

Marcel Gehrke

Data Understanding vs. Machine Training

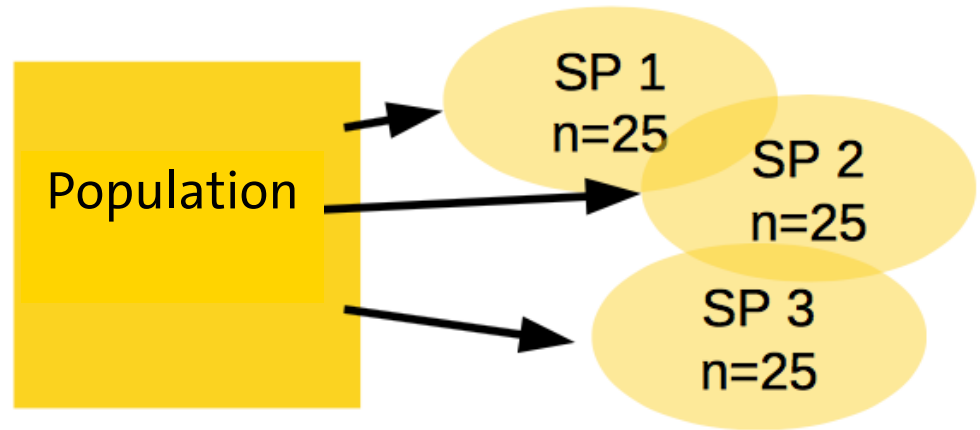
Statistical Basics

Statistical Basics

Definitions

Consideration of a Subset of the Data

- Data, also referred to as the population
- Subset of data, also known as **sample (SP)**



Consideration of a Subset of the Data

- Data, also referred to as population
- Subset of data, also known as sample (SP)

Definition	Population	Sample
		Subset of a population
Symbols	Greek	Latin
Mean	μ	\bar{x}
Standard deviation	σ	s $\hat{\sigma}$

Concept of Statistical Variable

A **statistical variable** assigns a value (**feature value**) to an attribute (**feature**) of a survey unit (**feature carrier**, object)

Examples

- Population: residents of the city of Hamburg
 - Feature carrier: a resident
 - Feature : gender
 - Feature value: male
- Population: days in a study period
 - Feature carrier: one day
 - Feature : Precipitation in Germany
 - Feature value: 1.5 cubic kilometres

Statistics

Descriptive statistics

- Describing data (also: subset of data)
 - Examples: mean, variance, ...
 - ... also referred to as "statistics"
- Searching for trends/patterns
 - Examples: Frequent item sets, association rules

Inductive statistics

- Objective: Generalisation of the description of a subset of the data to the population
- Conclusions about the population by collecting a "representative" sample

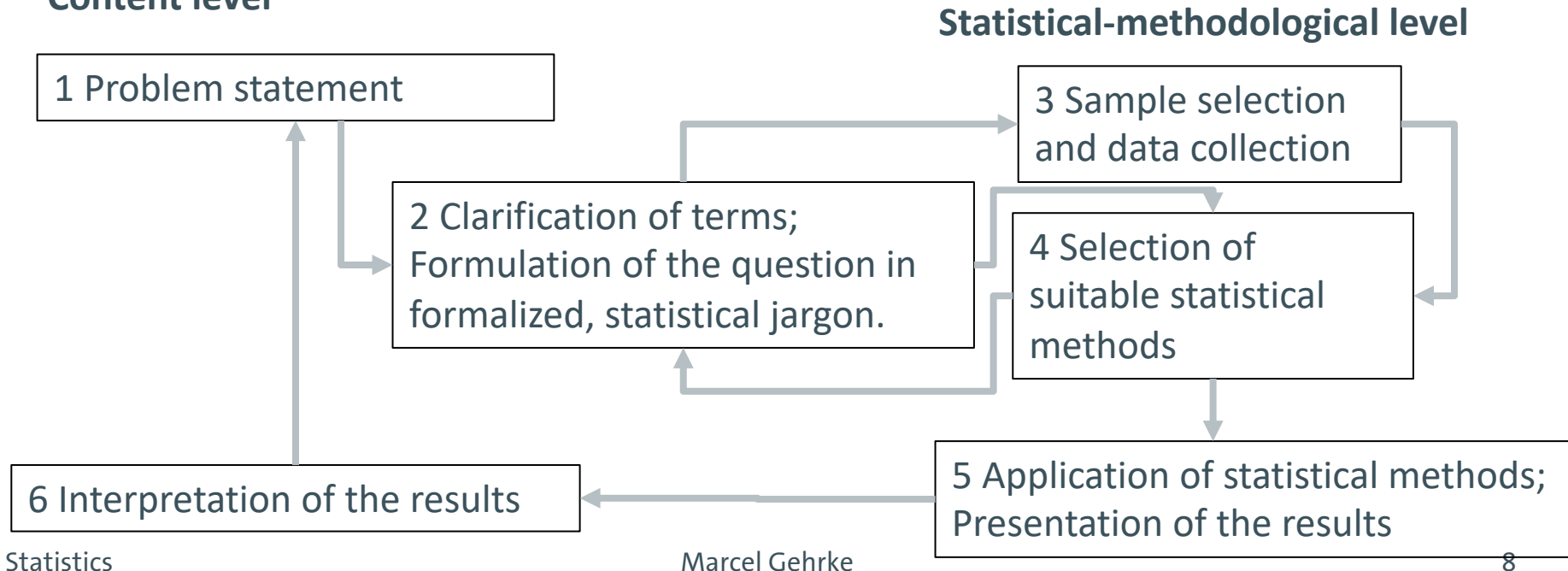
"Representative"

- Statements about a sample can be used to draw conclusions about the features of the population
- Take elements randomly from the population?
- The sample size should be "sufficient"
 - We will come back to this later

- For now: Not a formally defined concept; depending on the application, in many cases it is initially based on plausible arguments

Systematic Examination Procedure

Content level



Planning of Empirical Evaluation

Which **sampling unit** should be used?

- Which scaling/normalisation of the data?

Which **spatial** sampling pattern should be used?

- How should the area be divided for sampling?

Which **temporal** sampling pattern should be used?

- What are appropriate intervals?

Investigations/experiments are usually carried out to examine the influence of one or more factors on a variable.

Collection of Random Samples

Related samples

- E.g. repeated measurements on the same test object
- A sample taken at one point in time can influence a sample taken at another point in time

Unrelated samples

- Samples have no influence on each other
- E.g. different populations, comparison of different objects

Systematic Error/Trend (Bias)

- Occurring, usually disruptive systematic effect with a basic tendency, so that values deviate from the true events
- Examples
 - Estimating fish populations using nets with a specific mesh size: small fish can always escape
 - Catching mammals: some individuals are "trap happy", some are "trap shy"

Location Measurements - Mean Values

- Arithmetic mean

$$\bar{x}_{\text{arithm}} := \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- Geometric mean

The geometric mean of two numbers a and b gives the side length of a square that has the same area as the rectangle with side lengths a and b .

Relevant for among other things, logarithmic data. e.g. population growth.

$$\bar{x}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

- Harmonic mean

$$\bar{x}_{\text{harm}} = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

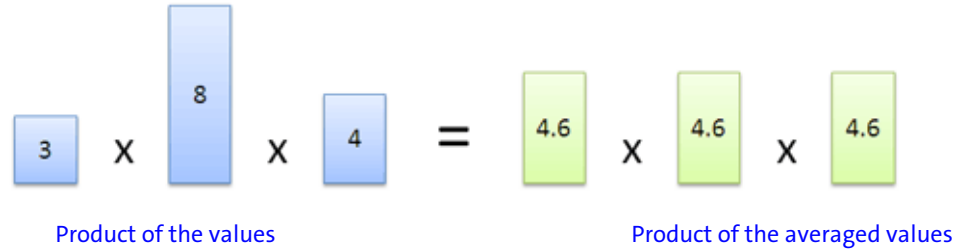
$$\frac{1}{\bar{x}_{\text{harm}}} = \frac{\frac{1}{x_1} + \dots + \frac{1}{x_n}}{n}$$

$$\log_a \bar{x}_{\text{geom}} = \frac{1}{n} \sum_{i=1}^n \log_a x_i,$$

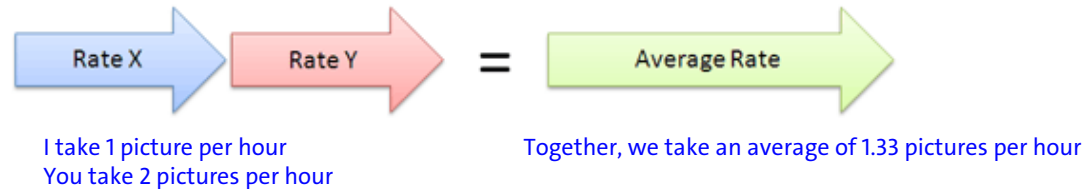
- $\min(x_1, \dots, x_n) \leq \bar{x}_{\text{harm}} \leq \bar{x}_{\text{geom}} \leq \bar{x}_{\text{arithm}} \leq \max(x_1, \dots, x_n)$.

Visualisation

Geometric Mean



Harmonic Mean



Further Location Measurements

Median (the value that is in the middle of a list of numerical values)

4, 1, 37, 2, 1 → Median = 2 (1, 1, 2, 4, 37)

Modal value

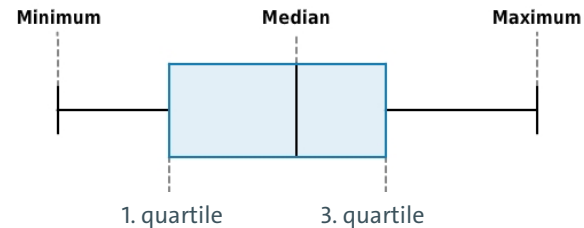
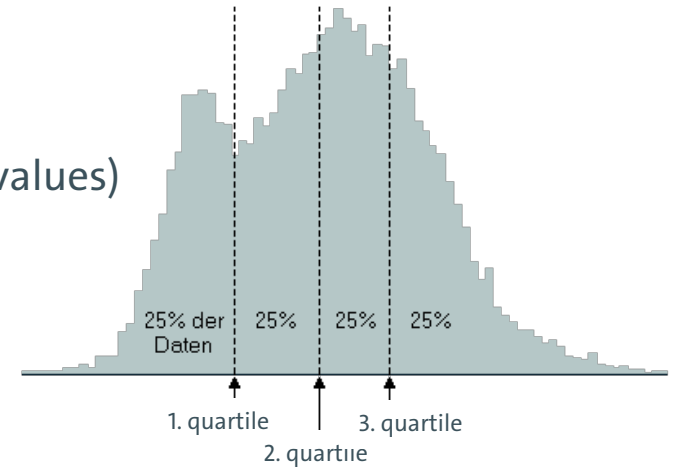
Most frequent value

2, 2, 3, 5, 5, 5, 9, 9, 15

Quantile, quartile

Ordered series of features is broken down into equal parts.

Cumulative frequencies



Data Types

Scale levels	Possible statements	Possible methods (e.g. position measure)	Beispiele
Nominal (no ordering of the data possible)	1. Equality and inequality (=, ≠) can be established	frequencies, relative frequencies, modal value	Favorite newspapers Sex Studies
Ordinal (order of magnitude possible, but distances without significance)	1. Equality and inequality 2. Ranking (<, >, =)	in addition: e.g. cumulative frequencies, median	Popularity Ranking Sequences
Interval (Spacing can be interpreted, but not the ratio of sizes)	1. Equality and Inequality 2. Ranking 3. Equality of differences	in addition: e.g. arithmetic mean	Date Temperature
Ratio (the features have an absolute zero point; the ratio can be interpreted)	1. Equality and inequality 2. Ranking 3. Equality of differences 4. Proportionality ($x_{11} = 3 * x_{12}$)	in addition: e.g. geometric mean	Age Price Size Nutritional value in calories Inflation

information content

Metric Variables

- Interval and ratio scales are often grouped together under the term **cardinal scale**.
- Features on this scale are then referred to as **metric**

Categorical Variables

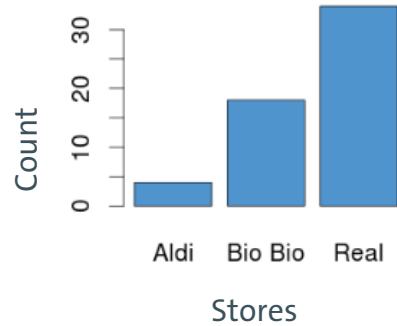
- Nominal scaled variables
- Ordinal scaled variables
- Variables created by categorising ordinal-scaled or metric variables (example: variable "income" with the categories "500-999 €", "1000-1499 €", etc.)

Scattering Measures

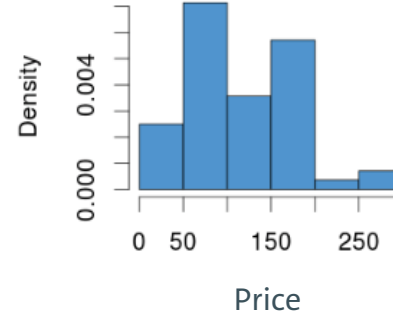
- **Range**
 - Maximum difference between underlying data
 - At least ordinal data required
- **Variance**
 - Mean square deviation of individual data values from the arithmetic mean
 - Units squared
- **Standard deviation**
 - The standard deviation is defined as the square root of the variance
 - The scattering measure has the same unit as the data and the mean value

Display of Data Properties

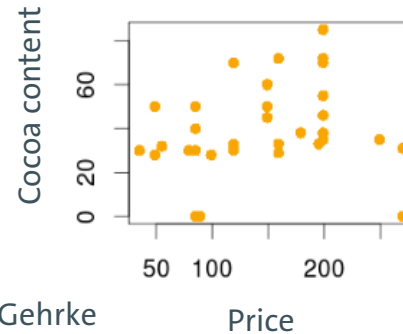
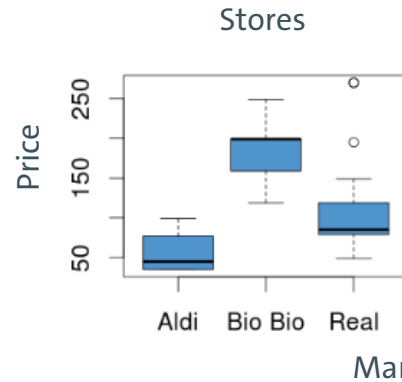
Column chart



Histogram



Box plot



Scatter plot

Presentation of Data

Barplot/Column Chart/Bar Chart

- Nominal and ordinal-scaled variables: Count

Histogram

- Ordinal-scaled or metric variables

Scatterplot

- For 2 variables
- Usually metric variables

Box plot

- Metric variables belonging to different categories

Relative Frequencies

- Histogram: Counter for number of occurrences
 - Frequency distribution
- Normalisation of numbers to $[0, 1]$ (scaling) results in **relative frequencies**
- Distribution usually considered in relation to relative frequencies

Statistical Basics

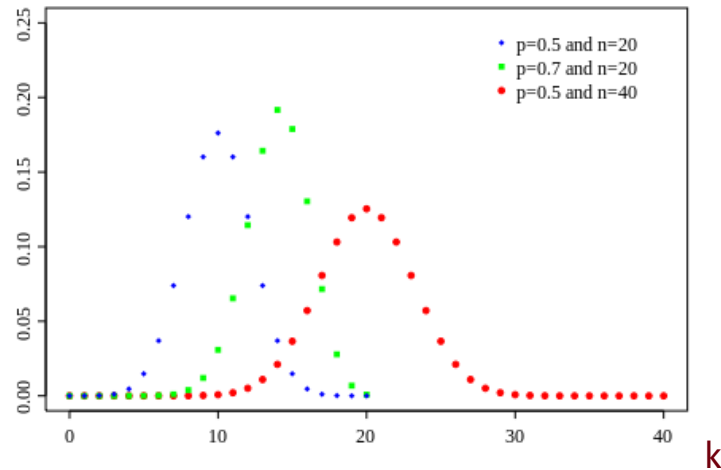
Distributions

Distributions

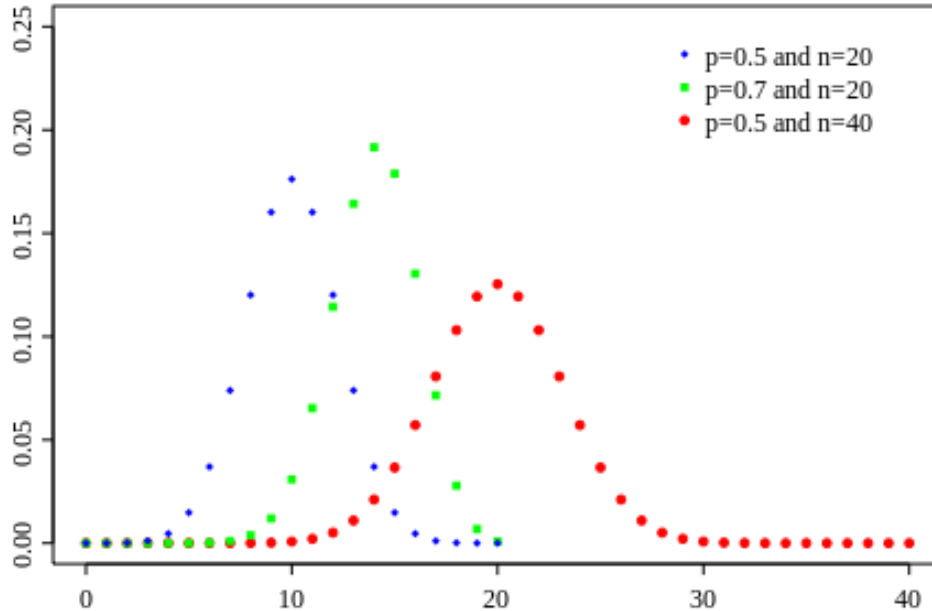
- Some distributions that occur naturally
 - Exponential distribution (we already covered this)
 - Cities (nominal) Number of inhabitants (metric)
 - Cities can be sorted by population
 - Binomial distribution
 - Normal distribution
- Description by function
 - $f: \text{population} \rightarrow [0, 1]$

Binomial Distribution

- Describes the number of successes in a series of similar and independent trials, each of which has exactly two possible outcomes: "success" or "failure"
 $B_{p,n}(k)$
- n = #trials
 p = #successful trials / #trials
- **Description**
of the relative frequency
of achieving **exactly** k successes
as a function $B_{p,n}(k)$



Task



How do we determine the relative frequency of achieving up to k successes?

$$\sum_{i=0}^k B_{p,n}(i)$$

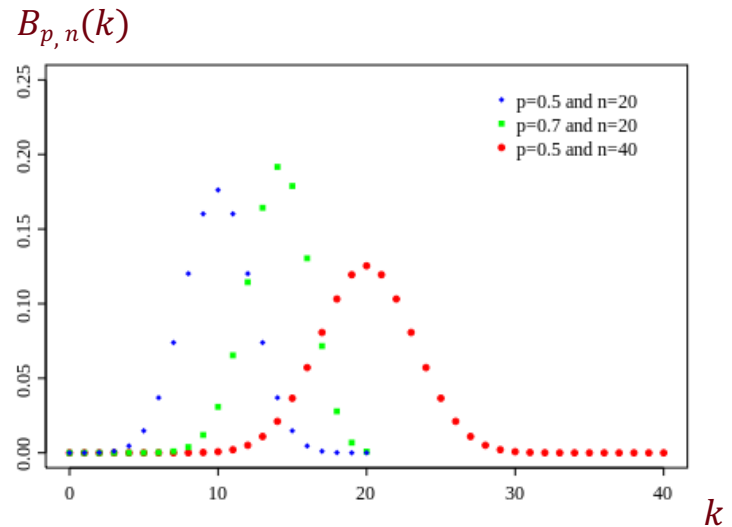
Cumulative frequency

Binomial Distribution

- Describes the number of successes in a series of similar and independent trials, each of which has exactly two possible outcomes: "success" or "failure"

- n = #trials
 p = #successful trials / #trials
- Description**
of the relative frequency
of achieving **exactly** k successes
as a function $B_{p,n}(k)$

- The following holds: $\sum_{i=0}^n B_{p,n}(i) = 1$

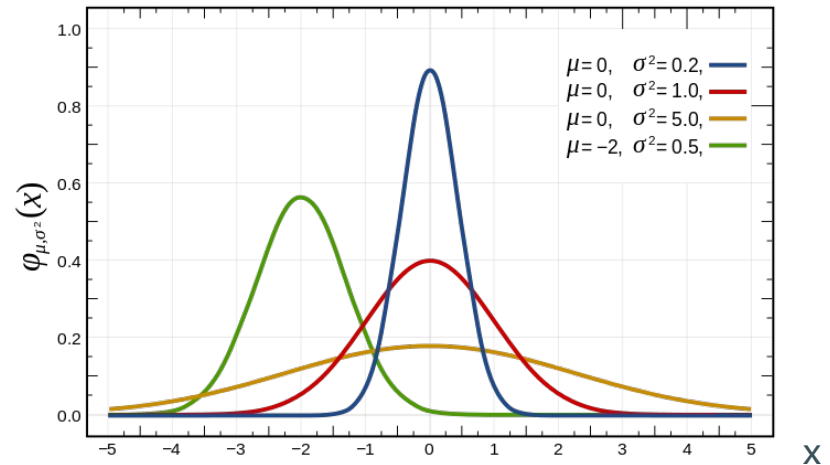


Normal distribution

- Base set: \mathbb{R}
- Measure of location: Mean
- Measure of dispersion: Variance
- Function for frequency distribution is in the continuous case
density function:

$$\int_{-\infty}^{\infty} \varphi_{\mu, \sigma^2}(x) dx = 1$$

In a normal distribution, the mean and median are equal

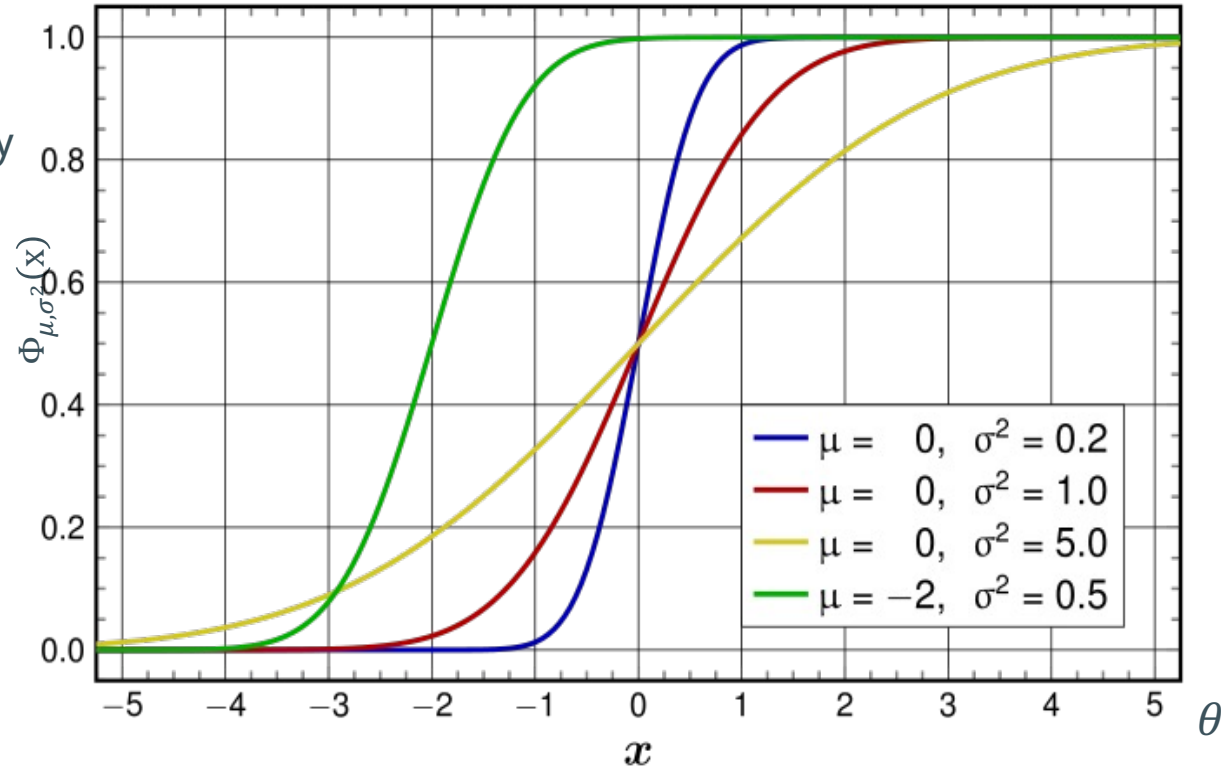


Distribution for Relative Frequency of

$$\varphi_{\mu, \sigma^2}(x) \leq \theta$$

$\Phi_{\mu, \sigma^2}(x)$ = area under the frequency distribution $\varphi_{\mu, \sigma^2}(x)$ from $-\infty$ to θ

→ so-called **distribution function**



From Relative Frequencies to Probabilities

Transition from relative frequencies to so-called probabilities as properties of the data-generating process

- Johann Bernoulli (1667–1748) and Pierre Laplace (1749–1822)
- Example: Probability of
of being male if you earn $\geq 400,000$ euros
- **But:** Even with large amounts of data, the probability of a property of the data-generating process is obviously only very roughly estimated by
#favourable cases / #possible cases

Consideration of the borderline case: #possible cases $\rightarrow \infty$

- Richard von Mises (c. 1883-1953)

Further development from 1930 onwards by Andrei Kolmogorov

Probability vs. Density Function

- Probability function
 - Probability for each feature/domain value
- Does not work with dense features
 - Probability for each individual value: 0
- Therefore, in this case: density function
- Use of density in distribution function
 - Determination of the probability that a certain event will occur up to x times
 - Distribution function for normal distribution

$$\Phi_{\mu, \sigma^2}(x) = \int_{-\infty}^x \varphi_{\mu, \sigma^2}(t) dt$$

- As $x \rightarrow \infty$ approaches 1

Normal Distribution

Density function

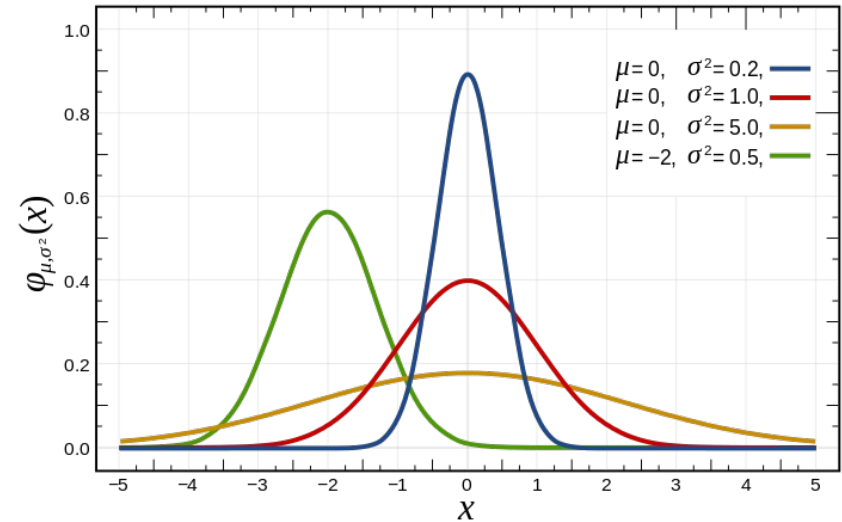
$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Standard normal distribution:

$\mu = 0$ and $\sigma = 1$

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

$\phi(x_0)$ Likelihood of x_0



Probability that when drawing from the population a value $\leq x$ is drawn

Statistical Basics

Hypothesis Testing

Hypothesis Test

Hypothesis: Chicks can recognise grains (circle) from birth and do not need to learn the shape of the feed first.

Experiment:

- Half of the circles and half of the triangles are provided for pecking (let's say 20 objects in total).
- If the assumption is true, $p_{\text{circle}} \gg 0.5$ should apply

Hypothesis H_0 :

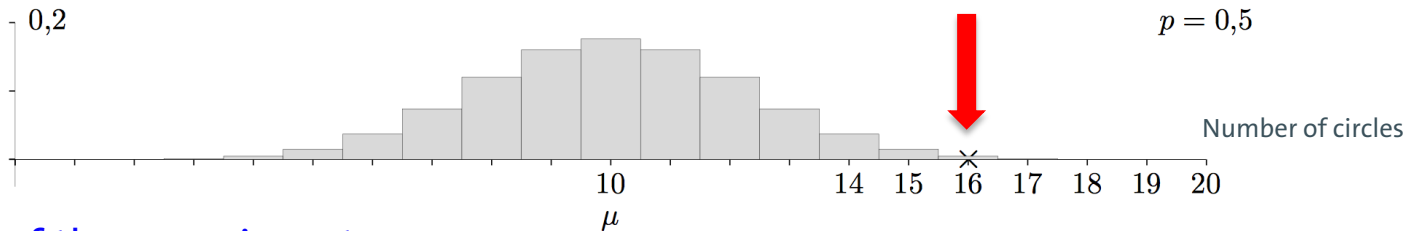
- Chicks do not distinguish between circles and triangles, $p_{\text{circle}} = 0.5$, mean value of the experiment should be 10, variance should be 2

Hypothesis H_1 :

- Chicks distinguish between circles and triangles, they peck more frequently at a circle

Experiment under Normal Distribution Assumption

If the assumption is false (i.e. H_0 is true),
then $p_{\text{Circle}} = 0.5$, mean value of $\mu=10$, $\sigma^2 = 2$ (empirical)



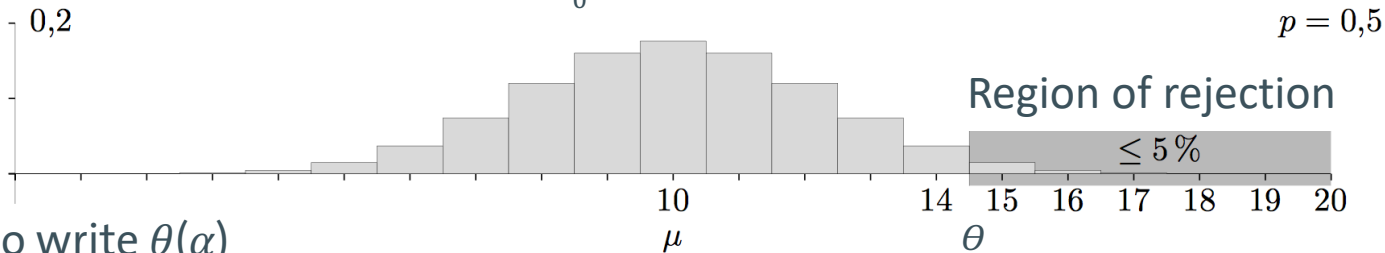
Outcome of the experiment:

- Outcome: On average, the chick pecks 16 times on the circle

Assumption: We want to minimise the probability of rejecting H_0 , even though it is true.

Rejection Region

- Objective: Keep the probability of error (rejection of H_0 , even though it is true) low
- Set error probability α to 0.05
- Determine θ such that $\Phi_{\mu, \sigma^2}(\theta) = \int_0^{\theta} \varphi_{\mu, \sigma^2}(x) dx = 0.95$



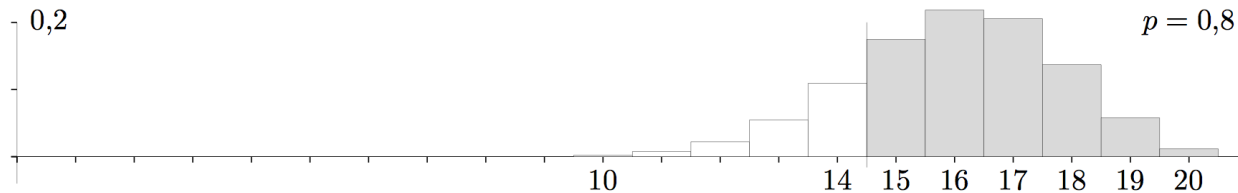
- We also write $\theta(\alpha)$
- If the test falls within the rejection region, there is a **significant deviation** and we speak of a **test with a significance level α**

Evaluation of the Experiment: Error Analysis

- The experiment falls within the rejection range for H_0
- Therefore: **Acceptance of the hypothesis H_1** as true
- Number of outputs with circle even 16
- Determine
$$\int_0^{16} \varphi_{\mu, \sigma^2}(x) dx = 0.979$$
- Probability of error only 0.021
- We say that $\alpha = 0.021$ (or 2.1%) and call this **a type 1 error**

Further Question

- Let us assume that we know the distribution $\mathcal{N}(\mu, \sigma^2)$ for the case that chicks have an innate ability to recognise grains.
- What is the probability that the chicks' ability would not be recognised?
- If chicks preferred circles with a probability of $p=0.8$, the result would be: $\beta = \Phi_{\mu, \sigma^2}(14) = 0,196 \approx 0.2$

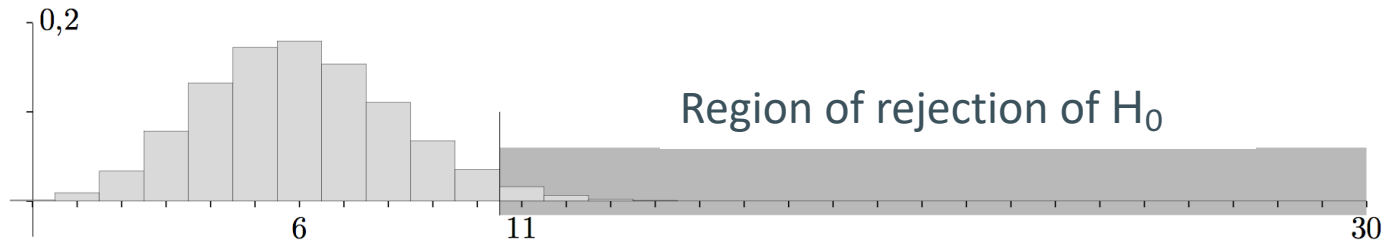


- The closer p gets to 0.5, the greater the **type 2 error** becomes.

Rejection Region on the Right

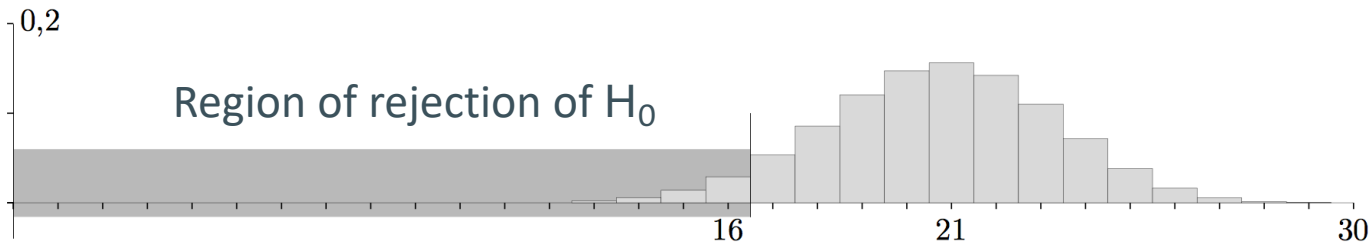
- Claim: A certain drug causes side effects in at most 20% of patients. We doubt this and test the null hypothesis at the 5% level.
- The sample size is $n = 30$.
- Choose H_0 if $p \leq \theta(\alpha)$ and H_1 if $p > \theta(\alpha)$

20% of 30 = 6



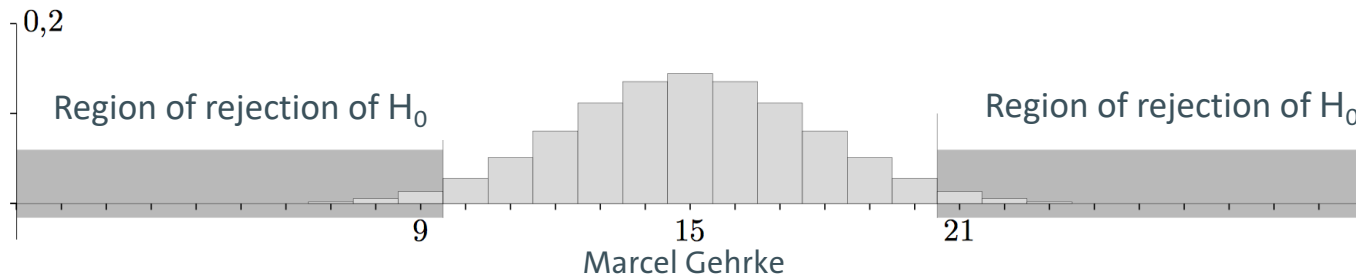
Rejection Area on the Left

- Claim: At least 70% of the cucumbers delivered meet the European curvature standard. We suspect the opposite and test at the 5% level.
- The sample size is $n = 30$
- Select H_0 if $p \geq \theta(1-\alpha)$ and H_1 if $p < \theta(1-\alpha)$



Rejection Area on Both Sides

- With random colouring, 50% of the series products should have a light tint. We want to detect deviations.
- The sample size is $n = 30$
- Select H_0 if $p \geq \theta(1-\alpha/2)$ and $p \leq \theta(\alpha/2)$;
 H_1 : otherwise







Type 1 and Type 2 Errors

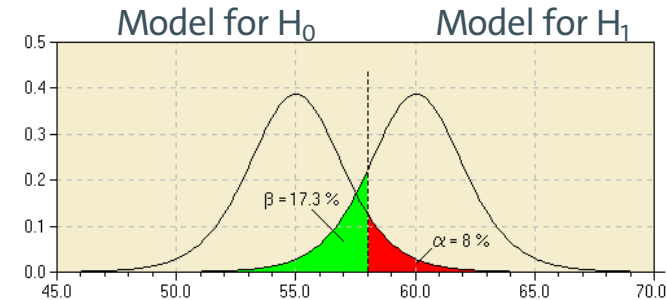
- Type 1: We reject H_0 , even though H_0 is true
 - If $\alpha=0.05$, then we reject H_0 in 5% of cases
 - Probability α with which we reject H_0 , i.e. make a type 1 error
- Type 2: We accept H_0 even though H_0 is false
 - The probability of making a Type 2 error is β
 - $1-\beta$ is then the probability of H_0 (correctly) NOT accepting
- However, different distributions are used as a basis; in general, the following applies: $\alpha \neq 1-\beta$

Possible Results of a Hypothesis Test

Reality

Result of hypothesis test

	H_0 true	H_1 true
H_0 true	Accurate $1 - \alpha$ 	Type 2 error β 
H_1 true	Type 1 error α 	Accurate $1 - \beta$ 



Summary: Hypothesis Testing

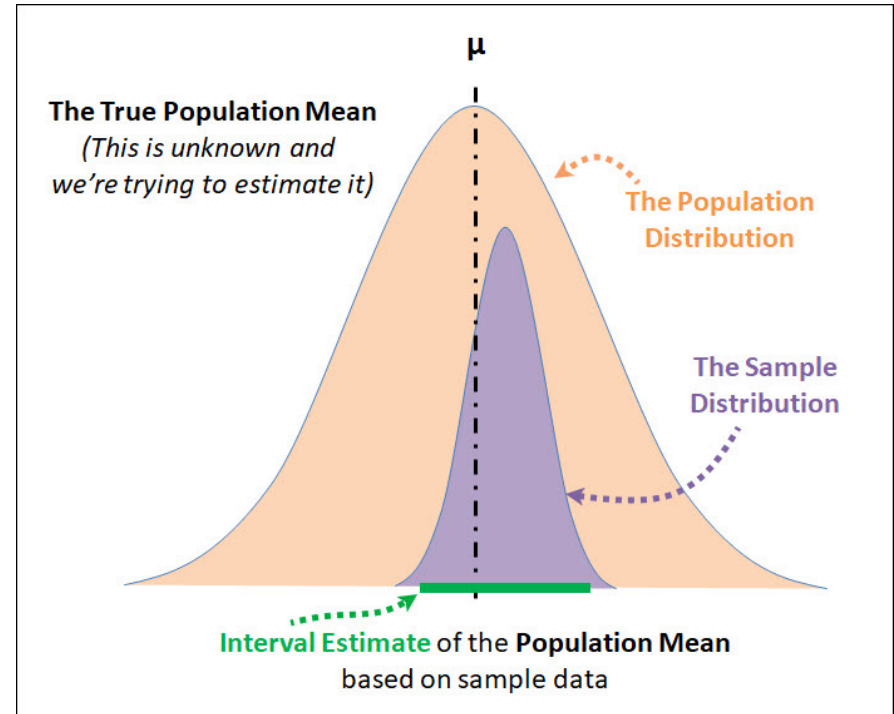
- To prove a hypothesis, one shows that the alternative hypothesis is extremely unlikely based on a test result.
- Which hypothesis is tested as the null hypothesis depends on the objective
- Important: The distribution assumption of the null hypothesis must be justified
- The parameters of the assumed distribution must be determined in a meaningful way
- How large should the sample be?
- How many data elements do we need to be able to make certain statements?

Statistical Basics

Estimators

Estimation of Parameters

- Evaluation of sample data
- Conclusions about the features of the population
- First, let us consider the normal distribution
 - Determining parameters from a sample



<https://cqeacademy.com>

Experiments, Random Variables, Distributions

Conducting experiments / evaluating data

- Determining feature values
- Values of statistical variables
- In the sense of drawing from a population: **random variable**

Example: Random variable X normally distributed

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- Standard normal distribution: $\mu = 0$ and $\sigma = 1$

Expectations Formally

- Expected value of random variable :

- Value that takes on average

- Discrete:

$$\mathbb{E}(X) = \sum_{i \in I} x_i p_i$$

where p_i is the relative frequency of occurrence of the value x_i

- Continuous:

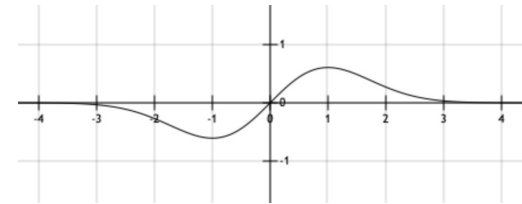
$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

f is the density distribution of X

- Notation sometimes also: $\mathbb{E}[X]$

Expected Value of the Standard Normal Distribution

- Density function of the standard normal distribution: $\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$
- The **expected value** of the standard normal distribution is 0
 - For $X \sim \mathcal{N}(0,1)$ the following holds
 - $\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot e^{-\frac{1}{2}x^2} dx = 0$
 - As integrand can be integrated and is point-symmetrical



Variance Formally

- Variance of random variable X :
 - Expected (squared) deviation from the mean value of X
 - Definition: $\text{Var}(X) := \mathbb{E}((X - \mathbb{E}(X))^2)$
 - Notation sometimes also: $\text{Var}[X]$
- $\text{Var}[X]$, if $X \sim \mathcal{N}(\mu, \sigma^2)$?

Multidimensional distributions

- We consider a two-dimensional distribution with **random variables X** and **Y**
- Let $\sigma_{X,Y} := Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$ be the covariance of X and Y

Correlation

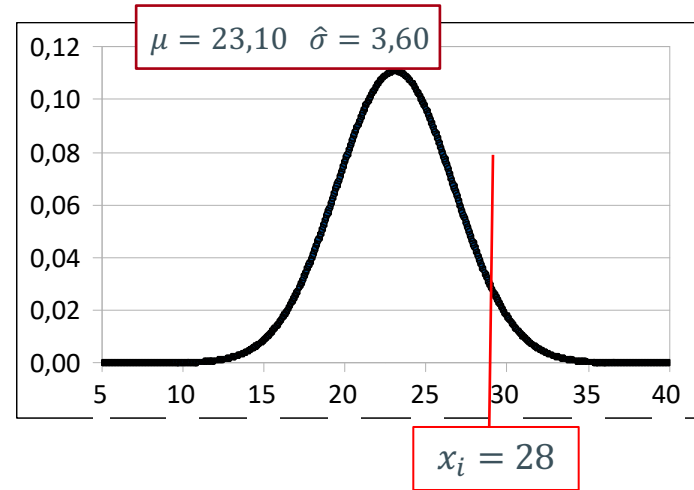
- Given two random variable X and Y
 - We call X and Y uncorrelated if $Cov(X, Y) = 0$
 - We call $\rho(X, Y)$ the correlation coefficient of X and Y with

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Interpretation of a Measured Value

Example $x_i = 28$

- Can only be interpreted for a given distribution
- x_i is above the arithmetic mean
- More precisely: x_i is more than one standard deviation above the arithmetic mean
- More precisely: What percentage of the total has values below/above 28?
- Standard score, also called z-score, helps to answer this question

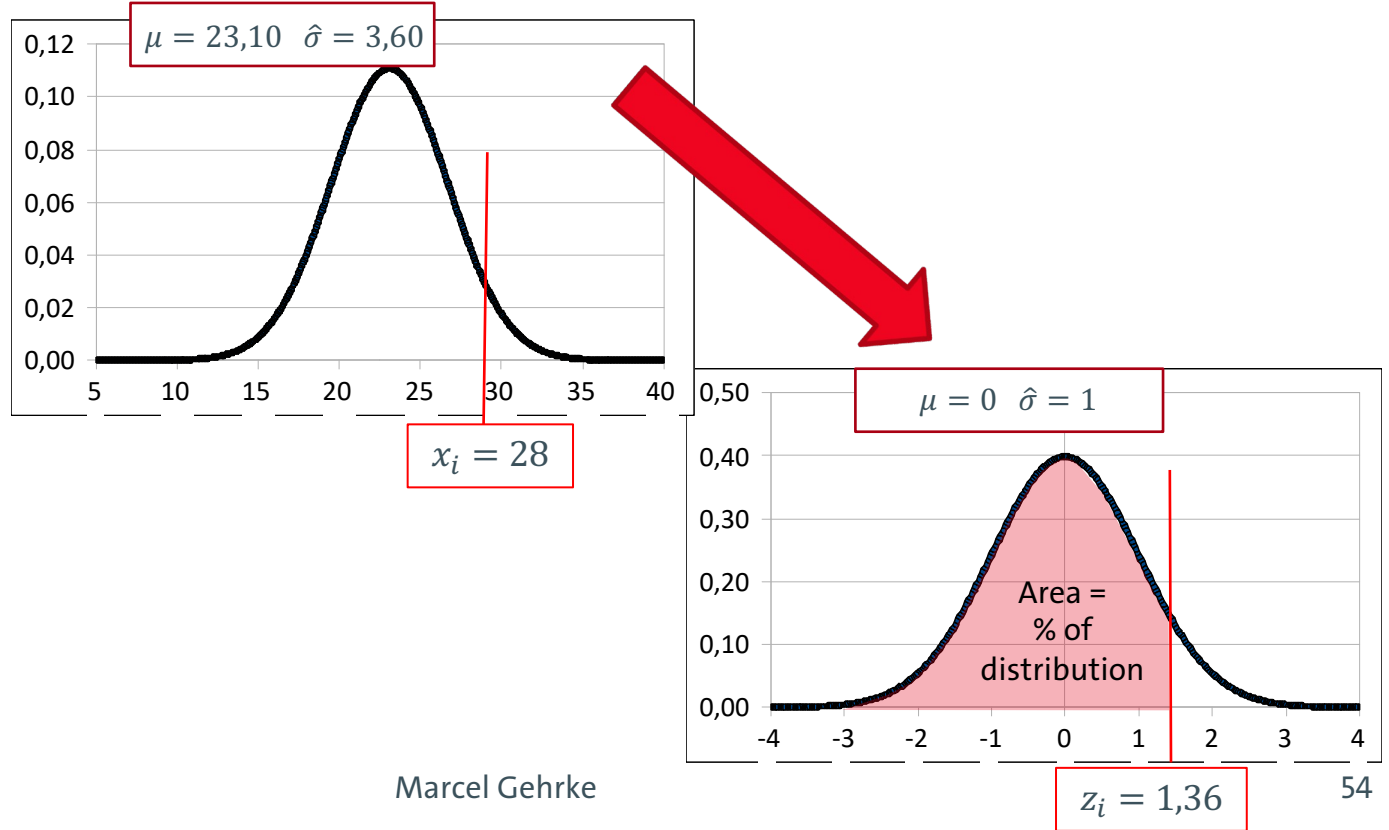


Z-score

- The z-score converts a normal distribution into a standard normal distribution.
- The z-score is carried out in two steps:
 - (1) First, the mean value is subtracted from each measured value.
 - (2) Then the result is divided by the standard deviation.

$$z_i = \frac{x_i - \bar{x}}{\hat{\sigma}}$$

z-score



Z-SCORE

- z-values can be easily interpreted using a z-table
- Tables for standard normal distribution always indicate the area under the curve to the left of a z-value
- The area indicates the proportion of the distribution whose values are less than or equal to the "critical" z-value
- Example:
 - $x_i = 28$
 - $z_i = 1,36$
 - $\text{Area}(z_i) = \Phi(z_i) = 0,91$
 - Proportion of z-values
 - 91% of the population have z-values less than or equal to 1,36
 - 91% of the population have x-values of 28 or below
 - Only 9% of the population have x-values greater than x_i

z-score

The z-table (standard normal distribution)

<i>z</i>	<i>Area</i>	<i>z</i>	<i>Area</i>	<i>z</i>	<i>Area</i>	<i>z</i>	<i>Area</i>
-3.00	0.00	-1.50	0.07	0.00	0.50	1.50	0.93
-2.90	0.00	-1.40	0.08	0.10	0.54	1.60	0.95
-2.80	0.00	-1.30	0.10	0.20	0.58	1.70	0.96
-2.70	0.00	-1.20	0.12	0.30	0.62	1.80	0.96
-2.60	0.00	-1.10	0.14	0.40	0.66	1.90	0.97
-2.50	0.01	-1.00	0.16	0.50	0.69	2.00	0.98
-2.40	0.01	-0.90	0.18	0.60	0.73	2.10	0.98
-2.30	0.01	-0.80	0.21	0.70	0.76	2.20	0.99
-2.20	0.01	-0.70	0.24	0.80	0.79	2.30	0.99
-2.10	0.02	-0.60	0.27	0.90	0.82	2.40	0.99
-2.00	0.02	-0.50	0.31	1.00	0.84	2.50	0.99
-1.90	0.03	-0.40	0.34	1.10	0.86	2.60	1.00
-1.80	0.04	-0.30	0.38	1.20	0.88	2.70	1.00
-1.70	0.04	-0.20	0.42	1.30	0.90	2.80	1.00
-1.60	0.05	-0.10	0.46	1.40	0.92	2.90	1.00

Z-SCORE

Interpretation of a value of a normally distributed feature

- Collection of a sample
 - Calculation of mean value and standard deviation
- Collection of features for person i
- Calculation of the z-value
- Looking up the size of the area below the z-distribution to the left of z_i
- The area $f(z_i)$ indicates the percentage of the population that has values less than or equal to z_i or x_i
- $1 - f(z_i)$ indicates the percentage of the population that has values greater than z_i or x_i

Percentile Ranks

- A percentile rank (PR) indicates what percentage of the population has values less than or equal to a critical value.

Task: IQ Score Analysis

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

Assumption: Normal distribution
 with $\mu = 100$; $\sigma = 15$
 What percentile rank corresponds to
 an IQ score of
 (a) 130; (b) 92.5; (c) 85; (d) 100; (e) 115?

z	Area	z	Area	z	Area	z	Area
-3.00	0.00	-1.50	0.07	0.00	0.50	1.50	0.93
-2.90	0.00	-1.40	0.08	0.10	0.54	1.60	0.95
-2.80	0.00	-1.30	0.10	0.20	0.58	1.70	0.96
-2.70	0.00	-1.20	0.12	0.30	0.62	1.80	0.96
-2.60	0.00	-1.10	0.14	0.40	0.66	1.90	0.97
-2.50	0.01	-1.00	0.16	0.50	0.69	2.00	0.98
-2.40	0.01	-0.90	0.18	0.60	0.73	2.10	0.98
-2.30	0.01	-0.80	0.21	0.70	0.76	2.20	0.99
-2.20	0.01	-0.70	0.24	0.80	0.79	2.30	0.99
-2.10	0.02	-0.60	0.27	0.90	0.82	2.40	0.99
-2.00	0.02	-0.50	0.31	1.00	0.84	2.50	0.99
-1.90	0.03	-0.40	0.34	1.10	0.86	2.60	1.00
-1.80	0.04	-0.30	0.38	1.20	0.88	2.70	1.00
-1.70	0.04	-0.20	0.42	1.30	0.90	2.80	1.00
-1.60	0.05	-0.10	0.46	1.40	0.92	2.90	1.00

IQ	z(IQ)	PR
130	2.0	98
92.5	-0.5	31
85	-1.0	16
100	0.0	50
115	1.0	84

Probabilities

- The z-table also makes it possible to make probability statements for specific intervals.
- What is the probability of an IQ value
 - (a) between 85 and 115; (b) between 70 and 130; (c) between 0 and 70;
 - (d) above 100

IQ	$z(IQ_1)$	$z(IQ_2)$	$p(z_1)$	$p(z_2)$	Δp
85 to 115	-1.0	1.0	.16	.84	.68
70 to 130	-2.0	2.0	.02	.98	.96
0 to 70	-6.7	-2.0	.00	.02	.02
> 100	0	∞	.50	1.0	.50

Probabilities

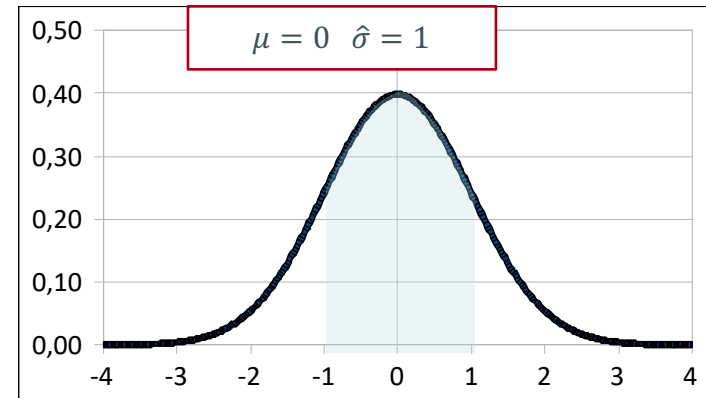
In general, the following applies to normally distributed features:

- 68.26% of values lie within the range:

$$\mu - 1,0 \cdot \sigma < x_i < \mu + 1,0 \cdot \sigma \quad \text{or} \quad -1,0 < z_i < 1,0$$

- 95.44% of values lie within the range:

$$\mu - 2,0 \cdot \sigma < x_i < \mu + 2,0 \cdot \sigma \quad \text{or} \quad -2,0 < z_i < 2,0$$

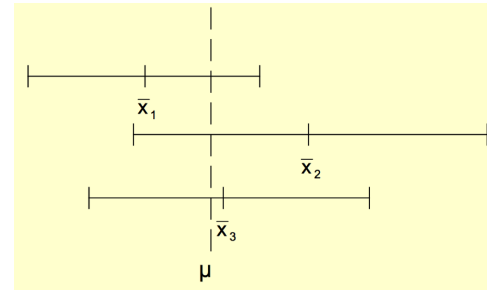


Sample Features

- We have learned about various sample features: e.g. mean, median, variance ("point estimators")
- In most cases, it is not the values for the specific **sample** that are of interest, but rather those for the underlying **population**
- The parameters from a sample are therefore used as **estimators** for the corresponding population parameters
- We expect that the larger a (representative) sample is, the more accurate the estimate will be

Sample Feature Distributions

- If samples are repeatedly taken from the same population, a new mean value is obtained for each sample
- If you collect a large number of samples, you also obtain many mean values
- Now you can consider the distribution of the resulting mean values
- This distribution is called **sampling feature distribution of the mean value**



Standard Error

- This "**distribution of mean values**" is itself normally distributed (if the feature is normally distributed).
- The **mean value of the sample feature distribution** for the mean values of the samples corresponds to the **mean value in the population**
- The **dispersion of the sample feature value distribution** is referred to as **the standard error** (of the mean value)
 - The standard error indicates how close an empirical sample mean is to the true population mean
 - This standard error of the mean can also be estimated from a single sample:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}_x^2}{n}}$$

- $\text{Var}(cx) = c^2 \cdot \text{Var}(x)$
- $\text{Var}(\sum x) = \sum \text{Var}(x)$

Justification: Standard Error of the Sample Mean

- Suppose a statistically independent sample of n observations x_1, x_2, \dots, x_n is taken from a statistical population with a standard deviation of σ (the standard deviation of the population). The mean value calculated from the sample, \bar{x} , will have an associated standard error on the mean, $\hat{\sigma}_{\bar{x}}$, given by: $\hat{\sigma}_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- The standard error on the mean may be derived from the variance of a sum of independent random variables, given the definition of variance and some properties thereof. If x_1, x_2, \dots, x_n is a sample of n independent observations from a population with mean \bar{x} and standard deviation σ ,
 - then we can define the total $T = \sum_{i=1}^n x_i$ and $\text{Var}(T) = \sum_{i=1}^n \text{Var}(x_i) = n\sigma^2$
 - the mean of these measurements \bar{x} (sample mean) is given by $\bar{x} = \frac{T}{n}$
 - the variance of the mean is then $\text{Var}(\bar{x}) = \text{Var}\left(\frac{T}{n}\right) = \frac{1}{n^2} \text{Var}(T) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$

https://en.wikipedia.org/wiki/Standard_error#Standard_error_of_the_sample_mean

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}_x^2}{n}}$$

Standard Error

Example: The motivation among the employees of a large company is to be determined . **10** employees are randomly selected and tested

- This results in a mean value of **60** with an estimated population variance of **90**

- What is the standard error of this mean?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{90}{10}} = \sqrt{9} = 3$$

- What would the standard error be if $\sigma^2=250$ and $n=10$?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{250}{10}} = \sqrt{25} = 5$$

- What would the standard error be if $\sigma^2=90$ and $n=90$?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{90}{90}} = \sqrt{1} = 1$$

Range Around the Mean

- The **standard error** is the standard deviation of the sample feature distribution
- Since the **sample feature distribution is normally distributed**, the probability that the true mean lies within a certain interval can be calculated
- Given $p = 0,68$, the population mean is at most one standard error away from the sample mean
- **Example:**
 - If $\bar{x} = 60$ and $\hat{\sigma}_{\bar{x}} = 3$, then for the population mean with $p = 0,68$, it holds that: $57 < \mu < 63$
- **Notation:** $P(\text{condition}) = p$ where $p \in [0, 1]$
- **Example:** $P(57 < \mu < 63) = 0,68$

Statistical Basics

Confidence Intervals

Confidence Intervals

- A **confidence interval** is a symmetrical range around the sample mean within which the population mean lies with a certain probability.

$$P(\bar{x} - 1,00 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 1,00 \cdot \hat{\sigma}_{\bar{x}}) = 0,682$$

$$P(\bar{x} - 2,00 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 2,00 \cdot \hat{\sigma}_{\bar{x}}) = 0,954$$

$$P(\bar{x} - 1,96 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 1,96 \cdot \hat{\sigma}_{\bar{x}}) = 0,95$$

$$P(\bar{x} - 2,57 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 2,57 \cdot \hat{\sigma}_{\bar{x}}) = 0,99$$

Confidence Interval

- The position and width of the confidence interval depend on the random confidence limits
- These depend on:
 - the sample size
 - the estimation function and its distribution, and
 - the so-called confidence level α
- **The width of the confidence interval** is an expression of the accuracy of the parameter estimate!
 - A higher **confidence level** (smaller α) leads to a widening of the confidence interval and ...
 - ... a larger **sample size** leads to a reduction in the confidence interval

Confidence Interval: Derivation

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a normally distributed random variable and let (X_1, \dots, X_n) be a mathematical sample from the population X .

- The **variance σ^2** of the normally distributed population is **known**.
A confidence estimate must be given for the unknown parameter μ .

As a point estimator for μ , we choose the arithmetic mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \text{ where } \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Confidence Interval: Derivation

- The probability that the absolute value of the estimation error is less than the **limit d** is given by $(1 - \alpha)$:

$$P(|\bar{X} - \mu| \leq d) = 1 - \alpha, \text{ where the estimation error is } \bar{X} - \mu$$

- Resolve absolute value :
 - Case: $\bar{X} - \mu \geq 0$: $\bar{X} - \mu \leq d \rightarrow \bar{X} - d \leq \mu$
 - Case: $\bar{X} - \mu \leq 0 = -(\bar{X} - \mu) \geq 0$: $-\bar{X} + \mu \leq d \rightarrow \bar{X} + d \geq \mu$
- Manipulation of terms:

$$P(|\bar{X} - \mu| \leq d) = P(\bar{X} - d \leq \mu \leq \bar{X} + d) = 1 - \alpha \quad (\text{symmetry of the ND density function})$$

Confidence Interval: Derivation

- To determine the value d , we **z-score** the random variable \bar{X} :
 - given
$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad z = \frac{x - \mu}{\sigma'} \quad \sigma' = \sqrt{\frac{\sigma^2}{n}}$$

(standard deviation σ = square root of variance)
- Z-scored:
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \quad Z \sim \mathcal{N}(0, 1)$$

Plan: **Insert** into $P(|\bar{X} - \mu| \leq d)$

Rewrite beforehand $\rightarrow P\left(\left|\frac{\bar{X} - \mu}{\sigma} \sqrt{n}\right| \leq \frac{d}{\sigma} \cdot \sqrt{n}\right)$

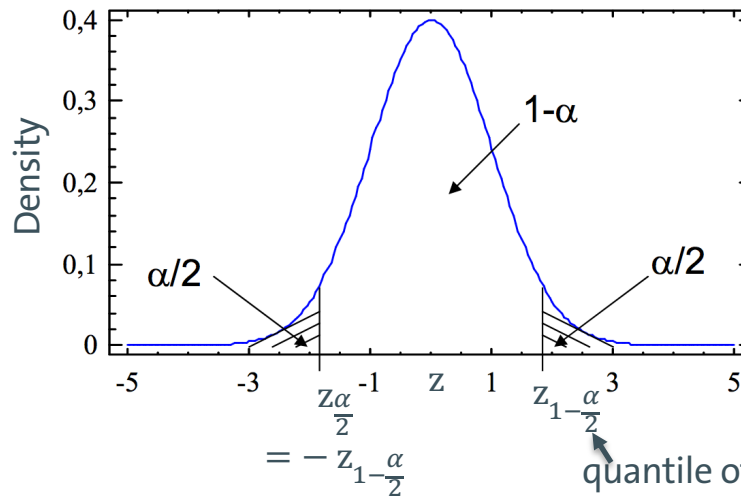
Confidence Interval: Derivation

- Given $P\left(\left|\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right| \leq \frac{d}{\sigma} \cdot \sqrt{n}\right) = 1 - \alpha$, Abbreviation: $z_{1-\frac{\alpha}{2}} = \frac{d}{\sigma} \cdot \sqrt{n}$
 - It follows that with $P(|\bar{X} - \mu| \leq d) = P(\bar{X} - d \leq \mu \leq \bar{X} + d)$: $P\left(\left|\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right| \leq \frac{d}{\sigma} \cdot \sqrt{n}\right)$
- $$= P(|Z| \leq z_{1-\frac{\alpha}{2}})$$
- $$\rightarrow d = \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$
- $$= P\left(\frac{z_{\alpha}}{2} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) \quad z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$$
- $$= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Confidence Interval: Derivation

- The confidence interval $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \right]$

thus covers the true parameter μ with the probability $(1 - \alpha)$



— Density function of the standard normal distribution

quantile of the standard normal distribution

Confidence Interval: Interpretation

- Each specific sample provides us with a realisation of the random variable \bar{X} and thus a realised confidence interval:

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \right]$$

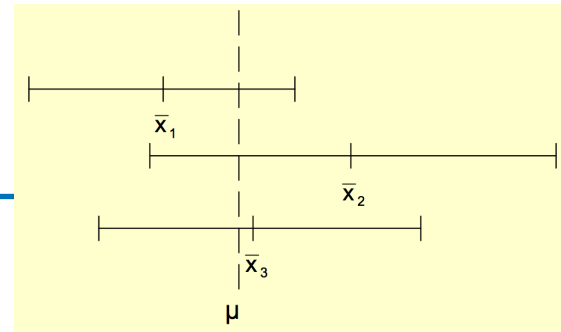
- The table contains some typical $z_{1-\frac{\alpha}{2}}$ -values (two-tailed queries) and $z_{1-\alpha}$ -values (one-tailed queries):

$1 - \alpha$	α	$z_{1-\frac{\alpha}{2}}$	$z_{1-\alpha}$
0.95	0.05	1.96	1.64
0.99	0.01	2.58	2.33
0.999	0.001	3.29	3.09

$$\Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

$$\Phi(z_{1-\alpha}) = 1 - \alpha$$

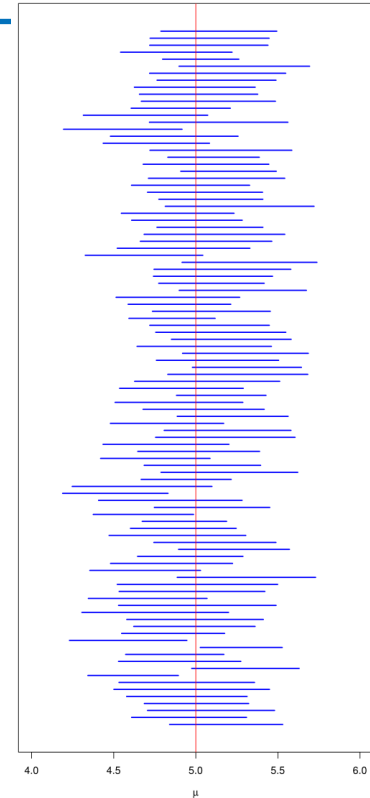
Confidence Interval: Comments



- The location of the specific confidence interval is determined by the specific sample
- At a confidence level of $(1 - \alpha) = 0,95$, this means that
 - In 95% of all cases, the confidence interval contains the unknown parameter of the population, and in 5% of cases it does not
 - In other words, if one claims k times that the unknown parameter lies within the confidence interval, one can expect an average $\alpha \cdot k$ error

Confidence Interval: Example

- Experiment
 - Normally distributed GG
 - $\mu = 5$
 - $\alpha = 0,05$
 - 100 samples
 - $n = 30$
- Calculate mean values and confidence intervals for each sample
 - 94 of the intervals overlap
 - 6 intervals do not
- Mean values of the 100 samples are normally distributed
 - Sample feature distribution



Confidence Interval: Comments

- The width of the confidence interval for the expected value μ is $2d$ and depends on α, n, σ and the distribution of the associated estimation function.

$$2d = 2 \cdot \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

- The larger α (n constant), the smaller the confidence interval
- The larger n , the smaller the confidence interval
- $2d$: Measure of the accuracy of the estimation of μ
- α a measure of the risk
- Planning of the sample size

Confidence Interval: Comments

- **Planning of the sample size**
 - Given:
 - half the width of the confidence interval d ,
 - variance σ^2
 - confidence level $(1 - \alpha)$
 - Required:
 - Sample size n

$$2d = 2 \cdot \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

$$d = \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

$$\sqrt{n} = \frac{\sigma}{d} \cdot z_{1-\frac{\alpha}{2}}$$

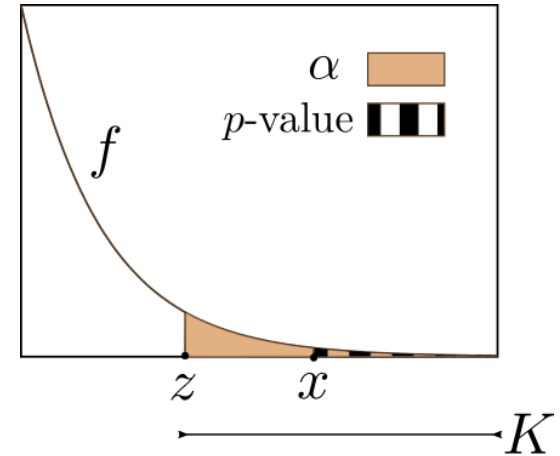
$$n = \frac{\sigma^2}{d^2} \cdot z_{1-\frac{\alpha}{2}}^2$$

Confidence Interval for Variance

- Similar considerations
- Derivation of the required sample size is also possible for this

P-value (One-sided Rejection Region)

- Hypothesis test H_0 vs. H_1
- How extreme is the the collected data calculated value of the test statistic?
- P-value = probability, given H_0 holds to obtain the specific or an more extreme value of the test statistic



For this realization x in the rejection range K , the p -value is less than α , or equivalently, the realization of the test statistics x is greater than the critical value z . Here f is the probability density of the distribution under the null hypothesis

In some publications, α is referred to as the p-value!

[Wikipedia]

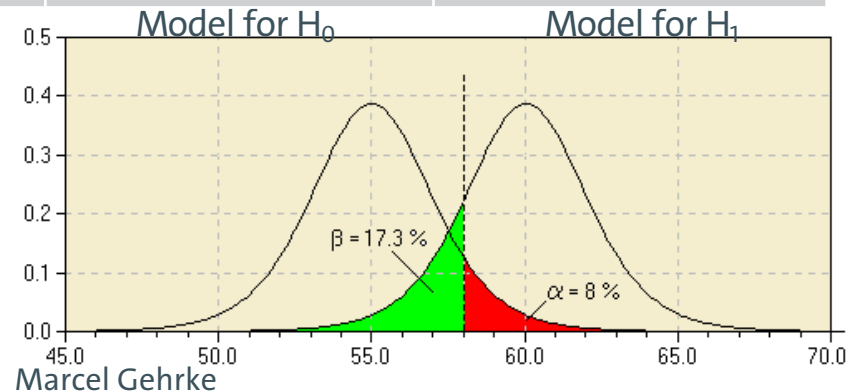
Not only α is Relevant

- We are also interested in:
 - **What is the probability** that a statistical test **will correctly reject** the **null hypothesis H_0** if the alternative hypothesis H_1 is true?
- Interpreted as **the "discriminatory power"** of the test
 - High selectivity of the test argues against, low selectivity for the null hypothesis H_0
- Objective:
 - Determine **rejection region A** such that the probability of rejecting a "false null hypothesis" H_0 , i.e. of **accepting the alternative hypothesis H_1** , is **as high as possible under the condition that H_1 is true**

Discrimination Power of a Test

		Reality	
		H_0 is true	H_1 is true
Decision of the test for H_0	Correct decision Probability: $1 - \alpha$	Type-2 error Probability: β
	... for H_1	Type-1 error Probability: α	Correct decision Probability: $1 - \beta$ (power)

- What discrimination power has a value of $1 - \beta$
- Choice of β -level?
- Model for H_1 required




Determinants of the Power

- The power $(1-\beta)$ increases:
 - with a growing difference of $(\mu_0 - \mu_1)$ (this means that a large difference between two subpopulations is less likely to be overlooked than a small difference)
 - with decreasing dispersion of characteristics σ
 - with increasing significance level α (unless β is specified)
 - with increasing sample size, as the standard error then becomes smaller: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
 - for single-sided tests compared to two-sided tests: For the two-sided test, a sample size of about 25% larger is needed to achieve the same power as for the single-sided test

Further Quality Criteria for Estimators

- Adherence to expectations
- Consistency
- Efficiency
- Reliability
- Validity
- Objectivity



This is only intended to provide a rough overview.

Statistical Basics

Corrected Sample Variance

Sampling Function

- In statistics, a sampling function, also known as sampling statistics or simply statistics, summarises information from a sample in a specific form as a function.
- Example of a sampling function: estimation function

▪ Notation:

Arithmetic mean	Sampling function
$\bar{x} := \frac{1}{n} (x_1 + x_2 + \dots + x_n)$	$\bar{X} := \frac{1}{n} (X_1 + X_2 + \dots + X_n)$

x_i : concrete values

Name of the estimation function

X_i : values still to be determined
(random variables)

$$\mathbb{E}(X) = \sum_{i \in I} x_i p_i$$

Concept of Unbiasedness

- An estimator is said to be **unbiased** if its expected value is equal to the true value of the parameter to be estimated
 - Estimation of μ of population by sample means

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

- If $x_i \sim \mathcal{N}(\mu, \sigma^2)$ is randomly drawn from the population, then $\mathbb{E}(\bar{x}) = \mu$
- Expected value of \bar{x} :

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(x_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu = \mu$$

- Sample mean, i.e. unbiased estimator of μ

Corrected Sample Variance

- Given sample values (x_1, \dots, x_n)
- Corrected sample variance

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

Why $n - 1$?
When using sample mean \bar{x} ?

- With sample mean

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

Corrected Sample Variance: Why $n - 1$?

- Sample values (x_1, \dots, x_n) are manifestations of the **independently identically distributed** random variables (X_1, \dots, X_n) with variance σ^2 and mean μ of the population
- Then, S_0^2 is an unbiased estimator for σ^2 and s_0^2 is an unbiased estimator for the variance

$$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

μ not \bar{x}

- The following holds

$$\mathbb{E}(S_0^2) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}((X_i - \mu)^2) = \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n} n \sigma^2 = \sigma^2$$

Corrected Sample Variance: Why $n - 1$?

- μ is usually unknown and estimated by

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

- As an estimation function, we obtain the following estimate for σ^2

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad \rightarrow \quad s_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Testing for unbiasedness via the expected value of S_1^2

$$S_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Corrected Sample Variance: Why $n - 1$?

$$\begin{aligned}\mathbb{E}(S_1^2) &= \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2\right) \\ &= \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right) \\ &= \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (X_i - \mu)^2 - 2 \sum_{i=1}^n (X_i - \mu)(\bar{X} - \mu) + n(\bar{X} - \mu)^2\right) \\ &= \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + n(\bar{X} - \mu)^2\right)\end{aligned}$$

Corrected Sample Variance: Why $n - 1$?

$$\begin{aligned} \frac{1}{n} \mathbb{E} \left(\sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + n(\bar{X} - \mu)^2 \right) &= \frac{1}{n} \mathbb{E} \left(\sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X} - \mu)(\bar{X} - \mu) + n(\bar{X} - \mu)^2 \right) \\ &= \frac{1}{n} \mathbb{E} \left(\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right) = \frac{1}{n} \left(\sum_{i=1}^n \mathbb{E}((X_i - \mu)^2) - n \cdot \mathbb{E}((\bar{X} - \mu)^2) \right) \\ &= \frac{1}{n} (n \cdot \text{Var}(X) - n \cdot \text{Var}(\bar{X})) = \text{Var}(X) - \text{Var}(\bar{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \end{aligned}$$

Corrected Sample Variance: Why $n - 1$?

- Result of $E(S_1^2)$

$$E(S_1^2) = \frac{n-1}{n} \sigma^2$$

- S_1^2 's estimation function not unbiased for σ^2

- Solution: multiply by $\frac{n}{n-1}$

- Unbiased estimation function for σ^2

$$S_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$S_1^2 = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Therefore, $E(S_1^2) = \sigma^2$ holds

We distinguish between:
empirical and corrected
variance

Statistical Basics

Quality Criteria of Estimators

Consistency and Efficiency of an Estimator

- An estimator is **consistent** if it becomes more and more accurate for increasingly larger samples
 - The estimate can be made as accurate as desired by increasing the sample size sufficiently
- The **efficiency** (effectiveness) of the estimate characterises the precision with which it estimates parameters
 - An estimation function is more efficient the smaller the dispersion (or variance) of the estimated values around the parameter
 - The greater the dispersion of the sample feature distribution, the lower the efficiency of the corresponding estimated value

Reliability

Accuracy of a test with multiple indicators or features (e.g. questionnaire) and, for example, averaging of the values determined for the sub-features

- **Internal (inner) consistency**
 - Do different summarised features (e.g. at different points in a questionnaire) measure the same thing?
- **(Temporal) stability**
 - Is the same thing measured at different points in time (when the test is repeated)?

Determining the Reliability of a Test

- **Retest reliability:**
 - Determining the statistical correlation between two **consecutive** measurements
 - A test is accurate if it produces the same result at several points in time
 - Correlation of the same questionnaire total score at different points in time with the same test subjects (not suitable for temporary features, e.g. mood)
- Recap:
 - Given two random variable X and Y
 - We call $\rho(X, Y)$ the correlation coefficient of X and Y with $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$

Determining the Reliability of a Test

- **Split-half reliability:**
 - Correlation between two halves of the items in a test
- **Cronbach's alpha** (measure of internal consistency):
 - Mean value of the correlations r between all individual items
 - Sufficient reliability: $r = 0.75$
 - Good reliability: $r = 0.90$

Further Quality Criteria

- **Validity:** Does a test measure what it is supposed to measure?
- **Objectivity:** Independence of the test results from the framework conditions (boundary conditions) and distorting third factors

Statistical Basics

Differential Hypotheses

Acknowledgements

- The following materials have been adapted from:
- Statistics lecture (WS08/09) from the Psychology programme at the University of Freiburg

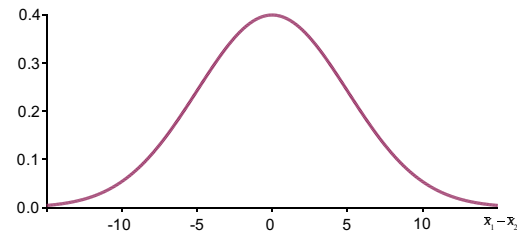
Differential Hypotheses

- Do women earn more than men?
 - Do the mean values of two groups differ?
 - Independent samples
- Is the mean salary higher after further training than before further training?
 - Does the mean value of a sample differ at two measurement points in time?
 - Dependent samples
- Is the mean IQ of a group above 100?
 - Does the mean value of a group differ from a specified value?
 - Test regarding group

Difference hypotheses: Independent samples

Do the mean values of two groups differ?

- Difference between the means of two samples: $\Delta_x = \bar{x}_1 - \bar{x}_2$
- H_0 : Difference irrelevant
- Estimate the density function for Δ_x if H_0 was true
- **Sample feature distribution:**
Distribution of mean differences under H_0
- How are empirical mean differences distributed when samples are taken very frequently?
- Distribution of mean differences for large samples is normally distributed



Standard Error of Feature Value Distribution

- For mean value:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}_x^2}{n}}$$

- For differences, it depends on the variances and sizes of the two subsamples:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_{x_1}^2}{N_{x_1}} + \frac{\hat{\sigma}_{x_2}^2}{N_{x_2}}}$$

- Required to interpret found differences in means

t-distribution

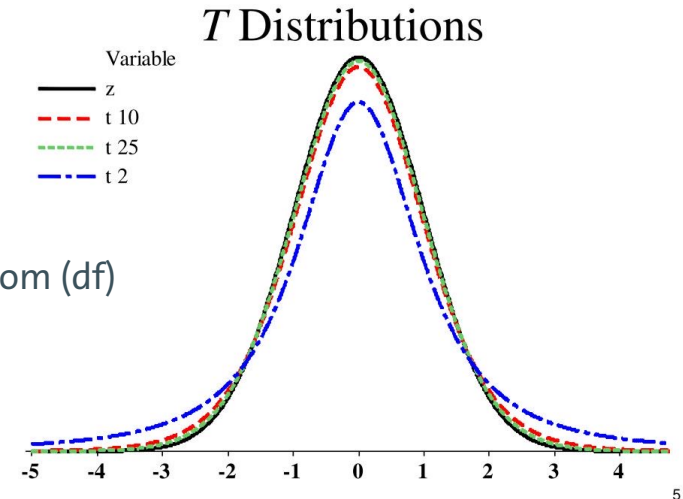
- Empirical (found) mean difference divided by standard error results in so-called **t-distribution**

$$t_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

- The exact form of the t-distribution depends on its degrees of freedom (df)

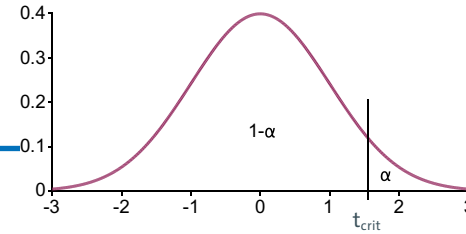
$$df = N_1 + N_2 - 2$$

- For $df > 120$, it is almost identical to the z-distribution (standard normal distribution)
- The smaller df , the narrower the peak of the t-distribution



The t-distribution

df	p=0.8	p=0.9	p=0.95	p=0.975	p=0.99	p=0.995
1	1.376	3.078	6.314	12.706	31.821	63.657
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1.372	1.812	2.228	2.764	3.169
20	0.860	1.325	1.725	2.086	2.528	2.845
30	0.854	1.310	1.697	2.042	2.457	2.750
40	0.851	1.303	1.684	2.021	2.423	2.704
50	0.849	1.299	1.676	2.009	2.403	2.678
60	0.848	1.296	1.671	2.000	2.390	2.660
70	0.847	1.294	1.667	1.994	2.381	2.648
80	0.846	1.292	1.664	1.990	2.374	2.639
90	0.846	1.291	1.662	1.987	2.368	2.632
100	0.845	1.290	1.660	1.984	2.364	2.626
200	0.843	1.286	1.653	1.972	2.345	2.601
1000	0.842	1.282	1.646	1.962	2.330	2.581



For a two-tailed test, t_{crit} must be selected so that a range of $\alpha/2$ is "cut off from the distribution".

Critical t-values:
 $\alpha = .05$, one-tailed, $df=100$:
 $t_{crit}(100) = 1.66$

$\alpha = .05$, two-tailed, $df=100$:
 $t_{crit}(100) = 1.98$

$\alpha = .01$, one-tailed, $df=100$:
 $t_{crit}(100) = 2.36$

Requirements for t-test Application

- (1) Variable has an **interval scale** (arithmetic mean is defined)
- (2) **Normal distribution** of the feature in the population
 - Can be tested (Kolmogorov-Smirnov test)
 - Not covered in depth here
- (3) **Homogeneity of variance**
 - "Equal" variances of the feature in both populations
 - "Variance of variance" small
 - Can be tested (Levene's test)
 - Not discussed in detail here
- (4) **Independence** of samples

Differential Hypotheses: **Dependent** Samples

- Drawing a feature carrier into the first sample influences the assignment of a feature carrier to the second sample
- Values from two samples are assigned **in pairs**.
 - Both subsamples are always the same size!
- **Repeated measurement**
 - The same feature is measured twice (or more) in the same individuals.
- **Parallelisation**
 - Two similar individuals are matched with each other
- **Matching**
 - Each person in sample 1 is assigned to a person in sample 2

Dependent Samples: Sample Calculation

- Does the attitude towards the subject of computer science change within the first 6 weeks of study?
- **Dependent variable:** Attitude towards studying computer science (value range 5 to 25)
- **Independent variable:** Measurement time (1st week vs. 6th week)

Test subject	1st week	Week 6
1	16	20
2	18	19
3	23	23
4	14	16
...
<i>mean</i>	19.67	18.98

Sample Calculation

- The difference between the measured values can be calculated for each person (change in opinions).

Ts	Week 1	Week 6	$D=x_2 - x_1$
1	16	20	4
2	18	19	1
3	23	23	0
4	16	14	-2

mean	19.67	18.98	.68

Hypotheses

- The statistical hypotheses of the **t-tests for dependent samples** refer to the **mean value of the differences** between all individuals
 - Advantage: It is now irrelevant whether there is large variance within the measurement points.
- Non-directional hypothesis:
 - $H_0 : \mu_d = 0$
 - $H_1 : \mu_d \neq 0$
- Directed hypothesis (1):
 - $H_0 : \mu_d \leq 0$
 - $H_1 : \mu_d > 0$
- Directed hypothesis (2):
 - $H_0 : \mu_d \geq 0$
 - $H_1 : \mu_d < 0$

Standard Error and t-value

- The standard error is required in order to evaluate the empirically determined difference

$$\hat{\sigma}_{\bar{x}_d} = \frac{\hat{\sigma}_{x_d}}{\sqrt{N}}, \text{ where } \hat{\sigma}_{x_d} = \sqrt{\frac{\sum_{i=1}^N (x_{di} - \bar{x}_d)^2}{N-1}} \quad \text{Based on corrected sample variance}$$

- The standard error can now be used to calculate an empirical normalised t-value

- Normalisation with regard to standard deviation $z_i = \frac{x_i - \bar{x}}{\hat{\sigma}}$
(cf. z-standardisation)

$$t_{df} = \frac{\bar{x}_d}{\hat{\sigma}_{x_d}}, \text{ where } df = N - 1$$

Standard Error and t-value

In the sample data set: $\bar{x}_d = 0.68$

$$\hat{\sigma}_{x_d} = 2.78$$

$$N = 60$$

■ This results in:

$$\hat{\sigma}_{\bar{x}_d} = \frac{2.78}{\sqrt{60}} = 0.36$$

$$t_{59} = \frac{0.68}{0.36} = 1.89$$

Critical t-value & Interpretation

$$t_{emp,59} = 1.89$$

$$t_{crit,59} = ?$$

- Open question \Rightarrow two-tailed test
- $\alpha = .05$

Interpretation:

- $t_{emp} < t_{crit}$
- Therefore: No significant difference!

df	p=0.8	p=0.9	p=0.95	p=0.975	p=0.99	p=0.995
1	1.376	3.078	6.314	12.706	31,821	63,657
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0.978	1,638	2,353	3,182	4,541	5,841
4	0.941	1,533	2,132	2.776	3.747	4.604
5	0.920	1.476	2,015	2.571	3.365	4,032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1,397	1,860	2,306	2,896	3,355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1,372	1,812	2.228	2.764	3.169
20	0.860	1,325	1.725	2.086	2.528	2.845
30	0.854	1,310	1,697	2.042	2.457	2.750
40	0.851	1,303	1,684	2,021	2,423	2,704
50	0.849	1.299	1.676	2.009	2.403	2.678
60	0.848	1,296	1.671	2.000	2.390	2.660
70	0.847	1,294	1,667	1,994	2.381	2.648
80	0.846	1,292	1,664	1,990	2.374	2.639
90	0.846	1,291	1,662	1,987	2,368	2,632
100	0.845	1,290	1,660	1.984	2.364	2,626
200	0.843	1,286	1,653	1.972	2.345	2.601
1000	0.842	1.282	1.646	1.962	2.330	2.581

One-group t-test

- Objective: Comparison of the mean value of a sample with a given (constant) value
- Examples:
 - Check whether a certain group of people differs in intelligence from the population mean (100)
 - Check whether the actual duration of study differs from the standard duration of study
 - Check whether the difference in reaction times under two conditions differs from zero

One-group t-test

Requirements

- Normal distribution of the feature
- Interval scale level of the feature
- The sample is random

One-group t-test

Statistical hypotheses

- Non-directional hypothesis:
 - $H_0 : \mu = c$
 - $H_1 : \mu \neq c$
- Directed hypothesis (1):
 - $H_0 : \mu \leq c$
 - $H_1 : \mu > c$
- Directed hypothesis (2):
 - $H_0 : \mu \geq c$
 - $H_1 : \mu < c$

Standard error and t-value

- Calculation of the standard error

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{N}}$$

- Calculation of the t-value

$$t(df = N - 1) = \frac{\bar{x} - c}{\hat{\sigma}_{\bar{x}_d}}$$

Example

- Is the IQ of children, classified as gifted, really above the population mean (100)?
- Hypotheses:
 - $H_0 : \mu \leq 100$
 - $H_1 : \mu > 100$
- Sample features for N=10 (assumptions):
 - Mean: 108.50
 - Standard deviation: 14.35

$$\hat{\sigma}_{\bar{x}} = \frac{14.35}{\sqrt{10}} = 4.54 \quad t(9) = \frac{108.5 - 100}{4.54} = 1.87$$

Example

- $t_{\text{emp}}(9) = 1.87$
- $t_{\text{crit}}(9) = ?$
 - Directed question \Rightarrow one-tailed test
 - $\alpha = .05$
- Interpretation:
 - $t_{\text{emp}} > t_{\text{crit}}$
 - H_0 is rejected

df	p=0.8	p=0.9	p=0.95	p=0.975	p=0.99	p=0.995
1	1.376	3.078	6.314	12.706	31,821	63,657
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0.978	1,638	2,353	3,182	4,541	5,841
4	0.941	1,533	2,132	2.776	3.747	4.604
5	0.920	1.476	2,015	2.571	3.365	4,032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1,397	1,860	2,306	2,896	3,355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1,372	1,812	2.228	2.764	3.169
20	0.860	1,325	1.725	2.086	2.528	2.845
30	0.854	1.310	1,697	2.042	2.457	2.750
40	0.851	1.303	1,684	2,021	2,423	2,704
50	0.849	1.299	1.676	2.009	2.403	2.678
60	0.848	1,296	1.671	2.000	2.390	2,660
70	0.847	1,294	1,667	1,994	2.381	2.648
80	0.846	1,292	1,664	1,990	2.374	2,639
90	0.846	1,291	1,662	1,987	2,368	2,632
100	0.845	1,290	1,660	1,984	2.364	2,626
200	0.843	1,286	1,653	1.972	2.345	2.601
1000	0.842	1.282	1.646	1.962	2.330	2.581

Summary of the 3 Types of t-tests

	Independent samples	Dependent samples	Single group t-test
Question			
Prerequisites			

Test Methods

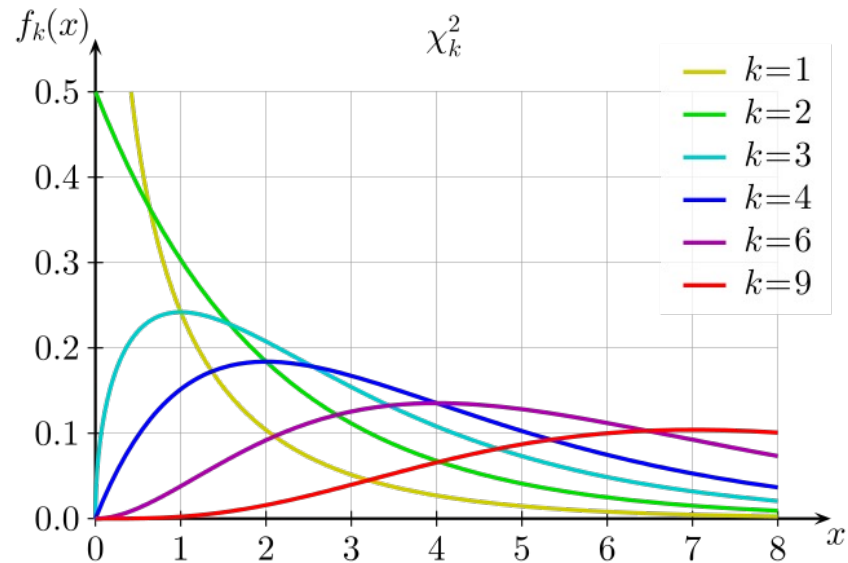
- Parametric methods
 - The variables involved must have the required distribution form (e.g. normal distribution for the t-test).
 - Interval scales required
 - But then good “explanatory power” (see „power of a test“)
- Non-parametric methods are used...
 - ⇒ For the analysis of ordinal or nominal scaled variables
 - ⇒ If the normal distribution assumption of the overall feature is violated
 - ⇒ Example: The sum of the squares of k normally distributed random variables is not normally distributed

Statistical Basics

Chi-squared Test

χ^2 distribution

If Z_1, \dots, Z_k are **independent, standard normal** random variables, then the sum of their squares, $X = \sum_{i=1}^k Z_i^2$, is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as $X \sim \chi^2(k)$ or $X \sim \chi_k^2$. The chi-squared distribution has one parameter: a positive integer k that specifies the number of degrees of freedom (the number of random variables Z_i being summed).



The χ^2 test

- The χ^2 test ("chi-square test") is used to compare observed and expected frequencies. It can be used when there are 1 or 2 **nominal-scale** independent variables.

	Feature		
	Char. 1	...	Char. k
Observed	$f_{b,1}$		$f_{b,k}$
Expected	$f_{e,1}$		$f_{e,k}$

Examples:

- Do young and old people suffer from a certain disease with equal frequency?
- Do highly anxious and low-anxious people provide help in an emergency situation with equal frequency?

The χ^2 test

Prerequisites for the χ^2 test (rules of thumb)

- (1) Less than $\frac{1}{5}$ of all cells have an expected frequency less than 5.
- (2) No cell has an expected frequency less than 1.

If requirements are not met \rightarrow other tests

The χ^2 test

χ^2 test – Example 1

- The aim is to test whether the distribution of men and women in a group deviates significantly from an equal distribution.
- $N = 76$ (women: 56; men: 20)
- Statistical hypotheses
 - $H_0 : \pi(\text{woman}) = \pi(\text{man})$
 - $H_1 : \pi(\text{woman}) \neq \pi(\text{man})$

$\pi(x) =$ Relative frequency with which feature value x occurs

The χ^2 test

Step 1:

- First, the expected frequencies according to the H_0 are calculated:
- Observed: $N_F = 56$; $N_M = 20$
- Expected: ???
 - Total number: 76
 - With an even distribution, men and women would be expected.

The χ^2 test

Step 2:

- Now the (empirical) χ^2 value is calculated:

$$\chi_{df=k-1}^2 = \sum_{i=1}^k \frac{(f_{b,i} - f_{e,i})^2}{f_{e,i}}$$

with:

- k : Number of levels of the two variables
- $f_{b,i}$: Observed frequency in the cell (i)
- $f_{e,i}$: Expected frequency in cell (i)

	Feature	
	Char. 1 ...	Char. k
Observed	$f_{b,1}$	$f_{b,k}$
Expected	$f_{e,1}$	$f_{e,k}$

The χ^2 test

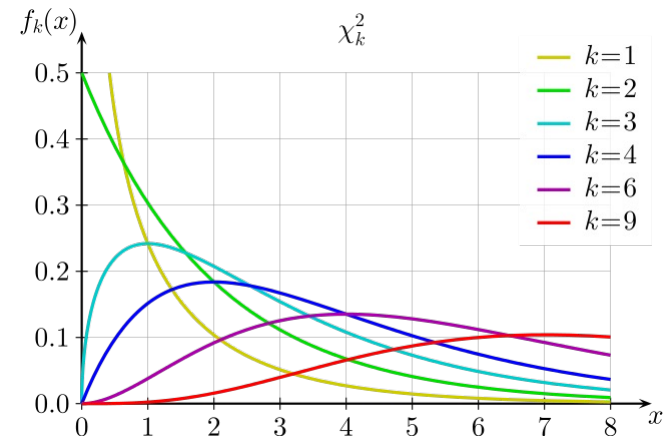
	Gender		
	Female	Male	
Observed	56	20	76
Expected	38	38	76

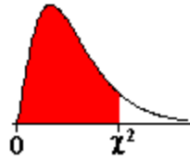
$$\chi_{df=k-1}^2 = \sum_{i=1}^k \frac{(f_{b,i} - f_{e,i})^2}{f_{e,i}}$$

$$\chi_{df=1}^2 = \frac{(56-38)^2}{38} + \frac{(20-38)^2}{38} = \frac{18^2}{38} + \frac{(-18)^2}{38} = 8.53 + 8.53 = 17.05$$

The χ^2 test

- Step 3: Compare the empirical χ^2 value with the critical χ^2 value.
- The critical χ^2 value depends on the degrees of freedom $k - 1$ and the selected α -level from a table for the χ^2 distribution





χ^2 table

Let's look for the χ^2 value, which covers 95% of all possible values of a χ^2 distributed random variable X^2 with a degree of freedom of 1. In the row for df-1, you will find the value $\chi^2 = 3.84$ in column $1-\alpha = 0.95$.

df	(red/dark) Area $(1 - \alpha)$								
	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84
4	4,88	5,39	5,99	6,74	7,78	9,49	11,14	13,28	14,86
5	6,06	6,63	7,29	8,12	9,24	11,07	12,83	15,09	16,75
6	7,23	7,84	8,56	9,45	10,64	12,59	14,45	16,81	18,55
7	8,38	9,04	9,80	10,75	12,02	14,07	16,01	18,48	20,28
8	9,52	10,22	11,03	12,03	13,36	15,51	17,53	20,09	21,95
9	10,66	11,39	12,24	13,29	14,68	16,92	19,02	21,67	23,59
10	11,78	12,55	13,44	14,53	15,99	18,31	20,48	23,21	25,19

The χ^2 test

- Step 3: Compare the empirical χ^2 value with the critical χ^2 value.
- For $\alpha=.05$, the result for $df=1$ is:

$$\chi_{emp}^2 = 17.05$$

$$\chi_{krit}^2 = 3.84$$

- The H_0 must be rejected; consequently, a difference can be proven.

The χ^2 test

χ^2 test – Example 2

Salary	Gender		
	Female	Male	
Low	23	14	37
High	35	6	41
	58	20	78

- Question: Is the relative frequency of high and low salaries the same for men and women?
- Statistical hypotheses
 - $H_0 : \pi(\text{salary} \mid \text{female}) = \pi(\text{salary} \mid \text{male})$
 - $H_1 : \pi(\text{salary} \mid \text{female}) \neq \pi(\text{salary} \mid \text{male})$

$\pi(x|y) =$ Relative frequency with which feature value x occurs when
 feature value y occurs

The χ^2 test

Step 1: First, the expected frequencies according to the H_0 are estimated from the marginal totals:

Observed:

Salary	Gender		
	Female	Male	
Low	23	14	37
High	35	6	41
	58	20	78

Expected:

Salary	Gender		
	Female	Male	
Low	27	10	37
High	31	10	41
	58	20	78

$$\begin{aligned}
 f_{e(i,j)} &= \frac{f_{b(i)} \cdot f_{b(j)}}{N} \cdot N \\
 &= \frac{f_{b(i)} \cdot f_{b(j)}}{N}
 \end{aligned}$$

$$\begin{aligned}
 b(1, \cdot) &= 37 & b(\cdot, 1) &= 58 \\
 N &= 78 \\
 (37 \cdot 58) / 78 &= 27
 \end{aligned}$$

The χ^2 test

Step 2: Now the (empirical) χ^2 value is calculated:

$$\chi^2_{df=(k-1)\cdot(l-1)} = \sum_{i=1}^k \sum_{j=1}^l \frac{(f_{b(i,j)} - f_{e(i,j)})^2}{f_{e(i,j)}}$$

with:

- k, l : Number of levels of the two variables
- $f_{b(i,j)}$: Observed frequency in the cell (i,j)
- $f_{e(i,j)}$: Expected frequency in cell (i,j)

The χ^2 test

Observed:

Salary	Gender		
	Female	Male	
Low	23	14	37
high	35	6	41
	58	20	78

Expected:

Salary	Gender		
	Female	Male	
Low	27	10	37
high	31	10	41
	58	20	78

$$\chi_{df=(k-1)(l-1)}^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(f_{b(i,j)} - f_{e(i,j)})^2}{f_{e(i,j)}}$$

$$\begin{aligned} \chi_{df=1}^2 &= \frac{(23-27)^2}{27} + \frac{(35-31)^2}{31} + \frac{(14-10)^2}{10} + \frac{(6-10)^2}{10} \\ &= 0,59 + 0,51 + 1,60 + 1,60 = 4,30 \end{aligned}$$

The χ^2 test

- Step 3: Compare the empirical χ^2 value with the critical χ^2 value.

df	(red/dark) Area ($1 - \alpha$)								
	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84

- For $\alpha=.05$, the result for $df=1$ is:

$$\chi_{emp}^2 = 4.30$$

$$\chi_{krit}^2 = 3.84$$

- The H_0 must be rejected; consequently, a difference can be proven.

Summary

- Non-parametric test procedures can be used if
 - a) the available data do not have an interval scale level or
 - b) the normal distribution assumption of parametric tests is violated.
- The χ^2 test checks whether observed and expected frequencies differ significantly from each other.

Statistics as an Important Science

- A great many statistical tests have been developed
- We can only shed light on the tip of the iceberg here (lifelong learning is necessary)
- Without statistical knowledge, one is lost in computer science
 - Common mistakes:
 - Only calculating mean values without also specifying variances
 - Using threshold values instead of (model-based) hypothesis tests