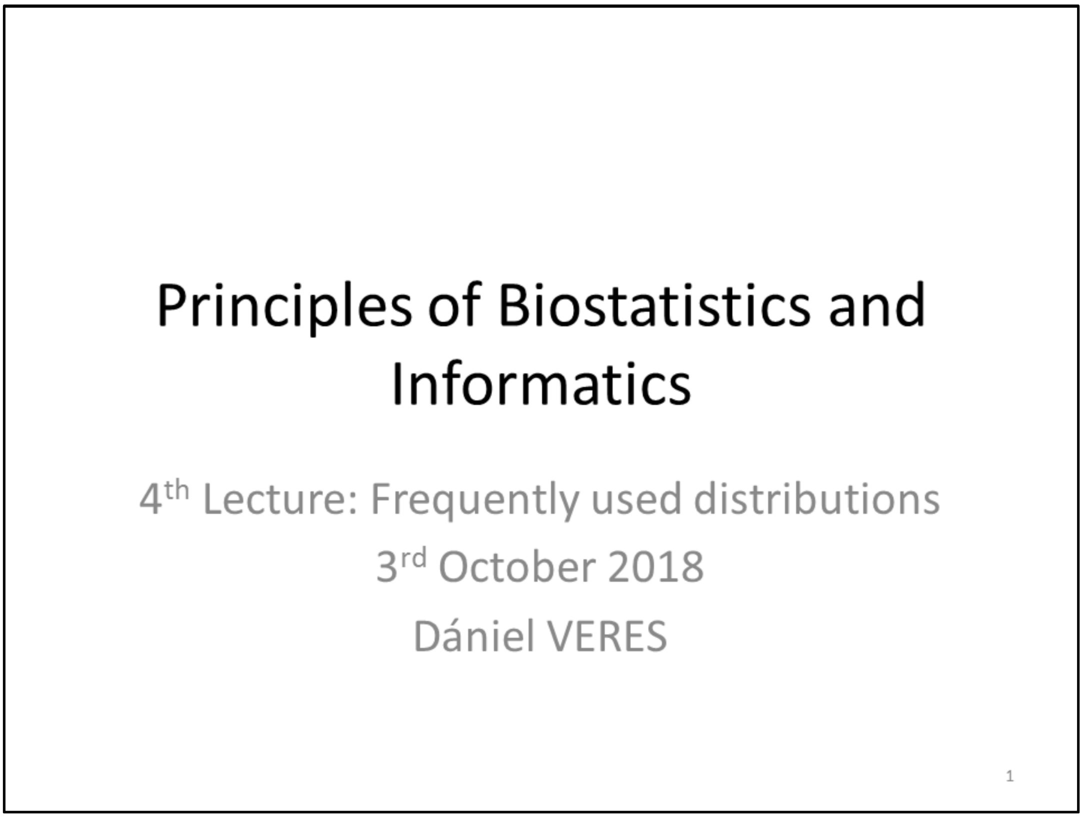


Principles of Biostatistics and Informatics

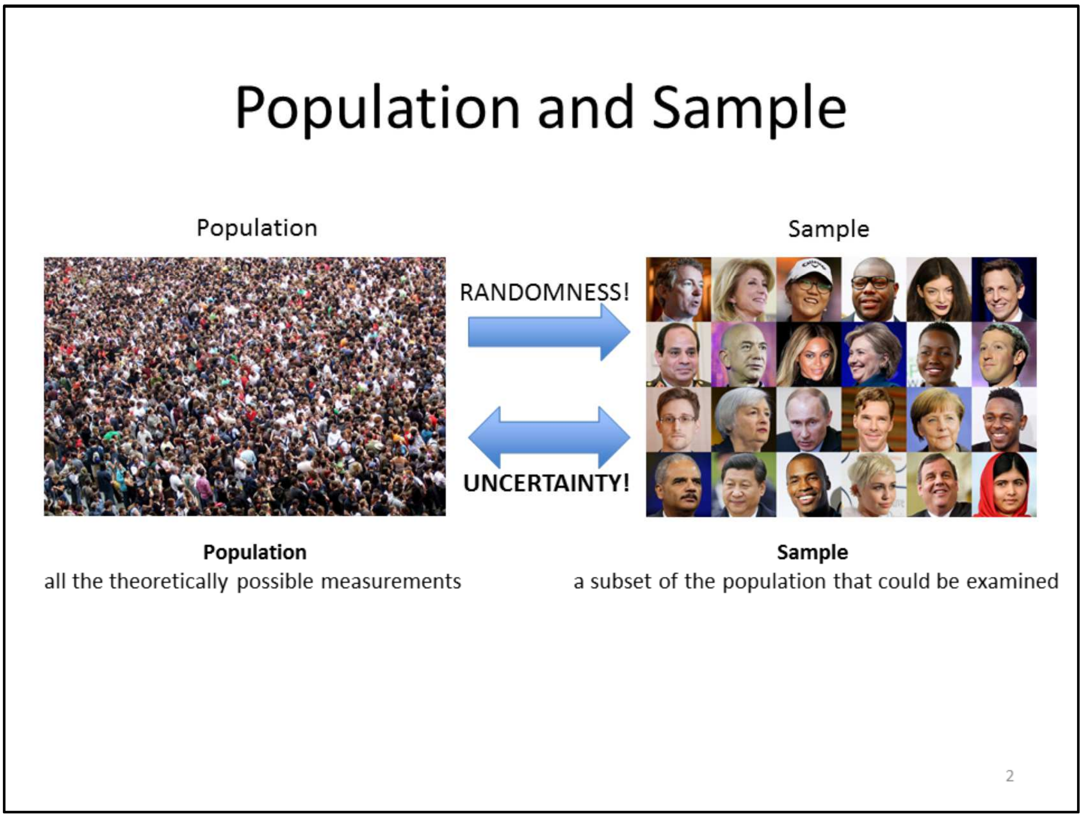
4th Lecture: Frequently used distributions

3rd October 2018

Dániel VERES

1

In this lecture we are discussing on theoretical distributions.
First we define theoretical distributions and their parameters: the expected value and the theoretical variance. After that we discuss on how to estimate/calculate probabilities using special theoretical distributions. We describe the uniform, binomial, Poisson, Gaussian, lognormal, exponential distributions with examples. We characterize the chi-square and the t-distributions. And finally some transformations on distributions will be discussed.



Population and Sample

Population

Sample

RANDOMNESS!

UNCERTAINTY!

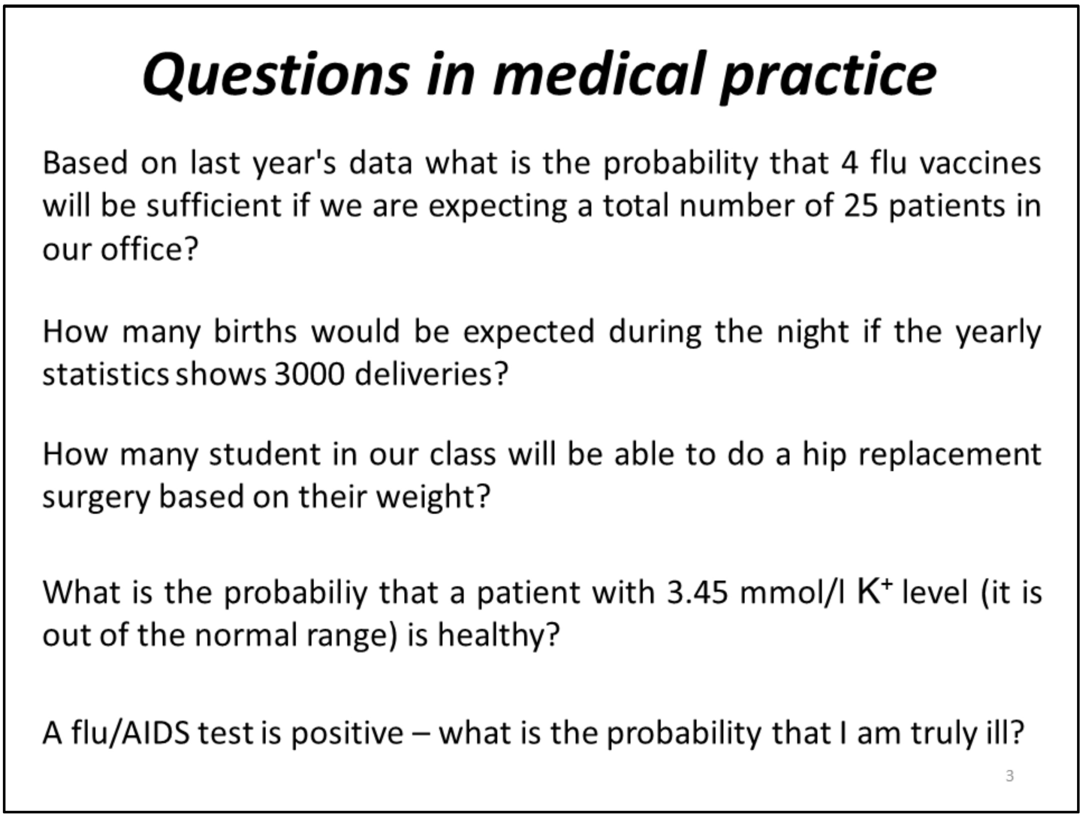
Population
all the theoretically possible measurements

Sample
a subset of the population that could be examined

2

In this slide I repeat again the main problematics of statistics: a theoretical population with all possible measurements (variables and their outcomes) is given but we have only a randomly chosen subset from it that called the sample. In inferential statistics we would like to make conclusions, estimate event counts in a population based on a sample, or in several cases we are interested in conclusions and event counts in a sample based on properties in the population. Now we will focus in the second case with the questions given in the next slide, but all the remarks will be true for the first case too.

As we discussed before we always have an uncertainty because of the sampling. The uncertainty in statistics is described by probabilities.



Questions in medical practice

Based on last year's data what is the probability that 4 flu vaccines will be sufficient if we are expecting a total number of 25 patients in our office?

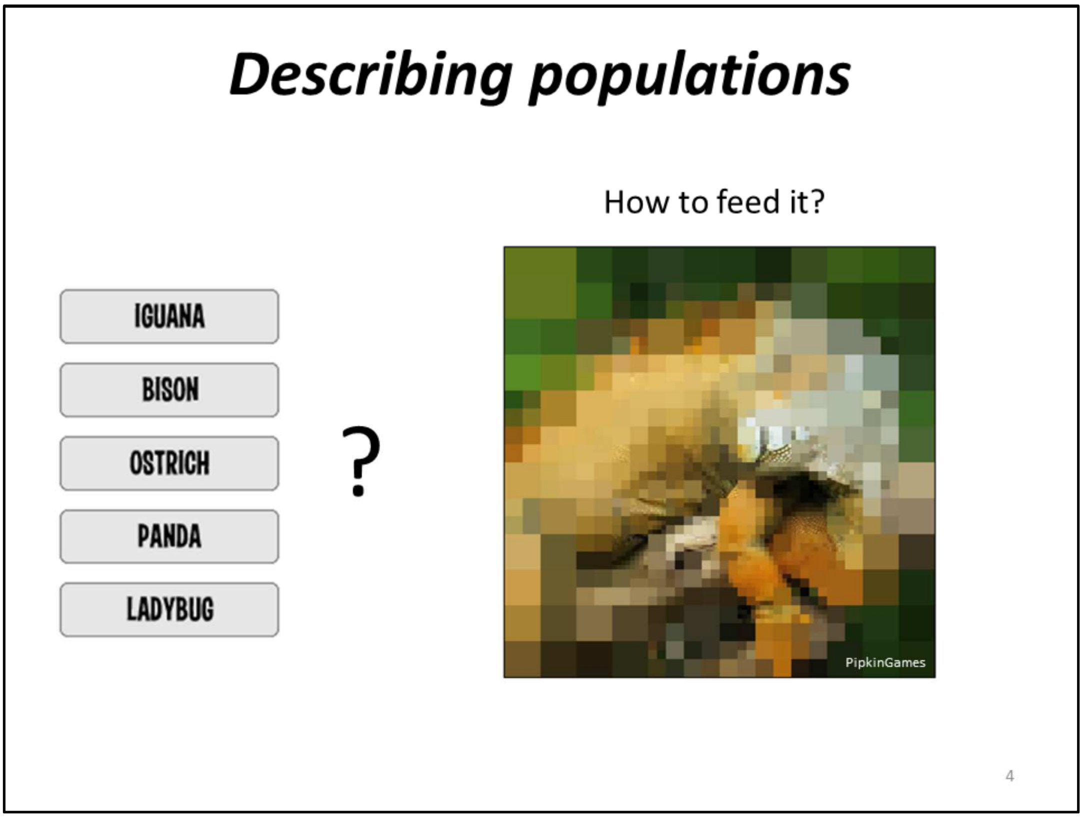
How many births would be expected during the night if the yearly statistics shows 3000 deliveries?

How many student in our class will be able to do a hip replacement surgery based on their weight?

What is the probabiliy that a patient with 3.45 mmol/l K^+ level (it is out of the normal range) is healthy?

A flu/AIDS test is positive – what is the probability that I am truly ill?

Here I gave some questions from medical statistics where we conclude (infer) on a property of a sample based on a population (but the population is estimated based on other samples). Eg. in the first question the „last year’s data” belongs to the population, the „25 patients” is our sample and the „flue vaccines” (needed) is the property.



Describing populations

How to feed it?

IGUANA

BISON

OSTRICH

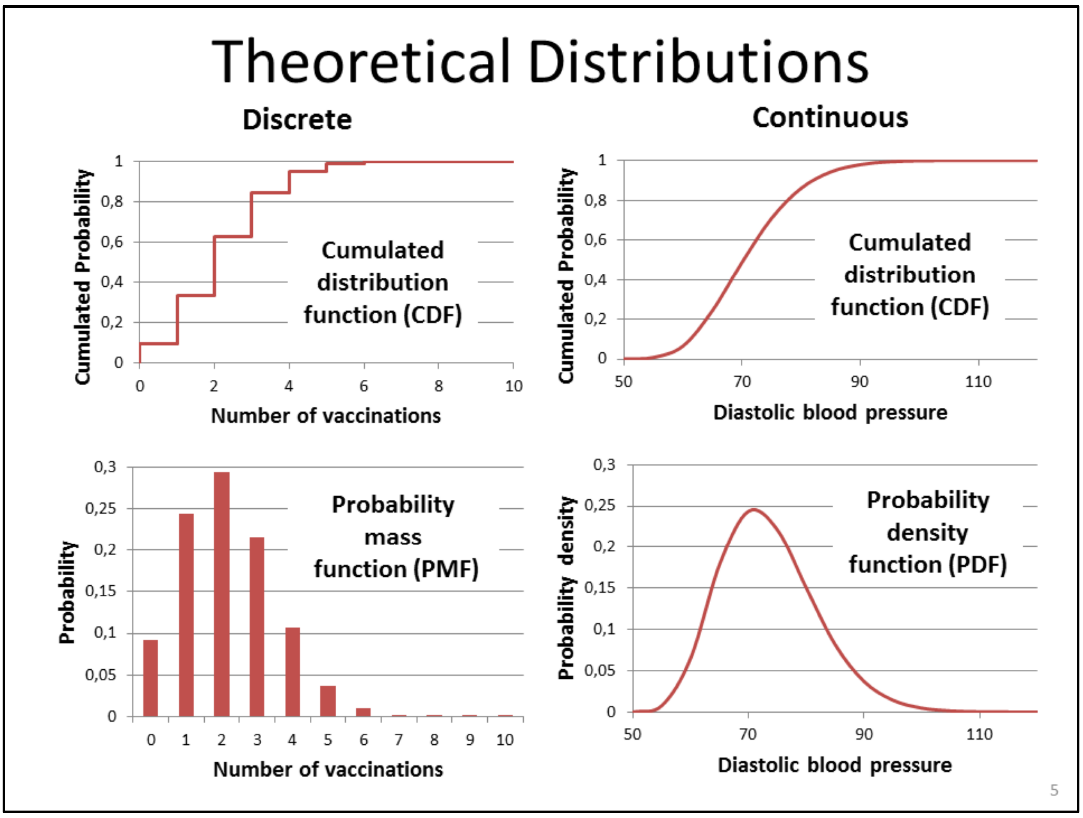
PANDA

LADYBUG

?

PipkinGames

For a good inference on a sample (or on a population – as you will see in the next lectures) we need to know how the population looks like.
To feed our not exactly known animal we need to know the species (the population) and the feeding behavior (a specific „parameter”) of the species.



Theoretical Distributions

Discrete

Continuous

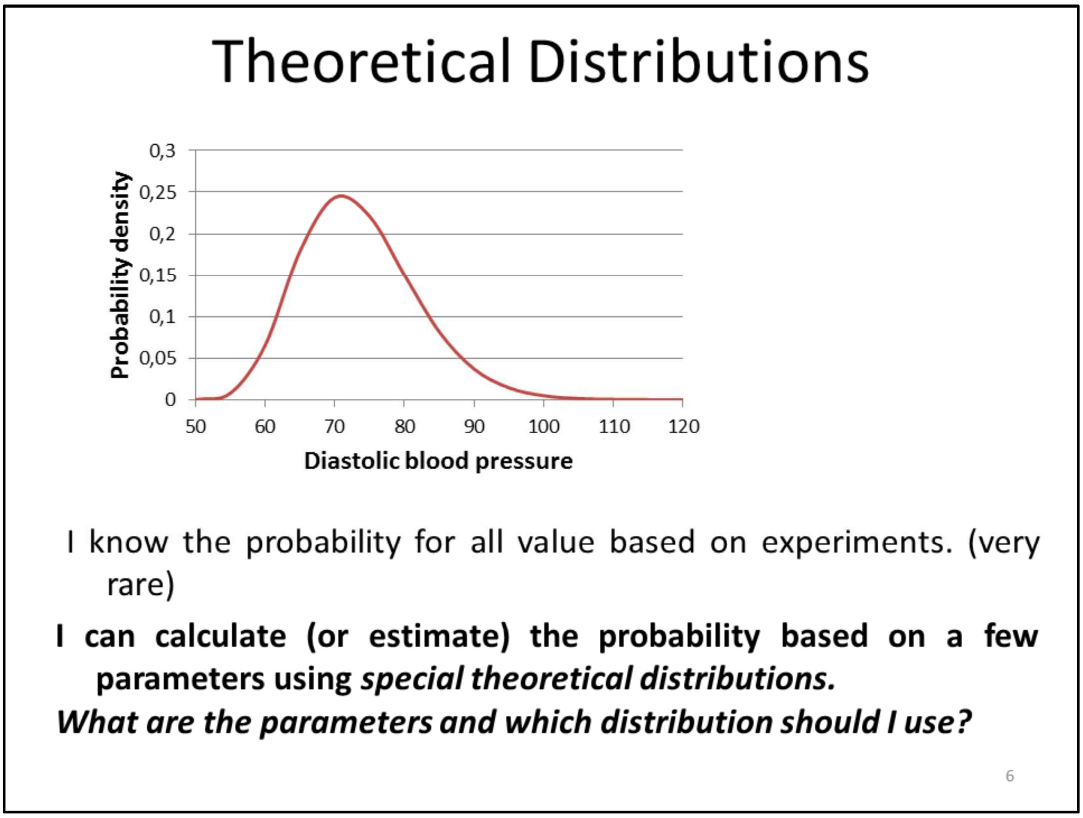
Cumulated distribution function (CDF)

Cumulated distribution function (CDF)

Probability mass function (PMF)

Probability density function (PDF)

The perfect description of a population is its distribution – the distribution of a population is called theoretical distributions. Theoretical distributions shows the probability for a given value, instead of frequencies – remember for the law of large numbers on relative frequencies.
The theoretical distributions categorized in two way. In one aspect we could say that our distribution could be continuous or discrete – as we have categorical and continuous variables. As we learn before we can create the cumulated and not cumulated distributions.



Theoretical Distributions

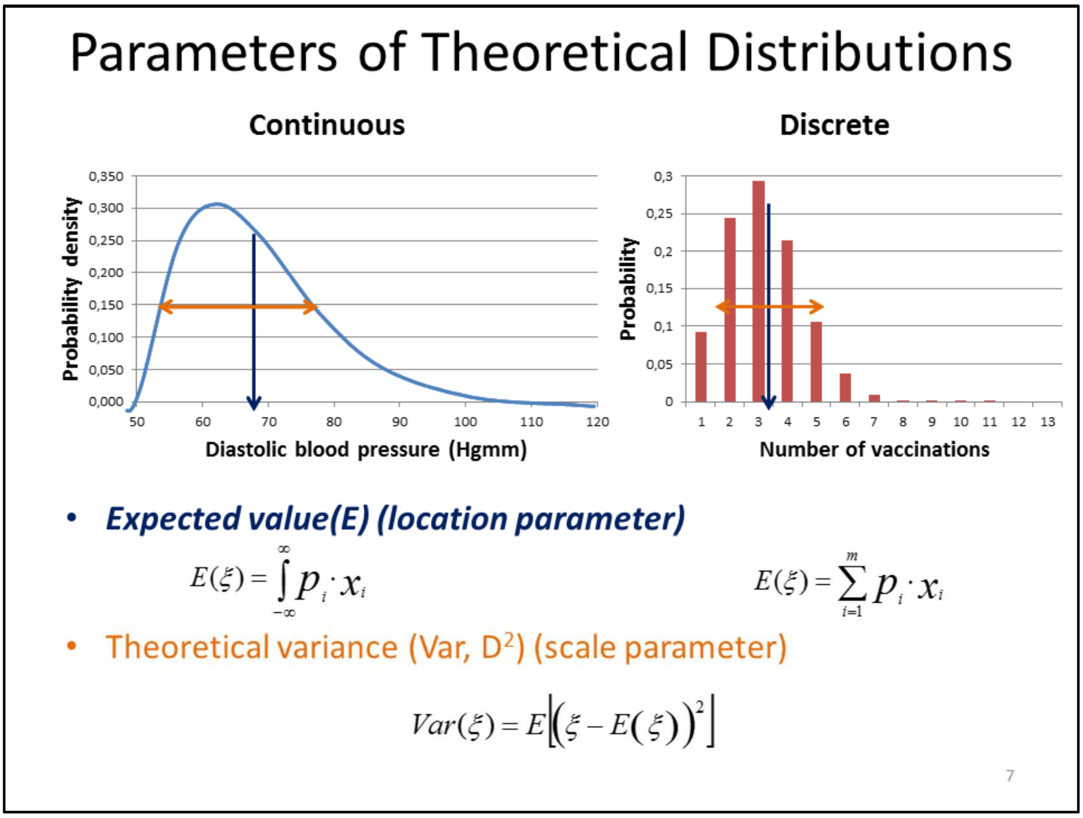
I know the probability for all value based on experiments. (very rare)

I can calculate (or estimate) the probability based on a few parameters using *special theoretical distributions.*

What are the parameters and which distribution should I use?

6

In a very rare cases we know the probabilities for every outcomes based on a large number of experiments.
But usually we can calculate or more often estimate the probabilities based on a few value (parameters) using *special distributions*. So the question is what are these parameters and which special distribution I should use for the given problem.



Parameters of Theoretical Distributions

Continuous

Discrete

- ***Expected value(E) (location parameter)***

$$E(\xi)=\int_{-\infty}^{\infty} p_i \cdot x_i \qquad\qquad E(\xi)=\sum_{i=1}^{m} p_i \cdot x_i$$

- Theoretical variance (Var, D^2) (scale parameter)

$$Var(\xi)=E\left[\left(\xi-E(\xi)\right)^2\right]$$

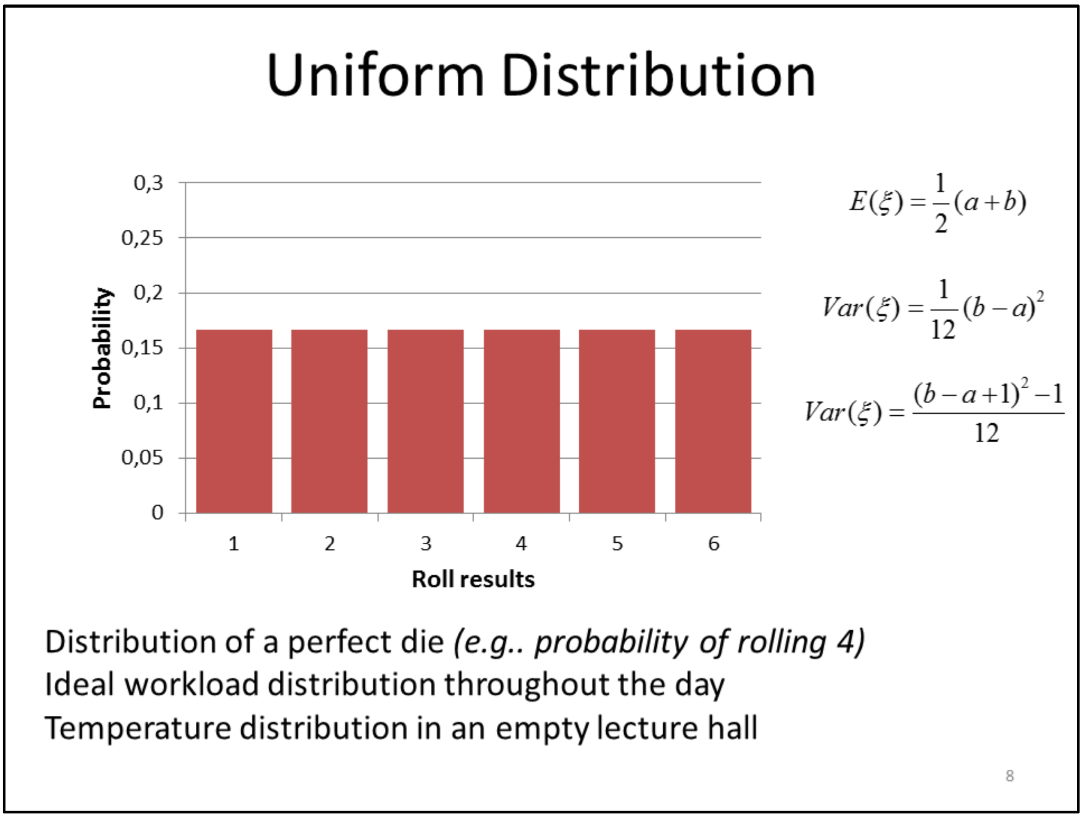
7

Theoretical distributions have similar parameters as mentioned in descriptive statistics. There is a parameter that describe the center of the distribution and an other one that describe the width of the distribution.
The first one called *expected value* (abbreviated with E), the second is the *theoretical variance* (Var). In the equation *x* is the given value and *p* is the probability of that value. The expected value calculated slightly differently for continuous and discrete variables.

As I showed in the lecture the **expected value is equal with the *mean of the population***. For continuous variable we use infinite small binwidth for summarization – that is the integral (∫).

This two indicator (the expected value and the variance) defines exactly the distribution that means knowing this indicators we could calculate the probability for all value.

Calculation of an expected value and theoretical variance has been shown in the lecture. (see attached excel file).



Uniform Distribution

$$E(\xi)=\frac{1}{2}(a+b)$$

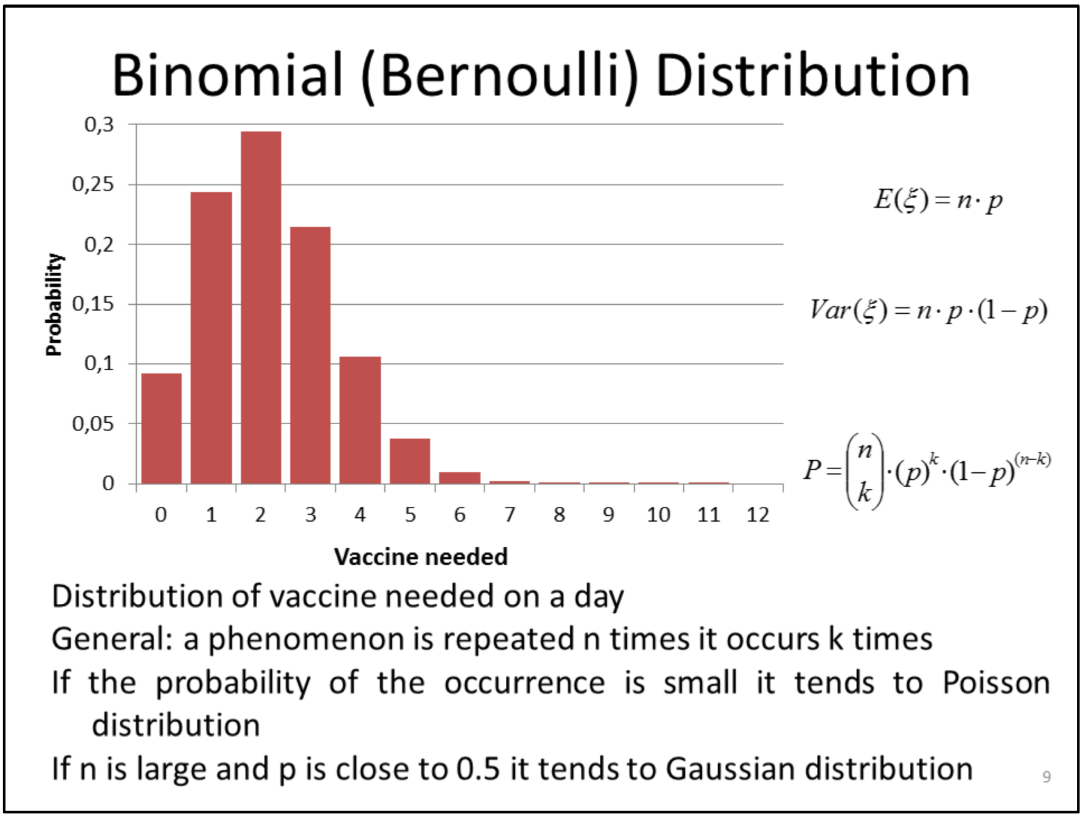
$$Var(\xi)=\frac{1}{12}(b-a)^2$$

$$Var(\xi)=\frac{(b-a+1)^2-1}{12}$$

Distribution of a perfect die *(e.g.. probability of rolling 4)*
Ideal workload distribution throughout the day
Temperature distribution in an empty lecture hall

8

Let’s see first the *uniform distribution*.
We have uniform distribution for example if we rolling a dice or we talk on the ideal workload or temperature distribution in an empty space.
For example using the uniform distribution we can calculate the probability of rolling 4 with a die.
The formula of the expected value and the variance is available in the formula collection.
Here *a* and *b* are the smallest and largest outcomes. In the case of a six sided die the expected value is 0.5*(1+6)=3.5.



Binomial (Bernoulli) Distribution

$$E(\xi) = n \cdot p$$

$$Var(\xi) = n \cdot p \cdot (1-p)$$

$$P = \binom{n}{k} \cdot (p)^{k} \cdot (1-p)^{(n-k)}$$

Distribution of vaccine needed on a day

General: a phenomenon is repeated n times it occurs k times

If the probability of the occurrence is small it tends to Poisson distribution

If n is large and p is close to 0.5 it tends to Gaussian distribution

9

The *binomial (or Bernoulli) distribution* is used in general if a phenomenon is repeated n times it occurs k times. It is called binomial because it handle the situation where we have two outcomes: an event occurs or not occurs.

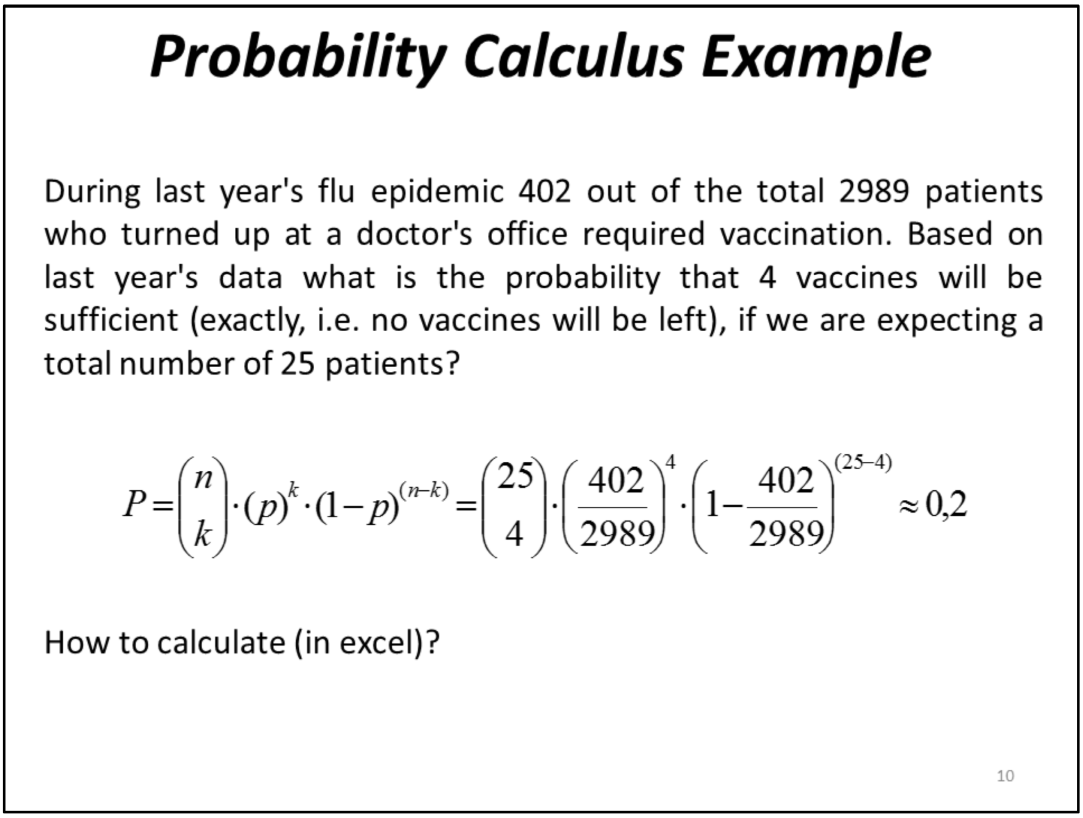
An example was the vaccination I mentioned before: During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines left) in a certain day, if we are expecting a total number of 25 patients?

For the calculation we need the expected value and the variance – or the parameters that describe it. E=n*p and Var=n*p*(1-p) so we need *n* (in our case the expected number of patient: 25) and *p (*we could estimate it using last years data: p=402/2989*)*.

Based on this values we can calculate the probability of *k* (in our case 4) using the equation of the distribution. This example calculation in excel was show in the lecture (see attached excel file).

If the probability of occurrence (p) is small the binomial distribution tends to a Poisson distribution.

If repetition (n) is large and the probability is close to 0.5 it tends to Gaussian distribution.



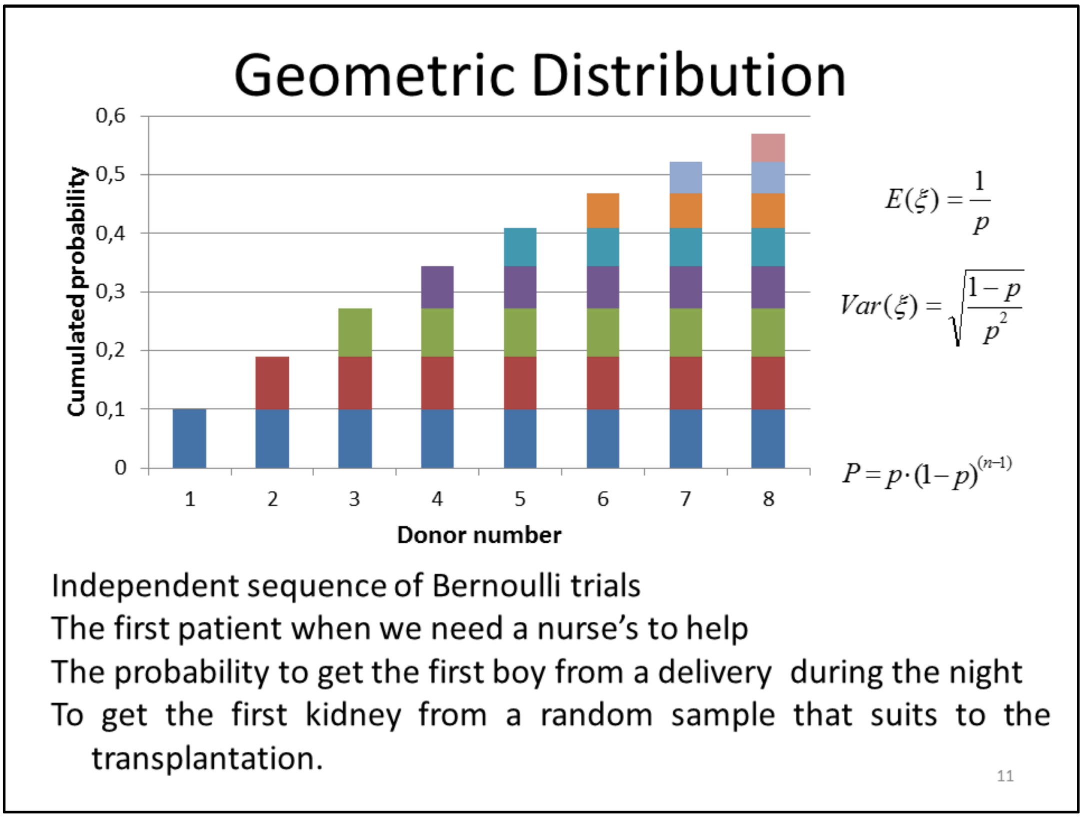
Probability Calculus Example

During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines will be left), if we are expecting a total number of 25 patients?

$$P=\binom{n}{k}\cdot(p)^{k}\cdot(1-p)^{(n-k)}=\binom{25}{4}\cdot\left(\frac{402}{2989}\right)^{4}\cdot\left(1-\frac{402}{2989}\right)^{(25-4)}\approx 0{,}2$$

How to calculate (in excel)?

An example calculation.
During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines left) in a certain day, if we are expecting a total number of 25 patients?
For answering the question we use the Bernoulli distribution’s (see later) equation. I show this equation to highlight that Bernoulli distribution is based on probability calculus.



Geometric Distribution

$$E(\xi) = \frac{1}{p}$$

$$Var(\xi) = \sqrt{\frac{1-p}{p^2}}$$

$$P = p \cdot (1-p)^{(n-1)}$$

Independent sequence of Bernoulli trials

The first patient when we need a nurse's to help

The probability to get the first boy from a delivery during the night

To get the first kidney from a random sample that suits to the transplantation.

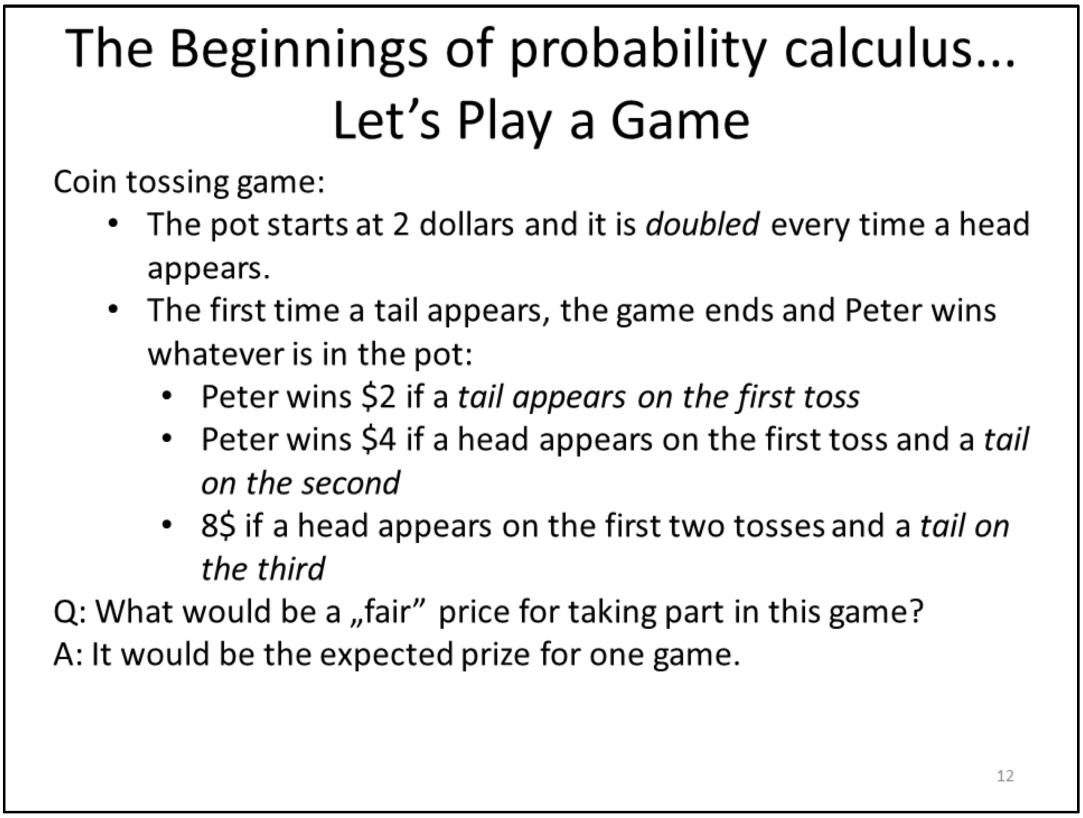
A geometric distribution is a special Binomial distribution. In this case we make independent sequence of Bernoulli trials.

A medical example: What is the probability to find the first kidney that suits to the transplantation if we check one, then two... random person.

What is the probability that we can examine the first patient without calling the nurse to help us? Or the probability that we couldn't examine the xth people without any help before. The probability to get the first boy from a delivery during the night.

In this graph I show the probability of this situation – this is a kind of a cumulative frequency distribution (we cumulate the probability that we have a boy at the first delivery (category 1); the probability we have a boy at the first delivery + the probability that the first is not a boy, but the second is a boy (category 2)...)

In the play of St. Petersburg paradox the prize of a single game follow geometric distribution too.



The Beginnings of probability calculus… Let’s Play a Game

Coin tossing game:

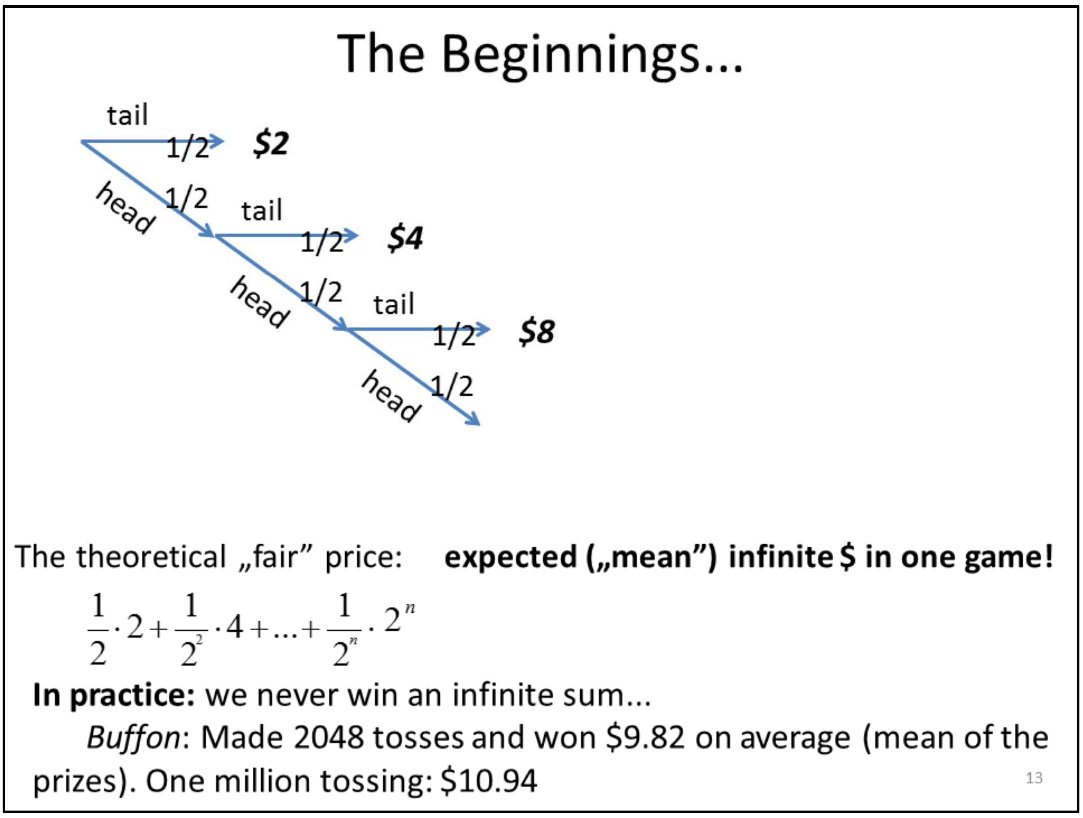
- The pot starts at 2 dollars and it is *doubled* every time a head appears.
- The first time a tail appears, the game ends and Peter wins whatever is in the pot:
  - Peter wins $2 if a *tail appears on the first toss*
  - Peter wins $4 if a head appears on the first toss and a *tail on the second*
  - 8$ if a head appears on the first two tosses and a *tail on the third*

Q: What would be a „fair” price for taking part in this game?
A: It would be the expected prize for one game.

Based on some opinion the beginnings of probability calculus was the Saint Petersburg paradox. It was a theoretical game publicated in 1715. A theoretical person – *Peter* – plays the game. The rules are the next.

1. The pot starts at 2 dollars and it is *doubled* every time a head appears.
2. The first time a tail appears, the game ends and Peter wins whatever is in the pot:
   - Peter wins $2 if a *tail appears on the first toss*
   - Peter wins $4 if a head appears on the first toss and a *tail on the second*
   - 8$ if a head appears on the first two tosses and a *tail on the third*
   - and so on

The question appears what would be a fair price for taking part in this game? It would be the expected prize (that Peter win) for one game.



The Beginnings...

The theoretical „fair" price: **expected („mean") infinite $ in one game!**

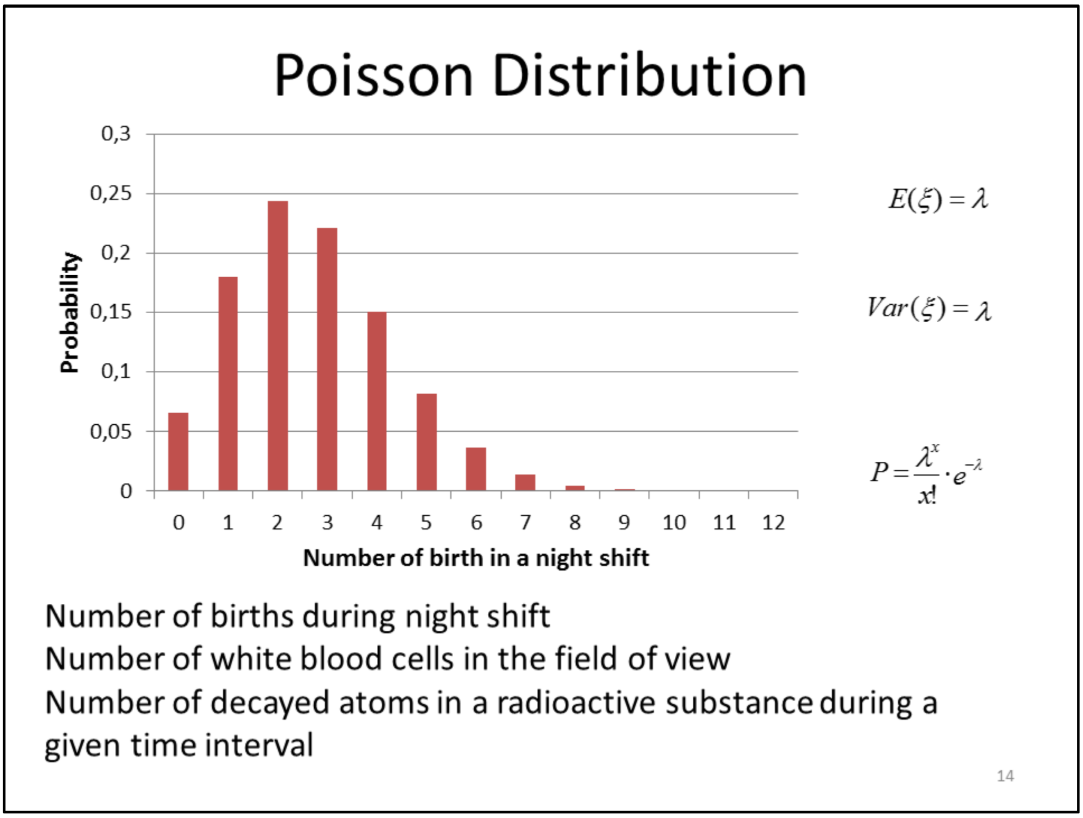
$$\frac{1}{2}\cdot 2+\frac{1}{2^2}\cdot 4+...+\frac{1}{2^n}\cdot 2^n$$

In practice: we never win an infinite sum...

Buffon: Made 2048 tosses and won $9.82 on average (mean of the prizes). One million tossing: $10.94

What will be the prize Peter win?
In the half of the cases the first result is tail – so Peter win ½*2 dollars in average.
In the half of the cases when the first toss is head (that is in the half of the cases) the second is tail – so we get ½*1/2*4 dollars in average and so on. Therefore it follows a geometric distribution.

Additional: We could calculate the expected prize in one single game based on the equation in the slide. Based on that the expected price is infinite in one game.
But in practice we realize the sum in one single game is never infinite! For example: *Buffon* (a famous mathematician) made 2048 tosses and won $9.82 on average (the mean of the prizes). With 1 million tossing we get $10.94 on average.
We could experience the mean of *the prize won in one single game is increasing and tends to the theoretical infinite if we play more and more games.*



Poisson Distribution

$$E(\xi)=\lambda$$

$$Var(\xi)=\lambda$$

$$P=\frac{\lambda^x}{x!}\cdot e^{-\lambda}$$

Number of births during night shift
Number of white blood cells in the field of view
Number of decayed atoms in a radioactive substance during a given time interval

14

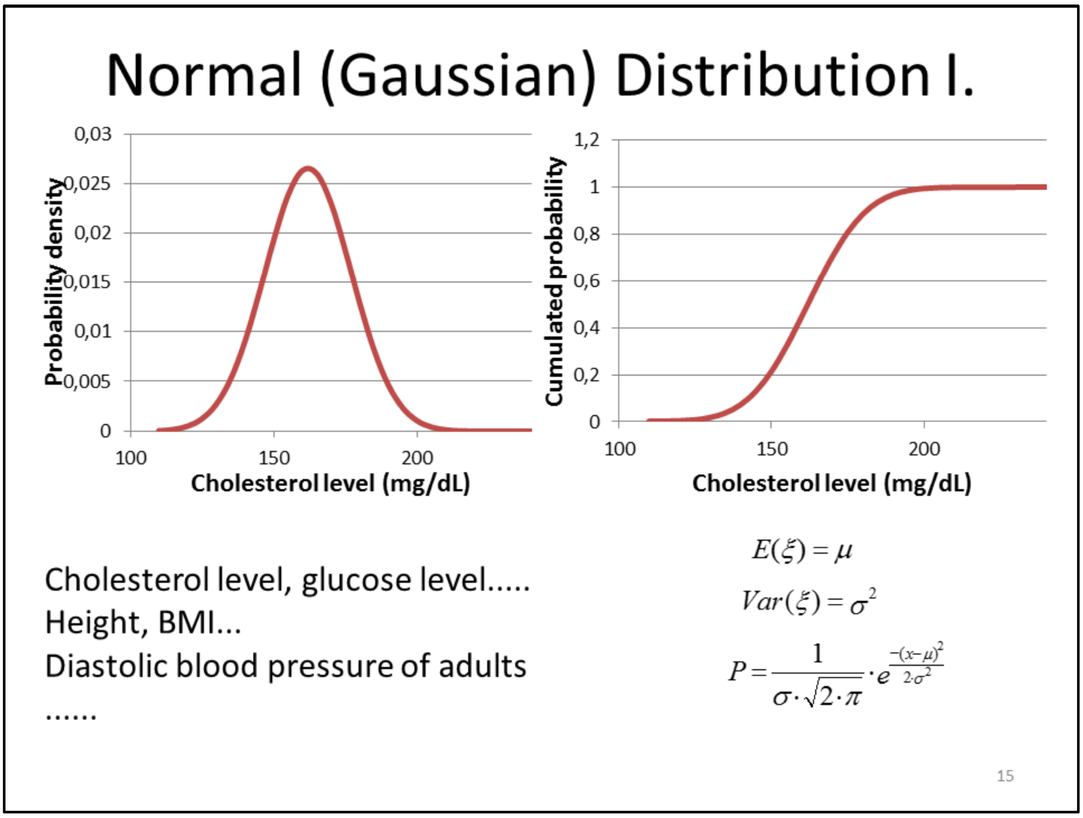
This distribution is a special binomial distribution too.

The Poisson distribution has a special attribution: the expected value and the variance are the same – so we need only 1 parameter to describe this distribution. It is useful to easier calculations.

For example based on this distribution we can calculate the probability that we have 3 birth during our night shift. Other examples that follows Poisson distribution are: Number of white blood cells in the field of view, number of decayed atoms in a radioactive substance during a given time interval.

In general: number of elements in a given time interval or volume..., if the *probability of the occurrence is small*.

The expected value (*λ*) derived from *n***p* (number of „repetition” * probability).



Normal (Gaussian) Distribution I.

Cholesterol level, glucose level.....
Height, BMI...
Diastolic blood pressure of adults
......

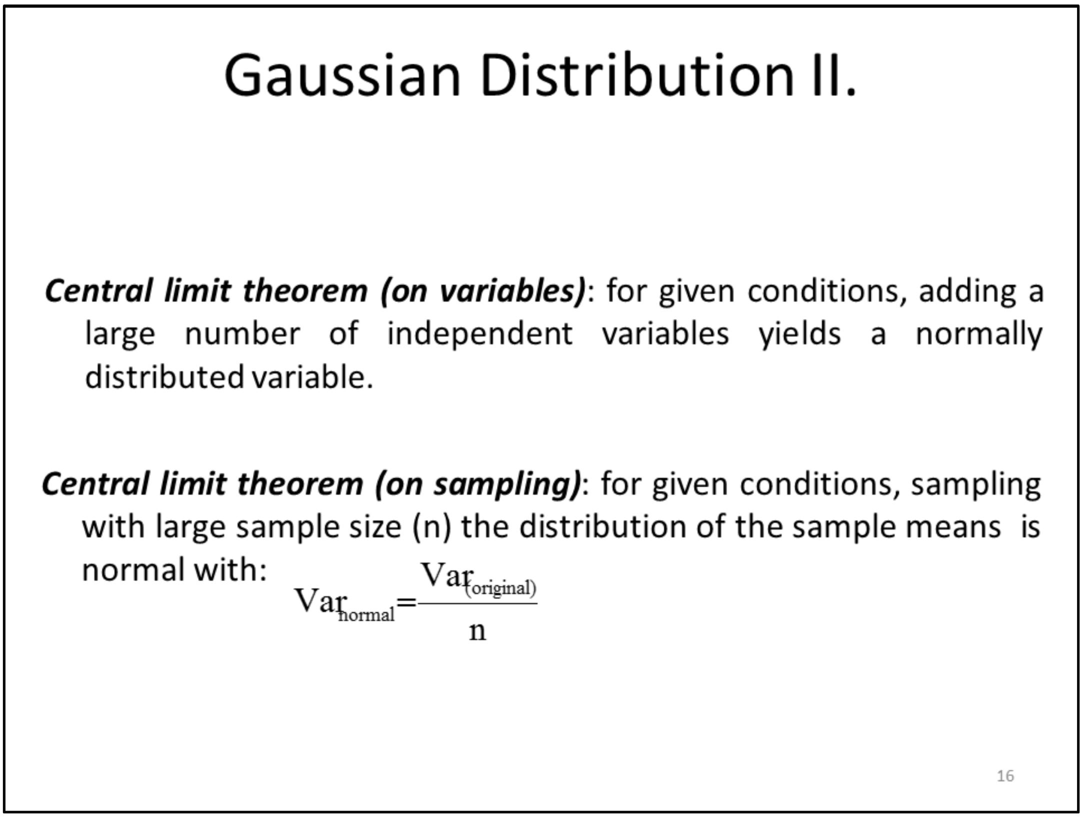
$$E(\xi) = \mu$$

$$Var(\xi) = \sigma^2$$

$$P = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}}$$

15

The normal or Gaussian distribution is the most common in medical practice.
In this slide I plotted both the relative distribution and cumulative frequency functions, because this is the most important distribution for us. As you see against the other mentioned distribution this is symmetric one. Both the skewness and kurtosis is 1.
The most of the variables in medical practice follows normal (Gaussian) distribution – e.g.. enzyme levels, height, body mass index (BMI), blood pressures...
Why?



Gaussian Distribution II.

Central limit theorem (on variables): for given conditions, adding a large number of independent variables yields a normally distributed variable.

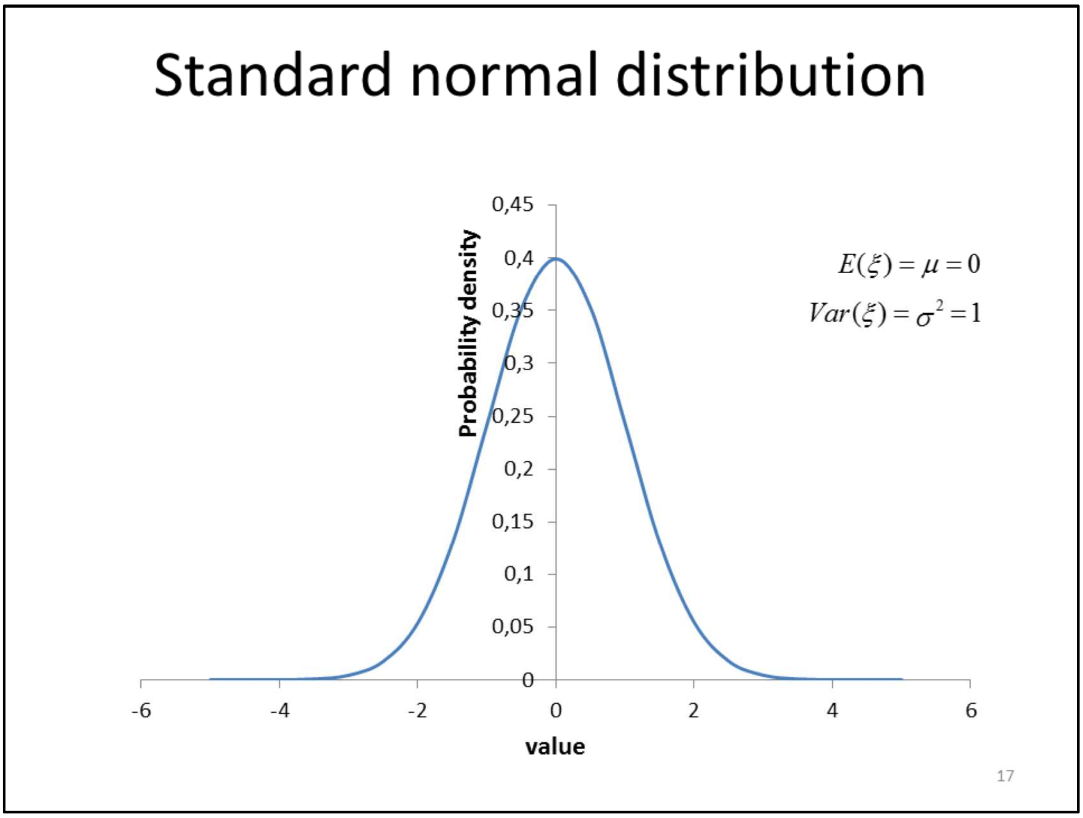
Central limit theorem (on sampling): for given conditions, sampling with large sample size (n) the distribution of the sample means is normal with:

$$\mathrm{Var}_{\mathrm{normal}}=\frac{\mathrm{Var}_{\mathrm{(original)}}}{\mathrm{n}}$$

The reason of why we have normal distribution in most of the variables in medical practice described by the central limit theorem. It says that summarizing large number of independent variables resulted a normally distributed variable. In medical practice most of the measure values affected by several factor: gens from father, gens from mother, nutrition, way of life...

An other important wording of this law: if you take a sample with a sample size n and n is large the distribution of the sample means is normal and its variance is Var(original)/n

Example given in the attached excel file.



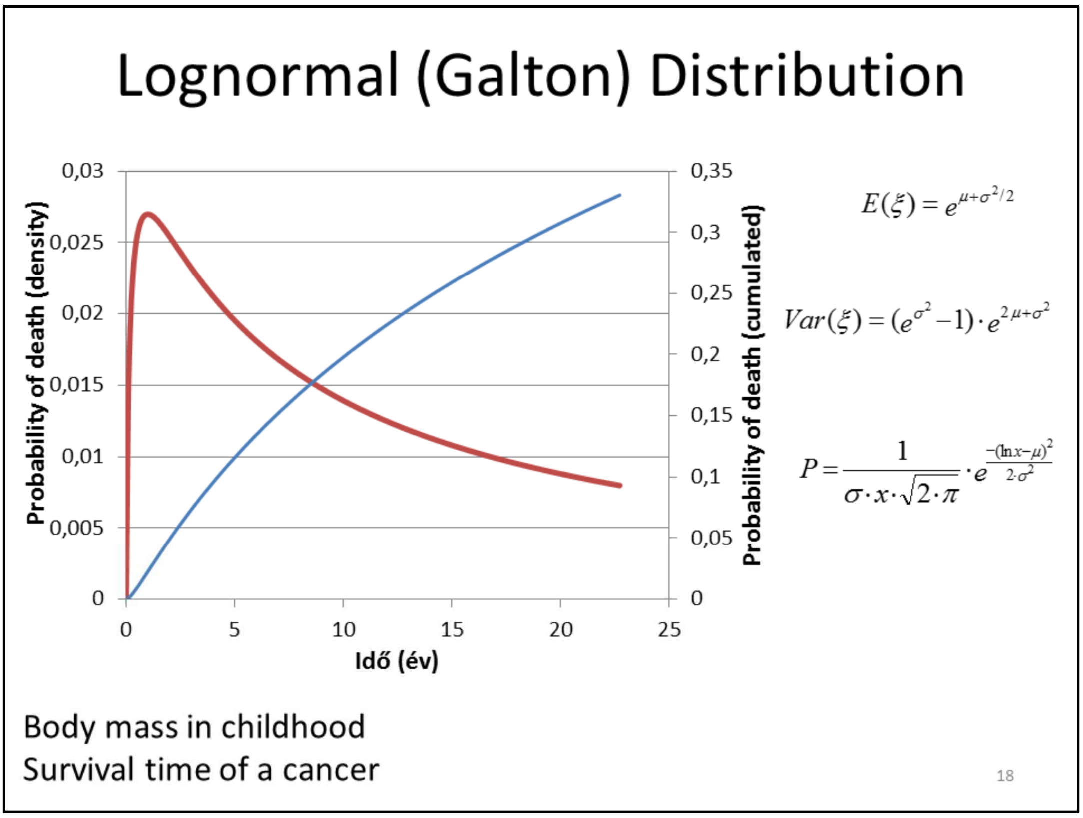
Standard normal distribution

$$E(\xi) = \mu = 0$$

$$Var(\xi) = \sigma^2 = 1$$

17

A special normal distribution is the standard normal distribution, where the expected value is 0 and the variance is 1.



Lognormal (Galton) Distribution

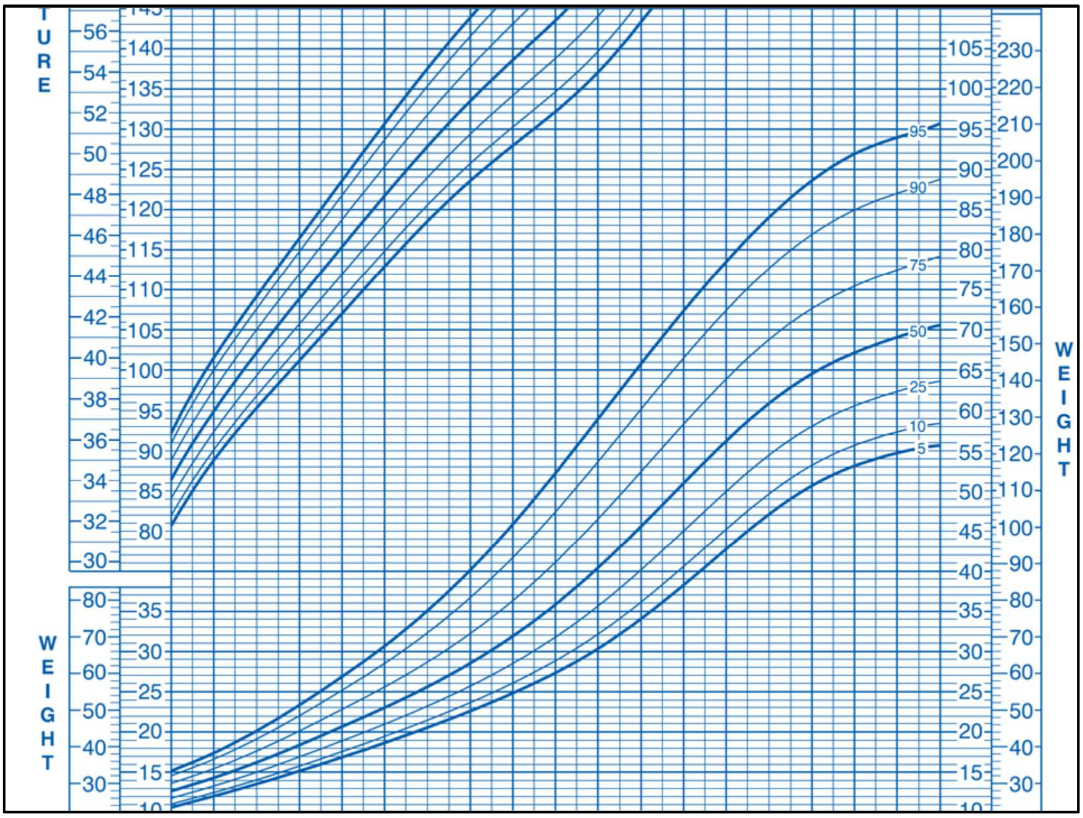
$$E(\xi) = e^{\mu+\sigma^2/2}$$

$$Var(\xi) = (e^{\sigma^2} - 1) \cdot e^{2\mu+\sigma^2}$$

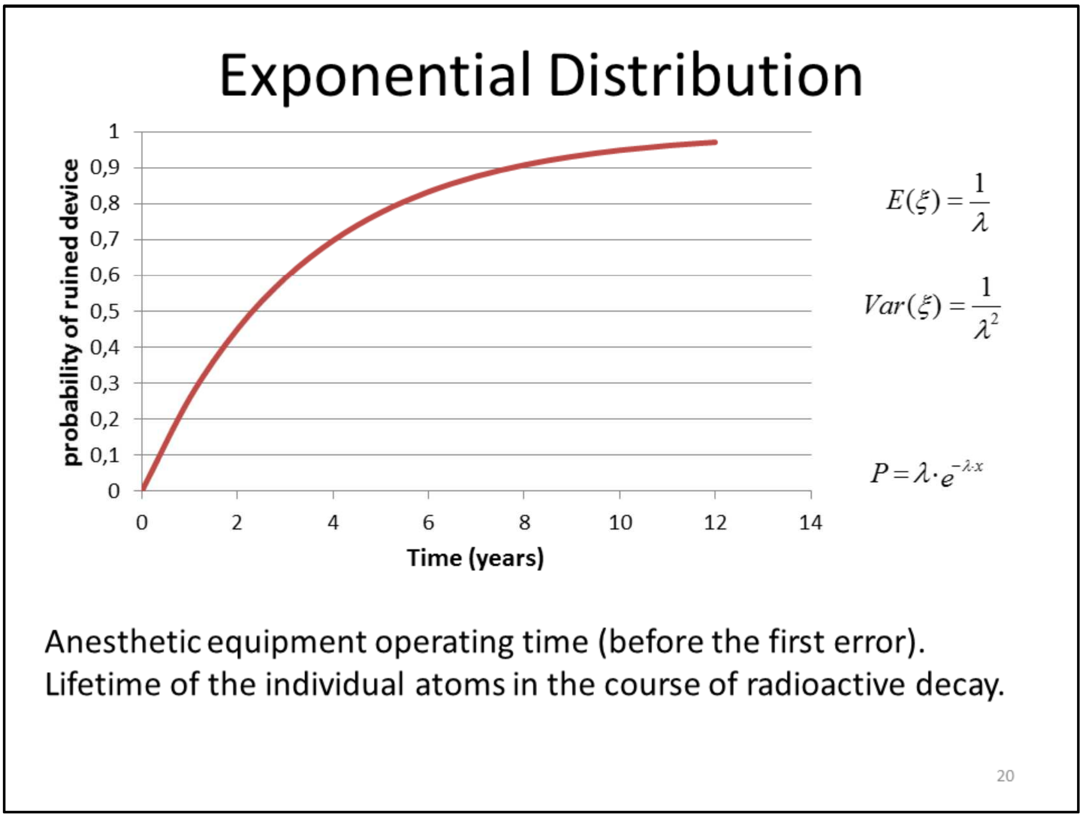
$$P = \frac{1}{\sigma \cdot x \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(\ln x - \mu)^2}{2 \cdot \sigma^2}}$$

Body mass in childhood
Survival time of a cancer

A common distribution in medical practice is the lognormal distribution. For example the body parameters (mass) in childhood, survival time of a cancer.
In general if the values of the variable are close to 0 and couldn't be negative instead of a normal distribution we get a lognormal distribution.



Here we can see the asymmetrically distribution. (eg. the distance between the 5th percentile and 10th percentile is different then the distance between the 95th and 90th percentile.)



Exponential Distribution

$$E(\xi) = \frac{1}{\lambda}$$

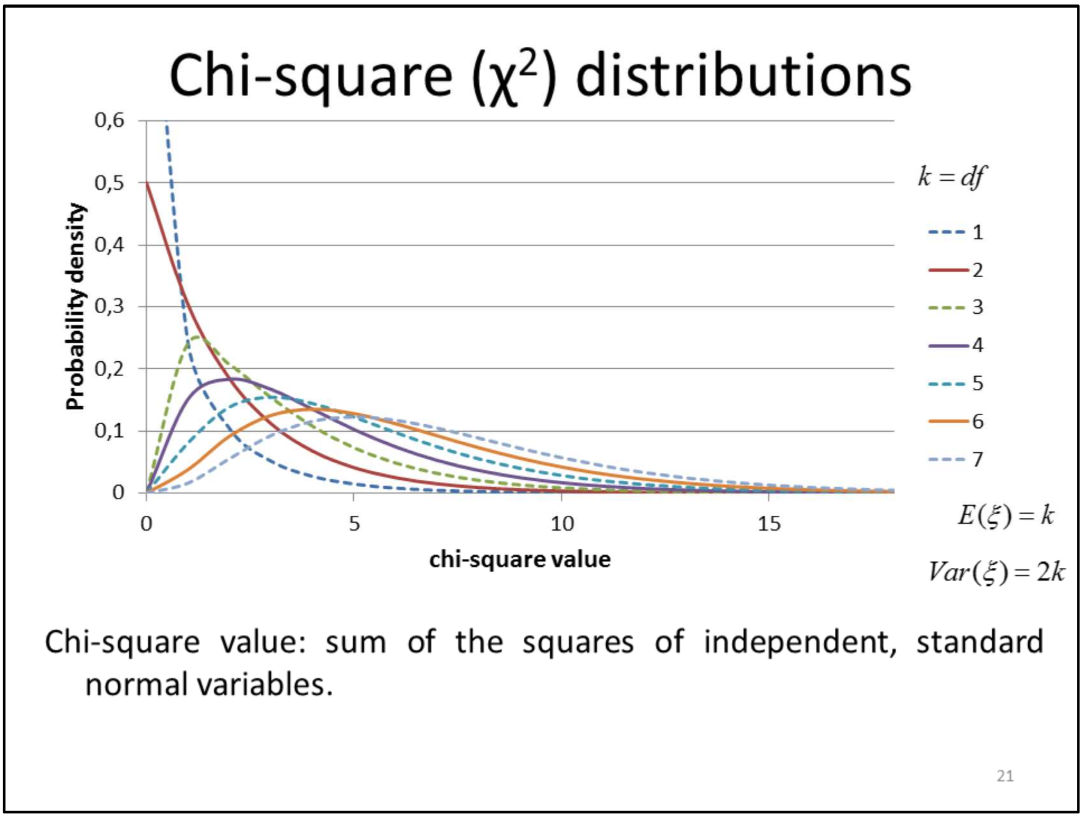
$$Var(\xi) = \frac{1}{\lambda^2}$$

$$P = \lambda \cdot e^{-\lambda x}$$

Anesthetic equipment operating time (before the first error).
Lifetime of the individual atoms in the course of radioactive decay.

20

The exponential distribution is well know in biophysics and has some appearance in medical practice too. I give you two example: anesthetic equipment operating time (before the first error) and the lifetime of the individual atoms in the course of radioactive decay.



Chi-square (χ^2) distributions

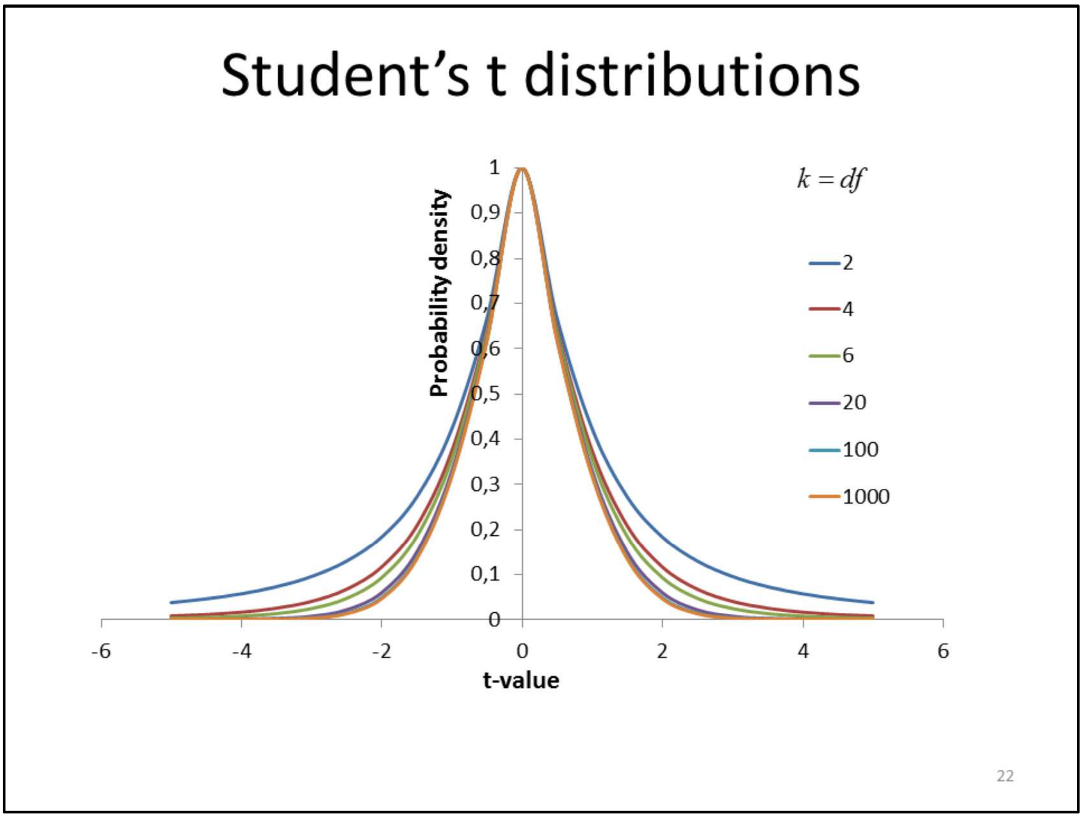
$E(\xi) = k$

$Var(\xi) = 2k$

Chi-square value: sum of the squares of independent, standard normal variables.

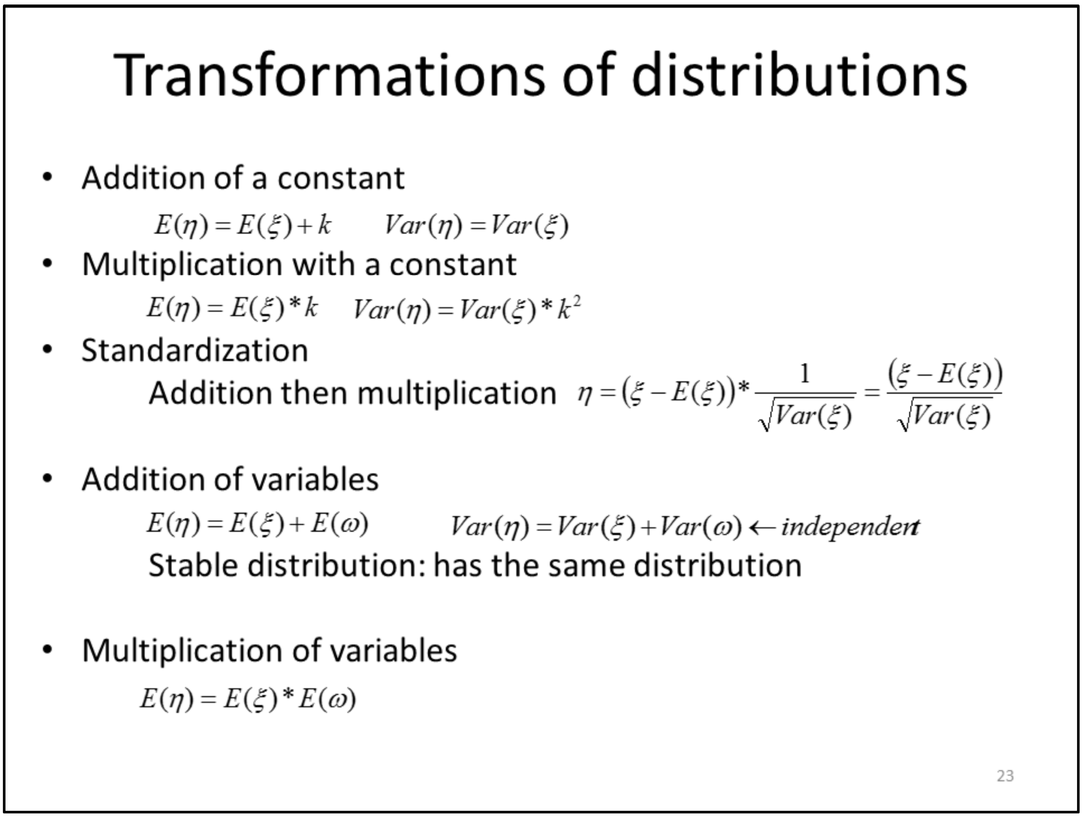
21

There are two other theoretical distribution that we have to mention:
The chi-square distribution and the (Student’s) t-distribution.
These distributions are often used in inferential statistics for hypothesis tests. This is a distribution family – the degrees of freedom (df) is the parameter that differentiate the distributions.
The chi-square value is the sum of the squares of independent, standard normal variables.
The expected value is 2*df.



Student’s t distributions

The Student’s t-distribution is one of the most frequently used distribution in inferential statistics.
This is a symmetric distribution family where the expected value is always 0, but the variances are different for different degrees of freedom.
If the df is infinite than the t-distribution = standard normal distribution, where the variance is 1.



Transformations of distributions

- Addition of a constant

  $E(\eta) = E(\xi) + k \qquad Var(\eta) = Var(\xi)$

- Multiplication with a constant

  $E(\eta) = E(\xi) * k \qquad Var(\eta) = Var(\xi) * k^2$

- Standardization

  Addition then multiplication $\eta = (\xi - E(\xi)) * \frac{1}{\sqrt{Var(\xi)}} = \frac{(\xi - E(\xi))}{\sqrt{Var(\xi)}}$

- Addition of variables

  $E(\eta) = E(\xi) + E(\omega) \qquad Var(\eta) = Var(\xi) + Var(\omega) \leftarrow independent$

  Stable distribution: has the same distribution

- Multiplication of variables

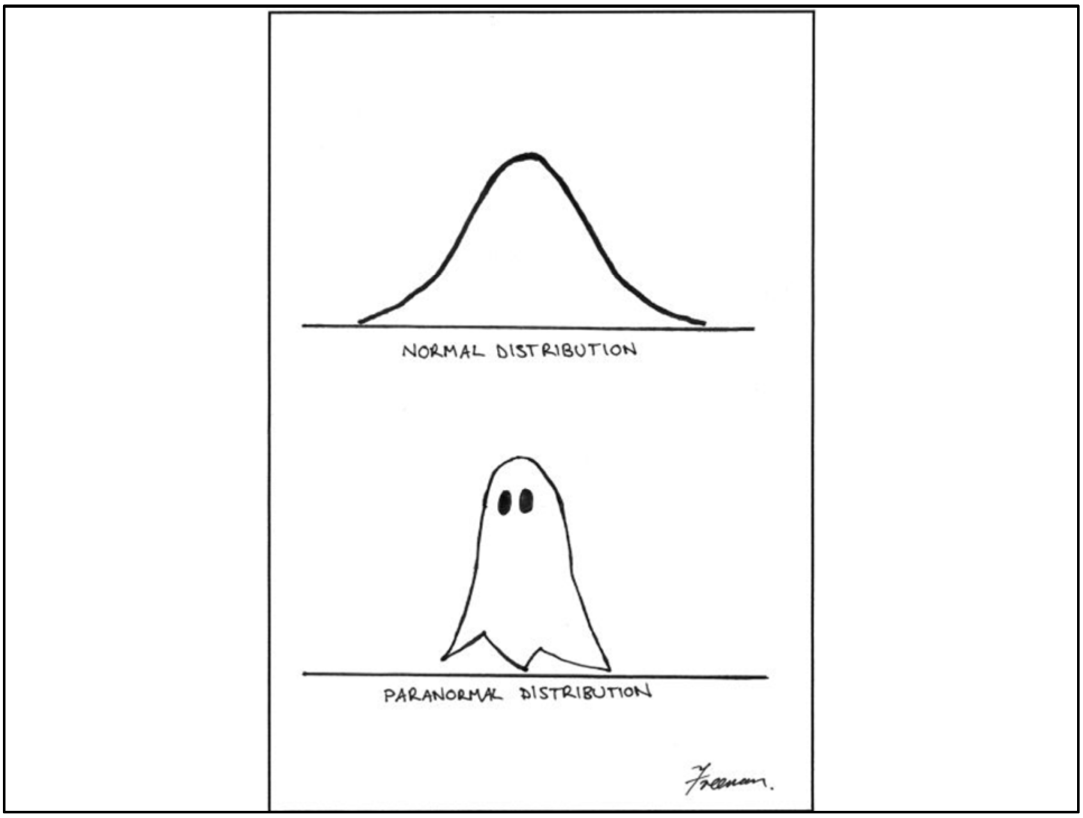
  $E(\eta) = E(\xi) * E(\omega)$

23

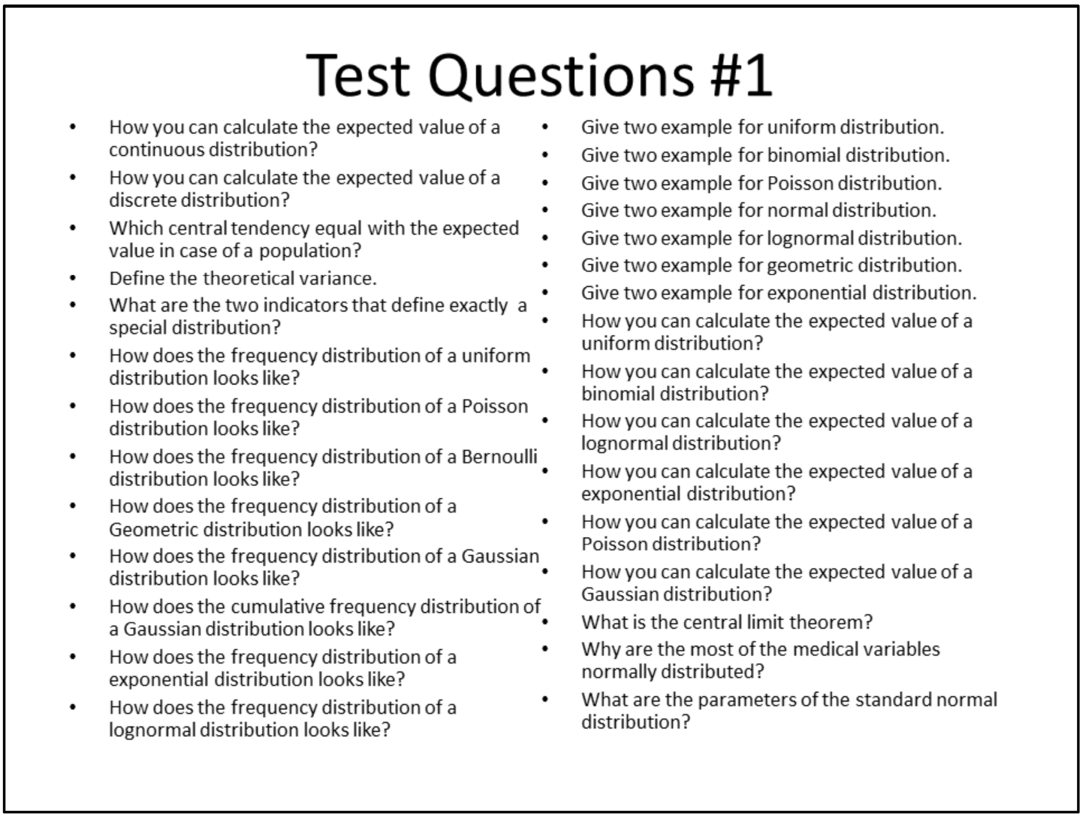
For an easier comparison we could transform our variables. Now I describe only a few transformation of normal distribution.

1. Addition of a constant
2. Multiplication with a constant
3. Standardization: sequence of addition then multiplication
4. Addition of variables
5. Multiplication of variables

See attached excel file.

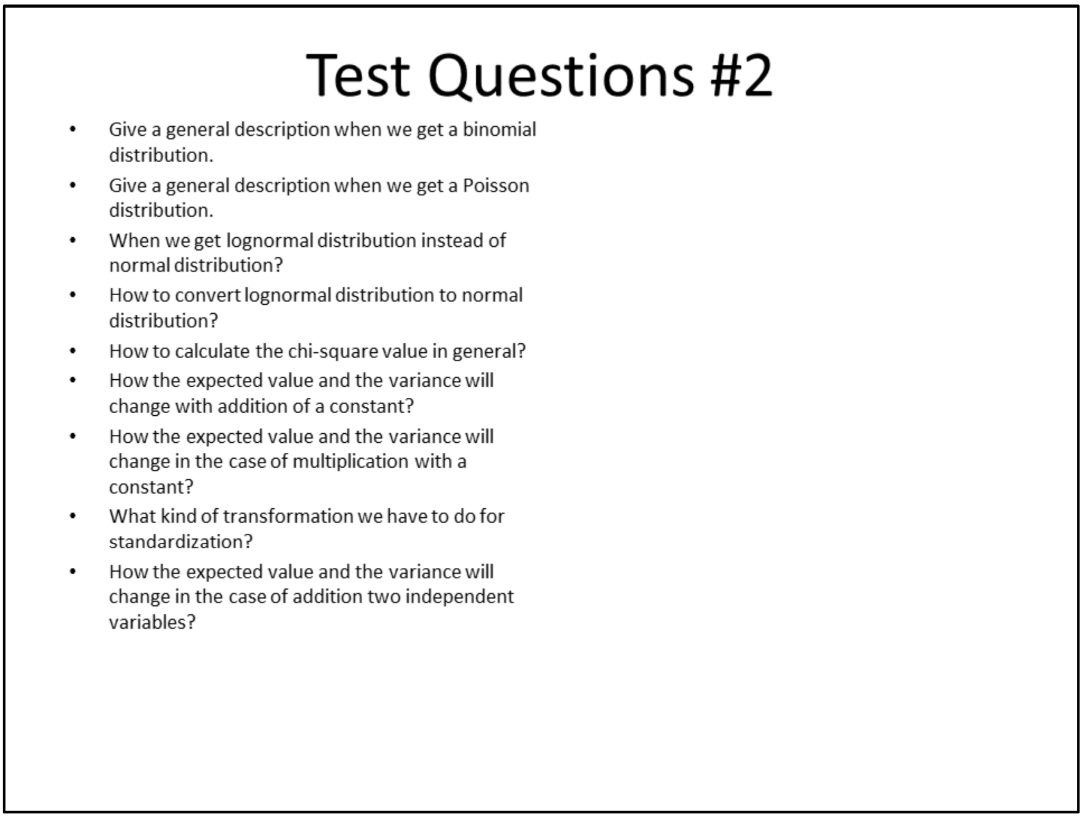


Not a statistical distribution….



Test Questions #1

- How you can calculate the expected value of a continuous distribution?
- How you can calculate the expected value of a discrete distribution?
- Which central tendency equal with the expected value in case of a population?
- Define the theoretical variance.
- What are the two indicators that define exactly a special distribution?
- How does the frequency distribution of a uniform distribution looks like?
- How does the frequency distribution of a Poisson distribution looks like?
- How does the frequency distribution of a Bernoulli distribution looks like?
- How does the frequency distribution of a Geometric distribution looks like?
- How does the frequency distribution of a Gaussian distribution looks like?
- How does the cumulative frequency distribution of a Gaussian distribution looks like?
- How does the frequency distribution of a exponential distribution looks like?
- How does the frequency distribution of a lognormal distribution looks like?
- Give two example for uniform distribution.
- Give two example for binomial distribution.
- Give two example for Poisson distribution.
- Give two example for normal distribution.
- Give two example for lognormal distribution.
- Give two example for geometric distribution.
- Give two example for exponential distribution.
- How you can calculate the expected value of a uniform distribution?
- How you can calculate the expected value of a binomial distribution?
- How you can calculate the expected value of a lognormal distribution?
- How you can calculate the expected value of a exponential distribution?
- How you can calculate the expected value of a Poisson distribution?
- How you can calculate the expected value of a Gaussian distribution?
- What is the central limit theorem?
- Why are the most of the medical variables normally distributed?
- What are the parameters of the standard normal distribution?



Test Questions #2

- Give a general description when we get a binomial distribution.
- Give a general description when we get a Poisson distribution.
- When we get lognormal distribution instead of normal distribution?
- How to convert lognormal distribution to normal distribution?
- How to calculate the chi-square value in general?
- How the expected value and the variance will change with addition of a constant?
- How the expected value and the variance will change in the case of multiplication with a constant?
- What kind of transformation we have to do for standardization?
- How the expected value and the variance will change in the case of addition two independent variables?