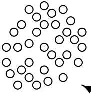
Statistics and Sensor Data Fusion

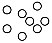
4. Inductive Statistics

Interplay Between Descriptive Statistics, Probability Calculus and Inductive Statistics

Probability Calculus

Goal: Mathematical treatment of random phenomena

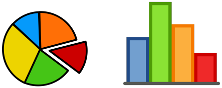




Data Sample

Descriptive Statistics

Goal: Suitable graphical and numerical re-presentation of the data



Inductive Statistics

Goal: To infer properties of an underlying population

Inductive Statistics

$\boxed{\textbf{Goal:}\ \text{To infer properties of an underlying population}}$

4.1 Basic Concepts of Inductive Statistics

Basic Concepts of Inductive Statistics

Background:

- Inductive statistics or statistical inference is the process of using data analysis to **infer properties** of an underlying probability distribution which describes a whole population.
- It is therefore applied if parameters or properties of a population are **unknown** and cannot be computed or determined directly.
- Information required to describe the properties of a population are for example
  - the **type** of the distribution, e.g. *normal distribution* $\mathcal{N}(\mu, \sigma^2)$
  - the **localization**, e.g. the value of the *mean* μ
  - the **dispersion**, e.g. the value of the *variance* σ^2

Basic Concepts of Inductive Statistics

Approach:

- The approach of inductive statistics is to draw conclusions about the population of interest based on **spot checks** (i.e. small subsets of the population).
- Conclusions based on the data sample provided by the spot check can be erroneous, because the sample at hand represents only a fraction of the population.
- Inferential statistics makes use of **probability theory** to assess the quality of the drawn conclusions based on the spot check or data sample.
- The **two main techniques** in inferential statistics are
  - Estimation
  - Testing (not covered in this course)

Basic Concepts of Inductive Statistics

The technique of estimation can be subdivided into two branches:

Parameter Estimation vs. Interval Estimation:

Based on a random sample of the underlying probability distribution

- ...an unknown **parameter** shall be estimated

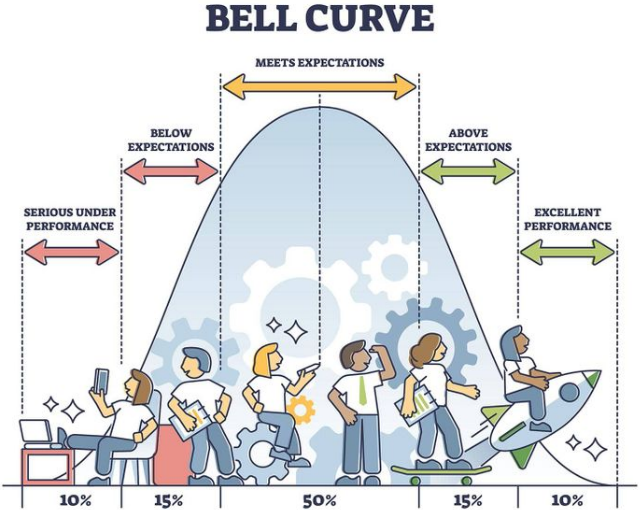
  $\rightarrow$ parameter estimation (Section 4.3)

- ...an **interval** shall be computed such that an unknown parameter lies within this interval with high probability

  $\rightarrow$ interval estimation (Section 4.4)

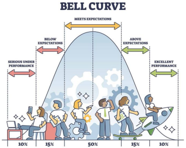
4.2 Central Limit Theorem

Central Limit Theorem



https://www.simplypsychology.org/normal-distribution.html

Central Limit Theorem



Central Limit Theorem – In a Nutshell:

Assumptions:

1. The (discrete or continuous) random variables $X_1, \ldots, X_n$ are **statistically independent** with mean $\mu_1, \ldots, \mu_n$ and variance $\sigma_1^2, \ldots, \sigma_n^2$, respectively.

2. The random variables $X_1, \ldots, X_n$ are either **identically** or **"nearly identically"** distributed (more specifically, certain constraints according to Ljapunow or Lindeberg hold).

Then it holds for large n:

The random variable X given by the **sum**

$$X = X_1 + \ldots + X_n$$

is approximately normally distributed, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$, where

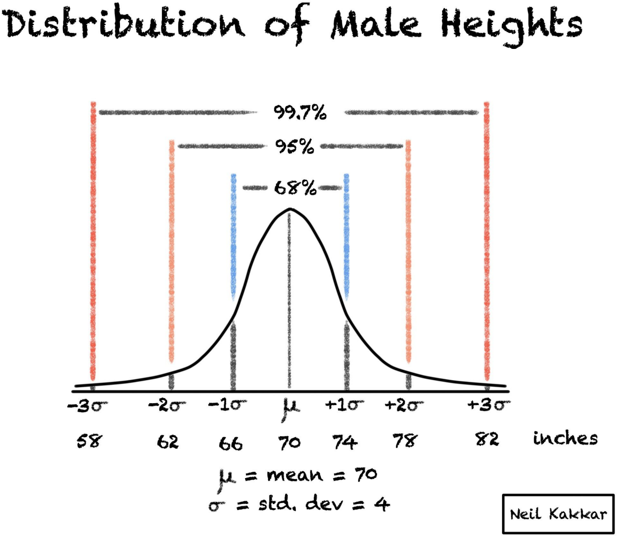
$$\mu = \mu_1 + \ldots + \mu_n \quad \text{and} \quad \sigma^2 = \sigma_1^2 + \ldots + \sigma_n^2$$

Central Limit Theorem

Central Limit Theorem – Some Examples:

- The **daily turnovers** of 30 logistic hubs of comparable size are distributed according to some unknown probability distribution. The **total turnover** is then approximately normally distributed.
- The **water consumption of a provincial town** per year is the sum of many individual consumptions that are usually not dominated by only a few households. Therefore the total water consumption is approximately normally distributed.
- The **monthly payoff of a big insurance company** is usually the sum of many individual payoffs and therefore approximately normally distributed.
- Many **biological phenomena** are influenced by a large sum of different factors and are therefore normally distributed.

Central Limit Theorem



https://www.freecodecamp.org/news/normal-distribution-explained/

Central Limit Theorem

Central Limit Theorem – Example:

- We consider the random experiment of **tossing a die** and performing n independent repetitions.
- The random variable X should count the total number of occurrences of the number "6" during these n repetitions.
- In consequence, we obtain $X \sim \mathrm{Bin}\left(n, \frac{1}{6}\right)$ with

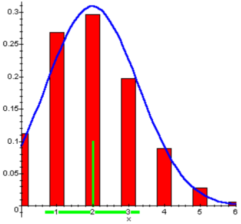
$$\mu = n \cdot p = \frac{n}{6} \quad \text{and} \quad \sigma^2 = n \cdot p \cdot (1 - p) = \frac{5n}{36}$$

- Because of the **central limit theorem**, for large n the random variable X is approximately normally distributed according to

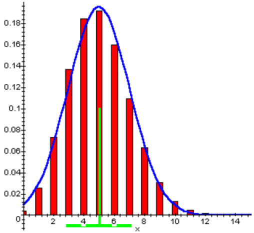
$$X \sim \mathcal{N}\left(\frac{n}{6}, \frac{5n}{36}\right)$$

Central Limit Theorem

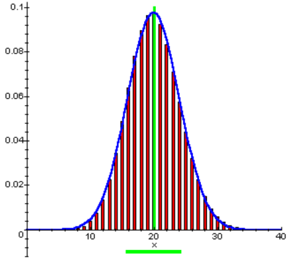
Central Limit Theorem – Example:



n=12



n=30



n=120

Approximation of the binomial distribution $\mathrm{Bin}\left(n, \frac{1}{6}\right)$ by the normal distribution $\mathcal{N}\left(\frac{n}{6}, \frac{5n}{36}\right)$ for increasing values of n.

4.3 Parameter Estimation

Parameter Estimation

Parameter Estimation:

In parameter estimation, one distinguishes between the estimate of an **unknown parameter** of a probability distribution based on a data sample, and the corresponding estimator or estimator function:

- A reasonable **estimate** for the mean or expectation value is

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

- The corresponding **estimator** is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{random variable}$$

Since the estimator $\bar{X}$ is built on the **random sample** $X_1, \ldots, X_n$ consisting of n i.i.d. random variables, it is itself a random variable.

Parameter Estimation

Two Important Estimators:

Parameter	Applied Estimator	Formula
Mean μ	Sample Mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance σ^2	Sample Variance	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

The bias of an estimator is the difference between the estimator's expected value and the true value of the parameter being estimated:

- An estimator with zero bias is called unbiased.
- Sample mean $\bar{X}$ and sample variance S^2 are **unbiased** since

$$E(\bar{X}) = \mu \quad \text{and} \quad E(S^2) = \sigma^2$$

Parameter Estimation

Exercise:

For civil speed control in a provincial town, there was a spot check of the velocity of cars driving through the city during lunch time.

From all the cars driving through town, five measurements were made leading to the five values (in km/h):

$$42, 31, 47, 44, 36$$

Calculate an estimate for the mean and the variance of the underlying probability distribution of car speeds.

Parameter Estimation

Parameter Estimation vs. Interval Estimation:

- Parameter or point estimation provides an **estimate** for an unknown parameter of an underlying probability distribution (e.g. expectation value μ or variance σ^2). However, point estimators do not provide an assessment of their accuracy.
- In turn, in interval estimation a **confidence interval** is computed which contains the true value of a population parameter with a **specified probability**.
- For the consistent construction of the confidence interval, the **distribution** of the estimator has to be known (at least approximately).
- We expect that there will be a **tradeoff** between the required probability and the size of the confidence interval.

Parameter Estimation

Central Limit Theorem Applied to the Sample Mean:

For i.i.d. random variables $X_1, \ldots, X_n$, the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an **estimator** for the unknown mean $\mu = E(X_1)$. It has the expectation value

$$E(\bar{X}) = E\left(\tfrac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n} \cdot n \cdot E(X_1) = \mu$$

and the variance

$$V(\bar{X}) = V\left(\tfrac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2} \cdot n \cdot V(X_1) = \frac{\sigma^2}{n}$$

Due to the **central limit theorem**, for large n the sample mean $\bar{X}$ is approximately normally distributed according to

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

4.4 Interval Estimation

Interval Estimation

Concept of a Confidence Interval:

For an **unknown parameter** $\Theta \in \mathbb{R}$ of an underlying probability distribution and a given level of significance α, we construct

- the lower confidence level $L = L(X_1, \ldots, X_n)$
- the upper confidence level $U = U(X_1, \ldots, X_n)$

with respect to a **random sample** $X_1, \ldots, X_n$, such that it holds

$$\boxed{P(L \leq \Theta \leq U) = 1 - \alpha}$$

With the introduced **random variables** L and U, the obtained random interval $[L, U]$ is called the confidence interval of the parameter $\Theta \in \mathbb{R}$ at the specified confidence level of $1 - \alpha$.

That means that the unknown parameter $\Theta \in \mathbb{R}$ should be covered by the confidence interval $[L, U]$ with the probability of $1 - \alpha$.

Interval Estimation

Confidence Interval – Interpretation:

- A **realization** $x_1, \dots, x_n$ of the random sample $X_1, \dots, X_n$ generates a specific confidence interval

  $$[L(x_1, \dots, x_n), U(x_1, \dots, x_n)] \subset \mathbb{R}$$

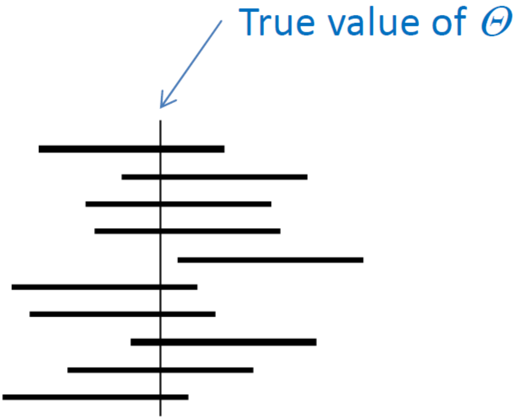
  where $L(x_1, \dots, x_n)$ and $U(x_1, \dots, x_n)$ are **realizations** of the random variables L and U.

- In turn, different realizations of the random sample will usually generate different specific confidence intervals.

- A level of significance $\alpha = 0.1$ means that on average 90% of all generated specific confidence intervals will **contain** the unknown parameter $\Theta \in \mathbb{R}$.

Interval Estimation

Confidence Interval – Graphical Illustration:



How to construct a confidence interval for a given confidence level?

Interval Estimation

For the construction of confidence intervals, we have to introduce the quantiles of a probability distribution:

Quantiles of a Probability Distribution:

For a (discrete or continuous) probability distribution characterized by the **cumulative distribution function** $F(x)$ and a specified level $\alpha \in (0, 1)$, the value $x_\alpha \in \mathbb{R}$ which fulfills the equation

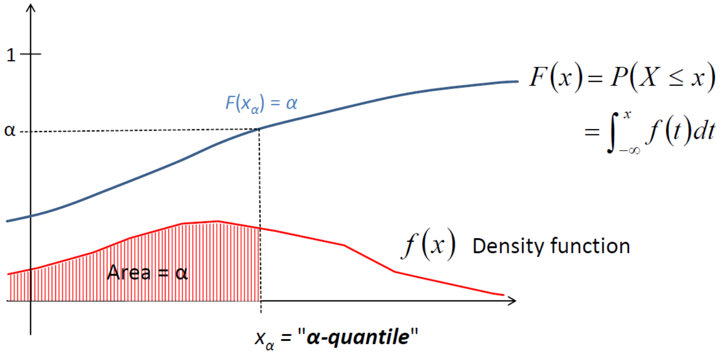
$$\boxed{F(x_\alpha) = P(X \leq x_\alpha) = \alpha}$$

is called α-quantile.

For the determination of α-quantiles, the type of the cumulative distribution function $F(x)$ must be known (at least approximately).

Interval Estimation

Quantiles of a Probability Distribution:



- A fraction of $\alpha \cdot 100\%$ of the probability **"lies left"** of x_α
- Vice versa, $(1 - \alpha) \cdot 100\%$ of the probability **"lies right"** of x_α
- In particular, $x_{0.5}$ is the median of the probability distribution

Interval Estimation

Quantiles of the Standard Normal Distribution:

- As a consequence of the **central limit theorem**, frequently a normal distribution can be assumed.
- Quantiles of the standard normal distribution $\mathcal{N}(0,1)$ can easily be retrieved **from tables**.
- In order to indicate explicitly that we work with the standard normal distribution, we will use the symbol u_α instead of the general x_α for the α-quantile of $\mathcal{N}(0,1)$.
- For the α-quantiles u_α of the standard normal distribution, it holds that

$$F_0(u_\alpha) = P(X \leq u_\alpha) = \int_{-\infty}^{u_\alpha} f_0(s)\, ds = \int_{-\infty}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}}\, ds = \alpha$$

Interval Estimation

First we construct symmetric confidence intervals for the mean:

Symmetric Confidence Intervals for the Mean (I):

We assume a normally distributed random variable

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

with unknown mean μ and **known variance** σ^2.

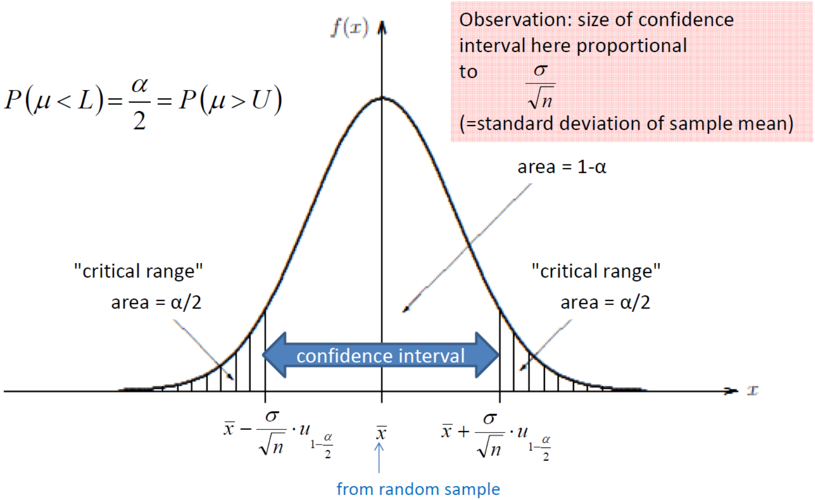
In this case, it holds for the **true value** $\mu \in \mathbb{R}$ that

$$P\Big(\underbrace{\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}}_{L} \leq \mu \leq \underbrace{\bar{X} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}}_{U}\Big) = 1 - \alpha$$

where $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ is the **sample mean** and $u_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$-quantile of the standard normal distribution.

Interval Estimation

Symmetric Confidence Intervals for the Mean (I):



Interval Estimation

Exercise:

The service times for $n = 12$ production orders in minutes were measured as follows:

$$514, 497, 508, 520, 497, 509, 520, 509, 503, 510, 497, 512$$

The service time is assumed to be normally distributed with a standard deviation of 9 minutes.

(a) Construct a symmetric confidence interval for the unknown mean $\mu \in \mathbb{R}$ at a level of significance $\alpha = 0.05$.

(b) How to choose the total number of measurements n if the confidence interval shall be reduced to a length of at most 8?

Interval Estimation

Now we want to construct symmetric confidence intervals for the mean in the case that also the variance is **unknown**:

Symmetric Confidence Intervals for the Mean (II):

In the case that also the variance σ^2 is unknown, it has to be replaced by its **estimate** s^2 obtained by the sample variance

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

The resulting probability distribution is the t-distribution of Student (i.e. William Gosset), where quantiles of the t-distribution with $n-1$ degrees of freedom are used. These quantiles can be found in **tables**, the α-quantile with n degrees of freedom is called $t_{(\alpha,n)}$.

By means of switching to the t-distribution, the decreasing accuracy caused by using the estimate s^2 instead of σ^2 is considered.

Interval Estimation

Some Background on the t-Distribution:

For large n, the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is approximately normally distributed according to

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

The **transformed random variable**

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

is distributed according to the standard normal distribution

$$Z \sim \mathcal{N}(0, 1)$$

with the α-quantiles u_α.

Interval Estimation

Some Background on the t-Distribution:

In the case that the variance σ^2 is unknown, we have to apply the sample variance

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

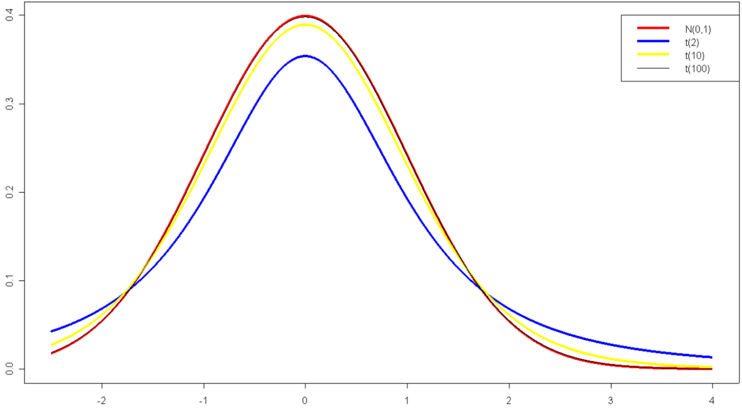
This results in the **new random variable**

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

which is distributed according to the t-distribution with $n-1$ degrees of freedom with the α-quantiles $t_{(\alpha,n-1)}$.

Interval Estimation

Standard Normal Distribution vs. t-Distribution:



Von Thomas Steiner - mit GNU R erzeugt (pt(x,dt=5)), Copyrighted free use,
https://commons.wikimedia.org/w/index.php?curid=36014931

Interval Estimation

Exercise:

The service times for $n = 12$ production orders in minutes were measured as follows (as in the previous exercise):

$$514, 497, 508, 520, 497, 509, 520, 509, 503, 510, 497, 512$$

The service time is assumed to be normally distributed with unknown variance.

Construct a symmetric confidence interval for the unknown mean $\mu \in \mathbb{R}$ at a level of significance $\alpha = 0.05$ and compare it to the confidence interval when the standard deviation is known ($\sigma = 9$).

Interval Estimation

A confidence interval has not to be centered symmetrically around the sample mean:

Non-Symmetric Confidence Intervals:

By using the confidence limits

$$L = -\infty \quad \text{or} \quad U = +\infty$$

we obtain the one-sided confidence intervals

$$(-\infty, U] \quad \text{or} \quad [L, +\infty)$$

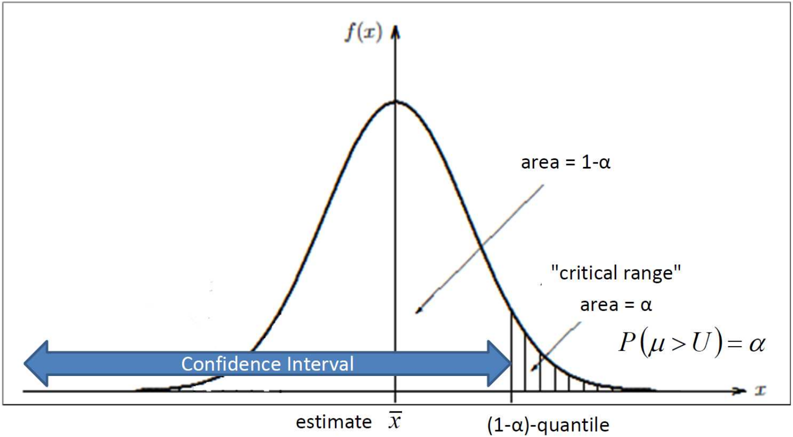
Because now the critical range is not divided by two anymore, the quantile for $1 - \alpha$ has to be chosen instead the one for $1 - \frac{\alpha}{2}$.

In the case of a normal distribution with known variance σ^2, the **lower confidence interval** for the mean $\mu \in \mathbb{R}$ is defined by

$$P\left(\mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha}\right) = 1 - \alpha$$

Interval Estimation

Lower Confidence Interval for the Mean:



Interval Estimation

We sum up **all possible cases** for one-sided confidence intervals for the mean both in the case of known and unknown variance:

One-Sided Confidence Intervals:

	Lower Confidence Interval	Upper Confidence Interval
σ known	$U = \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha}$	$L = \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha}$
σ unknown	$U = \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{(1-\alpha, n-1)}$	$L = \bar{x} - \frac{s}{\sqrt{n}} \cdot t_{(1-\alpha, n-1)}$

- Here, the expressions $\bar{x}$ and s denote the observed **realizations** of the sample mean $\bar{X}$ and the sample standard deviation S.
- The quantiles $u_{1-\alpha}$ of the standard normal distribution and $t_{(1-\alpha, n-1)}$ of the t-distribution can be found in tables.

Interval Estimation

Exercise:

Using the data of the previous exercise, construct a lower and an upper confidence interval for the mean $\mu \in \mathbb{R}$ in the case that the variance is unknown at a level of significance $\alpha = 0.05$.

4.5 Multivariate Parameter Estimation

Multivariate Parameter Estimation

In the case of **multivariate distributions**, we have to consider multivariate parameter estimation:

Maximum Likelihood Parameter Estimation for the Multivariate Normal Distribution:

We consider random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ according to

$$\boldsymbol{X}_i \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n$$

Both the mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ of $\mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are assumed to be unknown.

Based on observed realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the unknown parameters $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ should be estimated by maximizing the so-called likelihood function $\mathcal{L}$:

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i) \to \max_{\boldsymbol{\Theta}}!$$

Multivariate Parameter Estimation

Maximum Likelihood Parameter Estimation for the Multivariate Normal Distribution:

The maximum likelihood estimators for the mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ in the case of normally distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i.i.d. are

$$\hat{\boldsymbol{\mu}}_{ML}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$$

and

$$\hat{\boldsymbol{\Sigma}}_{ML}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{n} \boldsymbol{S} = \frac{1}{n} \underbrace{\sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T}_{\text{scatter matrix } \boldsymbol{S}}$$

Based on the realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the maximum likelihood estimates are given by $\hat{\boldsymbol{\mu}}_{ML}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\hat{\boldsymbol{\Sigma}}_{ML}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

Multivariate Parameter Estimation

Maximum Likelihood Parameter Estimation for the Multivariate Normal Distribution – Practical Aspects:

The estimation of the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ might lead to a singular matrix $\hat{\boldsymbol{\Sigma}}_{ML}$ with

$$\det(\hat{\boldsymbol{\Sigma}}_{ML}) = 0$$

The estimated matrix $\hat{\boldsymbol{\Sigma}}_{ML}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is built from n vectors $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$, from which due to

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

at most $n - 1$ are **linearly independent**. Therefore, the estimated matrix $\hat{\boldsymbol{\Sigma}}_{ML}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is **definitely singular** if $n \leq D$.

Multivariate Parameter Estimation

Unbiased Estimators:

The maximum likelihood estimator $\hat{\boldsymbol{\mu}}_{ML}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ for the mean vector $\boldsymbol{\mu}$ is unbiased, i.e.

$$E\big(\hat{\boldsymbol{\mu}}_{ML}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\big) = \boldsymbol{\mu}$$

In contrast, the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}_{ML}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ for the covariance matrix $\boldsymbol{\Sigma}$ is not unbiased.

An unbiased estimator for the covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{n-1}\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$$

With this definition, it holds that

$$E\big(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\big) = \boldsymbol{\Sigma}$$

Multivariate Parameter Estimation

Exercise:

Consider the observed realizations

$$\boldsymbol{x}_1 = \begin{pmatrix} 7 \\ 6 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \quad \boldsymbol{x}_3 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

of a bivariate random vector $\boldsymbol{X} = (X_1, X_2)^T$ and determine the unbiased estimates of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the underlying probability distribution.