Statistical analysis of data

Last updated: 02/22/2023 19:32:27

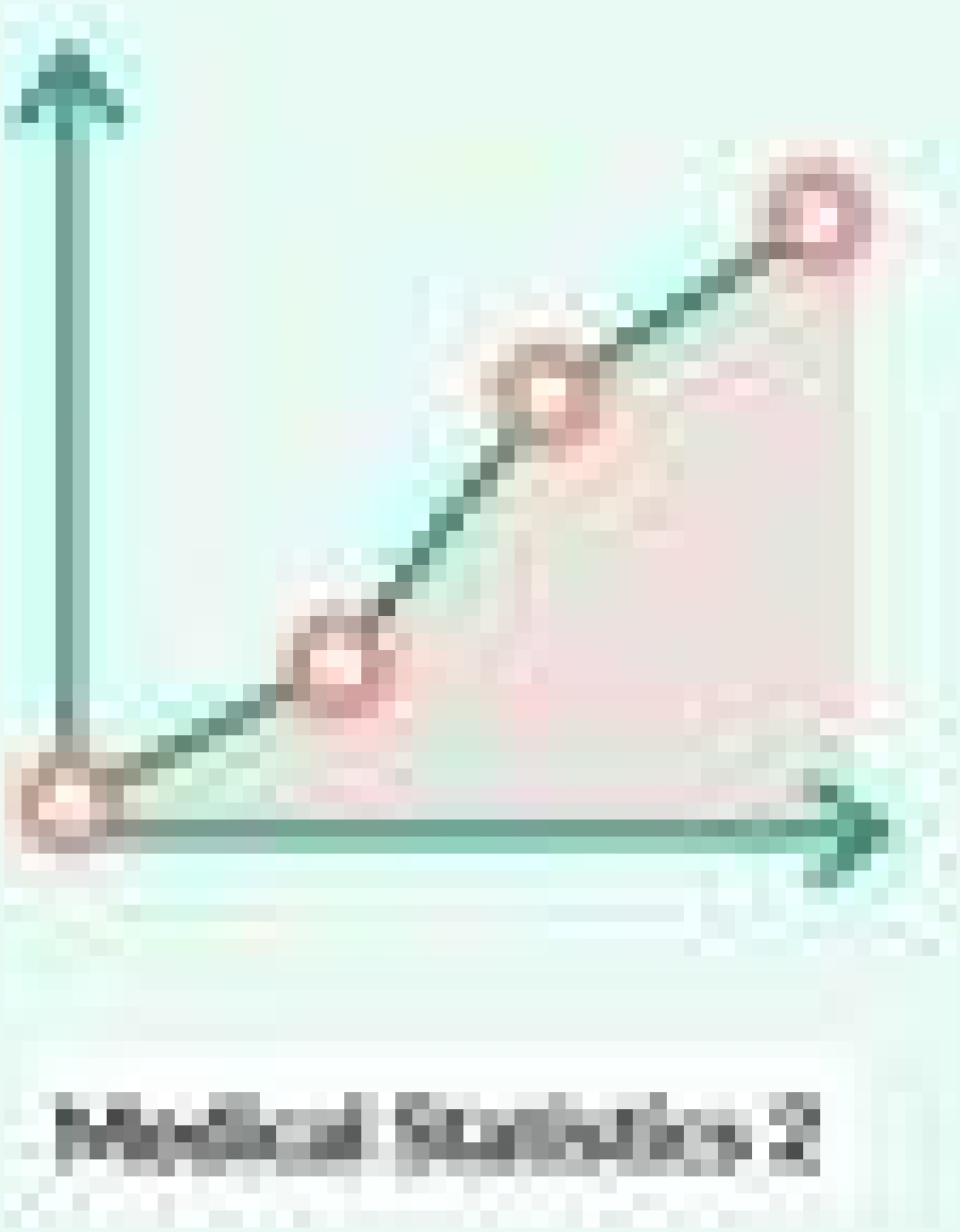
Summary

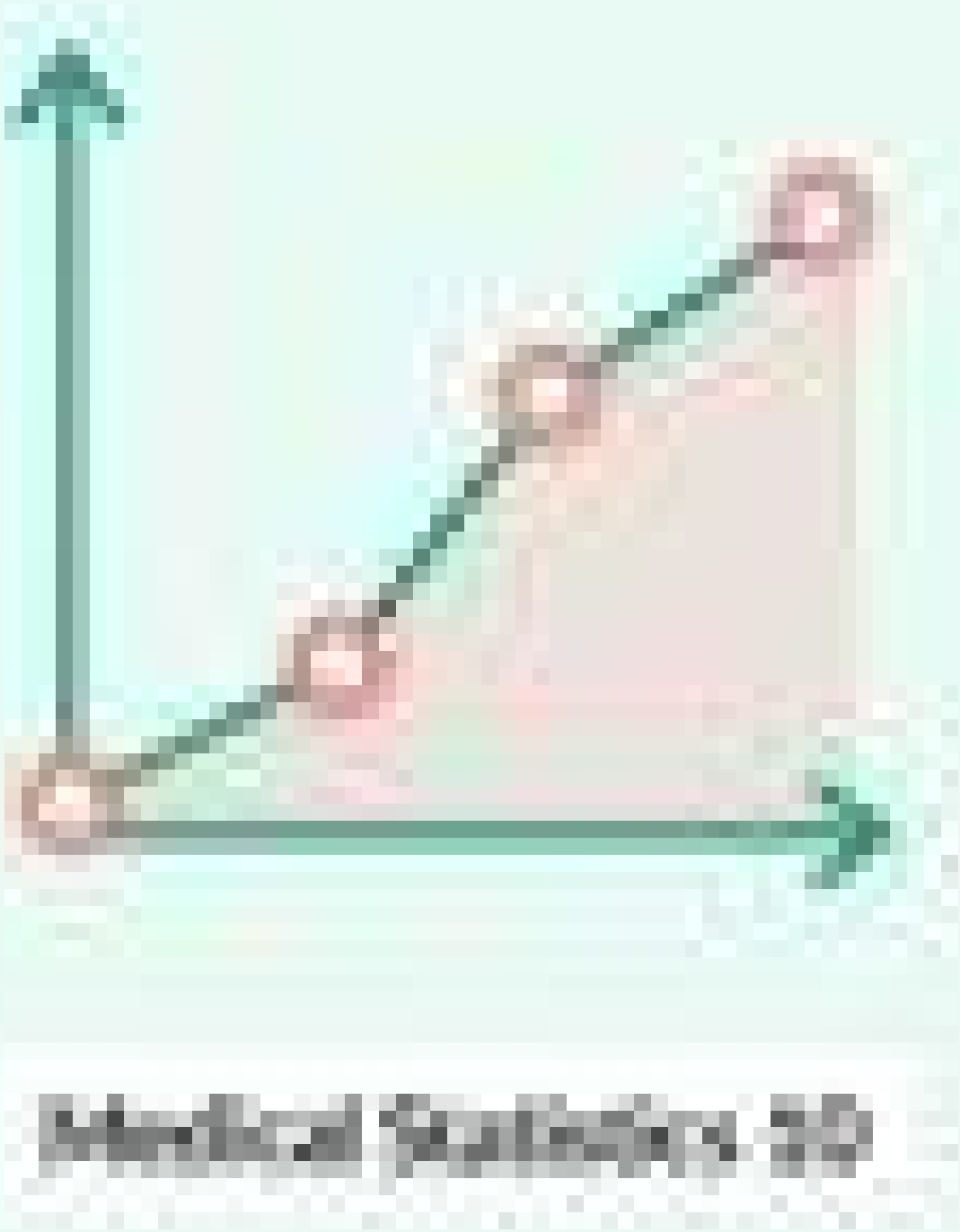
Statistical analysis is one of the principal tools employed in epidemiology, which is primarily concerned with the study of health and disease in populations and its clinical applications. Statistics is the science of collecting, analyzing, and interpreting data, and a good epidemiological study depends on statistical methods being employed correctly. At the same time, flaws in study design can affect statistics and lead to incorrect conclusions. Descriptive statistics measure, describe, and summarize features of a collection of data/sample without making inferences that go beyond the scope of that collection/sample. Common measures of descriptive statistics are those of central tendency and dispersion. Measures of central tendency describe the central distribution of data and include the mode, median, and mean. Measures of dispersion describe how data is distributed and include range, quartiles, variance, and deviation. The counterpart of descriptive statistics, inferential statistics, relies on data to make inferences that do go beyond the scope of the data collected and the sample from which it was obtained. Inferential statistics involves parameters such as sensitivity, specificity, positive/negative predictive values, confidence intervals, and hypothesis testing.

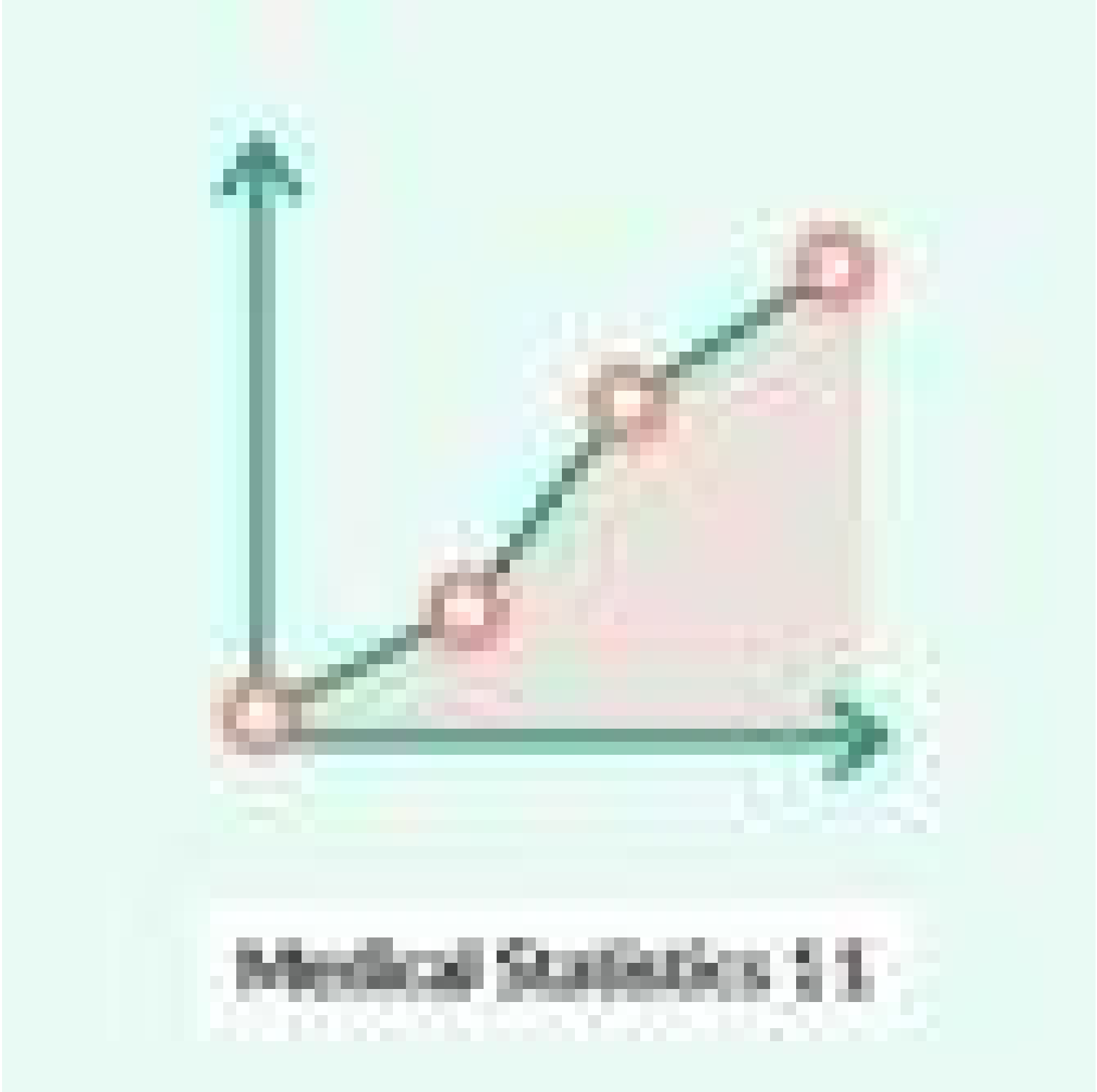
The values used to describe features of a sample or data set are called variables. Variables can be independent, in the sense that they are not dependent on other variables and can thus be manipulated by the researcher for the purpose of a study (e.g., administration of a certain drug), or dependent, in the sense that their value depends on another variable and, thus, cannot be manipulated by the researcher (e.g., a condition caused by a certain drug). Variables can furthermore be categorized qualitatively in categorical terms (e.g., eye color, sex, race) and quantitatively in numerical terms (e.g., age, weight, temperature).

Statistical analysis is used in all types of epidemiological studies, including the evaluation of diagnostic tests before approval for clinical practice. These rely on inferential statistics to draw conclusions from sample groups that can be applied to the general population.

See also "Epidemiology," "Interpreting medical evidence," and "Population health."







Descriptive and inferential statistics

- **Descriptive statistics**
  - Analysis of a sample group conducted in order to measure, describe, and summarize the data collected, but not to make inferences that go beyond the scope of that sample group
  - Employs measures of central tendency (mode, median, and mean) and measures of dispersion measures (range, quartiles, variance, and deviation)
- **Inferential statistics**: analysis of a sample group conducted in order to make inferences that go beyond the sample group.

Measures of central tendency and outliers

Measures of central tendency

- **Definition**: measures to describe a common, typical value of a data set (e.g., clustering of data at a specific value)
- **Approach**: The type of measure used depends on the sample size.

Measures of central tendency		
Measure	**Definition**	**Example**
Mean (statistics)	• The arithmetic average of the data set • Limitations: affected by extreme values (outliers)	• The sum of all the data divided by the number of values in the data set. (e.g., consider a data set of 3, 6, 11, 14, 16, 19. The mean value is 11.5 (= 69/6).
Median (statistics)	• The middle value of the data set that has been arranged in order of magnitude; it divides the upper half of the data set from the lower half • Not strongly affected by outliers or skewed data	• Uneven number of values: 3, 6, 11, 16, 19. The median value is the middle value = 11. • Even number of values: 2, 3, 5, 7, 9, 10. The median value is the average of the two middle values = (5+7)/2 = 6.
Mode (statistics)	• The most common value in a data set • Most resistant against outliers • Can be used to describe qualitative data.	• In a data set with the values "3, 6, 6, 11, 11, 11, 2, 2," the mode = 11.

Outlier

- **Definition**: a data point/observation that is distant from other data points/observations in a data set
- **Problem**
  - It is important to identify outliers, because outliers can indicate errors in measurement or statistical anomalies.
  - The mean is easily influenced by outliers
- **Approach**
  - Using a trimmed mean: calculate the mean by discarding extreme values in a data set and using the remaining values
  - Use the median or mode: useful for asymmetrical data; these measures are not affected by extreme values because they are based on ranks of data (median) or the most commonly occurring value (mode) rather than the average score of all values
  - Removing outliers can also distort the interpretation of data. It should be done with caution and with a view to reflecting the respective data set.
- **Regression to mean**: a phenomenon in which any measurement taken after the measurement of a random variable lying at the extreme (i.e., above or below the mean) is likely to be closer to the mean

Measures of dispersion

- **Definition**: measures the extent to which the distribution is stretched out

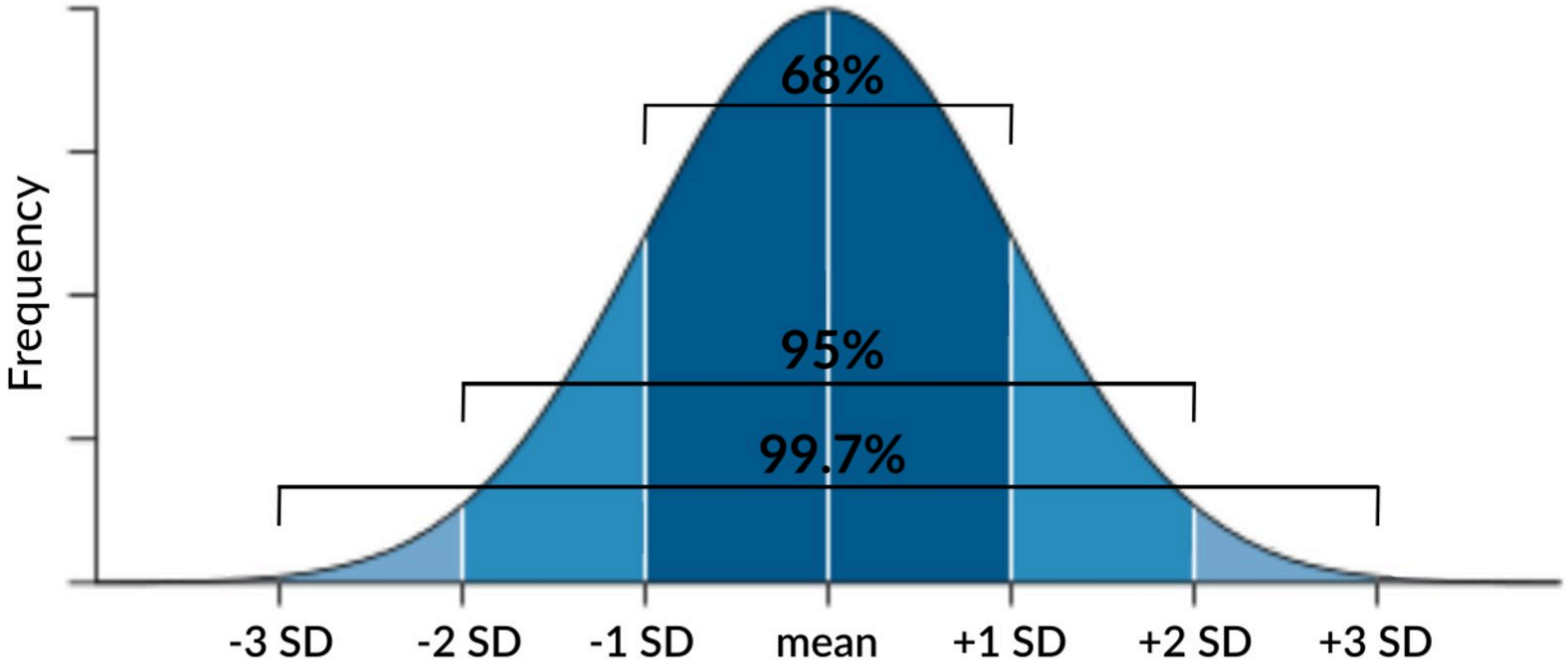
Measures of dispersion		
Measure	**Definition**	**Description**
Range (statistics)	• The difference between the largest and smallest value in a data set • Sensitive to extreme data values • Helps to identify an unusually wide or narrow data range, which may occur with data entry errors (e.g., data that actually belongs to another study population)	• In the data set "27, 3, 4, 9," the range is 24 (i.e., 27-3).
Interquartile range	• The range from the second to the third quartile. • Calculated by establishing the difference between the 75^{th} and 25^{th} percentile. • Less influenced by extreme data values (outliers)	• Calculated as the difference between the 75^{th} and 25^{th} percentile
Variance (statistics)	• The average of the squared deviations from the mean • Represented by σ^2 ("s~")	• σ^2 = sum of squared deviations from the mean divided by total number of observations • Calculated by subtracting the mean from each population data set value. Each difference is then squared and added together. Finally, the total sum is divided by n-1.

Standard deviation (SD)	• The square root of the variance
	• Describes the variability or dispersion of data in relation to its mean
	• Represented by σ (sigma)
		• σ = square root of variance
		• The standard deviation is calculated by first calculating the mean. The mean is subtracted from each population data set value. Each difference is squared and added together. The total sum is divided by the total number of data set values -1. The square root of this value is the standard deviation (σ).
		• **In a normal distribution**
		  ◦ 1 SD = 68% of the data set
		  ◦ 2 SD = 95% of the data set
		  ◦ 3 SD = 99.7% of the data set

Percentiles	• Division of the population data set into 100 equal parts
	• A percentile is the value below which a percentage of observations fall.
		• Percentiles are usually used to help evaluate children's growth.
		• If a child's weight is in the 25^{th} percentile for his or her age, this child's weight is heavier than 25% of children of the same age group, but lighter than 75% children of the same age group. For example:
		  ◦ 3^{rd} percentile (= 3/100-quartile): 3% of all values are smaller than this value.
		  ◦ 50^{th} percentile (= 50/100-quartile): 50% of all values are smaller than this value (median).
		  ◦ 97^{th} percentile (= 97/100-quartile): 97% of all values are smaller than this value.
		• Upper limit of normal (ULN): a value at the upper extreme of the reference range (95th percentile) of the target population.
		• Lower limit of normal (LLN): a value at the lower extreme of the reference range (5th percentile) of the target population.

Quartile	• One quarter of a data set	• Each quartile includes 25% of the population data set. ◦ First quartile (lower quartile): 25% of all values are smaller than this value.

Measures of dispersion		
Measure	**Definition**	**Description**
		◦ Third quartile (upper quartile): 75% of all values are smaller than this value.
Standard error of the mean	• The deviation of the sample mean from the population mean • Influenced by the standard deviation (e.g., a greater SD increases the chance of error) and the sample size (a smaller sample size will increase the chance of error)	• **SEM = standard deviation/√(sample size)**



Variables

- **Definition**
  - Variables: measured values of population attributes or a value subject to change
  - General population: the group from which the units of observation are drawn (e.g., all the patients in a hospital)
  - Unit of observation: the individual who is the subject of the study (e.g, inhabitant of a region, a patient)
  - Attribute: a character of the unit of observation (e.g., gender, patient satisfaction)
  - Attribute value
    - Variables can be qualitative (e.g., male/female) or quantitative (e.g., temperature: 10°C, 20°C) in nature
    - Quantitative variables can be discrete or nondiscrete (continuous) variables (see "Probability" below).
- **Types of variables**
  - Independent variable: a variable that is not dependent on other variables and can thus be manipulated by the researcher for the purpose of a study

- Dependent variable: a variable with a value that depends on another variable and therefore cannot be manipulated by the researcher

- **Types of quantitative variables**
  - Discrete variable: variables that can only assume whole number values
  - Continuous variable (nondiscrete variable): variables that can assume any real number value
- **Categorical variable** (nominal variable): variables that have a finite number of categories that may not have an intrinsic logical order
- **Variable scales**
  - Definition: types of measurement scales (categorized as categorical scales and metric scales)
  - Categorical scale (qualitative)
    - The distance (interval) between two categories is undefined.
    - Includes the nominal scale and ordinal scale
  - Metric scale (quantitative)
    - The distance between two categories is defined and the data can be ranked .
    - Includes the interval scale and ratio scale

Types of scales [1][2]

Types	Characteristics	Measure of central tendency	Measure of dispersion	Statistical analysis	Data illustration
Nominal scale	• Data cannot be ranked	• Mode • Absolute and relative frequency	• Not applicable	• Nonparametric tests (e.g., Mann-Whitney test)	• Pie chart • Bar graph
Ordinal scale	• Data can be ranked	• Median • Upper and lower quartiles • Percentiles	• Range • Interquartile range		• Pie chart • Bar graph • Box plot
Interval scale	• There is no natural zero point.	• Minimum/maximum values • Median • Upper and lower quartiles • Percentiles • Mean value	• Range • Interquartile range • Standard deviation • Variance	• Parametric tests (e.g., T-test)	• Pie chart • Bar graph • Box plot • Histogram • Scatter plot
Ratio scale	• There is a natural zero point.				

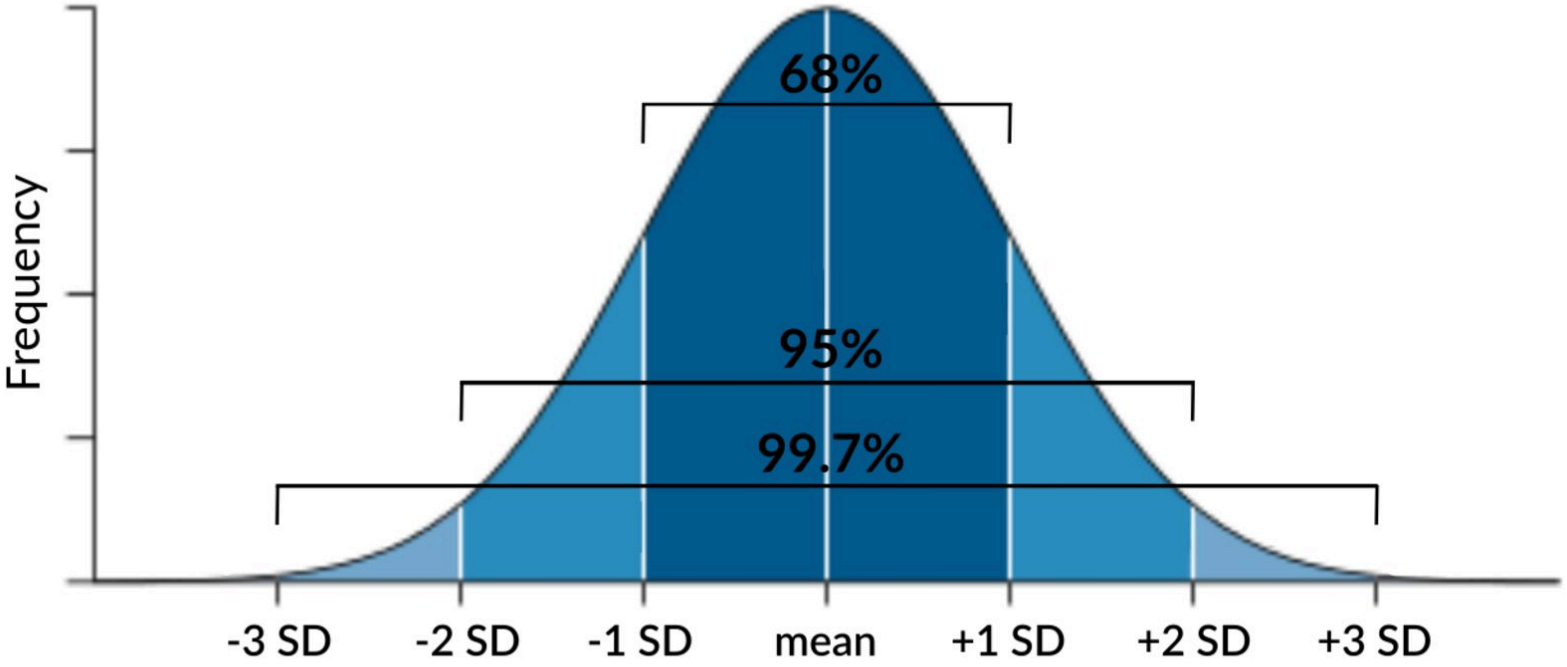
Distribution and graphical representation of data

Normal distribution (Bell curve, Gaussian distribution) [3]

- Normal distributions differ according to their mean and variance, but share the following characteristics:
  - The same basic shape
  - Unimodal distribution (i.e., one peak)
  - Asymptotic to the x-axis

- Symmetry (i.e., a symmetrical **bell curve**)
- The following assumptions about the data distribution can be made:
  - 68% of the data falls within 1 SD of the mean.
  - 95% of the data falls within 2 SD of the mean.
  - 99.7% of the data falls within 3 SD of the mean.
- Total area under the curve = 1
- All measures of central tendency are equal (**mean = median = mode**)

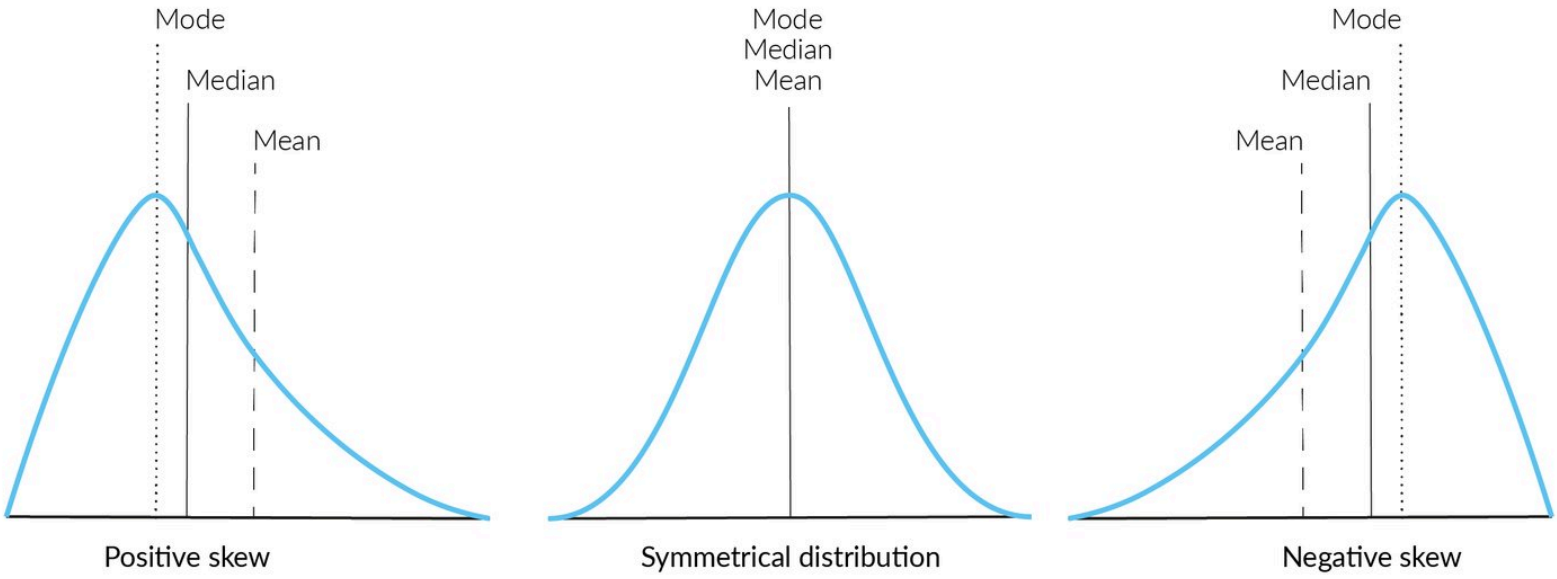
- **Standard normal distribution** (Z distribution): a normal distribution with a mean of 0 and standard deviation of 1



Nonnormal distributions

Types of nonnormal distributions		
	Description	**Meaning**
Bimodal distribution	• The data set has **two peaks** (peak = modal value).	• Two subgroups within the study population (e.g., the incidence rate of Hodgkin's lymphoma that has the first peak at 25–30 years and the second peak at 50–70 years)
Positively skewed distribution (right-skewed distribution)	• The data set has a peak on the left side and a **long tail on the right** (positive direction). • The mean falls closer to the right tail.	• **Mean > median > mode**

Types of nonnormal distributions		
	Description	**Meaning**
Negatively skewed distribution (left-skewed distribution)	• The data set has a peak on the right side and a **long tail on the left** (negative direction). • The mean falls closer to the left tail.	• **Mean < median < mode**



Standard normal value (Z-score, Z-value, standard normalized score)

- Enables the comparison of populations with different means and standard deviations
  - Standard normal value = (value - population mean) divided by standard deviation
  - A means of expressing data scores (e.g., height in centimeters or meters) in the same metric (specifically, in terms of units of standard deviation for the population)
  - Determines how many standard deviations an observation is above or below the mean

Recommended measures

Recommended measures according to distribution		
Distribution	**Measures of central tendency**	**Measure of spread**
Normal (symmetrical)	• Mean • Median • Mode	• Standard deviation
Skewed (asymmetrical)	• Median	• Range or interquartile range

Data illustration

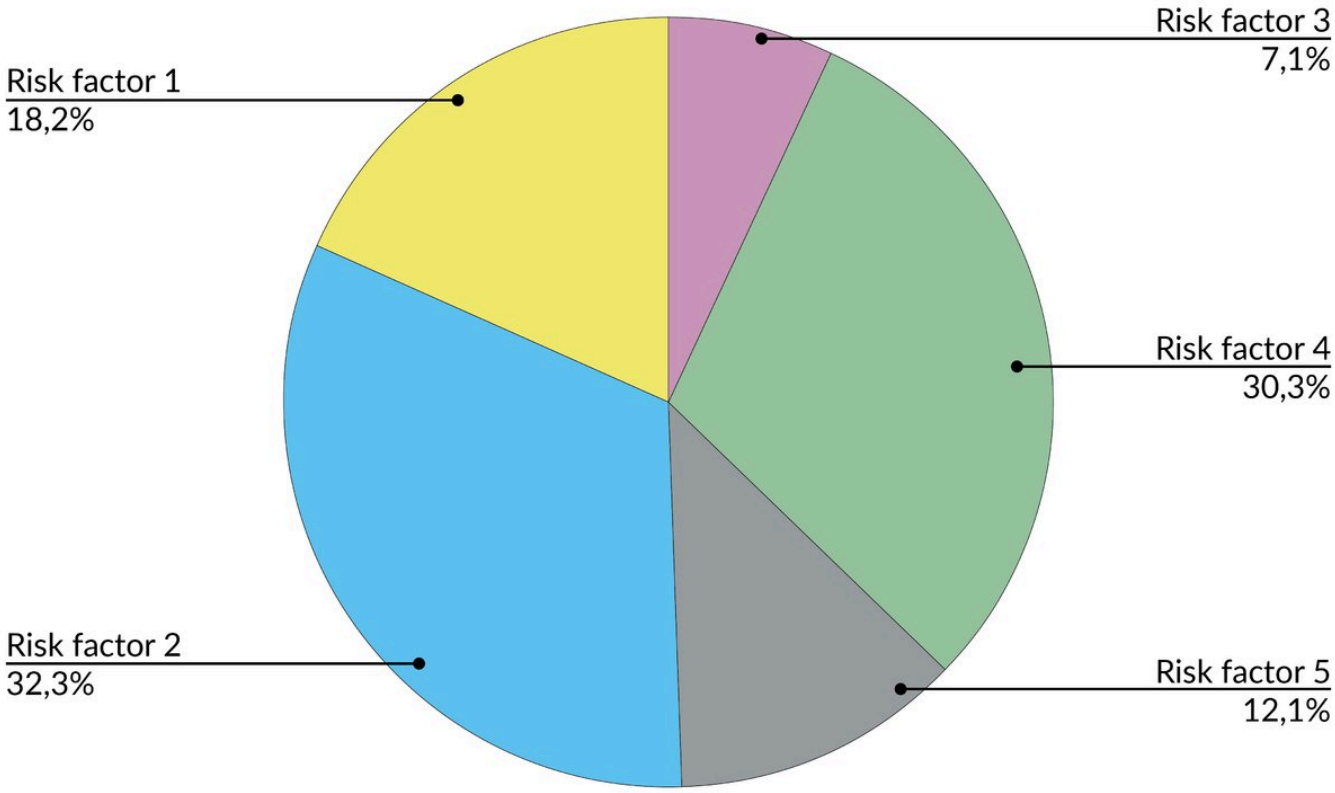
Categorical data

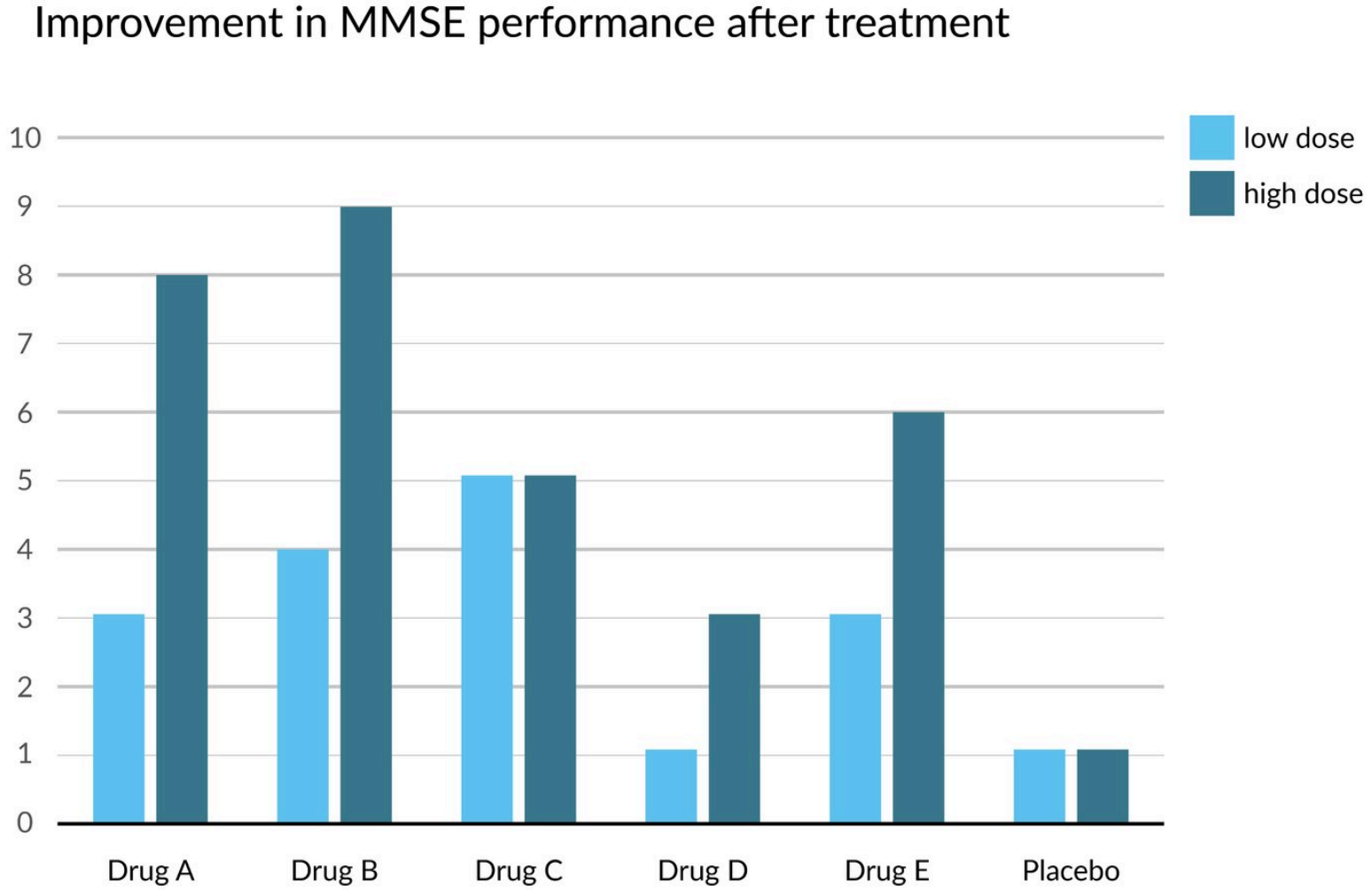
- **Frequency table**
  - Presents data values for each category in a table
  - Illustrates which values in a data set appear frequently
- **Pie chart**
  - Describes the frequency of categories in a circular graph divided into slices, with each slice representing a categorical proportion
  - Useful for depicting a small number of categories and large differences between them
- **Bar graph**
  - Describes the frequency of categories in bars separated from each other (the height/length of each bar represents a categorical proportion)
  - Useful for depicting many categories of information (compared to a pie chart)
  - Frequency can be expressed in absolute or relative terms.

Frequency table

Mean arterial pressure (mm Hg)	Number of patients
<60	3
60-69	12
70-79	13
80-89	17
90-99	10
≥100	8

Independent risk factors



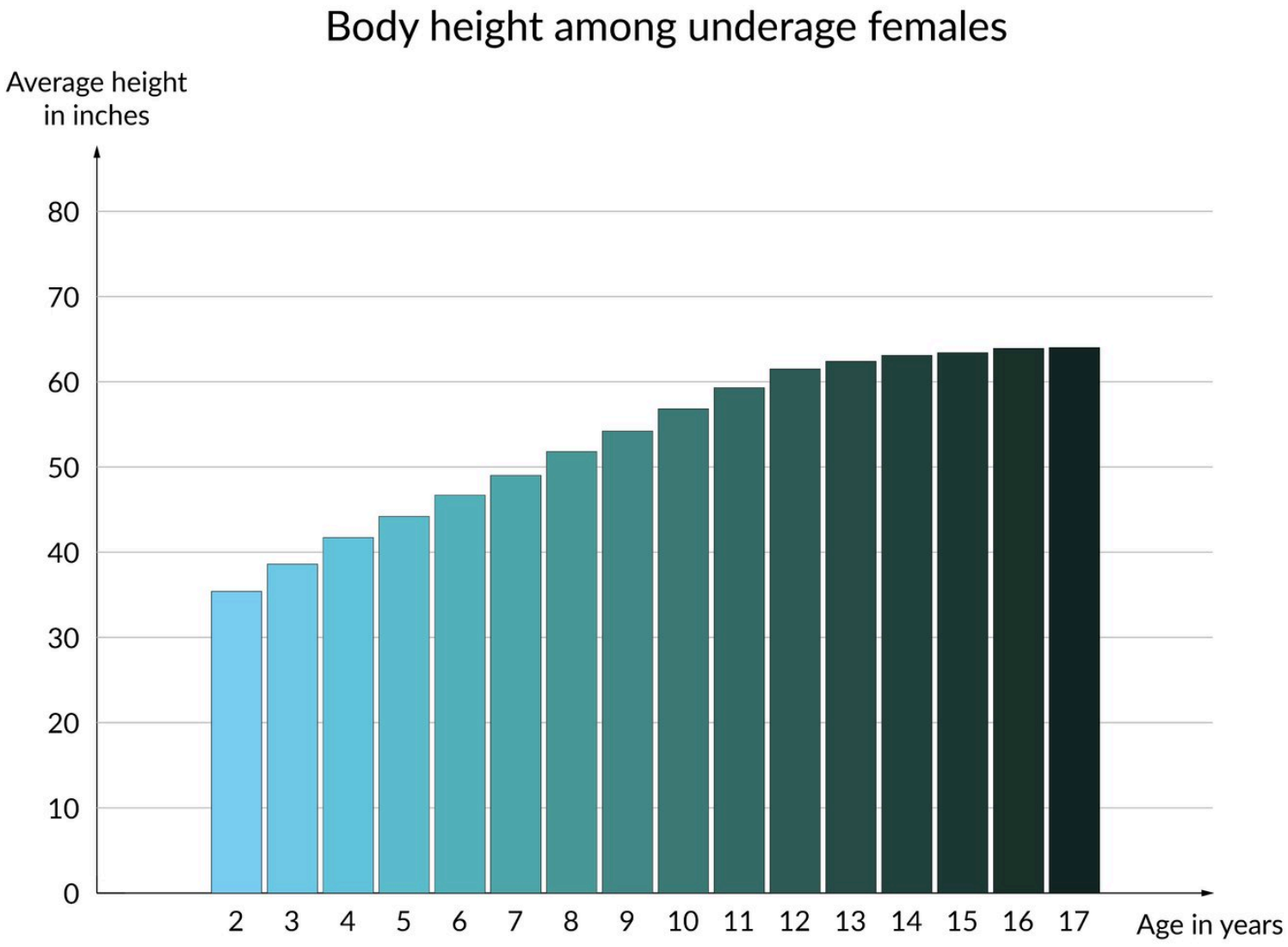


Continuous data

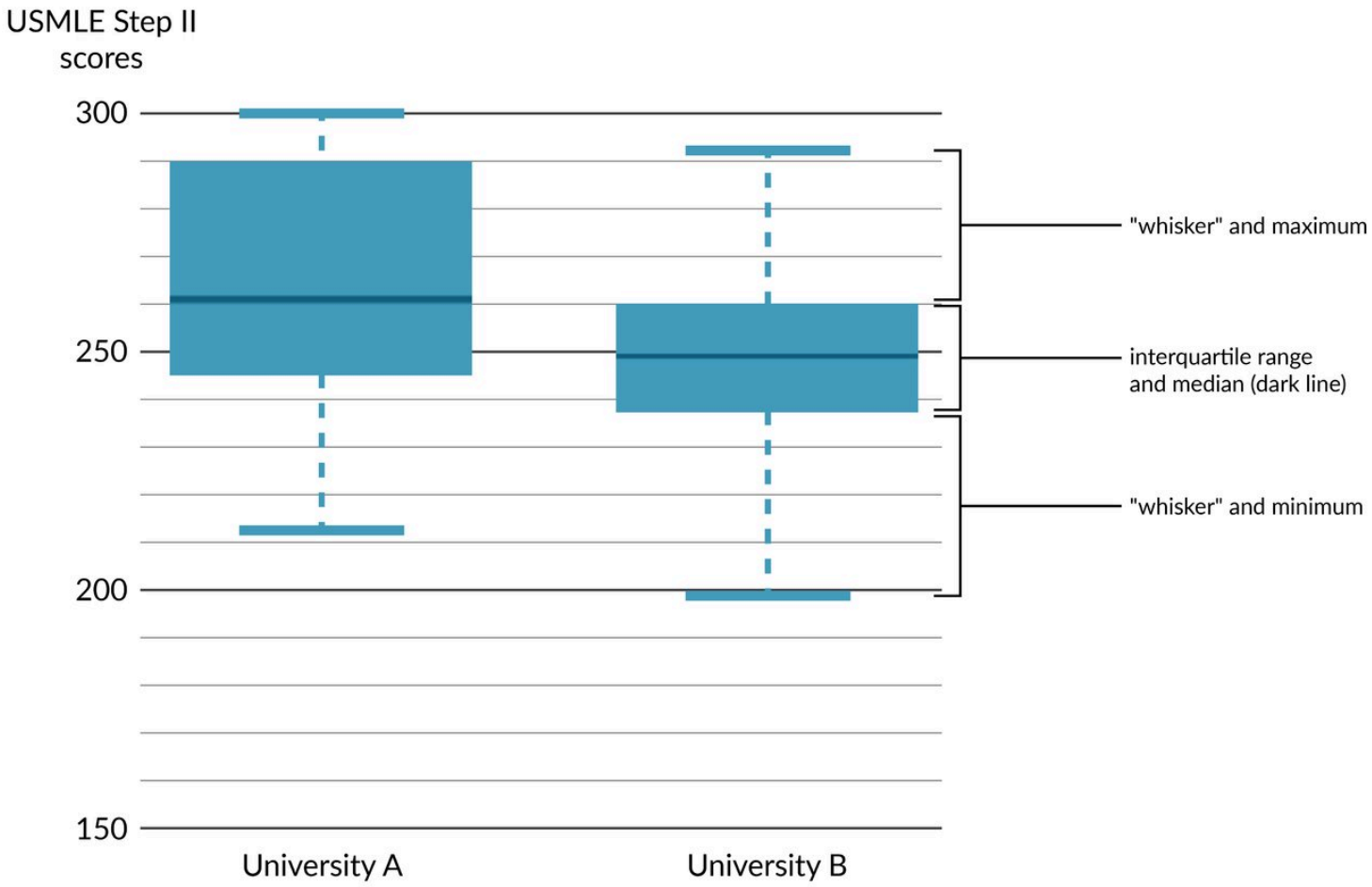
- **Histogram**
  - A histogram is similar to a bar graph but displays data on a metric scale.
  - The data is grouped into intervals that are plotted on the x-axis.
  - Useful for depicting continuous data
  - Similar to a bar chart, but differs in the following ways:
    - Used for continuous data
    - The bars can be shown touching each other to illustrate continuous data.
    - Bars cannot be reordered.
- **Box plot**
  - Quartiles and median are used to display numerical data in the form of a box.
  - Useful for depicting continuous data
  - Shows the following important characteristics of data:
    - Minimum and maximum values
    - First and third quartiles
    - Interquartile range
    - Median

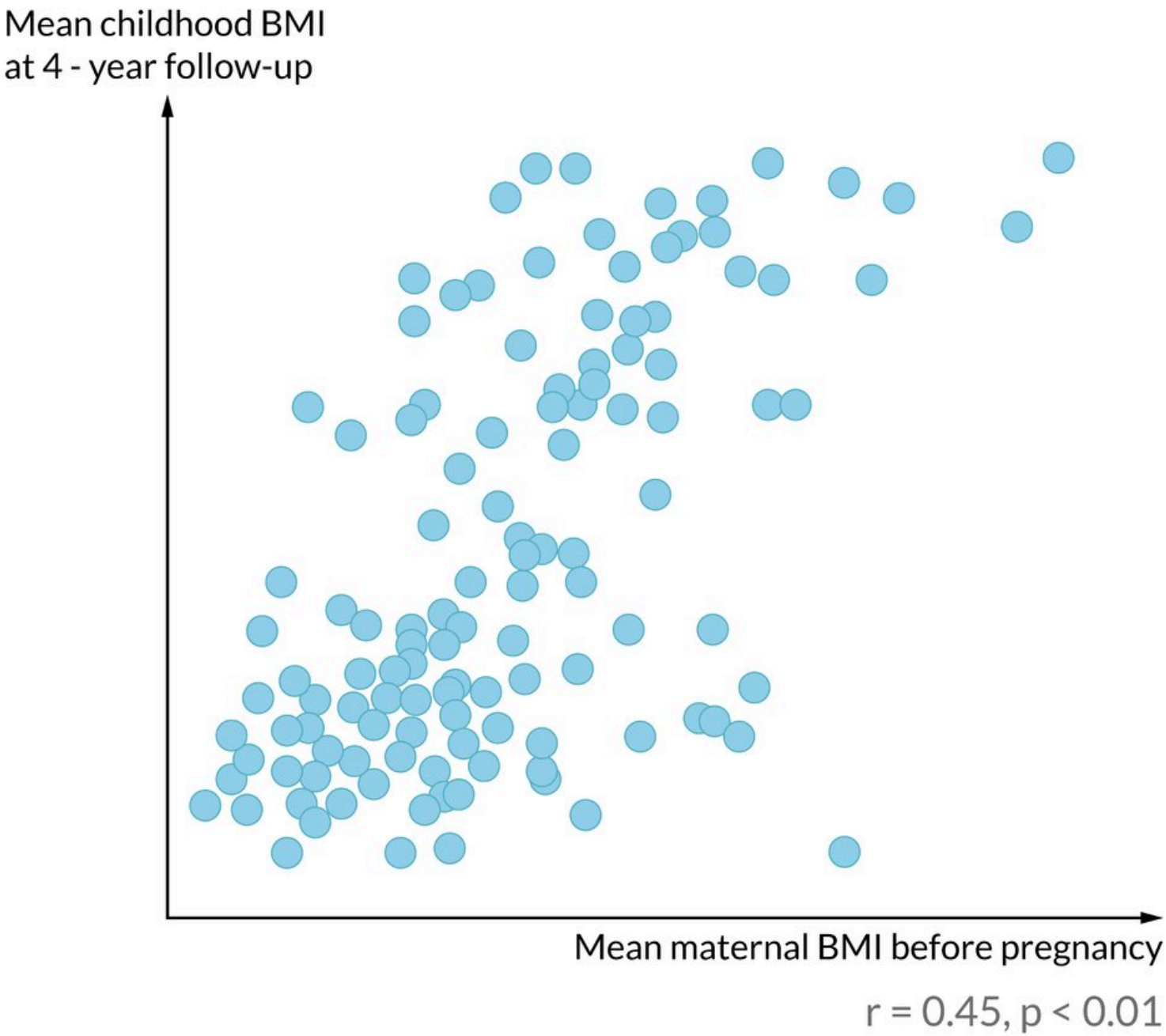
- Easily shows measures of central tendency, range, symmetry, and outliers at a glance

- **Scatter plot**
  - A graph used to display values for (typically) two variables of data, plotted on the horizontal (x-axis) and vertical (y-axis) axes using cartesian coordinates, which represent individual data values
  - Helps to establish correlations between dependent and independent variables
  - Helps to determine whether a relationship between data sets is linear or nonlinear



Comparison of USMLE Step II test results between two universities





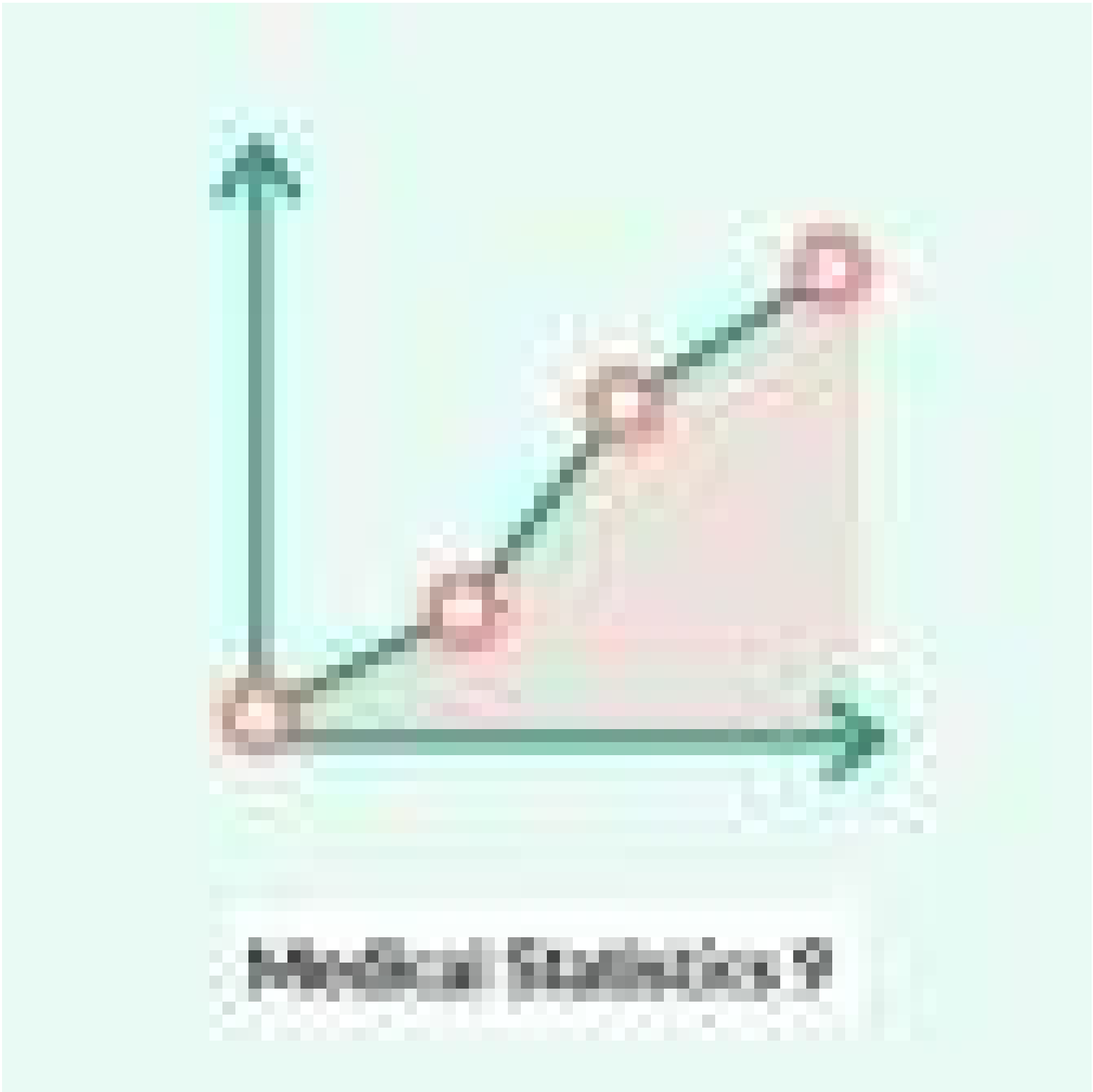
Hypothesis testing and probability

Hypothesis testing

Types of hypothesis

Two mutually exclusive hypotheses (null hypothesis and alternative hypothesis) are formulated.

- **Null hypothesis (H_0)**: The assumption that there is no statistically significant relationship between two measured variables (e.g., the exposure and the outcome) or no significant difference between two studied populations. Statistical tests are used to either reject or accept this hypothesis.
  - Null value
    - A number that corresponds to the null hypothesis
    - The null value is 1 for ratios (e.g., relative risk, odds ratio) and 0 for differences (e.g., attributable risk, absolute risk reduction).
- **Alternative hypothesis (H_1)**: The assumption that there is a relationship between two measured variables (e.g., the exposure and the outcome) or a significant difference between two studied populations. This hypothesis is formulated as a counterpart to the null hypothesis. Statistical tests are used to either reject or accept this hypothesis.
  - Directional alternative hypothesis (one-tailed): specifies the direction of a tested relationship
  - Non-directional alternative hypothesis (two-tailed): only states that a difference exists in a tested relationship (does not specify the direction)



Interpretation [4]

- **Correct result**
  - The null hypothesis is rejected when there is a relationship between two measured variables.
  - The null hypothesis is accepted when there is no relationship between two measured variables.
- **Type 1 error**
  - The **null hypothesis is rejected** when it is actually true and, consequently, the alternative hypothesis is accepted, although the observed effect is actually due to chance (false positive error).
  - Significance level (type 1 error rate): the probability of a type 1 error (denoted with "α")
    - The significance level is determined by the principal investigator before the study is conducted.
    - For medical/epidemiological studies, the significance level α is usually set to 0.05 (the lower α, the greater the statistical significance)
- **Multiple comparisons problem**
  - When multiple hypotheses are tested simultaneously with one data set (e.g., the data of a trial is tested for different outcomes), the probability of type 1 errors increases:
    - $P_1 = 1 - P_2$
    - $P_1 = 1 - (1 - \alpha)^m$
      - P_1 = at least one significant result of m test
      - P_2 = no significant results of m tests; α = type 1 error rate
    - Example: when testing 20 hypotheses on one set of data, each at a significance level of α = 0.05, the probability of obtaining at least one significant result by chance can be calculated as follows:
      - P(at least one significant result in 20 tests at α = 0.05) = $1 - (1 - 0.05)^{20}$)
      - P(at least one significant result in 20 tests at α = 0.05) ≈ 1 - 0.36 = 0.64)
      - For 20 independent hypotheses that are simultaneously tested on the same data, each at α = 0.05, the probability of a type 1 error in at least 1 test is approximately 64%.
  - There are methods to control the type 1 error rate for multiple comparisons; examples include:
    - Bonferroni correction: the α value is divided by the number of comparisons performed
    - False discovery rate: controls the proportion of false positives among the set of rejected hypotheses
  - Results that are not adjusted for multiple comparisons should not be used to infer treatment effects or make clinical decisions to avoid interpreting random results as significant.
- **Type 2 error**
  - The **null hypothesis is accepted** when it is actually false and, consequently, the alternative hypothesis is rejected even though an observed effect did not occur due to chance (false negative error).
  - Type 2 error rate: the probability of a type 2 error (denoted by "β")
  - Type 1 errors are inversely related to type 2 errors; The increase of one causes a decrease of the other.
- **Statistical power**: (1-β)
  - The probability of correctly rejecting the null hypothesis, i.e., the ability to detect a difference between two groups when there truly is a difference
  - Complementary to the type 2 error rate
  - Positively correlates with the sample size and the magnitude of the association of interest (e.g., **increasing the sample size** of a study would **increase its statistical power**)
  - Positively correlates with measurement accuracy
  - By convention, most studies aim to achieve 80% statistical power.
- **P-value**: the probability that the result of a given statistical test will be at least as extreme as the result actually observed, assuming that the null hypothesis is correct
  - Calculated using different statistical tests, depending on the type of data collected (e.g., parametric tests)

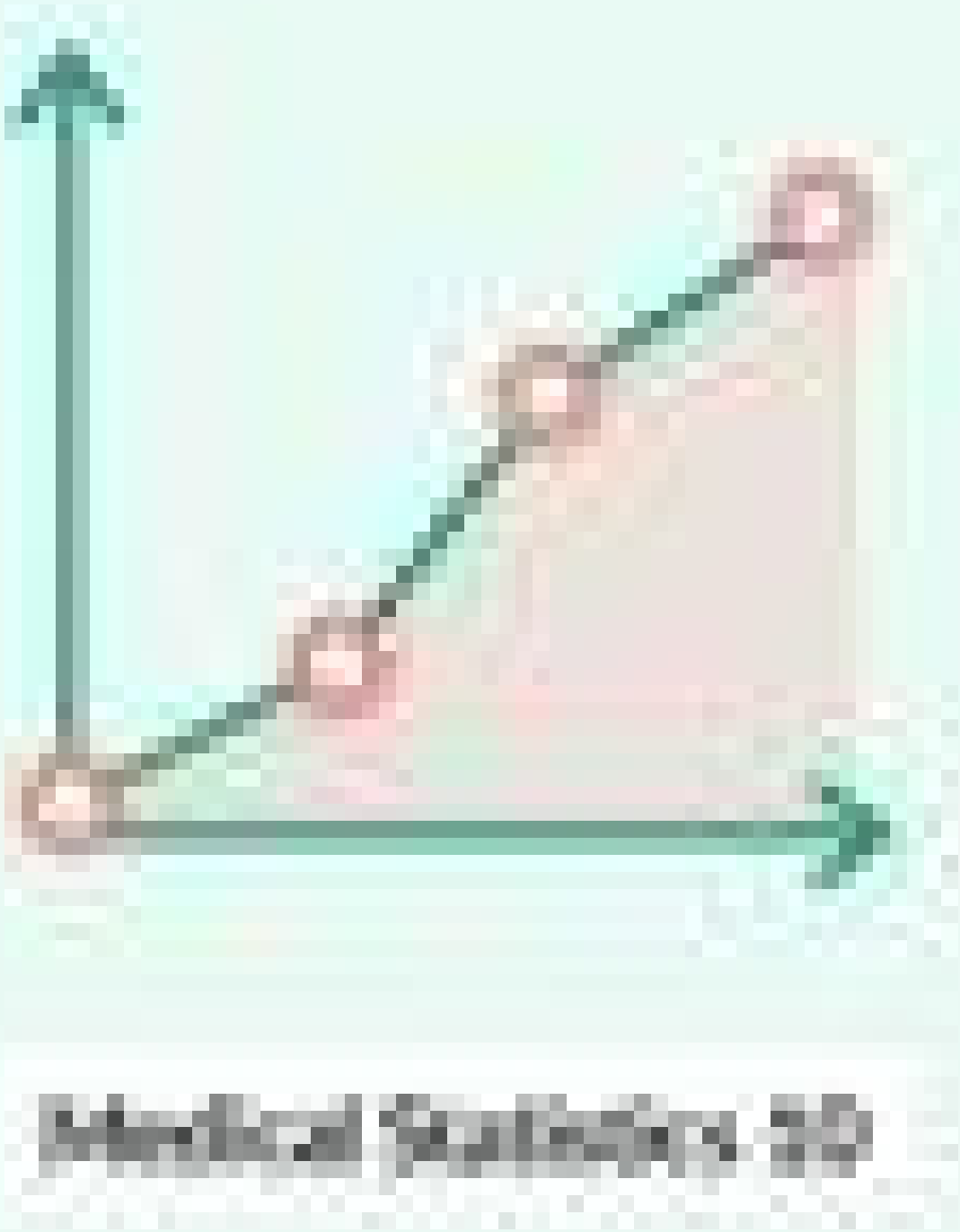
- Interpretation: The p-value is compared to the significance level (i.e., alpha or α-level), which is typically set at 0.05.
  - $p \leq \alpha$-level: The association is considered statistically significant and H_0 is rejected.
  - The p-value is not equivalent to the probability of H_0 being true, but rather to the probability of obtaining the same or more extreme results, assuming that H_0 is true.
  - The p-value cannot be used to prove H_1 but rather to prove that observed data is inconsistent with H_0.
  - When multiple comparisons are performed (e.g., ANOVA), the p-value must be compared to an adjusted α-level to ensure adequate statistical significance (e.g., α-level adjusted according to the Bonferroni correction)

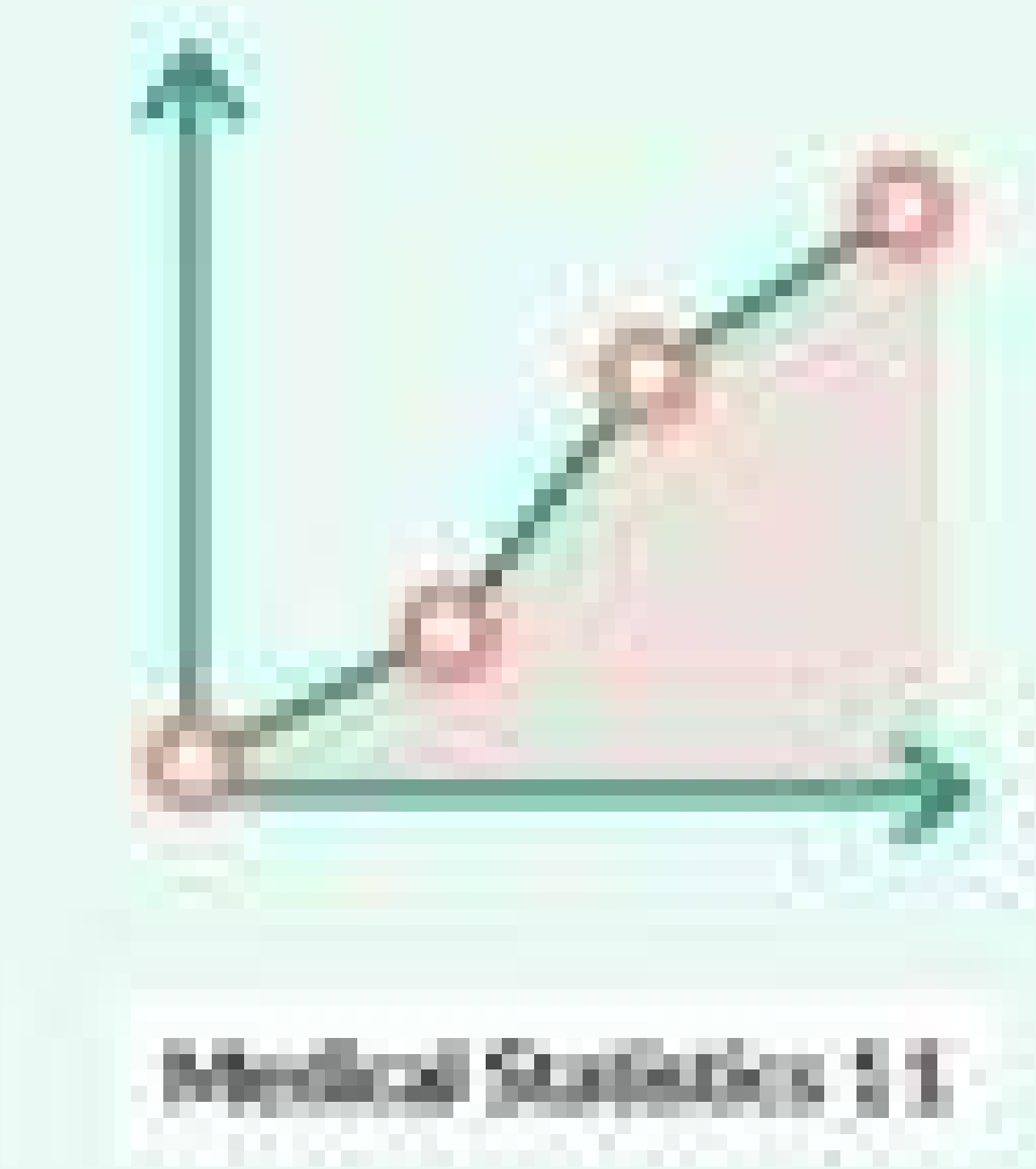
"The **A**ccusation is POSted **B**ut you NEGlect it!" (type I error (**A**lpha) is a false **pos**itive error and type II error (**B**eta) is false **neg**ative error)

"Statistical significance" does not mean "clinical significance."

Overview of errors

Overview		
Statistical test	**Null hypothesis (H_0) is true**	**Null hypothesis (H_0) is false**
Does not reject H_0	• 1-α	• Type 2 error (β)
Rejects H_0	• Type 1 error (α)	• Power (1-β)





Probability

- **Probability of an occurring event** (P)
  - Describes the degree of certainty that a particular event will take place (e.g., rolling a 6 is considered the event when tossing a die. When throwing a die, the probability of the event 6 occurring is 1/6)
  - P = number of favorable outcomes/total number of possible outcomes
- **Probability of an event not occurring** (Q)
  - The degree of certainty that a particular event will not take place (e.g., rolling a 6 is considered the event when tossing a die. When throwing a die, the probability of the event not occurring (rolling a “1”, “2”, “3”, “4”, or “5”) is 5/6)

- Q = number of unfavourable outcomes/total number of possible outcomes OR 1 - P

The actual probability of an event is not the same as the observed frequency of an event.

- **Probability of independent events**: The probability of event A is not contingent upon the probability of event B and vice versa. (e.g., eye-color and birthdays are two independent variables, with probability distributions independent of each other)
- **Conditional probability**: the probability of event A occurring given that event B has occurred
  - P(A|B) = P(A and B) / P(B)
    - P(B) = probability of event B
    - P(A and B) = probability of events A and B occurring simultaneously
    - Example: the probability of lung cancer in a smoker (A: occurrence of lung cancer; B: occurrence of smoking)
      - The underlying condition is that the individual is a smoker: P(B) = probability of being a smoker = number of smokers/total population
      - P(A and B) = probability of simultaneously being a smoker and having lung cancer = number of smokers with lung cancer/total population
      - Therefore, P(A|B) = the probability of lung cancer arising in a smoker = P(A and B)/P(B) = number of smokers with lung cancer/number of smokers
- **Multiplication rule**
  - P(A and B): the probability of events A and B occurring simultaneously
  - The multiplication rule is obtained by rearranging the formula for conditional probability.
    - For dependent conditions: P(A and B) = P(B) × P(A|B)
    - For independent conditions: P(A and B) = P(B) x P(A)
  - The multiplication rule can be applied to a decision tree (a visual representation of all possible outcomes) in order to calculate the probability of one of the branches (a particular outcome).
- **Addition rule**
  - P(A or B): the probability of either event A or B occurring
  - P(A or B) = P(A)+ P(B) − P(A and B): the probability of either event A or B occurring equals the sum of probabilities of events A and B minus the probability that they will occur simultaneously (nonmutually exclusive probability)
    - Example: the probability that an individual has a history of either myocardial infarction (event A) or stroke (event B) equals the probability of a history of myocardial infarction P(A) plus the probability of a history of stroke P(B) minus the probability of a history of both myocardial infarction and stroke P(A and B)
    - Example: meeting someone who has coronary artery disease (CAD) OR is obese : 0.3 + 0.4 - (0.3 × 0.4) = 0.58
  - If the events are mutually exclusive (mutually exclusive probability), P(A and B) = 0 and P(A or B) = P(A) + P(B)
    - Example: Since any individual can have only one blood group, the probability of having both groups A and B is 0, and the probability of having either blood group A or B is P(A) + P(B).
    - Example: Drawing an ace or a queen out of a deck with 52 cards: 4/52 + 4/52 = 8/52 = 2/13
- **Bayes theorem**
  - Bayes theorem is used to calculate conditional probabilities.
  - Bayes theorem describes the relationship between P(A|B) and P(B|A) :
    - P(A|B) = (P(B|A) × P(A)) / P(B)

Confidence interval

- **Overview**: Confidence intervals (CI) provide a way to test whether an effect size is statistically significant or not.
- **Definition**: the range of values that are highly likely to contain the true population measurement
  - Z scores for confidence intervals for normally distributed data (see Z score)
    - Z-score for a 95% confidence interval = 1.96
    - Z-score for a 97.5% confidence interval = 2.24

        - Z-score for a 99% confidence interval = 2.58
- **Formula**
    - The formula depends on the kind of data for which the confidence interval is calculated (e.g., means, proportions).
        - For confidence intervals for the mean: mean +/- Z score x (standard error of the mean)
        - For confidence intervals for the proportion: p +/- Z score x (√p x (1 - p)/n)
        - Requires the following values:
            - Confidence level (depends on the alpha level; if α-level = 5%, the confidence level is 95%))
            - Sample measurement
            - Standard error of the mean, which requires the sample size (a larger sample size lowers the standard error, resulting in more narrow confidence intervals) and standard deviation
- **Interpretation**
    - What kind of value does the confidence interval describe? Examples include a mean value (e.g., the height of students in a class), a difference between the means of two values (e.g., the difference between the mean height of students in class A and the mean height of students in class B), a relative risk (e.g., the relative risk of lung cancer in smokers vs. nonsmokers), and an odds ratio (odds of melanoma in residents of the Caribbean compared to residents in Vermont)
    - What is the confidence level?
        - Every alpha level has a corresponding CI of (1 - α)%.
            - An alpha level of 0.05 corresponds to a 95% confidence interval.
            - An alpha level of 0.01 corresponds to a 99% confidence interval.
    - Are several confidence intervals compared?
        - Nonoverlapping CIs **between two groups** signify a statistically significant difference.
        - Overlapping CIs may indicate that there is no statistically significant difference but can also occur with statistically significant differences.
    - What is the null value of the effect tested? 1 for ratios (e.g., relative risk, odds ratio) and 0 for differences (e.g., attributable risk, absolute risk reduction)
    - Does the CI include the null value?
        - If the CI includes the null value (i.e., 0 for differences or 1 for ratios), the result is not statistically significant, and the null hypothesis cannot be rejected.
        - If the CI does not include the null value, the result is statistically significant, and the null hypothesis can be rejected. If the results of a study are statistically significant, i.e., p-value < α-level, the associated CI does not include the null value.
    - How wide is the confidence interval?
        - The wider the CI, the less significant the findings of a given statistical test.
        - A larger sample size typically results in a narrower CI.
        - A narrow, statistically significant CI typically indicates a small p-value.

Statistical tests

Statistical significance vs. clinical significance [3][5]

- Significance (epidemiology): the statistical probability that a result did not occur by chance alone
    - Statistical significance
        - Describes a true statistical outcome (i.e., one that is determined by statistical tests) that has not occurred by chance
        - If the statistical significance is high, the probability that the results are due to chance is low.
    - Clinical significance (epidemiology)
        - Describes an important change in a patient's clinical condition, which may or may not be due to an intervention introduced during a clinical study
        - If the clinical significance is high, the intervention is likely to have had a great impact on patient outcome or measures.

- Statistical and clinical significance do not necessarily correlate. A study might have a high statistical significance but the tested intervention did not have any clinical impact on patient outcome.

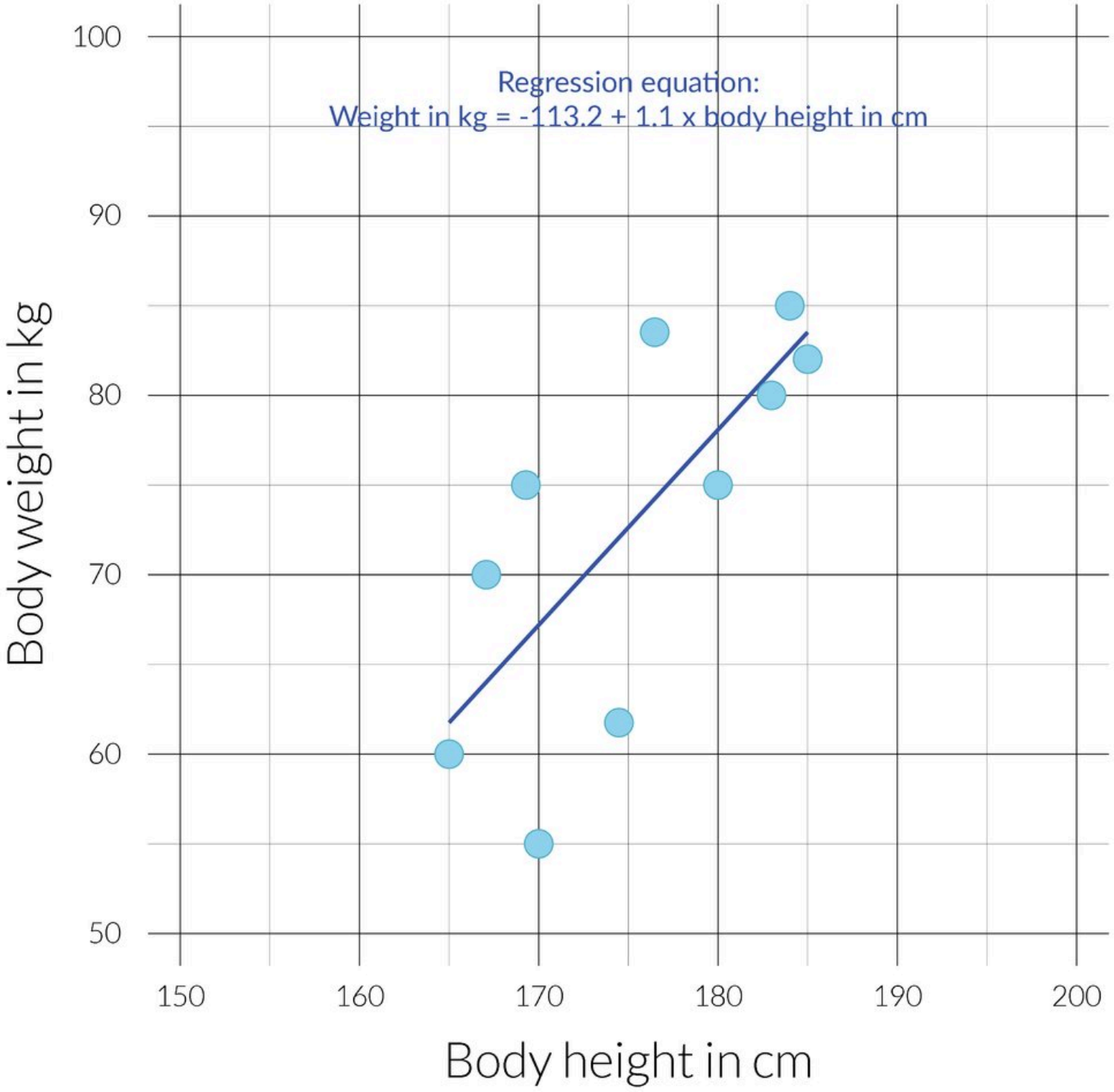
Correlation and regression

Correlation

- **Definition**
  - A measure of the linear statistical correlation between continuous variables
  - The correlation coefficient can lie anywhere between +1 and -1.
- **Example**: how does y change if x is changed?
- **Interpretation**: A correlation coefficient measures the strength (i.e., the degree) and direction (i.e., a positive or negative relationship) of a linear relationship (does not require causality).
  - Direction or relationship: can be positive or negative (which are identified by a plus or minus, respectively)
  - Strength of relationship
    - Perfect relationship: two variables are perfectly linear and the correlation coefficient is +1 or -1
    - No linear relationship: correlation coefficient is 0
  - See "Spearman correlation coefficient" and "Pearson correlation coefficient."

Regression (epidemiology)

- **Definition**: the process of developing a mathematical relationship between the dependent variable (the outcome; y) and one or more independent variables (the exposure; x)
- **Linear regression**: a type of regression in which the dependent variable is continuous
  - Simple linear regression
    - 1 independent variable is analyzed
    - If y has a linear relationship with an independent variable x, a graph plotting this relationship takes the form of a straight line (called regression line).
    - In the case of simple linear regression, the equation of the regression line is: y = mx + b, with m representing the slope of the regression line, y the dependent variable, x the independent variable, and b the y-intercept (the value of y where the line crosses the y-axis)
  - Multiple linear regression: > 1 independent variable is analyzed
- **Logistic regression**: a type of regression in which the dependent variable is categorical
  - Simple logistic regression: 1 independent variable is analyzed
  - Multiple logistic regression: > 1 independent variable is analyzed



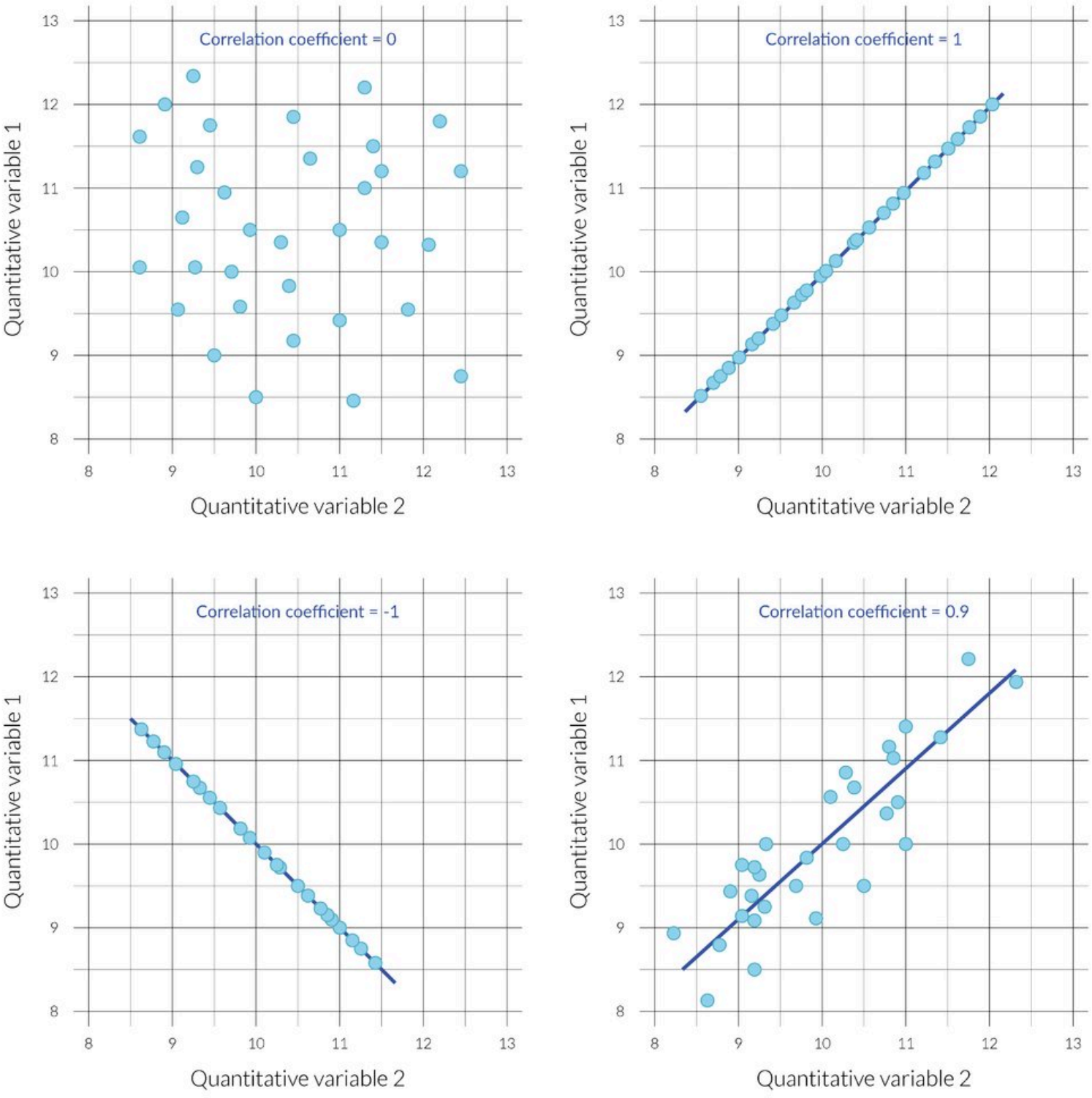
Parametric tests

Parametric tests are used to evaluate statistically significant differences between groups when the study sample has a normal distribution and the sample size is large.

- **Pearson correlation coefficient (r)**
  - Compares interval level variables
  - Calculates the estimated strength and direction of a relationship between two variables
  - Interpretation
    - r is always a value between -1 and 1.

        - A positive r-value = a positive correlation
        - A negative r-value = negative correlation
        - The closer the r-value is to 1, the stronger the correlation between the compared variables.
        - The coefficient of determination = r^2 (the coefficient may be affected by extreme values and indicates the proportion of a variable's variance that can be predicted by the variance of another variable)
- **T-test**
    - Calculates the difference between the means of two samples or between a sample and population or a value subject to change; especially when samples are small and/or the population or a value subject to change distribution is not known
    - Used to determine the confidence intervals of a t-distribution (a collection of distributions in which the standard deviation is unknown and/or the sample size is small)
    - One sample t-test
        - Interpretation
            - The t-value can be classified according a table that lists t-values and their corresponding quantiles based on the number of degrees of freedom (df) and the significance level (α value).
                - $|t|$ < tabular value of t_{df} $(1-\alpha/2)$: null hypothesis cannot be rejected
                - $|t|$ > tabular value of t_{df} $(1-\alpha/2)$: null hypothesis should be rejected
            - Alternatively, one may calculate the confidence intervals of the sample observations and check if the population mean (μ_0) falls within the range given by the confidence intervals.
        - Formula: t-value = (sample mean - population mean)/standard deviation) * √(n)
        - Prerequisite: normal distribution (the variance is known and depends on the degrees of freedom.)
        - Calculates whether a sample mean differs from the population mean (μ_0)
    - Two sample t-test
        - Calculates whether the means of two groups differ from one another
        - Prerequisites
            - Both sample groups are drawn from the same population and have the same (but unknown) variance.
            - The difference between the observations in the two groups approximately follows a normal distribution.
        - Formula: t-value = (mean difference between the two samples/standard deviation) * √(n)
        - Interpretation: The t-value is compared with a table of t-values in order to determine whether the difference is statistically significant (similar to the one sample t-test described above).
        - Unpaired t-test (independent samples t-test)
            - Two different groups are sampled at the same time
            - The difference between the means of a continuous outcome variable of these 2 groups is compared
            - The null hypothesis is that the mean of these two groups is equal
            - A statistically significant difference rejects the null hypothesis
        - Paired t-test (dependent samples t-test)
            - The same group is sampled at two different times
            - The difference between the means of a continuous outcome variable of this group is compared
            - The null hypothesis is that the group mean is equal at these two different times
            - A statistically significant difference rejects the null hypothesis
- **Analysis of variance (ANOVA)**
    - Calculates if there is a statistically significant difference between ≥ 2 groups by comparing their means
    - The aim is to determine whether there is a statistically significant effect of one or more independent variable(s) on a dependent variable (the mean).
    - Can be seen as an extension of the t-test (which can only be used for the analysis of two groups)

- One-way ANOVA: assesses if there is a statistically significant difference in the means of 1 independent variable (e.g., the mean height of women in clinics A, B, and C; the independent variable is the clinic, the dependent variable is the height)
- Two-way ANOVA: assesses if there is a statistically significant difference in the means of 2 independent variables (e.g., the mean height of women and men in clinics A, B, and C at a point in time; the independent variables are sex category and clinic, the dependent variable is the height)



T-test has **2 syllables** and differentiates between **2 groups**; ANOVA has more than **2 syllables** and can be used for 2 or more groups.

Nonparametric tests

Nonparametric tests are used to evaluate the statistically significant difference between groups when the sample has nonnormal distribution and the sample size is small.

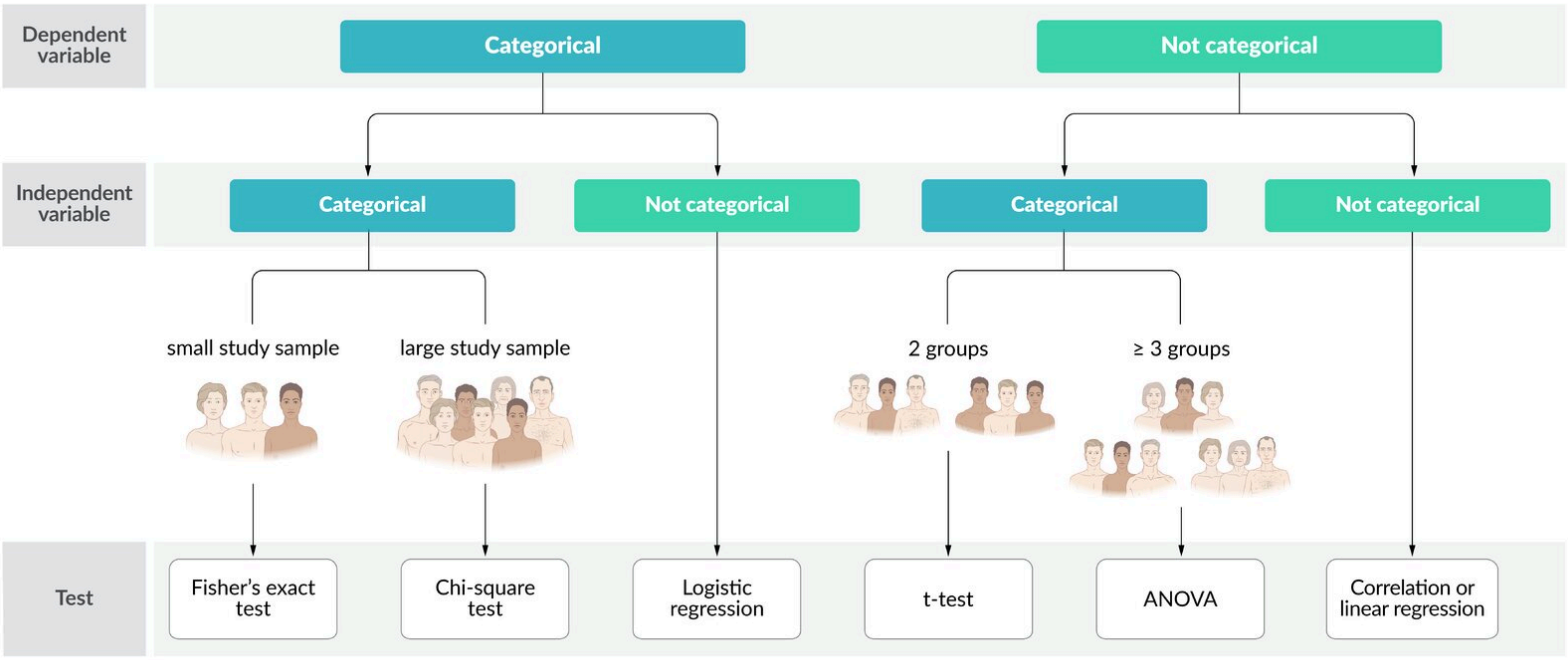
- **Spearman correlation coefficient**
  - Calculates the relationship between two variables according to their rank
  - Compares ordinal level variables
  - Interpretation
    - Extreme values have a minimal effect on Spearman coefficient.
    - Not precise because not all information from the data set is used.
  - See "Correlation."
- **Mann-Whitney U test**
  - Compares ordinal, interval, or ratio scales
  - Calculates whether two independently chosen samples originate from the same population and have identical distributions and/or medians
  - Example: comparing two groups of high school students – one with an average GPA of 4.2 and the other an average GPA of 3.2 – to determine if both came from the same larger group.
- **Wilcoxon test** (rank sum and signed rank)
  - Rank sum test: compares the means between groups of different sizes
  - Signed rank test: compares the means between pairs of scores that can be matched; substitute for the one-sample t-test when a pre-intervention measure is compared with a post-treatment measure and the null hypothesis is that the treatment has no effect
- **Kruskal-Wallis H test**
  - Extension of the Mann-whitney U test
  - Compares multiple groups by testing the null hypothesis (that there is no median difference between at least two groups)
- **Binomial test**
  - Examines whether the observed frequency of an event with binary outcomes (e.g., heads/tails, dead/alive) is statistically probable or not
  - Example: if a coin is tossed 20 times, it is likely to land on heads approx. 10 times. If only 9 heads come up, the result is still acceptable.
    - If only 6 heads come up ($p = 0.04$), one is left wondering whether the coin is biased
    - From a statistical perspective, the result of a coin toss should be questioned if fewer than 25% of the coin tosses result in heads because the probability of such an event is $< 2.5\%$

Categorical tests

Categorical tests are used to evaluate the statistically significant difference between groups with categorical variables (no mean values).

- **Chi-square test** (X^2 test)
  - Used to compare the distributions of two categorical variables.
    - The independent variable can be composed of ≥ 2 categorical groups (e.g., treatment groups).
    - The dependent variable (e.g., survival) is limited to 2 groups (true/not true).
  - A chi-square test compares proportions in two or more sets of categorical data to determine whether there is a statistically significant difference in the distribution (e.g., the proportion of patients with lung disease in clinics A, B, and C at a certain point in time or the proportion of individuals with diabetes in four different ethnic groups)
- **Fisher exact test**
  - Also calculates the difference between the frequencies in a sample but, unlike a Chi-square test, is used when the study sample is small
  - Also aims to determine how likely it was the outcomes occurred due to chance

"Chi-tegorical:" Chi-square test is used for categorical variables.


Dependent variable
Categorical
Not categorical
Independent variable
Categorical
Not categorical
Categorical
Not categorical
small study sample
large study sample
2 groups
≥ 3 groups
Test
Fisher's exact test
Chi-square test
Logistic regression
t-test
ANOVA
Correlation or linear regression