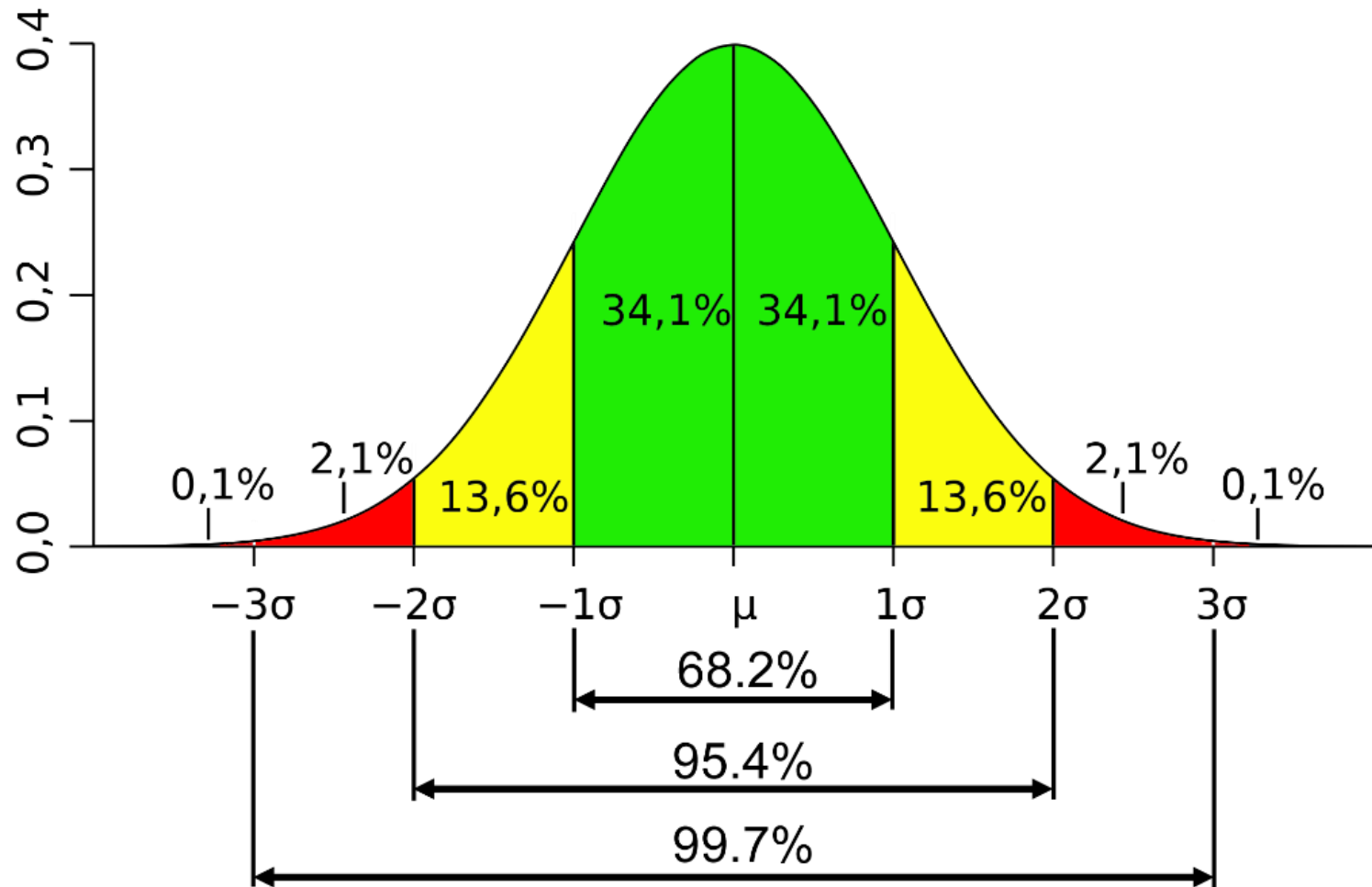
(cumulative) distribution function

For continuous numeric variable, it has two parameters: the center is the expected value  $\mu$ , and the width is the standard deviation (stdev)  $\sigma$ .

**Warning! The density function does NOT yield the probability, but the area under the curve does.**

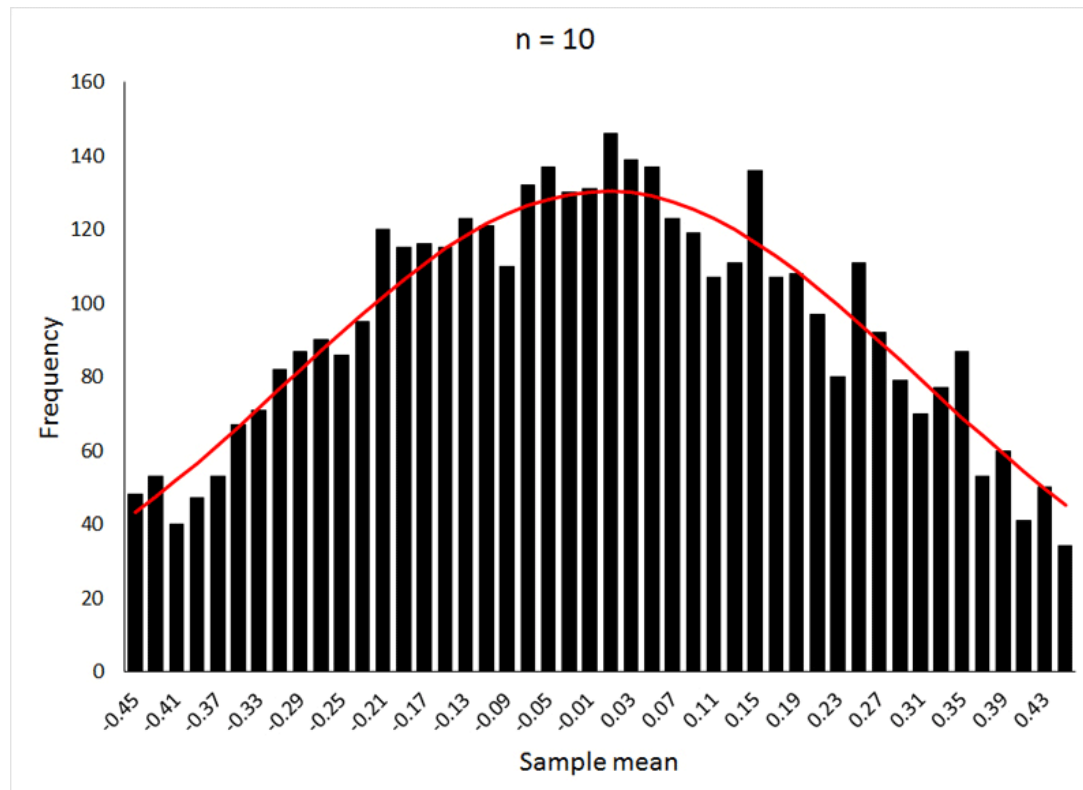


## Principal areas, ranges under the curve



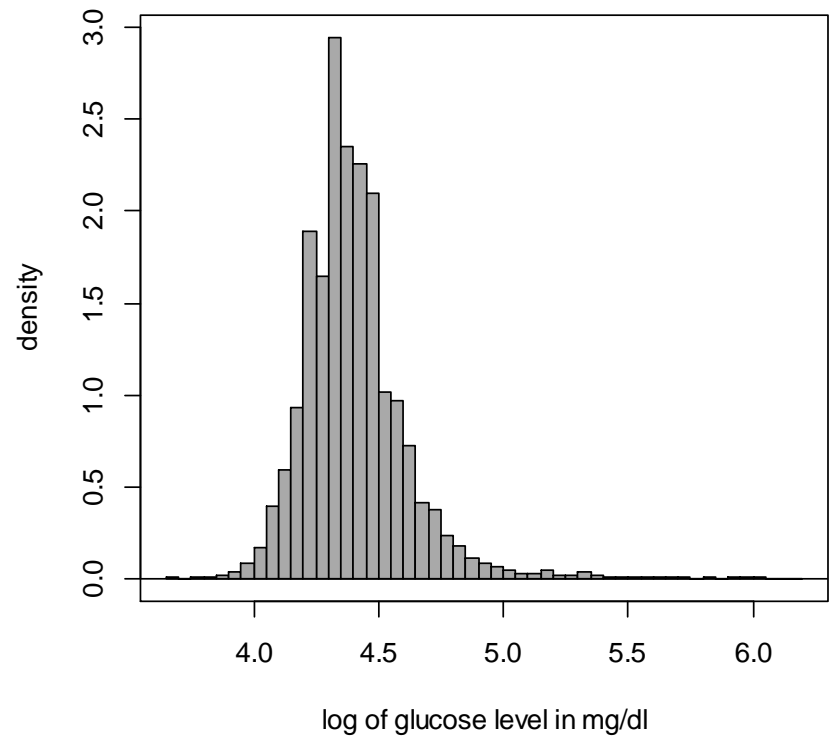
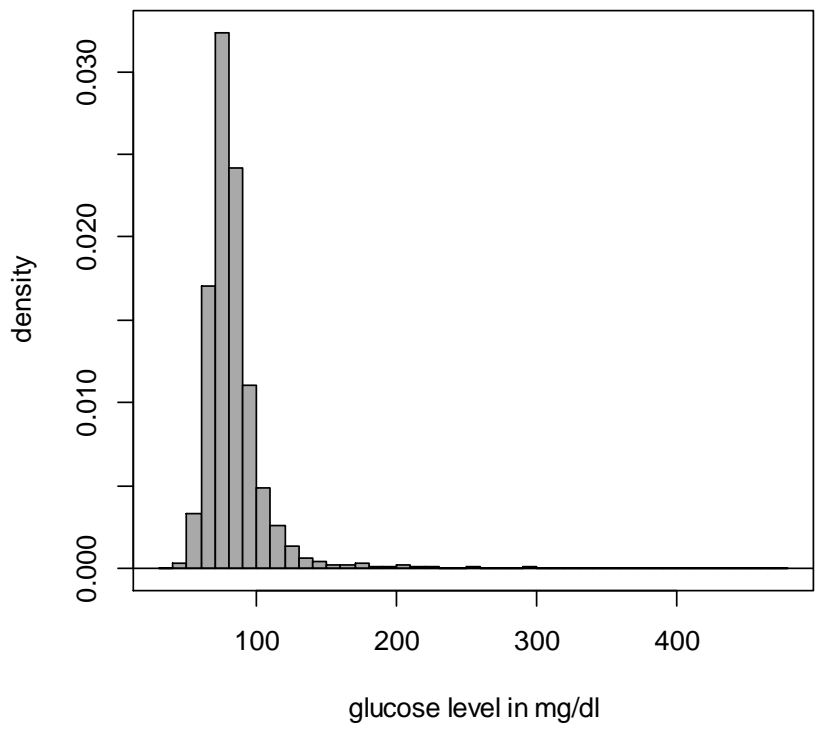
We often see distributions following a Gaussian pretty well.

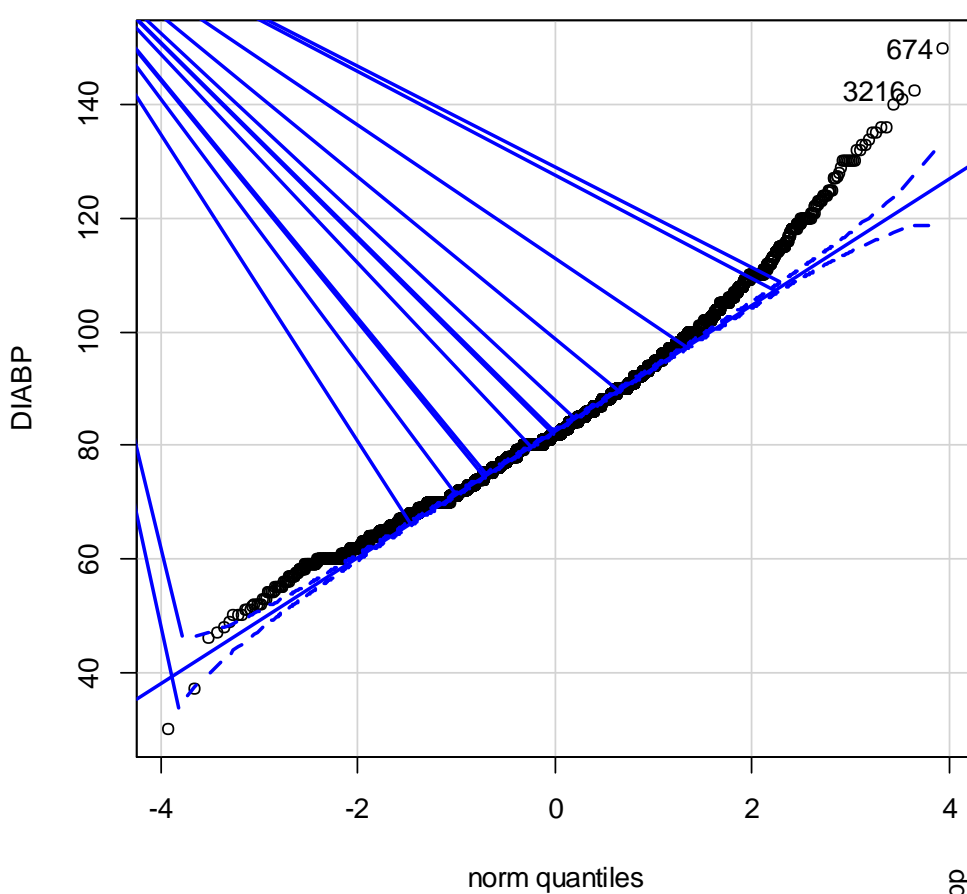
The goal of **estimations** is to get the  $(\mu, \sigma)$  values of the population from the (much smaller but representative) sample we have.



-> next lecture!

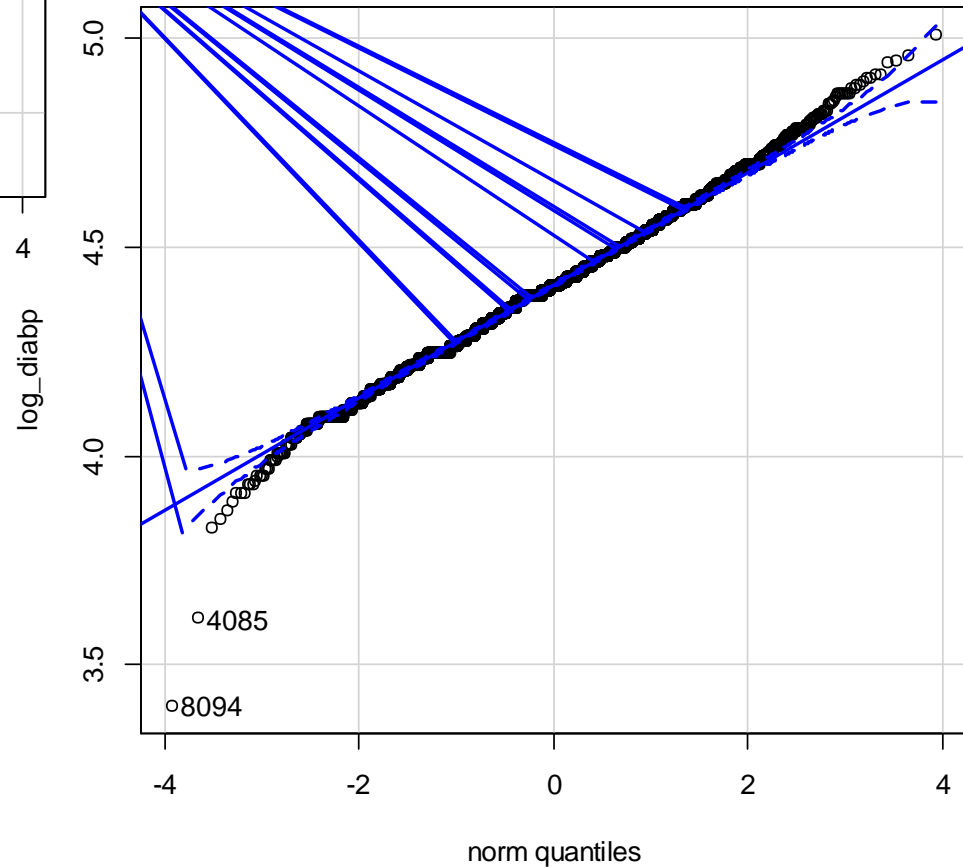
sometimes the data show a skewed distribution, then it is possible than the log of the original values follows a normal distribution. This situation is the *lognormal* distribution.





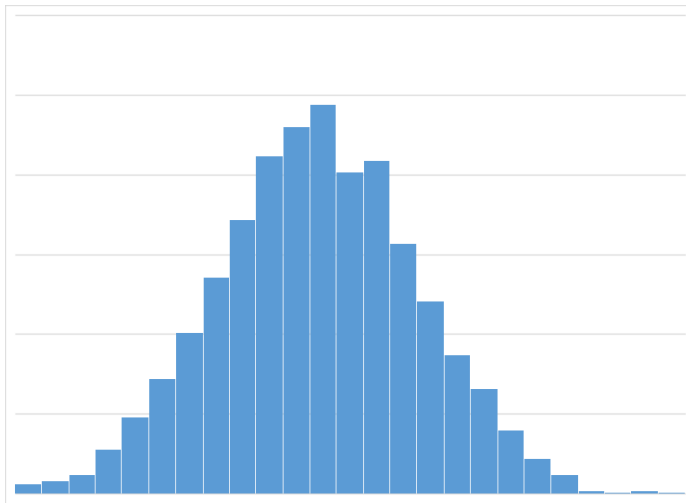
QQ plot  
(quantile-quantile plot)

in this plot we compare the theoretical quantiles with the actual ones to see if the data follow a specific distribution of interest.



## Central distribution theorem

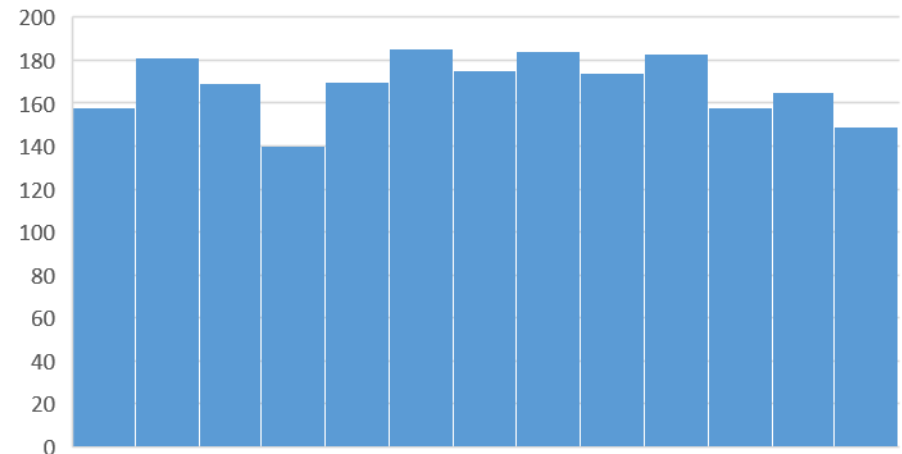
If we have a random variable which consists of a lot of (same distributed) independent random variables, then this common variable will follow a Gaussian. The more variables contribute at the same time, the closer it will be to a Gaussian.



The mean follows a gaussian

Averages of 30 variable samples

Distribution of the individual variables is even



The well known distributions can be used to forecast, or make decisions 😊

-> see hypothesis testing lectures

