Ing. Anna Horňáková, Ph.D.

Institute of Hygiene and Epidemiology of the 1st Faculty of Medicine

FIRST FACULTY OF MEDICINE
Charles University

STATISTICS IN EPIDEMIOLOGY

Descriptive statistics

- **Descriptive statistics** deals with the collection, arrangement, description and effective summarizing data sets.
- Ideally, statisticians compile data about the entire population (an operation called census). This may be organized by governmental statistical institutes. Descriptive statistics can be used to summarize the population data.
- When a census is not feasible, a chosen subset of the **population** called a **sample** is studied.
  - Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting. The descriptive statistics can be used to summarize the sample data.

Inferential statistics

- **Inferential statistics** draws the conclusions from the data from the part of the population (**sample**) and deduce general statements about the **population**. Conclusions can be loaded with errors -> there can be calculated, i.e. we can determine the degree of reliability.
- It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness, e.g.:
  - answering yes/no questions about the data (hypothesis testing),
  - estimating numerical characteristics of the data (estimation),
  - describing associations within the data (correlation),
  - modeling relationships within the data (for example, using regression analysis).

Population and sample

- **Population** (base set) – given by determining the elements:
  - by enumeration – has a finite range;
  - by delineating common features – infinite range;
- **Sample** (sampling set) – part of the population that we follow as part of the research:
  - representative = the structure corresponds to the structure of the population;
  - selective = gives a distorted picture of the studied population;
  - intentional = subjective, based on expert opinions and estimates;
  - random = "objective", elements from the population are selected completely randomly

Population and sample

- Random sampling:
  - simple random sampling: guarantees that each element of the population has an equal chance of being included in the set, e.g. a lottery;
  - mechanical (systematic): based on a certain predetermined arrangement of the population, e.g. choosing a step (selecting every third card from the card file);

Variables

- **quantitative** = expressed by a number
  - continuous – take on values of a certain interval: e.g. height, weight;
  - discrete – there are only finitely many possible states of the character: e.g. number of children;
- **qualitative** = expressed by the text
  - nominal – multiple mutually exclusive classes (without ordering): e.g. gender, marital status;
  - ordinal – incompatible but ordered classes: e.g. highest level of education, degree of pain;

Measures of central tendency

- **Measures of central tendency** = Position measures are descriptive measures that we use when, in a set of data (e.g., the results of an observation), we need to determine the value around which the data are centred, to establish some sort of "centre".

1) Mean (arithmetic, geometric, …)
2) Median
3) Mode

Arithmetic mean

- The average of the values in a selection is calculated by dividing the sum of all the values by the range of the selection (n). So, if we have n observations: $x_1,x_2,x_3,\ldots,x_n$, then we calculate the mean as follows: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- The arithmetic mean of $\bar{x}$:
  - we use when we can really add the numbers, i.e. the signs are quantitative, measured on a numerical scale;
  - should not be used for ordinal characters because of the arbitrariness of the choice of ordinal scale;
  - is sensitive to outliers.

Geometric mean

- The geometric mean $\bar{x}_G$ is calculated as the nth root of the product of the observations, i.e.: $\bar{x}_G = \sqrt[n]{x_1 x_2 \dots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Median

- If we have the observations arranged in ascending or descending order, then the **median** $\tilde{x}$ is the value that divides the observations into two groups of equal size.
- If we have an odd number of ordered observations, then the median is the middle one.
- For an even number, the median is usually the average of the two middle observations.

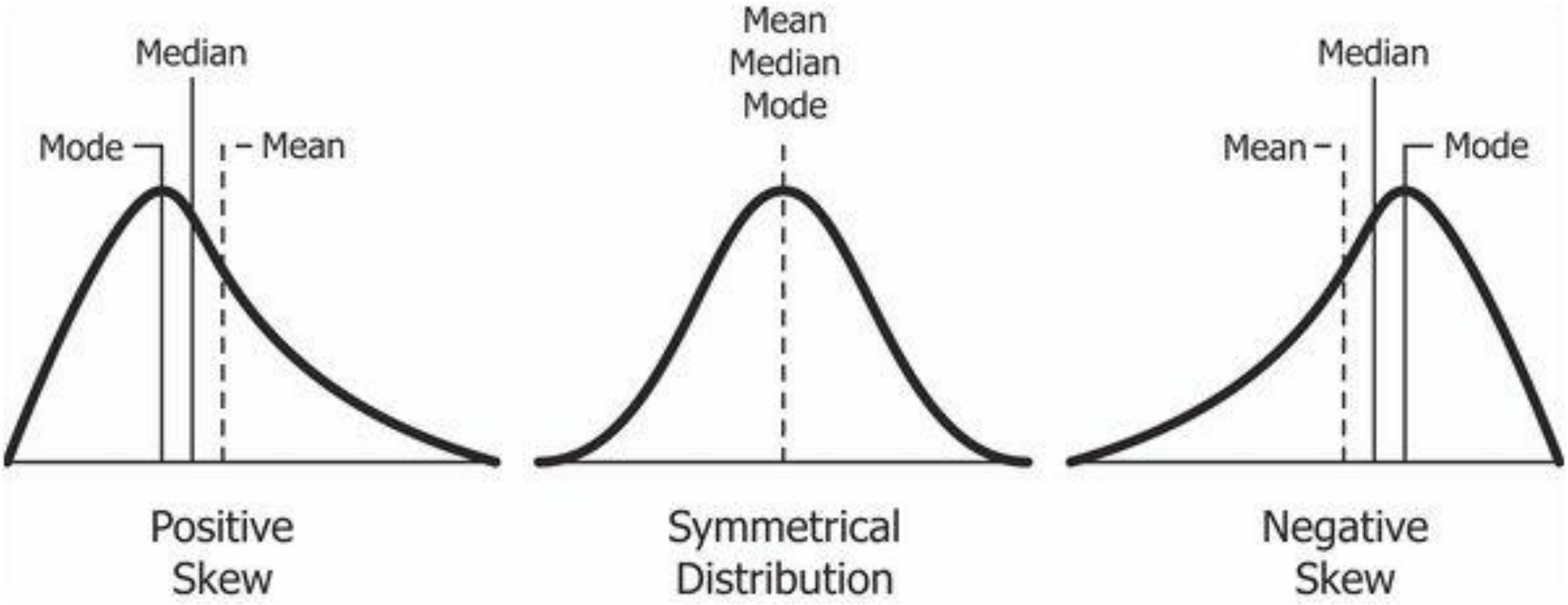
Median

- The median only uses information about the order of the values, and therefore it only makes sense to use it for quantitative and ordinal quantities.
- In pharmacology, the median is used under the name *50% effective dose* (ED50) or *50% lethal dose* (LD50) as a characterization of the efficacy of a product.

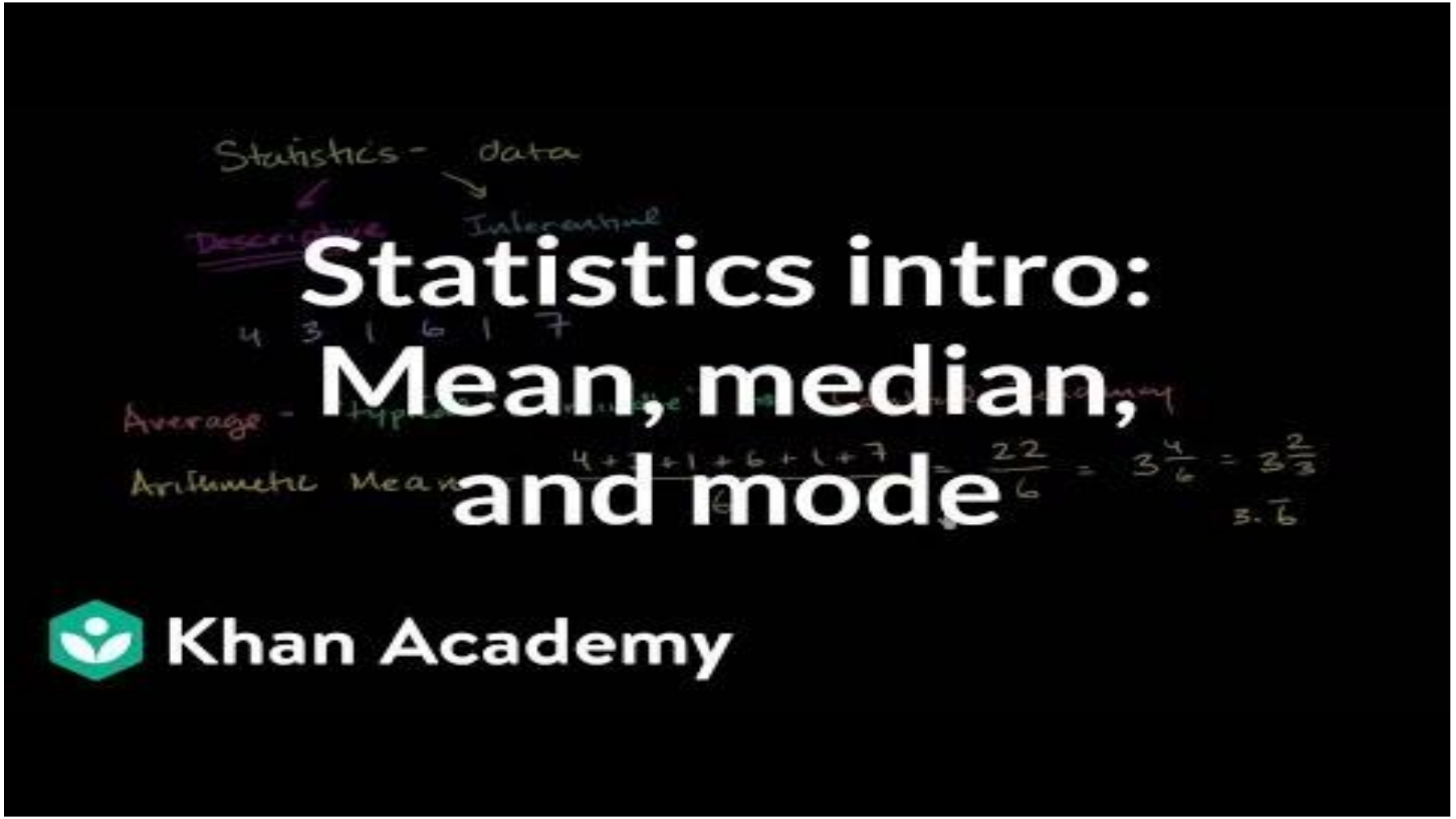
Mode

- The mode $\hat{x}$ is the value that occurs most frequently in the data set.
- The mode is:
  - Qualitative, especially nominal traits.
  - It is not affected by the values of all the elements in the sample.
  - It is mainly used to capture the most typical value of a trait, e.g. the duration of a particular disease or the age at which the disease of interest occurs.

Measures of central tendency



Measures of central tendency



Measures of variation

- **Measures of variation** = Measures of variation (dispersion) tell us how close or far apart the values of a variable are from each other. They assess the dispersion of the values of a statistical population around some mean value.
- Range
- Variance
- Standard deviation

Range

- The range (variation width R) is one of the measures of variation.
- It is the difference between the highest (x_{max}) and lowest (x_{min}) value in the data, $R = x_{\text{max}} - x_{\text{min}}$
- The range is a useful measure, but has the disadvantage that it depends on outliers, so it can give a very misleading picture (e.g. a sample can have a very large spread even though most observations will be close together).

Variance

- The **variance** s^2 is a measure of the dispersion about the mean arrived at by calculating the sum of the squared deviations from the mean and dividing by the sample size.
- However, when calculating the sample variance, we usually do not divide the sum of the squares of the deviations by n, but by $(n-1)$, because this gives a better estimate of the total variance of the population. The divisor $(n-1)$ is called the number of degrees of freedom of the variance.
- The general formula looks like this:

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

Standard deviation

- The sample standard deviation s is the square root of the variance. The standard deviation is in the same units as the original values.

$$s = \sqrt{s^2}$$

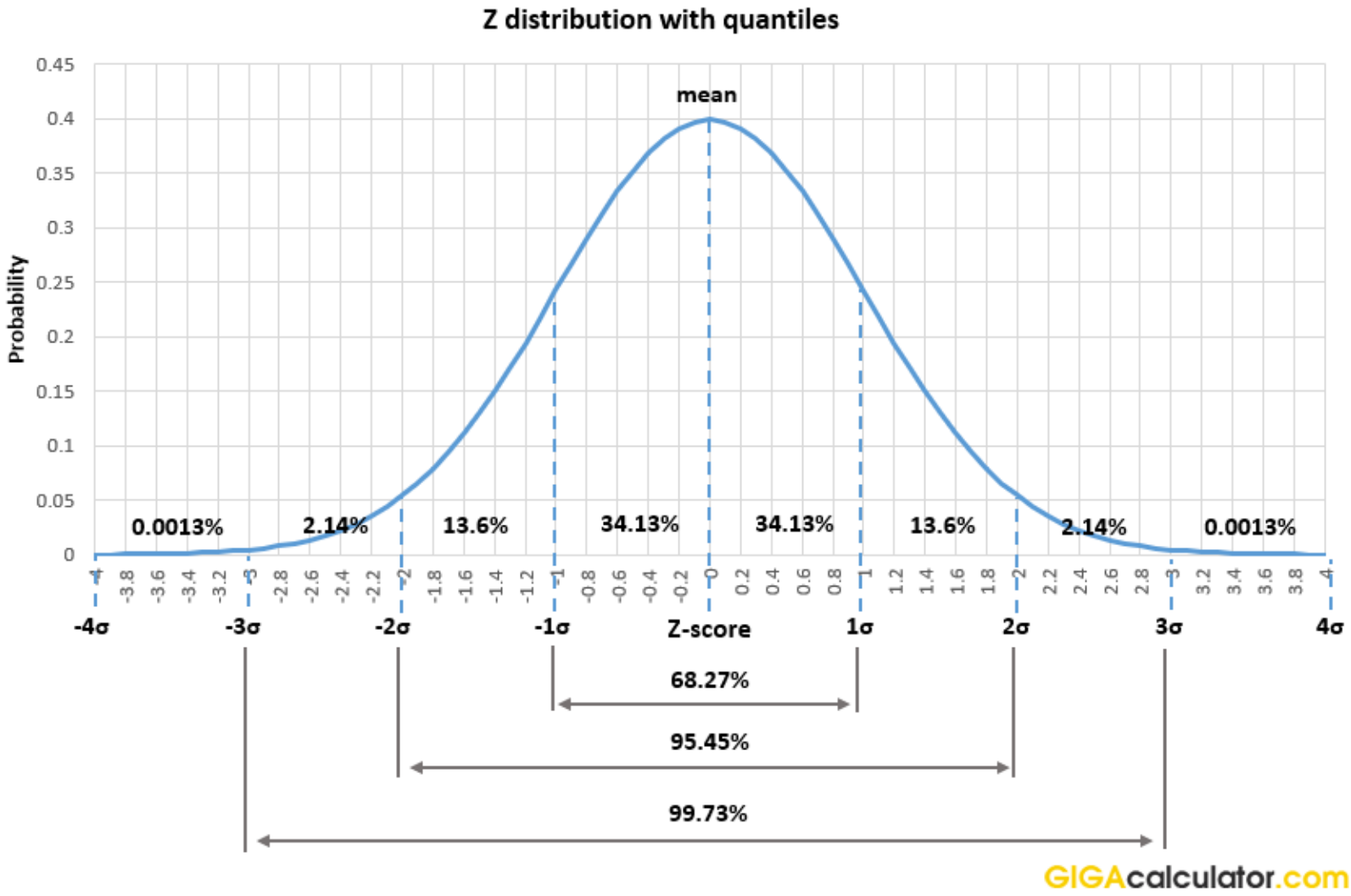
- The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values.
- A low standard deviation indicates that the data points tend to be close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

Measures of variation

Normal distribution

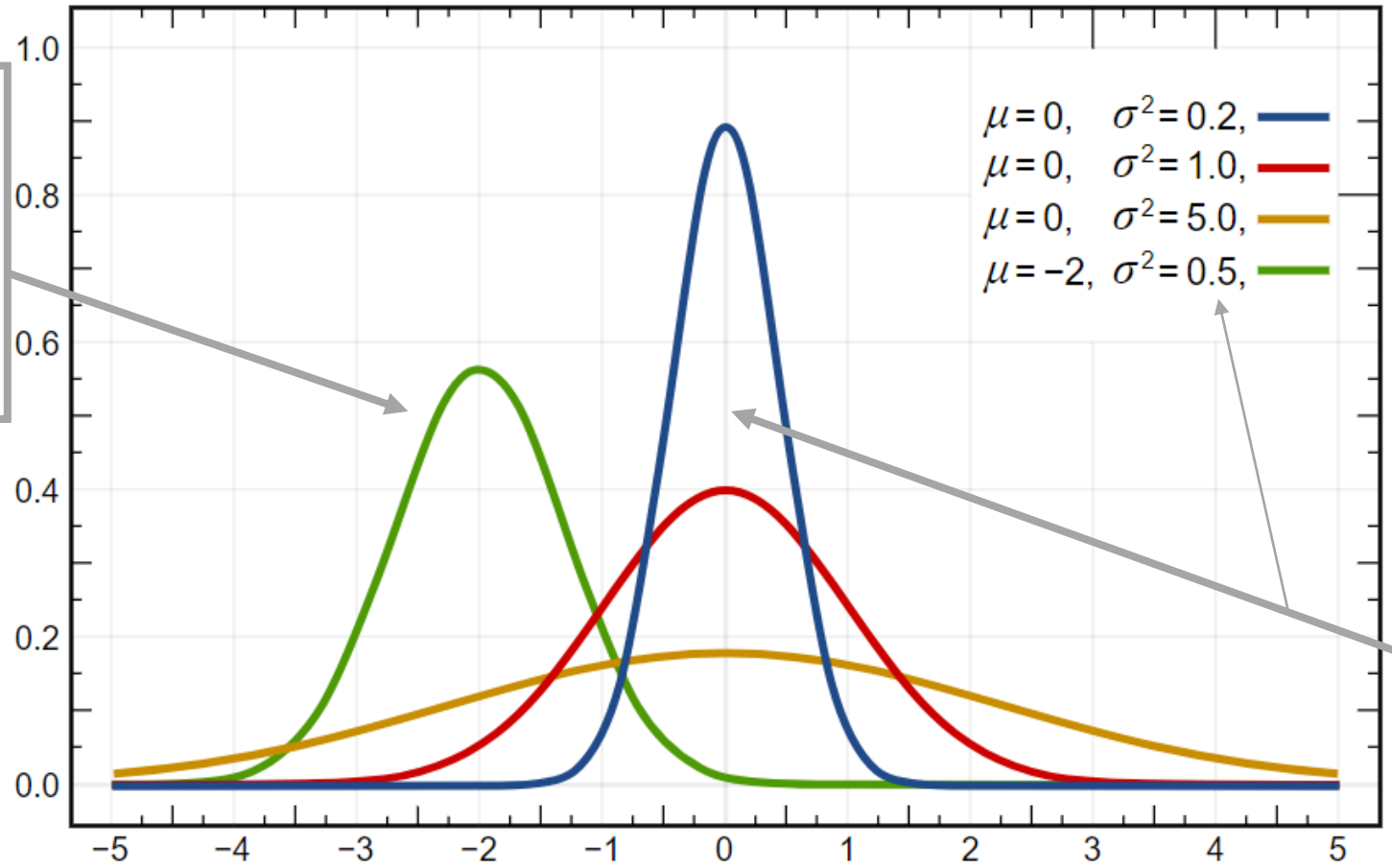
- Normal = Gaussian = bell-shaped curve
- The mean, mode and median are all equal
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

Standard normal distribution



Normal (Gaussian) distribution

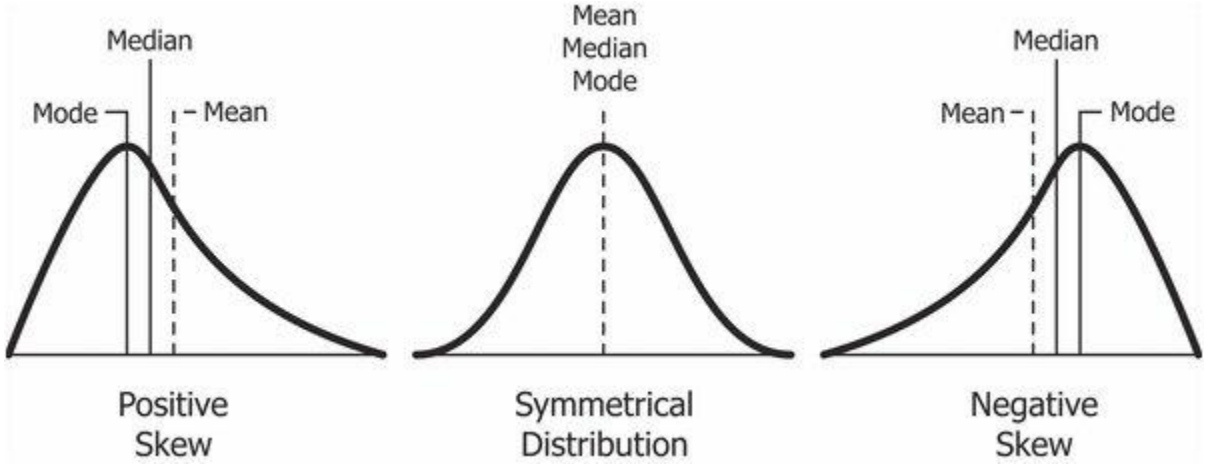
The mean of the random variable determines the distribution position on the x axis (green curve for $\mu = -2$)



The standard deviation determines the shape / extension of the graph.

Asymmetrical Distribution

- If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry. Symmetry means that one half of the distribution is a mirror image of the other half. For example, the normal distribution is a symmetric distribution with no skew. The tails are exactly the same.



Time for a break

Epidemiological methodology

- General methodology -> a statistical analysis -> it includes methods of clarification, causality, correct manipulation of qualitative and quantitative variables
- Auxiliary methods are based on other disciplines, e.g. clinical fields, bacteriology, virology, parasitology, etc.

Epidemiological methodology – aims:

- To study the history of the health of society, the dynamics of diseases and changes in their character, and on the basis of this knowledge to predict and prepare projects of further development.
- To measure health and disease under the appropriate conditions and defined concepts.
- To know the individual diseases and their course, to identify the relevant symptoms.
- To evaluate the activity and effectiveness of health services and measures.
- Search for the causes of illness and health conditions.

Types of epidemiological studies

- **Observational**
  - *descriptive* = describe the distribution of the disease in the population
  - *analytical* = explain the causes (factors, determinants) of the disease; serve to verify epidemiological hypotheses;

- *Interventional*: clinical, field studies

Descriptive studies

- Description of the general characteristics of the distribution of the disease in the population with a focus on:
  - particularities/specifics of persons and phenomena;
  - the time of occurrence (the characteristics of the site and the time at which it occurs);
  - other circumstances.
- It often uses demographic data (yearbooks, health service reports, World Health Organization information, health insurance statistics, ...).

Descriptive studies

1) **Case reports, case report series** -> study individuals

2) **Correlation studies = Ecological studies** -> all (or at least some) basic data (in particular the exposure to the risk factor and disease incidence) are surveyed *at population level, not at individual level*;

Descriptive studies

3) **Cross-sectional studies** -> detects the presence or absence of exposure to the observed risk factor and the occurrence of the disease at the same time in the past or in the presence; allows to estimate the percentage of patients and the percentage of people with a risk factor;

- *However, we can also test the size of the association between exposure and disease, but usually without the possibility to determine whether the exposure is preceded by a disease or vice versa => cross-sectional study, therefore, is often ranked between both descriptive and analytical studies.*

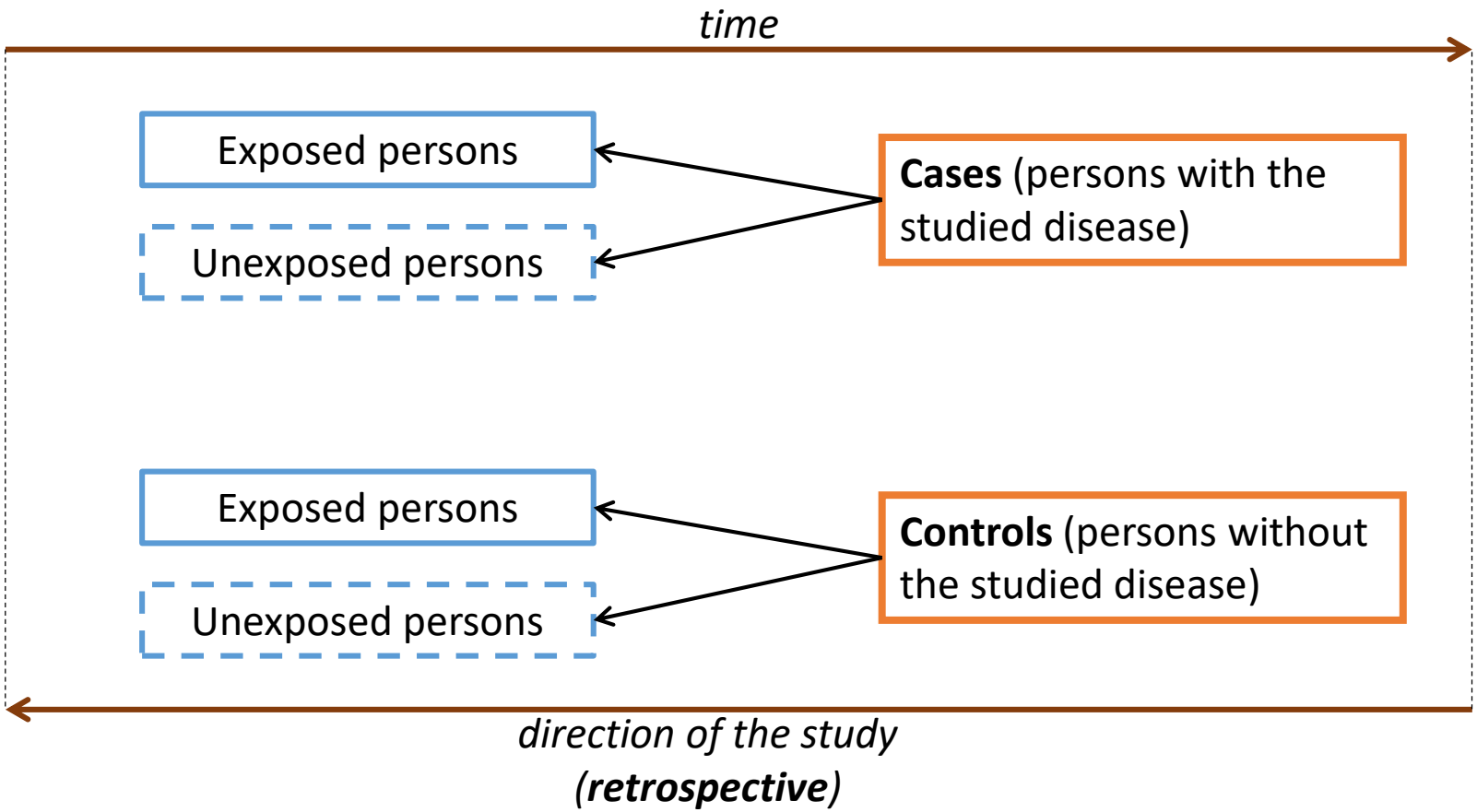
Analytical studies

- Sorts phenomena by sex, age, location, time -> ***retrospective*** or ***prospective***.
- It usually examines hypotheses built in the descriptive phase.
- Analytical epidemiology is a source of new hypotheses that are re-examined, followed by a return to the first stage for further information.
- Examples: case-control study, cohort study

Case-control study

- is observational analytical study into which persons are selected *according to the presence (cases) or absence (controls) of the of the studied disease*;
- is *retrospective study* -> for both cases and controls the data about their exposure are then collected;
- is suitable for the study of rare diseases;

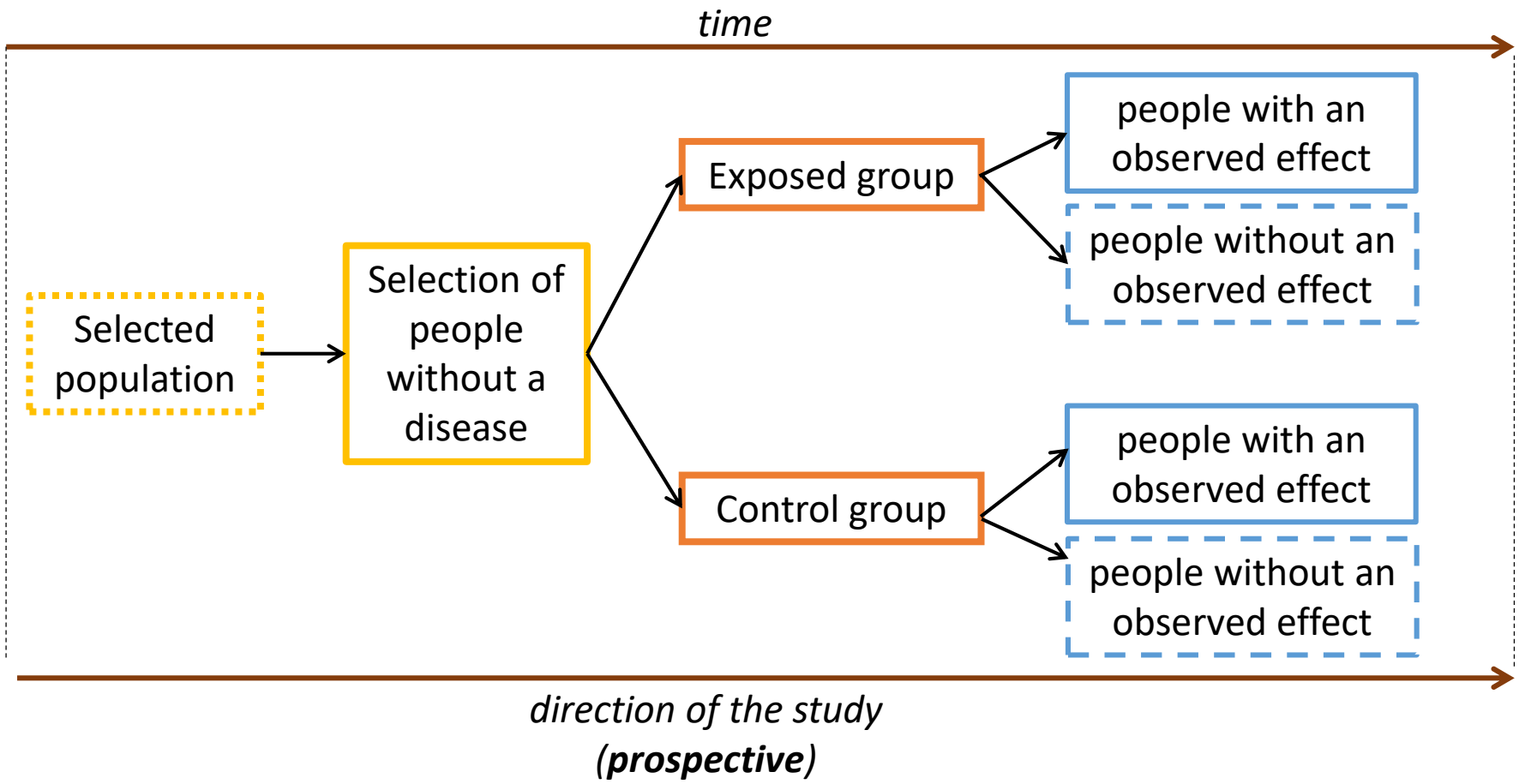
Case-control study



Cohort study

- is observational analytical study in which individuals are divided according to *the presence or absence of exposure at the beginning of the study*;
- at the time of exposure determination and entry into the study, no person included in the study is expected to show the underlying disease;
- is *prospective study* -> people in the study are followed for a longer period of time needed for the development and clinical manifestations of the disease;
- suitable for the study of rare exposures;

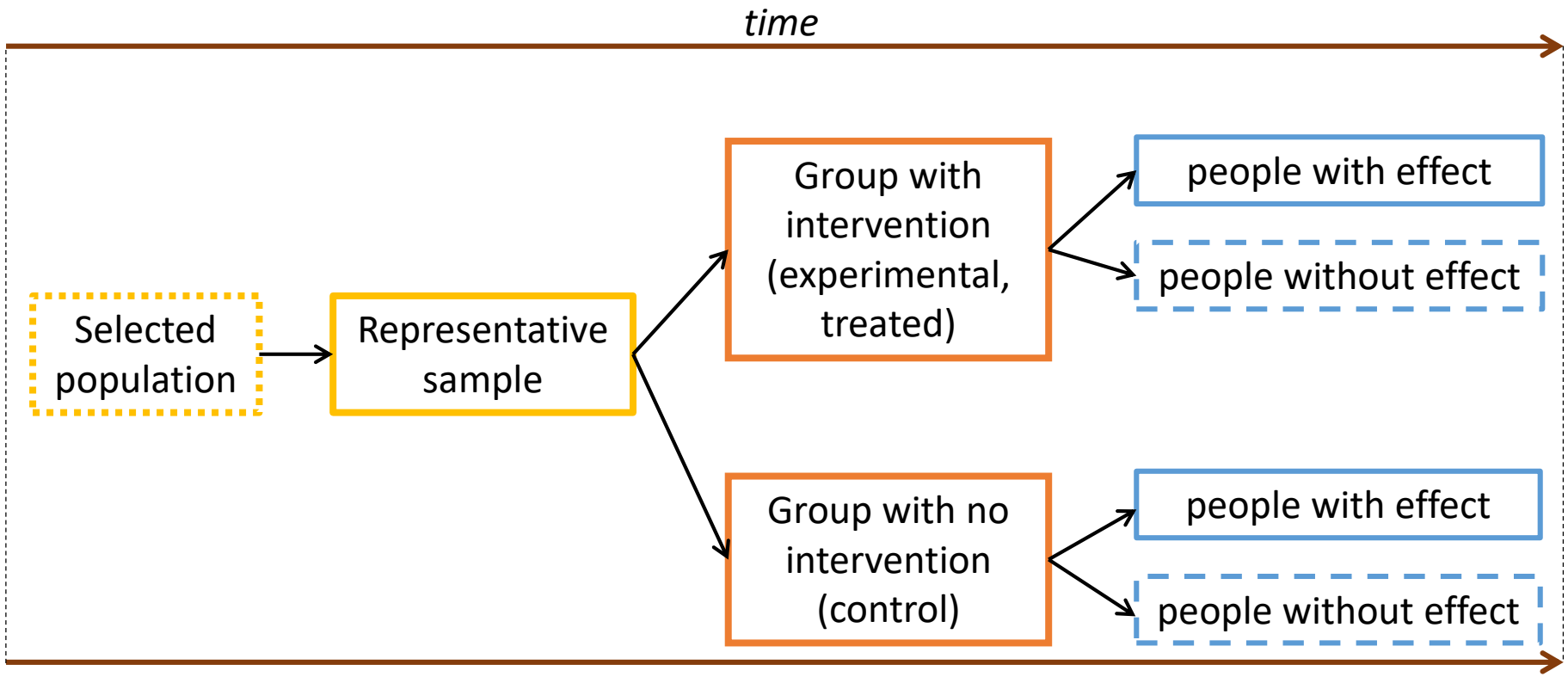
Cohort study



Interventional studies

- are under the direct control of the researchers who determine what exposure mode (treatment) will be subjected to who
- are *always prospective* because it start with assigning the exposure and are waiting for a reaction - sickness, symptom relief, etc.
- **Clinically controlled study**: Detects the effect of specific treatment on the patient in terms of suppressing symptoms, reducing the risk of death, disability or other complications.
- **Preventive (interventional) study**: Detects the effect of preventive measures on individuals without a given disease. It monitors the risk reduction.

Interventional studies



direction of the study
*(always **prospective**)*

Errors in epidemiological studies

- The goal of epidemiological research is to find out the true relationship between exposure and effect.
- The results of observational studies may be distorted by:
  - various external effects (confounding);
  - systematic errors in planning (errors in study design);
  - the way of data collection;
  - data analysis (selective or information bias).

Errors in epidemiological studies

- *Valid studies are considered to be those that do not suffer from these deficiencies.*
- The **validity** of the study is affected by the systematic errors, the **accuracy** of the study is determined by the size of the random errors.
- Measures of **reliability** are primarily important for what they reveal about the validity of a measure, and this concept is emphasized for both design and interpretation of reliability studies.

Bias

- Bias = Any **systematic error** that arises when collecting, analyzing, interpreting, publishing or checking the data and leads to conclusions that are systematically different from reality.
- **Recall bias** -> caused by the person under investigation in questionnaire surveys or controlled interviews;
  - Ability and willingness to realize the previous event (often for a long period of time) may vary depending on whether the person is a case or control.
- **Interviewer bias** -> can occur if the interviewer who conducts a controlled interview knows what is the hypothesis of the study;

Bias

- **Selection bias** -> e.g. in so-called *hospital controls*
- On the one hand, patients often do not represent the general population (they smoke more, drink more alcohol, are poorer, and live in worse conditions than the general population).
- On the other hand, there is less bias caused by the person being investigated, because even control is the patient.

Confounding factor

- The actual association is masked by a **confounding factor**. Ignoring the effect of confounding factor leads to erroneous estimates of the size of the effect.
- Socio-economic factors (such as education or income levels) or demographics (place of residence, maternal age during delivery) can well predict a wide range of diseases, although there is no direct biological link -> they are basically a surrogate, proxy factor, and therefore treated as misleading/confounding factors.

Confounding factor

- Distortion of association indicators (e.g. RR or OR) that arise because we have not reviewed other variables that are a risk factors for a studied phenomenon or disease.
- If a *confounding or disturbing factor* is present, the association observed is not a true association between exposure and consequence but is a link between the misleading factor and the exposure and consequence.

Measures of association

- **Association** -> the relationship between two or more variables (for example, risk factor and disease);

- Positive = there is a consistent relationship between the two variables and the values of both observed phenomena rise (or decrease) simultaneously;
- Negative = inverse relationship where the values of the variables move in the opposite direction (one value rises, the other decreases);

Measures of association

- Direct association = the relationship between two variables, without the participation of another factor;
- Indirect association = the apparent relationship, mediated by another (third, confounding) factor that is associated with both the risk factor and the disease.

Measures of association

- **Absolute risk** (AR) is the probability or chance of an event. It is usually used for the number of events (such as a disease) that occurred in a group, divided by the number of people in that group.
- Absolute risk is one of the most understandable ways of communicating health risks to the general public.

Measures of association

- **Relative risk** (RR) is the ratio of the probability of an event occurring (for example, developing a disease, being injured) in an exposed group to the probability of the event occurring in a comparison, non-exposed group.
  - A risk ratio of 1 means there is no difference in risk between the two groups.
  - An RR of < 1 means the event is less likely to occur in the experimental group than in the control group.
  - An RR of > 1 means the event is more likely to occur in the experimental group than in the control group.
- RR is used in cohort studies.

Measures of association

- **Odds ratio** (OR) is the ratio of the chances of a certain event occurring depending on the other event (events A and B). It thus quantifies the strength of the relationship between two quantities. OR works with dichotomous variables (taking on two states) and is most often used in case-control studies.

Measures of association

- **Attributable risk** (AR) is the difference in rate of a condition between an exposed population and an unexposed population.
  - Attributable risk is mostly calculated in cohort studies, where individuals are assembled on exposure status and followed over a period of time.
- **Population attributable risk** (PAR) is the reduction in incidence that would be observed if the population were entirely unexposed, compared with its current (actual) exposure pattern.

Time for a break

Correlation

- Dependence or association is any statistical relationship, whether causal or not, between two random variables or bivariate data.
- **Correlation** is any of a broad class of statistical relationships involving dependence, though in common usage it most often refers to how close two variables are to have a linear relationship with each other.

Correlation

- Correlations are useful because they can indicate a predictive relationship that can be exploited in practice.
- Example: an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather -> there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling.
- In general, **the presence of a correlation is not sufficient to infer the presence of a causal relationship** (i.e., correlation does not imply causation).

Correlation coefficients

- A **correlation coefficient** is a numerical measure of some type of correlation, meaning a statistical relationships between two values.
- There exist several types of correlation coefficients, each with its own definition and its own range of usability and characteristics.

Correlation coefficients

- They have in common that they assume values in the range from −1 to +1, where:

+1 indicates the strongest possible agreement

(direct dependence -> both values are growing together)

0 indicates the independence

–1 indicates the strongest possible disagreement

(indirect dependence -> if the values of one variable increase, the values of the second variable decrease)

Correlation coefficients

- **Pearson's correlation coefficient**
  - Measures statistical dependence on linear data (is parametric)
  - The correlation coefficient is highly influenced by outliers
  - The correlation coefficient is calculated using the standard deviations of the two variables and their **covariance** (= covariance measure of the interaction between variables)

Correlation coefficients

- **Spearman's correlation coefficient**
  - It is calculated from the order of the individual measurements of the two variables (nonparametric method)
  - It captures generally monotone (increasing / decreasing) relationships between variables, not only the linear relationship
  - It is resistant to outliers

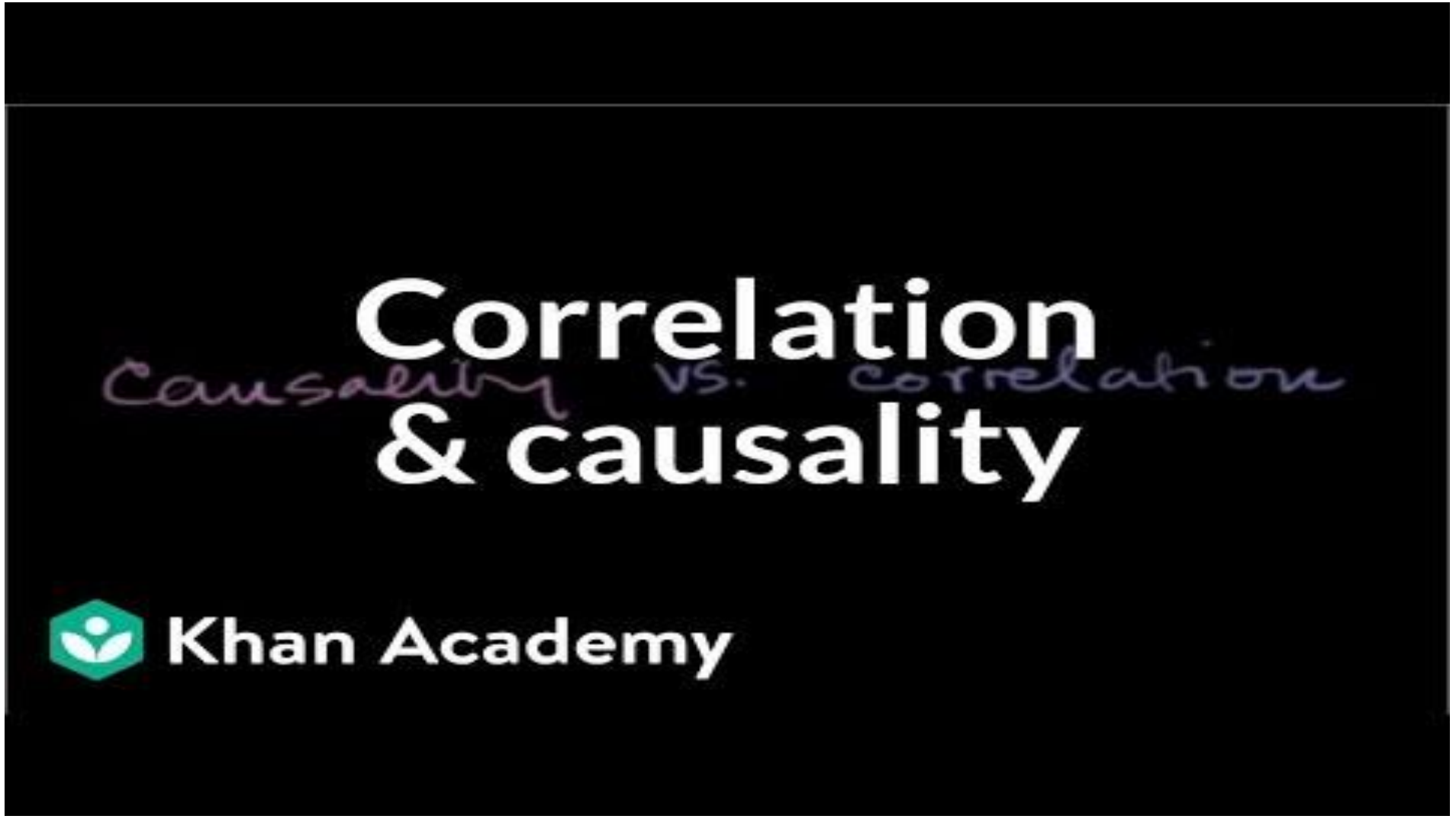
Causality

- **Causality** -> relationship between exposure to risk factor and health impact
- Confirmation of a ***causal relationship*** assumes the prior elimination of coincidence, bias, confounding factors and compliance with causality criteria (association strength, timeliness, biological justification, specificity, dose and effect relationship, coherence, experimental verification);

Causality

- *By statistical procedures, it can not be demonstrated whether or not the relation between variables is causal.*
- The greater the RR (risk ratio) or the OR (odds ratio), the more likely it is that the relationship between the exposure and the consequence is causal.
- For interpretation, in addition to the RR/OR values, it is always necessary knowledge of the confidence interval, range of selection, assessment of specificity, time sequence etc.

Causality


Correlation
Causality vs. Correlation
& causality
Khan Academy

Covariance

- **Covariance** is a measure of the joint variability of two random variables.
- If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, i.e., the variables tend to show similar behavior, *the covariance is positive*.
- If the greater values of one variable mainly correspond to the lesser values of the other, i.e., the variables tend to show opposite behavior, *the covariance is negative*.

Covariance

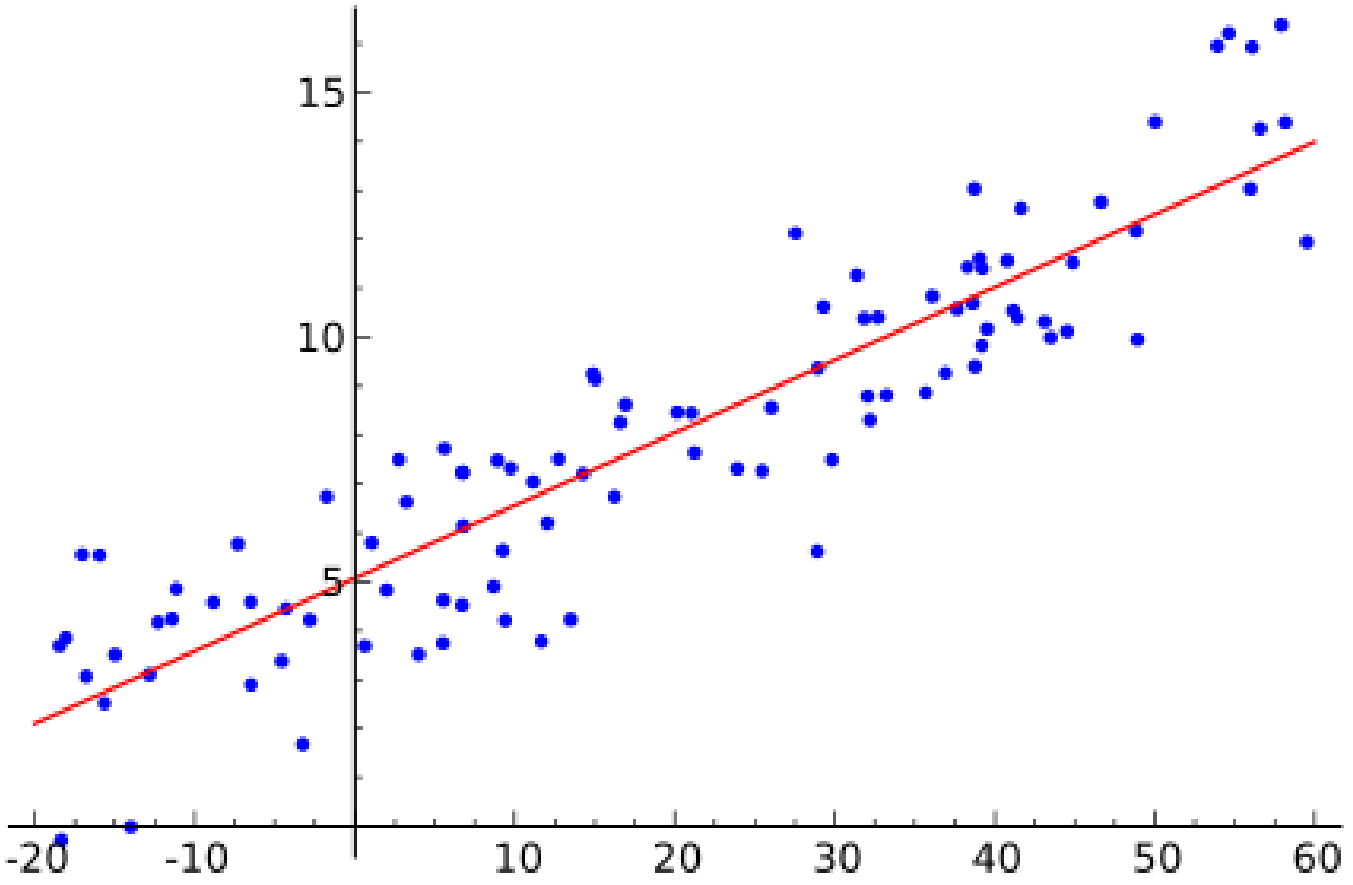
- The sign of the covariance therefore shows *the tendency in the linear relationship* between the variables.
- *The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables.*
- The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

Regression line

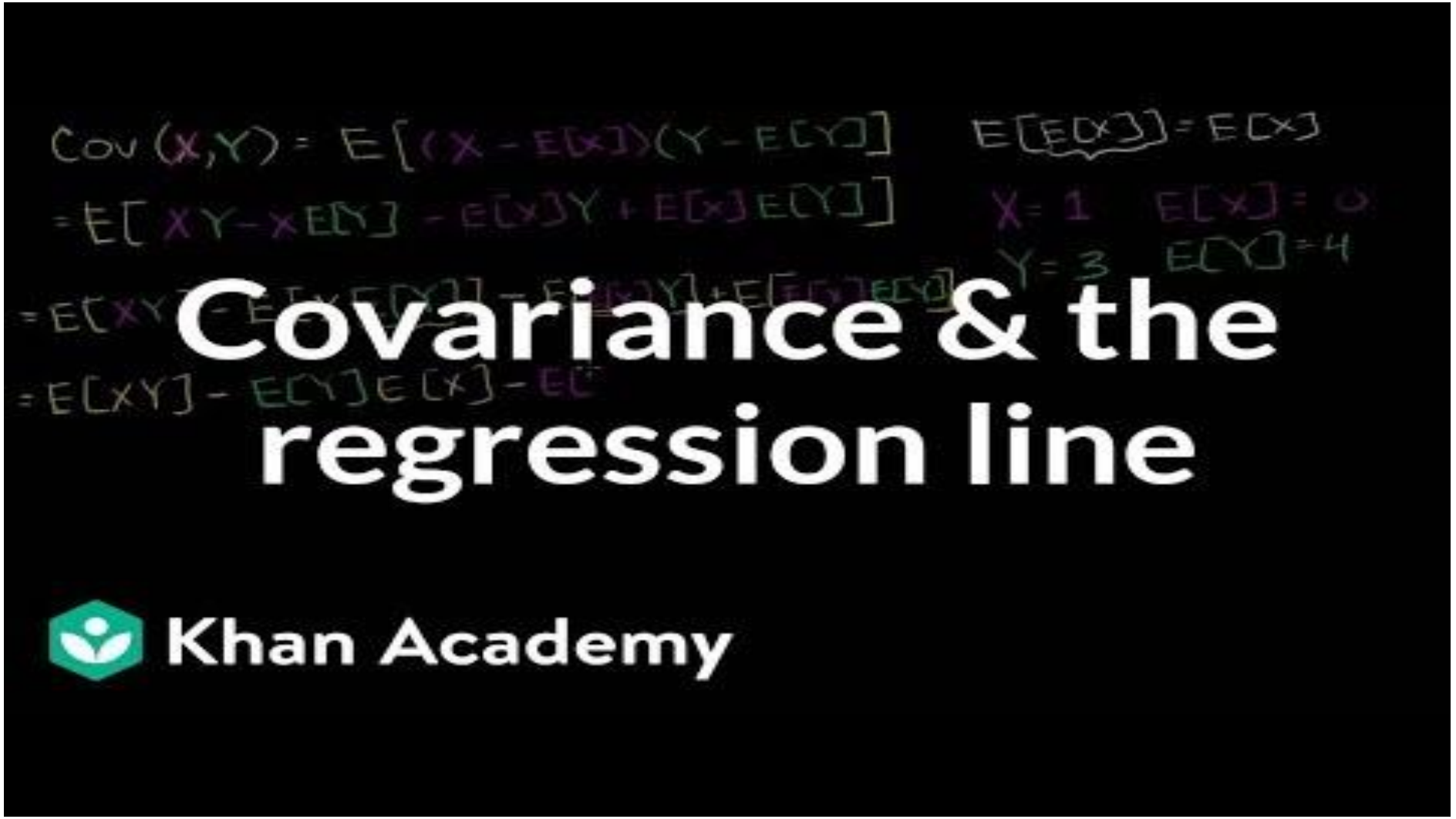
- **Regression analysis** -> for estimating the relationship between a dependent variable (or 'criterion variable') and one or more independent variables (or 'predictors')
- ***linear regression -> regression line***
- The regression line is such a line from which the distance of the points (representing the measured data) is as small as possible.
- One method for finding a regression line is *the least squares method*, i.e. the sum of the squares of the distances of the individual points from the line is minimal.

Regression line

- linear regression



Covariance and the regression line



Time for a break

Evaluation of the quality of tests

The result of a screening test	Disease		
	present	absent	
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
	TP+FN	FP+TN	N

TP = true positive = Sick people correctly identified as sick
FP = false positive = Healthy people incorrectly identified as sick
FN = false negative = Sick people incorrectly identified as healthy
TN = true negative = Healthy people correctly identified as healthy

Evaluation of the quality of tests

- **Sensitivity** measures the proportion of positives that are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

$$SE = \frac{TP}{TP+FN}$$

- **Specificity** measures the proportion of negatives that are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition).

$$SP = \frac{TN}{FP+TN}$$

Evaluation of the quality of tests

- **Prevalence** (pretest probability) of a disease is the proportion of patients with the target disorder in the population tested.

$$\frac{TP+FN}{N}$$

- **Positive predictive value** (PPV) is the probability that subjects with a positive screening test truly have the disease.

$$PPV = \frac{TP}{TP+FP}$$

- **Negative predictive value** (NPV) is the probability that subjects with a negative screening test truly don't have the disease.

$$NPV = \frac{TN}{FN+TN}$$

Evaluation of the quality of tests

- **Likelihood ratio** (LR)
- *LR+ (for positive test results)* – the probability of a person who has the disease testing positive divided by the probability of a person who does not have the disease testing positive

$$\mathrm{LR}^{+} = \frac{\mathrm{SE}}{1-\mathrm{SP}}$$

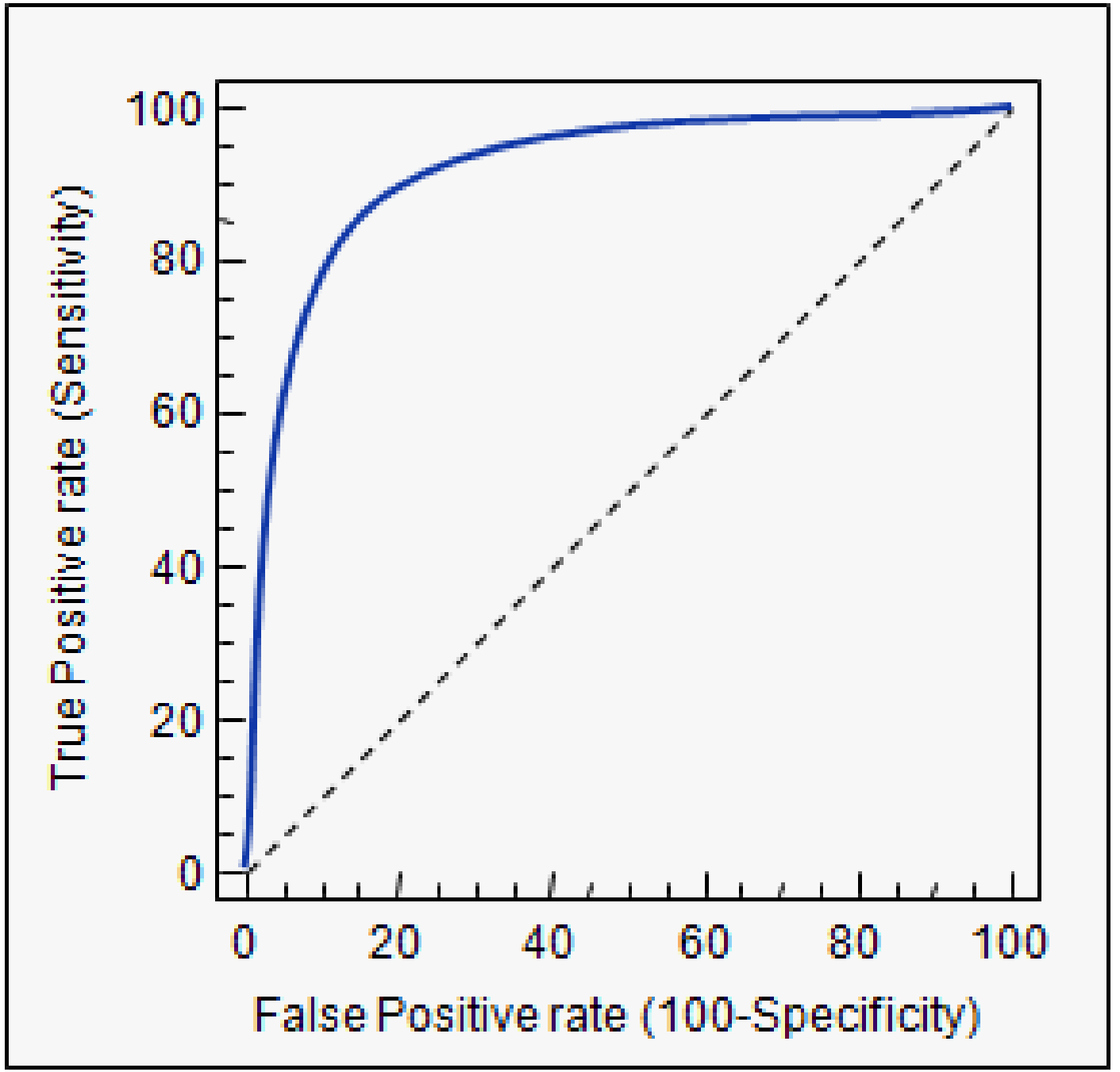
- *LR- (for negative test results)* – the probability of a person who has the disease testing negative divided by the probability of a person who does not have the disease testing negative

$$\mathrm{LR}^{-} = \frac{1-\mathrm{SE}}{\mathrm{SP}}$$

ROC curve

- Receiver operating characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- The ROC curve is created by plotting the true positive rate (TPR) also known as sensitivity against the false positive rate (FPR) also known as the fall-out calculated as (100 − specificity), at various threshold settings.
- Increasing the sensitivity of the test is only possible to the detriment of the decrease in specificity, and vice versa -> this is evidenced by clinical observations.

ROC curve


True Positive rate (Sensitivity)
100
80
60
40
20
0
0
20
40
60
80
100
False Positive rate (100-Specificity)

Measuring the frequency of disease

- The main task of descriptive epidemiology – determining the frequency with which the disease occurs in the population.
- The description also includes the dynamics of changes in this frequency in time and space.
- The unit of statistical investigation:
  - a person – as a carrier of a disease, the object of epidemiological research – a specific, unambiguously determined person
  - Illness -> more difficult to track than death -> information only on those sick who pass through a medical facility

Measuring the frequency of disease

- Determination of the unit of measurement
  - person as a carrier of the disease (number of HIV-infected people, number of diabetics)
  - case of illness (number of sore throats, flu)
  - other event related to illness – doctor's visit, hospitalization, incapacity for work, award of disability pension
- Defining a population
  - it is referred to as an exposed (risk) population
  - it is the population to which the given morbidity indicator applies
- Determination of time
  - specifying a moment or interval

Frequency indicators

- **Incidence** is a measure of the probability of occurrence of a given medical condition in a population within a specified period of time.
  - Although sometimes loosely expressed simply as the number of new cases during some time period, it is better expressed as a proportion or a rate with a denominator.

Frequency indicators

- **Prevalence** is the proportion of a particular population found to be affected by a medical condition (typically a disease or a risk factor such as smoking or seatbelt use) at a specific time.
  - It is derived by comparing the number of people found to have the condition with the total number of people studied and is usually expressed as a fraction, a percentage, or the number of cases per 10 000 or 100 000 people.
  - Prevalence is most often used in questionnaire studies.

Frequency indicators

- **Mortality** rate, or death rate, is a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time.
  - Mortality rate is typically expressed in units of deaths per 1,000 individuals per year.
  - It is distinct from "morbidity", which is either the prevalence or incidence of a disease, and also from the incidence rate (the number of newly appearing cases of the disease per unit of time).

Frequency indicators

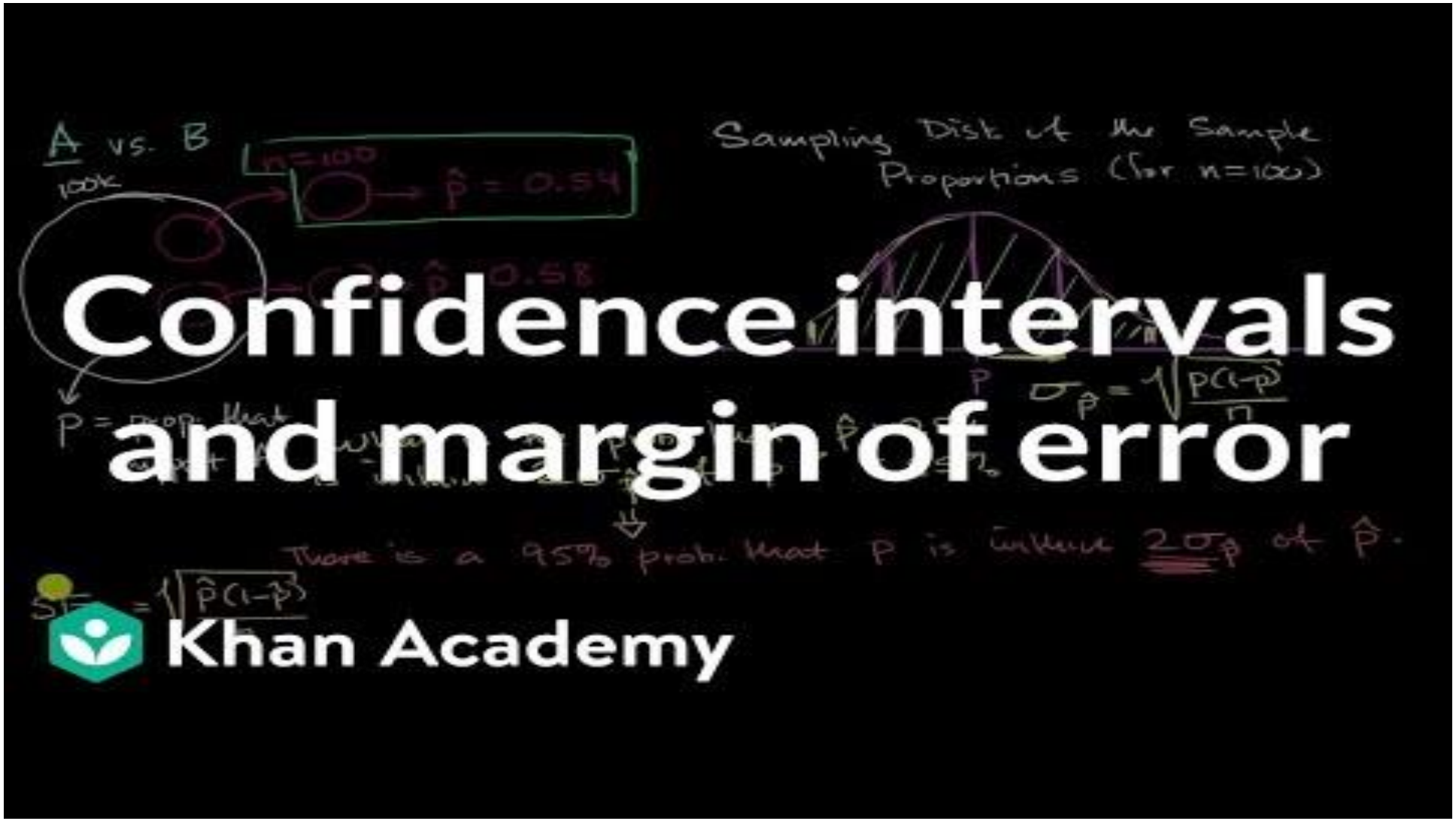
- **Morbidity** is a diseased state, disability, or poor health due to any cause. The term may refer to the existence of any form of disease, or to the degree that the health condition affects the patient.

  - Comorbidity is the simultaneous presence of two or more medical conditions.
  - In epidemiology and actuarial science, the term "morbidity rate" can refer to either the incidence rate, or the prevalence of a disease or medical condition.
  - This measure of sickness is contrasted with the mortality rate of a condition, which is the proportion of people dying during a given time interval.

Confidence interval

- A **confidence interval** is how much uncertainty there is with any particular statistic.
- Confidence intervals are often used with a margin of error. It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.
- *Confidence levels* are expressed as a percentage (for example, a 95% confidence level). It means that should you repeat an experiment or survey over and over again, 95 percent of the time your results will match the results you get from a population.

Confidence interval



Statistical hypothesis testing

- A standard statistical procedure involves the test of the relationship between two statistical data sets, or a data set and synthetic data drawn from idealized model.
- A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis of no relationship between two data sets.
- Rejecting or disproving the null hypothesis is done using statistical tests that quantify the sense in which the null can be proven false, given the data that are used in the test.

Statistical hypothesis testing

- **The testing process:**
  1) Formulate the relevant null and alternative hypotheses.
  2) Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%.
  3) Get the data.
  4) Decide which test is appropriate, and state the relevant test statistic.
  5) Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. Find the appropriate critical value in the tables.

Statistical hypothesis testing

6) Calculate the **p-value**. This is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed.

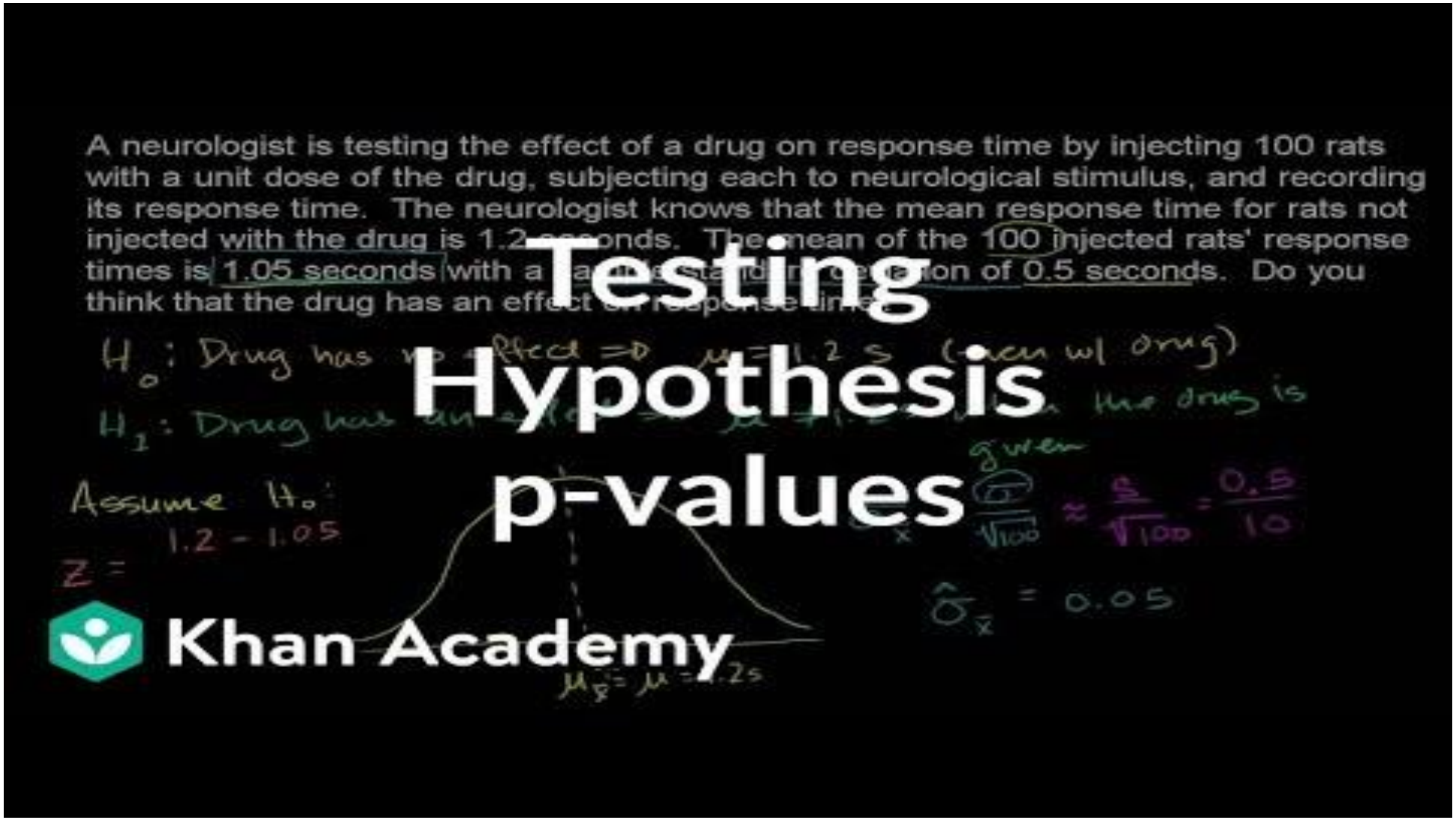
7) Make a **statistical decisions**. Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

Statistical hypothesis testing

- Working from a null hypothesis, two basic forms of error are recognized: Type I errors (null hypothesis is falsely rejected giving a "false positive") and Type II errors (null hypothesis fails to be rejected and an actual difference between populations is missed giving a "false negative").

	H_0 is true	H_1 is true
Accept null hypothesis	Right decision	Wrong decision Type II Error
Reject null hypothesis	Wrong decision Type I Error	Right decision

Statistical hypothesis testing



Ing. Anna Horňáková, Ph.D.

Institute of Hygiene and Epidemiology of the 1st Faculty of Medicine

FIRST FACULTY OF MEDICINE
Charles University

THANK YOU FOR YOUR ATTENTION

QUESTIONS?

CONTACT ME:
ANNA.HORNAKOVA@LF1.CUNI.CZ