Introduction to biostatistics

Hristo Todorov,
AK Prof. Dr. Susanne Gerber

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

JG|U

What statistics are (not) all about

"Data don't make any sense,
we will have to resort to statistics."

Introduction to statistics

Branches of statistics

- **Descriptive statistics**
  - Describe, summarize, order or graphically represent empirical data
- **Exploratory data analysis**
  - Identify patterns or structures in the data
- **Inferential statistics**
  - Predict, estimate and generalize about populations based on data derived from samples

Descriptive statistics

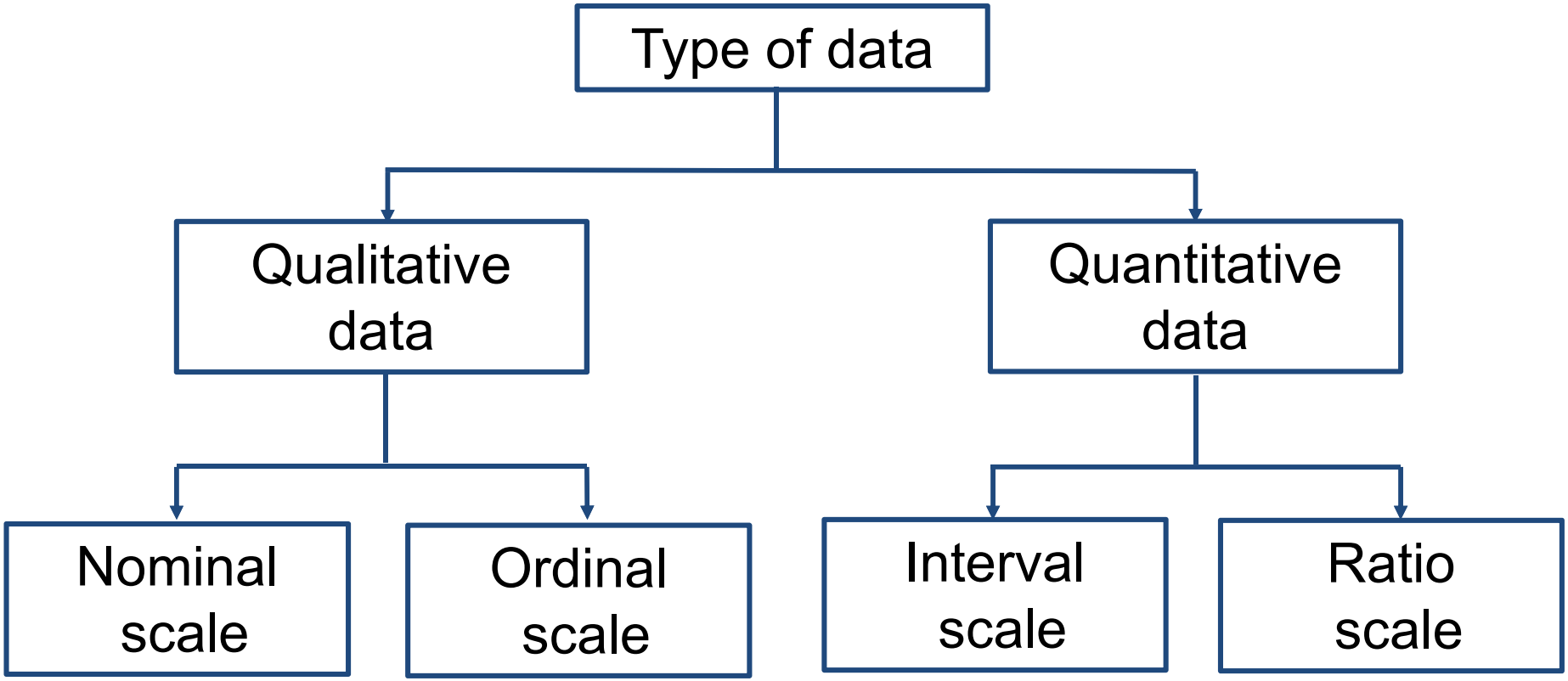
Descriptive statistics

Introduction to descriptive statistics

- Descriptive statistics provide the basis of quantitative data analysis
- The focus is to describe, summarize and graphically represent the data as it is without making any assumptions or generalizations
- Descriptive statistics provide measures to:
  - Describe the central tendency of data
    - Mean, median, modus
  - Describe the dispersion of the data
    - Variance, standard deviation, range
  - Describe how different data are related to each other
    - Correlation coefficient

Descriptive statistics

Measurement scales



Descriptive statistics

Measurement scales

- Nominal scale
  - The lowest level of measurement
  - Data belong to mutually exclusive categories
  - The only possible comparison between elements is “equal to” or “not equal to”
  - Examples include:
    - Blood type
    - Gender
    - Color of your eyes
    - Nucleotides in the DNA sequence
    - University field of studies

Descriptive statistics

Measurement scales

- Ordinal scale
  - Data belong to mutually exclusive categories which can be ordered (<,> comparisons allowed)
  - Differences between categories are not allowed
  - Examples include:
    - Exam grades (A, B, C, D)
    - Questionnaire options (“strongly agree”, “agree”, disagree”, strongly disagree”)
    - Levels of happiness, satisfaction, etc.
    - Some clinical scores

Descriptive statistics

Measurement scales

- Interval scale
  - A metric scale which allows building differences between values but not ratios
  - Examples include:
    - Celsius temperature
    - IQ score
    - Time on a clock
  - No “true” zero value is defined on an interval scale
    - 0 degrees Celsius does not mean there is no temperature
    - 0:00 does not mean time does not exist

Descriptive statistics

Measurement scales

- Ratio scale
  - A metric scale which allows building differences and ratios between values
  - A value of zero indicates non-existence
  - Examples include:
    - Age
    - Height
    - Weight
    - Number of children
    - Distance
    - Blood pressure

- Note: higher levels of measurements can be reduced to lower scales but not the other way around

Descriptive statistics

Measures of central tendency

- Central tendency, center or location of the data is a central or typical value

- Mode
  - The element if the data which appears most often
  - The mode can be determined for variables measured on any scale

  Out of 40 participants in a statistics course, 10 are studying bioinformatics, 25 are studying biomedicine and 5 are studying biochemistry. What is the mode for the the variable field of studies of the course takers?

Descriptive statistics

Measures of central tendency

- Median
  - The “middle” value in an ordered data set, half of the data are smaller and half of the data are larger than the median
  - Can be calculated for data measured at least on an ordinal scale

- Arithmetic mean
  - The arithmetic mean only makes sense for variables on a metric scale

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

Descriptive statistics

Measures of central tendency

- Mean or median?
- Consider a small sample containing the following values

$$\{3.5; 4; 2.5; 6; 7; 5.5\}$$

$$\bar{x} = \frac{3.5 + 4 + 2.5 + 6 + 7 + 5.5}{6} = 4.75$$

$$Median = 4.75$$

- Consider adding the value 500 to the set

$$\{3.5; 4; 2.5; 6; 7; 5.5; 500\}$$

$$\bar{x} = \frac{3.5 + 4 + 2.5 + 6 + 7 + 5.5 + 500}{7} = 75.5$$

$$Median = 5.5 - \ a\ robust\ statistic$$

Descriptive statistics

Measures of dispersion

- Empirical variance

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

- Sample variance

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

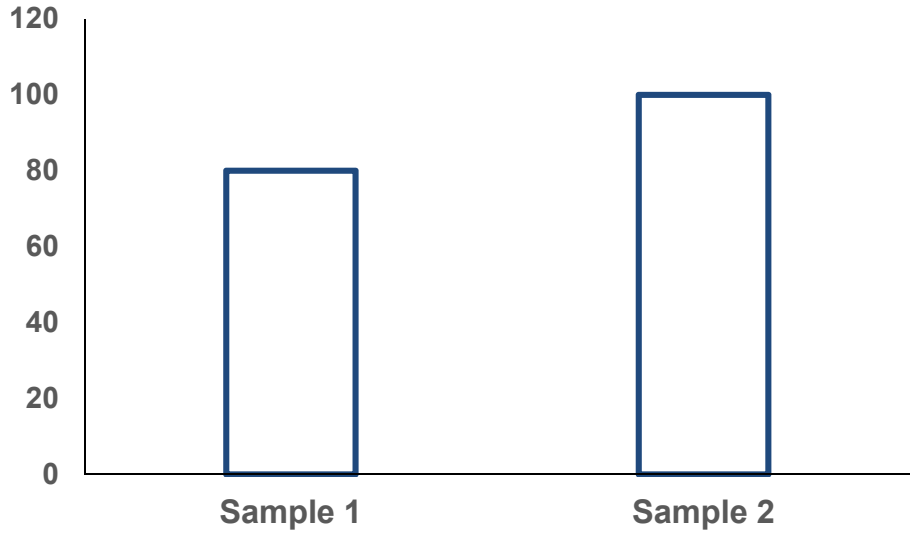
- Standard deviation

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

Descriptive statistics

Graphical representation of location and dispersion

- *Suppose that blood pressure was measured in two samples of 30 patients each. Mean diastolic blood pressure for the first sample turned out to be 80, whereas mean pressure in the second sample was 100 mmHg.*

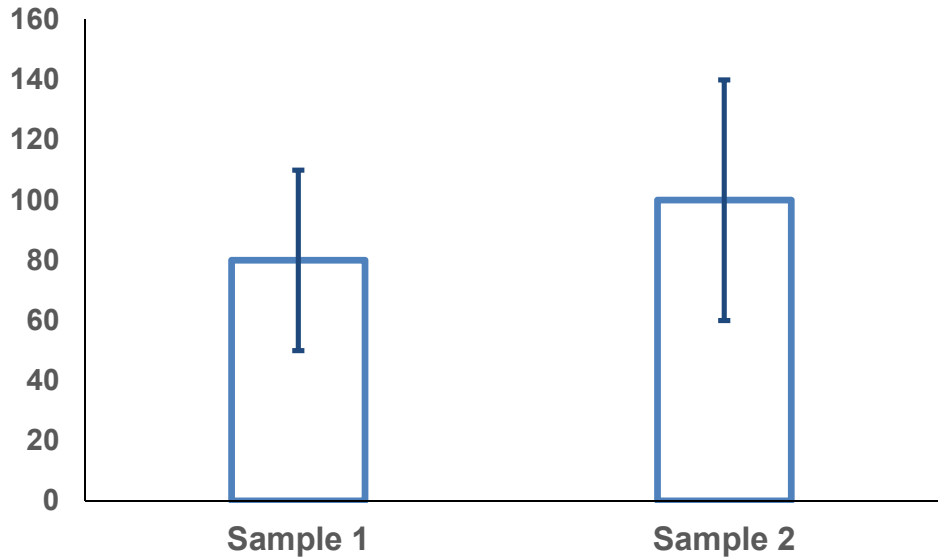


- *Do peaple in the second sample have higher diastolic blood pressure?*

Descriptive statistics

Graphical representation of location and dispersion

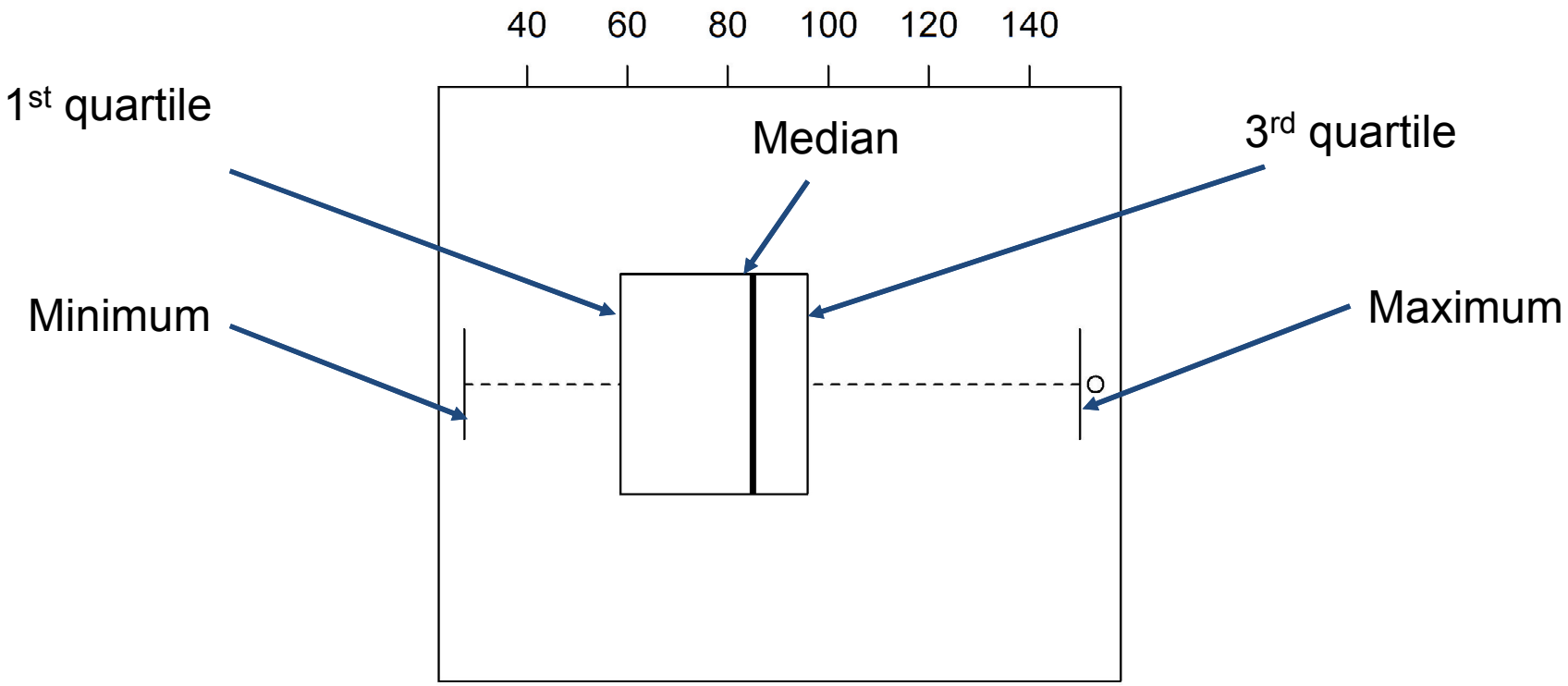
- *Suppose that blood pressure was measured in two samples of 30 patients each. Mean diastolic blood pressure for the first sample turned out to be 80 ± 30, whereas mean pressure in the second sample was 100 ± 40 mmHg.*



Descriptive statistics

Boxplot

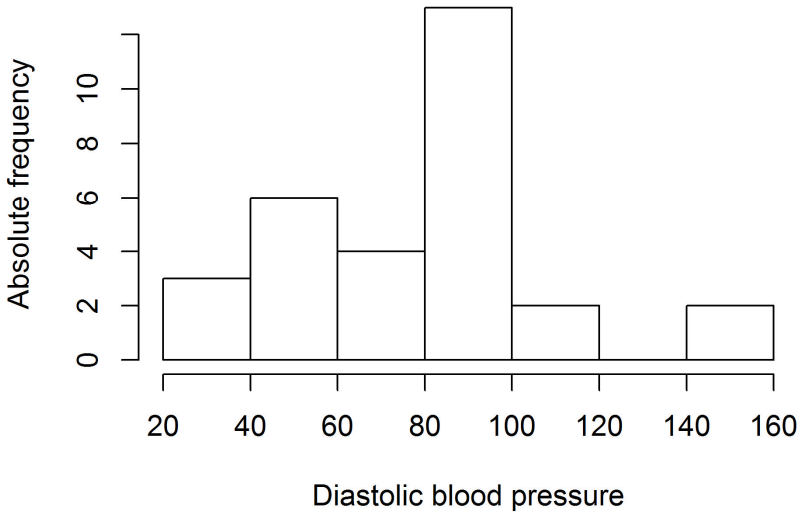
- Consider the first sample of 30 patients with mean diastolic blood pressure of 80, with a standard deviation of 30

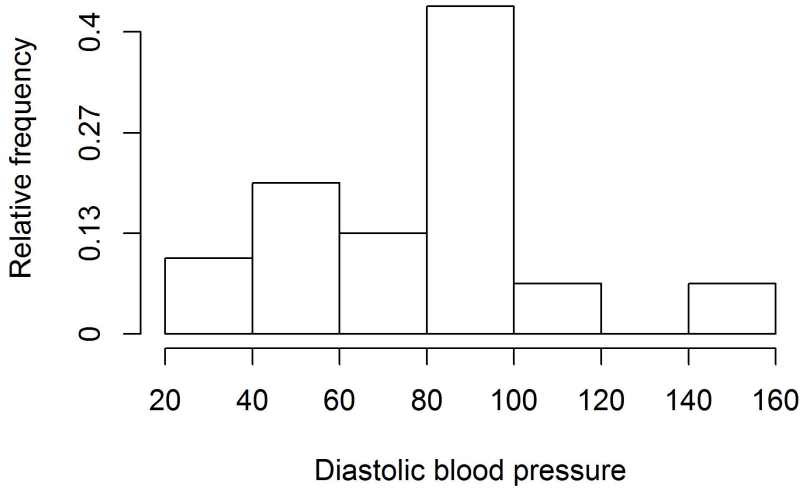


Descriptive statistics

Histograms

- Consider the first sample of 30 patients with mean diastolic blood pressure of 80, with a standard deviation of 30

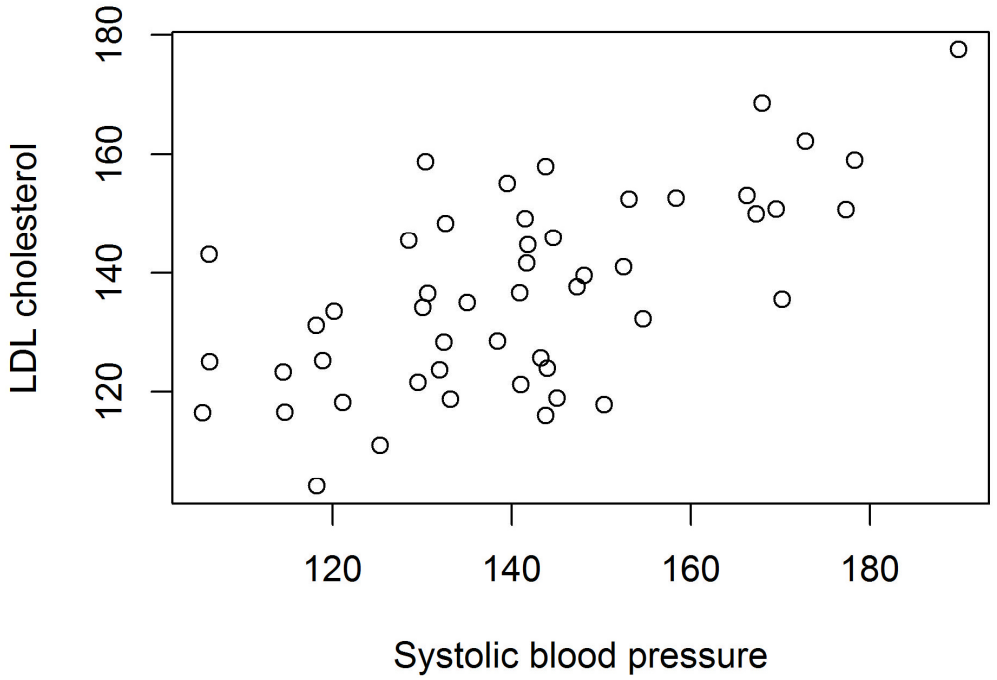




Descriptive statistics

Association between metric variables

- *In a study of 50 patients, systolic blood pressure and LDL cholesterol were measured. Mean blood pressure was 137.5 mmHg with a standard deviation of 18.27. Average LDL cholesterol was 134 mg/dL with a standard deviation of 14.14.*
  - *Is higher blood pressure associated with higher LDL cholesterol levels?*



Descriptive statistics

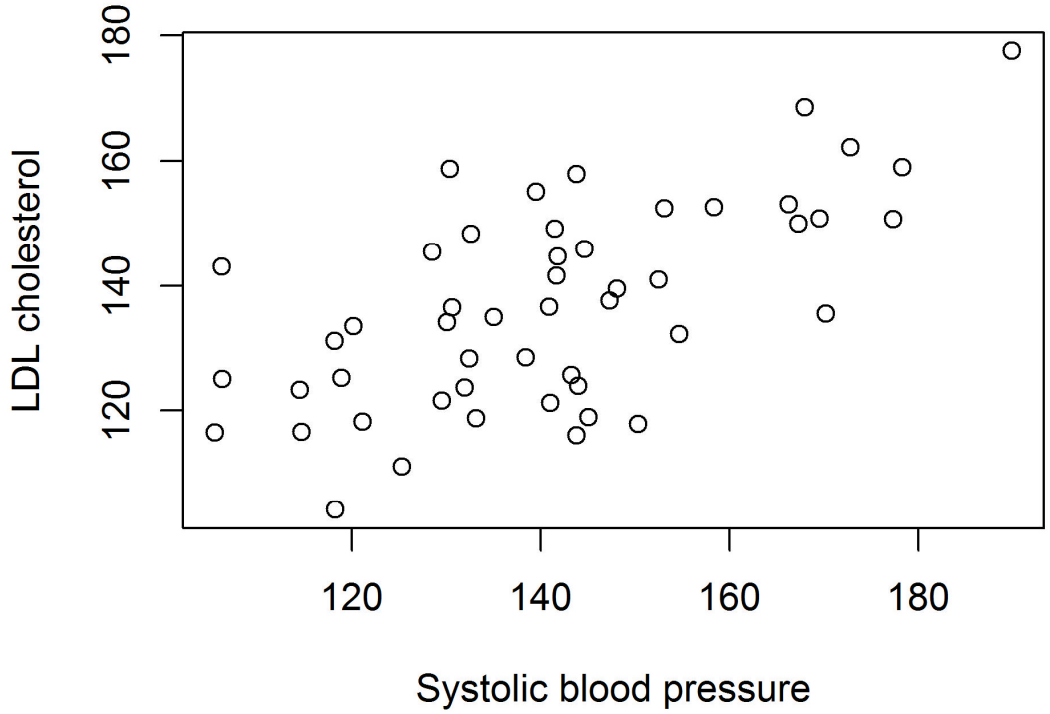
Association between metric variables

- Pearson correlation coefficient
  - Measures the linear relationship between metric variables
  - Values range from -1 to 1
    - -1 - perfect negative correlation
    - 0 – no correlation
    - +1 – perfect positive correlation

Descriptive statistics

Association between metric variables

- *In a study of 50 patients, systolic blood pressure and LDL cholesterol were measured. Mean blood pressure was 137.5 mmHg with a standard deviation of 18.27. Average LDL cholesterol was 134 mg/dL with a standard deviation of 14.14.*
  - *Is higher blood pressure associated with higher LDL cholesterol levels?*

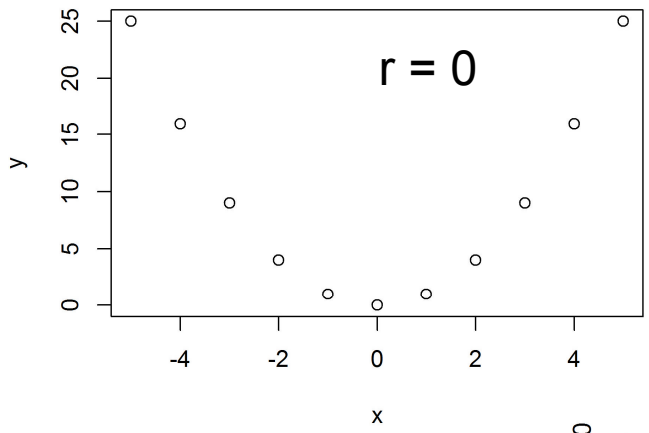


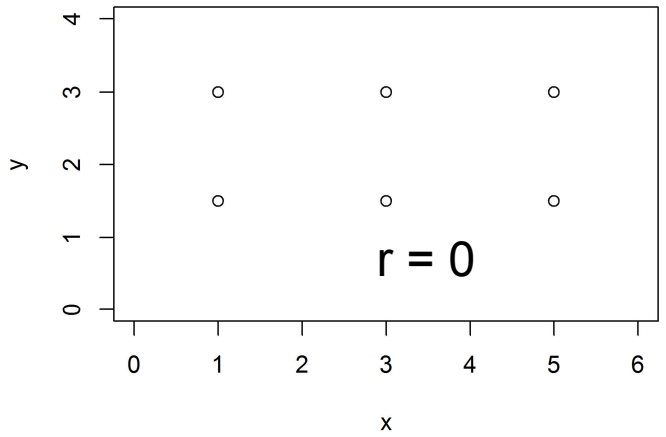
Pearson r = 0.64

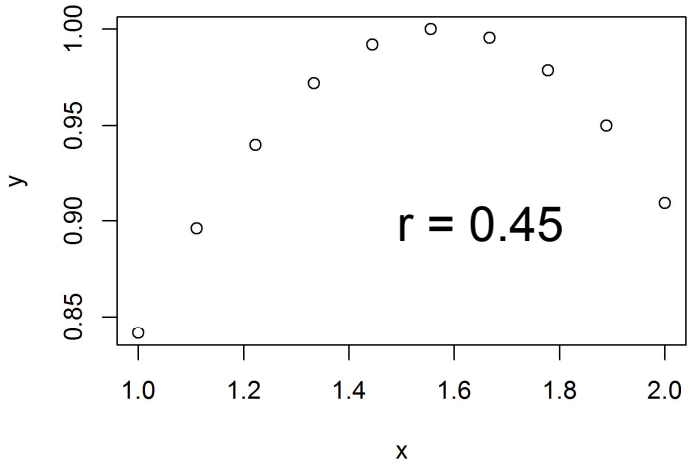
Descriptive statistics

Pearson correlation

- Pearson correlation coefficient is only appropriate if the relationship between the variables is linear







Descriptive statistics

Spearman correlation

- The Spearman correlation coefficient is an alternative to the Pearson correlation coefficient appropriate when:
  - The relationship between the variables is monotonically increasing or decreasing but not linear
  - Data are measured on an ordinal scale
- Values range from -1 to 1, 0 corresponds to no correlation
- Data are arranged in an increasing order and every data point is assigned a rank
- Spearman correlation belongs to the category of robust statistics because it is robust in the presence of outliers

Descriptive statistics

Pearson vs. Spearman correlation

- Influence of outliers

x	y
1	5
2	6
3	7
4	8
Pearson r = 1	
Spearman r = 1	

x	y
1	5
2	6
3	7
4	8
100	9
Pearson r = 0.72	
Spearman r = 1	

Descriptive statistics

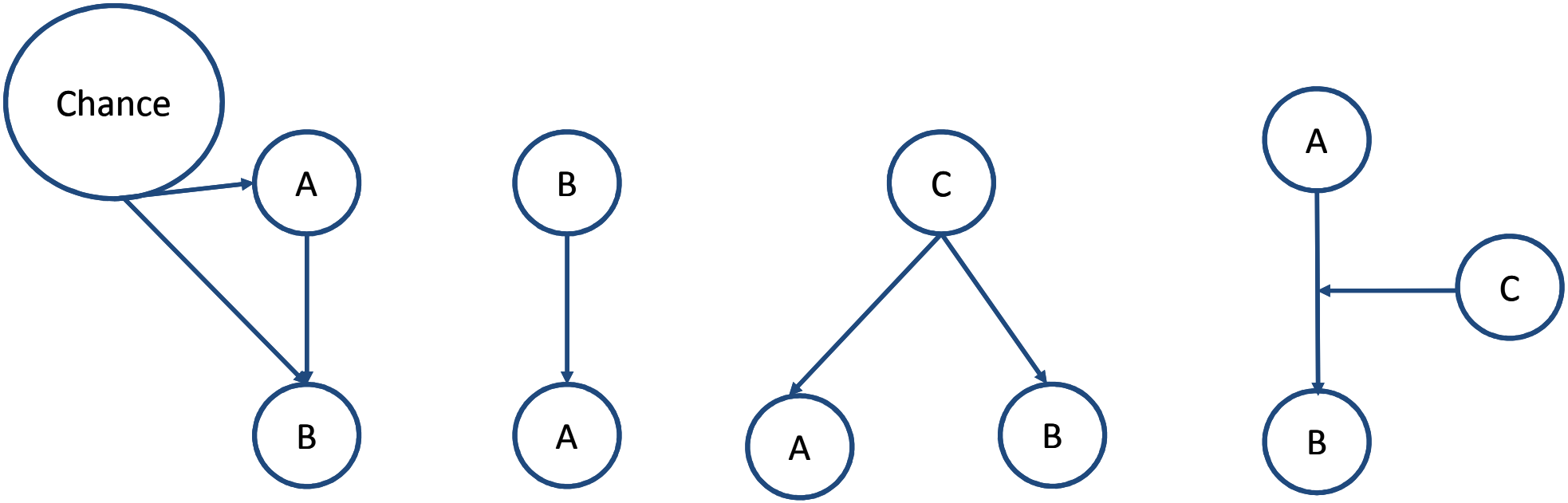
Correlation versus causality

- Causality is often the focus of research
  - We try to explain why phenomena occur and what causes them
- Caution: statistics only provides measures of association. The question of what is cause and what is consequence remains open.
- Causality, therefore, requires careful theoretical consideration and appropriate experimental designs
  - Causality can often be masked by random confounding factors or latent factors

Descriptive statistics

Correlation versus causality

- Consider two variables A and B which are highly correlated with each other.
- Different causal relationships are possible



Probability theory

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Probability theory

Random processes

- Random process – a phenomenon with an uncertain outcome. Examples include:
  - A coin toss or throwing a dice
  - The gender of an unborn child
  - Result from an exam
  - A scientific experiment
- The set of all possible outcomes of a random process is called a **sample space Ω.**
  - E.g. The colour of a gummy bear we take out of a bag with blue, red, yellow and green gummy bears is a random experiment with the following sample space:

  Ω = {„blue“, „red“, „yellow“, „green“}

Probability theory

Radom events

- The outcome of a **random experiment** is called a **random event**
  - **Simple event** – an event which contains only a single outcome

    E.g. A green gummy bear
  - An event is a unification of simple events

    E.g. A green gummy bear or a yellow gummy bear;

    Not a blue gummy bear
- Although the single outcome of a random process is uncertain, outcomes follow certain distributions if the random experiment is repeated many times
- The chance of a random outcome occurring is described by **the probability**

Probability theory

Frequency versus probability

- Consider the following random experiment:
  - *A pot contains 50 white and 50 black balls. What is the probability of drawing a white ball if we draw if replacement (we put the ball back in the pot after drawing it)*
  - *An R simulation was performed investigating relative frequency of white balls after repeating the experiment 10 000 times by drawing 1 to 10 000 balls.*

Probability theory

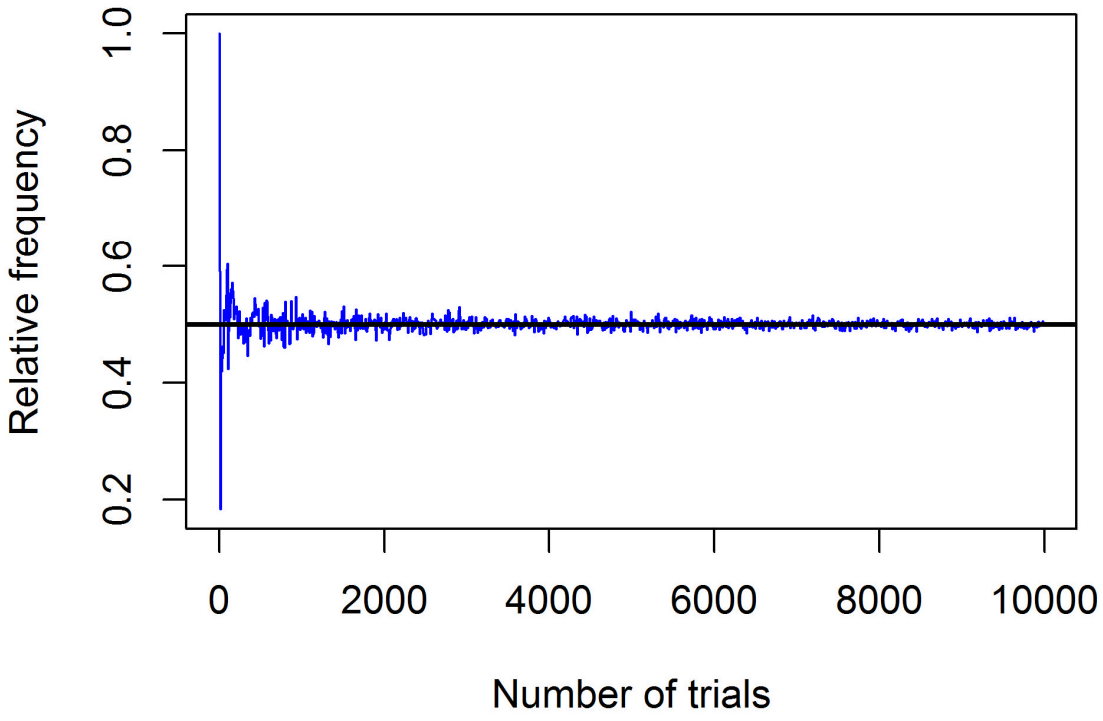
Frequency versus probability

Number of draws	Number of white balls	Relative frequency of white balls
1	1	1
11	2	0.182
101	61	0.604
501	238	0.475
1001	494	0.494
9991	4985	0.499

- The relative frequency approaches 0.5 as the number of trials increases

Probability theory

Frequency versus probability



- The **relative frequency** converges against a limit value as the number of trials **n** increases. The limit is called **probability**

Probability theory

Laplace probability

- In a Laplace experiment all simple events have the same probability
- Let **P(A)** describe the probability that event **A** occurs. The Laplace probability is then:

$$P(A) = \frac{Number\ of\ elements\ in\ A}{Number\ of\ all\ elements}$$

E.g. 50 students are taking part in a statistics course, 15 of them are studying bioinformatics, 25 are studying biomedicine and 10 are studying biophysics. What is the probability that a randomly selected student is studying biomedicine?

$$P = \frac{25}{50} = 0.5$$

Probability theory

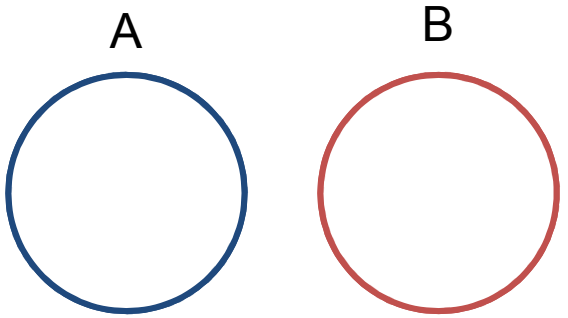
Rules for probabilities

- The probability must satisfy the following conditions:
  1. P(A) ≥ 0, the probability is always positive
  2. P(Ω) = 1, every random experiment must have an outcome, therefore the sum of the probabilities of all possible outcomes is 1
  3. P(A or B) = P(A) + P(B) if A and B are **disjoint**. Two events are disjoint if they cannot occur together

     E.g. There are red, blue and yellow balls in a pot and the probabilities of drawing a red, blue or yellow balls are 0.2, 0.5 and 0.3, respectively. What is the probability of drawing a red or a blue ball?

     P(red ball or blue ball) =

     P(red ball) + P(blue ball) = 0.2 + 0.5 = 0.7



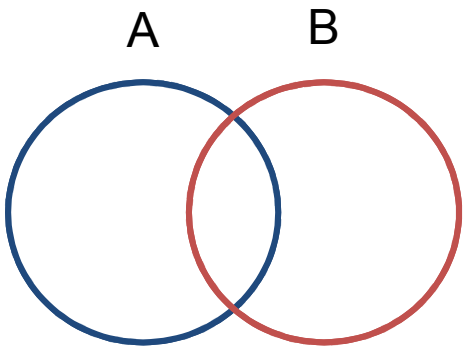
Probability theory

Rules for probabilities

- P(A or B) = P(A) + P(B) – P(A and B) if A and B are not disjoint

E.g. The probability of suffering from hypertension over the age of 65 is 60% and the probability of having a heart arrhythmia is 10 %. 30% of the population over 65 has both hypertension and a heart arrhythmia. What is the probability of having at least one of these conditions over the age of 65?

$$P(H\ or\ A) = P(H) + P(A) - P(H\ and\ A)$$
$$P(H\ or\ A) = 0.6 + 0.1 - 0.3 = 0.4$$



Probability theory

Independent events

- Two events are **stochastically independent** if the occurrence of one of the events does not influence the probability of the other event occurring
  - Flipping a coin multiple times
  - Throwing a dice multiple times
  - Drawing balls from a pot with replacement
- The probability that two **independent** events **A** and **B** occur simultaneously is the product of the respective probabilities for A and B:

$$P(A \text{ and } B) = P(A)P(B)$$

E.g. the probability two get first tail and then head after flipping a coin twice is:

$$P(first\ tail\ and\ then\ head) = P(tail)P(head) = 0.5 * 0.5 = 0.25$$

- What is the probability of getting head and tail after flipping a coin twice?

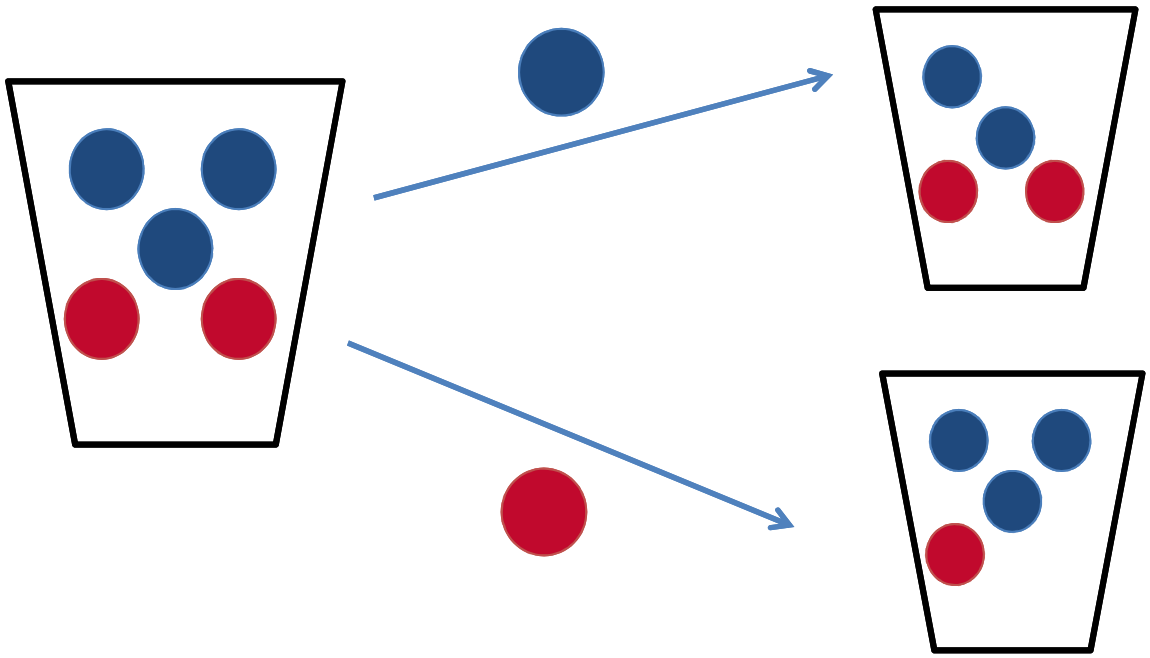
Probability theory

Dependent events

- Two events are **dependent** if the occurrence of one of the events influences the probability of the second event.

For example, suppose we are drawing balls from a pot without replacement. The pot contains 3 blue and 2 red balls. What is the probability that the second ball we draw is red?

- *Depends on the color of the first ball we draw.*

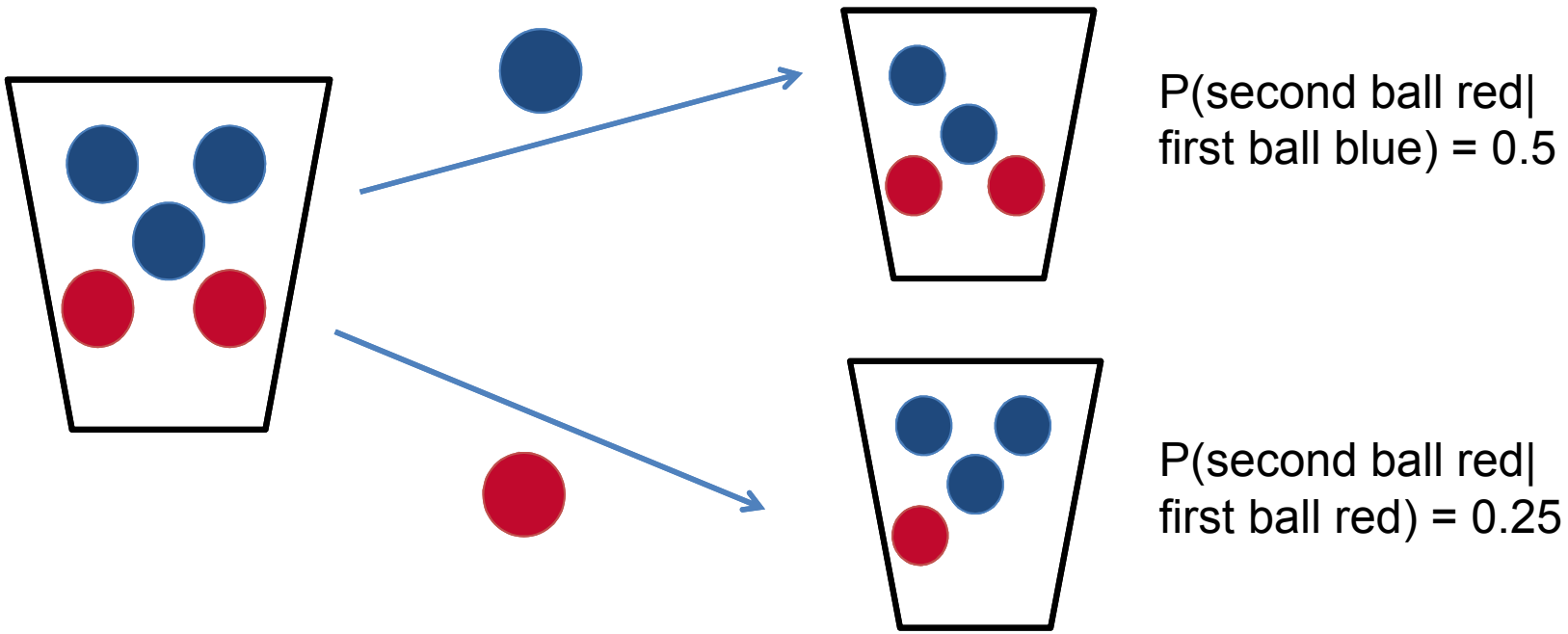


Probability theory

Conditional probability

- The **conditional probability** describes the probability of an event **A** when we know that an event **B** has occurred

$$P(A|B)$$



Probability theory

Conditional probability

- The probability that two **dependent events A** and **B** occur simultaneously:

$$\mathrm{P(A\ and\ B) = P(A|B)P(B) = P(B|A)P(A)}$$

  - The temporal sequence in which events occur is irrelevant

- The **conditional probability** is therefore:

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Probability theory

Conditional probability

Example: *A test T used to detect a specific disease D is positive in 99% of the cases when a patients truly suffers from the disease. Altogether, the test is positive in 3 % of tested patients and the prevalence of the disease in the investigated population is 1 %. What is the probability that a patient truly has the disease when the test result is positive?*

P(T) – probability that the test is positive

P(D) – prevalence of the disease

P(T|D) – probability that the test is positive when a patient truly has the disease

P(D|T) – probability that the patient truly suffers from the disease knowing that the test result is positive

Probability theory

Conditional probability

Example: *A test T used to detect a specific disease D is positive in 99% of the cases when a patients truly suffers from the disease. Altogether, the test is positive in 3 % of tested patients and the prevalence of the disease in the investigated population is 1 %. What is the probability that a patient truly has the disease when the test result is positive?*

P(T) = 0.03

P(D) = 0.01

P(T|D) = 0.99

P(D|T) = ?

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{0.99 * 0.01}{0.03} = 0.33$$

Random variables and probability distributions

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

JG|U

Random variables

Definition of a random variable

- A **random variable** is a numerical representation of a random phenomenon. The values of the random variable correspond to the outcomes of the random process
- The values of the random variable are called realizations
  - For example, flipping a coin is a random variable X with values 0 (in case of heads) and 1 (in case of tails)
  - The color of a ball we draw from a pot with blue, red and green balls is a random variable Y with values 0,1 and 2, respectively
  - The outcome of throwing a dice is a random variable Z with values {1,2,3,4,5,6}
- Notation:
  - Random variables – X,Y,Z
  - Realizations – x_i, y_i, z_i

JG|U

Random variables

Discrete random variables

- Random variables with a finite, countable set of possible values are called **discrete random variables**
  - The number of heads after flipping a coin n times
  - The number of children in a family
  - The gender of a randomly selected student from the class
  - The field of studies from a randomly selected student on the campus at a given time point

Random variables

Probability (mass) function

- Each outcome of a random experiment is associated with a specific probability, therefore each value of a random variable has a corresponding probability.
- Let X be a discrete random variable. Then the function

$$f(x_i) = P(X = x_i)$$

is called the **probability (mass) function** of X

- Following the rules for probabilities:
  - $0 \leq f(x_i) \leq 1$
  - $\sum_{i=1}^{n} f(x_i) = 1$

Random variables

Cumulative distribution function

- Let X be a discrete random variable. Then the function

$$F(x_i) = P(X \leq x_i)$$

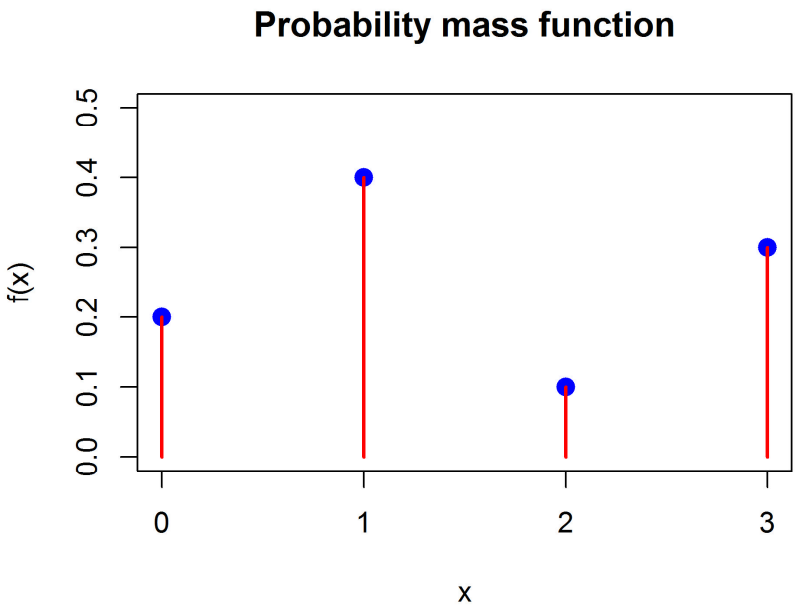
  is called the **cumulative disctribution function** of X.

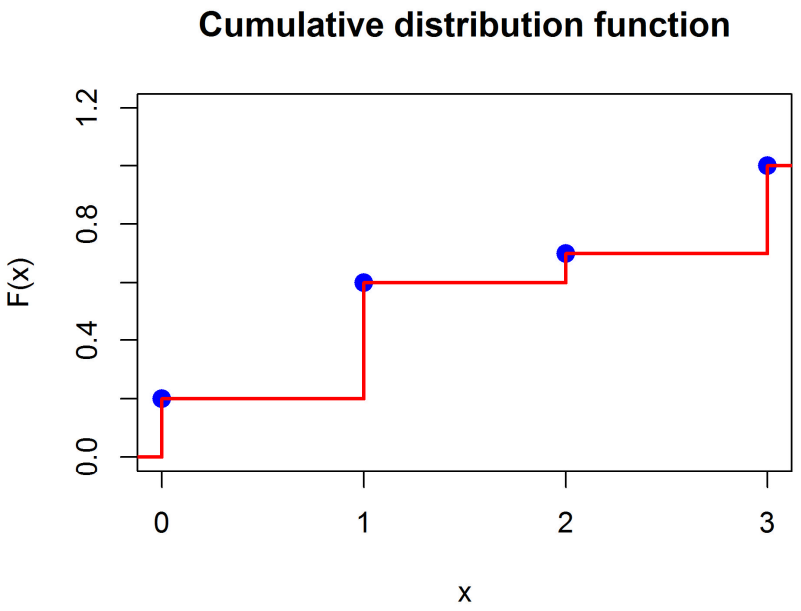
- Properties of F(x):
  - $0 \leq F(x) \leq 1$
  - A monotonically increasing function:
    - $F(x_i) < F(x_j)$ if $x_i < x_j$
  - $P(a < X \leq b) = F(b) - F(a)$

Random variables

Graphical representation of f(x) and F(x)

X	0	1	2	3
f(x)	0.2	0.4	0.1	0.3





Probability theory

Expected value of a discrete random variable

- Let **X** be a discrete random variable with a **probability function f(x).** The expected value E(X) is:

$$E(X) = \mu = \sum_{i=1}^{n} x_i f(x_i)$$

- E(X) is a sum of the values of X weighted by their probability
- E(X) is the value we would expect X to take if we repeat the experiment numerous times.

Probability theory

Expected value of a discrete random variable

X	0	1	2	3
f(x)	0.2	0.4	0.1	0.3

$$E(X) = \mu = \sum_{i=1}^{n} x_i f(x_i)$$

$$= 0 * 0.2 + 1 * 0.4 + 2 * 0.1 + 3 * 0.3$$

$$= 1.5$$

Probability theory

Continuous random variables

- A random variable which can take an infinite number of values in an interval [a; b] is called a **continuous random variable**
  - Height and weight of a sample of citizens
  - Blood pressure of patients in a study to test anti-hypertension therapy
  - The time needed to solve an exercise during an exam
  - The daily revenue of a supermarket

Probability theory

Probability density function

- Since there are infinitely many possible values for a continuous random variable, the probability that X takes exactly a specific value is 0

$$P(X = x) = 0$$

- Continuous random variables are characterized by their (**probability) density function f(x).**
- **f(x)** is used to specify the probability of X falling in a particular range of values

Probability theory

Cumulative distribution function

- Let X be a continuous random variable. The function

$$F(x) = P(X \leq x)$$

  is called **cumulative distribution function.**

- F(x) is the area under the density function curve

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

- Following the rules of probability, the total area under the density curve must be 1

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- $P(X \geq a) = 1 - P(X \leq a) = 1 - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a)$

Probability theory

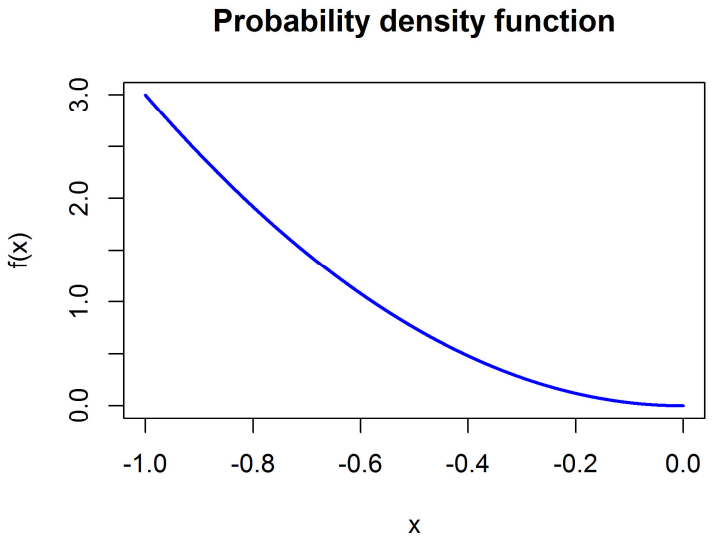
Graphical representation of f(x) and F(x)

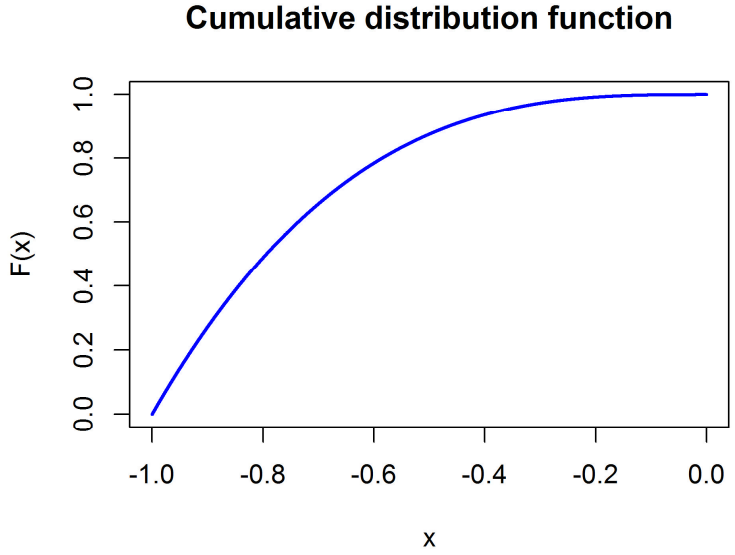
- Let X be a continuous random variable taking values from -1 to 0. The density function is given by:

$$f(x) = \begin{cases} 3x^2, x \in [-1; 0] \\ 0, otherwise \end{cases}$$

- The cumulative distribution function is therefore:

$$F(x) = \int_{-1}^{x} 3x^2 dx = x^3 + 1$$





Probability theory

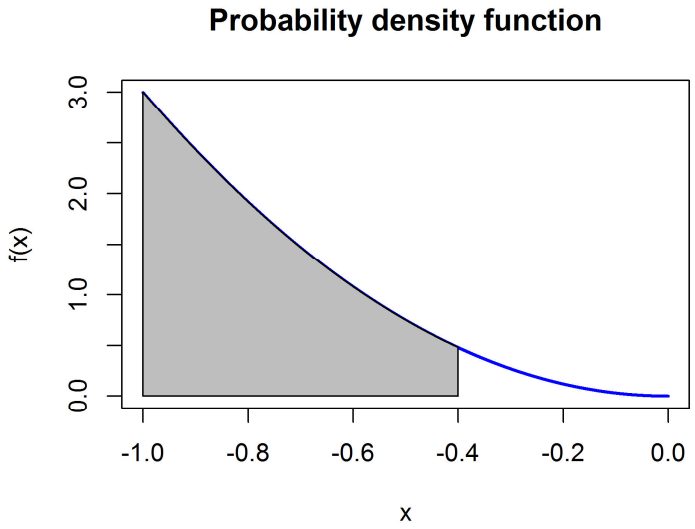
Graphical representation of f(x) and F(x)

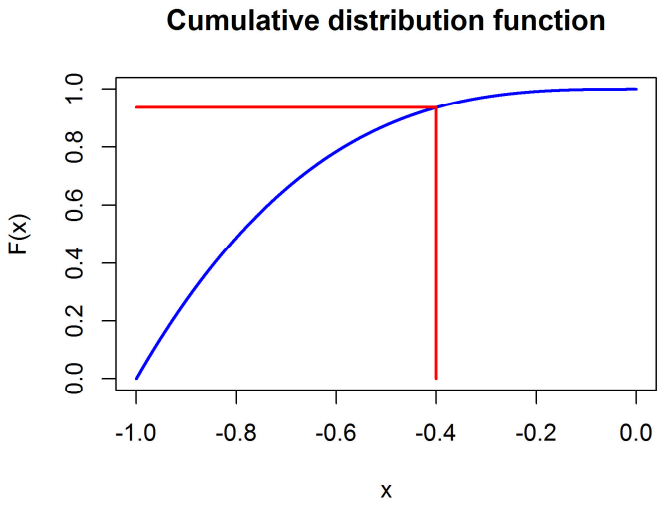
- Let X be a continuous random variable taking values from -1 to 0. The density function is given by:

$$f(x) = \begin{cases} 3x^2, x \in [-1; 0] \\ 0, otherwise \end{cases}$$

- What is the probability of X < -0.4?

$$P(X < -0.4) = \int_{-1}^{-0.4} 3x^2 dx = F(-0.4) = 0.934$$





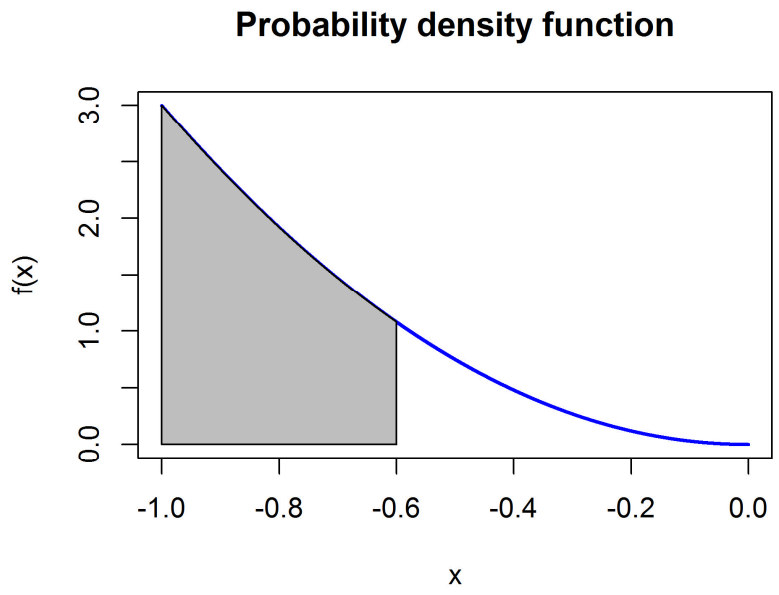
Probability theory

Graphical representation of f(x) and F(x)

- Let X be a continuous random variable taking values from -1 to 0. The density function is given by:

$$f(x) = \begin{cases} 3x^2, x \in [-1; 0] \\ 0, otherwise \end{cases}$$

- What is the probability that X ≥ 0.6
- P(X ≥ -0.6) = P(X > -0.6) = 1 – P(X < -0.6) = 1 – F(-0.6)



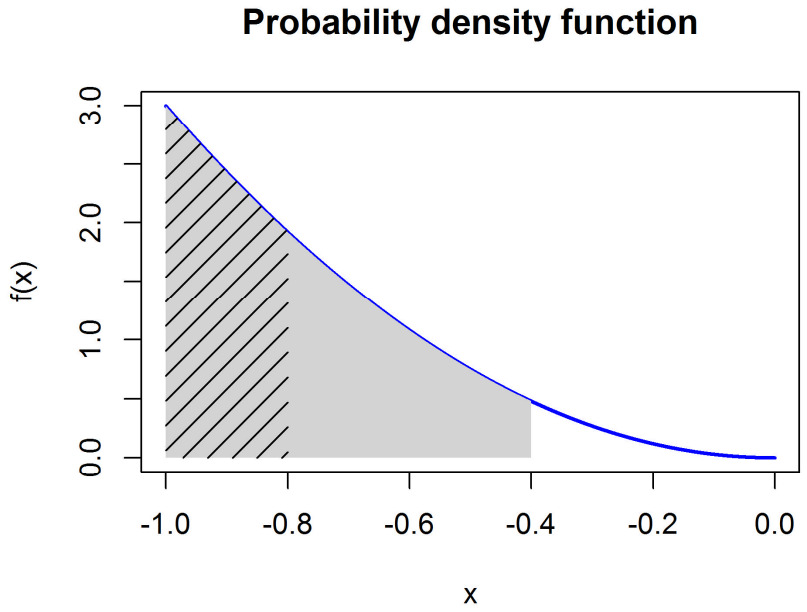
Probability theory

Graphical representation of f(x) and F(x)

- Let X be a continuous random variable taking values from -1 to 0. The density function is given by:

$$f(x) = \begin{cases} 3x^2, x \in [-1; 0] \\ 0, otherwise \end{cases}$$

- What is the probability that X lies between -0.8 and -0.4?
- P(-0.8 < X < -0.4) = F(-0.4) – F(-0.8)



Probability theory

Expected value of a continuous random variable

- Remember, E(X) for a discrete random variable is

$$E(X) = \mu = \sum_{i=1}^{n} x_i f(x_i)$$

- Since a continuous random variable can take infinitely many values, we cannot build a weighted sum of $xf(x)$ to calculate E(X)
- Instead, the expected value is the total area under the function $xf(x)$:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

Probability theory

Expected value of a continuous random variable

- Let X be a continuous random variable taking values from -1 to 0. The density function is given by:

$$f(x) = \begin{cases} 3x^2, x \in [-1; 0] \\ 0, otherwise \end{cases}$$

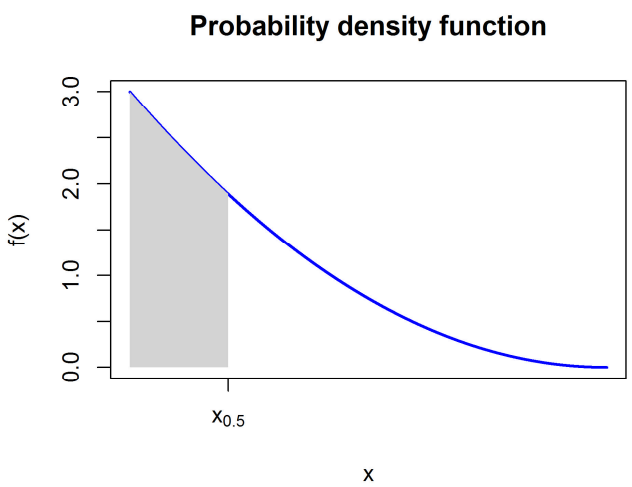
- The expected value is therefore:

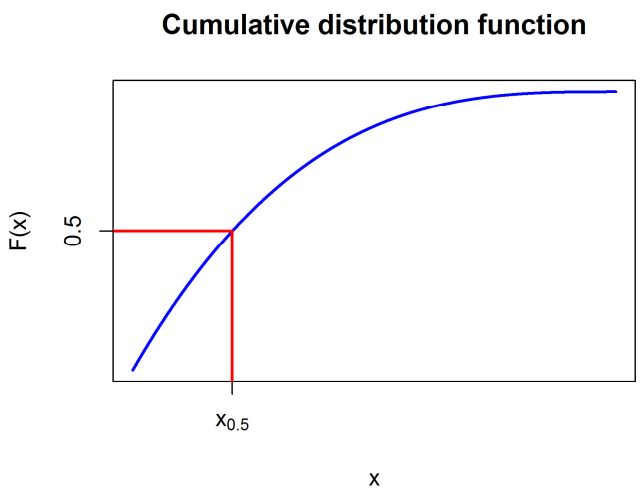
$$\int_{-1}^{0} x\, 3x^2 dx = -\frac{3}{4}$$

Probability theory

Quantiles

- Quantiles are values of a random variable corresponding to specific values of the cumulative probability function
  - $q_{0.1} = a \rightarrow F(a) = 0.1$
  - $q_{0.3} = b \rightarrow F(b) = 0.3$
  - $q_{0.5} = c \rightarrow F(c) = 0.5$
    - ➢ Values which are not exceeded with a specific probability or values corresponding to specific areas under the density function curve





Normal distrubution

Motivation for the normal distribution

- Many **empirical distributions** can be approximated by a normal distribution (e.g. height of adults, duration of pregnancy)
- Represents the **limit distribution** of many other probability distributions (e.g. binomial, poisson, t-distribution)
- **Sampling distributions** asymptotically approximate the normal distribution for large sample sizes
- The normal distribtution provides the theoretical basis for numerous models in statistics

Normal distribution

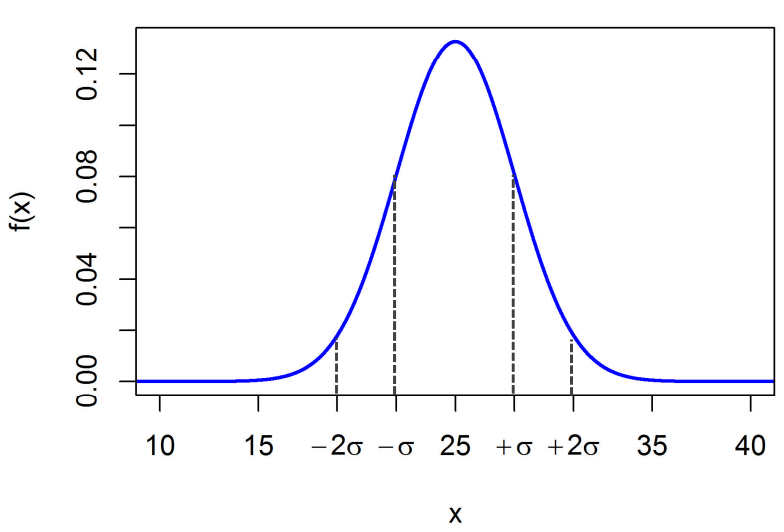
Features of the normal distribution

- Completely defined by its two paramers expected value μ and variance σ^2:

$$X \sim N(\mu; \sigma^2)$$

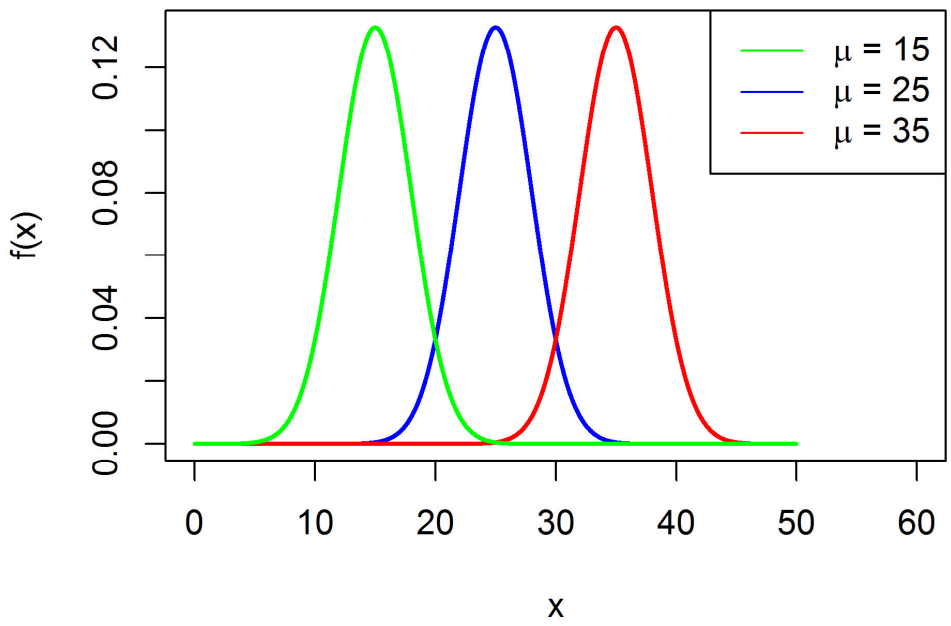
- Density function is unimodal with characteristic „bell“ shape
- Density is maximal at $X = \mu$ and symmetrical around μ with inflection points $+\sigma$ and $-\sigma$

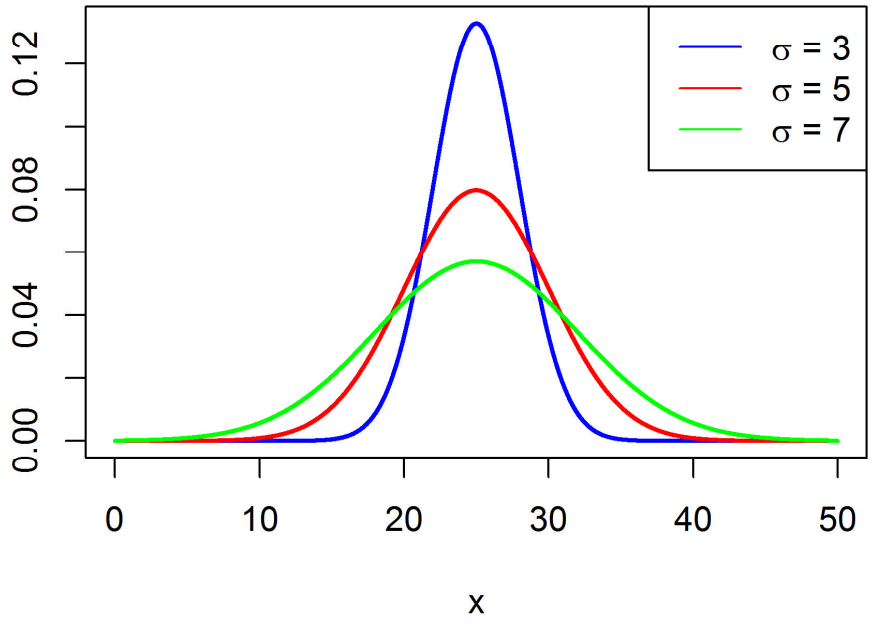
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$



Normal distribution

Normal distributions for different parameters





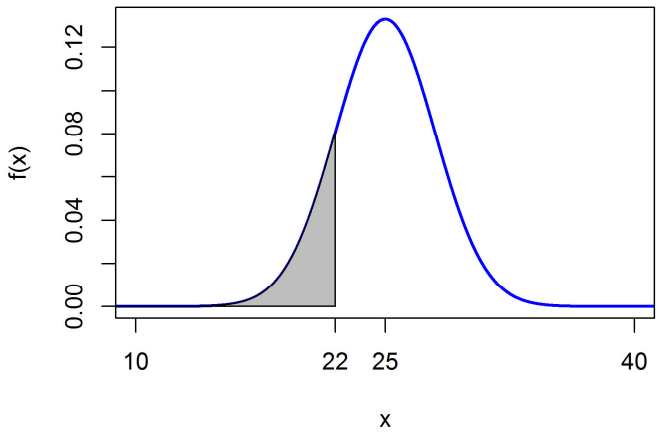
Normal distribution

Cumulative distribution function

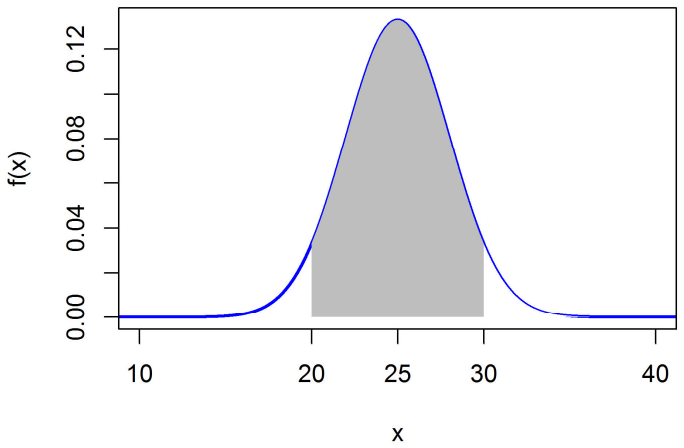
- F(a) = P(X ≤ a) – corresponds to the area under the density function curve
- Total area under the curve = 1, therefore P(X ≥ a) = 1 – P(X ≤ a)
- P(a ≤ X ≤ b) = F(b) – F(a)

$X \sim N(\mu = 25; \sigma = 3)$

$P(X \leq 22) = F(22) = 0.159$



$P(20 \leq X \leq 30) = F(30) - F(20) = 0.904$



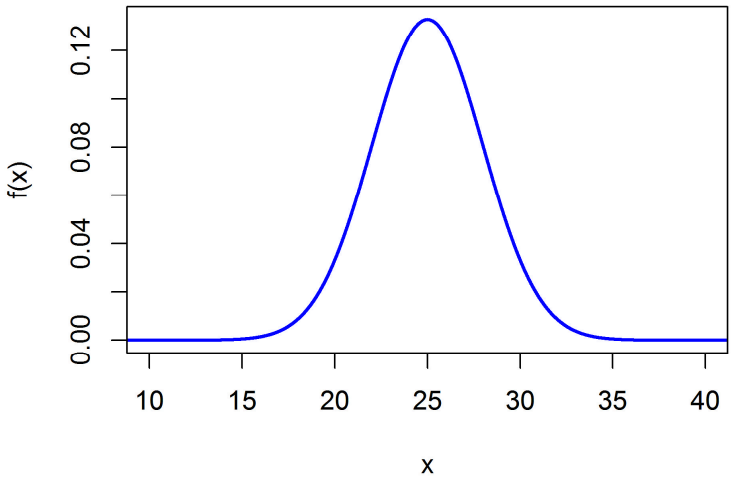
Normal distribution

Standard normal distribution

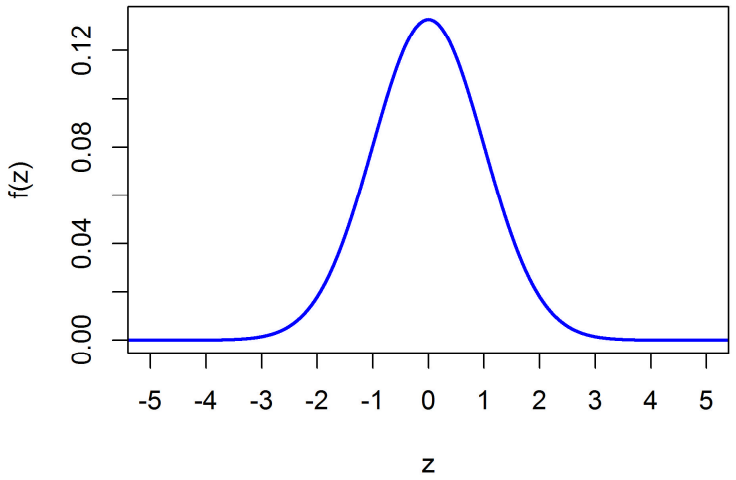
- Expected value E(Z) = μ = 0
- Variance Var(X) = $\sigma^2 = 1$
- Every normal distribution can be transformed in the standard normal distribution using the following transformation

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Normal distribution



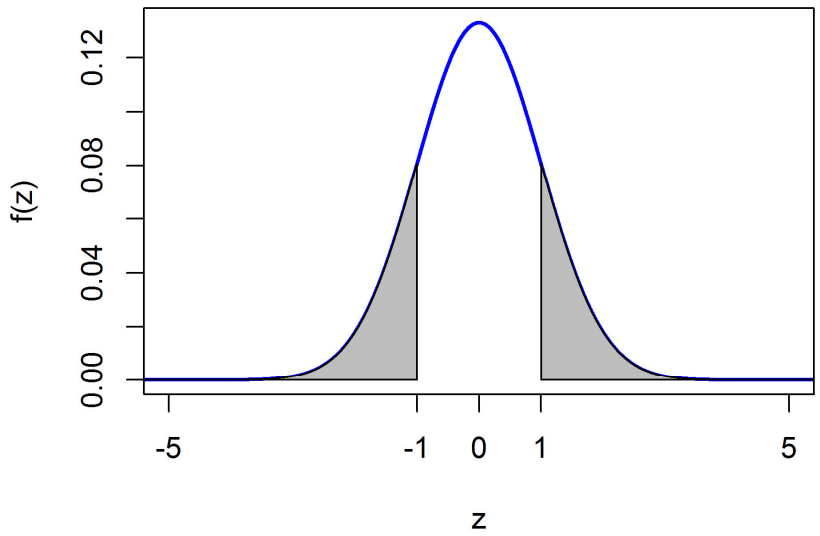
Standard normal distribution



Normal distribution

Cumulative distribution function

- $F(a) = P(X \leq a) = P(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma})$
- $\Phi(z)$ is the cumulative distribution function of the standard normal distribution.
- Values for $\Phi(z)$ are available in statistical tables
- Due to the symmetry of the distribution $\Phi(-z) = 1 - \Phi(z)$



$\Phi(1) = 1 - \Phi(-1)$
$= 1 - 0.159 = 0.841$

Normal distribution

Quantiles of the standard normal distribution

- Quantiles of the standard normal distribution with the corresponding cumulative distribution function values are summarized in statistical tables
- Since there are infinitely many normal distributions, they are usually transformed into a standard normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

Inferential statistics and hypothesis testing

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

JG|U

Inferential statistics

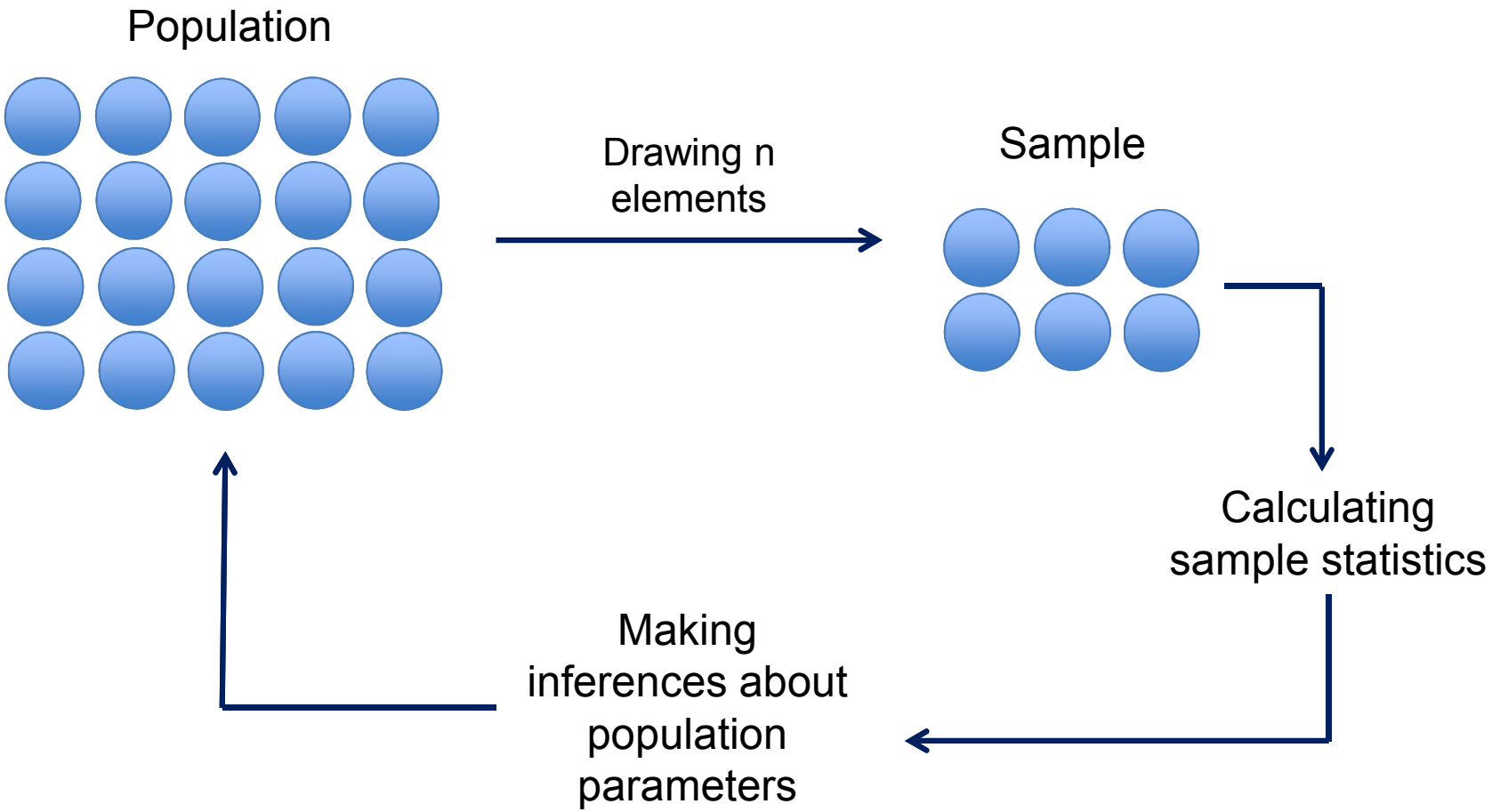
Population vs. sample

- **Statistical population –** a set of similar elements of interest in a reasearch question or experiment
  - *E.g. All patients undergoing cardiac surgery over the age of 65*
  - Usually it is not possible to investigate a whole population due to time or cost reasons

- **Statistical sample –** a subset of a statistical population chosen by following a specific stragegy or criterion
  - *E.g. All patients undergoing heart valve surgery at the University Medicine Mainz in the period between January 2017 and December 2017*

Inferential statistics

Population vs. sample



Inferential statistics

Population parameters and sample statistics

- Population parameters are unknown because the entire population is not available
- The measures which characterize samples are called statistics – directly calculated from the data
- Sample statistics are used to infer population parameters

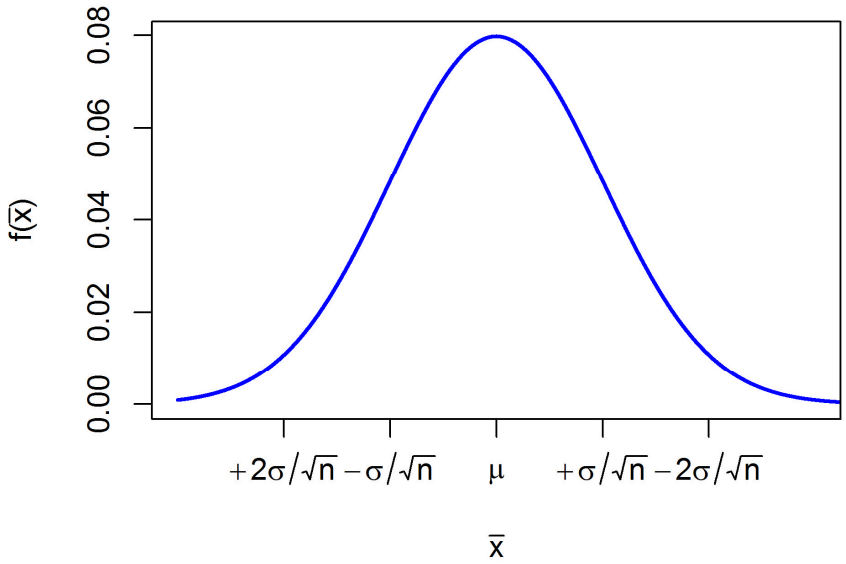
Population parameter	Sample statistic
Expected value μ	Sample mean $\bar{x}$
Variance σ^2	Sample variance s^2
Standard deviation σ	Sample standard deviation s
Correlation ρ	Correlation coefficient r

Inferential statistics

Sampling distributions

- Sample statistics are random variables because they are based on random samples
- Therefore, sample statistics have their own probability distributions called **sampling distributions**
  - *E.g. Let the height of Germans in cm be a normally distributed random variable **X** with mean **μ** and variance **σ²**. Let $\bar{X}$ be the mean height in a sample of size **n**.*
  - $\bar{X}$ is a random variable with the following distribution:

$$\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$$



Inferential statistics

Confidence intervals

- The sample mean, variance, standard deviation, etc. are so called point estimates for the respective population parameters because they provide only a single value
- The probability that the point estimate is equal to the true parameter is equal to 0 for continuous random variables
- Interval estimates provide a range of values which could contain the population parameter and are therefore more informative
- Confidence intervals are calculated from the sample data at pre-defined confidence level ***1 - α***

Inferential statistics

Confidence intervals

- Confidence interval for normally distributed parameters, e.g. ***μ***
  - Let C_u and C_o be the lower and upper limits of a 1-α confidence interval for μ

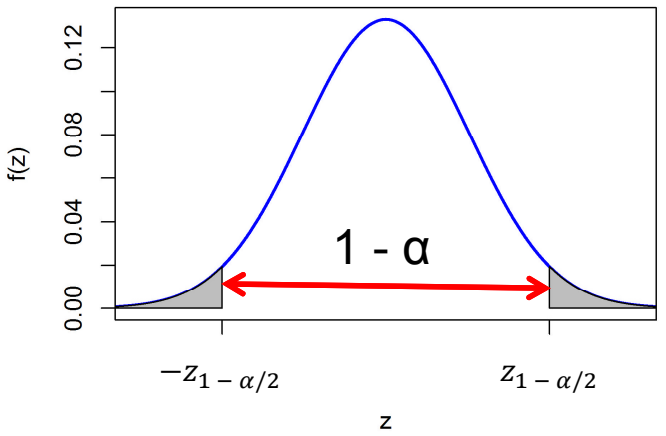
$$1-\alpha = P(C_u \leq \bar{X} \leq C_o)$$

$$1-\alpha = P(\frac{C_u-\mu}{\sigma}\sqrt{n} \leq \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \leq \frac{C_o-\mu}{\sigma}\sqrt{n})$$

$$1-\alpha = P(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \leq z_{1-\alpha/2})$$

$$1-\alpha = P(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Typical values for α include 0.05, 0.01, 0.001



Hypothesis test

Confidence intervals

Calculate the 95 % confidence interval for the average total bilirubin concentration of 2.7 mg/dL in a sample of 36 patients with moderate liver injury with a standard deviation of 0.4

$$1 - \alpha = 95\ \% = 0.95\ \rightarrow\ \alpha = 0.05$$

$$\pm z_{1-\alpha/2} = \pm z_{1-\frac{0.05}{2}} = \pm z_{0.975} = \pm 1.96$$

$$95\%\ CI = \bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$95\%\ CI = 2.7 \pm 1.96\frac{0.4}{\sqrt{36}}$$

$$95\%\ CI = [2.57; 2.83]$$

- Interpretation of CI
  - If confidence intervals were repeatedly calculated for many samples, then 1 - α of the intervals would countain the true population parameter

Inferential statistics

Hypothesis tests

- Researchers are often interested in yes/no questions regarding population parameters
- E.g. A new medicine to prevent cold was tested against a control treatment. In a small pilot experiment, 20 subjects were randomized to the control and 20 to the new treatment. 11 subjects in the control group got a cold compared to 7 subjects in the treatment group. *Does the new treatment prevent getting a cold?*

	Control	**Treatment**
Got a cold	11	7
Remained healthy	9	13

Inferential statistics

Hypothesis tests

- Results based on sample statistics initially only apply to the investigated sample
- Hypothesis tests are needed to decide if the observed results are due to population differences or simply the result of chance.
- The basis of hypothesis test is establishing the **hypothesis pair:**
  - **Null hypothesis:** Represents the status quo, absence of a difference – usually the hypothesis we want to disprove

    H_0: The new therapy does not reduce the risk of getting a cold.
  - **Alternative hypothesis:** Represents the alternative to H_0 – usually represents the desired outcome

    H_1: The new therapy prevents getting a cold.

Inferential statistics

Hypothesis tests

- The null and alternative hypothesis make assumptions about an **unknown** population parameter **Θ** (e.g. population mean **μ** or population proportion **p**)
- Types of hypothesis pairs:
  - H_0: $\Theta = \Theta_0$ vs. H_1: $\Theta \neq \Theta_0$: **two-tailed** test
    - *E.g. Expression levels of gene X under condition A are different compared to condition B*
  - H_0: $\Theta = \Theta_0$ vs. H_1: $\Theta < \Theta_0$: **left-tailed** test
    - *E.g. Therapy A reduces blood pressure compared to therapy B*
  - H_0: $\Theta = \Theta_0$ vs. H_1: $\Theta > \Theta_0$: **right-tailed** test
    - *E.g. Therapy A increases survival time in cancer patients compared to control.*

Inferential statistics

Hypothesis tests

- At the end of a specific test we always retain one of the hypotheses and reject the other. However, we never know if this decision is correct.
- What we do instead is try to reduce the probability of a false decision
- The following possibilities exist:

	Decision for a test	
Reality	Reject H_0	Reject H_1
H_0 is true	Type I error (α)	Correct decision
H_1 is true	Correct decision	Type II error (β)

Inferential statistics

Hypothesis tests

- Researchers typically control the type I error probability, each statistical test is performed at a predefined **significance level α.**
- Typical values include α = 0.05, α = 0.01, α = 0.001
- In order to decide which hypothesis to retain, a **test statistic T** is calculated.
- **T** is a function of the sample statistic, e.g. the sample mean $\bar{X}$
  - Remember that hypothesis are defined for the unknown population parameters
- **T** is therefore a random variable with a probability distribution
- In order to decide which hypothesis to retain/reject, the distribution of **T** under the **null hypothesis** is evaluated.
- The **null hypothesis** is rejected if **T** exceeds critical values which depend on **α.**

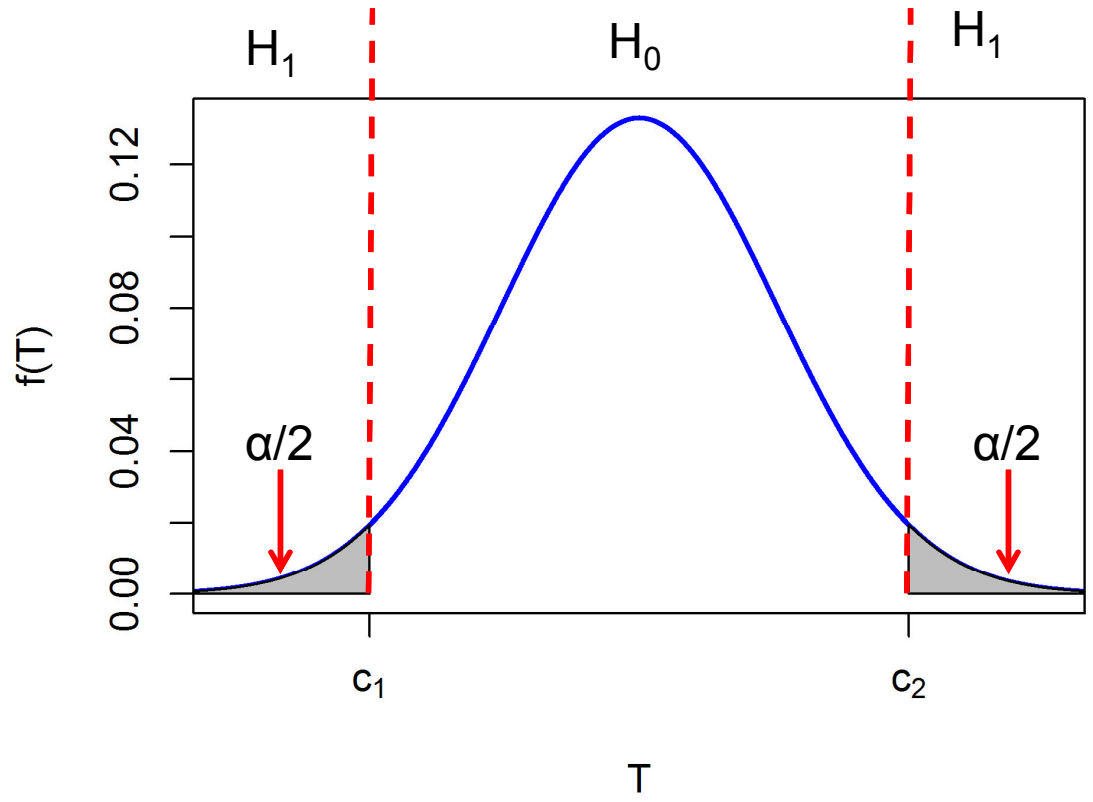
JG|U

Inferential statistics

Hypothesis tests

Distribution of the test statistic **T,** two-tailed test

- Reject H_0 if $T < c_1$ or if $T > c_2$
- c_1 and c_2 are the so called critical values

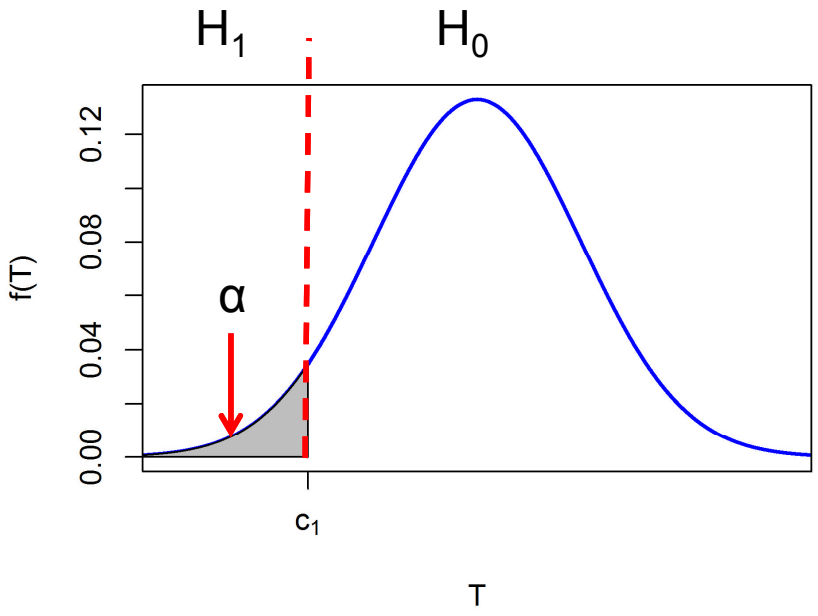


Inferential statistics

Hypothesis tests

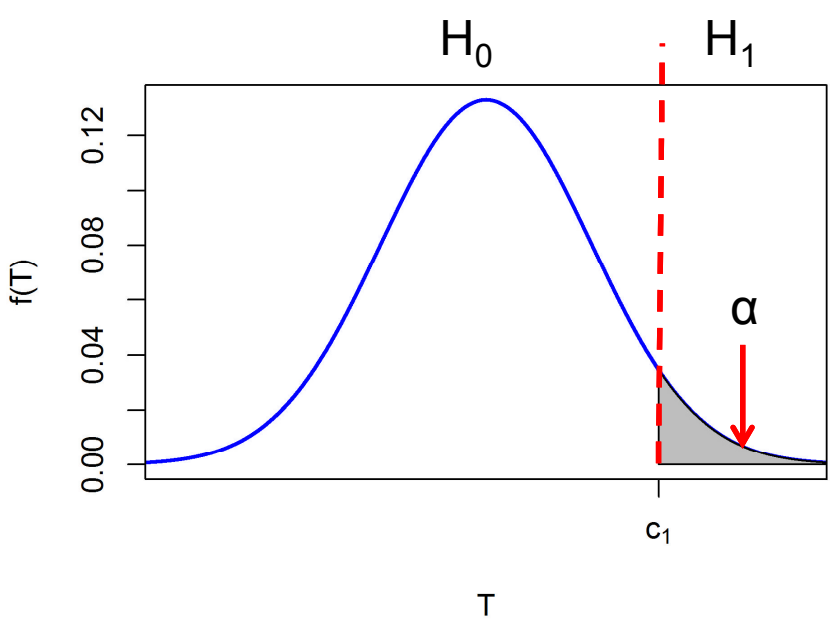
Distribution of the test statistic **T,** left- and right- tailed tests

Left-tailed test



Reject H_0 if $T < c_1$

Right-tailed test



Reject H_0 if $T > c_1$

Hypothesis tests

Steps in hypothesis testing

1. Evaluation of test assumptions
2. Definition of the null and alternative hypothesis
3. Definition of significance level
4. Calculation of the test statistic
5. Definition of the rejection region based on the probability distribution of the test statistic, critical values specification
6. Comparison of the test statistic with critical values
7. Decision to retain/reject null hypothesis and interpretation

Hypothesis tests

One sample z test

- The one sample z-test investigates if a population mean μ significantly deviates from a given value μ_0 when the population variance is known.

E.g. The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. Does the result for this year's exam represent a systematic change or only random fluctuation?

→ Investigate hypothesis with the help of the one sample z-test using the 7-point scheme outlined on the previous slide

Hypothesis tests

One sample z test example

The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. Does the result for this year's exam represent a systematic change or only random fluctuation?

1. Evaluation of test assumptions
   - Normally distributed random variable and known population variance
2. Definition of the null and alternative hypothesis
   - Null hypothesis: The average note of biomedicine students in statistics is 2.2

     $H_0: \mu = 2.2$
   - Alternative: The average note of biomedicine students in statistics is not 2.2.

     $H_1: \mu \neq 2.2$ ⟶ two-sided test

Hypothesis tests

One sample z test example

The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. Does the result for this year's exam represent a systematic change or only random fluctuation?

3. Definition of significance level
   - α = 0.05 (type I error = 5 %)
4. Calculation of the test statistic

$$T = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$$

$\bar{X}$ = 1.8, μ_0 = 2.2, σ = $\sqrt{0.9}$, n = 49

$$T = \frac{1.8 - 2.2}{\sqrt{0.9}}\sqrt{49} = -2.951$$

Hypothesis tests

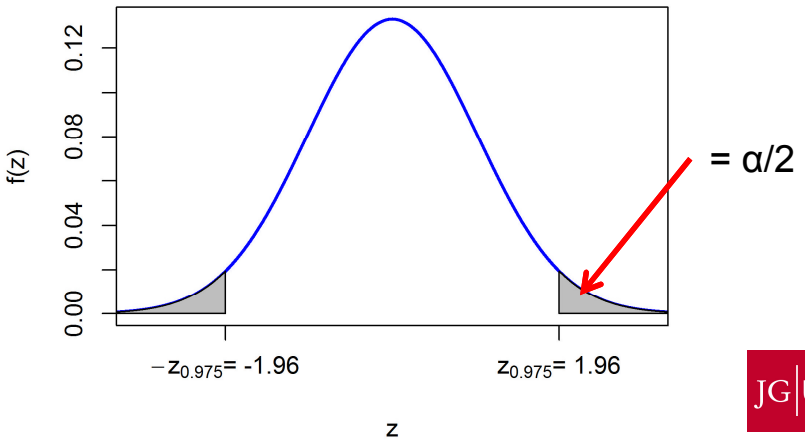
One sample z test example

The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. Does the result for this year's exam represent a systematic change or only random fluctuation?

5. Definition of the rejection region based on the probability distribution of the test statistic, critical values specification

- When assumptions of the test are met, the test statistic follows a standard normal distribution, $T \sim \mathcal{N}(0,1)$.
- Critical values for a two sided test:

$$\pm z_{1-\alpha/2} = \pm z_{1-0.025} = \pm z_{0.975} = \pm 1.96$$



Hypothesis tests

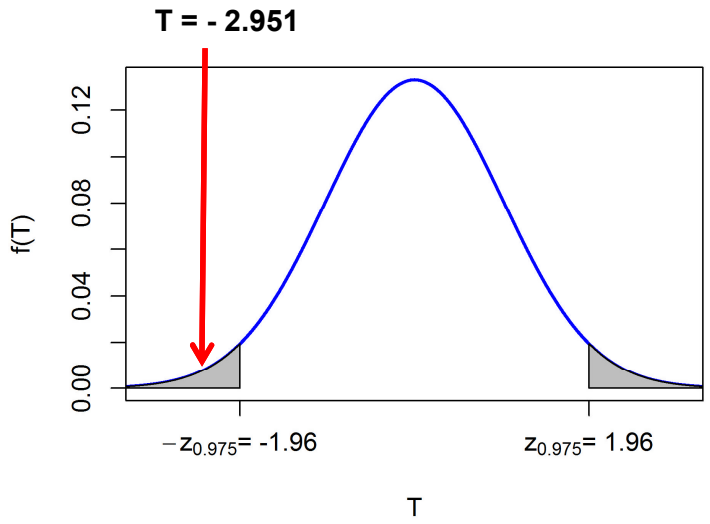
One sample z test example

The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. Does the result for this year's exam represent a systematic change or only random fluctuation?

6. Comparison of the test statistic with critical values

$T = -2.951 < -z_{0.975} = -1.96$

- Test statistic exceeds the critical value



Hypothesis tests

One sample z test example

The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. Does the result for this year's exam represent a systematic change or only random fluctuation?

7. Decision to retain/reject null hypothesis and interpretation
   - Null hypothesis is rejected
   - Interpretation:

     The average note of biomedicine students in statistics this year is significantly different from the 5 year-average.

Hypothesis tests

Power of a test

- Usually, the desired outcome of a hypothesis test is to find sufficient empirical evidence to reject the null hypothesis.
- Even if H_1 is true, we still might fail to reject H_0 (Type II error: false negative)
- Type II errors can have serious consequences especially in the field of clinical research
- Power = 1 – β, probability to reject a false H_0
- Ways to increase power of a test:
  - Increase the significance level
  - Increase sample size

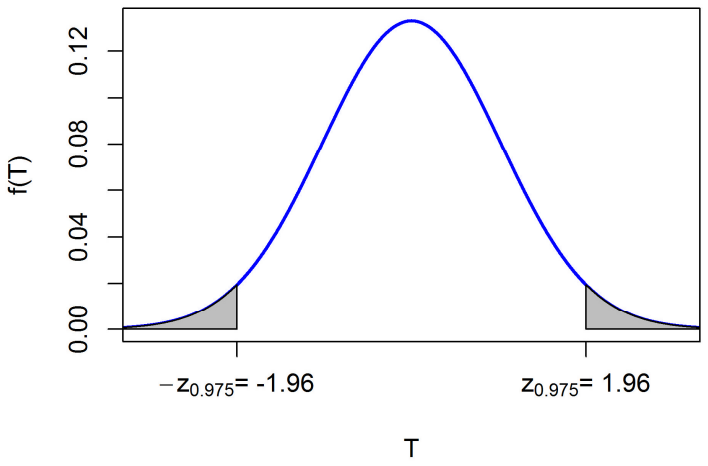
Hypothesis tests

Power of a test

- Power can only be calculated for specific values of the parameters under the alternative hypothesis.
- *The average note in the final biostatistics exam of biomedicine students at the University Mainz from the last 5 years was 2.2 with a known variance of 0.9. The average note of the 49 students who took the exam this year was 1.8. What is the power of a two-sided z-test given the alternative hypothesis that $\mu_1 = 1.9$?*

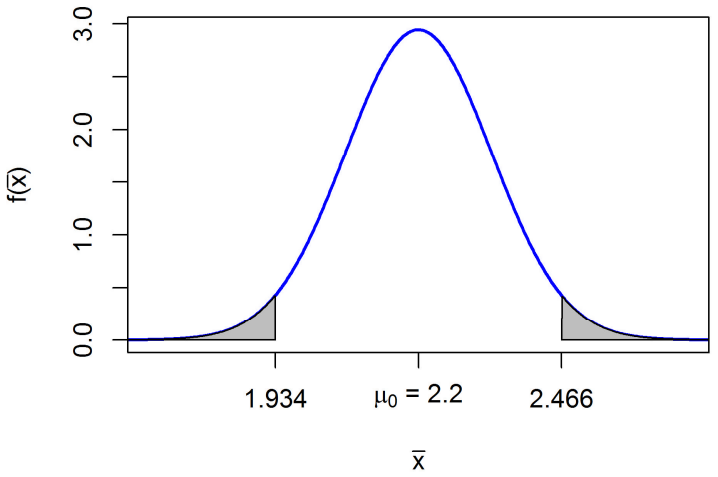
$$\text{Power} = P(T < -z_{0.975} \text{ or } T > z_{0.975} \mid \mu_1 = 1.9) = ?$$

Distribution of T under H_0



$$\bar{x}_{crit} = \pm \frac{z_{0.975}\sigma}{\sqrt{n}} + \mu_0$$

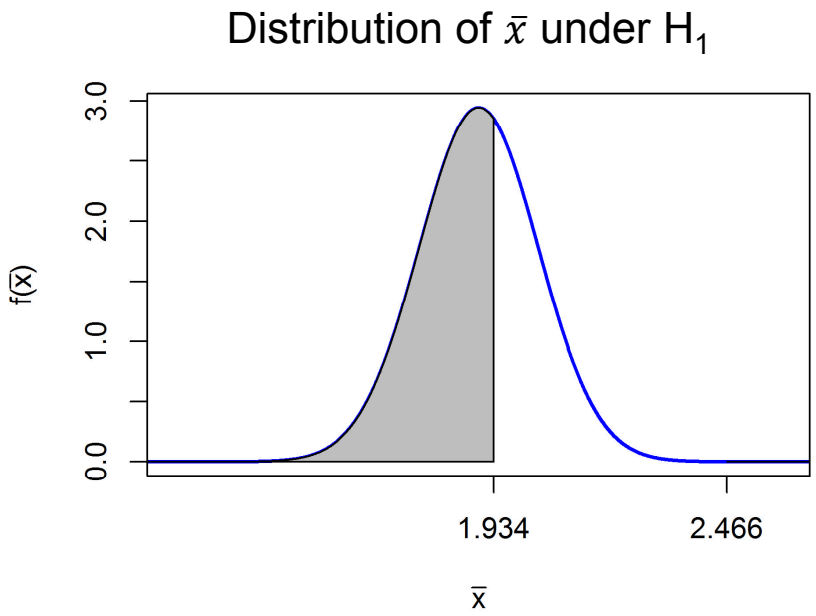
Distribution of $\bar{x}$ under H_0



Hypothesis tests

Power of a test

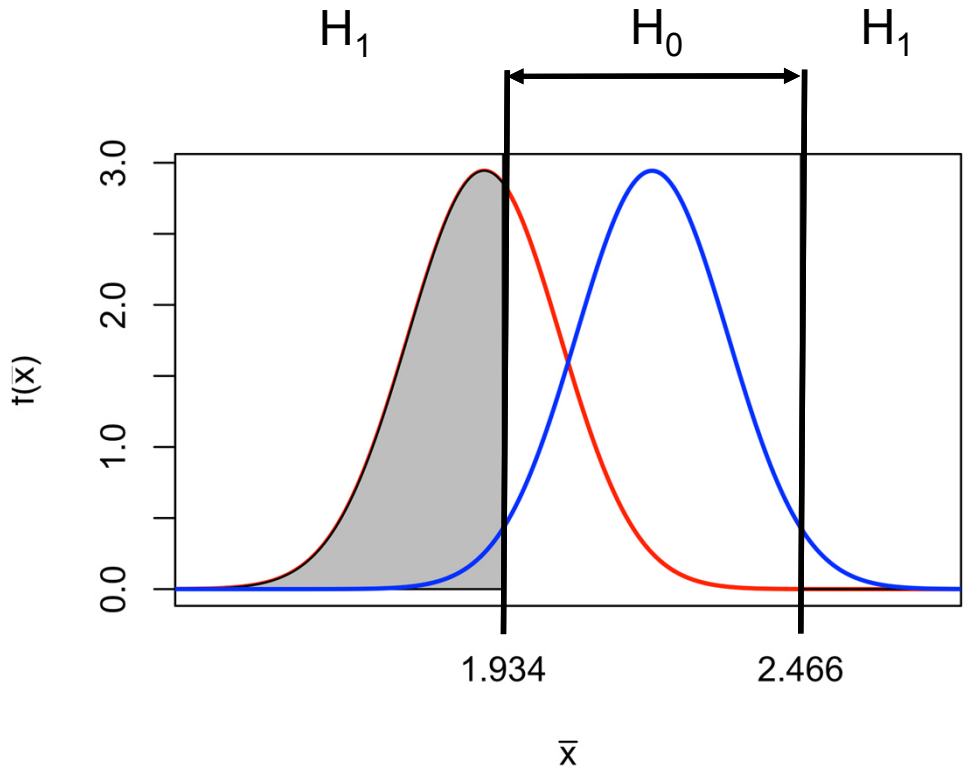
- Reject H_0 if $\bar{x} < 1.934$ or $\bar{x} > 2.466$
- Power = $P(\bar{x} < 1.934 \text{ or } \bar{x} > 2.466 \mid \mu_1 = 1.9) = 0.599$



Hypothesis tests

Power of a test

- Reject H_0 if $\bar{x} < 1.934$ or $\bar{x} > 2.466$
- Power = $P(\bar{x} < 1.934 \text{ or } \bar{x} > 2.466 \mid \mu_1 = 1.9) = 0.599$



Hypothesis tests

Effect size

- Rejecting a null hypothesis does not automatically imply that results have practical relevance
- In large samples, even trivial differences might be statistically significant
- This calls for standard measures to evaluate effect size, e.g.:
  - Standardized mean difference
  - Percentage of variance explained by the model
- Effect sizes should be reported with results from significance tests

Hypothesis tests

Effect size example

- Cohen's *d* is a measure of effect size evaluating standardized mean difference
  - $Cohen's\ d = \frac{|\mu_1 - \mu_2|}{\sigma}$
- Reference values:
  - d = 0.2 – small effect
  - d = 0.5 – medium effect
  - d > 0.8 – large effect

- *E.g. The average note of biomedicine students from this year's exam* $\bar{x} = 1.8$ *was shown to be significantly different from the 5-year average of 2.2. What is the effect size?*

$$d = \frac{|1.8 - 2.2|}{\sqrt{0.9}} = 0.422$$

Hypothesis tests

Power and effect size in controlled experiments

- Controlled clinical experiments are associated with significant costs and time
- Failing to detect a true alternative hypothesis could be detrimental
- From a statistical point of view, studies are planned to maximize the chances of proving a study hypothesis which is also of practical relevance:
  - Sample size is based on the level of power desired to be achieved, typically 80%
  - Effect size is based on theoretical knowledge or pilot investigations