

Lecture 4

Model layouts

Normal distribution as a statistical model

Application of model layouts

Overview of model layouts

Annotation

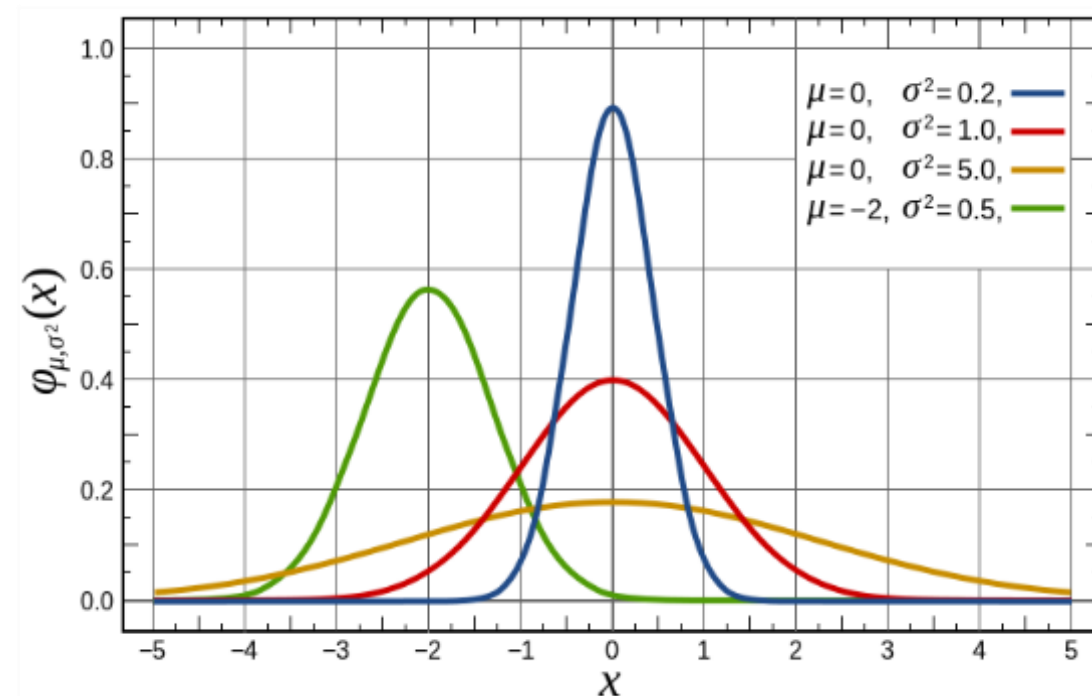
- The classical procedure of statistical analysis is to identify the type and characteristics of the model distribution of the data based on a sample of the target population, use its mathematical model to describe reality, and generalize the results to the target population under evaluation.
- The use of this approach is only possible if the real data match the model distribution, otherwise there is a risk of obtaining misleading results.
- The most classical model distribution from which many statistical analyses are derived is the normal distribution, also known as the Gaussian curve.

All models are wrong but some are useful.

George Box, 1978

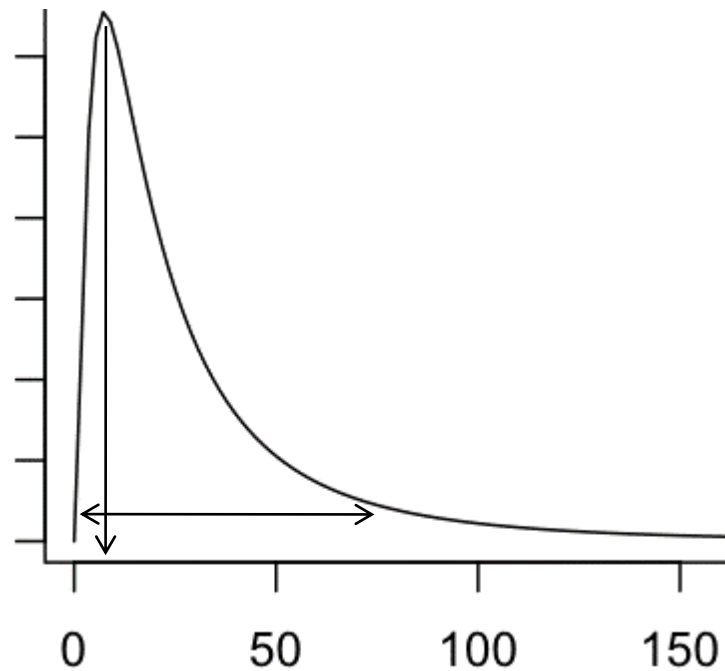
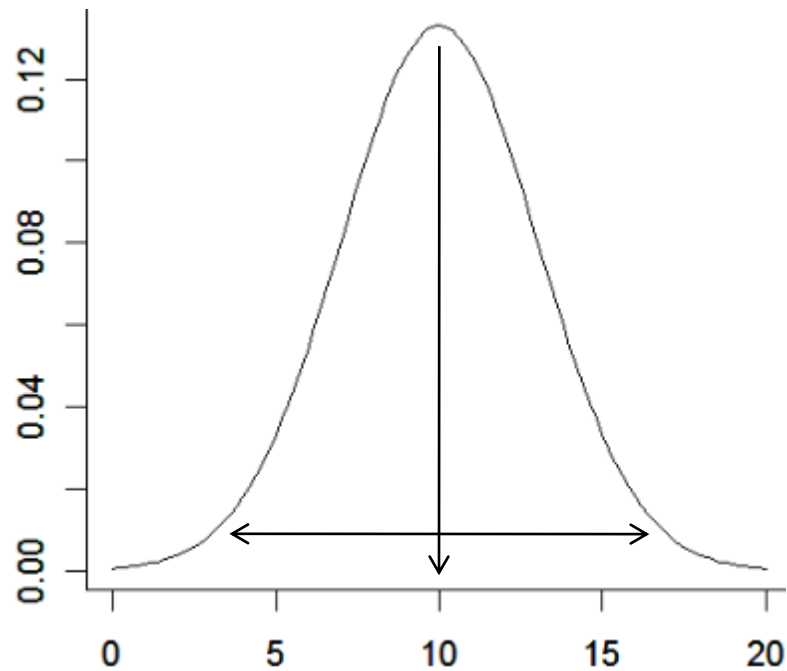
Normal distribution

- The most classic model distribution from which many statistical analyses are derived is the normal distribution, also known as the Gaussian curve.
- It describes the probability distribution of a continuous random variable: e.g., height in a population, measurement error...
- It is completely described by two parameters:
 - μ - mean value
 - σ^2 - dispersion
 - Designation: $N(\mu, \sigma^2)$
- Normality is a key assumption of many statistical methods
- There are a number of tests and graphical methods to verify normality



Description of the distribution of quantitative data: what do we want to describe about the data?

- Quantitative data - centre of gravity and range of observed values.



Calculation of the characteristics of the normal distribution: mean

- μ - average of the distribution (target population)
- \bar{x} - average of the sampled data distribution (estimate of the average of the target population)
- The mean can be calculated from arbitrary quantitative data, but only in some situations can it be considered an indicator of the mean of the data (symmetric, normal distribution of data)
- Outliers and data asymmetries significantly affect the result of the average calculation

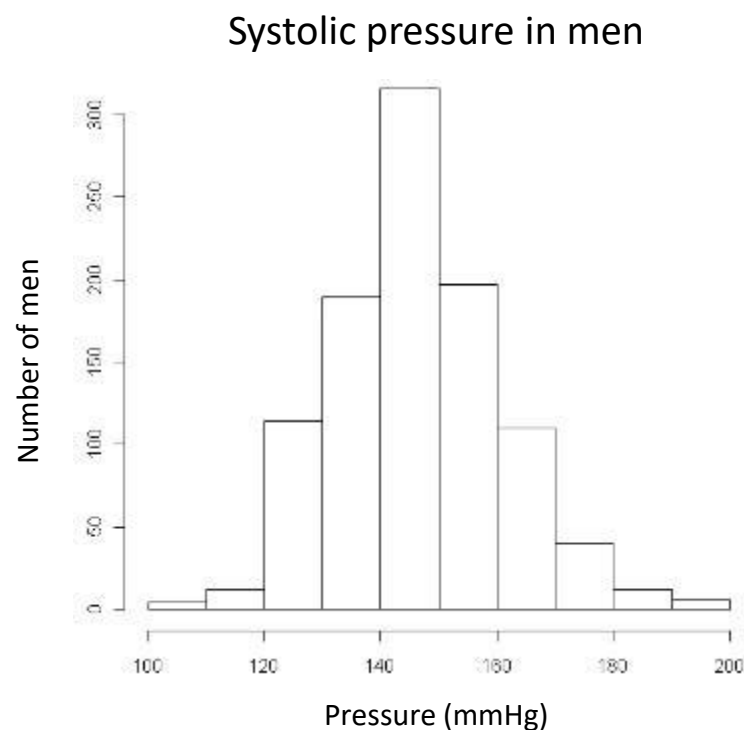
N=5

Object	Value
x_1	5
x_2	3
x_3	4
x_4	7
x_5	2

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{21}{5} = 4,2$$

Average vs. median

- If we have symmetric data, the result of calculating the mean and median is similar.
- Everything is OK.



➡ Diameter = 149.9 mmHg

➡ Median = 150.0 mmHg

Average vs. median

- If we do not have symmetric data, the result of calculating the mean and median is different.
- It's not OK. Calculating the average is inappropriate at this time!

- Example 1: grading at school

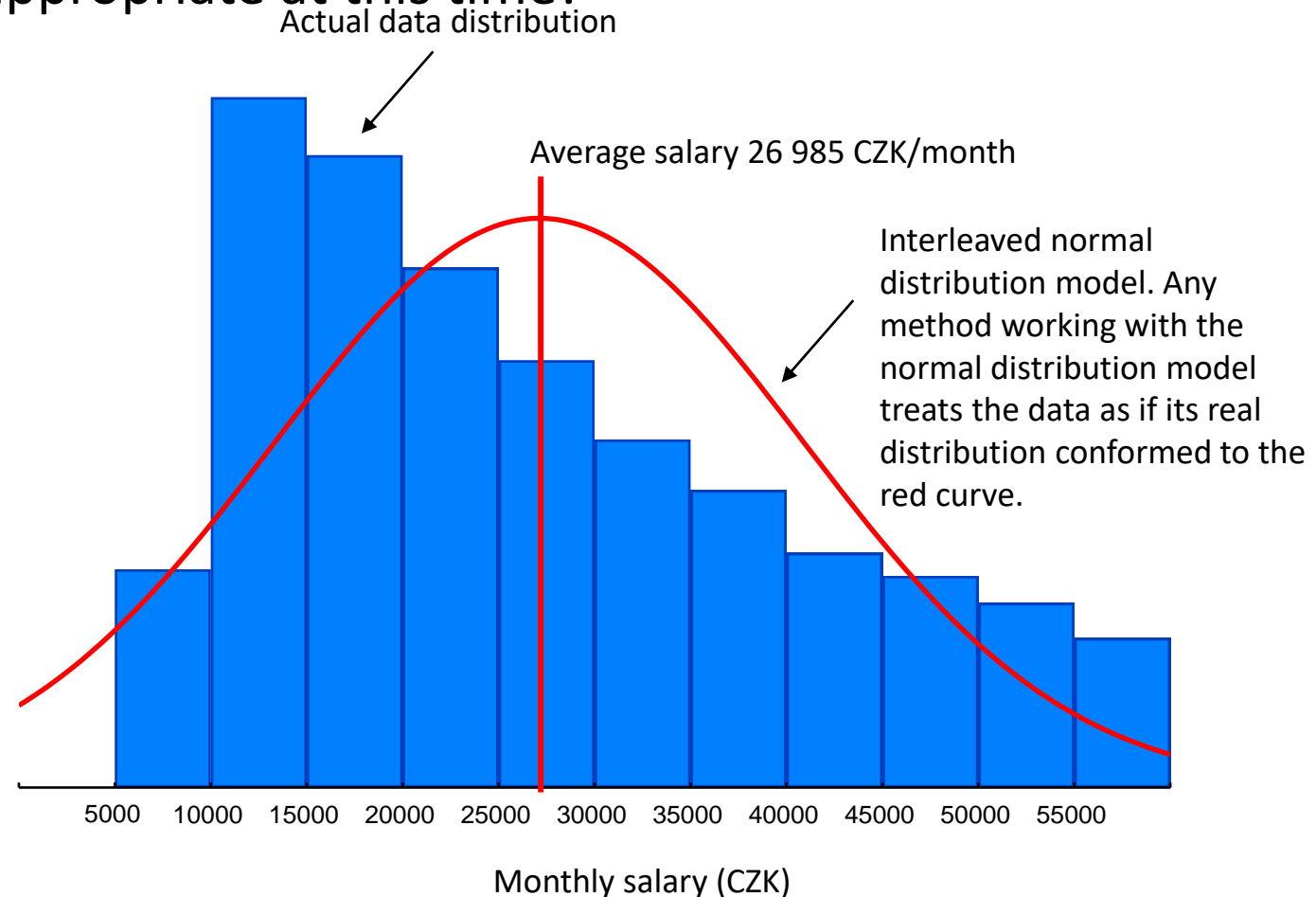
- Student A: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 5

Average = 1.35 Median = 1.00

- Student B: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2

Average = 1.13 Median = 1.00

- Example 2: salary in the Czech Republic



Description of the "centre of gravity" - position measure

- Let's have observed values: x_1, x_2, \dots, x_n
- Let's sort them by size: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

- **Minimum and maximum** - the smallest and largest observed values give us an idea of where we are on the x-axis.
- **Average** - characterizes the value around which the other observed values fluctuate. It is the physical image of the centre of gravity of equally massive points on the x-axis.
- **Median** - this is the middle observed value. It divides the observed values into two halves, half of the values are smaller and half of the values are larger than the median.

$$x_{\min} = x_{(1)}$$

$$x_{\max} = x_{(n)}$$

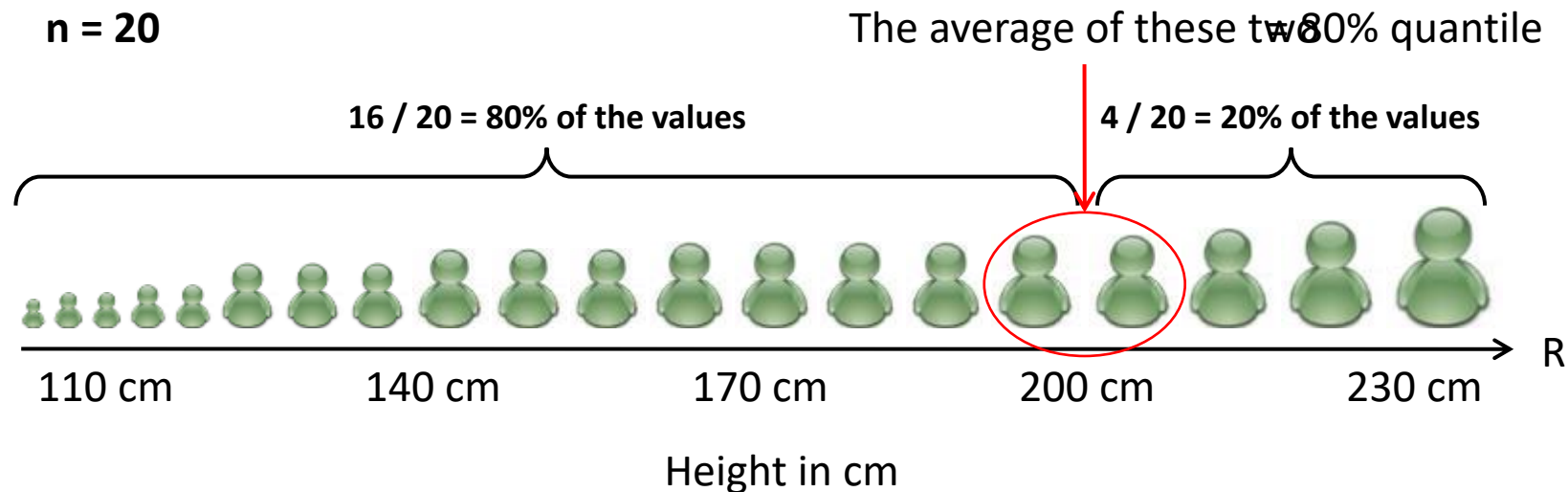
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\tilde{x} = x_{((n+1)/2)} \quad \text{for } n \text{ odd}$$

$$\tilde{x} = \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}) \quad \text{for } n \text{ even}$$

The concept of quantile

- In layman's terms, a quantile can be defined as a number on the real axis that divides the observed data into two parts: the $p\%$ quantile divides the data into $p\%$ values and $(100-p)\%$ values.
- We have a set of 20 people whose height we measure. We want to find the 80% quantile of the observed data set.

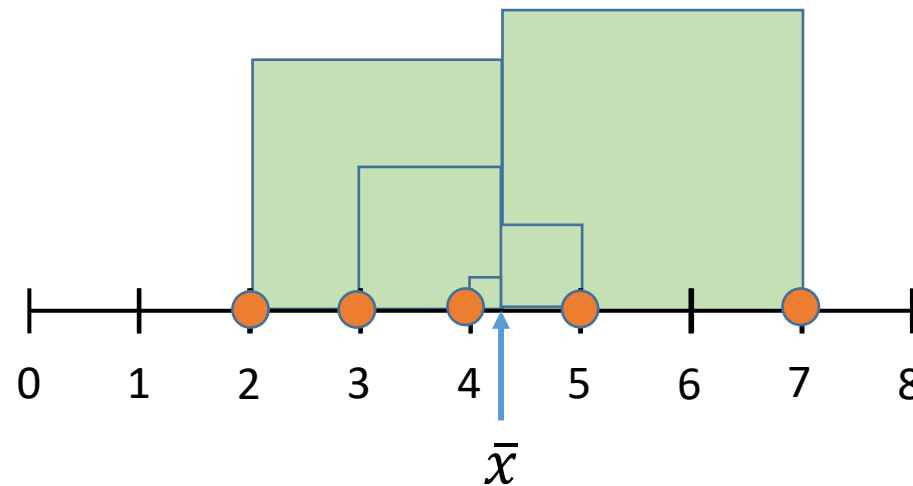


Calculation of the characteristics of the normal distribution: variance and standard deviation

- σ^2 - variance of the distribution (target population)
- s^2 - variance of the sampled data distribution (estimation of the variance of the target population)

N=5

Object	Value
x_1	5
x_2	3
x_3	4
x_4	7
x_5	2



$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1} = \frac{14,8}{4} = 3,7$$

$$s = \sqrt{s^2} = \sqrt{3,7} = 1,92$$

- Standard deviation (s, SD=standard deviation) = square root of the variance (easier to interpret)
- N-1 or N ? The N-1 division is a sample variance calculation, the N division is for the whole population (by exception)

Description of the "range" - the degree of variability

- The simplest characteristic of the variability of the observed data is the range of values (range) = maximum - minimum. It is easily influenced by atypical (outlying) values.
- **The quantile range** is defined by the p% quantile and the (100-p)% quantile and is less affected by outliers. A special case is the quartile spread, which covers 50% of the observed values.
- **Variance** - the average square of the deviation from the mean. Highly influenced by outliers.

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

- **Standard deviation** - the square root of the variance. The advantage of standard deviation is that it has the same units as the observed data.
- **Coefficient of variance** - the ratio of standard deviation to the mean (for a normal distribution, 95% of values should fit within the mean ± 3 SD), if the SD is greater than 1/3 of the mean, negative values in the distribution are theoretically likely - an indicator of problems with the normality of the data

Normal distribution: the effect of outliers on descriptive statistics

- The aim is to determine the average cholesterol level of a selected population of men (values in mmol/l)

Correct data	6.3	<div><div>Average value</div><div>6,32</div><div>Standard deviation</div><div>1,34</div><div>Which characteristic will increase more significantly?</div><div>Average or standard deviation?</div></div>	<div><div>Average value</div><div>?</div><div>Standard deviation</div><div>?</div></div>	<div><div>Incorrect data</div></div>	6.3
	7.6				7.6
	6.3				6.3
	9.1				9.1
	4.2				4.2
	5.8				5.8
	5.65				5.65
	6.3				6.3
	8.6				8.6
	6				6
	6.2				6.2
	6.7				6.7
	4.6				4.6
	6.25				6.25
	6.3				6.3
	4.04				4.04
	6.3				6.3
	9.1				9.1
	6.3				6.3
	5.2				5.2
	6.4				64
	5.75				5.75

Normal distribution: the effect of outliers on descriptive statistics

- The aim is to determine the average cholesterol level of a selected population of men (values in mmol/l)



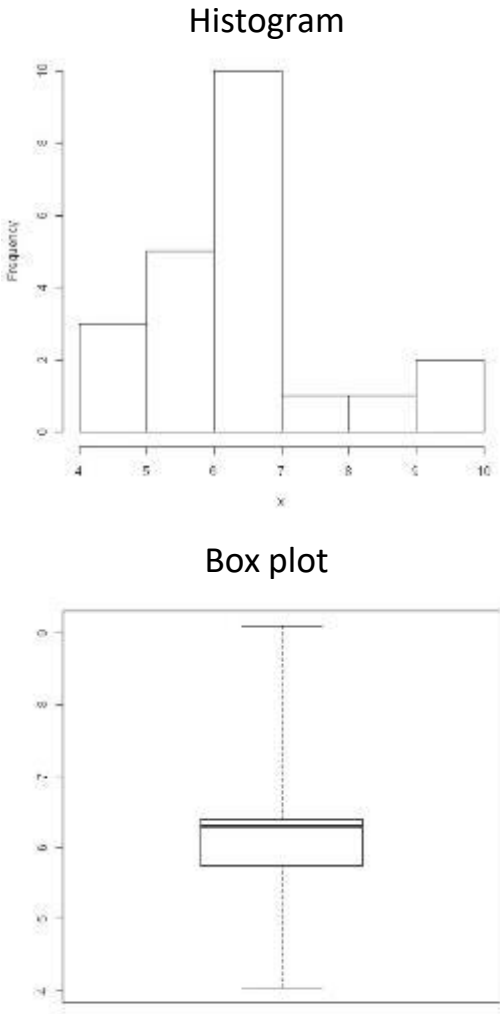
Identification of outliers

- On smaller files, visualization is sufficient.
- On larger datasets, visualization and descriptive statistics are not possible without.
- Graphical identification: using histogram and box plot.
- Identification using descriptive statistics: comparison of median and mean.

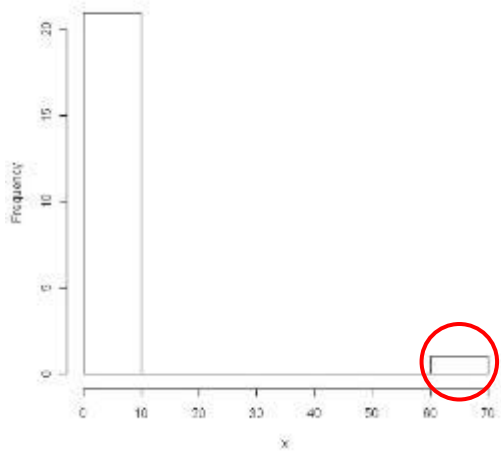
Identifying outliers - example

Correct data

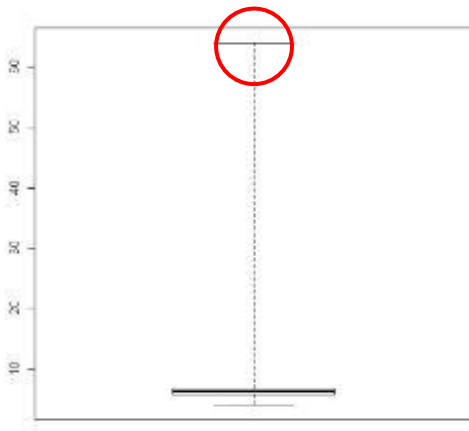
6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
6.4
5.75



Histogram



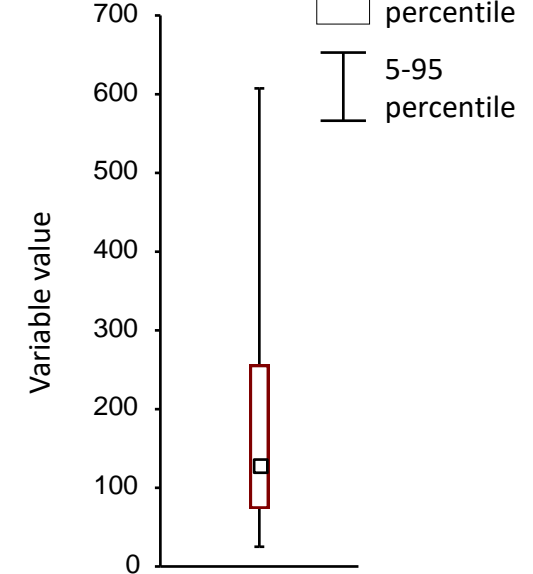
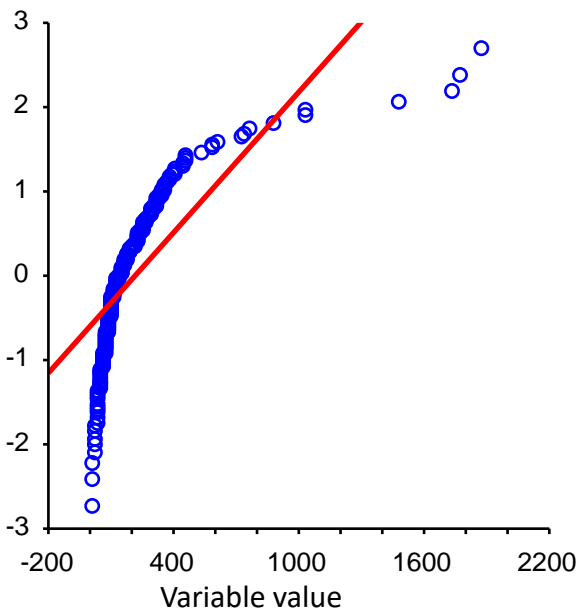
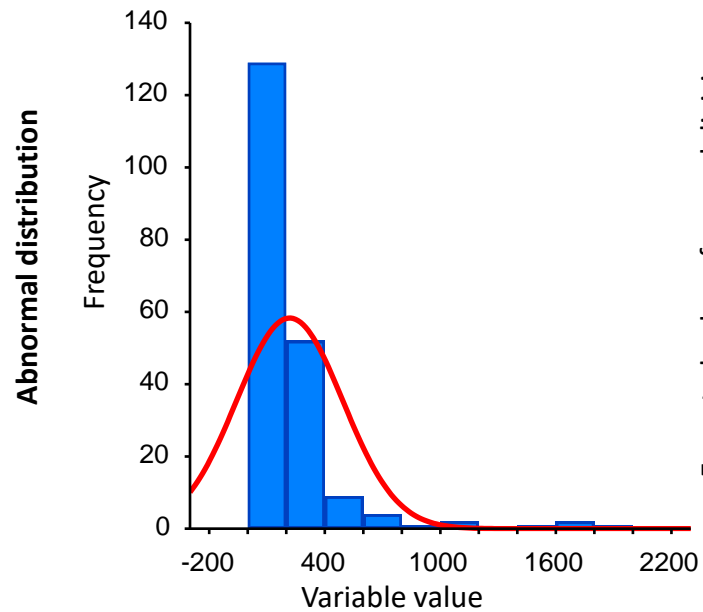
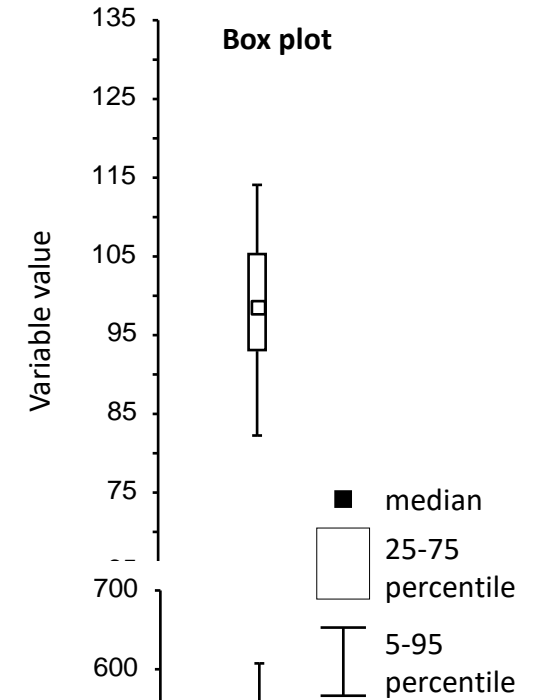
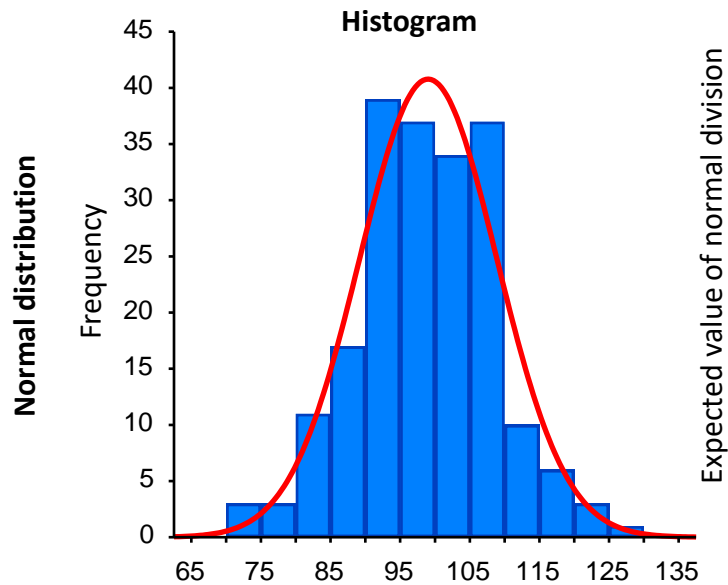
Box plot



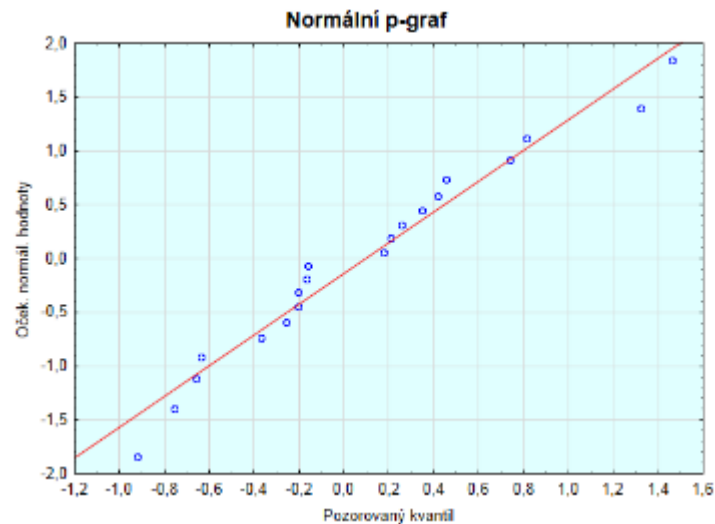
Incorrect data

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
64
5.75

Visual assessment of normality

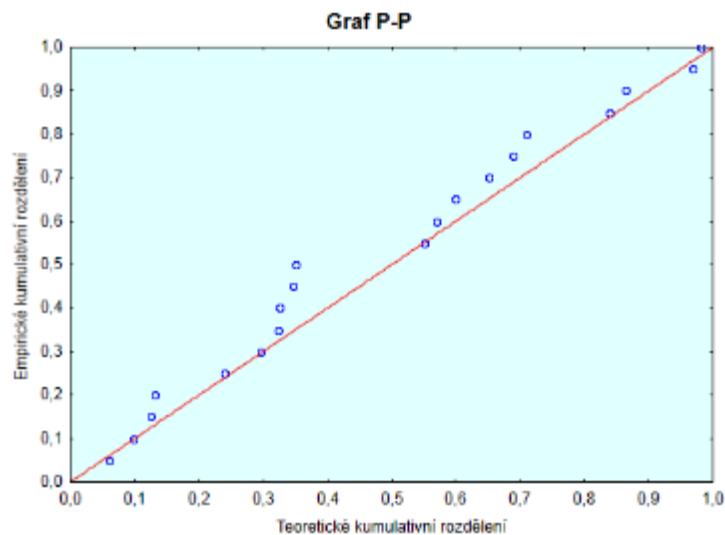
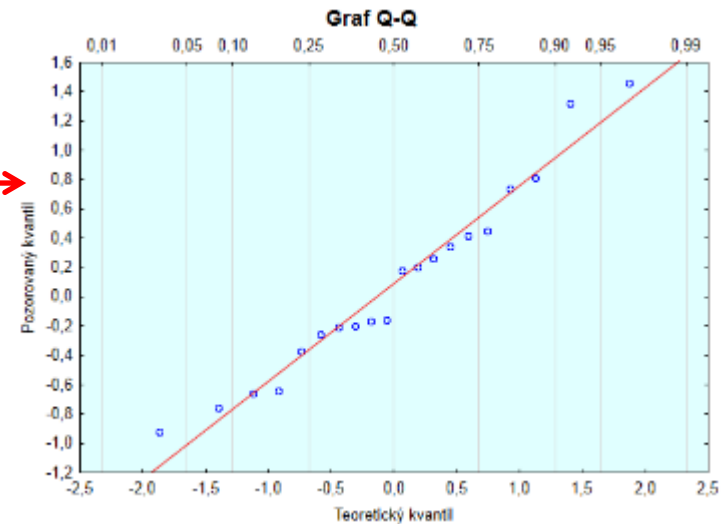


Difference between N-P, Q-Q, P-P chart



???

- Axle replacement only
- Observed and theoretical quantile shown



- Cumulative distribution plotted

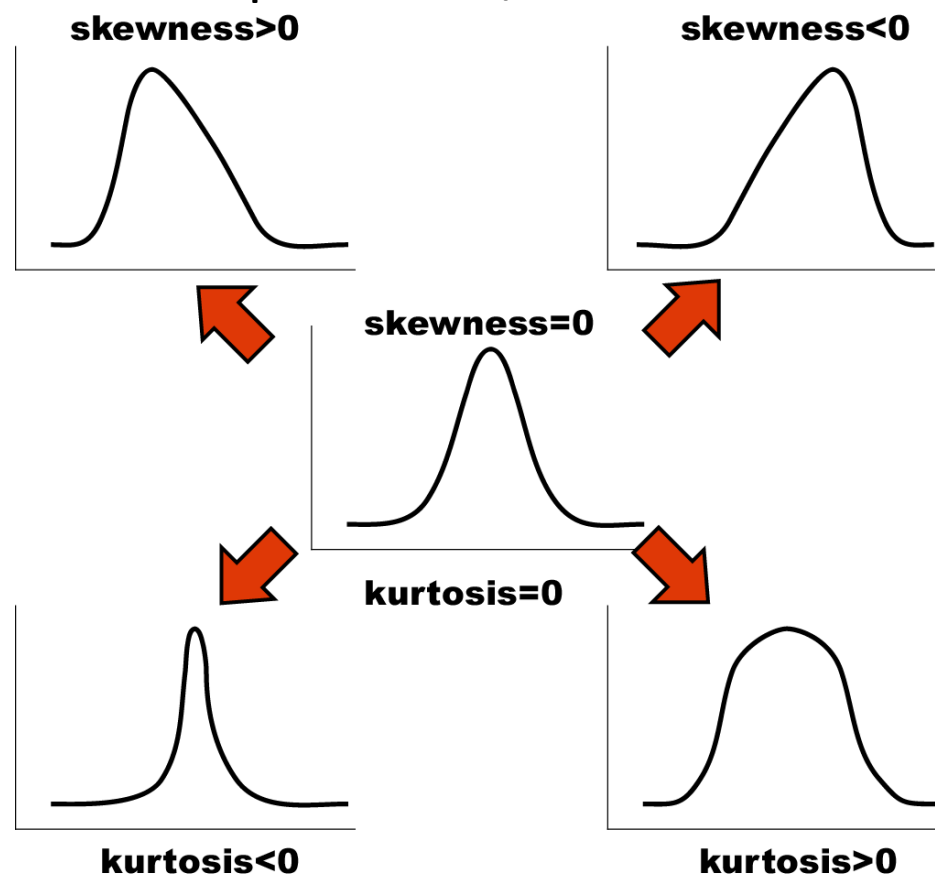
REMEMBER:

If the data come from a normal distribution, then the points will lie around the line

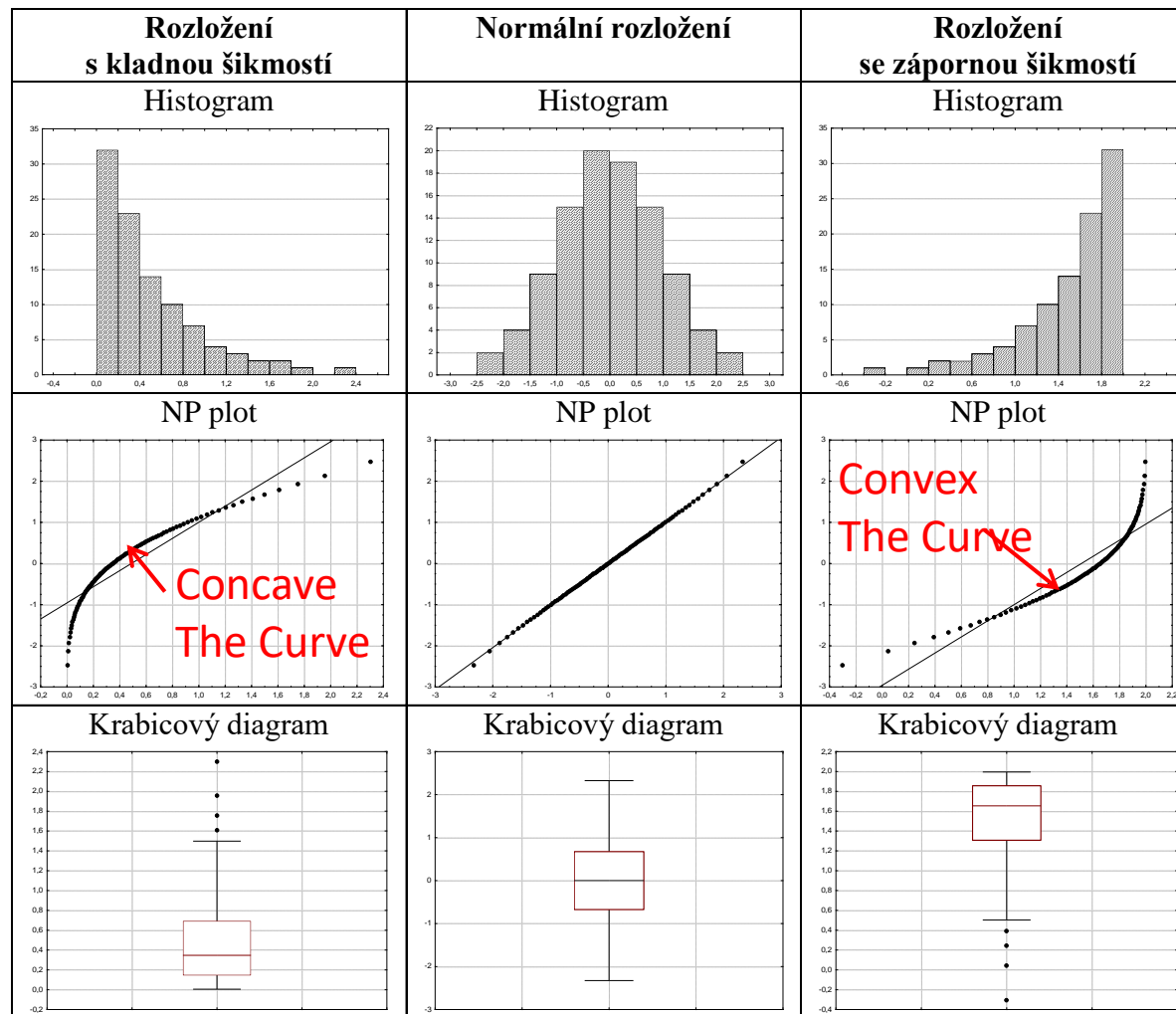


Layout shape indicators

- **Skewness** - indicator of the "skewness" of the distribution, asymmetry of the distribution
- **Kurtosis** - an indicator of the "peakiness/flatness" of the distribution



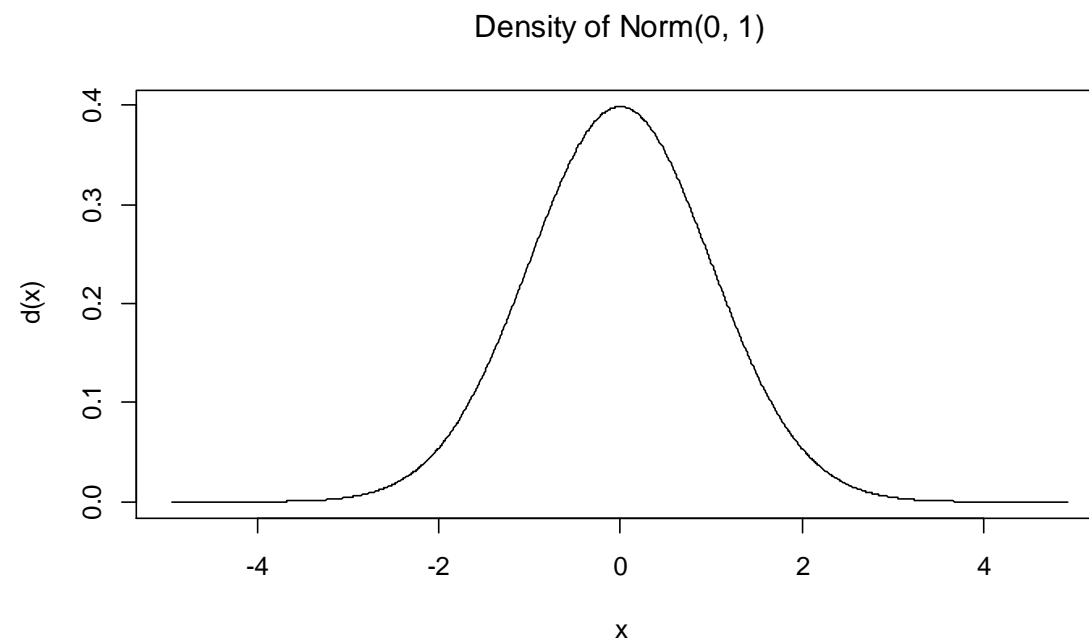
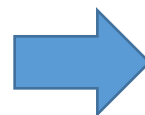
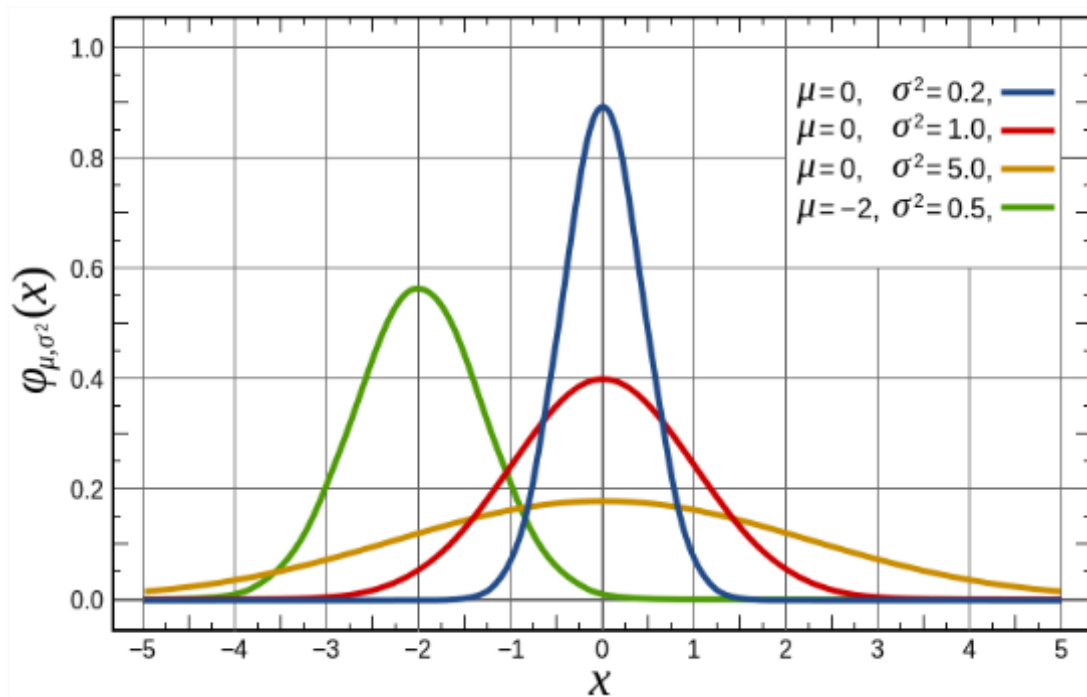
How does data asymmetry show up in diagnostic graphs?



Teaching materials: Computational
Statistics, RNDr. Marie Budíková, Dr., 2011

Standard normal distribution

- Special case of normal distribution with $N(\mu=0, \sigma^2=1)$ - standardized form used:
 - in statistical calculations
 - to compare the extremes/averages of values for variables with different ranges or units
 - Simple interpretation - basic values to remember



Conversion to standard normal distribution

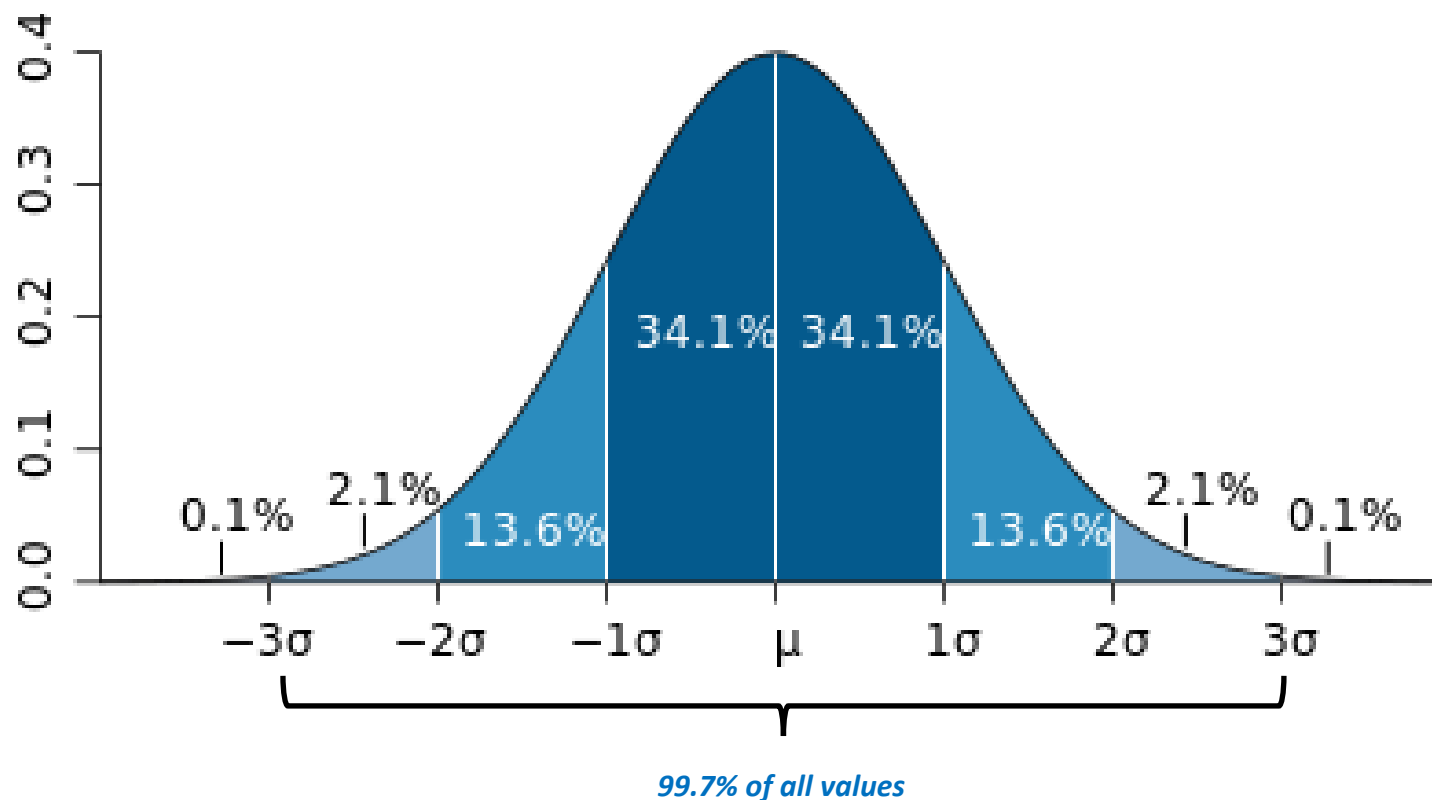
- So-called Z scores - in addition to statistical calculations, used e.g. in diagnostic scores (osteoporosis) or for comparing extremes/averages of variables with different ranges or units (e.g. pollutant measurements)
- Use in the calculation of standardized characteristics (e.g. covariance -> correlation coefficient)
- Used in multivariate analysis to achieve equal weighting of different variables in the calculation
- Tabulated form -> use in calculations

Object	Value	Standardised value (z)
x_1	5	0.42
x_2	3	-0.62
x_3	4	-0.10
x_4	7	1.46
x_5	2	-1.14
diameter	4,2	0
s	1,92	1

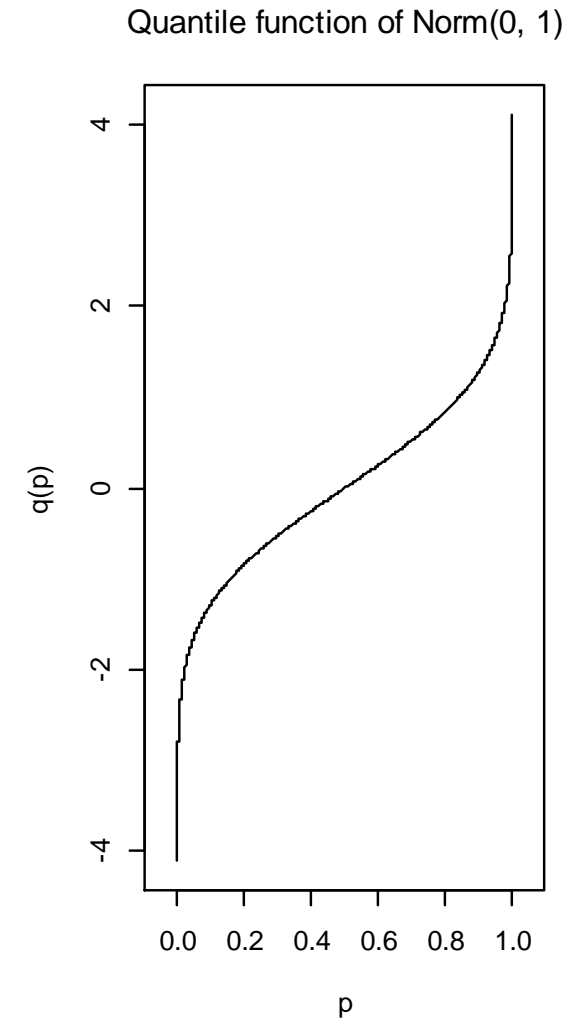
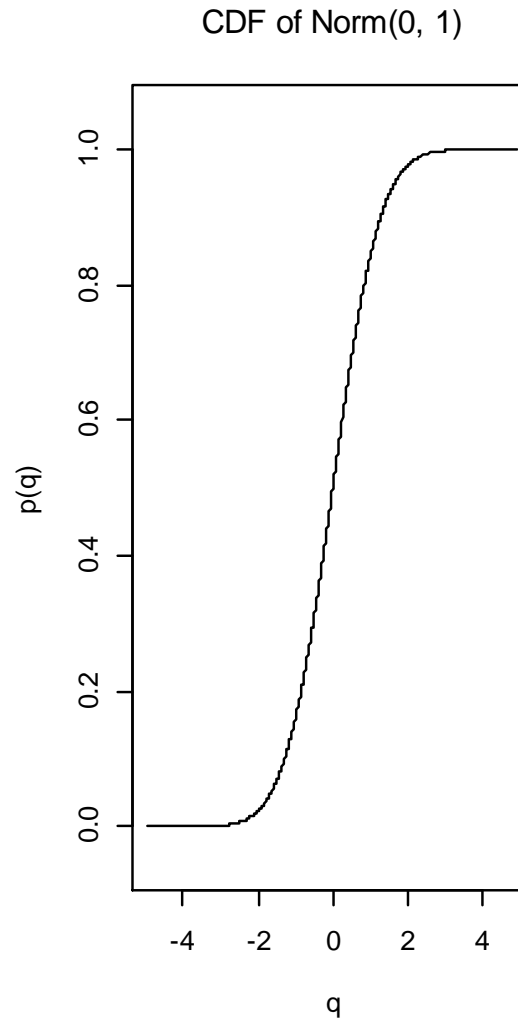
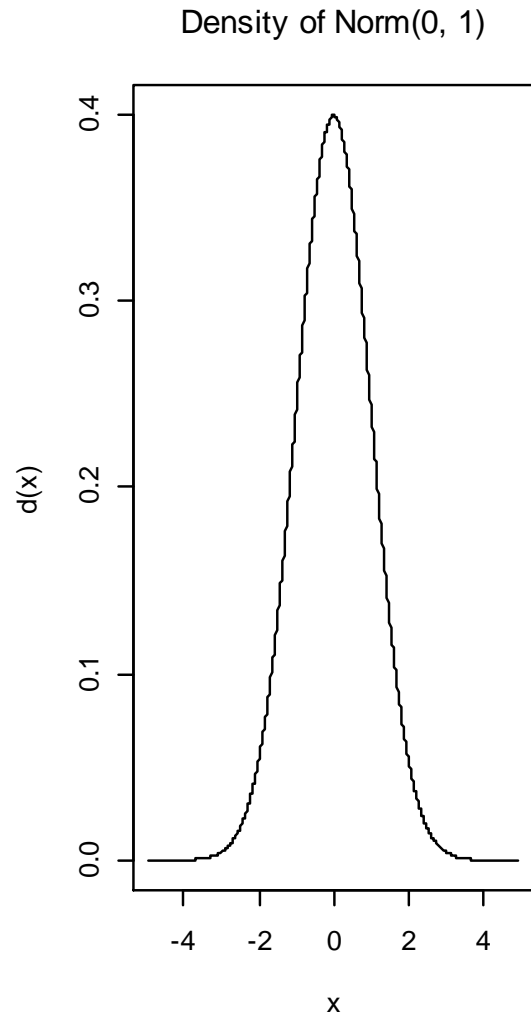
$$Z_i = \frac{x_i - \mu}{\sigma}$$

The 3 sigma rule

- Within $\mu \pm 3\sigma$, 99.7 % of all values should occur
- Useful to know for indicative assessment of the extent of the data
- For variables that cannot be negative use for indicative normality assessment

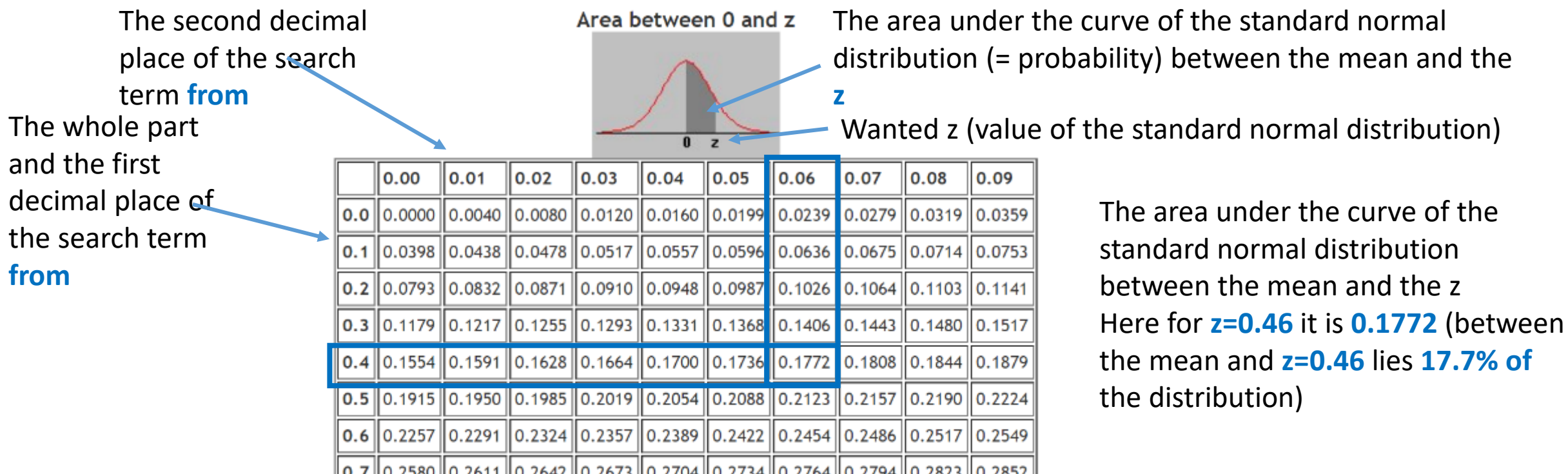


Standardized normal distribution and its characteristics



Statistical tables

- Clear expression of the distribution function for model distributions
- In the pre-computer era a basic aid, now mainly of educational importance
- <http://www.statsoft.com/Textbook/Distribution-Tables> (also needed for the exam)

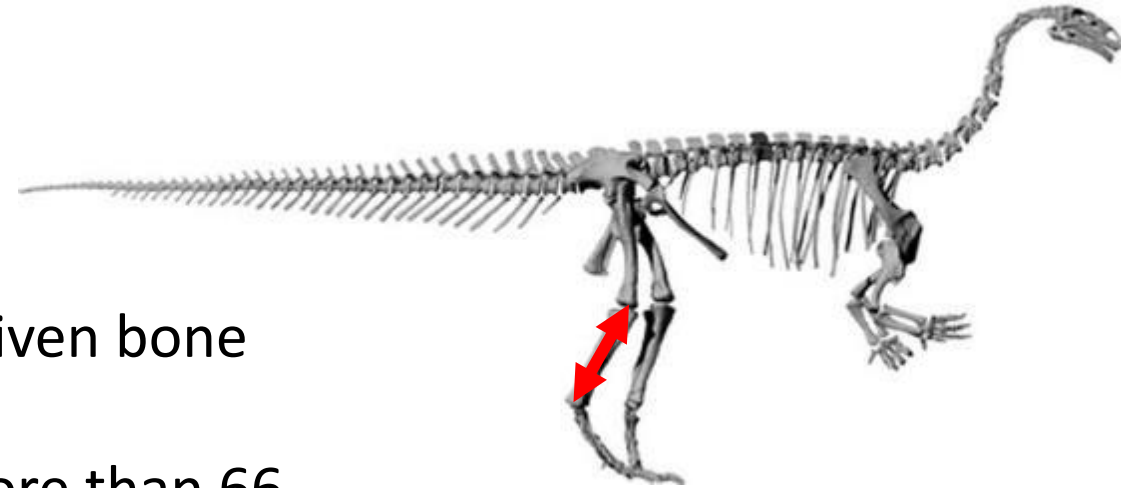


Use of statistical models

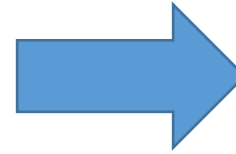
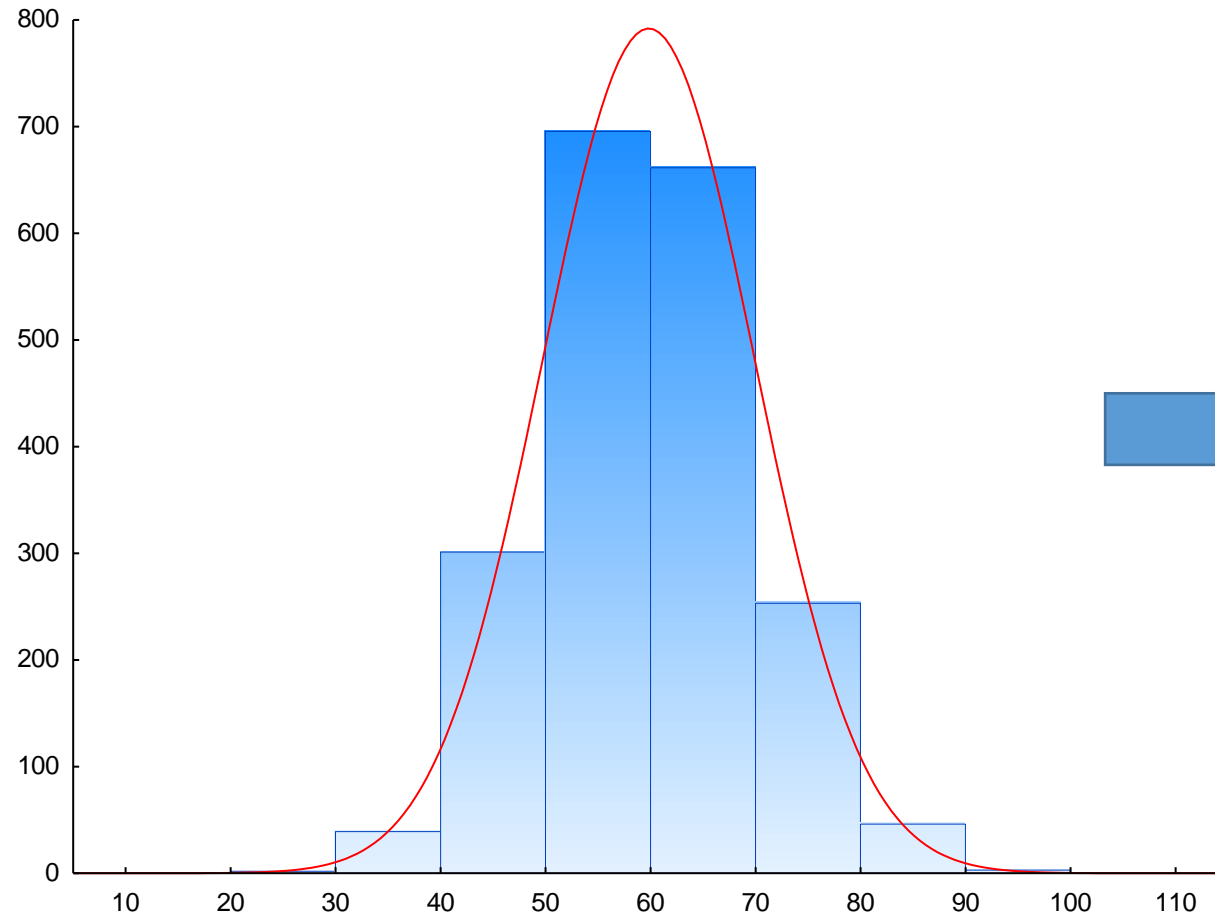
1. We have a trait in the population that we want to replace with a statistical model for analysis purposes (de facto we do this every time we calculate the mean, which we consider to be an indicator of the mean)
2. Verify the assumption that the trait is distributed according to the given model = **Is the selected model valid?** E.g. visual assessment of normality or testing it.
3. Calculate the characteristics of the model (mean and standard deviation in the case of a normal distribution)
4. Convert to the standard form of the model (standard normal distribution in the case of a normal distribution)
5. Use known properties of the distribution to answer the questions (distribution function, its values in statistical tables)

Example of application of the normal distribution model

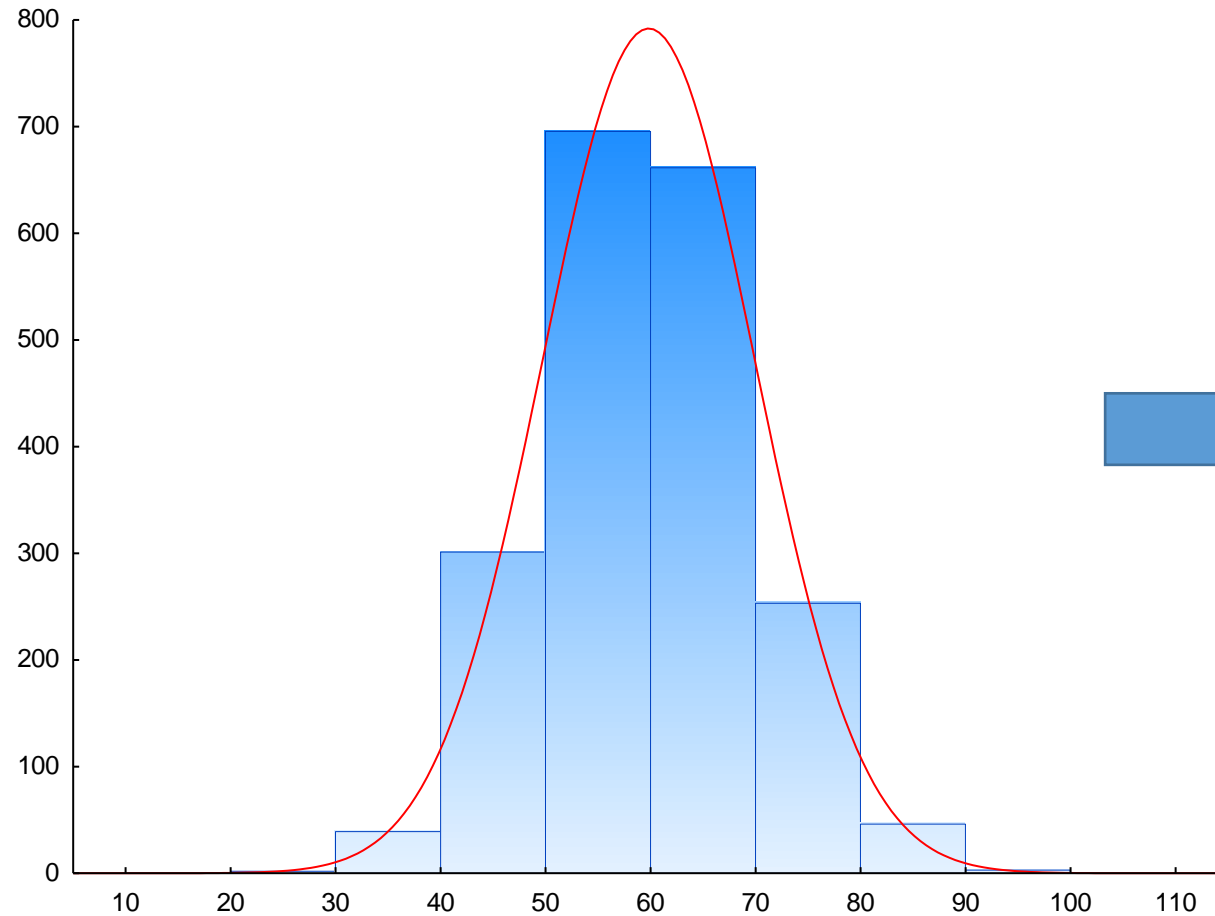
- We've got data from a bone survey of a prehistoric animal
 - $N=2\ 000$
 - Average length = 60 cm
 - Standard deviation = 10 cm
- Research questions:
 - What is the probability that the size of a given bone would exceed 66 cm?
 - How many of the bones were probably more than 66 cm in length ?
 - What proportion of the bones lay between 60 cm and 66 cm in length ?



Validation of data distribution and selection of statistical model



Validation of data distribution and selection of statistical model



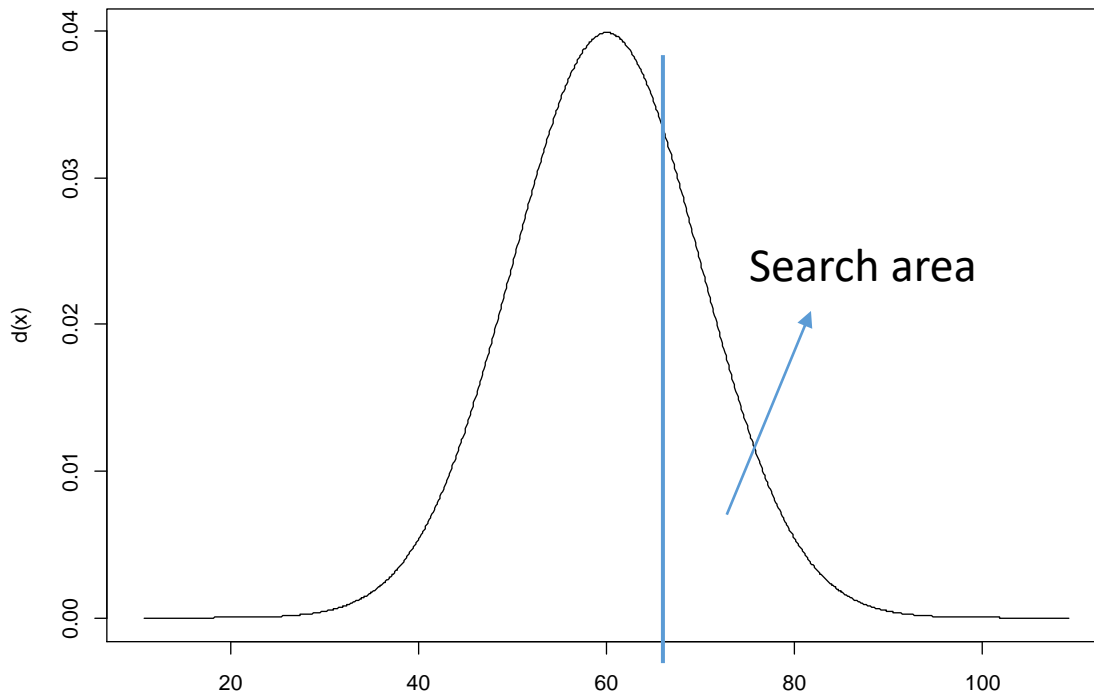
The assumption of a normal distribution of data seems justified.

What is the probability that the size of a given bone would exceed 66 cm?

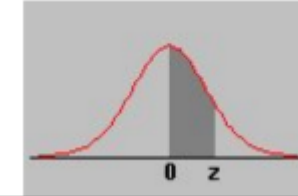
- Conversion of the search value to the standardized form of the normal distribution

$$z = \frac{x - \mu}{\sigma} = \frac{66 - 60}{10} = 0,6$$

Density of Norm(60, 10)



Area between 0 and z



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852

$$P(x > 66) = 1 - P(x \leq 66) = 1 - P\left(\frac{x - m}{s} \leq \frac{66 - 60}{10}\right) = 1 - F(0,6) = 0,27425$$

Application of the normal distribution model

- How many of the bones were probably more than 66 cm in length ?

$$P(x > 66) * n = 0,27425 * 2000 = 548$$

- What proportion of the bones lay within the range x of 60 cm to 66 cm in length ?

$$P(60 < x < 66) = P\left(\frac{60-60}{10} < Z < \frac{66-60}{10}\right) = F(0,6) - F(0) = 0,22575$$

- 22.6% of the bones lie in the range 60-66cm

Brief overview of model distributions I

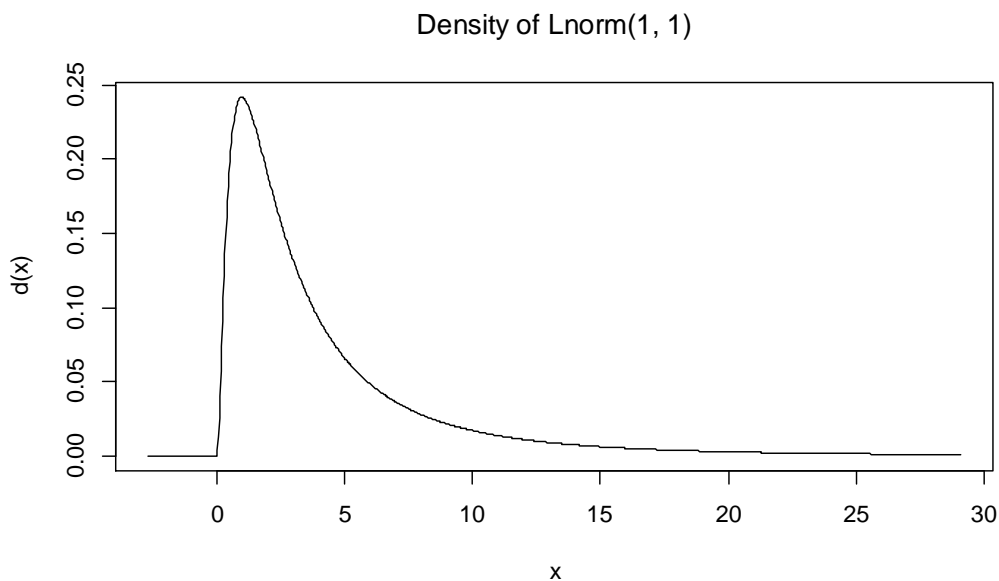
Layout	Parameters	Brief description
<u>Normal</u>	Diameter (μ) Scatter ($\omega\tau\eta$) ²	Symmetric function describing the interval frequency density; the most likely values are the mean values of the trait in the population.
<u>Log-normal</u>	Median Geometric diameter Scatter ($\omega\tau\eta$) ²	An interval frequency density function which, after logarithmic transformation, takes the form of a normal distribution.
Weibull	α - shape parameter β - value range parameter	By varying the parameter a , the distribution of survival time of, e.g., a stressed organism can be modelled. The distribution is also used as a model to estimate LC_{50} or EC_{50} in toxicity tests.
Steady	Median Geometric diameter Scatter ($\omega\tau\eta$) ²	An interval frequency density function which, after logarithmic transformation, takes the form of a normal distribution.
Triangular	$f(x) = [b - \text{ABS}(x - a)] / b^2$ $a - b < x < a + b$	Probability functions for the type of distribution where the mean values are significantly more likely than the marginal values. It allows flexible modelling of distribution
Gamma	Distribution function parameters: α - shape parameter β - value range parameter	functions of various shapes. For example, χ^2 distribution is a Gamma distribution. Gamma distribution with $a = 1$ is known as the exponential distribution.

Brief overview of model layouts

Layout	Parameters	Brief description
Beta	Distribution function parameters: α - shape parameter β - value range parameter	Probability function for a variable limited in range to the interval [0; 1]. It is mathematically more complicated, but very flexible in describing changes in the values of the variable in a bounded interval.
<u>Student's</u>	Degrees of freedom - considers sample size Average Scatter	Simulates the normal distribution for smaller samples of numbers. For larger sets ($n > 100$), it marginally approaches the normal distribution.
<u>Pearson's</u>	Degrees of freedom - considers sample size	It is mainly used to compare the frequencies of phenomena in two or more categories. It is used to model the distribution of the variance estimate of normally distributed data.
<u>Fisher-Snedecoroff</u>	Double degrees of freedom - considers the size of two samples	It is used to test the values of means - F test to compare two sample variances; F test, ANOVA etc.

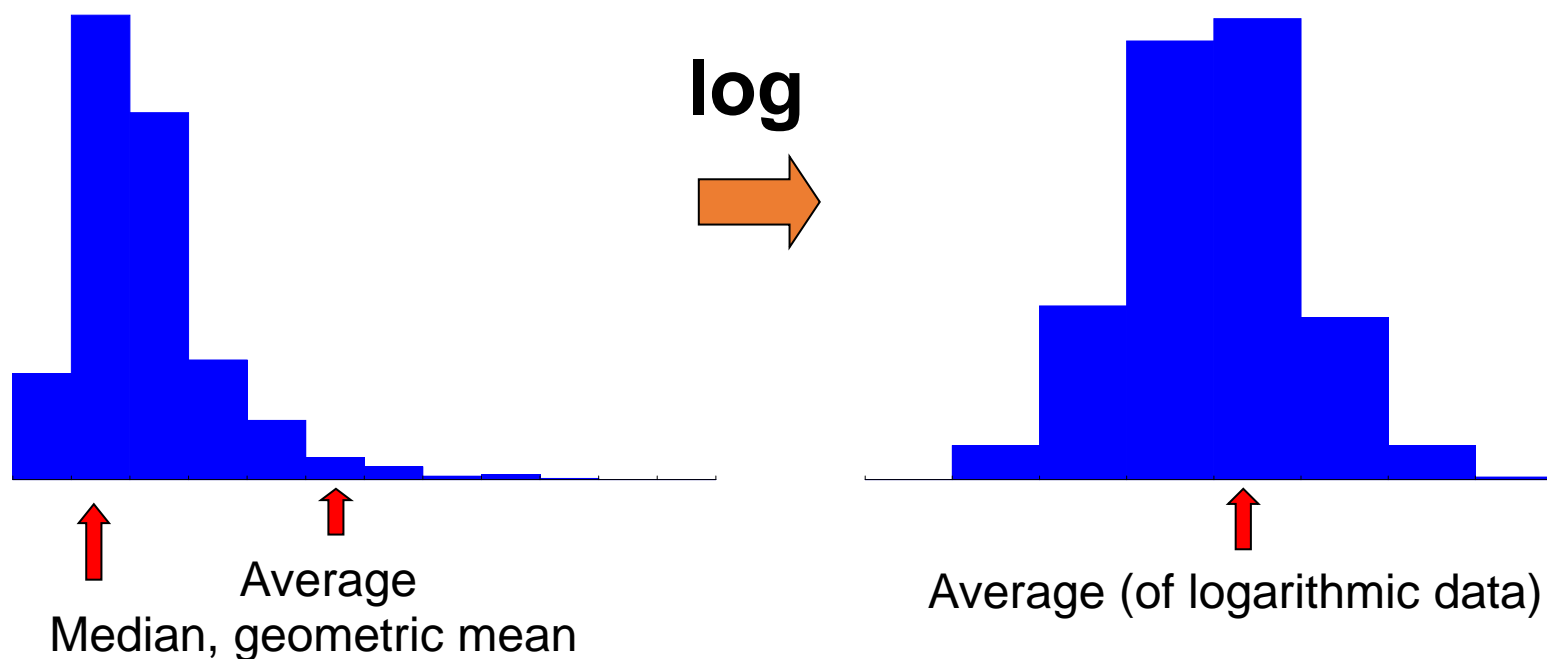
Lognormal distribution

- Asymmetrically distributed data - very common in biology (but also elsewhere, e.g. salaries)
- Along with the normal distribution, the most common model
- The geometric mean as an indicator of the centre is associated with the distribution

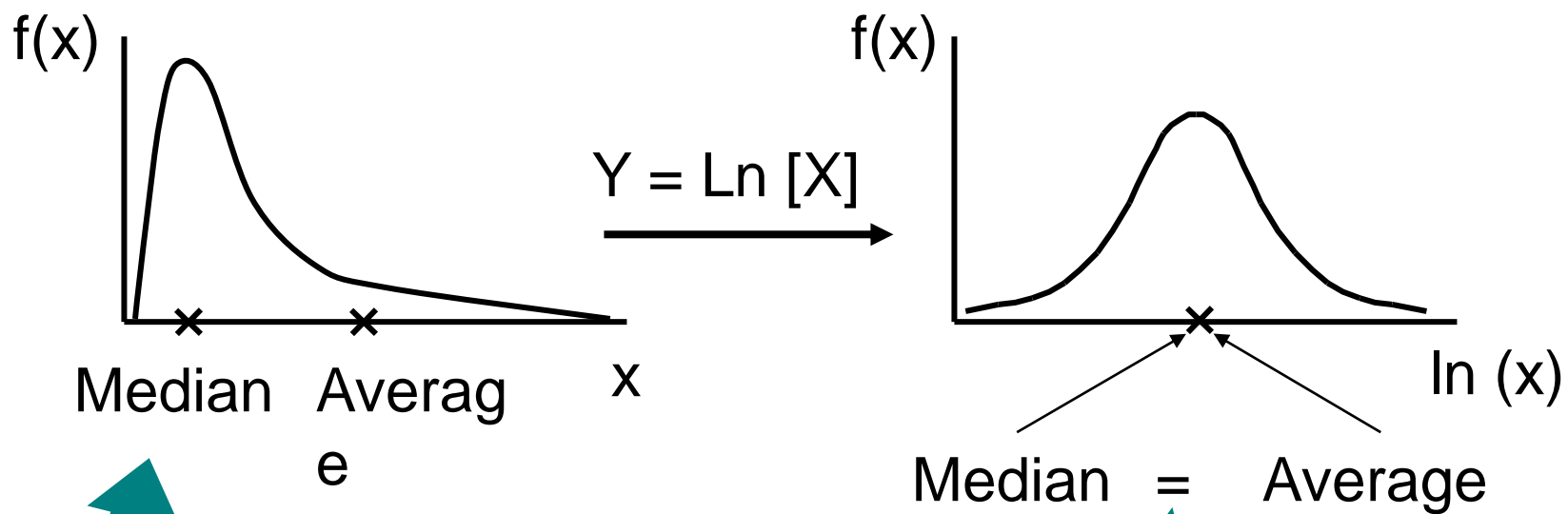


Logarithmic transformation

- Geometric mean - the antilogarithm of the mean of log-log data, it is suitable for left asymmetric data (lognormal distribution), which are very common in biology, its value basically corresponds to the median
- Such asymmetric data can be converted by logarithmic transformation to normal distribution



Geometric diameter



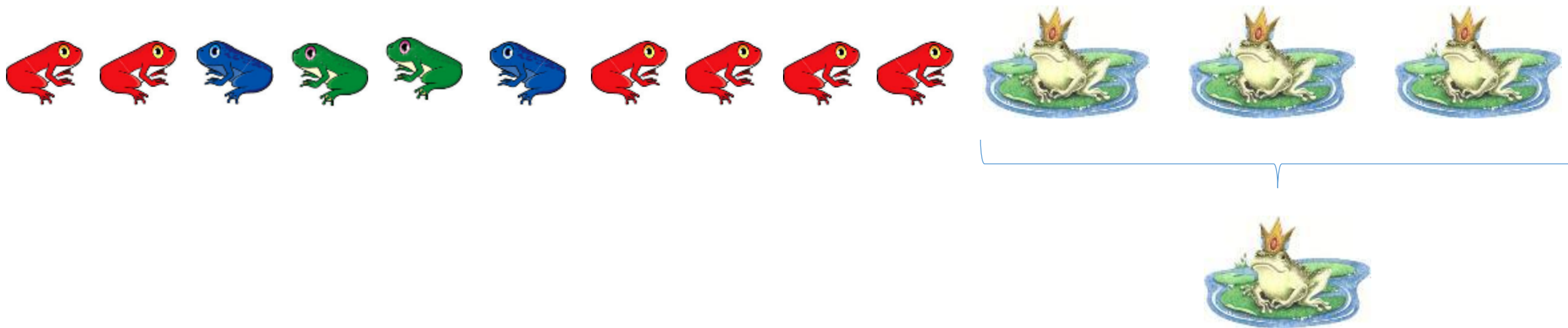
EXP (Y) = Geometric diameter X

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}$$

$\bar{\Psi} \pm \text{Standard error}$

Degrees of freedom

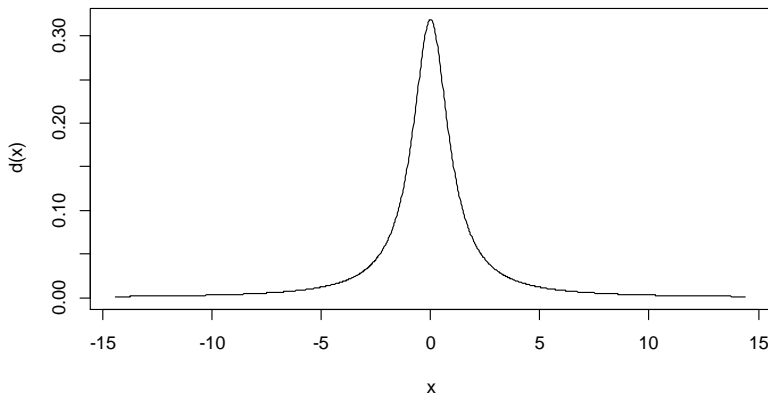
- Independent units of information
- Associated with the number of objects or groups in the data
- They decrease with the calculation of each summary statistic (= subtract the dependent statistics from the total number of dependent statistics)



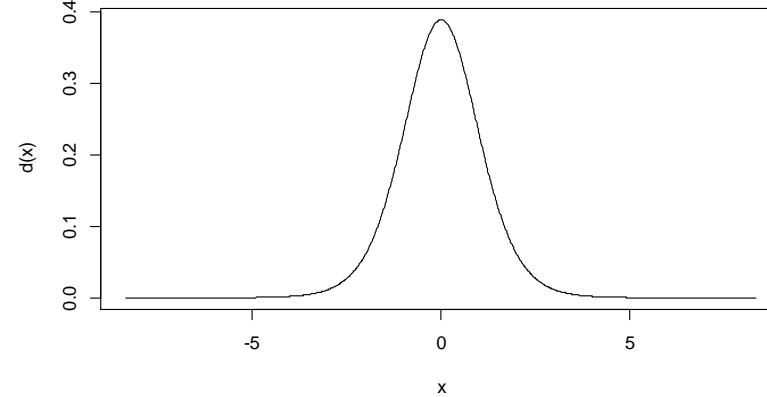
Student's division

- For a more realistic description of reality than the normal distribution allows
- Degrees of freedom - in relation to sample size

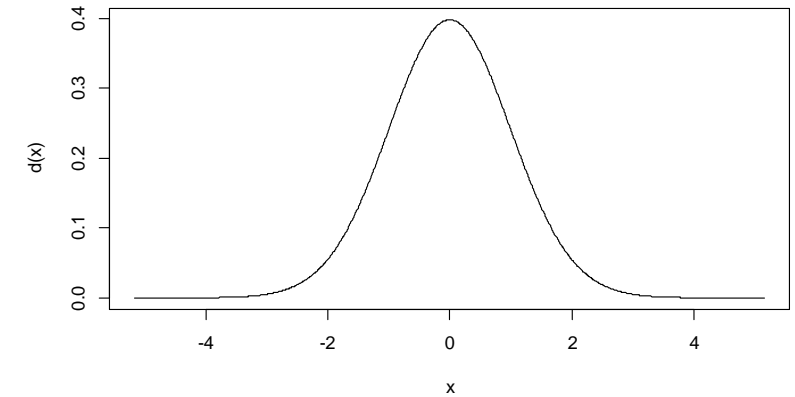
Density of $Td(1, 0)$



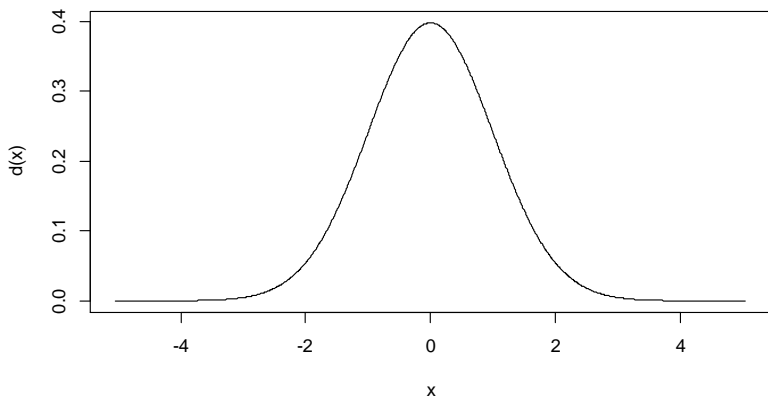
Density of $Td(10, 0)$



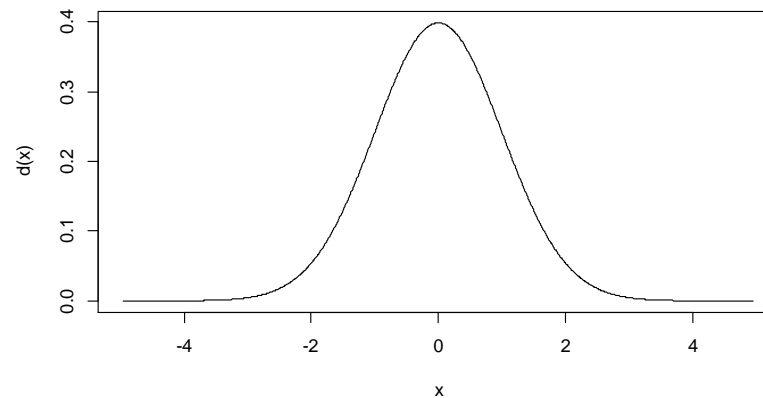
Density of $Td(100, 0)$



Density of $Td(200, 0)$



Density of $Td(1000, 0)$



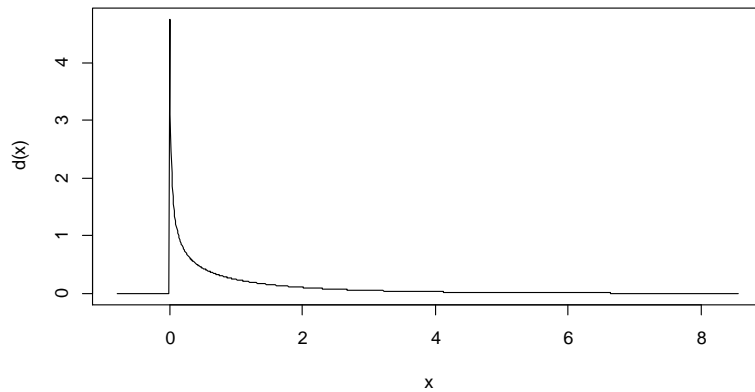
William Sealy Gosset

Publications under the pseudonym Student
t division based on experiments with yeast

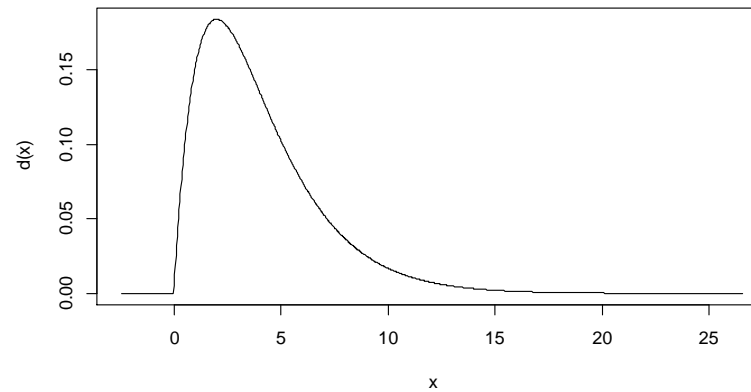
Pearson (Chi-square) distribution

- For data that can never be negative in principle
- Shape affected by degree of freedom
- Expected and observed counts, variances
- Often in genetics

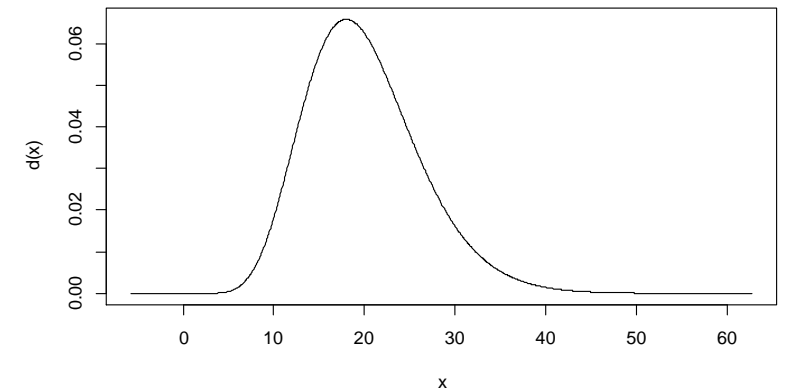
Density of Chisq(1, 0)



Density of Chisq(4, 0)



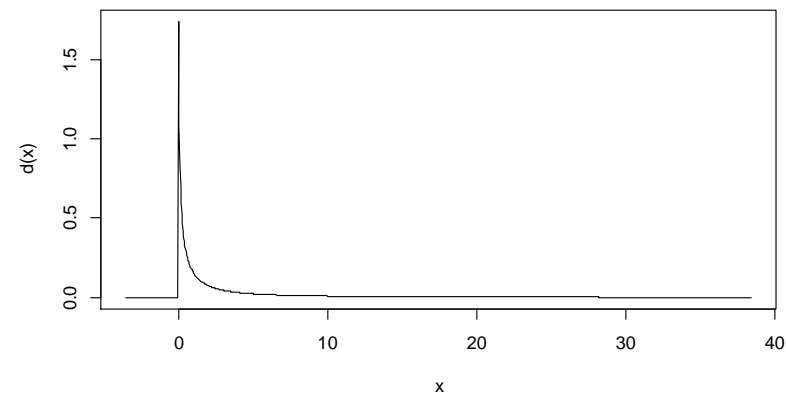
Density of Chisq(20, 0)



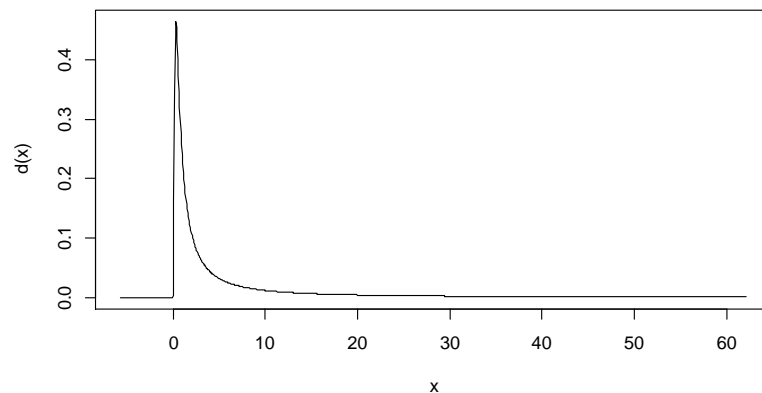
Fisher-Snedecor distribution

- For data that can never be negative in principle
- Typically a ratio of two variances - use in many, especially more advanced statistical tests
- Two different degrees of freedom

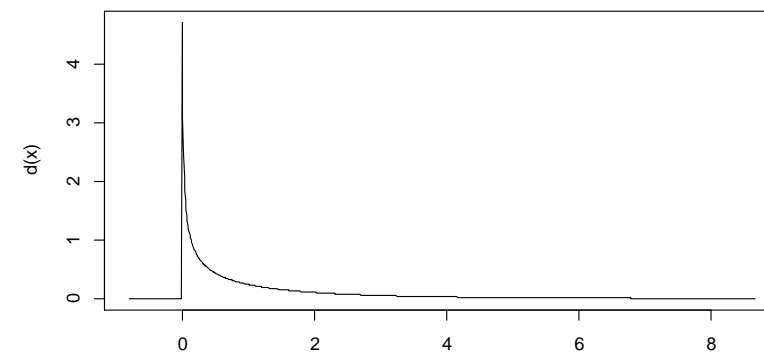
Density of $Fd(1, 1, 0)$



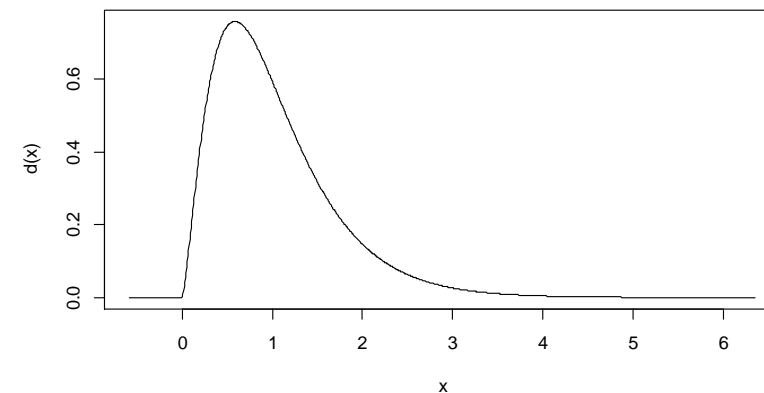
Density of $Fd(100, 1, 0)$



Density of $Fd(1, 100, 0)$



Density of $Fd(5, 100, 0)$



Data Transformation - Legitimate Layout Adjustment

- **Basic types of transformations lead to normality of distribution or homogeneity of variance**
- **Logarithmic transformation**
- The logarithmic transformation is very suitable for data with outliers at the upper end of the range. When comparing the means of multiple data sets, the situation where the standard deviation changes proportionally with increasing mean, and thus the variables have the same coefficient of variation even though they have different means, is indicative of this transformation.
- In such a situation, the logarithmic transformation not only weakens the asymmetry of the original distribution, but also brings a higher homogeneity of the variance of the variables. The natural logarithm is most often used for the transformation and if there are zero values in the original data set, the operation $Y = \ln(X+1)$ is appropriate.
- When the average of the logarithmized data (i.e. the average logarithm) is transformed back to the original values, the result is not the arithmetic but the geometric mean of the original data.

Data Transformation - Legitimate Layout Adjustment

- **Basic types of transformations lead to normality of distribution or homogeneity of variance**
- **Square root transformation**

- The transformation is suitable for variables having a Poisson distribution, i.e. variables expressing the total number of occurrences of a certain (rather rare) phenomenon in n independently repeated experiments. More generally, this type of transformation may be recommended in the case of normalization of data of the type of number of individuals (cells, etc.). It is a transformation:

- $$Y = \sqrt{x} \quad \text{or} \quad Y = \sqrt{x+1} \quad Y = \sqrt{x} + \sqrt{x+1}$$

- Transformations with the value 1 added are efficient if X takes very small or zero values. A situation indicating the appropriateness of the square root transformation is also the proportionality of the sample variance and the mean, that is, in general if $s^2x = k$ (sample mean).

Data Transformation - Legitimate Layout Adjustment

- **Basic types of transformations lead to normality of distribution or homogeneity of variance**
- **Arcsin transformation**
- The so-called angular transformation - very suitable for data of the type of proportions of occurrence of a certain phenomenon (trait) among n evaluated individuals - i.e. for data having a binomial distribution. If a trait occurs r times among n possibilities (individuals, repetitions), then the relative frequency of its occurrence can be expressed as $p = r/n$ with a variability of $p(1-p)/n$. The Arcsin transformation removes proportions close to 0 or 1 from the data sets, effectively reducing the variability of the mean estimates. However, the transformation is not able to remove the variability caused by the different number of repetitions in each variant, in which case weighted transformations of the data may be recommended. A very common form of this transformation is:
$$Y = \arcsin \sqrt{p}$$
- - i.e. the transformation of the shares into values whose sine is equal to the square root of the original values. If the total number of individuals (repetitions) among which the occurrence of a trait is monitored is $n < 50$, then very efficient empirical measures can be recommended to transform the proportions close to 0 or 1. For this case, zero proportions can be replaced by $1/4n$ and 100% proportions by $(n-1/4)/n$. If there are more extreme values between the values (less than 0.2 and greater than 0.8), a transformation can be recommended:

$$Y = \frac{1}{2} \left[\arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}} \right]$$

Lecture 5

Making estimates

Point and interval estimates

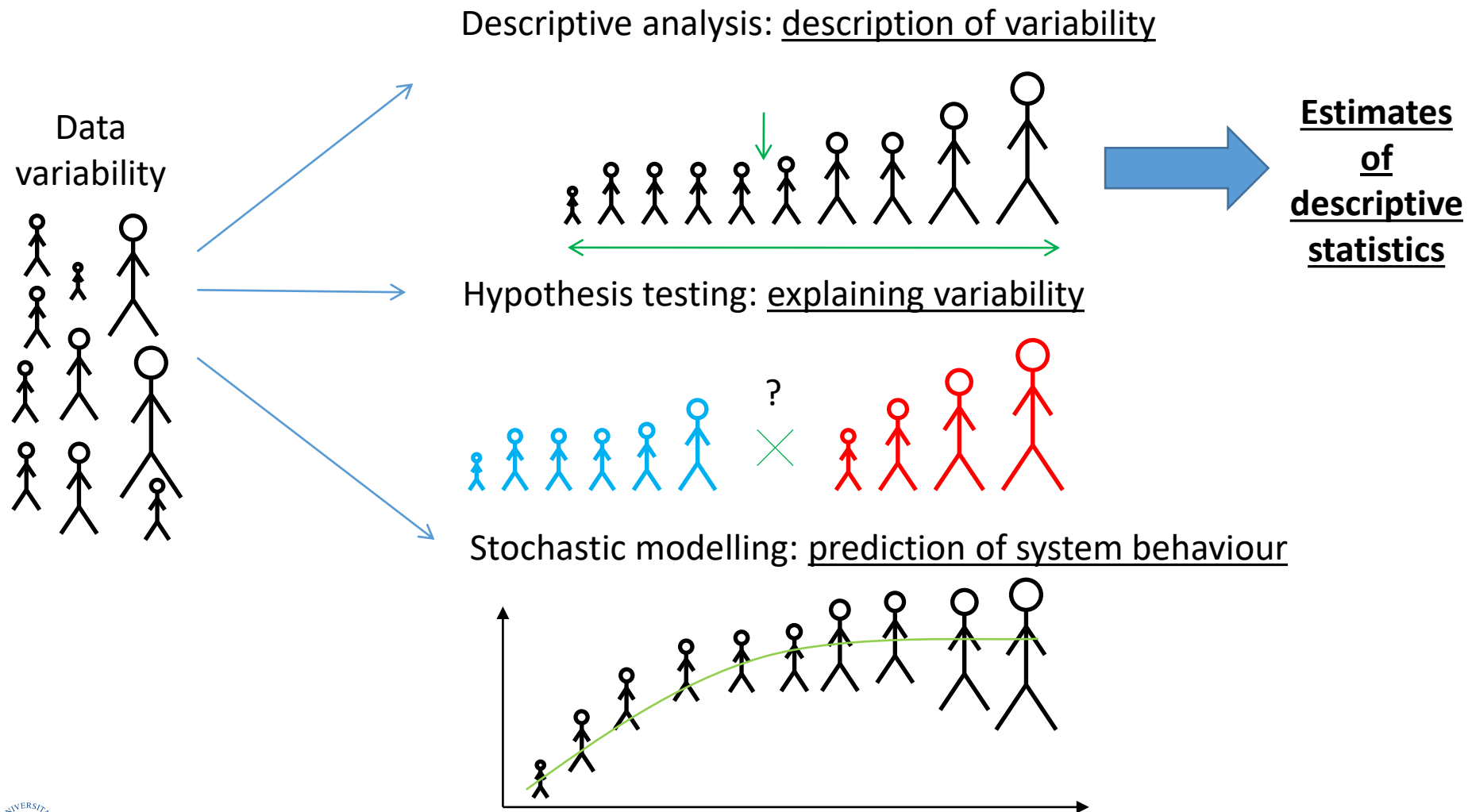
The meaning of the confidence interval

Annotation

- The two basic approaches to statistical evaluation are data description and hypothesis testing.
- When describing the data, it is important to note that the descriptive statistics obtained from the sample are not the actual value in the target population, but only an estimate of it.
- The precision of the estimate depends on both the variability of the data and the sample size; if the entire target population were sampled, the resulting descriptive statistic would already be a precise value, not an estimate.
- Estimates and associated confidence intervals are a universal statistical procedure and can be computed to any descriptive statistic.

Working with variability in data analysis

- There are three main approaches to dealing with variability in data analysis



Point estimate of descriptive statistics

- By calculating the descriptive statistics of the sample, we obtain the so-called point estimate



Point estimate of mean,
standard deviations



Is that enough?



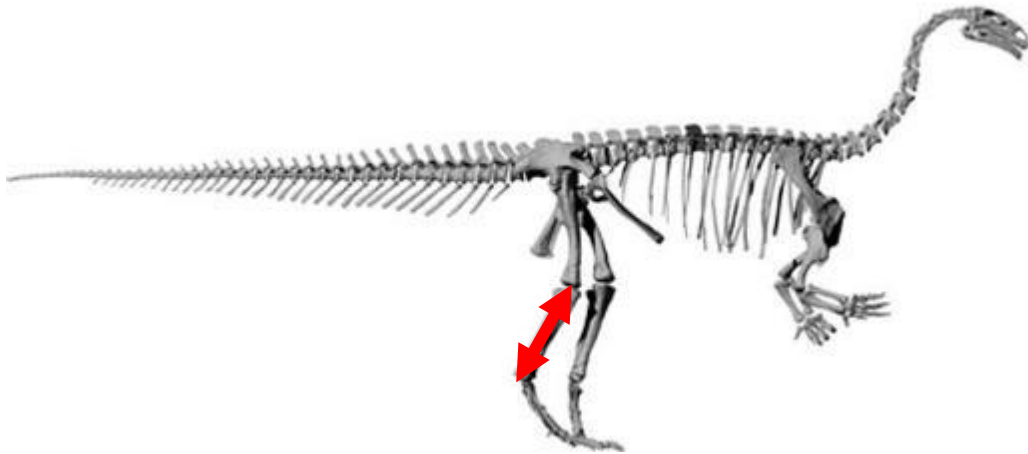
It is not, we do not take into account
the influence of chance that was
applied during the sampling !!!

Interval estimation

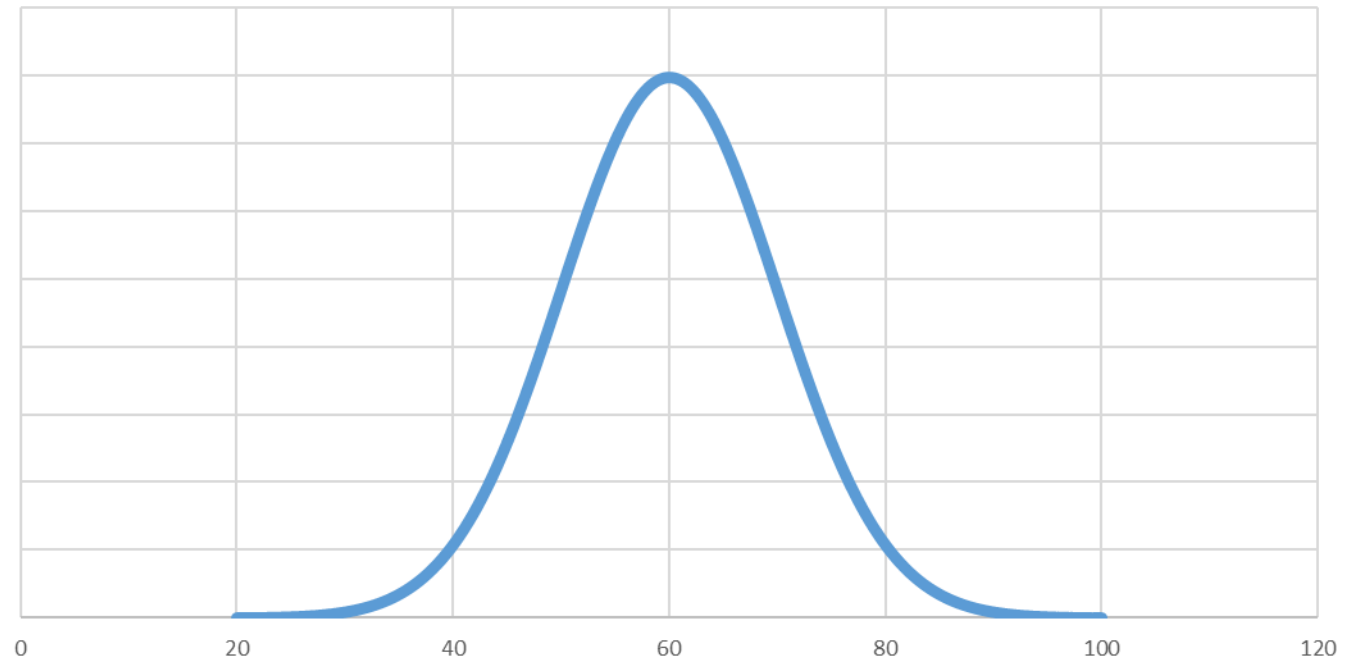
- Point estimation is the first step in the statistical description of the data.
- What does one number tell us? Study 1 may publish the number x_1 , study 2 the number x_2 . Which is more correct, better, more accurate?
- **The point estimate is inadequate by itself to describe the probability distribution parameter of a random variable.**
- We are interested in the accuracy (reliability) of the point estimate.

What is the importance of interval estimation and its reliability?

- We are sampling a population of animals and want to estimate the average value of the observed variable
- Average length in the population = 60, standard deviation = 10 (we don't actually know these values)



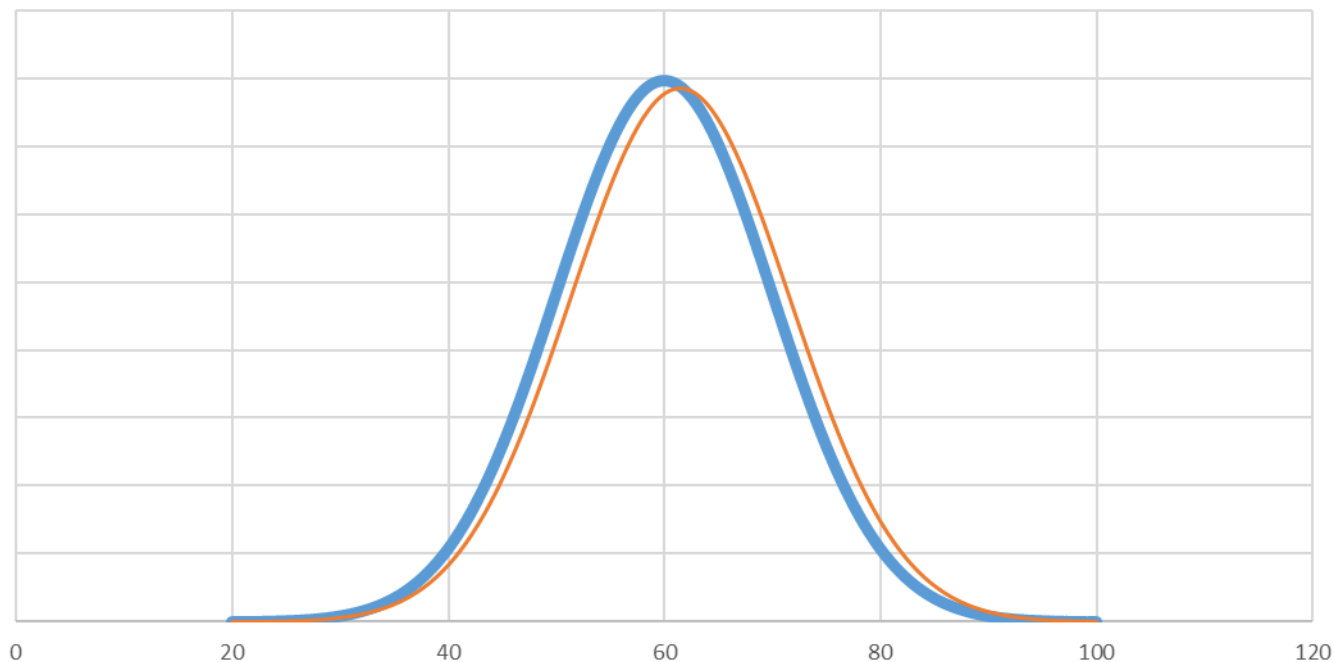
Population: mean = 60, standard deviation = 10



We perform a sampling of size $N = 100$.

One sampling

- There is only a low probability that the sample exactly matches the population of interest



Population: mean = 60, standard deviation = 10

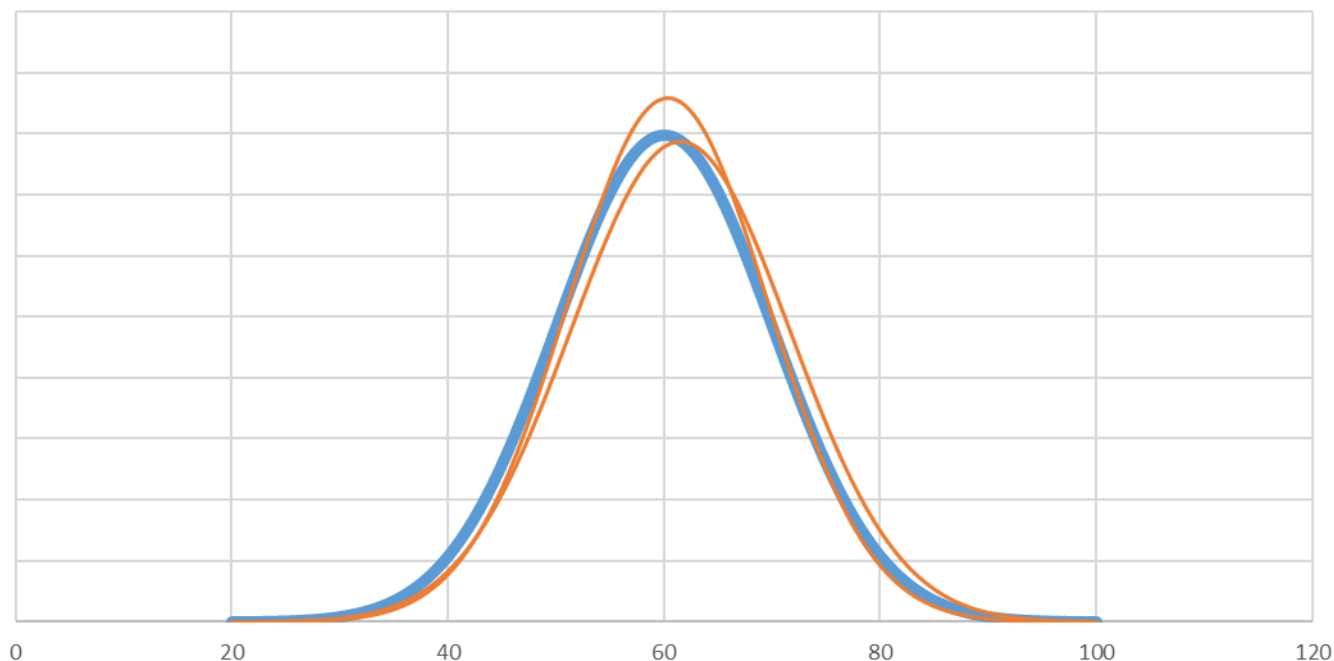
Sample 1: mean = 61.5, standard deviation = 10.1



How would further
sampling work?

Two sampling

- There is only a low probability that the sample exactly matches the population of interest



Population: mean = 60, standard deviation = 10

Sample 1: mean = 61.5, standard deviation = 10.1

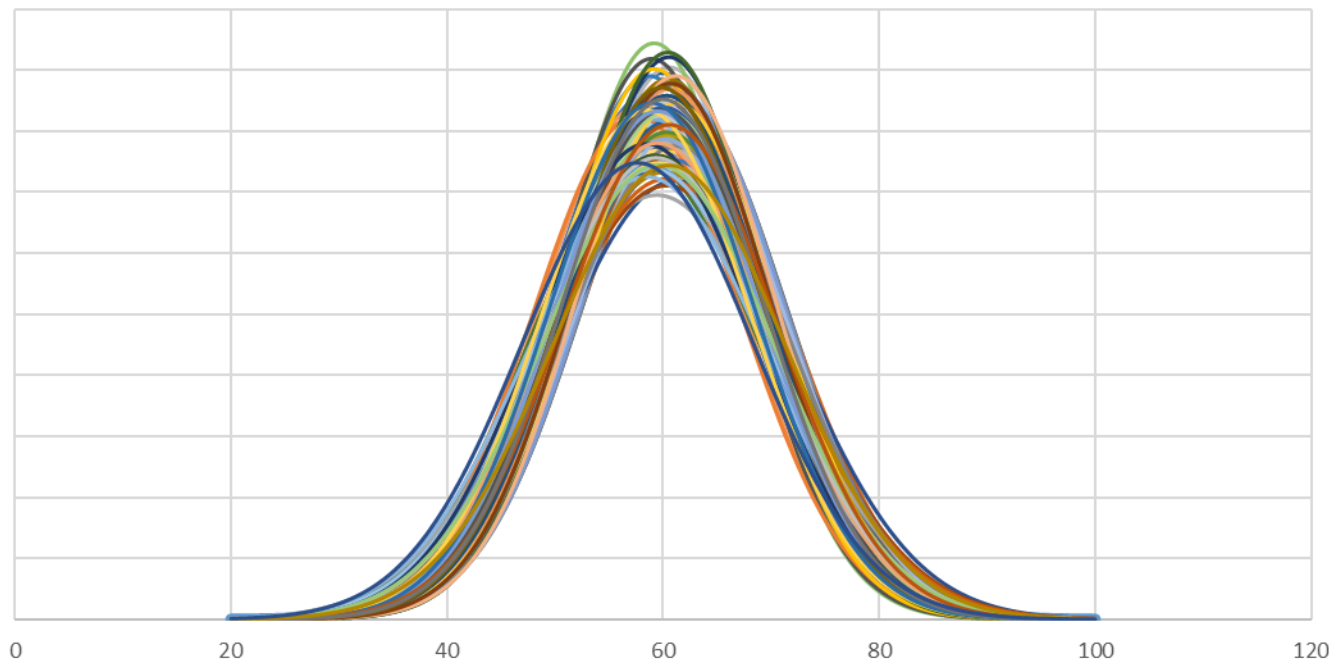
Sample 2: mean = 60.4, standard deviation = 9.3



How would further
sampling work?

One hundred samplings

- There is only a low probability that the sample exactly matches the population of interest



Population: mean = 60, standard deviation = 10

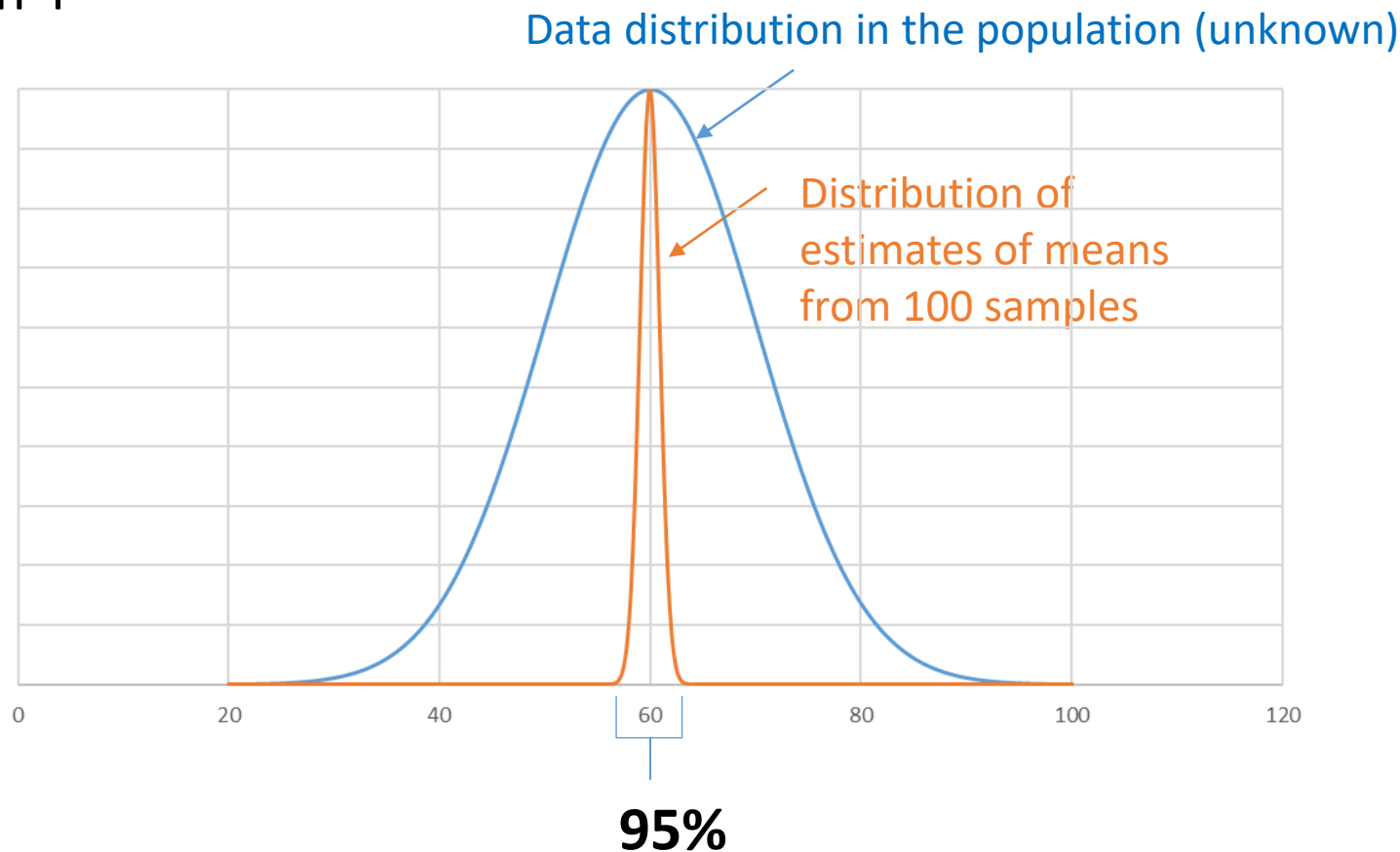
By resampling, we obtained different versions of the point estimate simulating how different population sampling would have turned out for a given sample size.



How would the next sampling turn out?
Are we able to describe it in terms of probability
= the **estimate at the next sampling will end up with a certain probability in a certain range of values?**

Estimation confidence interval I

- Estimates of the mean from individual samples produce a distribution of mean estimates
- If we know the distribution, we can easily determine the range in which the specified percentage of values lies = the probability of encountering an estimate of the mean in this range when sampling
- The most commonly used 95% range = **95% confidence interval**
- How to calculate it?



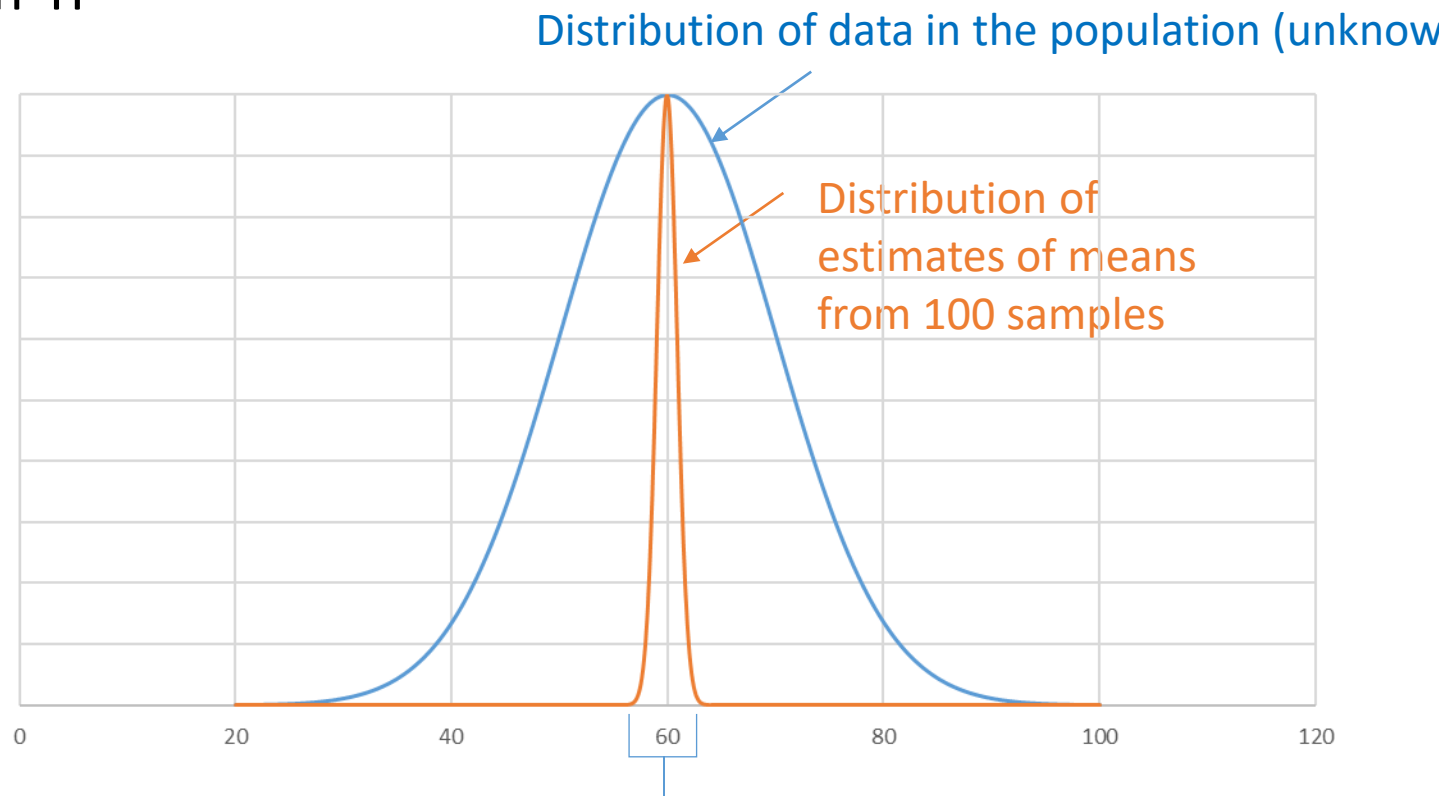
Population: mean = 60, standard deviation = 10

Samples (N = 100): mean = 59.9, standard deviation of mean estimates = 0.93

???

Estimation confidence interval II

- How to calculate it?
- Empirically: 2.5% and 97.5% quantile
- According to the model distribution:
 - The estimates of the averages have a normal distribution
 - Mean 95% values are bounded by the mean $\pm 1.96 \cdot \text{standard deviation}$
- Note: the described method of calculating the confidence interval is used only in computer simulations, not in real sampling (here for educational reasons)



Population: mean = 60, standard deviation = 10

Samples (N = 100): mean = 59.9, standard deviation of mean estimates = 0.93

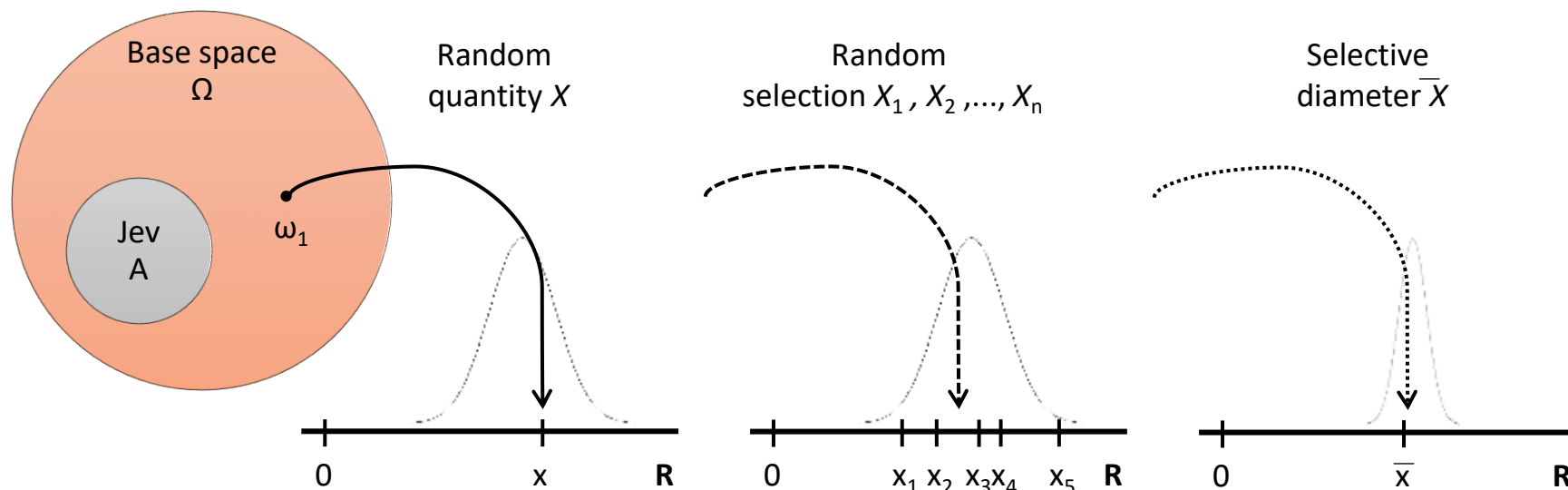
Mean error of the mean estimate (standard error, s.e., SE, $\sigma_{\bar{x}}$)

Probabilistic behaviour of a random variable

- In classical statistical calculations, the confidence interval is derived from a single sample based on knowledge of the model distribution of estimates of the statistic (e.g. the mean)
- Two characteristics reflect the properties of a single-number distribution: the mean and the variance. The square root of the variance is the standard deviation (SD).
- The following applies:
 - Individual realizations of the random variable show variability (according to SD).
 - Any statistic (e.g. the mean) is also a random variable as a transformation of random variables. Therefore, it also has a probability distribution.
 - Individual realizations of the statistic over different random samples also show variability (again proportional to SD).
 - S.E. - standard error

Example - sample average

- In the case of the mean, its estimates are describable by a normal distribution model
- The normal distribution is described by the mean (the actual estimate of the mean) and the standard deviation of the estimates (to distinguish it from the sample standard deviation, in this case it is called the mean error of the mean estimate)



SD and SE

- The standard deviation (SD) is not the standard error of the descriptive statistic (SE)!
- The standard deviation (SD) is a reflection of the variability of the random variable in the population under study.
- The standard error (SE) is a reflection of the accuracy of descriptive statistics as an estimate of the mean of a random variable.
- Note the difference between SD and SE in articles and books - tables and graphs!
- **What does the size of the SE (and therefore the width of the confidence interval) depend on?**

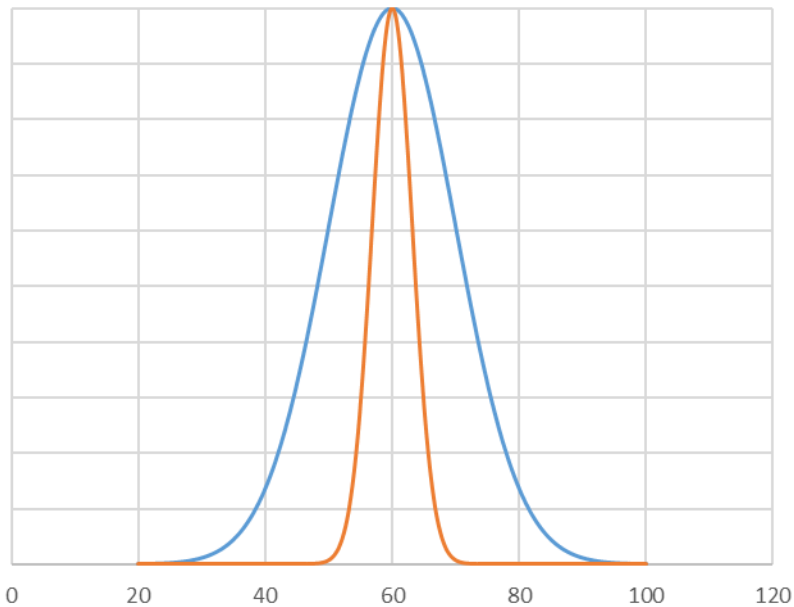
SD and SE

- The standard deviation (SD) is not the standard error of the descriptive statistic (SE)!
- The standard deviation (SD) is a reflection of the variability of the random variable in the population under study.
- The standard error (SE) is a reflection of the accuracy of descriptive statistics as an estimate of the mean of a random variable.
- Note the difference between SD and SE in articles and books - tables and graphs!
- **What does the size of the SE (and therefore the width of the confidence interval) depend on?**
 - On the sample size
 - The variability (standard deviation) of the variable being assessed in the population
- The SD of the population is given by reality, but the sample size is in our hands = **by changing the sample size we can change the width of the confidence interval !!!!**

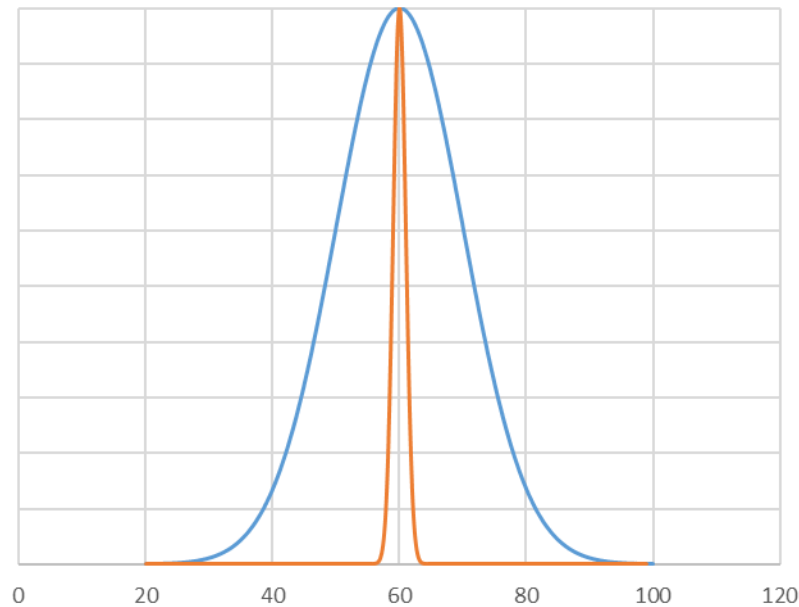
Example - confidence interval for different sample sizes

- We are sampling a population of animals and want to estimate the mean value of the variable of interest - we are trying different sample sizes
- Average length in the population = 60, standard deviation = 10 (we don't actually know these values)

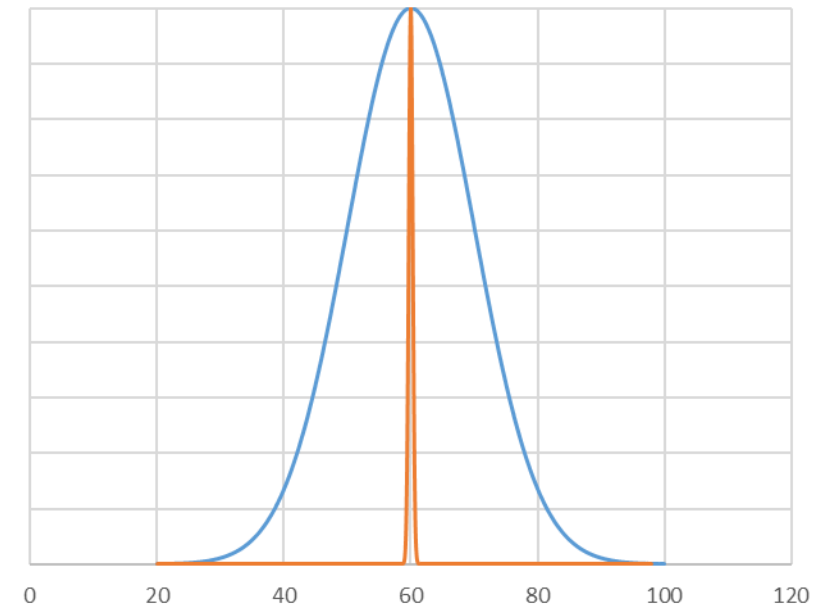
N = 10



N = 100



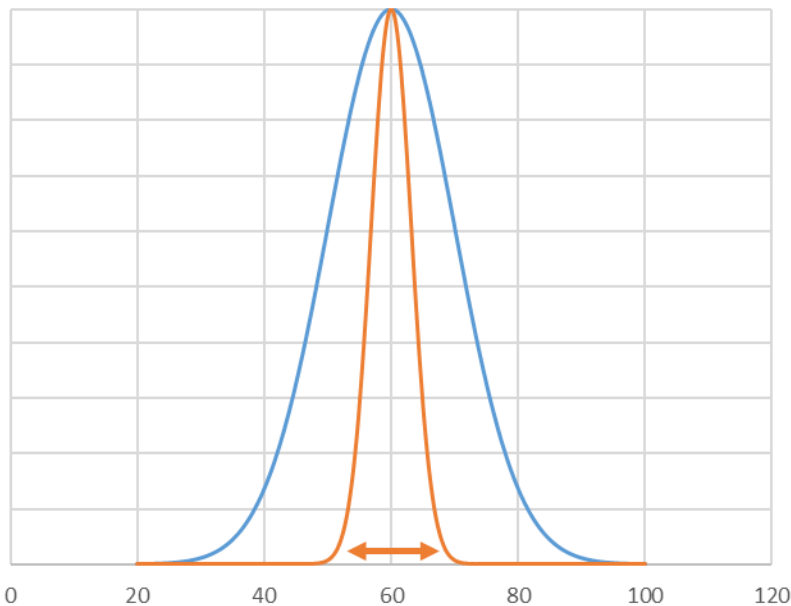
N = 1000



Example - confidence interval for different sample sizes

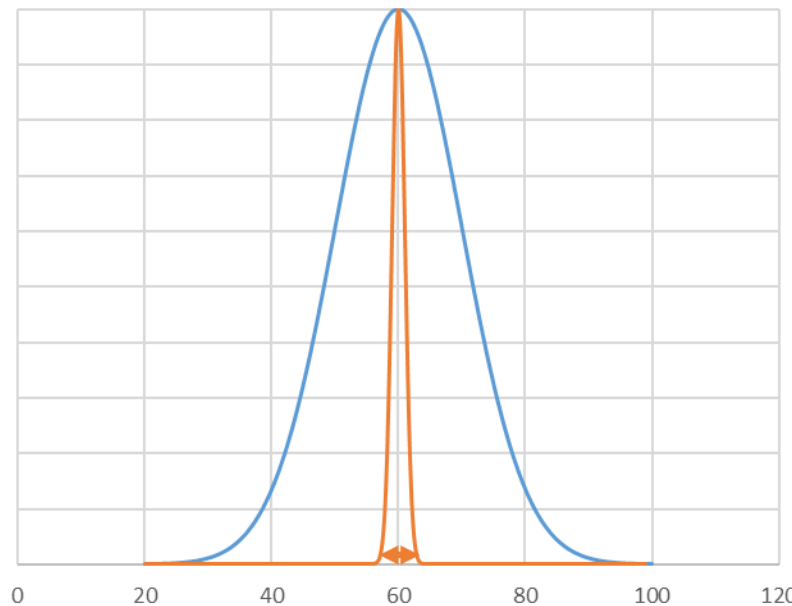
- We are sampling a population of animals and want to estimate the mean value of the variable of interest - we are trying different sample sizes
- Average length in the population = 60, standard deviation = 10 (we don't actually know these values)

N = 10



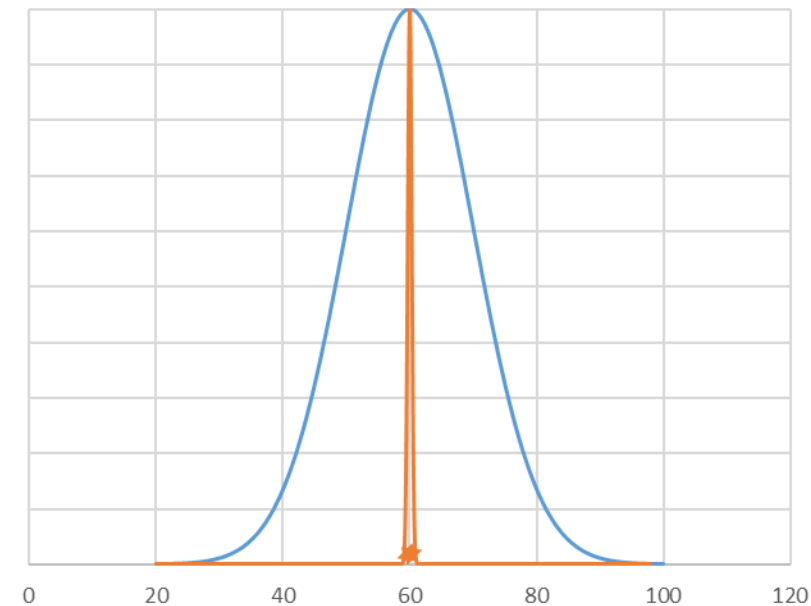
95% IS = 53.8 - 66.2

N = 100



95% IS = 58.0 - 62.0

N = 1000



95% IS = 59.4 - 60.6

General formula for calculating the confidence interval

- Confidence intervals can be calculated to estimate any descriptive statistic (mean, standard deviation, percentage, correlation coefficient, regression coefficient, odds ratio, etc.)
- For a given descriptive statistic, we need to know the corresponding model distribution of its estimate
- General equation for calculating confidence interval boundaries (in some cases it can be more complicated - asymmetric confidence intervals, different equations for lower and upper boundaries):

Point estimate ± quantile of model distribution * mean estimation error

↑
E.g. sample diameter

↑
In the case of the mean and 95% confidence interval this is 2.5% and 97.5% quantile of normal distribution
= ± 1.96

↑
In the case of the average, it is calculated as:
$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Calculation of the average estimate

- Point estimate of the sample mean
- Mean error of the mean estimate

$$\bar{x}$$
$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

- Confidence interval

$$\bar{x} - t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}}$$

$$\mu: \bar{x} \pm t_{1-\alpha/2}^{v=N-1} \frac{s}{\sqrt{N}}$$

$$\mu: \bar{x} \pm t_{1-\alpha/2}^{v=N-1} s_{\bar{x}}$$

t - Student's distribution (used instead of the normal distribution for small sample sizes)

v - degrees of freedom, here calculated as N-1

What is it?

$t_{1-\alpha/2}^{v=N-1}$

The quantile of the model distribution, α , denotes the proportion of cases we do not want to include in the interval, here for a 95% confidence interval $\alpha = 5\%$, so we are looking for the 97.5% quantile of the student's distribution

Statistical t-distribution tables

- In contrast to the normal distribution tables, we must also take into account the degrees of freedom
- For this reason, the table is constructed only for selected probability values

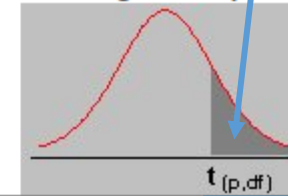


William Sealy Gosset

Publications under the pseudonym Student
t division based on experiments with yeast

We look for the value of **t** (= quantile of the distribution) for a given area (**probability**) and **degrees of freedom**

t table with right tail probabilities



Probability (area under the curve), most commonly 0.025 ($2 \times 0.025 = 0.05$)

df \ p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869

Degrees of freedom

Estimating the mean and its confidence interval - Example 1

- We are sampling a population of animals and want to estimate the average value of the observed variable
- Sample: N = 10, mean (point estimate) 61.5, standard deviation 10.1

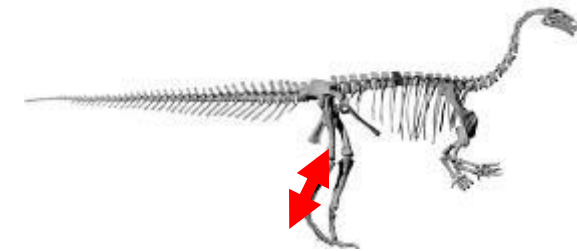
- **What is the 95% confidence interval?**

- Mean estimation error $s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{10,1}{\sqrt{10}} = 3,207$

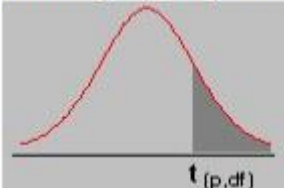
- Quantile of model distribution for $\alpha=0.05$ (1-0.95)
=2,262

- 95% confidence interval - calculation
= 7,256

- 95% confidence interval - result
61,5 (54,2 - 68,7)



t table with right tail probabilities



df \ p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809

- **When resampling with N=10, the estimate of the mean with probability 0.95 will lie in the range (54.2 - 68.7)**

Estimating the mean and its confidence interval - Example 2

- We are sampling a population of animals and want to estimate the average value of the observed variable
- Sample: N = 100, mean (point estimate) 61.5, standard deviation 10.1

- **What is the 95% confidence interval?**

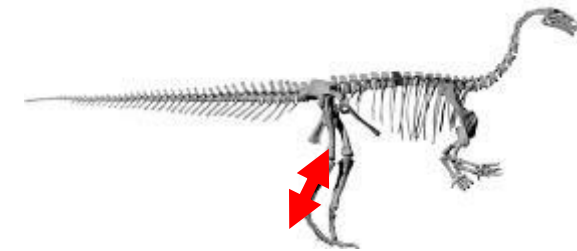
- Mean estimation error $s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{10,1}{\sqrt{100}} = 1,014$

- Quantile of model distribution for $\alpha=0.05$ (1-0.95)
=1,960

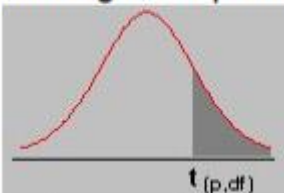
- 95% confidence interval - calculation
= 1,988

- 95% confidence interval - result
61,5 (59,5 - 63,5)

- **With a resampling of N=100, the estimate of the mean with probability 0.95 will lie in the range (59.5 - 63.5)**



t table with right tail probabilities



df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Confidence interval for variance estimation

- Example of an asymmetric confidence interval; the model distribution is the Pearson (chi-squared distribution)

- **For dispersion**

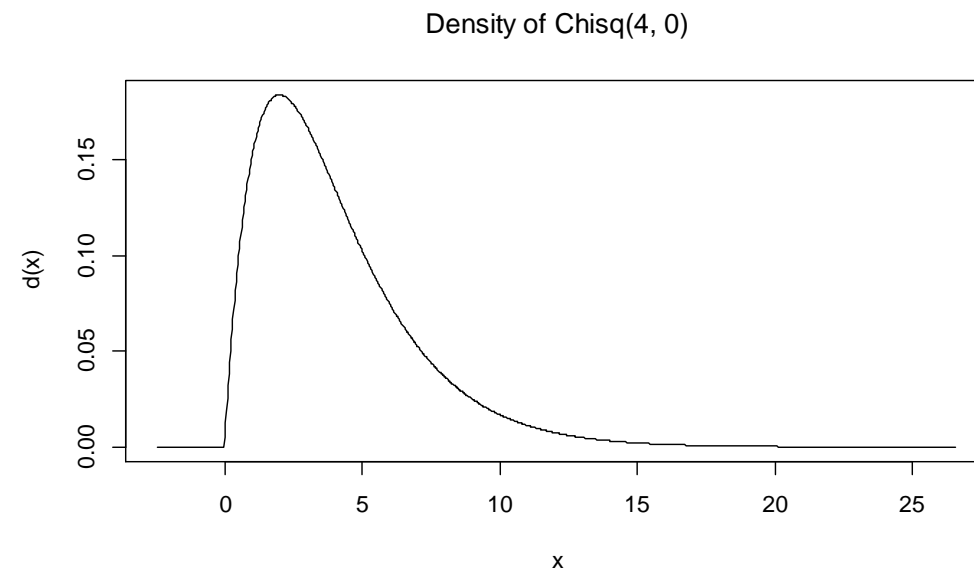
$$\frac{(N-1)s^2}{\chi^2_{\alpha/2, \nu=N-1}} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi^2_{1-\alpha/2, \nu=N-1}}$$

- **For the standard deviation**

$$\sqrt{\frac{(N-1)s^2}{\chi^2_{\alpha/2, \nu=N-1}}} \leq \sigma \leq \sqrt{\frac{(N-1)s^2}{\chi^2_{1-\alpha/2, \nu=N-1}}}$$

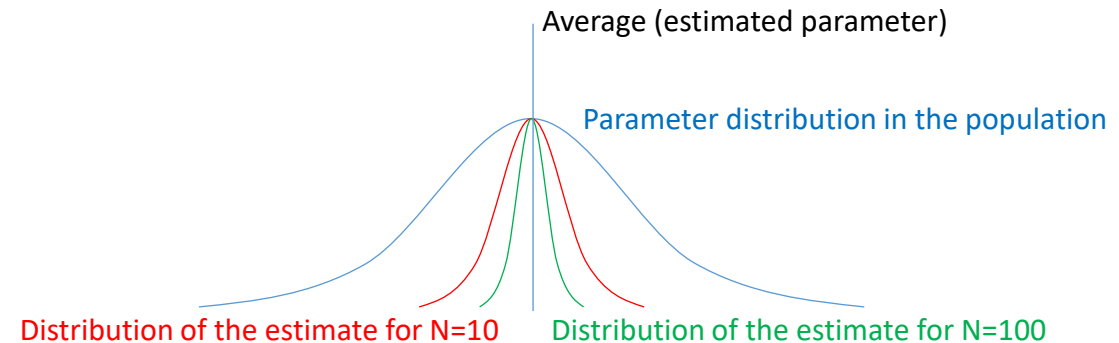
- **For the mean estimation error of the mean**

$$\sqrt{\frac{(N-1)s^2}{N\chi^2_{\alpha/2, \nu=N-1}}} \leq \frac{\sigma}{\sqrt{N}} \leq \sqrt{\frac{(N-1)s^2}{N\chi^2_{1-\alpha/2, \nu=N-1}}}$$



The confidence interval concept and its interpretation: a summary

- When calculating an estimate of a descriptive statistic, we are interested not only in its intrinsic value (point estimate) but also in its confidence range
- The confidence interval depends on:
 - Sample sizes
 - Data variability
 - Required reliability
- Confidence intervals can be calculated for any statistic (mean, standard deviation, correlation, percentage, etc.)
- The confidence interval provides a clue as to how "reliable" our results are and how likely they are to be achieved repeatedly
- The 95% confidence interval is the range of values within which we are 95% likely to be within when repeating the study
- The claim that the true population mean lies within the 95% confidence interval with 95% probability is not true, we do not know the true population mean !!!



Note on the confidence interval

- The confidence interval only accounts for variability due to random selection, it does not account for sources of systematic bias.
- **Examples:**
 - Measurements of pollutant concentration or blood pressure may be systematically biased by an old meter ("technical bias").
 - Pollutant concentration measurements can be systematically biased by selecting only clean or only contaminated sites ("selection bias")
 - Blood pressure measurements can be systematically biased by enrolling only a certain group of people in a study ("selection bias")

Basics of hypothesis testing

The principle of statistical hypothesis testing

Test statistics and statistical significance

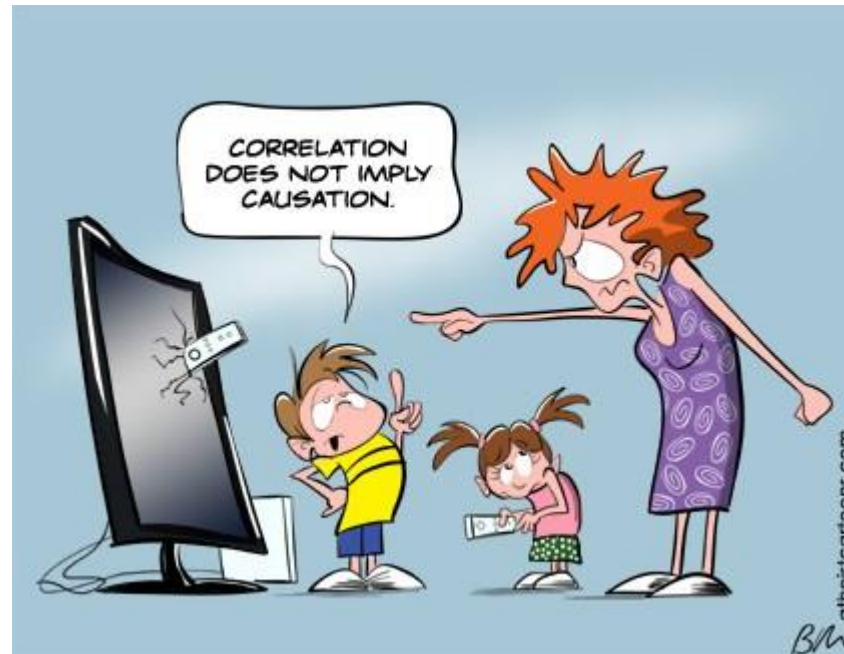
Statistical testing errors

Annotation

- Hypothesis testing is the second main direction of statistical analysis after descriptive statistics. In testing, we pose hypotheses that we seek to confirm or reject with some probability.
- The so-called null hypothesis can best be described as a situation where we assume the influence of chance (the difference between groups is pure chance, the relationship between two variables is pure chance, etc.), the alternative hypothesis assumes the influence of a non-random factor.
- The result of a statistical test is basically the probability of whether the phenomenon being evaluated is random or not, when a certain threshold is crossed (most often less than 5% probability that the phenomenon is just a coincidence) we declare that the probability of coincidence is low enough for us to declare the phenomenon non-random
- Statistical significance is influenced by the sample size and thus is only an indication to declare e.g. the difference between two groups of patients as truly significant. In an ideal situation, it is necessary that the difference is significant not only statistically (=non-random) but also practically (=not just an artifact of sample size).

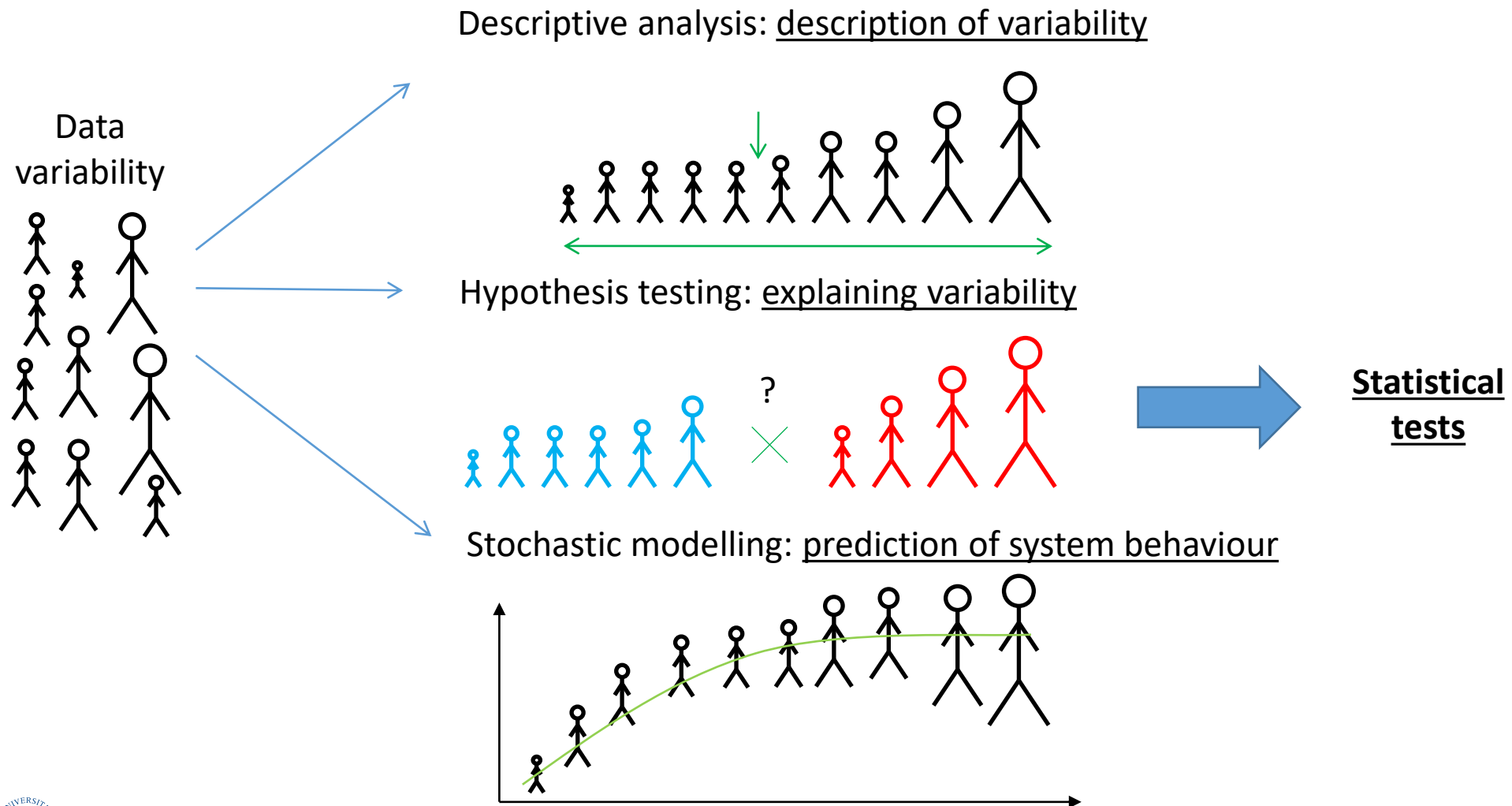
Statistical testing is not proof of causality !!!!

- The result of statistical testing does not mean causal proof or non-proof of a relationship, it is only an indication for our decision making.



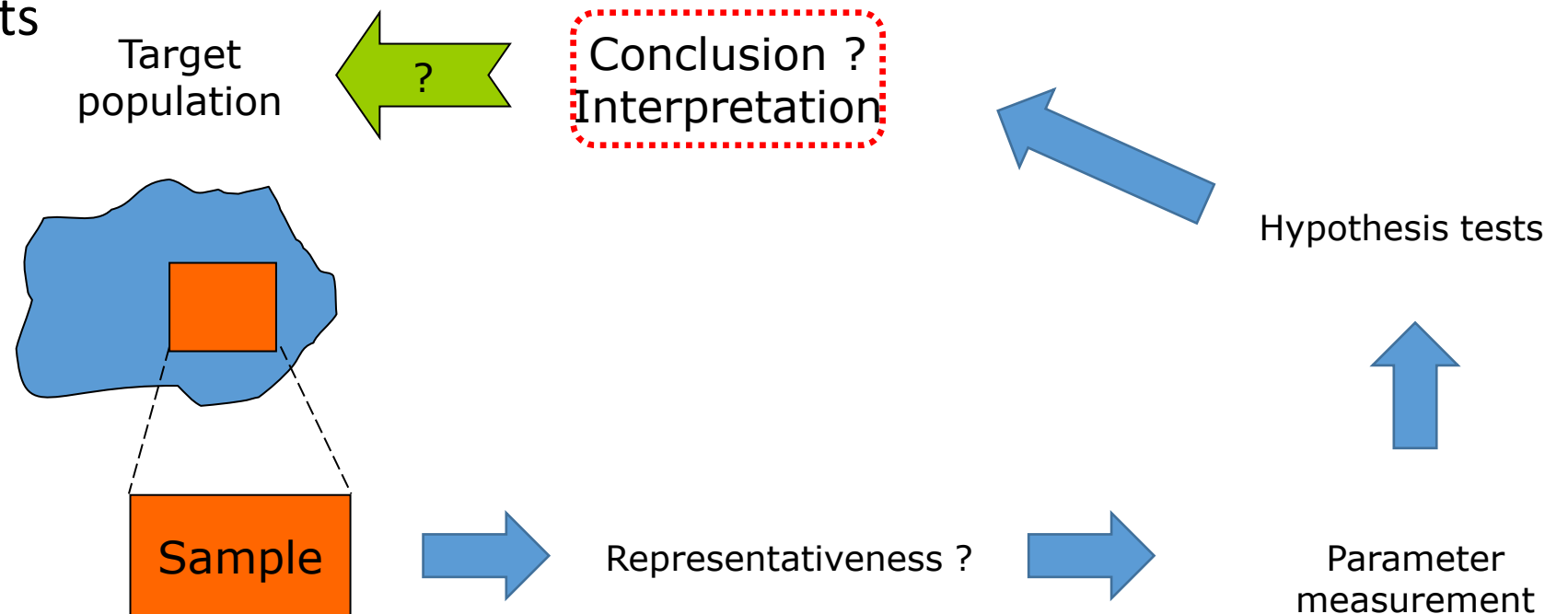
Working with variability in data analysis

- There are three main approaches to dealing with variability in data analysis



The principle of hypothesis testing

- Formulation of the hypothesis
- Selection of the target population and a representative sample from it
- Measurement of monitored parameters
- Use of the corresponding test ~~Test~~ → conclusion
- Interpretation of results



Determination of the hypothesis

- **Null hypothesis** - a statement about unknown properties of the probability distribution of the observed random variable (trait, characteristic) concerning the target population.
- The null hypothesis takes the form $H_0: \theta = \theta_0$
- The null hypothesis generally says that there is no difference, or that the difference is so small that it can be considered random -> so the basic question of testing is "how do we define what is "sufficiently" random for us?"
- **Alternative hypothesis** - a statement about the unknown properties of the probability distribution of the observed random variable that contradicts the validity of the null hypothesis. It defines what situation occurs when the null hypothesis is not valid.
- The alternative hypothesis takes the form:
 - $H_1: \theta \neq \theta_0$
 - $H_1: \theta < \theta_0$
 - $H_1: \theta > \theta_0$

Examples of hypothesis

- Do sites near human settlements differ from sites in protected reserves in terms of pollution levels?

Levels of pollution at sites near settlements: θ_1 $H_0 : \theta_1 = \theta_2$

Levels of pollution at sites in reserves: θ_2 $H_1 : \theta_1 \neq \theta_2$

- Is the effect of lowering systolic blood pressure with a new antihypertensive the same in hypertensives who smoke as in hypertensives who do not smoke?

Mean effect size for smokers: θ_1 $H_0 : \theta_1 = \theta_2$

Mean effect size for non-smokers: θ_2 $H_1 : \theta_1 < \theta_2$

Why does the null hypothesis express the absence of an effect?

- The null hypothesis reflects the fact that something did not happen or did not manifest
→ is usually stated as the opposite of what we want to prove by experiment.
- **The null hypothesis is constructed so that we can use the observed values to disprove it.**
- In order to reject the validity of the null hypothesis, we only need to find one example where the null hypothesis does not hold - this example should be our random sample (our observed data).
- Rejecting the null hypothesis is easier than confirming the null hypothesis.

Hypothesis testing

- Hypothesis testing deals with deciding the validity of stated hypotheses based on observed data.
- We test the validity of the hypotheses using a **statistical test** - a decision rule that assigns to each random selection just one of two possible decisions - H_0 do not reject or H_0 reject.

Statistical test

- Hypothesis testing is based on data.
- **The hypothesis being tested corresponds to a statistical test**, or a test statistic, which allows to test the validity of the null hypothesis.
- **A test statistic** is a formula based on observed data with a probability distribution, so it itself has a **probability distribution**. The probability distribution of a test statistic under the validity of H_0 is referred to as the "null distribution".

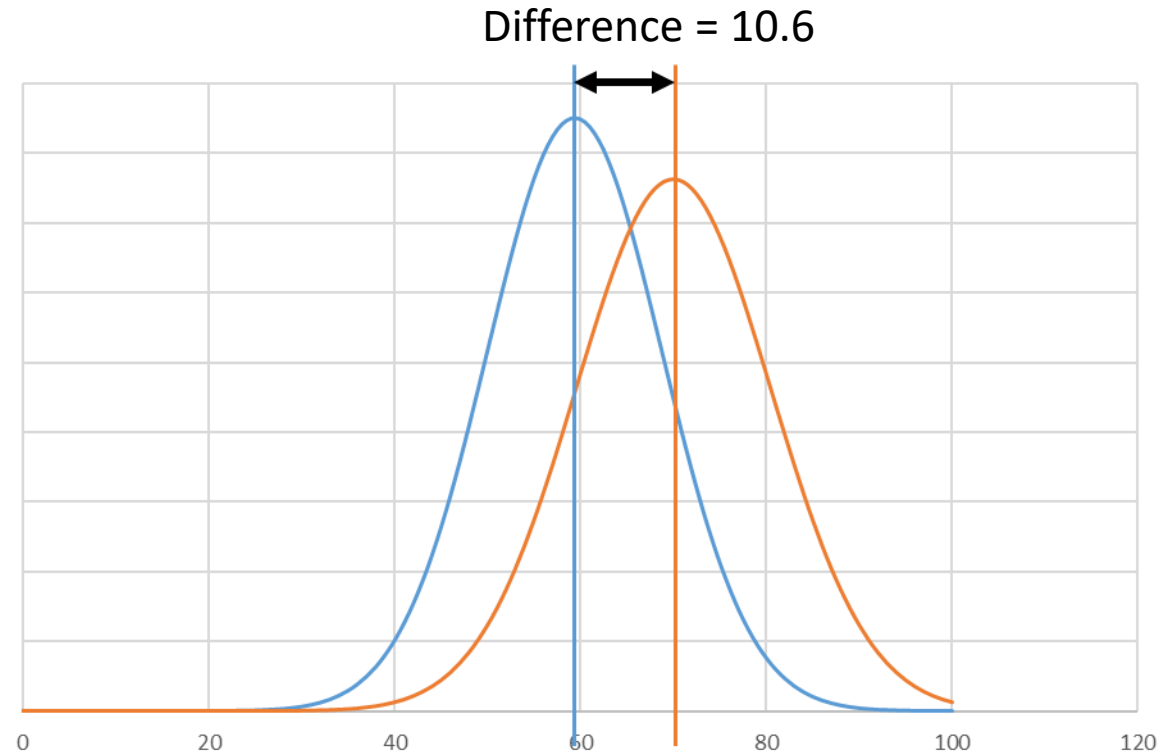
Statistical testing procedure

- We formulate the null hypothesis H_0 (the observed effect is zero)
- We formulate an alternative hypothesis H_A (the observed effect is different between groups) The alternative hypothesis for parametric tests can be two-sided or one-sided.
- The hypothesis must be stated so that we can select and calculate the so-called test statistic (e.g., the hypothesis about averages is likely to be solved using a t-test whose test statistic has a t distribution)
- We calculate the value of the test statistic based on the observed values
- We compare the calculated test statistic with its distribution (= distribution of random differences), assess the randomness of the difference and conclude to reject/not reject H_0

What does the value of the test statistic depend on?

- We have two groups of values, each described by its size, mean and standard deviation
 - what influences the significance of the difference in their means?

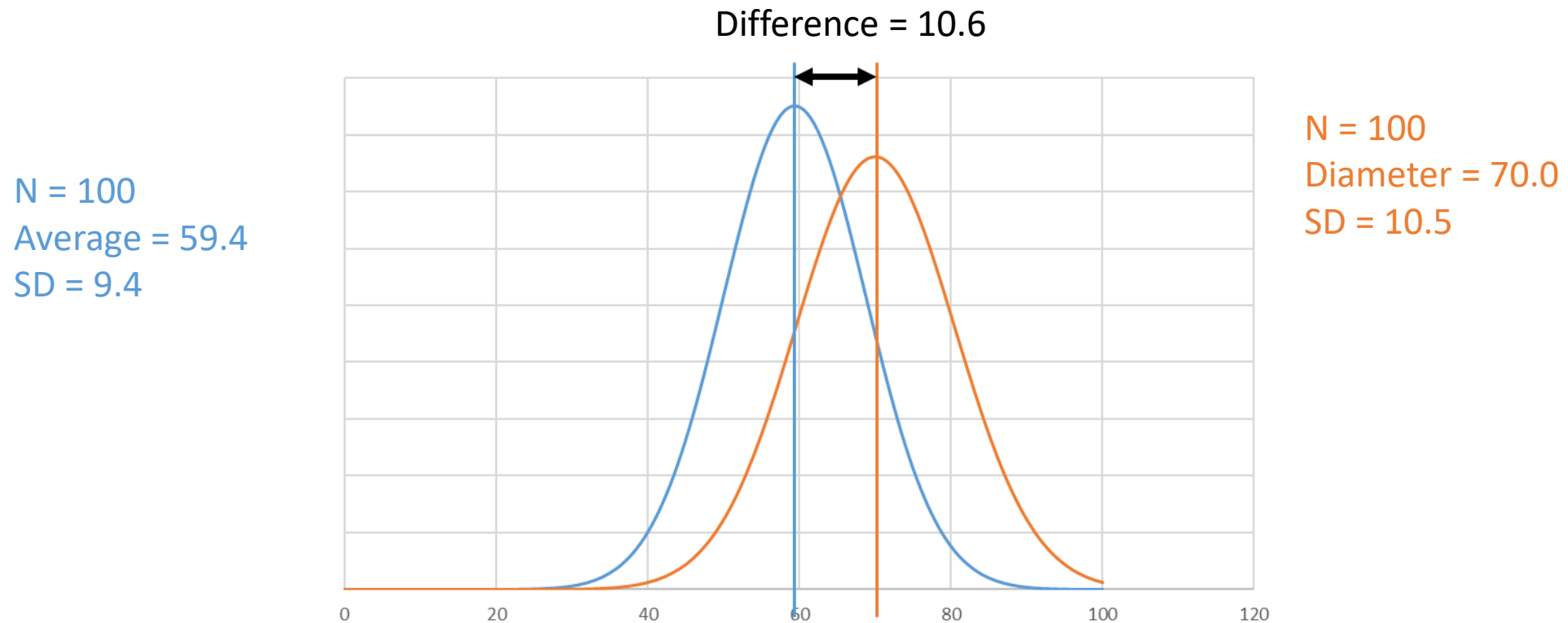
N = 100
Average = 59.4
SD = 9.4



N = 100
Diameter = 70.0
SD = 10.5

What does the value of the test statistic depend on?

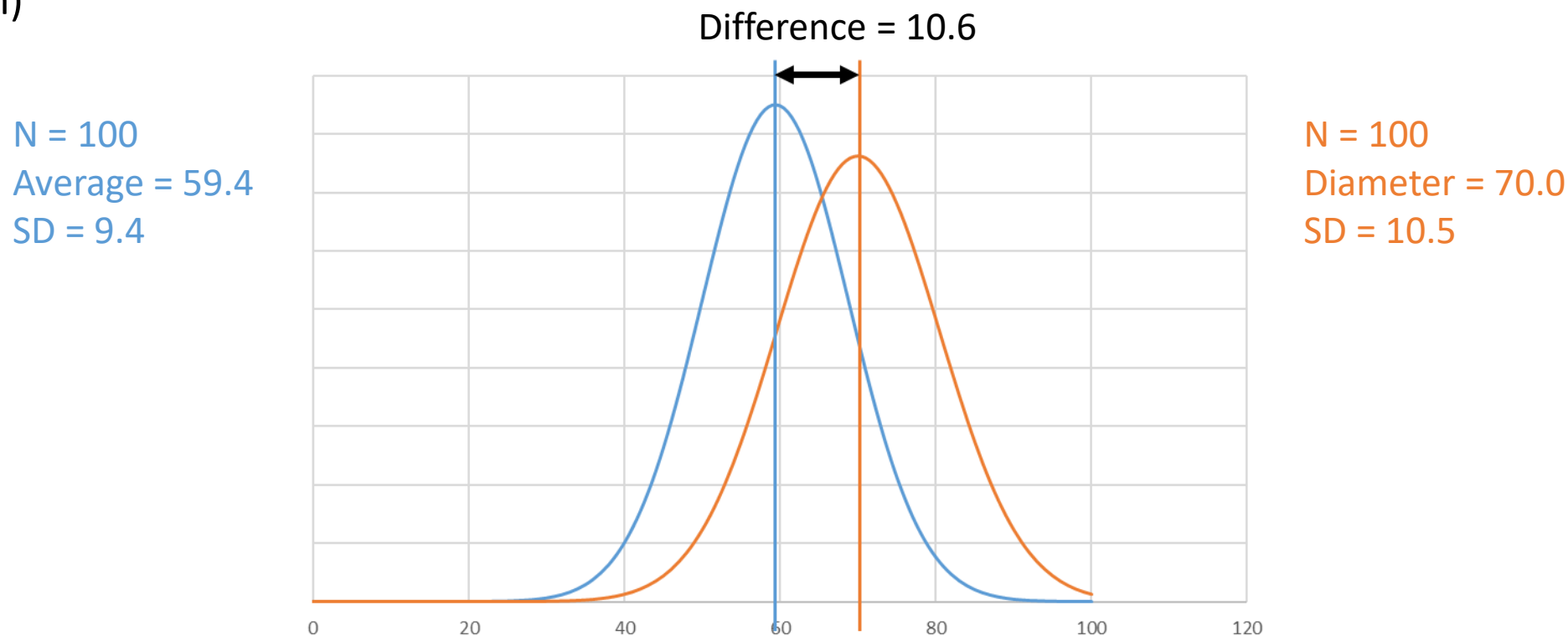
- We have two groups of values, each described by its size, mean and standard deviation - what influences the significance of the difference in their means?



- Sample size (larger sample = greater significance) and standard deviation (greater variability = less significance) - affect the confidence with which we estimate the means being compared
- On the size of the difference between the averages being compared (larger difference = greater significance)

Test statistics

- The test statistic combines the magnitude of the difference with other characteristics of the data (sample size, variability, etc.); it is actually a difference weighted by other characteristics
- The value of the test statistic is related to the significance of the difference
- For a final decision on the significance of the difference, it is necessary to compare the test statistic with its distribution of random differences (= what would be the distribution of this statistic if the difference were random)



Two ways of obtaining the distribution of the test statistic

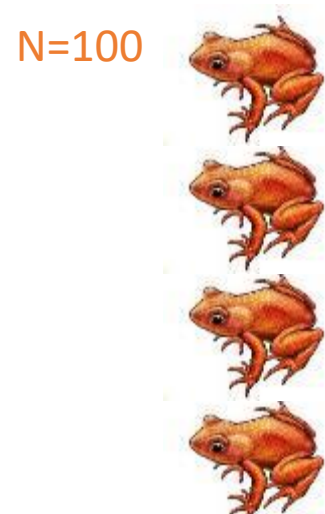
- The test statistic represents the distribution of random differences and can be obtained in two ways
- **Approximation to the model distribution**
 - "standard" procedure, advantage of easy calculation, sensitive to non-compliance with data distribution assumptions
 - Different tests have their distributions of random differences described by different model distributions (e.g., t-test using t-distribution, goodness-of-fit test using Pearson (chi-square) distribution)
- **Permutation methods**
 - The distribution of random differences is obtained by computer simulation of either all possible or a specified number of random situations
 - Suitable for small sample sizes or situations where approximation to model distributions is not possible
 - Demanding on computing power (currently less and less of a problem)
 - Educatively illustrative

Methods of testing

- Testing H_0 against H_A at the α significance level can be done in three different ways:
 1. The critical region (labelled W) or the rejection region H_0 ,
 2. Confidence interval,
 3. P-value.

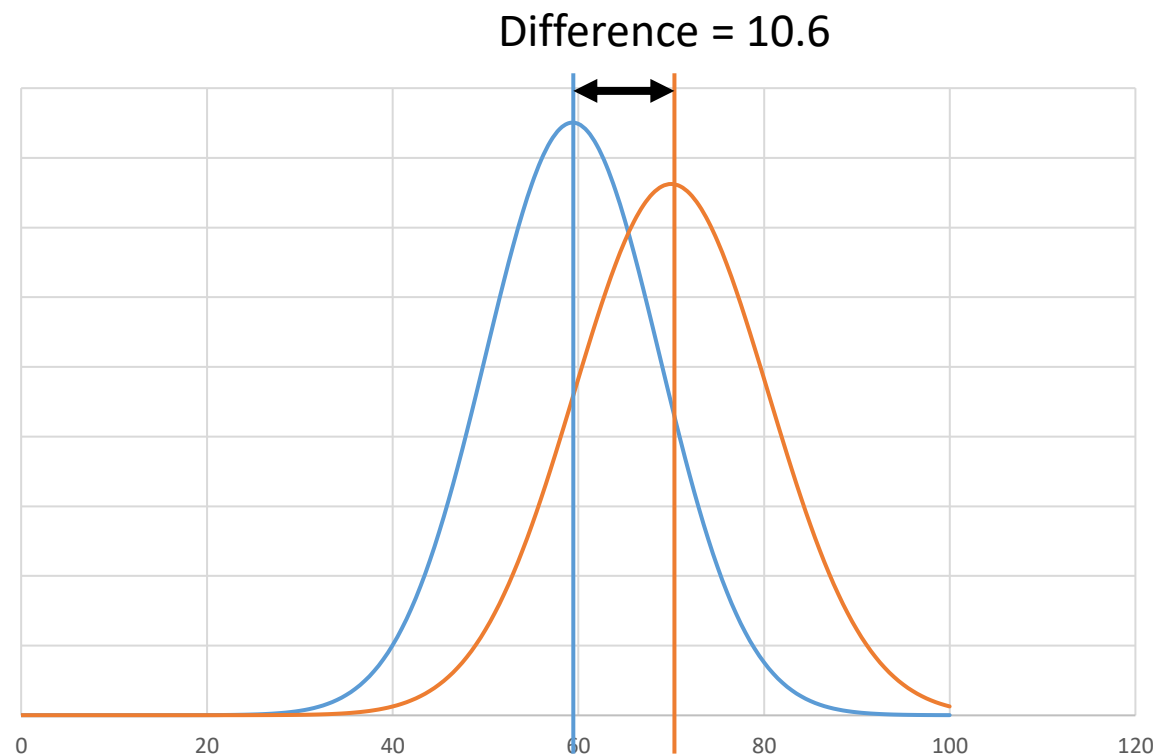
Example: permutation testing

We assessed the size of two species of frogs, sampling 100 individuals of each species.



The difference ???

N = 100
Average = 59.4
SD = 9.4

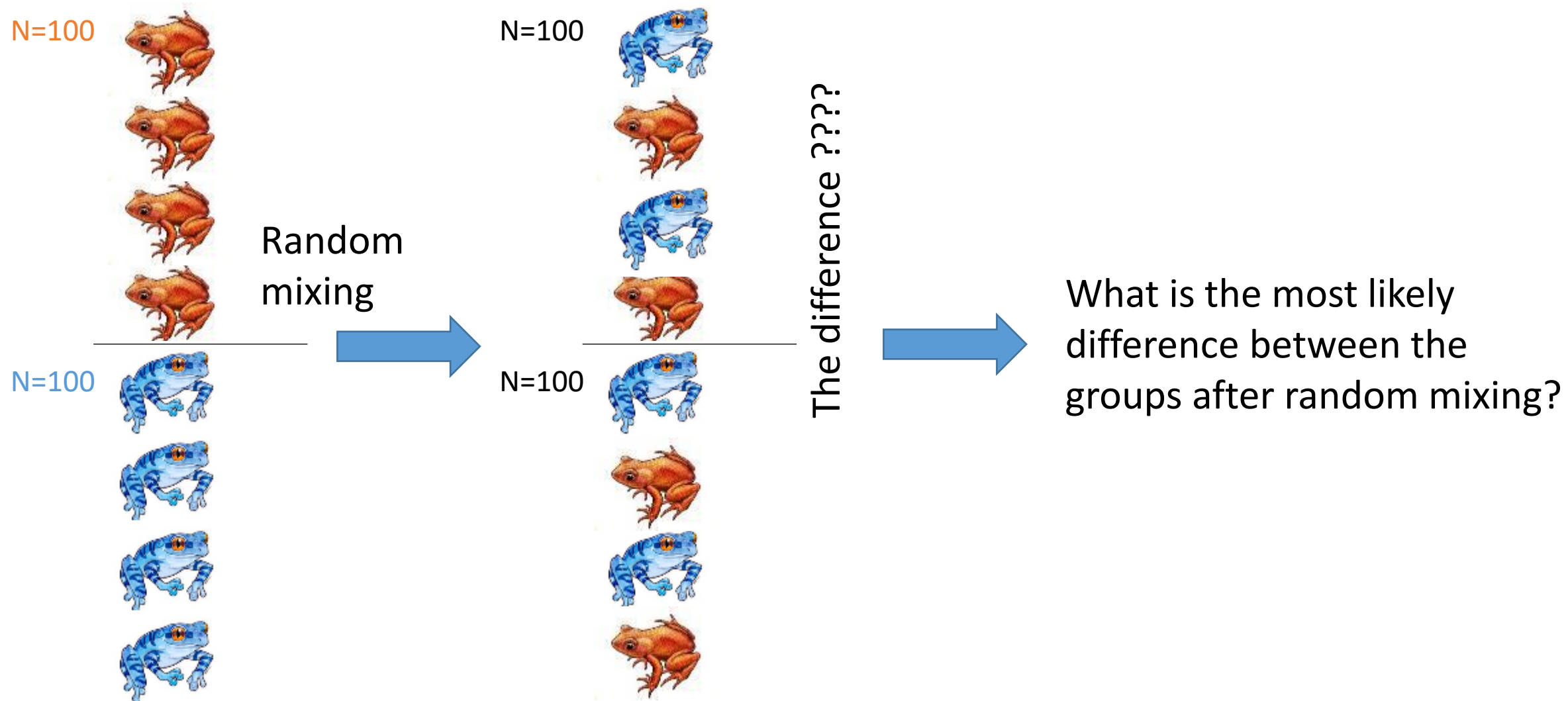


N = 100
Diameter = 70.0
SD = 10.5

How to determine if the observed difference is due to chance?
Let's simulate it !!!!

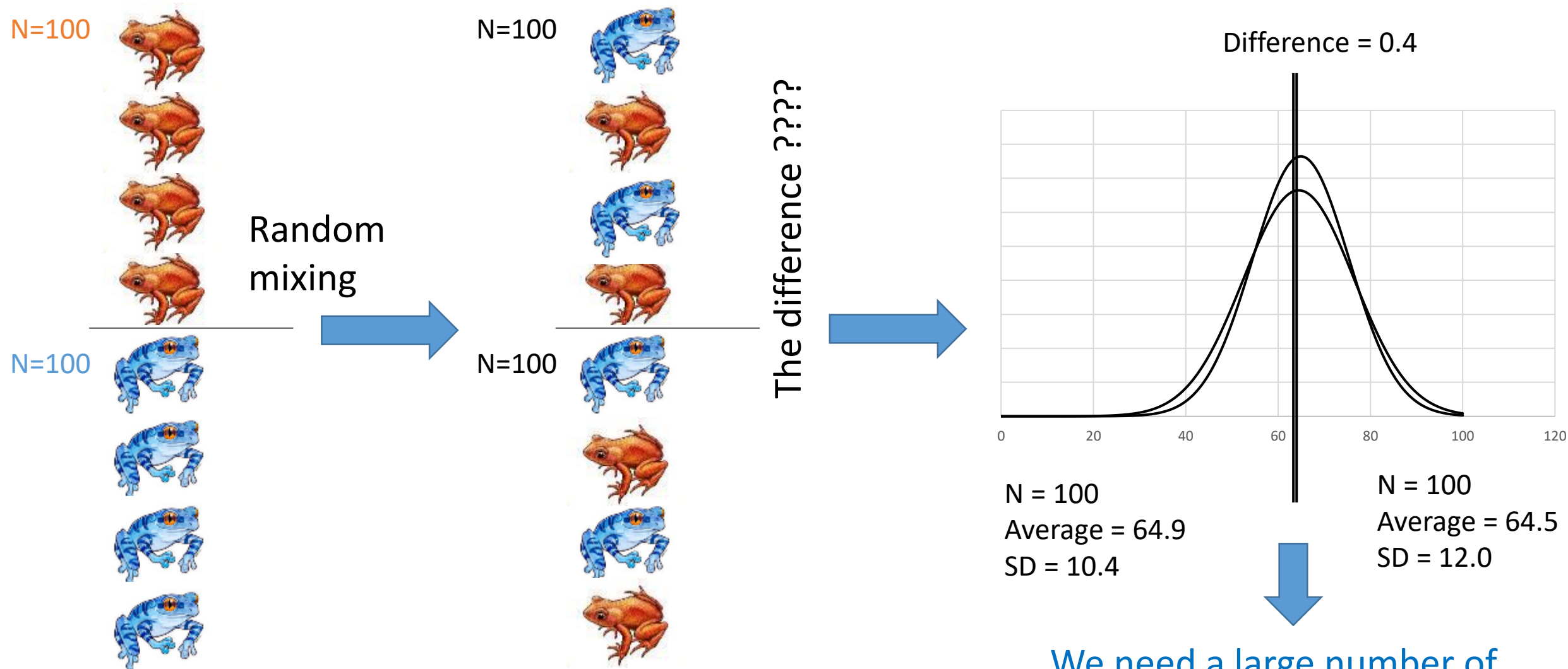
Example: permutation testing

We assessed the size of two species of frogs, sampling 100 individuals of each species.



Example: permutation testing

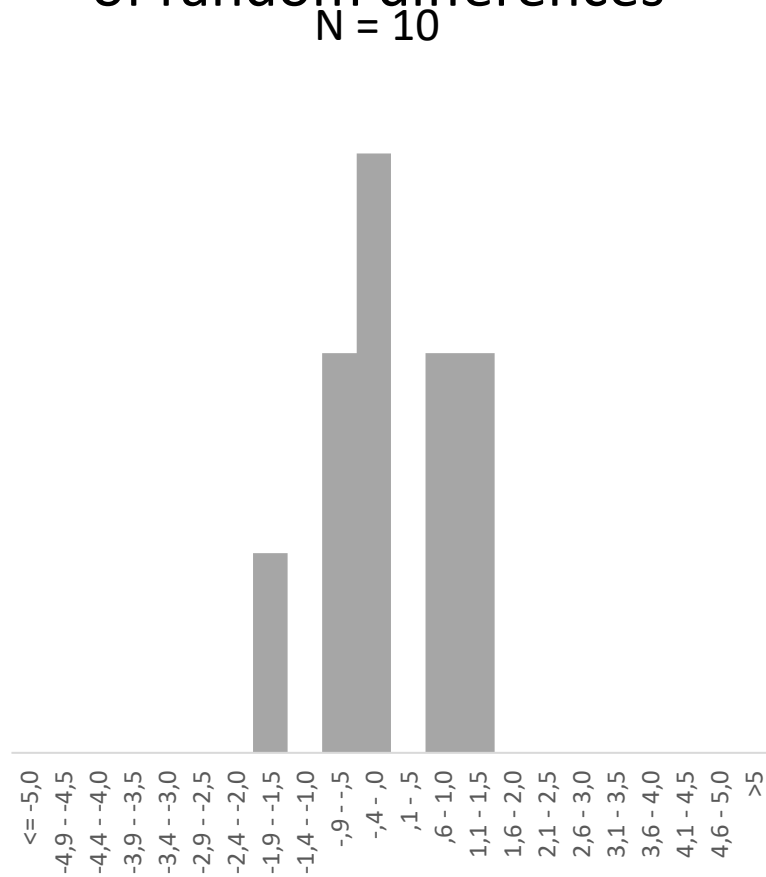
We assessed the size of two species of frogs, sampling 100 individuals of each species.



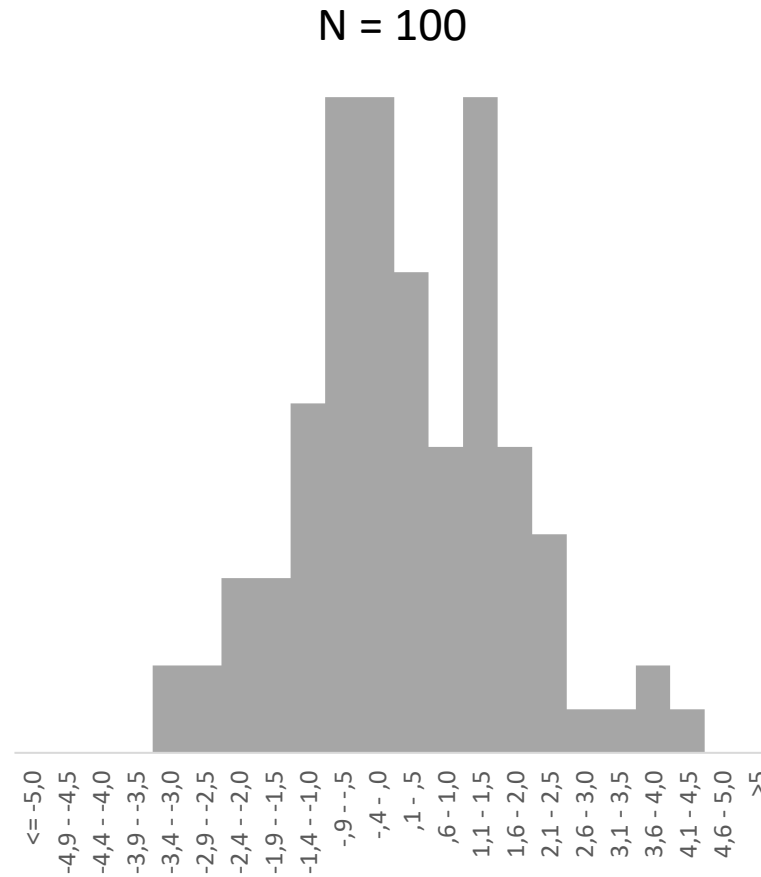
We need a large number of permutations to stabilize the result.

Results with different number of permutations

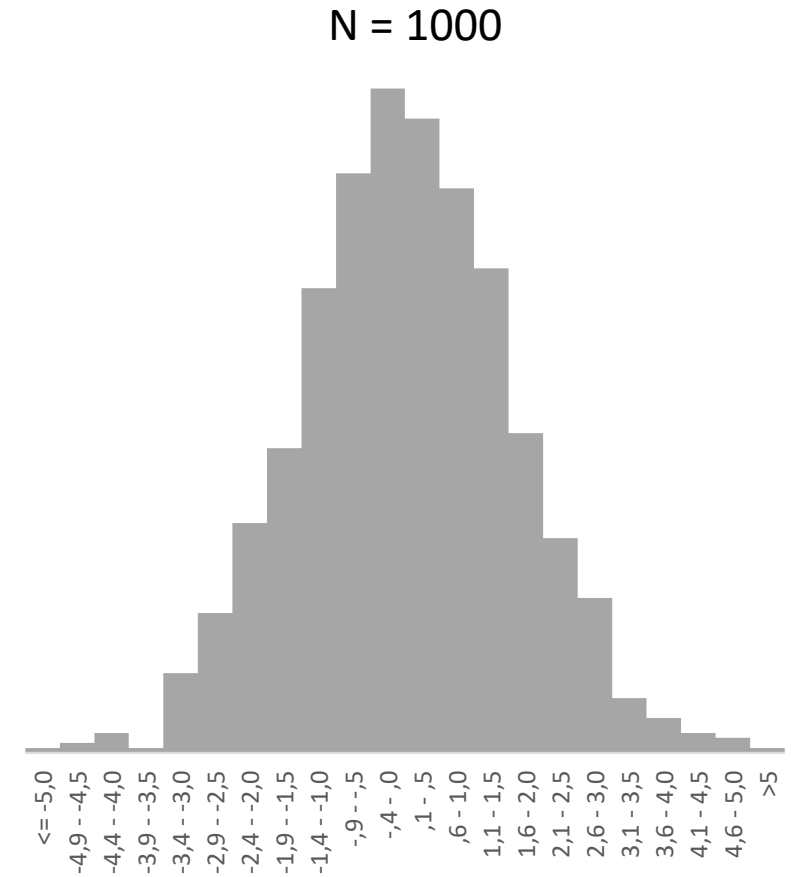
- As the number of permutations increases, we observe the formation of a distribution of random differences



Random differences



Random differences



Random differences

Random differences vs. observed difference

N=100



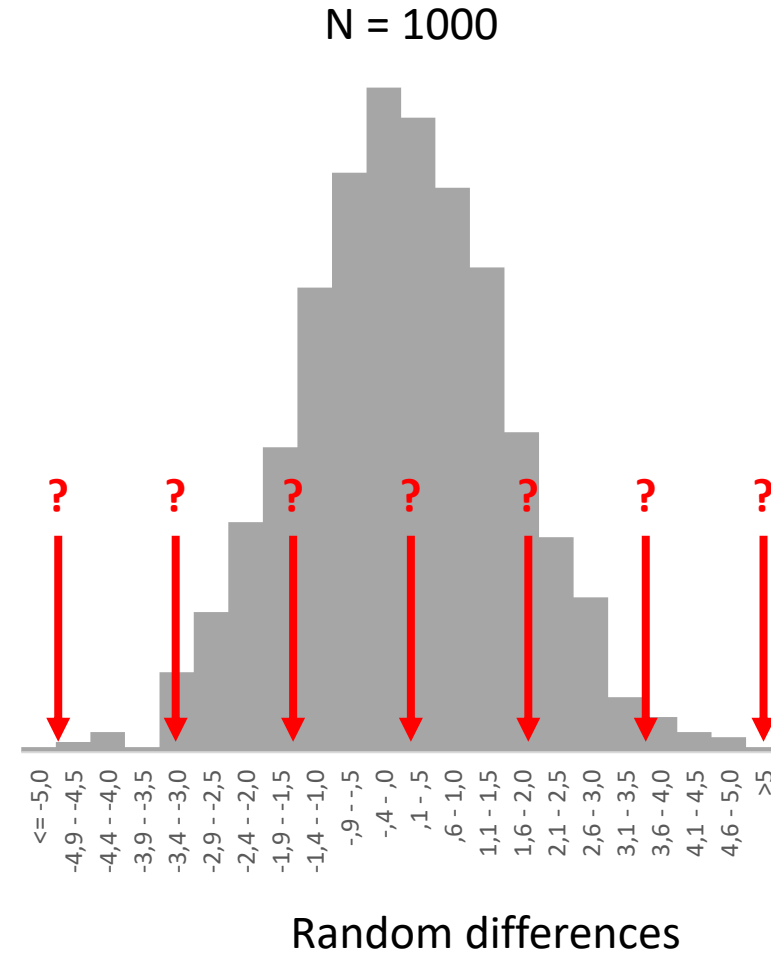
N=100



Difference = 10.6



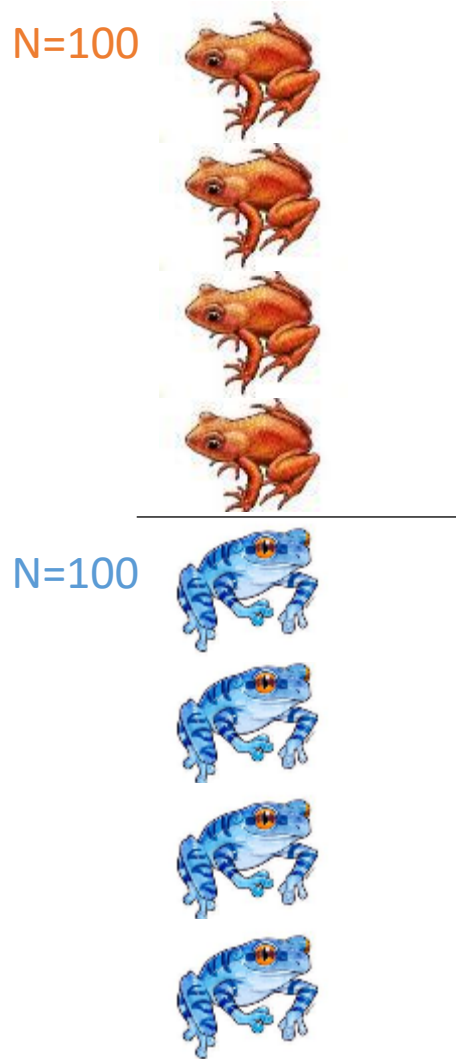
- We compare the real difference with the distribution of random differences



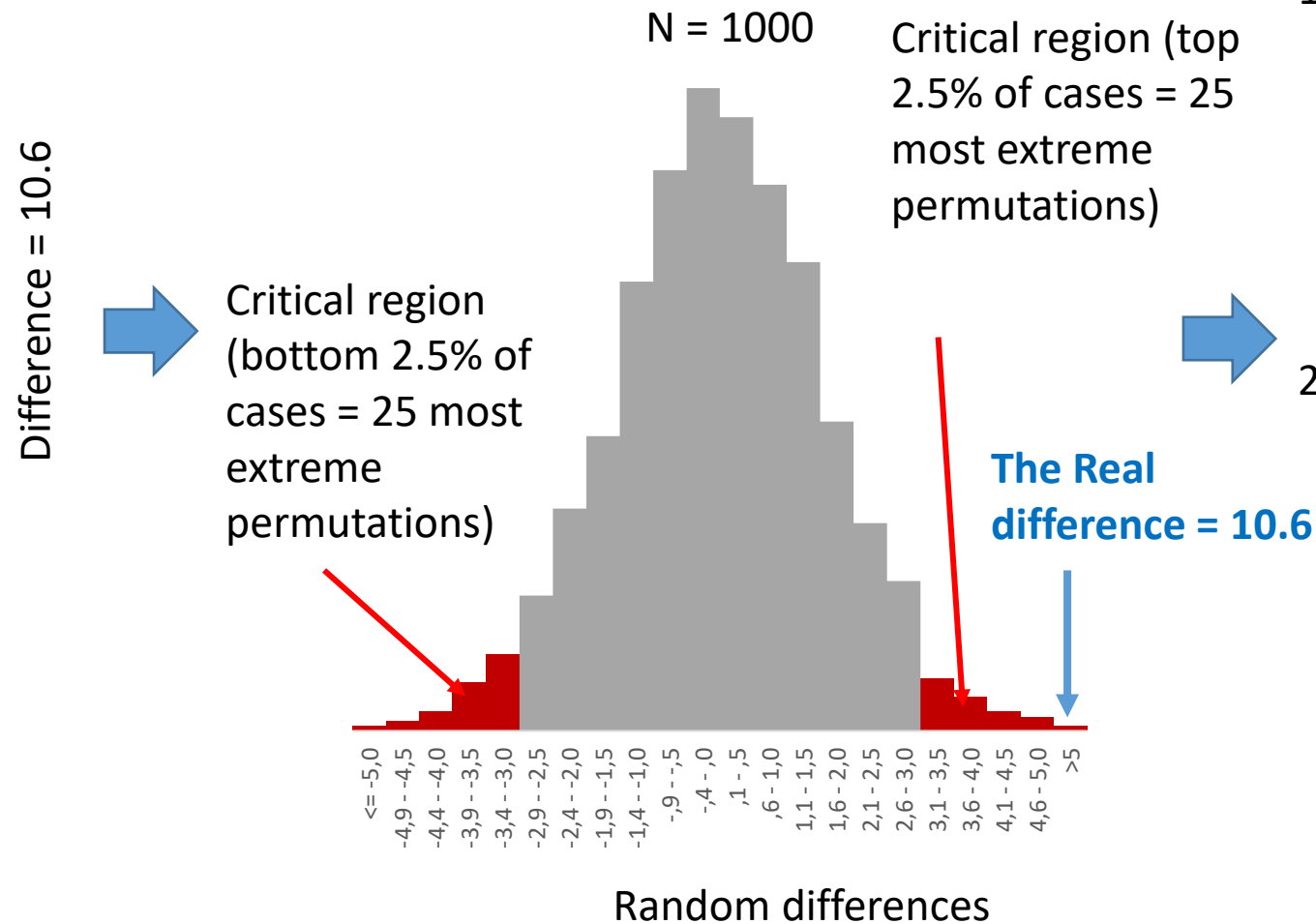
Random difference distribution and its use for testing

- Determine the critical region of the test statistic = what probability of the observed difference occurring by chance we can accept when rejecting the null hypothesis (i.e., declaring that the difference is not random)
- The most commonly used critical region of test statistics leading to a probability of random difference of 0.05 or 0.01 (called the **level of statistical significance, it is not a law of nature, only an assumption**)
- We compare our actual difference with the distribution of random differences and the specified critical region of this statistic
- If the true difference lies in the critical region, we say that at a given significance level we reject the null hypothesis
- For a given value of the test statistic, we are also able to determine the exact probability with which there are random differences greater than our observed difference = the probability that our observed difference is a mere coincidence

Statistical significance of the observed difference

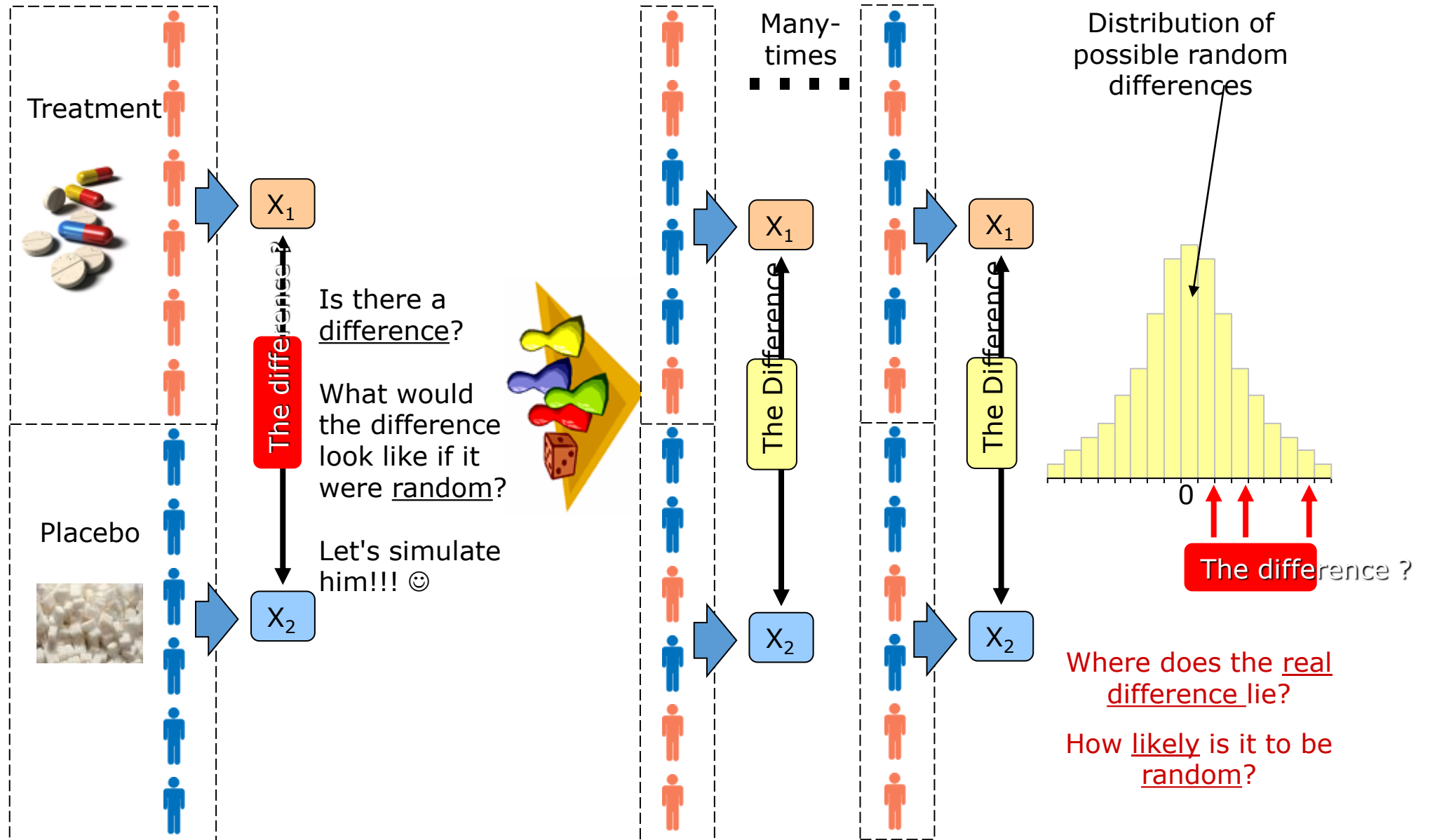


- The level of statistical significance will be 0.05 (5%)



- The real difference lies in the critical region of the test statistic = **we reject the null hypothesis that the means of the two groups are the same**
- There is only one random difference due to permutations greater than the actual difference = the probability that by chance alone there is a difference greater than the one we observed is $1/1000 = 0.001$ = the **statistical significance of the difference we observed is $p=0.001$** .

What does a random difference mean? Summary.



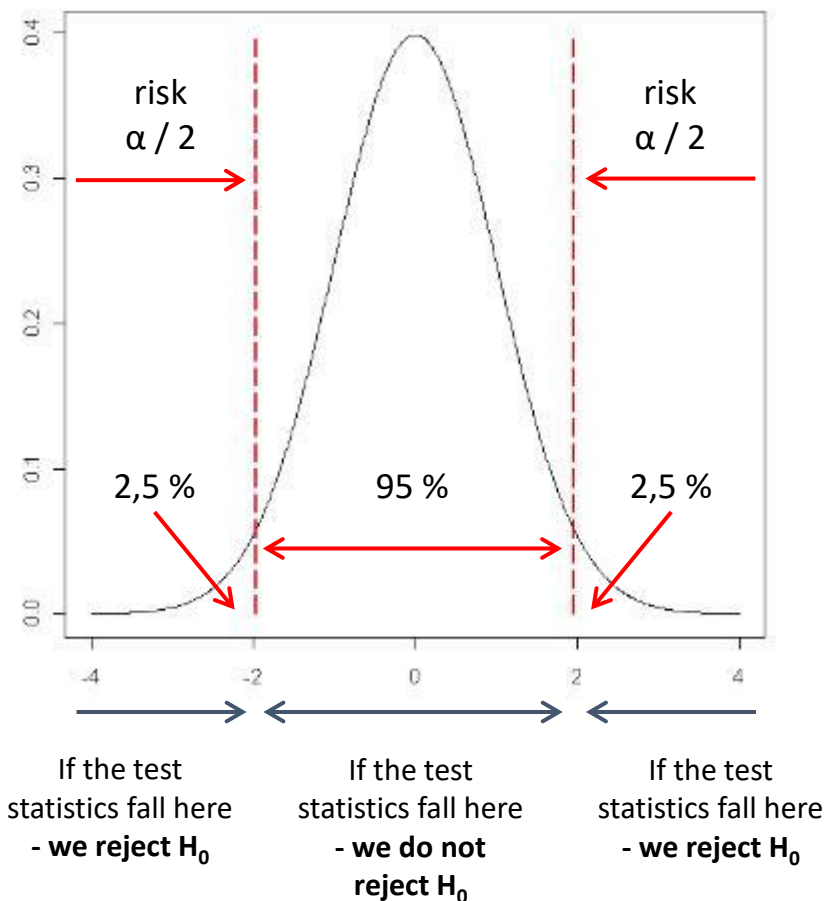
Rejection/non-rejection of the null hypothesis

- The value of the test statistic is compared with the quantile (critical value) of its distribution corresponding to the chosen significance level α .
- If the observed value of the test statistic represents a more extreme (less likely) value within the distribution corresponding to the null hypothesis than the critical value (quantile) corresponding to the chosen risk α , then the null hypothesis is rejected.

Rejection/non-rejection of the null hypothesis

Two-sided test at $\alpha = 0.05$

$$H_0 : \theta_1 = \theta_2 \quad H_1 : \theta_1 \neq \theta_2$$



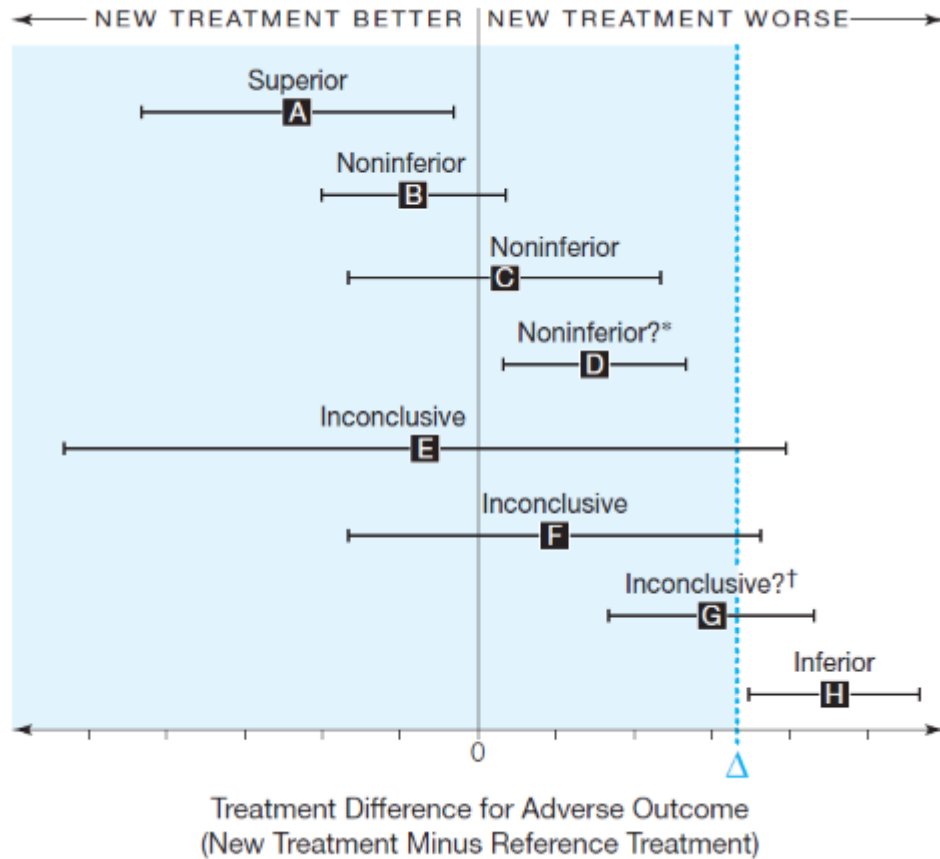
Distribution of random differences:

- Either the relevant model distribution
- Or the result of a simulation

Rejection of the null hypothesis:

- Our test statistics fall into the critical field of
- The derived exact value of p is smaller than the p associated with the critical region

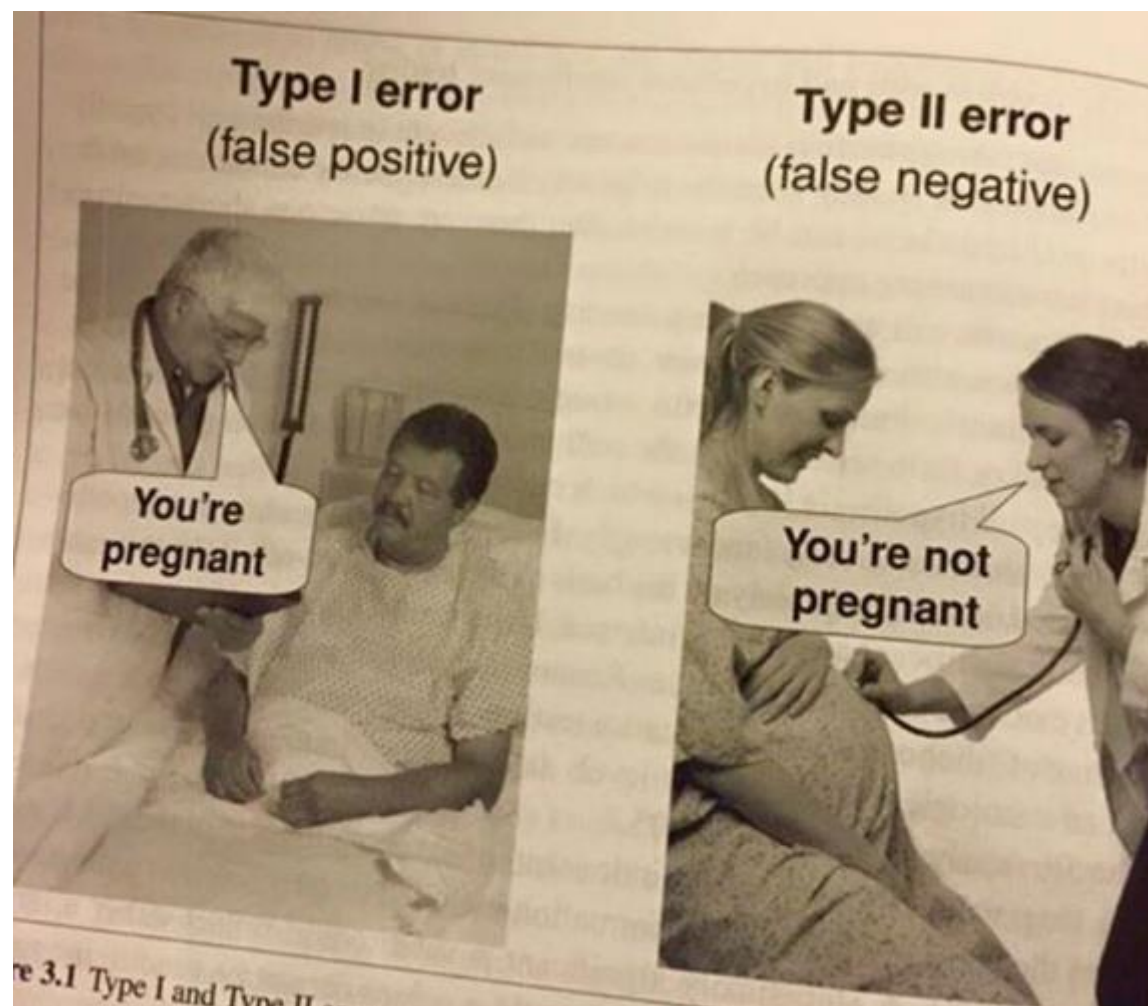
Testing with confidence intervals



- The principle of confidence interval testing is to calculate a confidence interval for a given difference or measure of the relationship between variables and compare it with a reference value (e.g. 0 in the case of a difference).
- If the interval does not contain this reference value, it is equivalent to demonstrating the statistical significance of the difference at that significance level (95% confidence interval is equivalent to a significance level of 0.05)

Source: Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006 Mar 8;295(10):1152-60.
Statistics and Informatics Services Group, Department of Reproductive Health and Research, World Health Organization, Geneva.

Possible errors in hypothesis testing



What can happen when making a decision

- Given the null hypotheses we have four possible outcomes of the decision process:

Decision	Reality	
	H_0 applies	H_0 does not apply
H_0 will not be rejected	correct acceptance of valid null hypotheses	Type II error
H_0 we reject	Type I error	correct rejection of an invalid null hypothesis

- We can make two mistakes when making a decision, we can make two wrong judgments.

Analogy with the judicial process

- We honor the presumption of innocence = we assume the null hypothesis is true.
- **We require evidence to prove guilt = we want to show, based on data, that the null hypothesis does not hold.**
- If we have insufficient evidence, the percentage of those convicted who are innocent will increase = **Type I error**, but at the same time the percentage of those convicted who are actually guilty will also increase = **correct rejection of the invalid null hypothesis**.
- If we demand a lot of evidence, the percentage of innocents who will be acquitted will increase = **correct acceptance of a valid null hypothesis**, but at the same time the percentage of guilty who will be acquitted will also increase = **Type II error**.

Probability of the results of the decision-making process

Decision	Reality	
	H_0 applies	H_0 does not apply
H_0 will not be rejected	the right decision $P = 1 - \alpha$	Type II error $P = \beta$
H_0 we reject	Type I error $P = \alpha$	the right decision $P = 1 - \beta$

- As can be seen from the analogy with the lawsuit, one cannot simultaneously minimize α and β . In practice, it is necessary to guard α more closely \rightarrow we set a maximum threshold for α in advance (the "level of significance" of the test) and minimize β under this condition.

What does "failing test statistics" mean

- If the value of the test statistic is greater than the risk quantile α , then two situations could occur:
 - 1. either H_0 is valid and we have observed an unlikely phenomenon**
 - 2. or H_0 does not apply**
- We work with risk α , i.e., unlikely events are part of our risk, so in this case we choose option 2 and reject H_0 .

Statistical test errors as a consequence of our decision

- Statistical significance itself means only the probability that the observed difference or relationship between variables is due to chance
- At the point where we make a decision based on this probability that the null hypothesis is invalid, we accept the probability (corresponding to a given statistical significance) that this decision is wrong and that the null hypothesis is in fact valid (the difference is due to chance alone)
- Every decision we make to reject the null hypothesis contains a Type I error snake



P-value

- The p-value expresses the probability under the validity of H_0 , with which we would obtain the same or a more extreme value of the test statistic (given, of course, the one-sidedness or two-sidedness of the test).
- Thus, the lower the p-value of the test, the less likely the test indicates that the null hypothesis is true. In other words, if we get a p-value "close to zero" when we evaluate the statistical test (again, two thresholds are accepted by default: 5% and 1%), this means that our null hypothesis has very little support in the observed data and can be rejected.

P-value

- The resulting p-value is then compared to the chosen significance level α , with the null hypothesis being rejected when the p-value of the test falls below this level.
- Thus, it can be said that the moment the risk of a false positive result associated with rejecting the null hypothesis falls below a selected level (e.g. 5% or 1%), then we reject it.
- The p-value can be understood as a numerical indicator of the validity or invalidity of the null hypothesis expressed on a probability scale. And like any indicator, the p-value can indicate a bad outcome, because we must always be aware that we are at risk of both a Type I and a Type II error.

The strength of the test

- We denote the probability of type II error by β .
- $1 - \beta$ is called the power of the test and expresses the probability that we reject H_0 at the moment when H_0 is indeed invalid.
- We try to optimize the power of the test while maintaining the significance level of the test $\alpha \rightarrow$ principle of calculating the experimental sample size before the study
- Optimizing test power and sample size in advance is not trivial; we can run into many problems - biological limits, ethical limits, financial limits.

Factors affecting the strength of the test

- **Sample size:** the more observations (information about the validity of the null hypothesis), the greater the power of the test. As with confidence intervals, the power of the test increases with the square root of n .
- **Effect size:** the size of the difference in the unknown parameters also affects the power of the test. It is always easier to identify a large effect size as significant, e.g. a large difference in the mean prostate volume of two populations. Conversely, it is harder to show a smaller effect (smaller difference) as significant.
- **Data variability:** data variability increases the variability of the estimates, making the decision on H_0 difficult. The more variable the observed values are, the more data will be needed to accurately estimate the effect size (difference).
- **Significance level:** if we lower the significance level of the test (e.g. choose 0.01 instead of 0.05), it will be more difficult to reject $H_0 \rightarrow$ the power of the test will decrease.