Statistical Methods of Data Analysis

W. Waltenberger
wolfgang.waltenberger@gmail.com

based on the slides of the year 2018 von R. Frühwirth

VU 260014 (Uni) / SV 142.340 (TU)

October 19, 2023

Syllabus

Part 1: Descriptive Statistics

Part 2: Probabilities

Part 3: Random Variables and Distributions

Part 4: Point Estimators

Syllabus

Part 5: Confidence Intervals

Part 6: Testing Hypotheses

Part 7: Regression and Linear Models

Part 8: Bayes Statistics

Part I

Descriptive Statistics

Overview Part 1

Section 1: Introduction

Subsection: Basic Terminology

Basic Terminology

Definition of Statistics

- The collection and storage of data, e.g. by statistical offices.
- The mathematical analysis of data, e.g. the calculation of measures and ratios, the estimation of unknown parameters, the testing of hypotheses

Descriptive Statistics

- Description of existing data by measures, tables, graphs.

Basic Terminology

Inductive (or Inferential) Statistics

- Investigation of regularities and causes that are behind the data and (partially) explain the data.
- Exploratory data analysis: the goal is to generate hypotheses for theorizing. gain
- Confirmatory data analysis: aim is to test existing theories, e.g. by estimating parameters or testing hypotheses.

Subsection: Features and Scale Types

Features and Scale Types

Qualitative Features

- **binary** (yes/no). Example: EU citizenship.
- **categorical** (classification). Example: single/divorced/married/widowed.
- **ordinal** (rank). Example: Grades 1–5.

Quantitative Features

- **discrete** (integer). Example: counting.
- **continuous** (real-valued). Example: measuring process.

Features and Scale Types

Scale Types

- **Nominal Scale**: Numerical values are only designations for mutually exclusive categories.
- **Ordinal Scale**: Order of numbers is essential.
- **Interval Scale**: Order and differences between values are interpretable in a meaningful way, the zero point is arbitrarily fixed.
- **Ratio Scale**: Order, differences and ratios are meaningfully interpretable, there is an absolute zero point.

Features and Scale Types

Example

- The marital status of a person is coded by numbers (1=single, 2=married, 3=divorced, 4=widowed). Nominal scale.
- A team's standing in the championship is indicated by its rank in the league. Ordinal scale.
- The years (2007, 2008, ...) form an interval scale, since the zero point is arbitrary.
- The Celsius scale of temperature is an interval scale, since the zero point is arbitrarily fixed.
- The Kelvin scale of temperature is a ratio scale, since the zero point is physically fixed.
- The height of a person is given in cm. A ratio scale, the zero value is natural.

Features and Scale Types

Example

The following data matrix D compiles several features of eight individuals.

Number	gender	age	education
1	1	34	2
2	2	54	1
3	2	46	3
4	1	27	4
5	1	38	2
6	1	31	3
7	2	48	4
8	2	51	2

Gender: 1=F, 2=M, Age: in years. Education: 1=compulsory school, 2=high school, 3=bachelor's, 4=master's.

Subsection: Statements and Frequencies

Statements and Frequencies

The Concept of a Statement

- We call the set of objects of study the **population** Ω.
- A **statement** $A(x)$ is a statement about properties of the elements $x \in \Omega$.
- For each $x \in \Omega$, $A(x)$ must be either **true** or **false**.

Example

Let $A(x)$ be the statement "x is female". Then $A(1)$ is true and $A(2)$ is false.

Example

Let $B(x)$ be the statement "x is over 50 years old". Then $B(2)$ is true and $B(6)$ is false.

Statements and Frequencies

Linking Statements

Let A and B be two statements.

Symbol	Name	Meaning
$A \cup B$	disjunction	A or B (or both)
$A \cap B$	conjunction	A and B (both A and B)
A'	negation	not A (the opposite of A)
$A \subseteq B$	implication	from A follows B

Statements and Frequencies

Definition (Absolute Frequency)

Let $A(x)$ be a statement about $x \in \Omega$. The **absolute frequency** $h(A)$ of A is the number of elements of Ω for which $A(x)$ is true.

Example

If $A(x)$ is the statement "The person $x \in D$ has at least a bachelor's degree". Then $h(A) = 4$.

Statements and Frequencies

Definition (Relative Frequency)

Let $A(x)$ be a statement about $x \in \Omega$. The **relative frequency** $f(A) = h(A)/n$ of A is the absolute frequency $h(A)$ divided by the total number $n = |\Omega| \in \mathbb{N}$ of elements.

Example

If $B(x)$ is the statement "The person $x \in D$ is older than thirty years". Then $f(B) = 7/8$.

Statements and Frequencies

Special Statements

- $A = \emptyset$: A is never true, $h(A) = f(A) = 0$.
- $A = \Omega$: A is always true, $h(A) = n, f(A) = 1$.

Arithemetic Laws for Frequencies

Addition Law

$$A \cap B = \emptyset \implies \begin{cases} h(A \cup B) = h(A) + h(B) \\ f(A \cup B) = f(A) + f(B) \end{cases}$$

Statements and Frequencies

Sieve Formula

$$h(A \cup B) = h(A) + h(B) - h(A \cap B)$$

Example

33% of a bank's customers have a home loan, 24% have a loan to finance consumer goods, 11% have both. What is the percentage of customers who have neither a housing loan nor a consumer goods loan?

Section 2: One-dimensional Features

1 Introduction

2 One-dimensional Features
- Graphical Representation
- Kernel Density Estimator
- Empirical Measures
- Boxplot and Empirical Distribution Function
- Examples

3 Two-dimensional Features

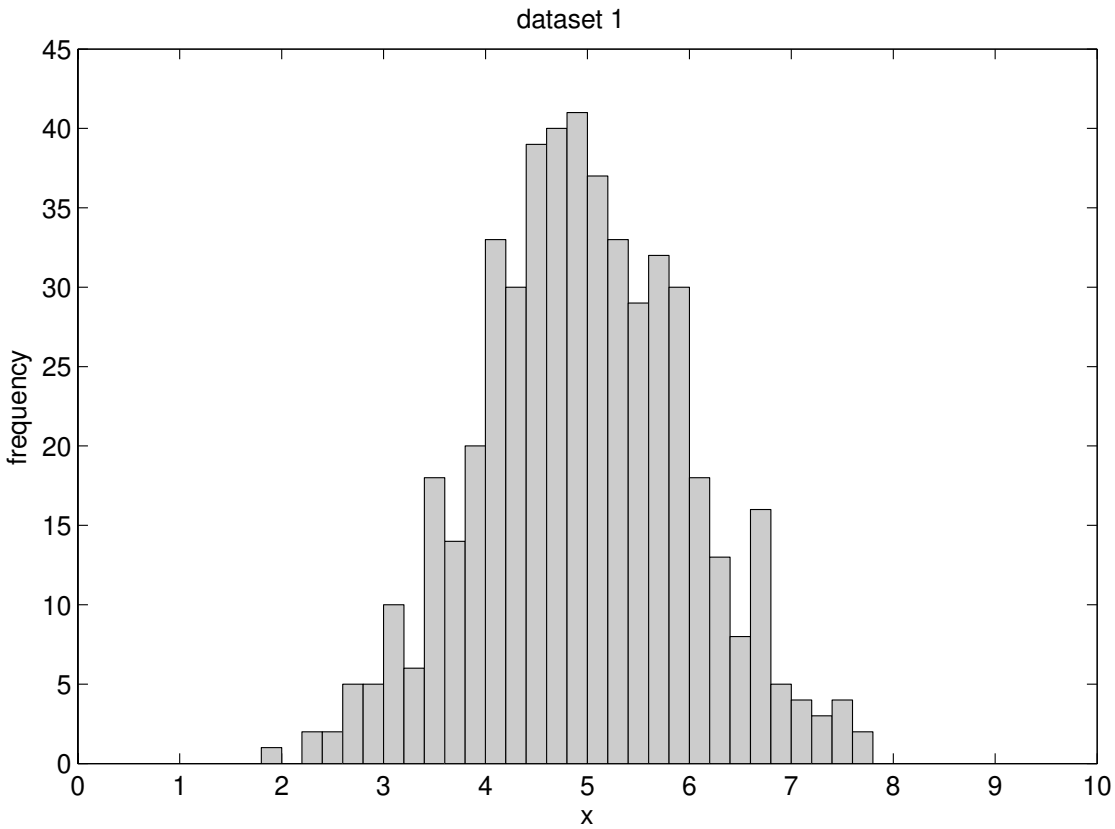
Subsection: Graphical Representation

Graphical Representation

- A picture is worth a thousand words!
- Graphical representations of data sets are therefore extremely popular and useful.
- Qualitative variable: frequency table, pie chart, bar chart.
- Quantitative variable: grouped frequency table, histogram, boxplot, empirical distribution function

Graphical Representation

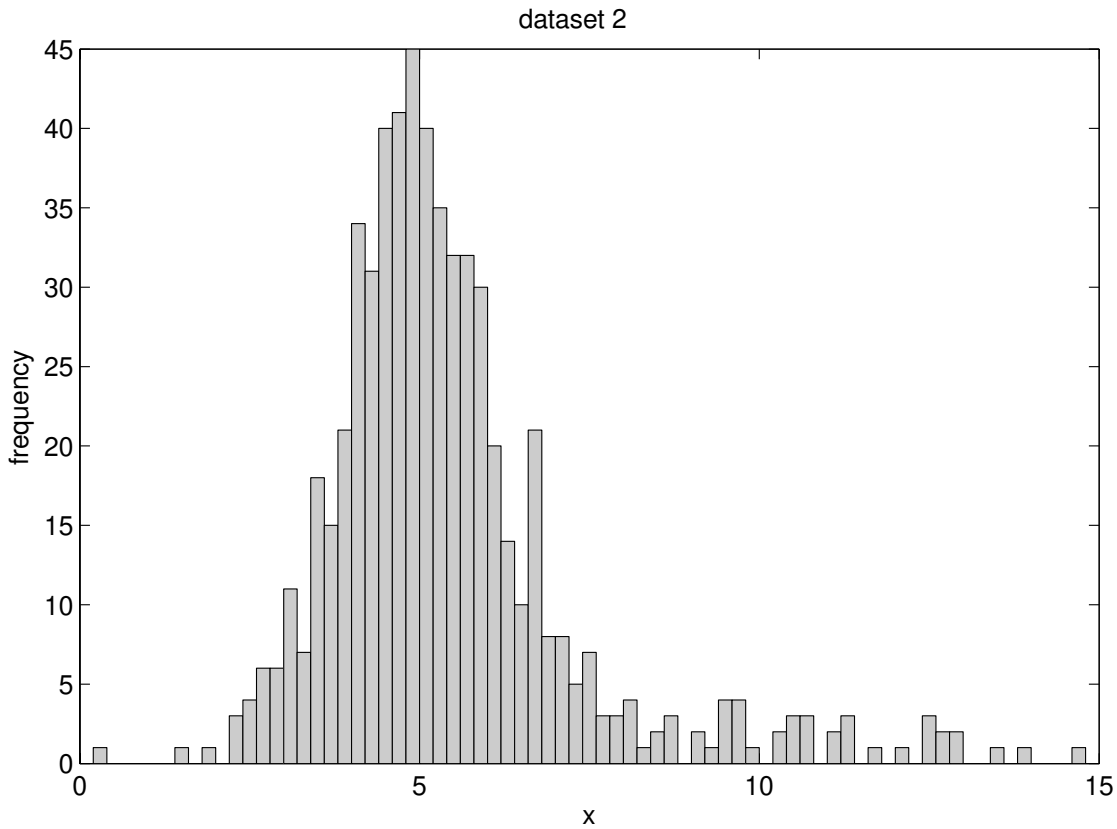
- Dataset 1 (500 normally distributed values):



histogram

Graphical Representation

- record 2 = record 1 + contamination (100 values):



histogram

Graphical Representation

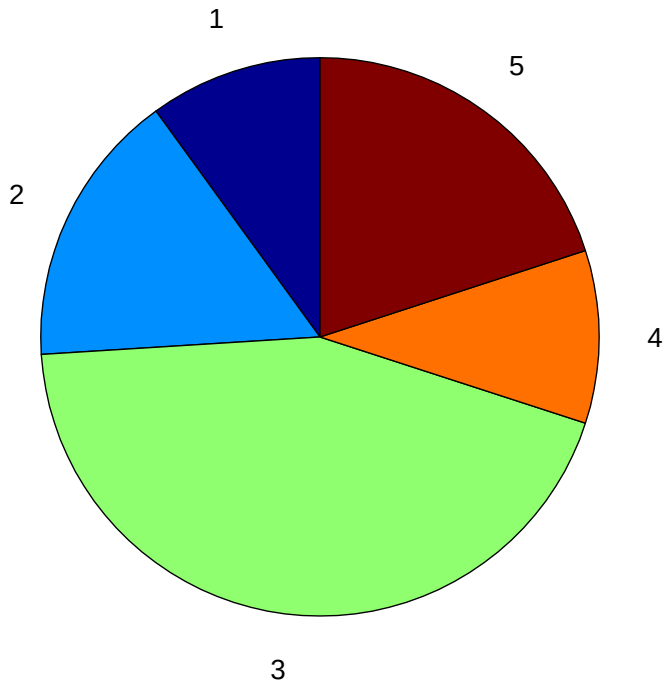
- Dataset 3 (50 exam notes):

Grade k	$h(k)$	$f(k)$
1	5	0.10
2	8	0.16
3	22	0.44
4	5	0.10
5	10	0.20
	50	1.00

frequency table

Graphical Representation

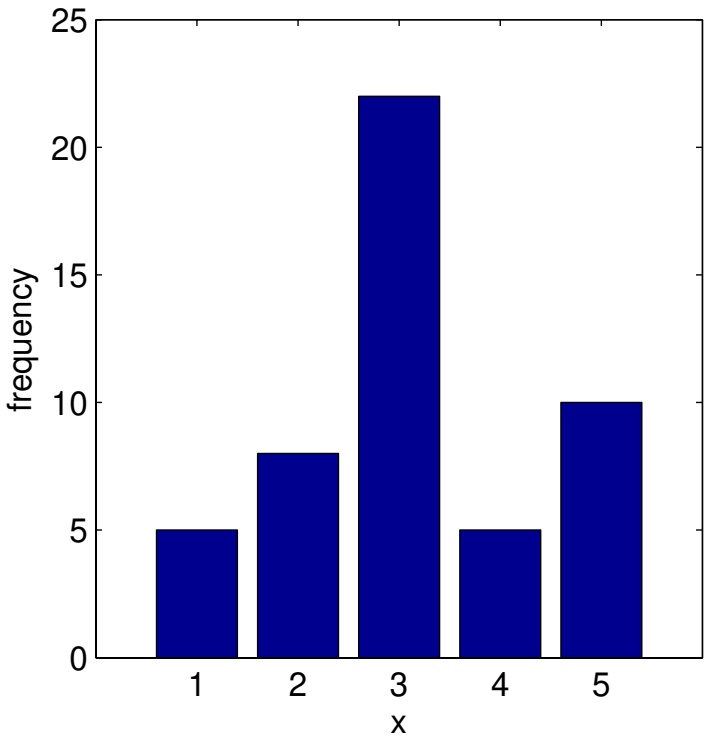
- Dataset 3 (50 exam notes):



pie chart

Graphical Representation

- Dataset 3 (50 exam notes):



bar chart

Subsection: Kernel Density Estimator

1 Introduction

2 One-dimensional Features

- Graphical Representation
- Kernel Density Estimator
- Empirical Measures
- Boxplot and Empirical Distribution Function
- Examples

3 Two-dimensional Features

Kernel Density Estimator

- The frequency distribution (histogram) can be smoothed with a kernel density estimator.
- The density of the observed feature is thereby approximated by a sum of kernels $K(\cdot)$:

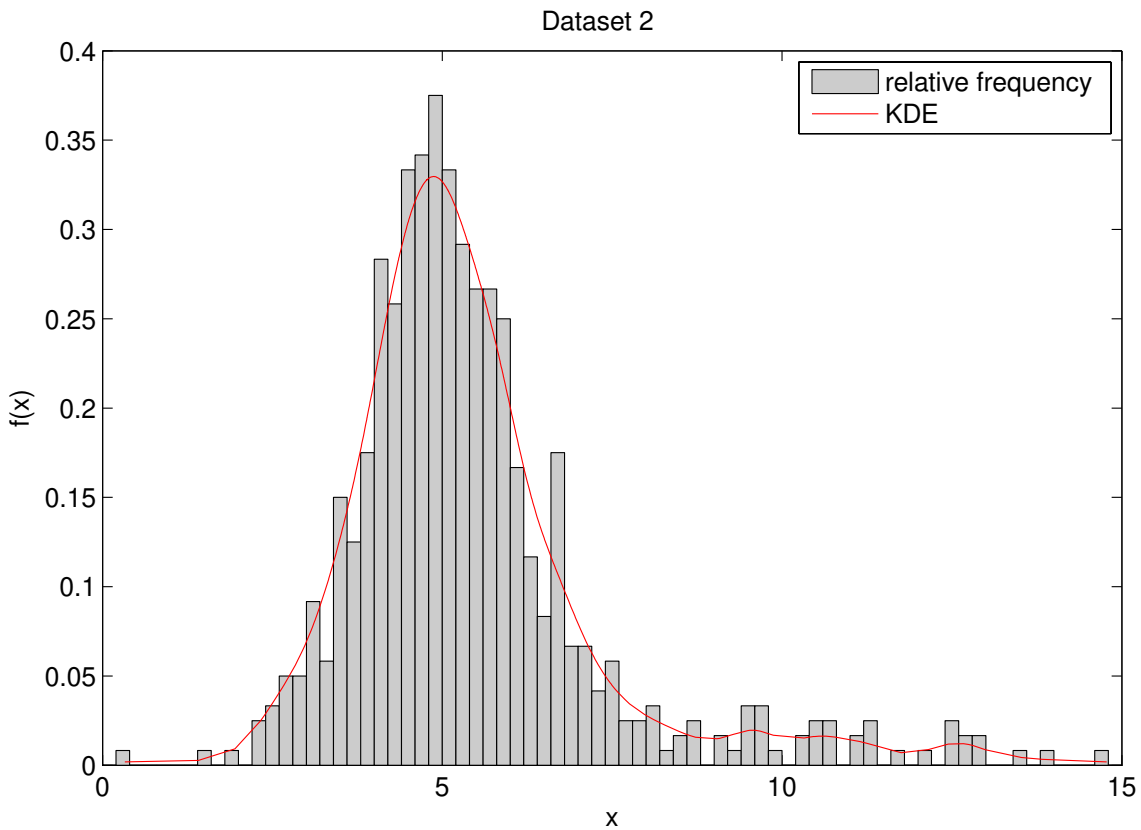
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- h is the bandwidth of the kernel density estimator.
- A popular kernel is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Kernel Density Estimator

- record 2:



Smoothing of histogram by kernel density estimators.

PYTHON: `sklearn.neighbors.KernelDensity`

Subsection: Empirical Measures

1 Introduction

2 One-dimensional Features

- Graphical Representation
- Kernel Density Estimator
- Empirical Measures
- Boxplot and Empirical Distribution Function
- Examples

3 Two-dimensional Features

Empirical Measures

- Data lists are often extensive enough that their content is to be summarized by a few empirical measures. Which measures are useful depends on the type of feature.
- Some measures start from the ordered data list $x_{(1)}, \ldots, x_{(n)}$.
- We distinguish between measures of location, dispersion, and skewness.
- A measure of location (or position) indicates the value around which the data is concentrated.
- A measure of dispersion indicates how large the fluctuations of the data are around their central value.
- A skewness measure indicates how symmetrical the data are about their central value.

Empirical Measures

Measures of Location

Definition (Positional Measure)

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a data list. The function $\ell(\boldsymbol{x})$ is called a position measure for $\boldsymbol{x}$ if the following holds true:

- $\ell(a\boldsymbol{x} + b) = a\ell(\boldsymbol{x}) + b$
- $\min(\boldsymbol{x}) \leq \ell(\boldsymbol{x}) \leq \max(\boldsymbol{x})$

- Meaningful position measures indicate the "typical" or "central" value of the data list.
- Different position measures are useful depending on the scale.

Empirical Measures

The Mean Value

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

- Meaningful for ordinal, interval and ratio scales.
- The mean minimizes the following function:

$$\bar{x} = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - x)^2$$

PYTHON: `xbar=numpy.mean(x)`

Empirical Measures

Median of Data

$$\tilde{x} = x_{(n/2)}$$

- The median divides the ordered data list into two equal parts.
- Useful for ordinal, interval and ratio scales.
- The median minimizes the following function:

$$\tilde{x} = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - x|$$

PYTHON: `xmed=numpy.median(x)`

Empirical Measures

- The median is a special case of a more general term, the quantile.

α-quantile

$$Q_\alpha = x_{(\alpha n)}$$

- The α-quantile divides the ordered data list in the ratio $\alpha : 1 - \alpha$.
- Meaningful for ordinal, interval, and ratio scales.

PYTHON: `qa=numpy.quantile(x,alpha)`

- Q_0 is the smallest value, Q_1 is the largest value in the data list. $Q_{0.5}$ is the median.
- The five **quartiles** $Q_0, Q_{0.25}, Q_{0.5}, Q_{0.75}, Q_1$ form the **five point summary** of the data list.

PYTHON: `fps=numpy.quantile(x,[0,0.25,0.5,0.75,1])`

Empirical Measures

LMS (Least Median of Squares)

LMS value is the midpoint of the shortest interval containing $h = \lfloor n/2 \rfloor + 1$ data points

.

- The LMS value is extremely insensitive to erroneous or atypical data.
- The LMS value minimizes the following function:

$$\tilde{x} = \underset{x}{\operatorname{argmin}} \operatorname{med}_{i=1}^{n} (x_i - x)^2$$

- A related measure of position is the "shorth", the mean of all data in the shortest interval containing h data points.

Empirical Measures

Mode

Mode is the most common value of a data list

- Useful mainly for qualitative features.
- For quantitative features, the mode can be determined from the kernel density estimator.

PYTHON: `xmode=scipy.stats.mode(x)`

HSM (Half-sample mode)

- Determine the shortest interval containing $h = \lfloor n/2 \rfloor + 1$ data points.
- Repeat the process on the data in that interval until there are two data points left.
- The HSM value is the average of these last two data points.

Dispersion Measures

Definition (Measure of Dispersion)

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a data list. The function $\sigma(\boldsymbol{x})$ is called a measure of dispersion for $\boldsymbol{x}$ if the following holds:

- $\sigma(\boldsymbol{x}) \geq 0$
- $\sigma(a\boldsymbol{x} + b) = |a|\,\sigma(\boldsymbol{x})$

- Meaningful measures of dispersion quantify the deviation of the data from their central value.
- Measures of dispersion are invariant under a shift of the data.
- Different measures of dispersion are useful depending on the scale and use case.

Empirical Measures

The Standard Deviation

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

- Useful for interval and ratio scales.
- The standard deviation has the same dimension as the data.
- The square of the standard deviation is called **variance of the data**.

PYTHON: `xstd=numpy.std(x)`

PYTHON: `xvar=numpy.var(x)`

Empirical Measures

Interquartile Distance

$$IQR = Q_{0.75} - Q_{0.25}$$

- The interquartile range is the length of the interval containing the central 50% of the data.
- Useful for ordinal, interval and ratio scales.

PYTHON: `xiqr=scipy.stats.iqr(x)`

Empirical Measures

LoS (Length of the Shorth)

LoS is the length of the shortest interval containing $h = \lfloor n/2 \rfloor + 1$ data points

- Meaningful for ordinal, interval and ratio scales.

Skewness Measures

Definition (Measure of Skewness)

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a data list. The function $\mathfrak{s}(\boldsymbol{x})$ is called a skewness measure for $\boldsymbol{x}$ if the following holds:

- $\mathfrak{s}(a\boldsymbol{x} + b) = \operatorname{sgn}(a)\,\mathfrak{s}(\boldsymbol{x})$
- $\mathfrak{s}(\boldsymbol{x}) = 0$ if $\exists b : \boldsymbol{x} - b = b - \boldsymbol{x}$

- Sensible skewness measures quantify the asymmetry of the data.
- Skewness measures are invariant under translation of the data.
- Different skewness measures are useful depending on the scale.

Empirical Measures

The Skewness

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

- The skewness γ is equal to 0 for symmetric data.
- If $\gamma < 0$, the data is said to be **skewed towards the left**.
- If $\gamma > 0$, the data is said to be **right skewed**.
- Useful for interval and ratio scales.

PYTHON: `xgamma=scipy.stats.skew(x)`

Empirical Measures

Skewness Coefficient

$$SK = \frac{R - L}{R + L}$$

where $R = Q_{0.75} - Q_{0.5}, L = Q_{0.5} - Q_{0.25}$.

- SK is between -1 $(R = 0)$ and $+1$ $(L = 0)$.
- The skewness coefficient is equal to 0 for symmetric data.
- If $SK < 0$, the data is called **left skewed**.
- If $SK > 0$, the data is called **right skewed**.
- Useful for ordinal, interval and ratio scales.

Subsection: Boxplot and Empirical Distribution Function

1 Introduction

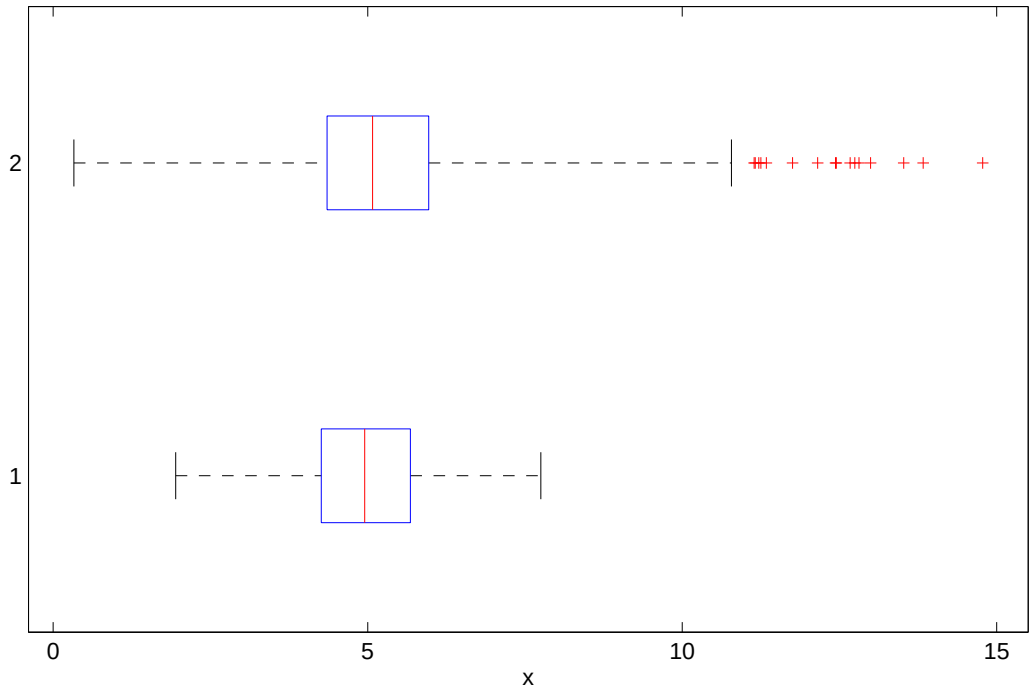
2 One-dimensional Features

- Graphical Representation
- Kernel Density Estimator
- Empirical Measures
- Boxplot and Empirical Distribution Function
- Examples

3 Two-dimensional Features

Boxplot and Empirical Distribution Function

- The **Boxplot** is the graphical representation of the five point summary.
- Comparison of data set 1 and data set 2:



PYTHON: `matplotlib.pyplot.boxplot(x)`

Boxplot and Empirical Distribution Function

- For ordinal scales "onwards" it is useful to order the data.
- The frequency table can be supplemented by *cumulative* frequencies.
- Data set 3 (50 exam scores):

Grade k	$h(k)$	$H(k)$	$f(k)$	$F(k)$
1	5	5	0.10	0.10
2	8	13	0.16	0.26
3	22	35	0.44	0.70
4	5	40	0.10	0.80
5	10	50	0.20	1.00

frequency table with cumulative frequencies

Boxplot and Empirical Distribution Function

- The graphical representation of the cumulative frequencies is called the **empirical distribution function** of the data list.

Definition (Empirical Distribution Function)

The empirical distribution function $F_n(x)$ of the data list $\boldsymbol{x} = (x_1, \ldots, x_n)$ is the proportion of the data that are less than or equal to x:

$$F_n(x) = f(\vec{x} \leq x).$$

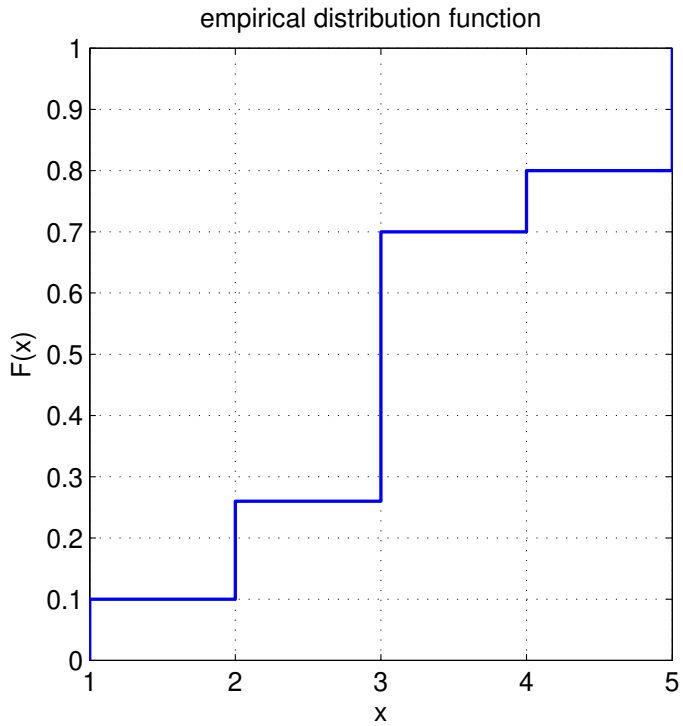
- If $x_i \leq x < x_{i+1}$, then:

$$F_n(x) = f(x_1) + \cdots + f(x_i).$$

- F_n is a step function. The step points are the data points, the step heights are the relative frequencies of the data points.

Boxplot and Empirical Distribution Function

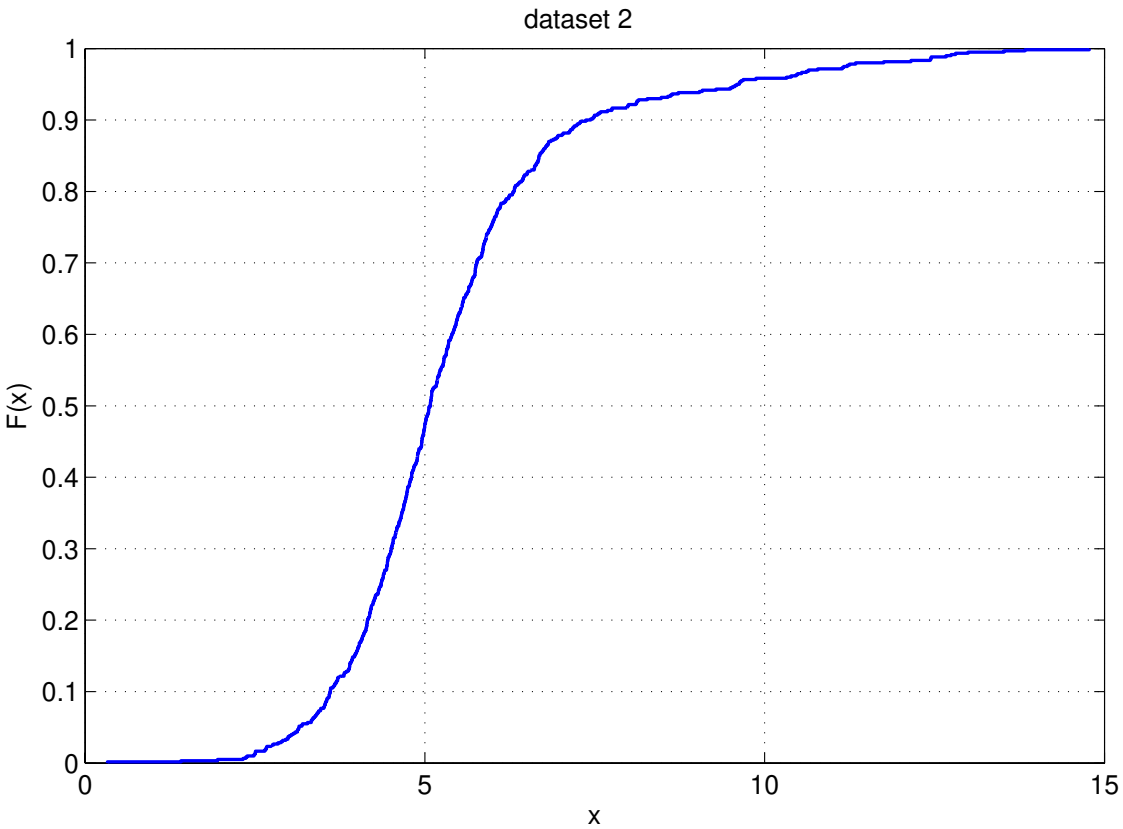
- record 3: (50 exam notes):



empirical distribution function

Boxplot and Empirical Distribution Function

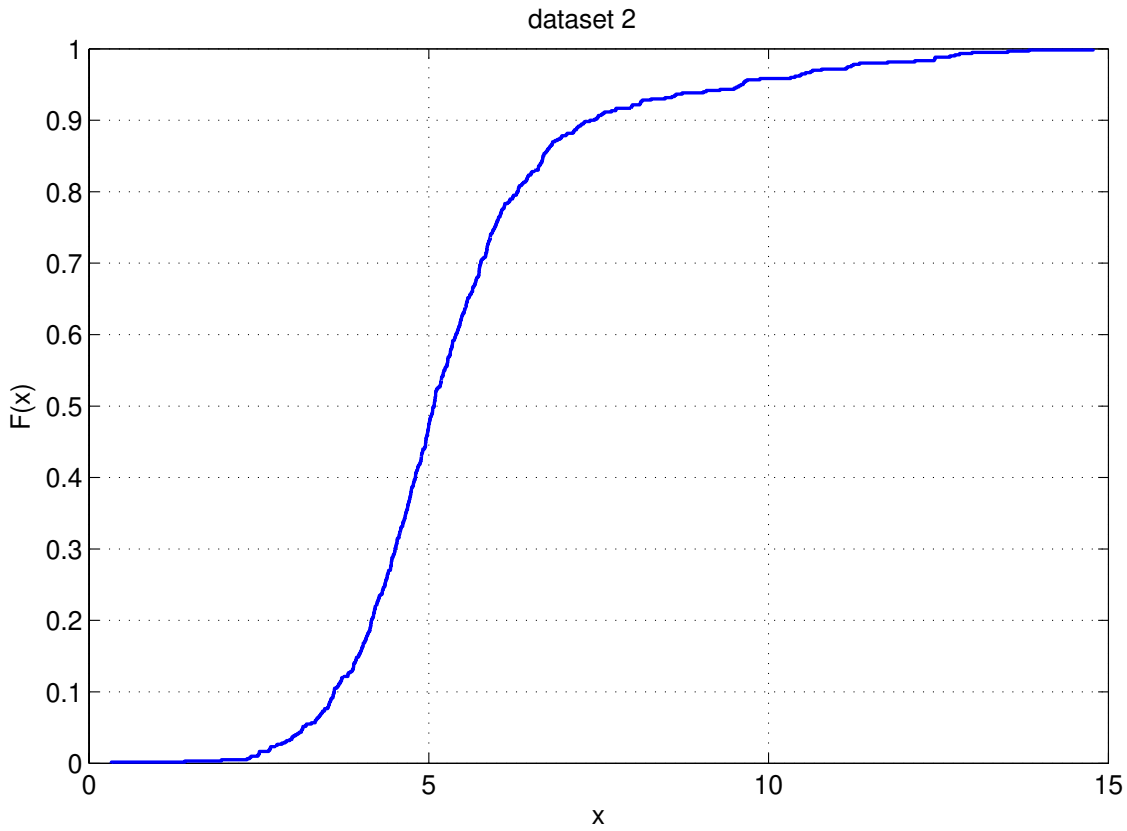
- dataset 2 (500 values + contamination):



empirical distribution function

Boxplot and Empirical Distribution Function

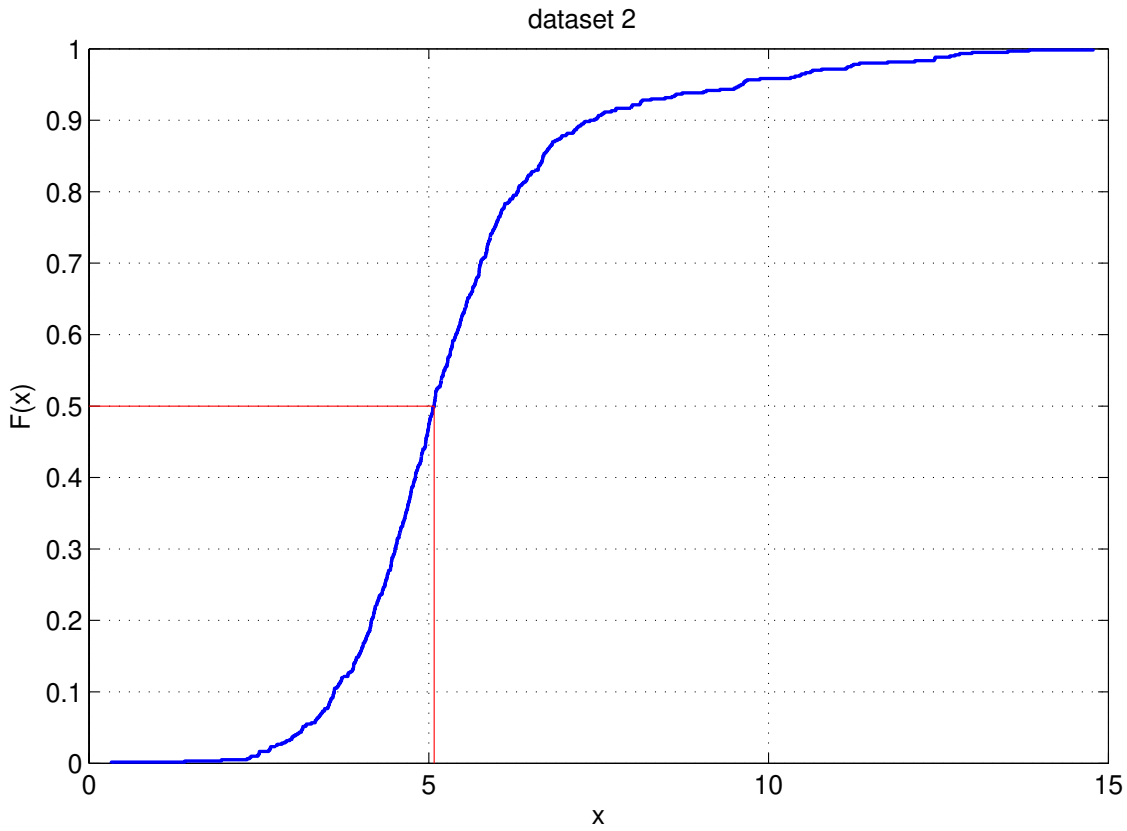
- Quantiles can be easily read off the empirical distribution function.
- Median of data set 2:



empirical distribution function

Boxplot and Empirical Distribution Function

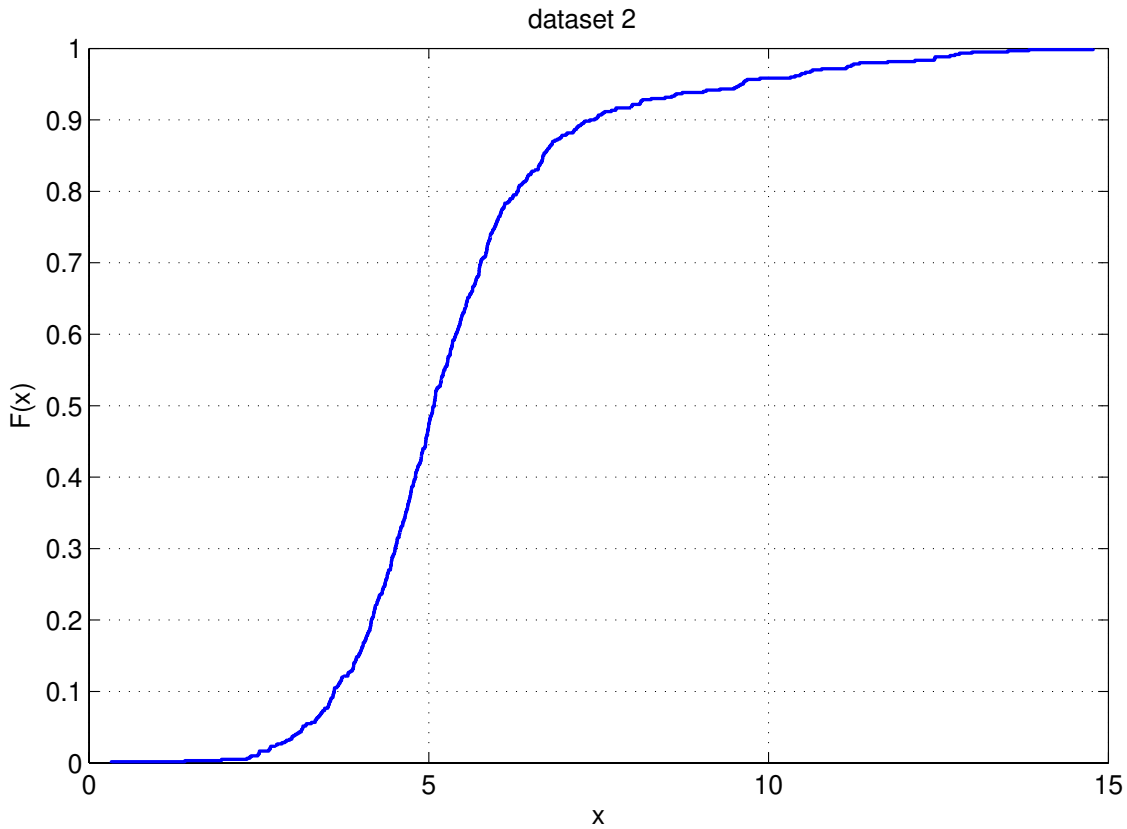
- Quantiles can be easily read off the empirical distribution function.
- Median of data set 2:



empirical distribution function

Boxplot and Empirical Distribution Function

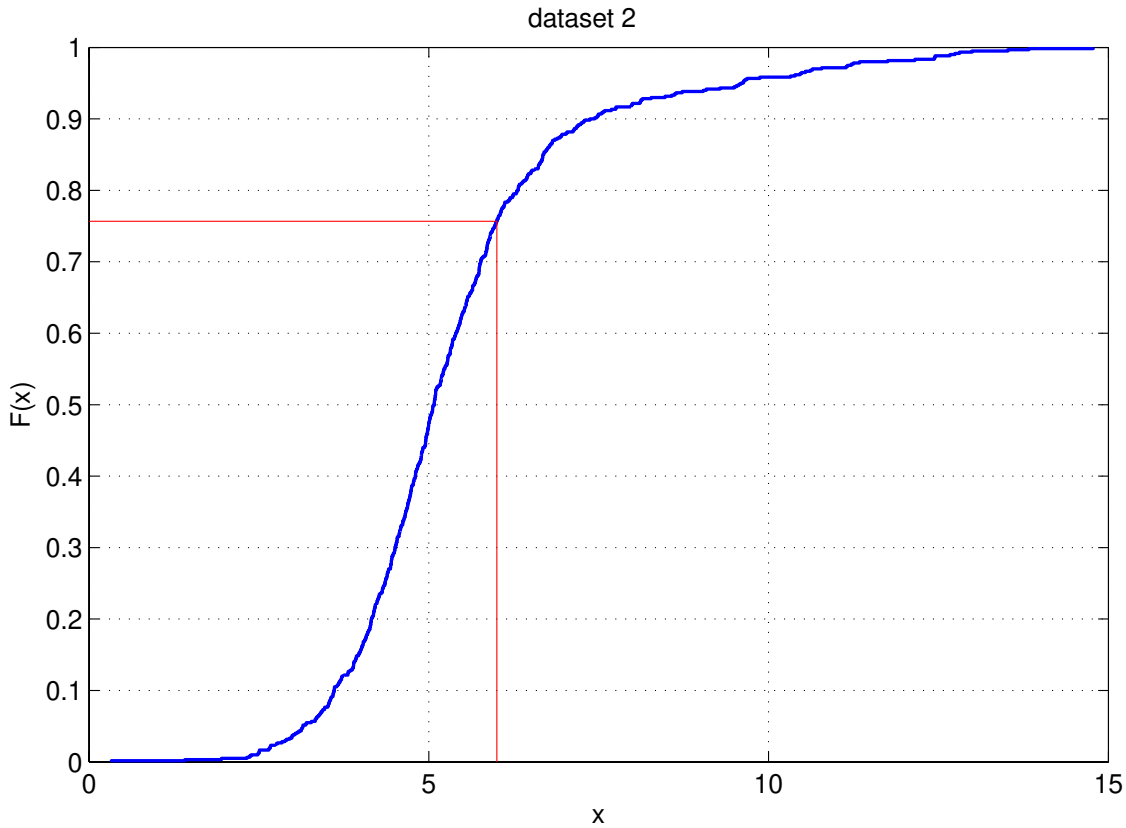
- Lower tail and upper tail frequencies can also be read off easily.
- What proportion of the data is less than or equal to 6?



empirical distribution function

Boxplot and Empirical Distribution Function

- Lower tail and upper tail frequencies can also be read off easily.
- What proportion of the data is less than or equal to 6?



empirical distribution function

Subsection: Examples

1 Introduction

2 One-dimensional Features

- Graphical Representation
- Kernel Density Estimator
- Empirical Measures
- Boxplot and Empirical Distribution Function
- Examples

3 Two-dimensional Features

Examples

- dataset 1: Symmetric, 500 values

- Measures of location:

Mean:	4.9532
Median:	4.9518
LMS:	4.8080
Shorth:	4.8002
HSM:	5.0830

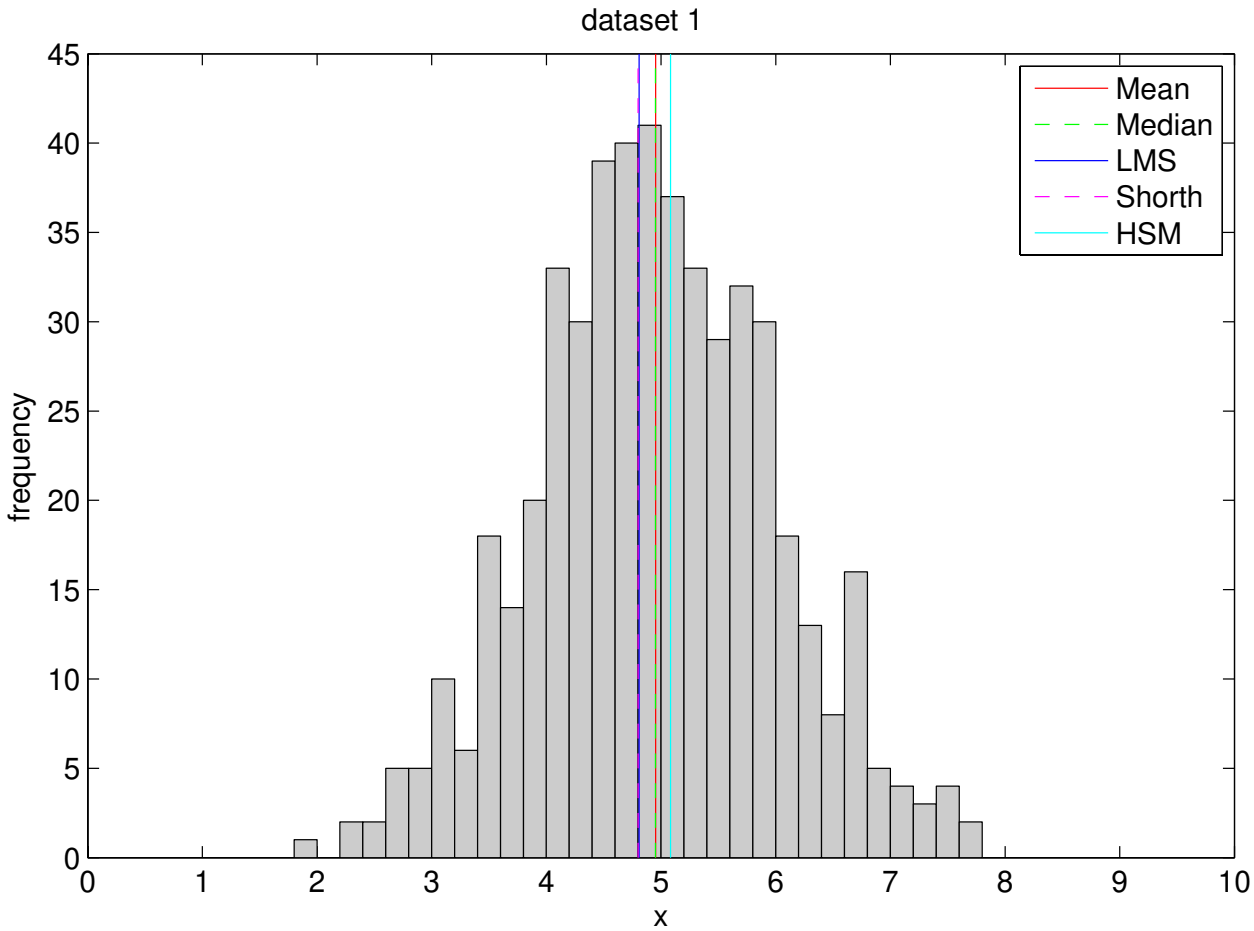
- Measures of Skewness:

Skewness:	0.0375
Skewness coefficient:	0.0258

- Measures of dispersion:

Standard deviation:	1.0255
Interquartile range:	1.4168
Length of the Shorth:	1.3520

Examples



dataset 1: mean, median, LMS, Shorth, HSM.

Examples

- dataset 2: dataset 1 + contamination (100 values)

- Measures of location:

Mean:	5.4343
Median:	5.0777
LMS:	5.1100
Shorth:	5.0740
HSM:	4.9985

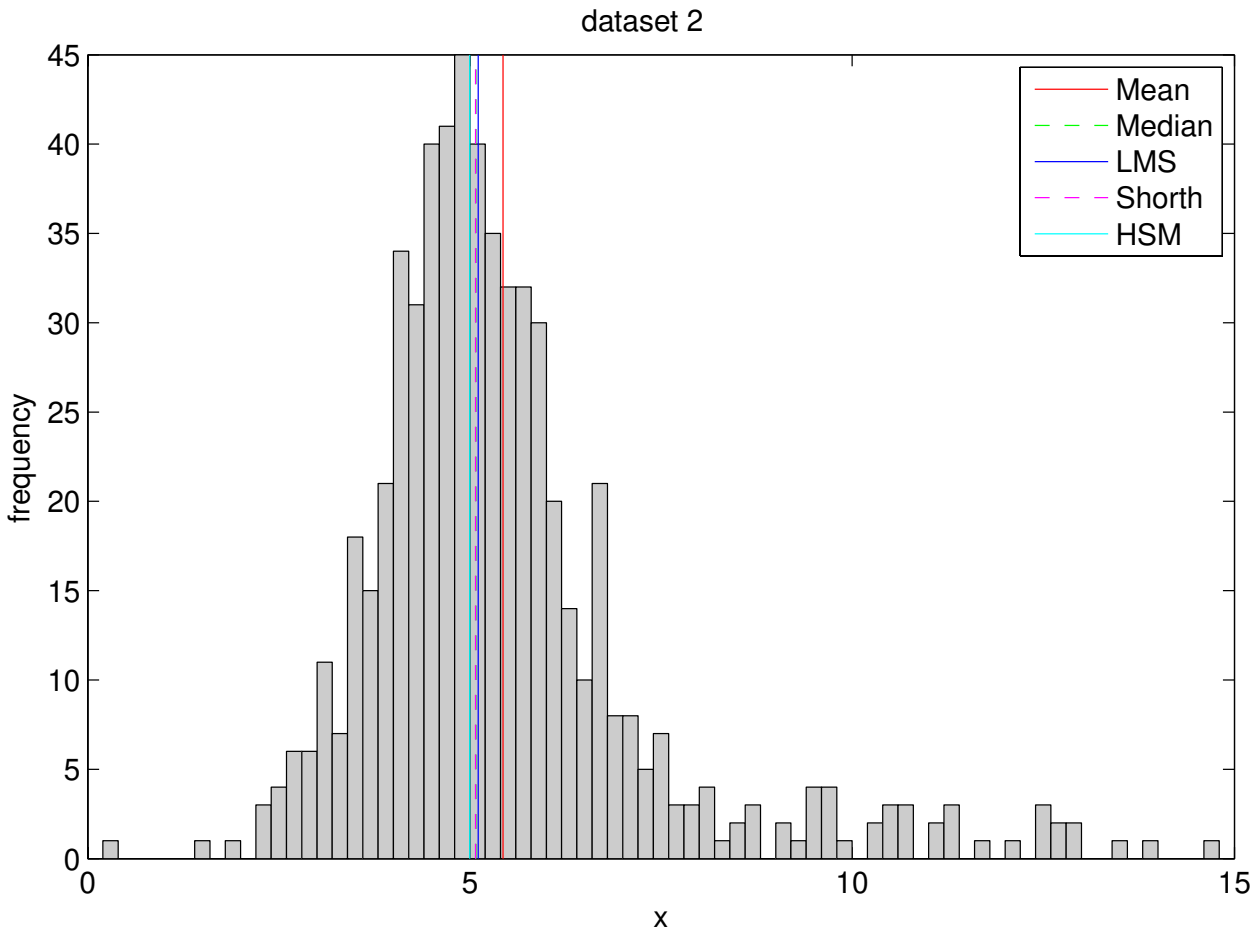
- Measures of skewness:

Skewness:	1.7696
Skewness coefficient:	0.1046

- Measures of dispersion:

Standard deviation:	1.8959
Interquartile range:	1.6152
Length of the Shorth:	1.5918

Examples



dataset 2: mean, median, LMS, Shorth, HSM.

Examples

- record 3: 50 exam grades

- Measures of position:

  Mean: 3.14

  Median: 3.0

  Mode: 3.0

- Measures of dispersion:

  Standard deviation: 1.20

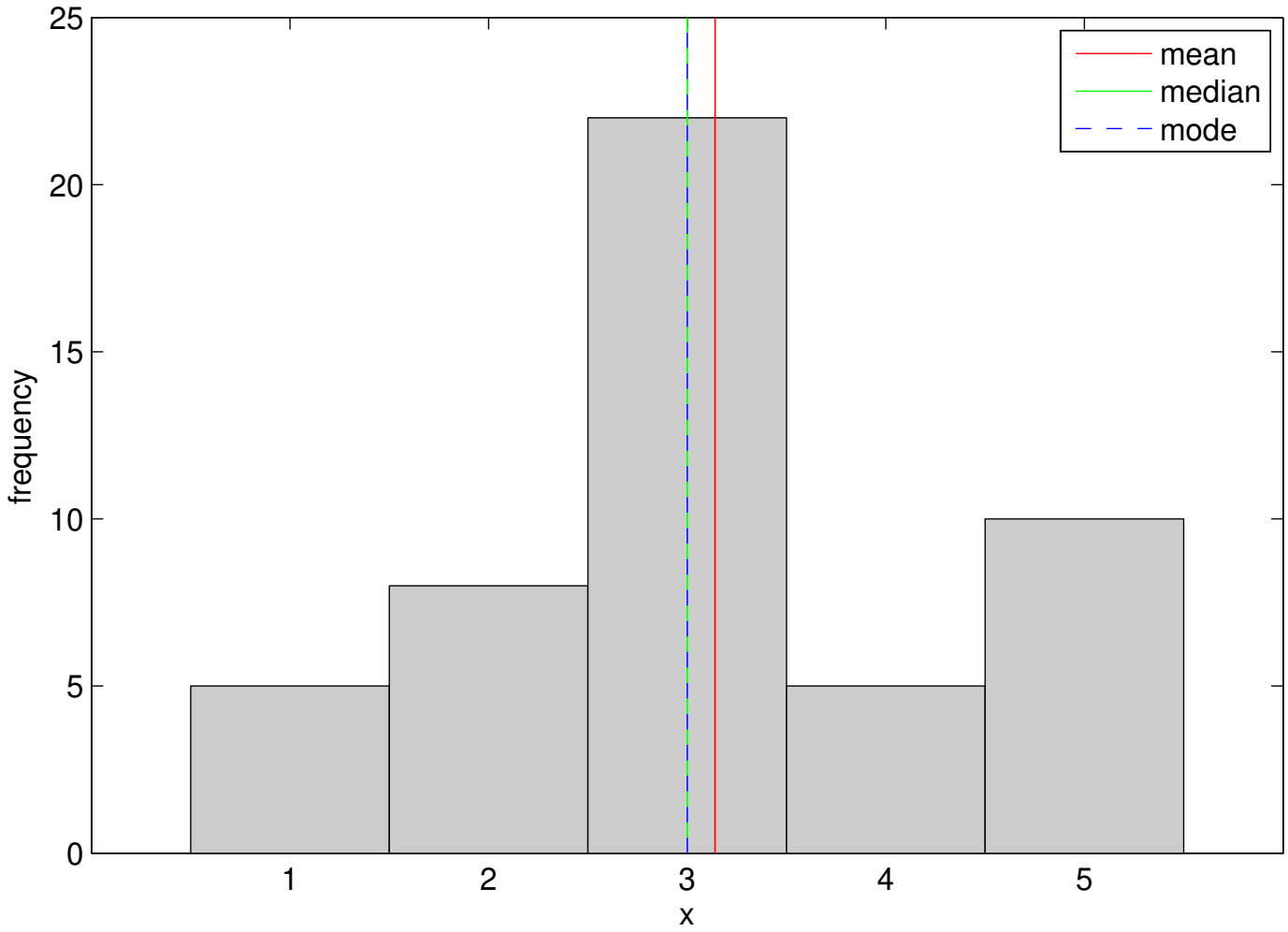
  Interquartile range: 1.75

- Measures of skewness:

  Skewness: 0.0765

  Skewness coefficient: 0.14

Examples



dataset 3: mean, median, mode.

Section 3: Two-dimensional Features

Two-dimensional Features

- Often two or more features of an object **are observed simultaneously**.
- Examples:
  - height and weight of a person
  - age and income of a person
  - education and gender of a person
- The correlation between the two features is additional information.

Subsection: Qualitative Features

Qualitative Features

- We first consider two binary features A and B.
- The frequency of the occurrence of A and B can be expressed in an **four-field table** or **contingency table**.
- example:

  A=*"The person is female"*
  B=*"The person is a smoker"*

- Four-field table for 1000 people:

	B	B'	
A	228	372	600
A'	136	264	400
	364	636	1000

Qualitative Features

- General structure of a four-field table:

	B	B'	
A	$h(A \cap B)$	$h(A \cap B')$	$h(A)$
A'	$h(A' \cap B)$	$h(A' \cap B')$	$h(A')$
	$h(B)$	$h(B')$	n

- Row and column sums are the frequencies of the expressions A, A' and B, B'.

Qualitative Features

- The four-field table can be converted to **relative frequencies** using division by n:

	B	B'	
A	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
A'	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	1

- row and column sums are the relative frequencies of the expressions A, A' and B, B'.

Qualitative Features

- The correlation of the two features can be measured by **four-field correlation**:

Four-field Correlation

$$\rho(A,B) = \frac{f(A \cap B) - f(A)f(B)}{\sqrt{f(A)f(A')f(B)f(B')}}$$

- It is always true: $-1 \leq \rho(A,B) \leq 1$
- If $\rho(A,B) > 0$, A and B are called **positively coupled**.
- If $\rho(A,B) < 0$, A and B are called **negatively coupled**.

Qualitative Features

- The sign of $\rho(A, B)$ indicates the **direction** of the coupling.
- The magnitude of $\rho(A, B)$ indicates the **strength** of the coupling.
- Specifically:

$$A = B \implies \rho(A, B) = 1$$
$$A = B' \implies \rho(A, B) = -1$$

- An existing coupling is no proof of a causal relationship!
- The coupling can also result from a common cause for both features.

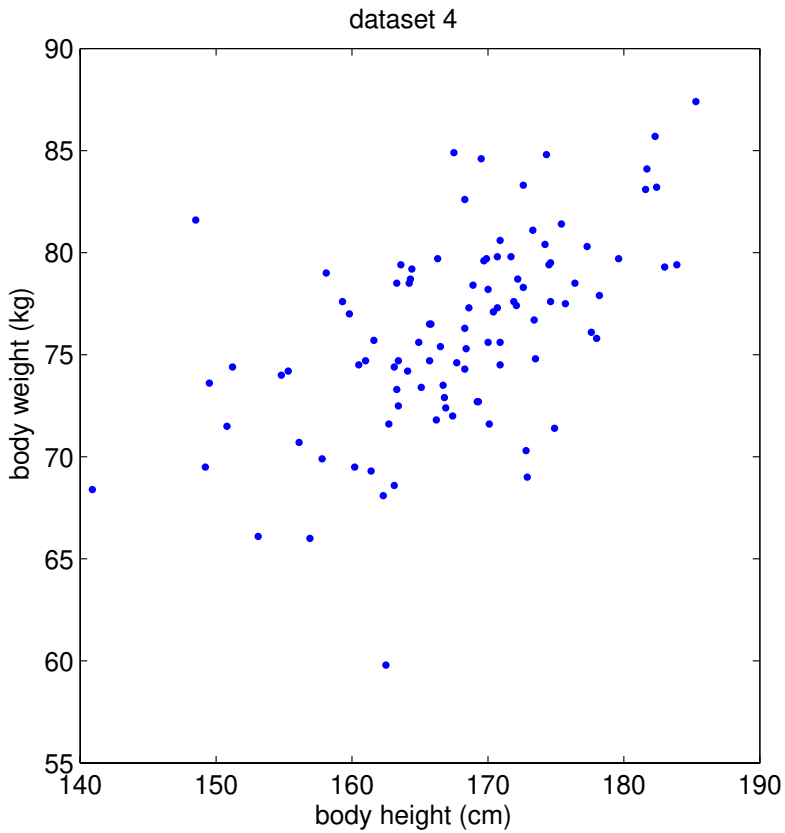
Subsection: Quantitative Features

Quantitative Features

- Preferred representation of two-dimensional features: Scatter Plot.
- Each point corresponds to one object.
- The observed features determine the position of the point in the x-y-plane.
- Higher dimensional features can be represented by histograms and scatter plots. Of course, some of the information is lost in the process.

Quantitative Features

- dataset 4: height and weight of 100 individuals



scatter plot

PYTHON: `matplotlib.pyplot.scatter(x,y)`

Quantitative Features

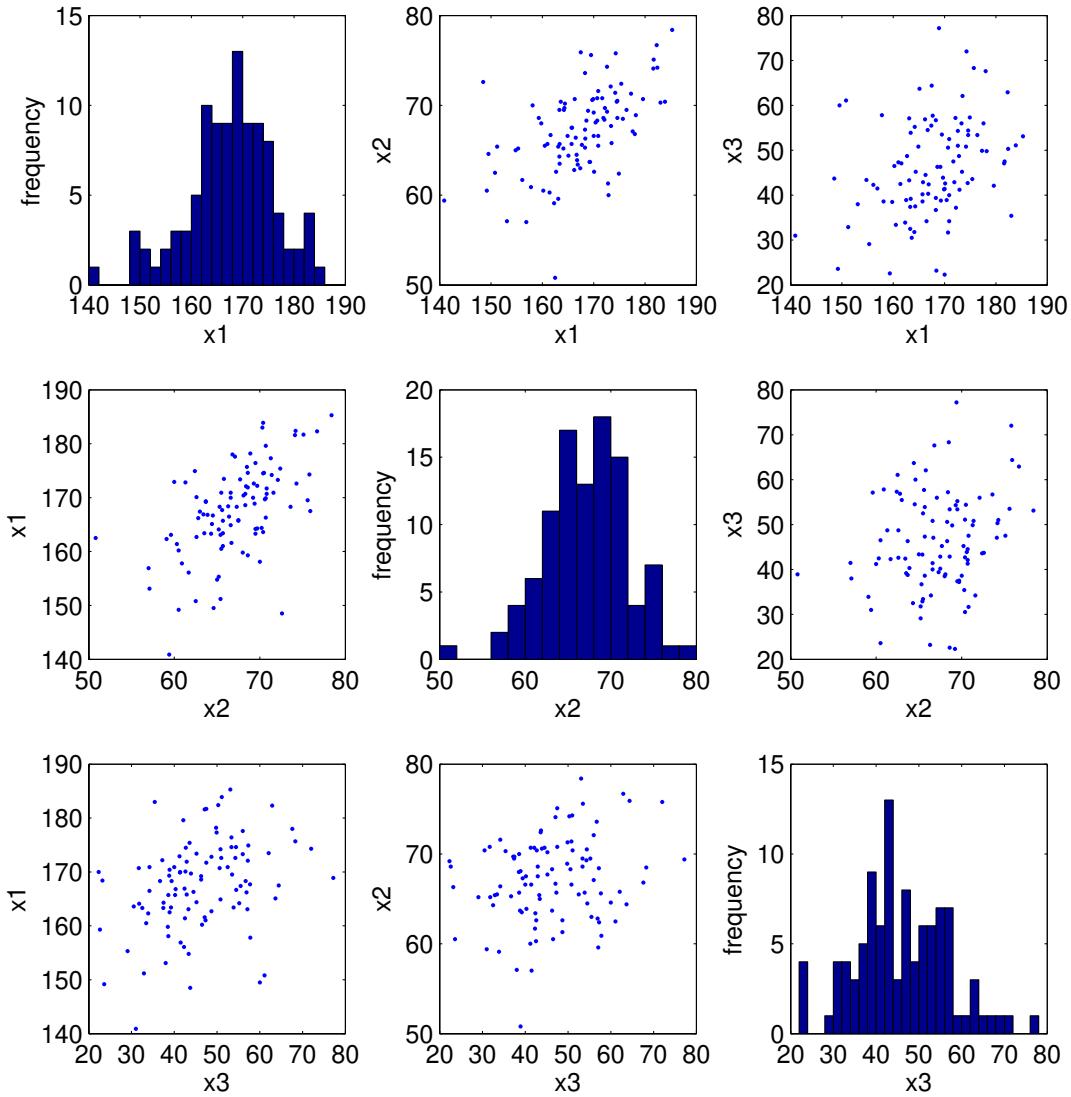
- dataset 5: height, weight and age of 100 people.

  Feature x_1: height (in cm)

  Feature x_2: weight (in kg)

  Feature x_3: age (in years)

Quantitative Features



Subsection: Correlations

1 Introduction

2 One-dimensional Features

3 Two-dimensional Features

- Qualitative Features
- Quantitative Features
- Correlations

Correlations

Properties of scatter plot

- $(\bar{x}, \bar{y})$ is the center of the point cloud.
- The projection of the scatter plot onto the x-axis results in the scatter plot of the data list $x_1, \ldots, x_n$.
- The projection of the point cloud onto the y-axis results in the point diagram of the data list $y_1, \ldots, y_n$.

- From the scatter plot of dataset 4, it can be seen that **tendentially** larger height is associated with larger weight.
- There is obviously a relationship between the two features x and y, which is also intuitively completely clear.

Correlations

- We need a **measure** for this correlation.
- A useful measure is the **empirical correlation coefficient**.
- Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a bivariate sample.
- We compute the **standard scores**:

$$z_{x,i} = \frac{x_i - \bar{x}}{s_x}, \quad z_{y,i} = \frac{y_i - \bar{y}}{s_y}$$

- We recall that

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- The empirical correlation coefficient is the **mean of the products** of the standard scores.

Correlations

Definition (Empirical Correlation Coefficient)

The **empirical correlation coefficient** r_{xy} is defined as

$$r_{xy} = \frac{1}{n} \sum_{i=1}^{n} z_{x,i} z_{y,i} = \frac{1}{n} \left(z_{x,1} z_{y,1} + \cdots + z_{x,n} z_{y,n} \right)$$

- It is always true that:

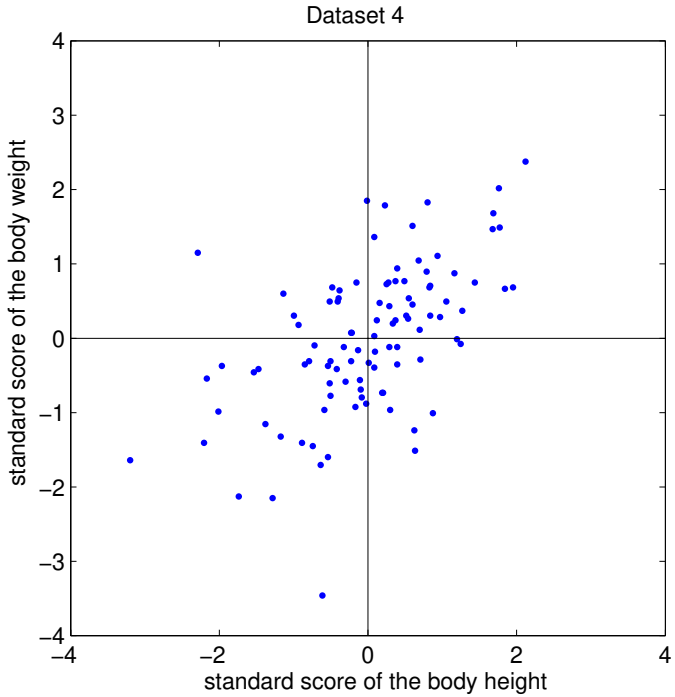
$$-1 \leq r_{xy} \leq 1$$

Correlations

- r_{xy} is positive if many products are positive, i.e. many pairs of of standard scores have the same sign.
- This is the case when the pairs of standard scores are predominantly in the 1st or 3rd quadrant.
- x and y are then called **positively correlated**.
- r_{xy} is negative if many products are negative, i.e. many pairs of standard scores have different sign.
- This is the case when the pairs of standard scores are predominantly in the 2nd or 4th quadrant.
- x and y are then called **negatively correlated**.

Correlations

- Scatter plot of standard scores from dataset 4:



- Obviously x and y are positively correlated, since most of the points are in the 1st and 3rd quadrants.
- $r_{xy} = 0.5562$

Correlations

- A positive correlation does not necessarily imply a causal relationship.
- The positive correlation can also be caused by a common cause or a parallel trend.

Example

The number of children born correlates with the number of storks in Austria in the last 30 years. Why?

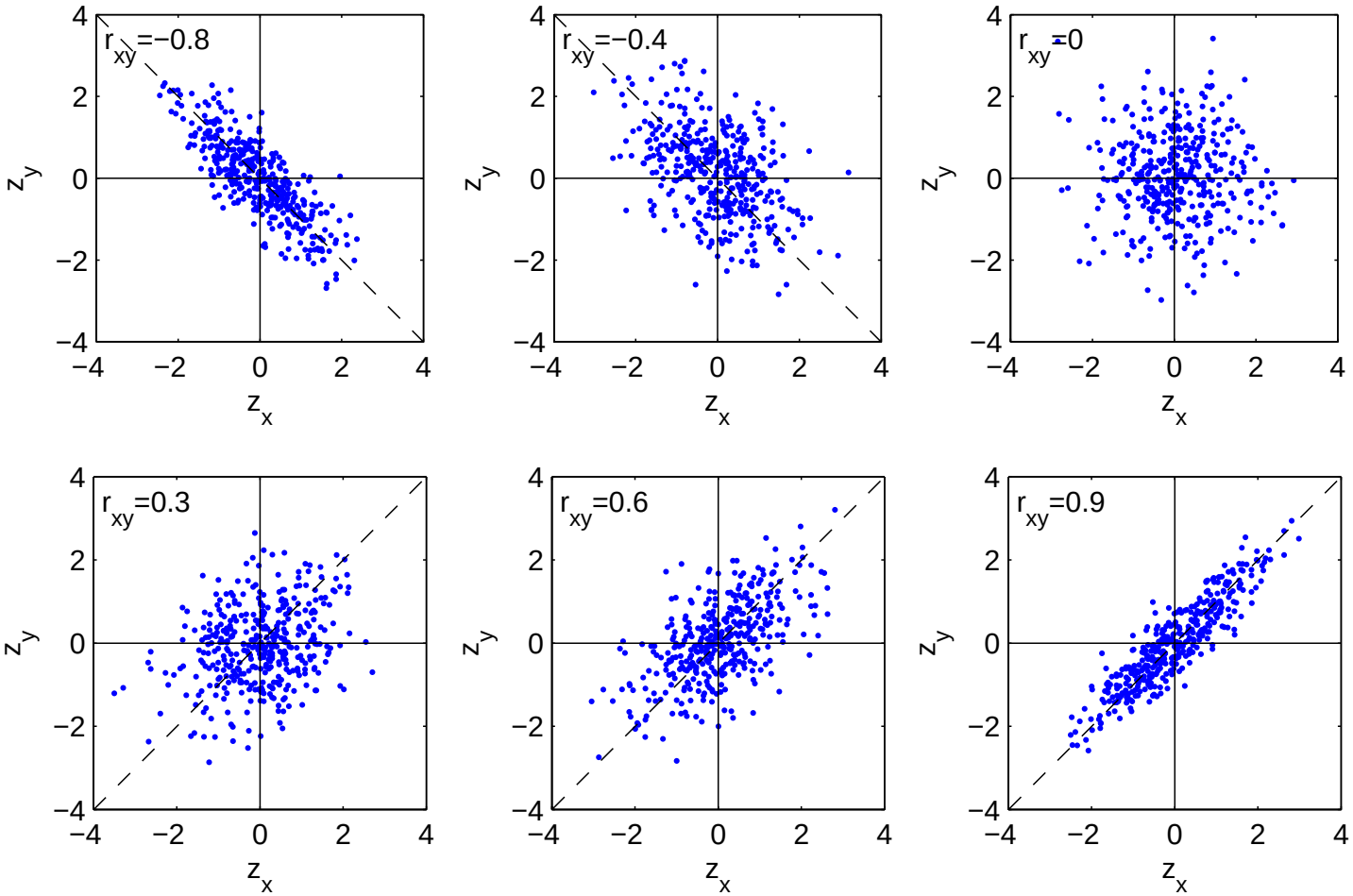
Example

We see a positive correlation between the price of butter and the price of bread in the last 50 years. Why?

Example

Over the last few hundred years, we observe an anti-correlation between the number of pirates on the world's oceans and global CO_2 emissions. Why?

Correlations

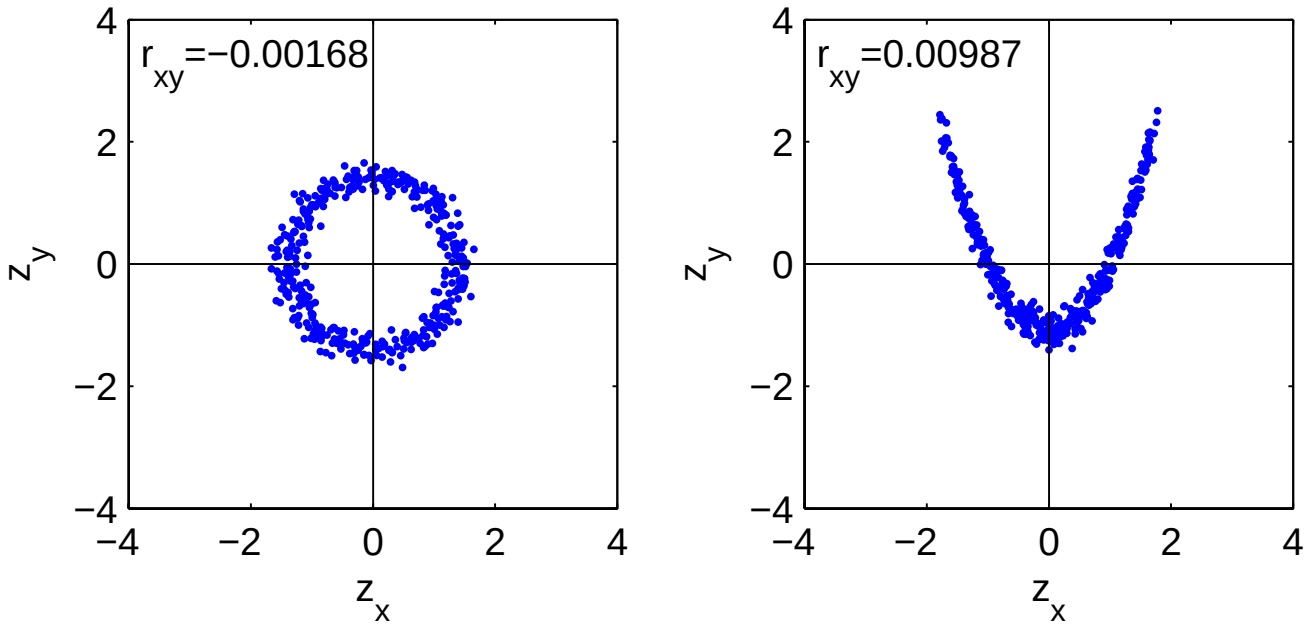


standard scores with different correlation coefficients.

Correlations

- The correlation coefficient measures the **correlation** of the data.
- The correlation indicates the "binding" of the point cloud to a rising or falling **line**, the **major axis**.
- Thus, the correlation indicates the extent of the **linear** coupling.
- If there is a strong but **nonlinear** relationship between x and y, the correlation may still be very small.

Correlations



Nonlinear relation between x and y

Correlations

- The correlation coefficient can also be calculated directly from the sample:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Definition (Covariance)

The quantity

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

is called the **covariance of the data**

Part II

Probabilities

Overview Part 2

Section 4: Introduction

Introduction

- The specific outcome of an experiment generally cannot be accurately predicted; however, the possible outcomes are known.
- Several reasons:
  - The observed objects are a random selection (sample) from a larger population.
  - The observed process is in principle indeterministic (quantum mechanics).
  - The observed process is practically indeterministic: lack of knowledge of the initial state (roulette), chaotic system (hydrodynamics, psychology).
- In addition, measurement errors can add stochasticity (randomness) to the result.

Introduction

- The goal of probability theory is to assign **probabilities** to sets of outcomes, called events.
- Two interpretations of probabilities exist.

Introduction

Frequentist Interpretation

- The probability of an event is its frequency when the experiment is repeated very often under the same conditions.
- The statistics based on this interpretation is called 'frequentist'.

Example

The probability of the output '1' when rolling a dice is the limit of the frequency for a large number of rolls.

Subjective Interpretation

- The probability of an output is a statement about the belief of the person giving the probability.
- The statistics based on this interpretation is called 'Bayesian'.

Example

'The probability that it will rain tomorrow is 40 percent' is a statement about the belief of the person making this statement.

Introduction

- In practice, the transition between the two approaches is often fluid.
- In many cases the results are identical, only the interpretation is different.
- The Bayesian approach is more comprehensive and flexible.
- The frequentist approach is often simpler, but more limited. Its interpretation is typically clearer.

Section 5: Events

Subsection: The Sample Space

The Sample Space

- The notion of a (random) **event** is fundamental to statistics.
- Concretely: the **outcome of an experiment**, the result of which **cannot be predicted exactly**.
- The set Ω of all possible outcomes is called **sample space**.
- The sample space Ω can be discrete (finite or countably infinite) or continuous (uncountably infinite).

The Sample Space

Example

- In roulette, there are 37 possible outcomes. The sample space is discrete and finite.
- If a radioactive source is monitored, the number of decays per second is in principle unbounded. The sample space is discrete and countably infinite.
- The waiting time between two decays can take any value. The sample space is continuous and uncountably infinite.

Subsection: The Event Algebra

The Event Algebra

Definition (Event)

An **event** E is a subset of the sample space Ω. An event E **occurs** if E contains the outcome $\omega \in \Omega$ of the experiment.

Example

The roll of a dice has the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event E (even number) is the subset

$$E = \{2, 4, 6\}$$

E occurs when an even number is thrown.

The Event Algebra

Definition (event algebra)

The set of all events of the sample space Ω is called the event algebra $\Sigma(\Omega)$.

- In the finite or countably infinite case, **any** subset can be considered as an event. The event algebra is called **discrete**.
- In the uncountably infinite case, certain pathological (non-measurable) subsets must be excluded. The event algebra is called **continuous** or **real-valued**.
- Like statements, two events $A \in \Sigma$ and $B \in \Sigma$ can be logically **connected**.

The Event Algebra

Linking Events

Disjunction

Symbol	Name	Meaning
$A \cup B$	disjunction	A or B (or both)

conjunction

Symbol	Name	Meaning
$A \cap B$	conjunction	A and B (both A and B)

Negation

Symbol	Name	Meaning
A'	negation	not A (the opposite of A)

The Event Algebra

- With these operators, Σ is a **Boolean algebra**: distributive complementary lattice with zero- and one-elements.
- The null element $\mathbf{0} = \emptyset$ is the **impossible event**.
- The one element $\mathbf{1} = \Omega$ is the **sure event**.
- An event consisting of only one possible result is called an **elementary event**.

The Event Algebra

- If Ω is (countably or uncountably) infinite, one requires that also countably many unions and intersections of events can be formed.
- The event algebra is then a so-called σ-algebra.
- If in the continuous case $\Omega = \mathbb{R}$, then the event algebra $\Sigma(\Omega)$ is the smallest σ-algebra containing all intervals.

Subsection: Repeating Experiments

Repeating Experiments

- The sample space of a dice roll is

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Consequently, the event algebra $\Sigma(\Omega)$ has six elementary events:

$$e_1 = \{1\}, e_2 = \{2\}, e_3 = \{3\}, e_4 = \{4\}, e_5 = \{5\}, e_6 = \{6\}$$

and a total of $2^6 = 64$ events (subsets of Ω).

- The sample space of a double roll is the Cartesian product $\Omega \times \Omega$:

$$\Omega \times \Omega = \{(i, j) | i, j = 1, \ldots, 6\}$$

The ordered pair (i, j) means i on the first throw, j on the second throw. Consequently, the event algebra $\Sigma(\Omega \times \Omega)$ has 36 elementary events e_{ij}:

$$e_{11} = \{(1, 1)\}, \ldots, e_{36} = \{(6, 6)\}$$

Repeating Experiments

- Similarly, in the case of rolling n times, the sample space is n times the Cartesian product $\Omega \times \Omega \times \ldots \times \Omega$.

Example (event algebra of double throw)

Examples of events in the event algebra of the double roll are:

6 on the first throw:	$\{(6,1),(6,2),\ldots,(6,6)\}$
6 on the second throw:	$\{(1,6),(2,6),\ldots,(6,6)\}$
Both throws equal:	$\{(1,1),(2,2),\ldots,(6,6)\}$
Sum of throws equal to 7:	$\{(1,6),(2,5),\ldots,(6,1)\}$

Repeating Experiments

Example (Repeated Bernoulli Experiment)

An experiment that has only two possible outcomes is called a **Bernoulli experiment** or **alternative experiment**. There are two outcomes, 1 ("success") and 0 ("failure").

An alternative experiment that is executed n times is described by a sample space with 2^n possible outcomes, namely all sequences of the form $(i_1, \ldots, i_n)$ with $i_j = 0$ or 1.

Often, however, only the **frequency** of the occurrence of 1 (or 0) is of interest. This situation is more comprehensively described by only $n + 1$ outcomes: 1 occuring $0, 1, 2, \ldots$ or n times. If the event E_1 denotes a single occurrence of 1, then E_1 is the union of multiple elementary events of the original event algebra:

$$E_1 = \{(e_1, e_0, \ldots, e_0), (e_0, e_1, e_0, \ldots, e_0), \ldots, (e_0, \ldots, e_0, e_1)\}$$

An example is an n-fold coin flip.

Section 6: Probabilities

Subsection: Probability Measures

Probability Measures

Definition (probabilitymeasure)

Let Σ be an event algebra, A and B events in Σ, and P a mapping from Σ into $\mathbb{R}$. P is called a **probability measure** if the following holds:

1. positivity $\quad P(A) \geq 0, \ \forall A \in \Sigma$
2. additivity: $\quad A \cap B = \mathbf{0} \implies P(A \cup B) = P(A) + P(B)$
3. normalization $\quad P(\mathbf{1}) = 1$

Probability Measures

Definition (probability space)

If Σ is a σ-algebra, which must be assumed for infinite event spaces, one requires for countable J:

4. σ-additivity: $A_i \in \sigma,\ i \in J;\ A_i \cap A_j = \mathbf{0},\ i \neq j \implies P(\bigcup_{i \in J} A_i) = \sum_{i \in J} P(A_i)$

Σ is then called normalized, and (Σ, P) a **probability space**. P is also called a **probability distribution**.

Probability Measures

Calculating laws for probability

If (Σ, P) is a probability space, then:

- $P(\mathbf{0}) = 0$
- $P(\mathbf{1}) = 1$
- $0 \leq P(A) \leq 1, \ \forall A \in \Sigma$
- $P(A') = 1 - P(A), \ \forall A \in \Sigma$
- $A \subseteq B \implies P(A) \leq P(B), \ \forall A, B \in \Sigma$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B), \ \forall A, B \in \Sigma$
- If Σ has at most countably many elementary events $\{e_i \mid i \in I\}$, then $\sum_{i \in I} P(e_i) = 1$.

Probability Measures

- In a discrete event algebra, the probability of an event is equal to the sum of the probabilities of the elementary events whose union it is.
- Therefore, in this case, a probability measure is **uniquely determined** by the values it assigns to the elementary events.
- On the other hand, any positive function defined on the set of elementary events that satisfies the normalization condition can be uniquely continued to a probability measure.
- Thus, one can define infinitely many distributions on a discrete event algebra Σ.

Probability Measures

- In a continuous event algebra, the probability of each elementary event is 0.
- Therefore, the probability of an event can no longer be determined by summation.
- Instead, a **density function** $f(x)$ is given which assigns a nonnegative value $f(x)$ to each elementary event x.
- The probability of an event A is determined by **integration** over the density:

$$P(A) = \int_A f(x)\,\mathrm{d}x$$

- The density function is **normalized** to 1:

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$$

Subsection: Law of Large Numbers

Law of Large Numbers

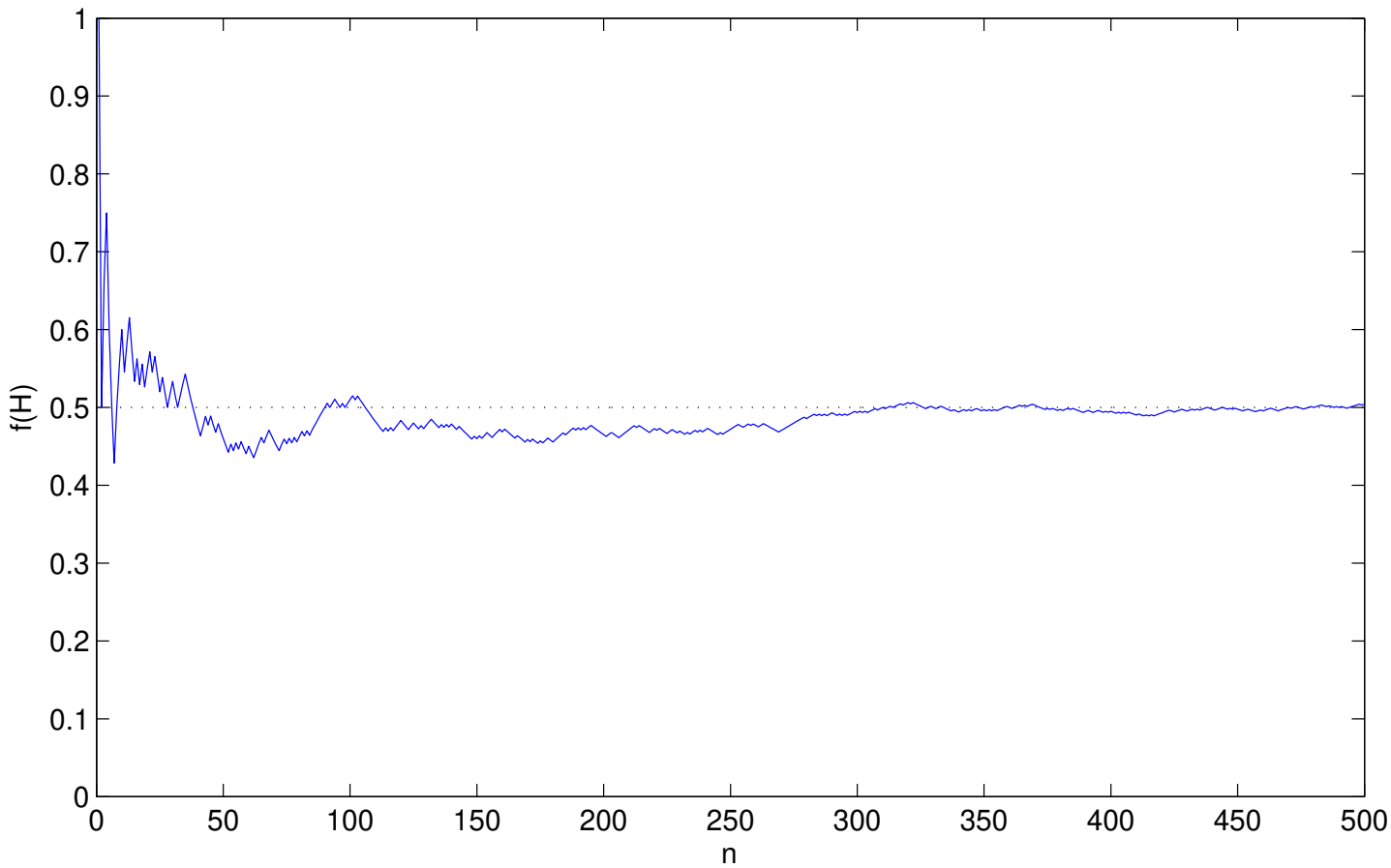
- We consider a simple random experiment: coin toss.
- Two possible outcomes: Heads (H), Tails (T).
- Assumption: symmetric coin, H and T equally probable
- Experiment is repeated n times

n	$h_n(H)$	$f_n(H)$	$\lvert f_n(H) - 0.5 \rvert$
10	6	0.6	0.1
100	51	0.51	0.01
500	252	0.504	0.004
1000	488	0.488	0.012
5000	2533	0.5066	0.0066

frequency table

MATLAB: `make_coin`

Law of Large Numbers



evolution of the relative frequency of H

Law of Large Numbers

- The relative frequency of the event K seems to strive towards the limit value 0.5.
- This limit is called the **probability** $P(K)$.

Empirical Law of Large Numbers

$$\lim_{n\to\infty} f_n(K) = P(K)$$

- The mathematical problem of this definition lies in the fact that the existence of the limit cannot be seen a priori and in fact need not exist in the classical analytic sense, but only in a broader, statistical sense.

Section 7: Conditional Probability

Subsection: Coupling and Conditional Probability

Coupling and Conditional Probability

- We now consider two events A and B that can occur in an experiment.
- **Question**: Is there a relationship between the events?
- Such a connection is called a **coupling**.
- **Positive Coupling**: The more often A occurs, the more often B tends to occur.
- **Negative Coupling**: The more often A occurs, the less often B also tends to occur.
- We can quantify the meanings of 'often' and 'rarely' with the frequency table.

Coupling and Conditional Probability

- The frequency of the occurrence of A and B can be summarized in a **four-field table** or **contingency table**.
- Example:

  A=*'The person under study is female'*
  B=*'The person under study has diabetes'*

- Four-field table for 1000 persons:

	B	B'	
A	19	526	545
A'	26	429	455
	45	955	1000

Coupling and Conditional Probability

- **Ordinary relative frequencies** are related to the size n of the entire data set:

$$f(A \cap B) = \frac{h(A \cap B)}{n}$$

- **Conditional relative frequencies** are related to the occurrence of the other feature:

$$f(A|B) = \frac{h(A \cap B)}{h(B)} = \frac{f(A \cap B)}{f(B)}$$

- $f(A|B)$ is called the conditional relative frequency of A given B.

Coupling and Conditional Probability

- The four-field table results in the following conditional relative frequencies:

$$f(A|B) = \frac{19}{45} = 0.422, \quad f(A|B') = \frac{526}{955} = 0.551$$

- Thus, it can be assumed that the two features are coupled.
- $f(A|B) > f(A)$ indicates a positive coupling, $f(A|B) < f(A)$ indicates a negative coupling.

Coupling and Conditional Probability

- If the data originate from a random experiment, then combinations of events can also be assigned probabilities.
- Probability table:

	B	B'	
A	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
A'	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
	$P(B)$	$P(B')$	1

- By the empirical law of large numbers, these probabilities are the limits of the respective relative frequencies.

Coupling and Conditional Probability

- The conditional relative frequencies converge towards a limit with $n \to \infty$:

$$f_n(A|B) = \frac{f_n(A \cap B)}{f_n(B)} \to P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Definition (Conditional probability)

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

is called the **conditional probability of A under condition B**, provided $P(B) \neq 0$.

Coupling and Conditional Probability

Example (The symmetric cube)

If the cube is completely symmetric, the elementary events $e_i = \{i\}$ are assigned equal probabilities:

$$P(e_i) = \frac{1}{6},\ 1 \leq i \leq 6$$

We define the following events:

$$O = \{1, 3, 5\},\ E = \{2, 4, 6\}$$

Then for example

$$P(e_1|O) = \frac{P(e_1 \cap O)}{P(O)} = \frac{P(e_1)}{P(O)} = \frac{1}{3}$$

$$P(e_1|E) = \frac{P(e_1 \cap E)}{P(E)} = \frac{P(\mathbf{0})}{P(E)} = 0$$

Example (Continuation)

$$P(O|e_1) = \frac{P(e_1 \cap O)}{P(e_1)} = \frac{P(e_1)}{P(e_1)} = 1$$

$$P(e_1 \cup e_3|O) = \frac{P((e_1 \cup e_3) \cap O)}{P(O)} = \frac{P(e_1 \cup e_3)}{P(O)} = \frac{2}{3}$$

$$P(e_1 \cup e_2|O) = \frac{P((e_1 \cup e_2) \cap O)}{P(O)} = \frac{P(e_1)}{P(O)} = \frac{1}{3}$$

Coupling and Conditional Probability

- The

Product Formula

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

follows immediately from the definition of conditional probabilities; as well as the formula for

Inverse Probability

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

- Both formulas are also valid for relative frequencies!

Subsection: Bayes' Theorem

Bayes' Theorem

Definition (Decomposition)

The events $B_1, B_2, \ldots, B_m$ form a **decomposition** of the sample space Ω if the following holds:

1. Non-overlapping: $B_i \cap B_j = \emptyset, i \neq j$
2. Completeness: $B_1 \cup B_2 \cup \ldots \cup B_m = \Omega$

Theorem

If the events $B_1, B_2, \ldots, B_m$ form a decomposition of the sample space Ω, it the follows that:

$$P(B_1) + P(B_2) + \ldots + P(B_m) = P(\Omega) = \mathbf{1}$$

Bayes' Theorem

- Let $B_1, \ldots, B_m$ be a decomposition. It follows that:

Total Probability

$$P(A) = P(A|B_1)P(B_1) + \ldots + P(A|B_m)P(B_m)$$

Example

A factory produces resistors with 10 kΩ (35% of production), with 22 kΩ (45%) and with 47 kΩ (20%). After two years of continuous operation, 98% of the 10 kΩ resistors are still functional, 96% of the 22 kΩ resistors, and 92% of the 47 kΩ resistors. What proportion of all resistors is still functional after two years?

Bayes' Theorem

- Let $B_1, \ldots, B_m$ be a decomposition. It follows that:

Bayes' Theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}$$
$$= \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \ldots + P(A|B_m)P(B_m)}$$

- $P(B_i)$ will be the **a-priori** probability of B, $P(B_i|A)$ is called the **a-posteriori** probability.

Bayes' Theorem

Example

A company buys components from two suppliers, where the share of the first one is 65% is. From experience, the scrap rate of supplier 1 is equal to 3% and that of supplier 2 is equal to 4%.

1. What is the total scrap percentage?
2. What is the probability that a flawless part comes from supplier 2?
3. What is the probability that a defective component comes from supplier 1?

Bayes' Theorem

Example

A component is supplied by four companies, and 20% comes from company 1, 30% comes from company 2, 35% comes from company 3, and 15% comes from company 4. The probability that the component fails in test operation within 24 hours is 0.02 for company 1, 0.015 for company 2, 0.025 for company 3, and 0.02 for company 4. A component fails in test operation after 16 hours. The probability that it comes from company i is to be calculated by means of Bayes' theorem.

Subsection: Independence

Independence

- Two events are **positively coupled** if.

$$P(A|B) > P(A) \quad or \quad P(A \cap B) > P(A)P(B)$$

- Two events are **negatively coupled** if.

$$P(A|B) < P(A) \quad P(A \cap B) < P(A)P(B)$$

- If neither positive nor negative coupling is present, then A and B are mutually **independent**.

Independence

Definition (Independence)

Two events A and B are called **stochastically independent** if

$$P(A \cap B) = P(A)P(B)$$

The events $A_1, A_2, \ldots, A_n$ are called independent if:

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot \ldots \cdot P(A_n)$$

For $n > 2$, pairwise independence of any two events A_i and A_j is not generally sufficient!

Independence

Example

We consider flipping a fair coin twice (heads H/tails T). The possible outcomes are $\Omega = \{HH, HT, TH, TT\}$. We define the events:

$$E_1 = \{HH, HT\} \ldots \text{head on first throw}$$
$$E_2 = \{HH, TH\} \ldots \text{head on second throw}$$
$$E_3 = \{HH, TT\} \ldots \text{even number of heads}$$

Then for all $i \neq j$.

$$P(E_i \cap E_j) = \frac{1}{4} = P(E_i) \cdot P(E_j)$$

but

$$P(E_1 \cap E_2 \cap E_3) = \frac{1}{4} \neq \frac{1}{8} = P(E_1) \cdot P(E_2) \cdot P(E_3)$$

Independence

- If A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.
- The four-field table for two independent events:

	B	B'	
A	$P(A)P(B)$	$P(A)P(B')$	$P(A)$
A'	$P(A')P(B)$	$P(A')P(B')$	$P(A')$
	$P(B)$	$P(B')$	1

Independence

- The coupling can be measured by the **four-field correlation**:

Four-field Correlation

$$\rho(A,B) = \frac{P(A \cap B) - P(A)P(B)}{\sqrt{P(A)P(A')P(B)P(B')}}$$

Properties of Four-field Correlation

- $-1 \leq \rho(A,B) \leq 1$
- $\rho(A,B) = 0 \iff A$ and B independent
- $\rho(A,B) > 0 \iff A$ and B positively coupled
- $\rho(A,B) < 0 \iff A$ and B negatively coupled

Independence

- The sign of $\rho(A, B)$ indicates the **direction** of the coupling.
- The magnitude of $\rho(A, B)$ indicates the **strength** of the coupling.
- Specifically:

$$A = B \implies \rho(A, B) = 1$$
$$A = B' \implies \rho(A, B) = -1$$

- An existing coupling is no proof for a causal relationship!
- The coupling can also result from a common cause for both events.

Independence

- Two events can be considered independent if there is no connection whatsoever between them. The occurrence of one event cannot affect the probability of the other.
- Two elementary events are never independent, since their $\cap$-connection is always the impossible event.
- Two elementary events are even highly 'mutually dependent' since the occurrence of one precludes the occurrence of the other with complete certainty.
- If E_1 and E_2 are two independent events of a probability space (Σ, P), then E_1 and E_2', E_1' and E_2, and E_1' and E_2' are also independent.

Independence

Example (rolls with two distinguishable dice)

There are 36 elementary events $e_{ij} = \{(i,j)\},\ 1 \leq i,j \leq 6$. The event E_i^1 of rolling an i on the first roll is thus composed of:

$$E_i^1 = e_{i1} \cup e_{i2} \cup \ldots \cup e_{i6} \text{ and analogously}$$
$$E_j^2 = e_{1j} \cup e_{2j} \cup \ldots \cup e_{6j}$$

Clearly, $E_i^1 \cap E_j^2 = e_{ij}$.
If we can assume that all elementary events are **equally probable**, then:

$$P(E_i^1) = \frac{1}{6},\ P(E_j^2) = \frac{1}{6}$$
$$P(E_i^1 \cap E_j^2) = P(e_{ij}) = \frac{1}{36} = P(E_i^1) \cdot P(E_j^2)$$

Independence

Example (Continuation)

Thus, in this case, the elementary events of the simple throw are also **equally probable** and the two partial throws are **independent**.
Conversely, if we assume that for both partial throws the elementary events are equally probable, and that E_i^1 and E_j^2 are independent for all i and j, then the e_{ij} are equally probable.
If the partial rolls are not independent, then the e_{ij} **are no longer** equally probable, **despite** e_i and e_j being equally probable! An example of this is the 'rolling' a very large dice, which can be rotated by a mere 90^o each time. E.g. the elementary event e_{34} is then no longer possible and must therefore be assigned a probability of 0.

Independence

Example (Repetition of an alternative experiment)

The event algebra has 2^n elementary events, namely sequences of the form $(i_1, \ldots, i_n)$, $i_j = 0$ or 1. If the repetitions are independent, and p denotes the probability of 1 occurring, the probability of a sequence

$$P(\{(i_1, \ldots, i_n)\}) = p^{n_1}(1-p)^{n_0}$$

where n_0 or n_1 denotes the number of occurrences of 0 or 1, respectively. Clearly, $n_0 + n_1 = n$.

Part III

Random Variables and Distributions

Overview Part 3

Section 8: Random Variables

Subsection: Basic Terms

Basic Terms

Definition (Random Variable)

A mapping X:

$$\omega \in \Omega \mapsto x = X(\omega) \in \mathbb{R}$$

that assigns a real number to each element ω of the sample space Ω is called a (one-dimensional or univariate) **random variable**.

- If Ω is finite or countably infinite, any mapping X is allowed.
- If Ω is uncountably infinite, X must be a **measurable** mapping.
- Since the value of a random variable depends on the outcome of the experiment, we can assign probabilities to the possible values.

Basic Terms

- If the random variable X takes only finitely or countably infinitely many values, it is called **discrete**.
- If the random variable X takes a continuum of values, it is called **continuous** or **real-valued**.

Example

The mapping that assigns the number of eyes i to the elementary event e_i when rolling the dice is a discrete random variable. Of course, the mapping $e_i :\longrightarrow 7 - i$ would also be a discrete random variable.

Example

The mapping assigning a lifetime x to the decay of a particle is a continuous random variable.

Subsection: Discrete Random Variable

Discrete Random Variable

- Discrete random variables are often the result of **counting operations**.
- Discrete random variables occur in many areas: for example, counting events in a fixed time interval (Poisson distribution) or counting successes in repeated alternative trials (binomial distribution),
- In the following, we assume that the values of a discrete random variable are nonnegative integers. This is not a restriction because any countable set of real numbers can be mapped bijectively to (a subset of) $\mathbb{N}_0$.
- The event algebra is the power set (set of all subsets) P of $\mathbb{N}_0$.

Discrete Random Variable

- If a probability measure P is defined on the event algebra $\Sigma(\Omega)$, then a probability measure can also be defined on the power set P of $\mathbb{N}_0$ using the random variable X.

Definition (Distribution of a Discrete Random Variable)

Let $\Sigma(\Omega)$ be a discrete event algebra. The discrete random variable $X : \Omega \mapsto \mathbb{N}_0$ induces a probability measure on $\mathbb{N}_0$ by means of.

$$P_X(\{k\}) = P(X^{-1}(k)) = P(\{\omega | \boldsymbol{X}(\omega) = k\})$$

P_X is called the **distribution** of X, and it is discrete.

Discrete Random Variable

Example

We assign the number 0 to the even numbers of the cube, and the number 1 to the odd numbers:

$$X : \omega \mapsto \mod(\omega, 2)$$

The distribution of X is then given by

$$W_X(0) = W(X^{-1}(0)) = W(\{2,4,6\}) = \frac{1}{2}$$
$$W_X(1) = W(X^{-1}(1)) = W(\{1,3,5\}) = \frac{1}{2}$$

Discrete Random Variable

Example

We assign the sum of the numbers of eyes to the outcome of a double roll:

$$X : (i, j) \mapsto i + j$$

The values of X are the natural numbers from 2 to 12. The distribution of X is then given by

$$W_X(k) = W(X^{-1}(k)) = \sum_{i+j=k} W(\{(i,j)\}) = \begin{cases} \frac{k-1}{36}, & k \leq 7 \\ \frac{13-k}{36}, & k \geq 7 \end{cases}$$

Discrete Random Variable

- The numbers $P_X(k)$ can be regarded as function values of a spectral function f_X:

$$f_X(x) = \begin{cases} P_X(k), & \text{if } x = k \\ 0, & \text{else} \end{cases}$$

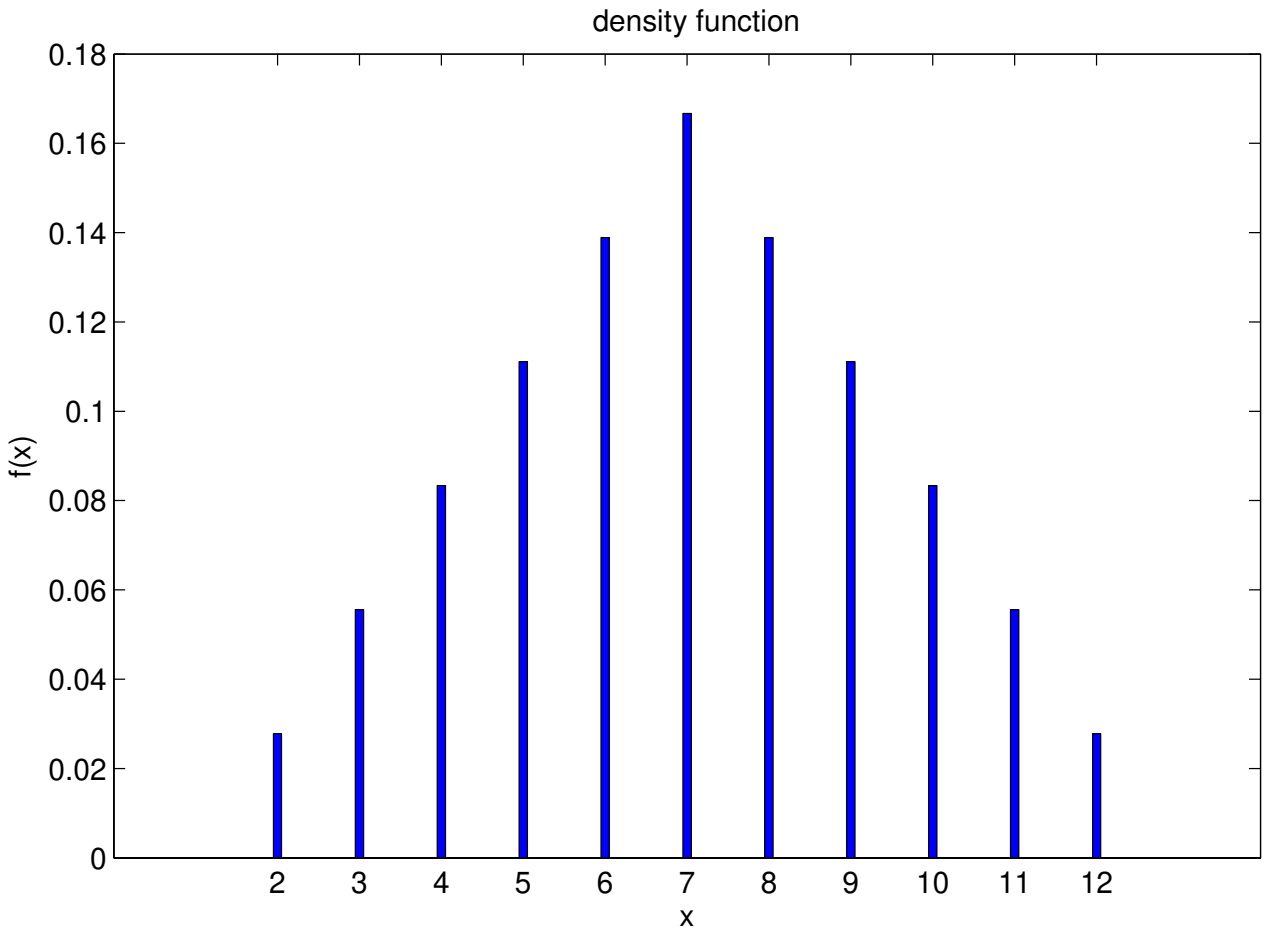
Definition (Discrete Density Function)

The function $f_X(x)$ is called **probability density function** or density of a random variable X for short

.

Discrete Random Variable

- The density of the random variable $X = i + j$:



Discrete Random Variable

- The probability $P_X(E)$ of an event E can be conveniently calculated using the density of X:

$$P_X(E) = \sum_{k \in E} f_X(k)$$

Definition (Discrete Distribution Function)

If X is a discrete random variable, the **distribution function** F_X of X is defined by:

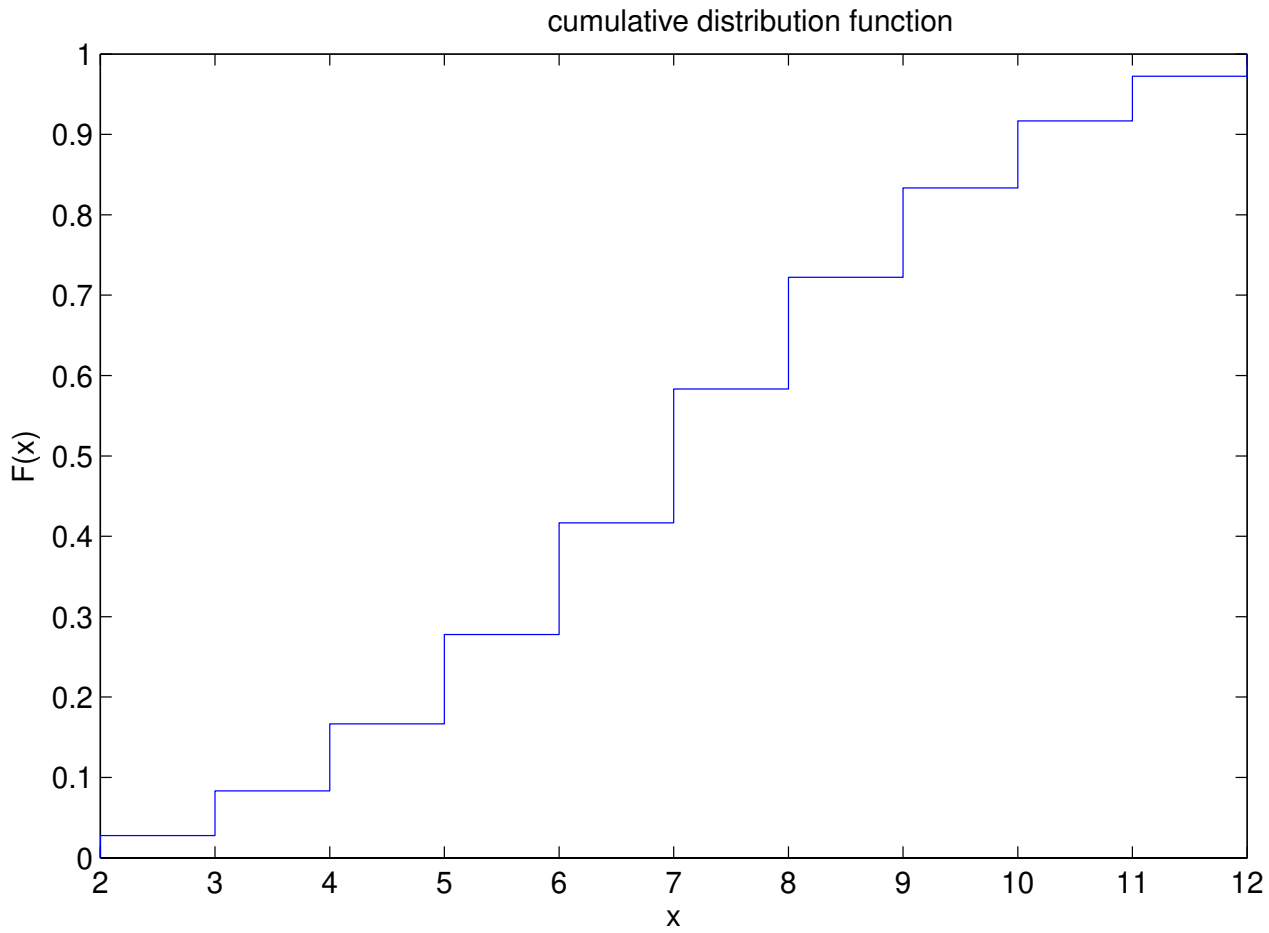
$$F_X(x) = P(X \leq x)$$

Obviously,

$$F_X(x) = \sum_{k \leq x} f_X(k) = \sum_{k \leq x} P_X(\{k\})$$

Discrete Random Variable

- The distribution function of the random variable $X = i + j$:



Discrete Random Variable

Properties of a Discrete Distribution Function F

- F has a step in all points within the range of values.
- The height of the step at point k is equal to $f_X(k)$
- $0 \leq F(x) \leq 1, \ \forall x \in \mathbb{R}$
- $x \leq y \implies F(x) \leq F(y), \ \forall x, y \in \mathbb{R}$
- $\lim_{x \to -\infty} F(x) = 0; \ \lim_{x \to \infty} F(x) = 1$
- The probability that r falls within the interval $(a, b]$ is $F(b) - F(a)$:

$$P(r \leq a) + P(a < r \leq b) = P(r \leq b) \implies$$
$$P(a < r \leq b) = F(b) - F(a)$$

Discrete Random Variable

Important Discrete Distribution Families

- Poisson distribution $\mathrm{Po}(\lambda)$
- Bernoulli or alternative distribution $\mathrm{Al}(p)$
- Binomial distribution $\mathrm{Bi}(n, p)$
- Hypergeometric distribution $\mathrm{Hy}(N, M, n)$

Subsection: Continuous Random Variable

Continuous Random Variable

- So far, only random variables defined on discrete event algebras were treated.
- Let us now drop this restriction, i.e. let us permit uncountably many elementary events. This is necessary, if we wish to describe any types of measurements.
- A function X defined on such an uncountable set of elementary events can take on arbitrary real values.

Continuous Random Variable

Definition (Continuous Distribution Function)

Let (Σ, P) be a probability space over a countable outcome space Ω. Let X be a random variable, i.e., a (measurable) function of Ω in $\mathbb{R}$. The function F_X, defined by:

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

is called the **distribution** function of X. The probability that X falls within an interval $(x, x + \delta x]$ is then:

$$P(x < X \leq x + \Delta x) = F_X(x + \Delta x) - F_X(x) = \Delta F_X.$$

Continuous Random Variable

Properties of a Continuous Distribution Function

- $0 \leq F(x) \leq 1, \ \forall x \in \mathbb{R}$
- $x_1 \leq x_2 \implies F(x_1) \leq F(x_2), \ \forall x_1, x_2 \in \mathbb{R}$
- $\lim_{x \to -\infty} F(x) = 0; \ \lim_{x \to \infty} F(x) = 1$

Definition (quantile)

Let $F_X(x)$ be a continuous distribution function. The value x_α, for which

$$F_X(x_\alpha) = \alpha, \quad 0 < \alpha < 1$$

holds, is called the α-**quantile** of the distribution of X. The function

$$x = Q_X(\alpha) = F_X^{-1}(\alpha), \quad 0 < \alpha < 1$$

is called the **quantile function** of the distribution of X.

Continuous Random Variable

Definition (quartile)

The quantiles to the values $\alpha = 0.25, 0.5, 0.75$ are called **quartiles**. The quantile to the value $\alpha = 0.5$ is called the **median** of the distribution.

- Quantiles can also be defined for discrete distributions, in which case they are not always unique.

Continuous Random Variable

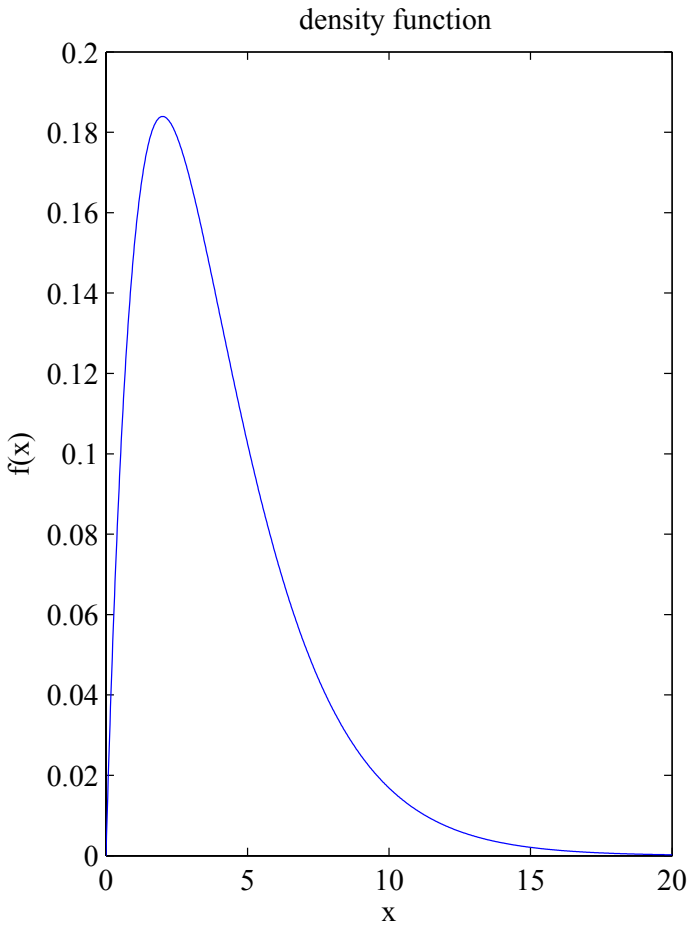
Definition (Continuous density function)

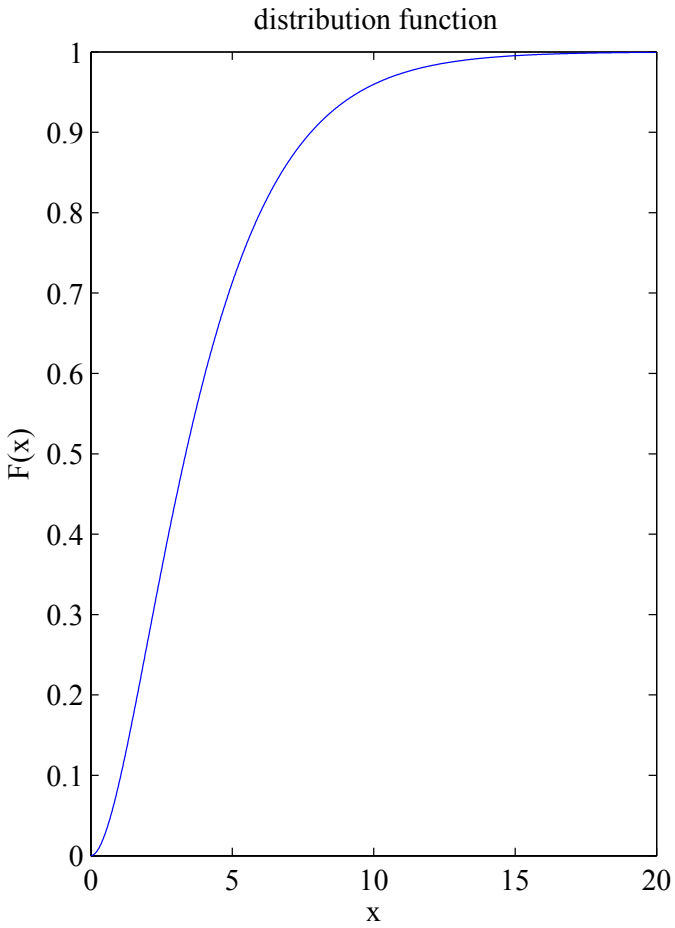
If F_X is differentiable, X is called a **continuous random variable**. For the distribution of X, according to the main theorem of integral calculus:

$$P_X(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\,\mathrm{d}x$$

where $f_X(x) = F'_X(x)$. The derivative of the distribution function, the function f_X, is called the **probability density function** or, again, the density of X for short.

Continuous Random Variable





Continuous Random Variable

- The probability measure P_X is called the distribution of X. It is defined on an event algebra Σ consisting of sets of real numbers and containing at least all intervals and their unions as elements.
- Analogously to discrete random variables, the probability P_X of a set $M \in \Sigma$ can be easily specified using density:

$$P_X(M) = \int_M f_X(x)\,\mathrm{d}x$$

- The probability of a single point is always equal 0:

$$P_X(\{x\}) = \int_x^x f_X(x)\,\mathrm{d}x = 0$$

Continuous Random Variable

- Therefore

$$P_X((x_1, x_2]) = P_X((x_1, x_2)) = P_X([x_1, x_2]).$$

- More generally, a statement about continuous random variables is obtained by replacing summation with integration in a statement about discrete random variables.
- E.g. for a discrete density f it is true that:

$$\sum_{k \in \mathbb{N}_0} f(k) = 1$$

similarly, for a continuous density f:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

Continuous Random Variable

Important continuous distribution families

- Normal distribution $\mathrm{No}(\mu, \sigma^2)$
- Gamma distribution $\mathrm{Ga}(a, b)$, with the special cases
  - Exponential distribution $\mathrm{Ex}(\tau)$
  - Chisquare distribution $\chi^2(n)$
- Student or t distribution $\mathrm{t}(n)$
- Beta distribution $\mathrm{Be}(a, b)$
- Uniform distribution $\mathrm{Un}(a, b)$

Subsection: Independence

Independence

Definition (Independent Random Variable)

Two random variables X and Y are called independent if for all pairs (A, B) of events:

$$P(X \in A \cap Y \in B) = P_X(A) \cdot P_Y(B)$$

Definition (Joint Distribution Function)

The joint distribution function F_{XY} of X and Y is defined by.

$$F_{XY}(x, y) = P(X \leq x \cap Y \leq y)$$

Independence

Definition (Common Density)

If X and Y are discrete random variables, their joint density is defined by.

$$f_{XY}(x, y) = P(X = x \cap Y = y)$$

If X and Y are continuous random variable, their joint density is defined by

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}}{\partial x \, \partial y}$$

Independence

Theorem

Two random variables X and Y are independent if and only if their joint distribution function is the product of their individual distribution functions:

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$$

Theorem

Two random variables X and Y are independent if and only if their joint density is the product of their individual density functions:

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$$

Subsection: Convolution

Convolution

Definition (Convolution)

Let X and Y be two **independent** random variables and $Z = X + Y$ be their sum. The distribution of Z is called the **convolution** of the distributions of X and Y.

Example

An experiment measures the lifetime X of an unstable particle, with a measurement error Y. If X and Y are independent, the distribution of the observation $Z = X + Y$ is the convolution of the distributions of X and Y, respectively.

Convolution

Theorem

Let X and Y be two **independent** random variable and $Z = X + Y$ their sum. Then the density of Z is the **convolution product** of the densities of X and Y.

- If X and Y are discrete, then:

$$f_Z(n) = \sum_k f_X(n-k) f_Y(k)$$

- If X and Y are continuous, then:

$$f_Z(z) = \int_{\mathbb{R}} f_X(z-y) f_Y(y)\,\mathrm{d}y$$

It should be noted that the effective integration range may depend on z.

Section 9: Moments

Subsection: Expectation

Expectation

Definition (Expectation)

Let X be a (discrete or continuous) random variable with density $f(x)$. Further, let g be any continuous real or complex function. Let $\mathsf{E}_X[g] = \mathsf{E}[g(X)]$ be defined by:

$$\mathsf{E}[g(X)] = \sum_{k \in \mathbb{N}_0} g(k) f(k) \quad \text{or} \quad \mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) \, \mathrm{d}x$$

$\mathsf{E}_X[g] = \mathsf{E}[g(X)]$ is called the **expectation** of $g(X)$.

Expectation

Definition (expectation of random variables)

If $g(x) = x$, then $\mathsf{E}[g(X)] = \mathsf{E}[X]$ is called the **expectation** or the **mean** of X.

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \text{ but } \mathsf{E}[X] = \sum_{k \in \mathbb{N}_0} k\, f(k)$$

Properties of the Expectation

- $\mathsf{E}[c] = c,\ c \in \mathbb{R}$
- $\mathsf{E}[aX + b] = a\mathsf{E}[X] + b$
- $\mathsf{E}[X_1 + X_2] = \mathsf{E}[X_1] + \mathsf{E}[X_2]$
- X_1 and X_2 independently $\implies \mathsf{E}[X_1 X_2] = \mathsf{E}[X_1] \cdot \mathsf{E}[X_2]$

Expectation

- The expectation is a **location parameter**.
- The expectation need not exist. An example is the Cauchy distribution with the density

$$f(x) = \frac{1}{\pi(1+x^2)}, \; x \in \mathbb{R}$$

In this case, the median is a suitable location parameter.

Expectation

Definition (moments)

Let X be a random variable. The expectation of $g(x) = (x-a)^k$, if it exists, is called k-**th moment of** X **around** a. The k-th moment around 0 is denoted by μ'_k. The k-th moment around the expected value $\mathrm{E}[X]$ is called **central moment** μ_k.

- The expectation is the first moment around 0.
- The central moments $\mu_1, \ldots, \mu_k$ can be calculated from the moments around 0 $\mu'_1, \ldots, \mu'_k$, and vice versa.
- Even if all moments of a distribution exist, it is still **not** uniquely determined.

Subsection: Variance

Variance

Definition (Variance)

The second central moment μ_2 is called the **variance** of X, denoted by $\mathrm{var}[X]$. The root of the variance is called the **standard deviation** of X, denoted by $\sigma[X]$.

- The standard deviation is a scale parameter that describes the width of the distribution.
- The standard deviation has the same dimension as the values of the random variables.
- Variance and standard deviation are (like all central moments) invariant against translations.
- The variance need not exist. An example is the distribution with density

$$f(x) = \frac{1}{(2+x^2)^{3/2}}, \; x \in \mathbb{R}$$

Variance

Properties of the Variance

- $\mathrm{var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$
- $\mathrm{var}[aX + b] = a^2 \mathrm{var}[X]$
- For independent $X_1, \ldots, X_n$:

$$\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{var}[X_i]$$

Subsection: Skewness

Skewness

Definition (Skewness)

The reduced third central moment $\gamma = \mu_3/\sigma^3$ is called the **skewness**.

- The skewness measures the asymmetry of a distribution. If the skewness is positive (negative), the distribution is called right skewed (left skewed). For symmetric distributions, it is 0.

Skewness

Definition (Kurtosis)

The reduced fourth central moment $\kappa = \mu_4/\sigma^4$ is called the **curvature** or **kurtosis**.

- The kurtosis of the normal distribution is 3.
- Distributions with higher kurtosis have relatively stronger margins than the normal distribution.

Subsection: Error Propagation

Error Propagation

Affine Transformations

- Let X be a random variable with density $f(x)$ and $Y = aX + b$.
- Then the density $g(y)$ of Y is equal to

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$$

- Furthermore.

$$\begin{aligned} \mathsf{E}[Y] &= a\mathsf{E}[X] + b \\ \mathsf{var}[Y] &= a^2 \mathsf{var}[X] \\ \gamma[Y] &= \mathrm{sgn}(a)\gamma[X] \\ \kappa[Y] &= \kappa[X] \end{aligned}$$

Error Propagation

Nonlinear Transformations

- Let X be a random variable with density $f(x)$ and $Y = h(X)$.
- If $h(x)$ is bijective, the density $g(y)$ of Y is equal to

$$g(y) = \left| \frac{\mathrm{d}h^{-1}}{\mathrm{d}y} \right| f\left(h^{-1}(y)\right)$$

- The expectation and variance of $Y = h(X)$ can be calculated **approximately** using the Taylor expansion of $h(x)$.

Error Propagation

- With the development place x_0 applies in linear approximation

$$h(x) \approx h(x_0) + h'(x_0)(x - x_0)$$

- With the choice $x_0 = \mathsf{E}[X]$ follows

Theorem

$$\mathsf{E}[h(X)] \approx h(\mathsf{E}[X])$$
$$\mathsf{var}[h(X)] \approx h'(\mathsf{E}[X])^2 \cdot \mathsf{var}[X] \quad \text{(linear error propagation)}$$

- If the evolution is continued to the 2nd order, the improved approximation is obtained.

Theorem

$$\mathsf{E}[h(X)] \approx h(\mathsf{E}[X]) + \frac{1}{2} h''(\mathsf{E}[X]) \cdot \mathsf{var}[X]$$

Section 10: Functions of Random Samples

Subsection: Basic Terminology

Basic Terminology

- Let $X_1, \ldots, X_n$ be independent random variables that all have the same distribution F.
- They then form a **random sample** of the distribution F.
- A random variable

$$Y = h(X_1, \ldots, X_n)$$

  is called a **sampling function**.
- In many cases, moments or the distribution of Y are to be determined.

Subsection: Sample Mean

Sample Mean

Definition (Sample Means)

The **sample mean** $\overline{X}$ of the sample $X_1, \ldots, X_n$ is defined by.

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

Moments of the Sample Mean

If F has the mean μ and the variance σ^2, the following holds true

- $\mathsf{E}[\overline{X}] = \mu$
- $\mathsf{var}[\overline{X}] = \frac{\sigma^2}{n}$

Subsection: Sample Variance

Sample Variance

Definition (Sample Variance)

The **sample variance** S^2 of the sample $X_1, \ldots, X_n$ is defined by.

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

Moments of the Sample Variance

If F has variance σ^2, the following holds.

$$\mathsf{E}[S^2] = \sigma^2$$

If the fourth central moment μ_4 of F exists, holds

$$\mathsf{var}[S^2] = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$$

Subsection: Sample Median

Sample Median

Definition (Sample Median)

The **sample median** $\tilde{X}$ of the sample $X_1, \ldots, X_n$ is defined by.

$$\tilde{X} = \begin{cases} X_{((n+1)/2)}, & n \text{ odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right), & n \text{ even} \end{cases}$$

Moments of the Sample Median

If F has median m and density f, holds:

- $\lim_{n\to\infty} \mathsf{E}[\tilde{X}] = m$
- $\lim_{n\to\infty} \mathsf{var}[\tilde{X}] = \frac{1}{4\,n f^2(m)}, \quad \text{if } f(m) > 0$

Part IV

Point Estimators

Overview Part 4

Section 11: Point Estimators

Subsection: Basic Terminology

Basic Terminology

- A point estimator is a sampling function designed to provide as accurate an approximation as possible for an unknown distribution parameter ϑ:

$$T = g(X_1, \ldots, X_n)$$

- The function $g(x_1, \ldots, x_n)$ is called the estimator.
- The construction of reasonable point estimators for a parameter ϑ is the task of estimation theory.
- Many point estimators are possible for a parameter ϑ. However, a "'good'" point estimator should meet certain requirements.

Basic Terminology

Definition (Bias)

A point estimator T for the parameter ϑ is called **unbiased** if for all values of ϑ:

$$\mathsf{E}_\vartheta[T] = \vartheta$$

T is called **asymptotically unbiased** if:

$$\lim_{n\to\infty} \mathsf{E}_\vartheta[T] = \vartheta$$

- If the unknown parameter is equal to ϑ, then the expectation of the point estimator is equal to ϑ.
- An unbiased point estimator has random deviations from the true value ϑ, but no systematic bias.

Basic Terminology

Definition (MSE)

The **mean squared error (MSE)** of a point estimator T for parameter ϑ is defined by:

$$\mathrm{MSE}[T] = \mathrm{E}_\vartheta[(T-\vartheta)^2]$$

Definition (MSE-consistency)

A point estimate T for parameter ϑ is called **consistent in the mean square (MSE-consistent)** if holds:

$$\lim_{n\to\infty} \mathrm{MSE}[T] = 0$$

Basic Terminology

Definition (MSE-efficiency)

A point estimator T_1 is called **MSE-efficient** than the point estimator T_2 if for all admissible ϑ holds:

$$\mathrm{MSE}[T_1] \leq \mathrm{MSE}[T_2]$$

Definition (Efficiency)

An unbiased point estimator T_1 is called **efficient** than the unbiased point estimator T_2 if for all admissible ϑ holds:

$$\mathrm{var}[T_1] \leq \mathrm{var}[T_2]$$

An unbiased point estimator T is called **efficient** if its variance takes the smallest possible value.

Basic Terminology

Definition (Fisher Information)

Let $X_1, \ldots, X_n$ be a sample with joint density $g(x_1, \ldots, x_n|\vartheta)$. The expectation

$$I_\vartheta = \mathsf{E}\left[-\frac{\partial^2 \ln g(X_1, \ldots, X_n|\vartheta)}{\partial \vartheta^2}\right]$$

is called the **Fisher information** of the sample.

Theorem of Rao and Cramèr

Let $X_1, \ldots, X_n$ be a sample with joint density $g(x_1, \ldots, x_n|\vartheta)$. The variance of an unbiased estimator T for the parameter ϑ is bounded below by:

$$\mathsf{var}[T] \geq 1/I_\vartheta$$

Subsection: Sample Moments and Sample Median

Sample Moments and Sample Median

Theorem

- Let $X_1, \ldots, X_n$ be a sample from the distribution F with expectation μ. Then the sample mean $\overline{X}$ is an unbiased point estimator of μ.
- If F has finite variance σ^2, then.

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n}$$

- In this case, $\overline{X}$ is MSE consistent.

Sample Moments and Sample Median

Theorem

- Let $X_1, \ldots, X_n$ be a sample from the distribution F with expectation μ and variance σ^2. Then the sample variance S^2 is an unbiased point estimator of σ^2.
- If F has finite fourth central moment μ_4, then.

$$\mathrm{var}(S^2) = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}$$

- In this case S^2 is MSE-consistent.

Sample Moments and Sample Median

Theorem

- Let $X_1, \ldots, X_n$ be a sample from the continuous distribution F with median m and density f. Then the sample median $\tilde{X}$ is an asymptotically unbiased point estimator of m.
- For symmetric F, $\tilde{X}$ is unbiased.
- The sample median $\tilde{X}$ has asymptotically the variance

$$\text{var}(\tilde{X}) \approx \frac{1}{4nf(m)^2}$$

- The sample median is MSE-consistent provided $f(m) > 0$.

Section 12: Normally Distributed Data

Subsection: The Gauß or Normal Distribution

The Gauß or Normal Distribution

The Normal Distribution $\mathrm{No}(\mu, \sigma^2)$

- The **normal distribution** is one of the most important families of distributions in science and engineering. We denote it by $\mathrm{No}(\mu, \sigma^2)$.
- Its density is:

$$f_{\mathrm{No}}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The normal distribution with $\mu = 0$, $\sigma = 1$ is called **standard normal distribution**. Its density is often notated as $\varphi(x)$.
- The distribution function $\Phi(x)$ is not representable by elementary functions.
- The mode M (the maximum of the density function) and the median m are at $x = \mu$.

The Gauß or Normal Distribution

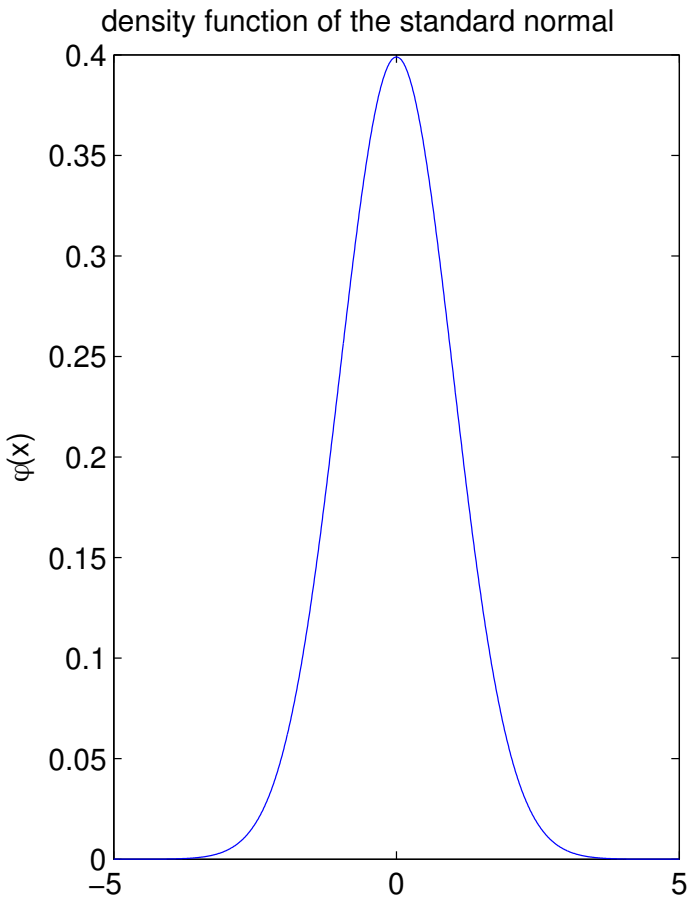
The moments and CF of the normal distribution

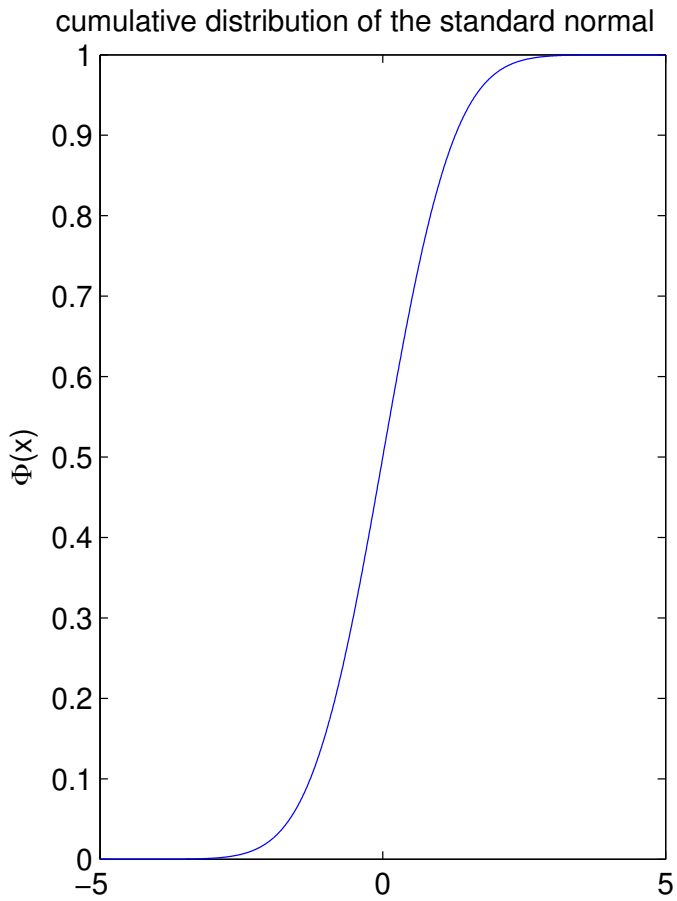
Let $X \sim \mathrm{No}(\mu, \sigma^2)$. Then holds:

- $\mathrm{E}[X] = \mu$
- $\mathrm{var}[X] = \sigma^2$
- $\gamma[X] = 0$
- $\kappa[X] = 3$
- $\varphi_X(t) = \exp\left(\mathrm{i}\mu t - \sigma^2 t^2/2\right)$

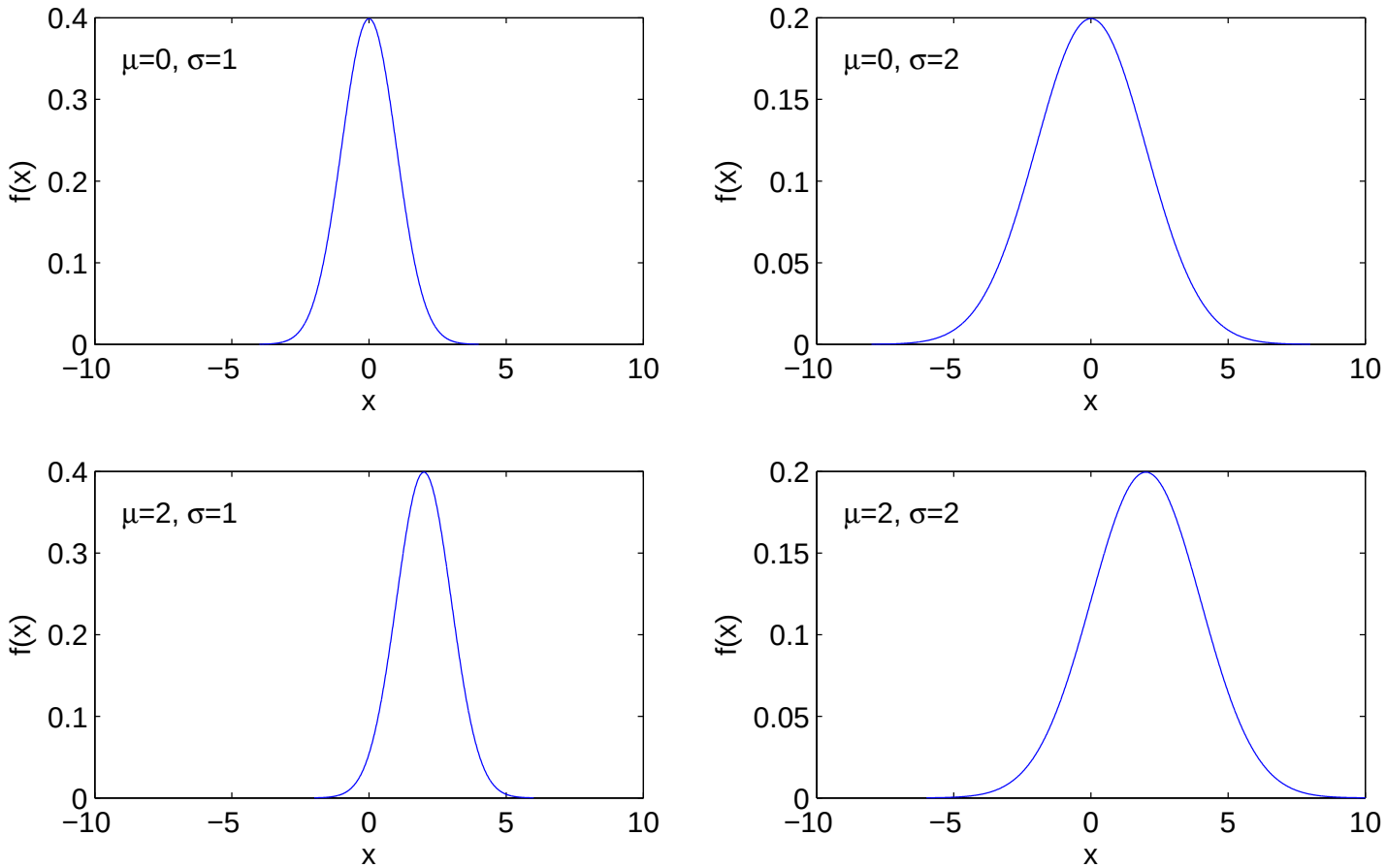
- The α quantile of the standard normal distribution is denoted by z_α.
- The α quantile of $\mathrm{No}(\mu, \sigma^2)$ is equal to $\mu + \sigma z_\alpha$.

The Gauß or Normal Distribution



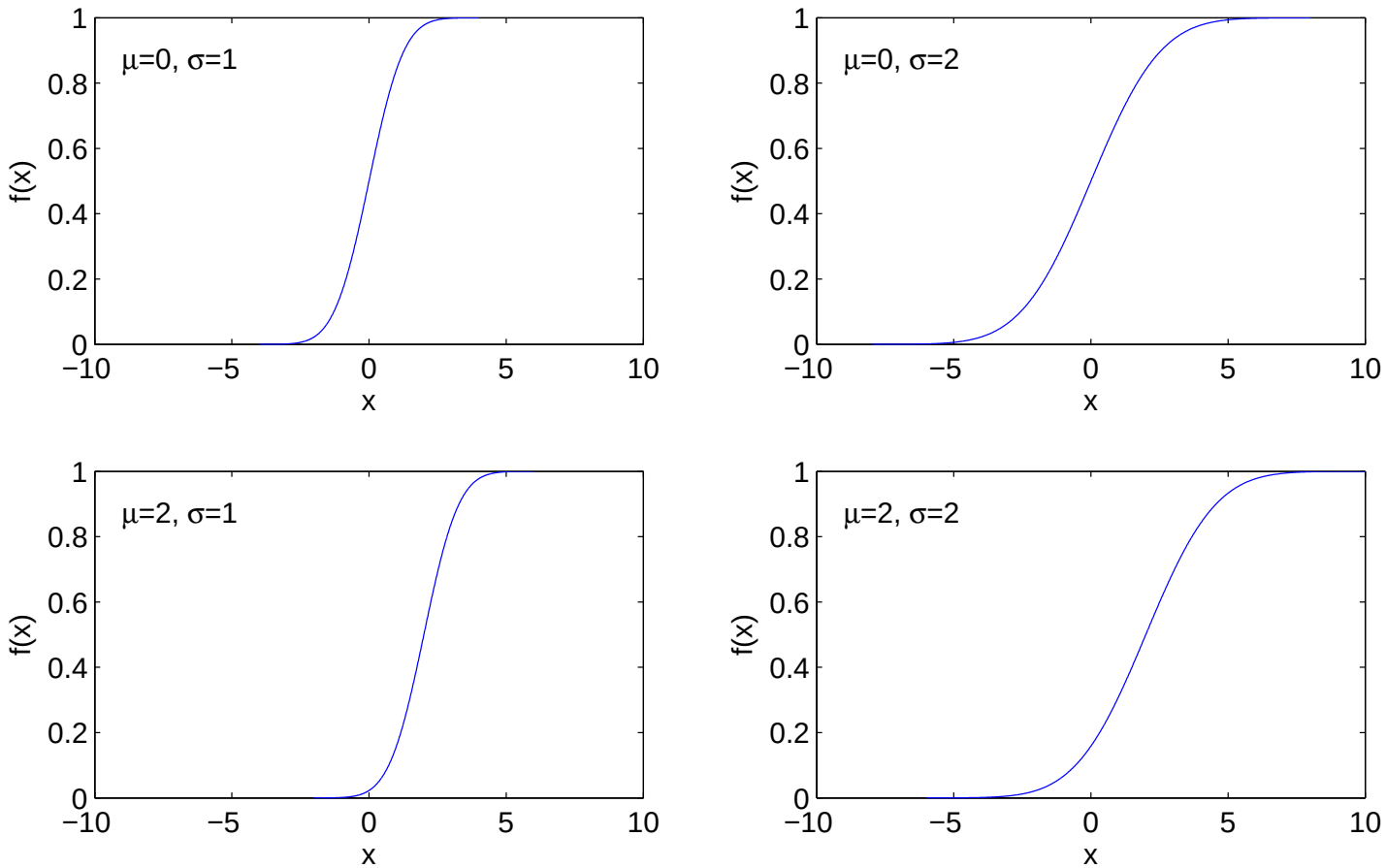


The Gauß or Normal Distribution



four normal distributions (density functions)

The Gauß or Normal Distribution



four normal distributions (distribution functions)

The Gauß or Normal Distribution

- One of the reasons for the great importance of the normal distribution is the **central limit theorem**. In its simplest formulation it reads:

Central Limit Theorem

Let $X_1, \ldots, X_n$ be a sample from the distribution F with mean μ and variance σ^2. Then the distribution of

$$U_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

for $n \to \infty$ converges to the standard normal distribution.

- Clearly, if F is a normal distribution, then U_n is exactly standard normally distributed for all n.

The Gauß or Normal Distribution

Definition (StandardScore)

Let X be a normally distributed random variable with mean μ and variance σ^2. Then

$$Z = \frac{X - \mu}{\sigma}$$

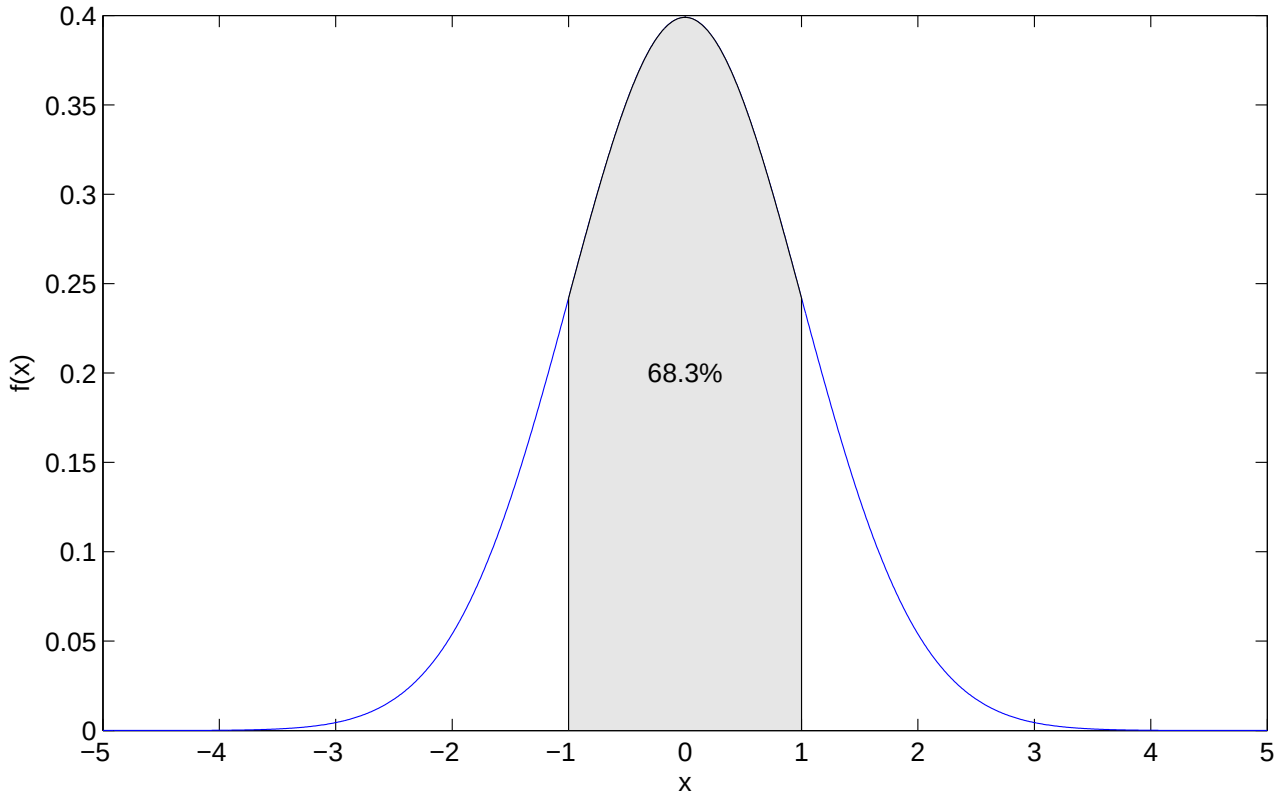
the **standard score** of X.

Theorem

If X is normally distributed, then the standard score Z is standard-normally distributed.

The Gauß or Normal Distribution

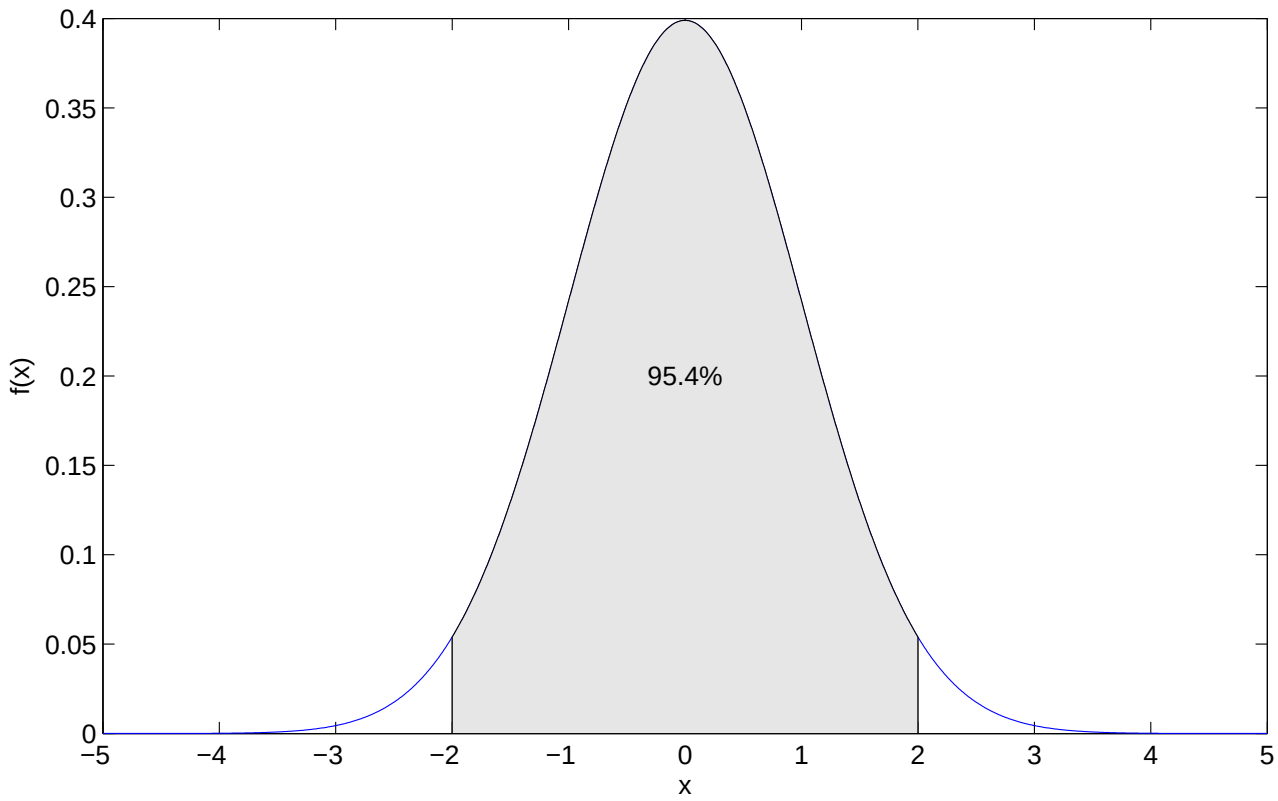
- Probability contents of the standard normal distribution



$$W(-1 \leq X \leq 1)$$

The Gauß or Normal Distribution

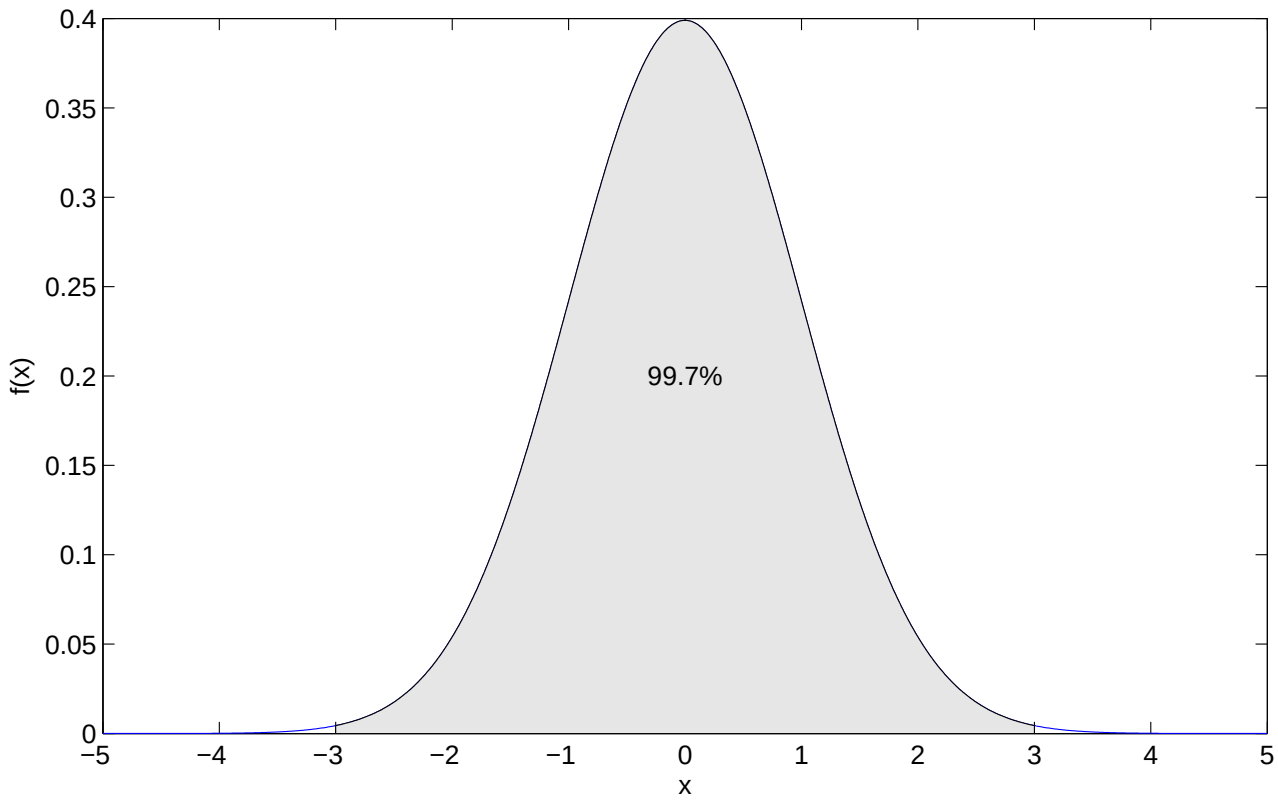
- Probability contents of the standard normal distribution



$$W(-2 \leq X \leq 2)$$

The Gauß or Normal Distribution

- Probability contents of the standard normal distribution



$$W(-3 \leq X \leq 3)$$

The Gauß or Normal Distribution

Fisher information of a normally distributed sample

Let $X_1, \ldots, X_n$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$. Then holds:

- $I_\mu = \frac{n}{\sigma^2}$
- $I_{\sigma^2} = \frac{n}{2\sigma^4}$

Subsection: Estimate of the Mean

Estimate of the Mean

Theorem

Let $X_1, \ldots, X_n$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$ and $\overline{X}$ be the sample mean. Then:

- $\overline{X}$ is normally distributed according to $\mathrm{No}(\mu, \sigma^2/n)$.
- $\overline{X}$ is an unbiased and consistent estimator of μ.
- $\overline{X}$ is efficient.

Subsection: Estimate of the Variance

Estimate of the Variance

Theorem

Let $X_1, \ldots, X_n$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$ and S^2 be the sample variance. Then holds:

- $(n-1)S^2/\sigma^2$ is χ^2-distributed with $n-1$ degrees of freedom.
- $\mathsf{E}[S^2] = \sigma^2$
- $\mathsf{var}[S^2] = \dfrac{2\sigma^4}{n-1}$
- S^2 is an unbiased and consistent estimator of σ^2.
- S^2 is asymptotically efficient.

Estimate of the Variance

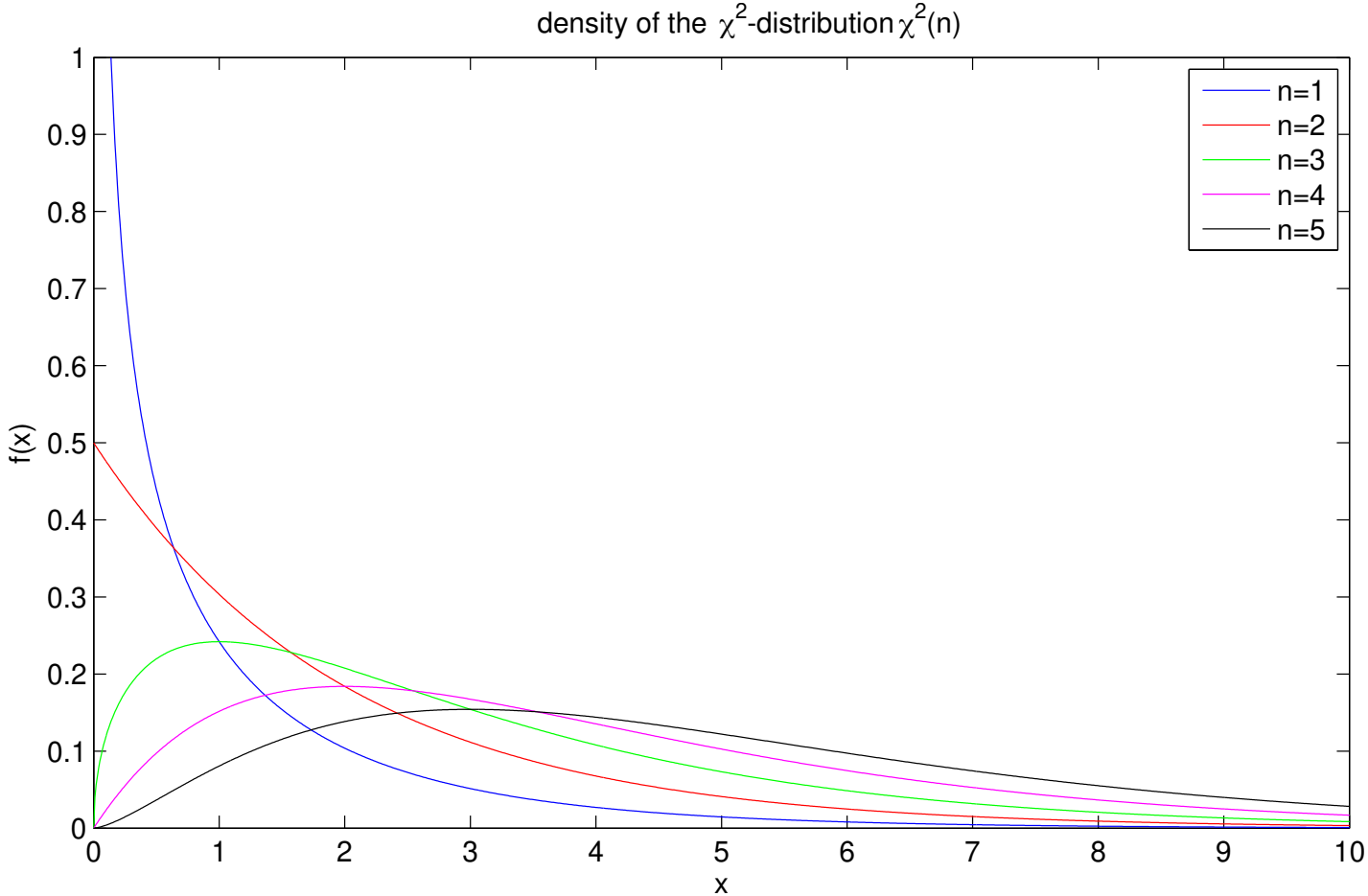
Excursion: The χ^2 distribution $\chi^2(n)$

- The density of the $\chi^2(n)$ distribution with n degrees of freedom is:

$$f_{\chi^2}(x;n) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{n/2-1} e^{-x/2} \cdot I_{[0,\infty)}(x)$$

- It is the gamma distribution $\mathrm{Ga}(n/2, 2)$.
- The $\chi^2(2)$ distribution is the exponential distribution $\mathrm{Ex}(2)$.
- The mode of the density is at $x = \max(n-2, 0)$.
- The density of the $\chi^2(1)$ distribution has a pole at $x = 0$.
- For $n \to \infty$, the $\chi^2(n)$ distribution converges to the normal distribution $\mathrm{No}(n, 2n)$.
- If $X_1, \ldots, X_n$ are independent and standard normally distributed, then $Y = \sum_{i=1}^{n} X_i^2$ is $\chi^2(n)$-distributed with n degrees of freedom.

Estimate of the Variance



Estimate of the Variance

The moments and the CF of the $\chi^2(n)$ distribution

Let $X\chi^2(n)$. Then holds:

- $\mathrm{E}[X] = n$
- $\mathrm{var}[X] = 2n$
- $\gamma[X] = \sqrt{8/n}$
- $\kappa[X] = 3 + 12/n$
- $\varphi_X(t) = (1 - 2\mathrm{i}t)^{-n/2}$

- The α quantile of the $\chi^2(n)$ distribution is denoted by $\chi^2_{\alpha;n}$.

Estimate of the Variance

Excursion: The gamma distribution $\mathrm{Ga}(a, b)$

- The density of the gamma distribution is:

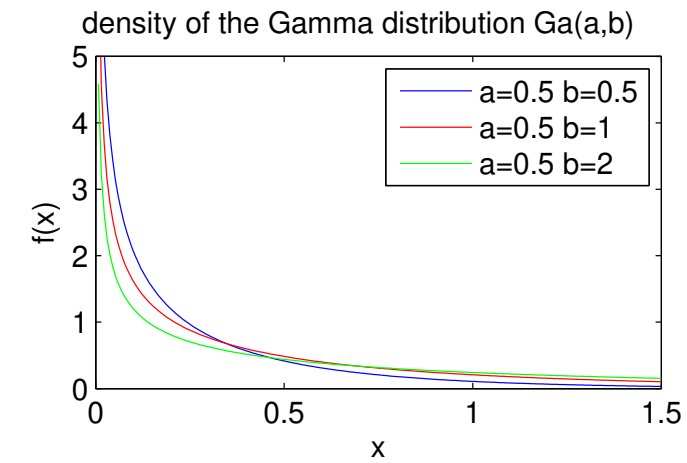
$$f_{\mathrm{Ga}}(x; a, b) = \frac{x^{a-1} e^{-x/b}}{b^a \Gamma(a)} \cdot I_{[0,\infty)}(x)$$

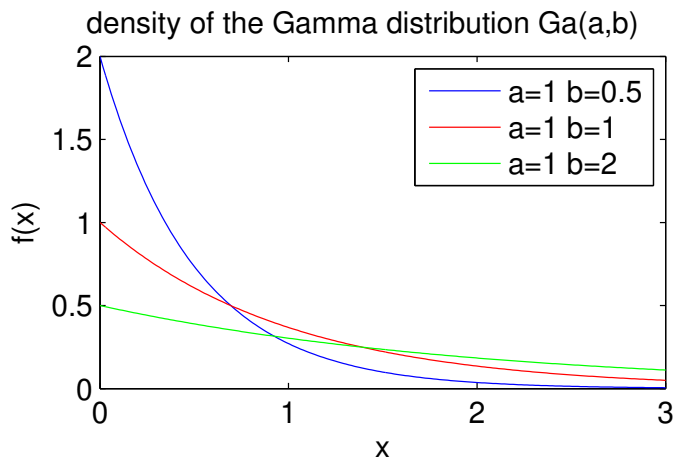
- Its distribution function is the regularized incomplete gamma function:

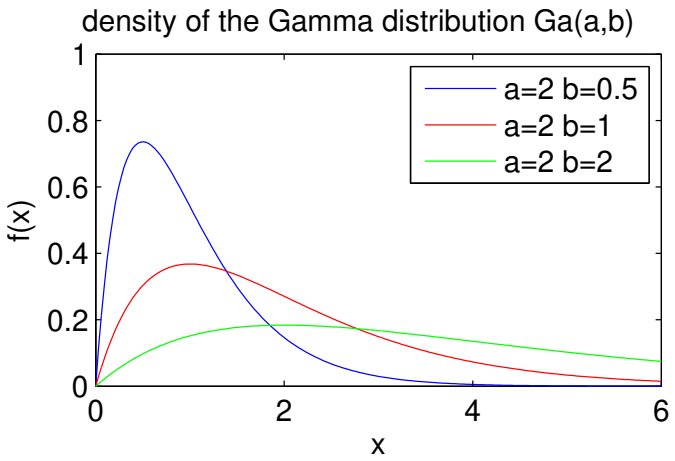
$$F_{\mathrm{Ga}}(x; a, b) = \int_0^x \frac{t^{a-1} e^{-t/b}}{b^a \Gamma(a)} \, \mathrm{d}t = \frac{\gamma(a, x/b)}{\Gamma(a)}$$

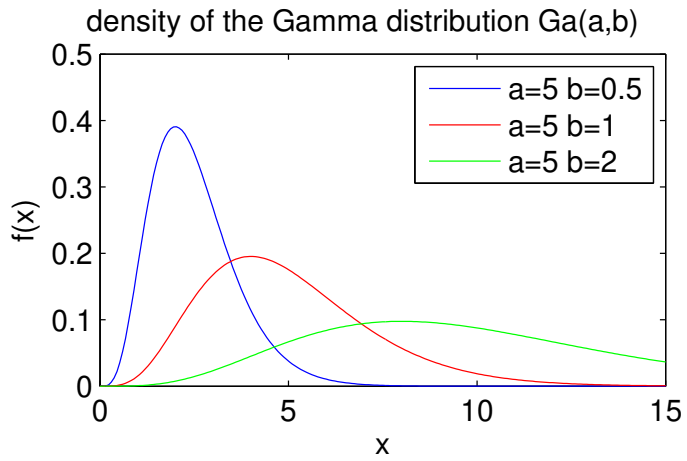
- The mode M is at $m = (a-1)b$ when $a \geq 1$.
- The α quantile of the $\mathrm{Ga}(a, b)$ distribution is denoted by $\gamma_{\alpha;a,b}$.

Estimate of the Variance









Estimate of the Variance

The moments and CF of the $\mathrm{Ga}(a, b)$ distribution

let $X \sim \mathrm{Ga}(a, b)$. The following is then true:

- $\mathsf{E}[X] = ab$
- $\mathsf{var}[X] = ab^2$
- $\gamma[X] = 2/\sqrt{a}$
- $\kappa[X] = 3 + 6/a$
- $\varphi_X(t) = (1 - \mathrm{i}bt)^{-a}$

Other properties of the $\mathrm{Ga}(a, b)$ distribution

- If $X \sim \mathrm{Ga}(a, b)$, then $cX \sim \mathrm{Ga}(a, cb)$
- $\gamma_{\alpha;a,b} = b\,\gamma_{\alpha;a,1}$
- If $X_1 \sim \mathrm{Ga}(a_1, b)$ and $X_2 \sim \mathrm{Ga}(a_2, b)$ are independent, $Y = X_1 + X_2 \sim \mathrm{Ga}(a_1 + a_2, b)$

Subsection: Estimate of the Median

Estimate of the Median

Theorem

Let $X_1, \ldots, X_n$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$ and $\tilde{X}$ be the sample median. Then holds:

- $\mathrm{E}[\tilde{X}] = \mu$
- $\mathrm{var}[\tilde{X}] \approx \frac{2\,\pi\sigma^2}{4\,n} \approx 1.57 \cdot \frac{\sigma^2}{n}$

Thus, for large n, the variance of $\tilde{X}$ is more than 50 percent larger than the variance of $\overline{X}$.

- However, there are distributions for which the sample median has a smaller variance than the sample mean, as shown in the following example.

Estimate of the Median

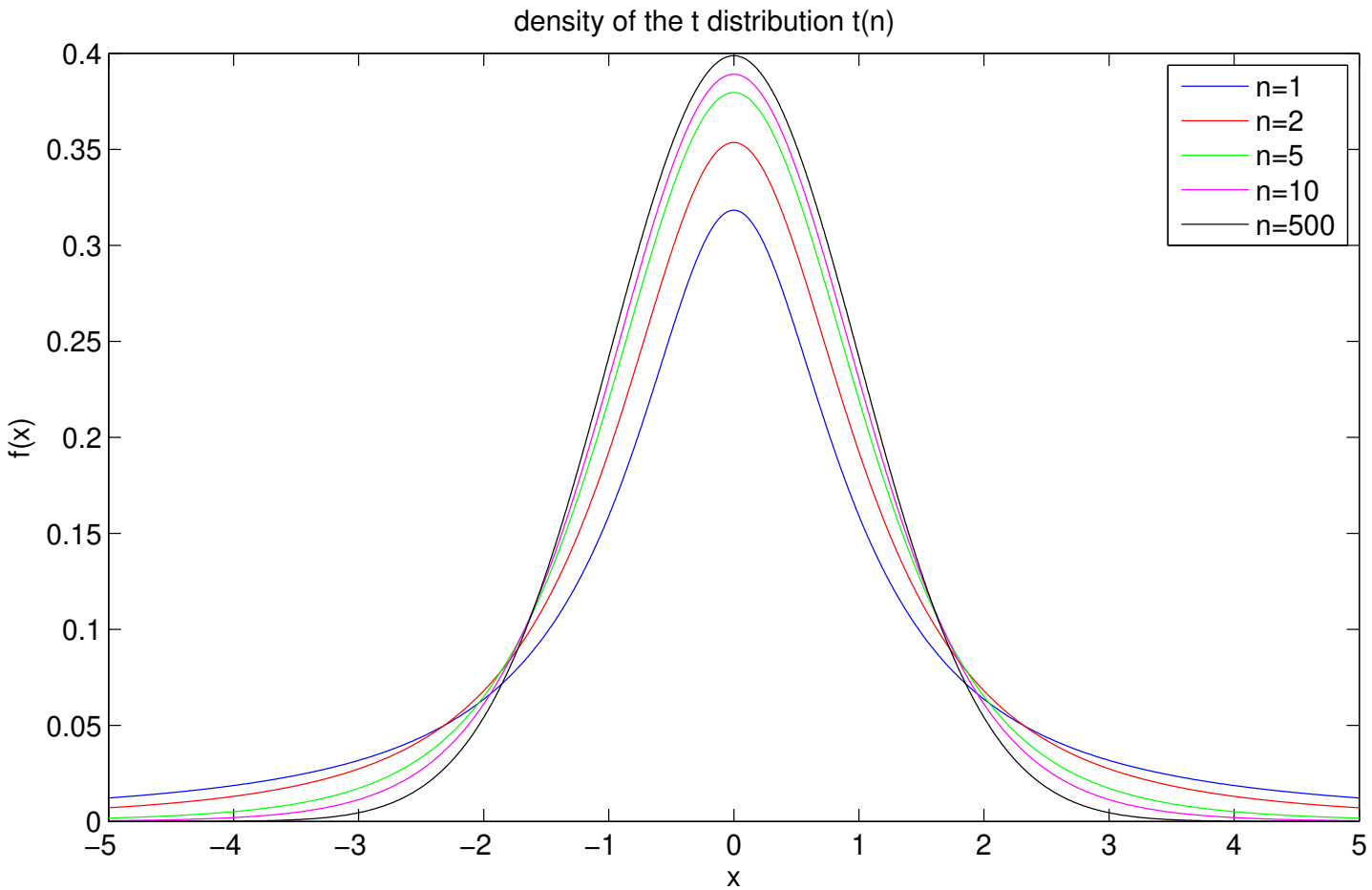
Exkurs: The t distribution $\mathrm{t}(n)$

- The density of the $\mathrm{t}(n)$ distribution with n degrees of freedom is:

$$f_{\mathrm{t}}(x;n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

- The $\mathrm{t}(1)$ distribution is also called the Cauchy or (in particle physics) Breit–Wigner distribution.
- The k-th moment μ_k exists only for $k < n$.
- For $n \to \infty$, the $\mathrm{t}(n)$ distribution converges to the standard normal distribution.

Estimate of the Median



Estimate of the Median

The moments of the $\mathrm{t}(n)$ distribution

Let $X \sim \mathrm{t}(n)$. Then holds:

- $\mathrm{E}[X] = 0, \quad n > 1$
- $\mathrm{var}[X] = \dfrac{n}{n-2}, \quad n > 2$
- $\gamma[X] = 0, \quad n > 3$
- $\kappa[X] = \dfrac{3n-6}{n-4}, \quad n > 4$

- The α quantile of the $\mathrm{t}(n)$ distribution is denoted by $\mathrm{t}_{\alpha;n}$.

Estimate of the Median

Example

Let $X_1, \ldots, X_n$ be a sample from the t distribution $\mathrm{t}(3)$. The variance of $\overline{X}$ is equal to.

$$\mathrm{var}(\overline{X}) = \frac{3}{n}$$

The variance of $\tilde{X}$ for large n is equal to

$$\mathrm{var}(\tilde{X}) = \frac{1}{4\,n f(0)^2} = \frac{1.8506}{n} \approx 0.62 \cdot \frac{3}{n}$$

Thus, it is almost 40 percent smaller than the variance of $\overline{X}$.

Section 13: Exponentially Distributed Data

Subsection: The Exponential Distribution

The Exponential Distribution

The Exponential Distribution $\mathrm{Ex}(\tau)$

- The **exponential** or **waiting time distribution** is an important family of distributions in science and engineering. We denote it by $\exp(\tau)$.
- Its density is:

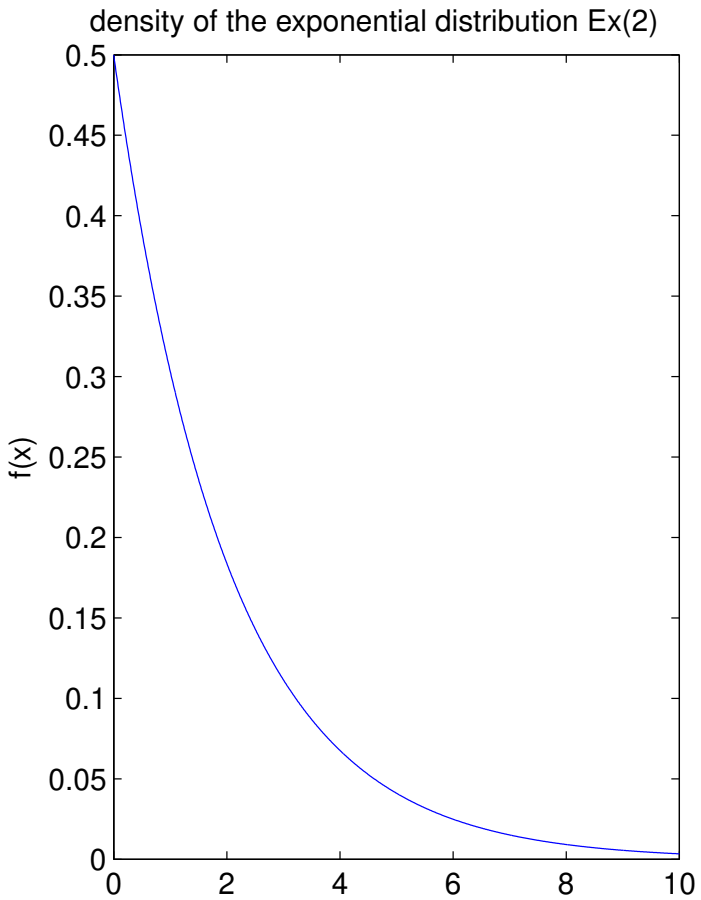
$$f_{\mathrm{Ex}}(x;\tau) = \frac{1}{\tau}\, e^{-x/\tau}$$

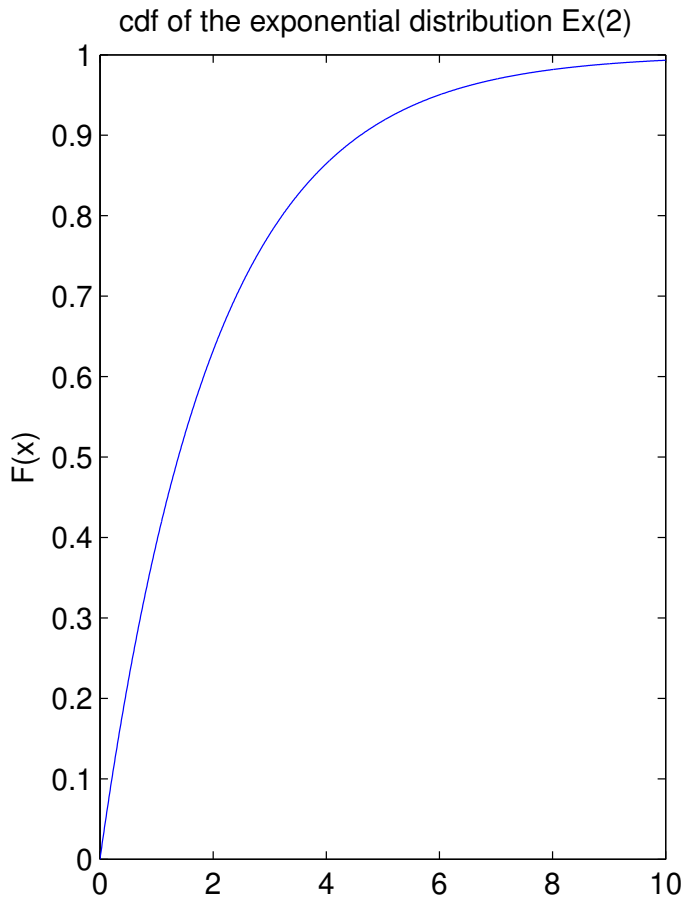
- Its distribution function is:

$$F_{\mathrm{Ex}}(x;\tau) = 1 - e^{-x/\tau}$$

- The exponential distribution is a special case of the **gamma distribution**: $\mathrm{Ex}(\tau)$ is equal to $\mathrm{Ga}(1,\tau)$.
- The mode M is at $x = 0$.

The Exponential Distribution





The Exponential Distribution

- The α quantile of the exponential distribution $\mathrm{Ex}(\tau)$ is:

$$\gamma_{\alpha;1,\tau} = -\tau \ln(1-\alpha)$$

- Consequently, the median of the distribution is at $x = \tau \ln 2$.
- The exponential distribution is the distribution of a waiting time **without memory**: The distribution is independent of the time at which the timing starts.
- This behavior is typical for physical processes like the radioactive decay of an atom or the decay of an unstable elementary particle.

The Exponential Distribution

The moments and the CF of the exponential distribution

let $X \sim \mathrm{Ex}(\tau)$. Then holds:

- $\mathrm{E}[X] = \tau$
- $\mathrm{var}[X] = \tau^2$
- $\gamma[X] = 2$
- $\kappa[X] = 9$
- $\varphi_X(t) = (1 - \mathrm{i}\tau t)^{-1}$

Fisher information of an exponentially distributed sample

Let $X_1, \ldots, X_n$ be a sample from the exponential distribution $\mathrm{Ex}(\tau)$. Then holds:

$$I_\tau = \frac{n}{\tau^2}$$

Subsection: Estimate of the Mean

Estimate of the Mean

Theorem

Let $X_1, \ldots, X_n$ be a sample from the exponential distribution $\mathrm{Ex}(\tau)$ and $\overline{X}$ be the sample mean. Then holds:

- $\overline{X}$ is gamma distributed according to $\mathrm{Ga}(n, \tau/n)$.
- $\overline{X}$ is an unbiased and consistent estimator of τ.
- $\overline{X}$ is efficient.
- $\overline{X}$ is asymptotically normally distributed with mean $\mu = \tau$ and variance $\sigma^2 = \tau^2/n$.

Section 14: Poisson Distributed Data

Subsection: The Poisson Process

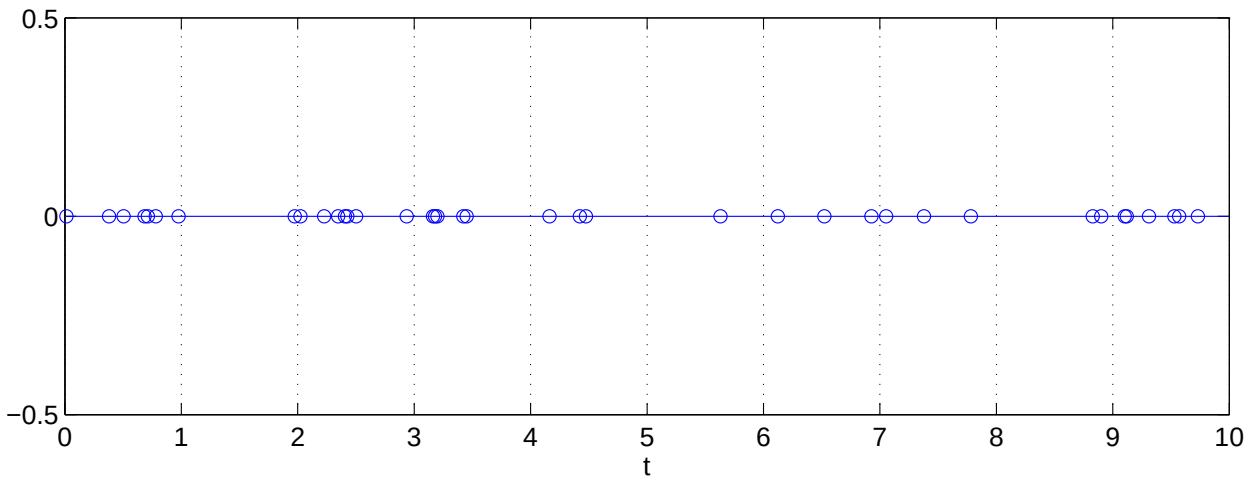
The Poisson Process

- We observe a process in which certain events occur at random times.
- If the waiting times between events are independent and exponentially distributed with mean τ, we speak of an **Poisson process** with intensity $\lambda = 1/\tau$.
- The number X of events per time unit is then independent and **Poisson-distributed** according to $\mathrm{Po}(\lambda)$.
- Conversely, if the number X of events per unit time is independent and Poisson-distributed according to $\mathrm{Po}(\lambda)$, then the waiting times are exponentially distributed with mean $\tau = 1/\lambda$.
- In a Poisson process of intensity λ, the number of events per time interval of length T is again Poisson distributed according to $\mathrm{Po}(\lambda T)$.
- The sum of two Poisson processes is again a Poisson process. Its intensity is the sum of the intensities of the summands.

The Poisson Process

Examples of Poisson processes

- The number of decays per unit time in a (large) radioactive source.
- The number of particles per unit time in cosmic radiation.
- The number of pixel errors on a TFT display.
- The number of rare events per time unit (insurance claims, suicides, accidents).



The Poisson Process

The Poisson distribution $\mathrm{Po}(\lambda)$

- The density of the Poisson distribution is:

$$f_{\mathrm{Po}}(k;\lambda) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad k \in \mathbb{N}_0$$

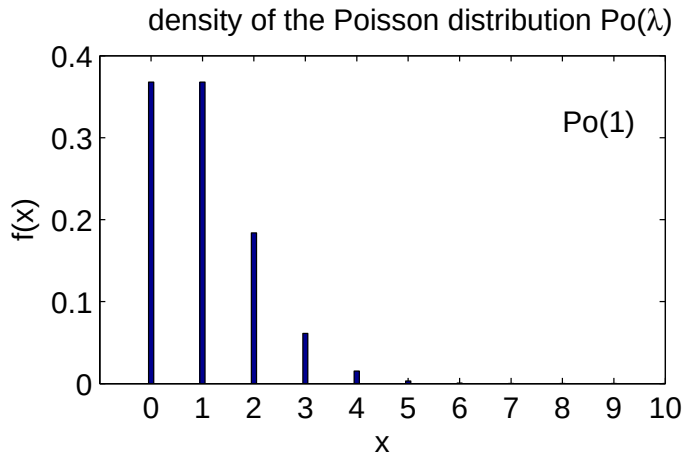
- The mode M is equal to

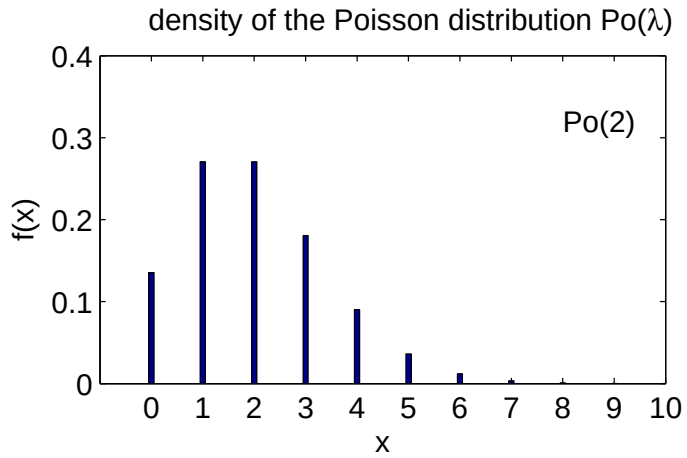
$$M = \begin{cases} \lfloor \lambda \rfloor, & \text{if } \lambda \notin \mathbb{N} \\ \lambda \text{ and } \lambda - 1, & \text{if } \lambda \in \mathbb{N} \\ 0, & \text{if } \lambda = 0 \end{cases}$$

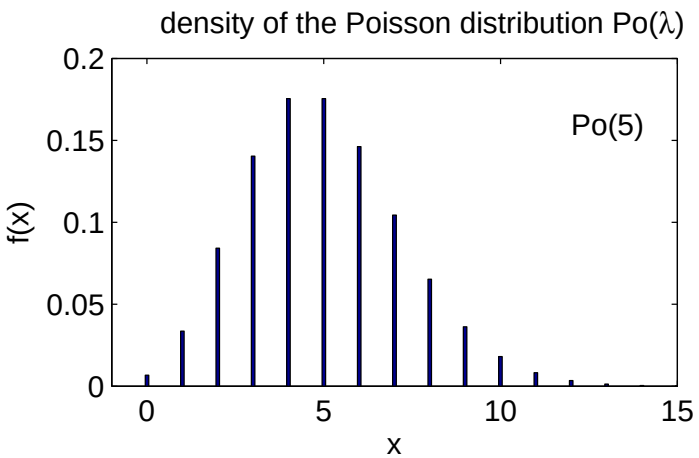
- The distribution function can be expressed by the distribution function of the gamma distribution $\mathrm{Ga}(k+1, 1)$:

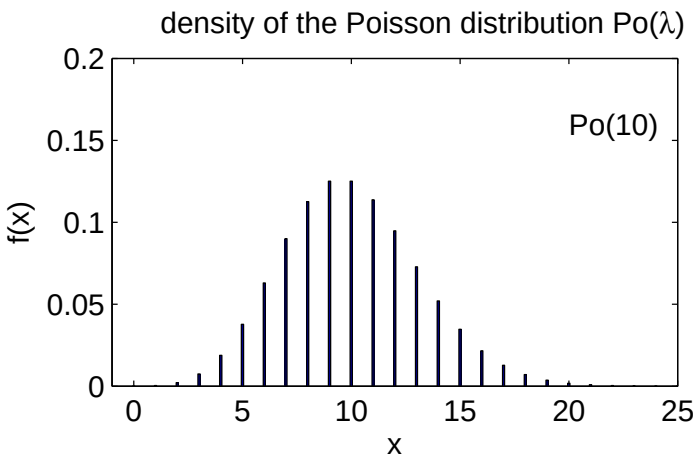
$$F_{\mathrm{Po}}(k;\lambda) = 1 - F_{\mathrm{Ga}}(\lambda; k+1, 1)$$

The Poisson Process









The Poisson Process

The moments and the CF of the Poisson distribution

let $X \sim \mathrm{Po}(\lambda)$. Then holds:

- $\mathsf{E}[X] = \lambda$
- $\mathsf{var}[X] = \lambda$
- $\gamma[X] = 1/\sqrt{\lambda}$
- $\kappa[X] = 3 + 1/\lambda$
- $\varphi_X(t) = \exp\left(\lambda\left(e^{\mathrm{i}t} - 1\right)\right)$

Fisher information of a Poisson distributed sample

Let $X_1, \ldots, X_n$ be a sample from the Poisson distribution $\mathrm{Po}(\lambda)$. Then holds:

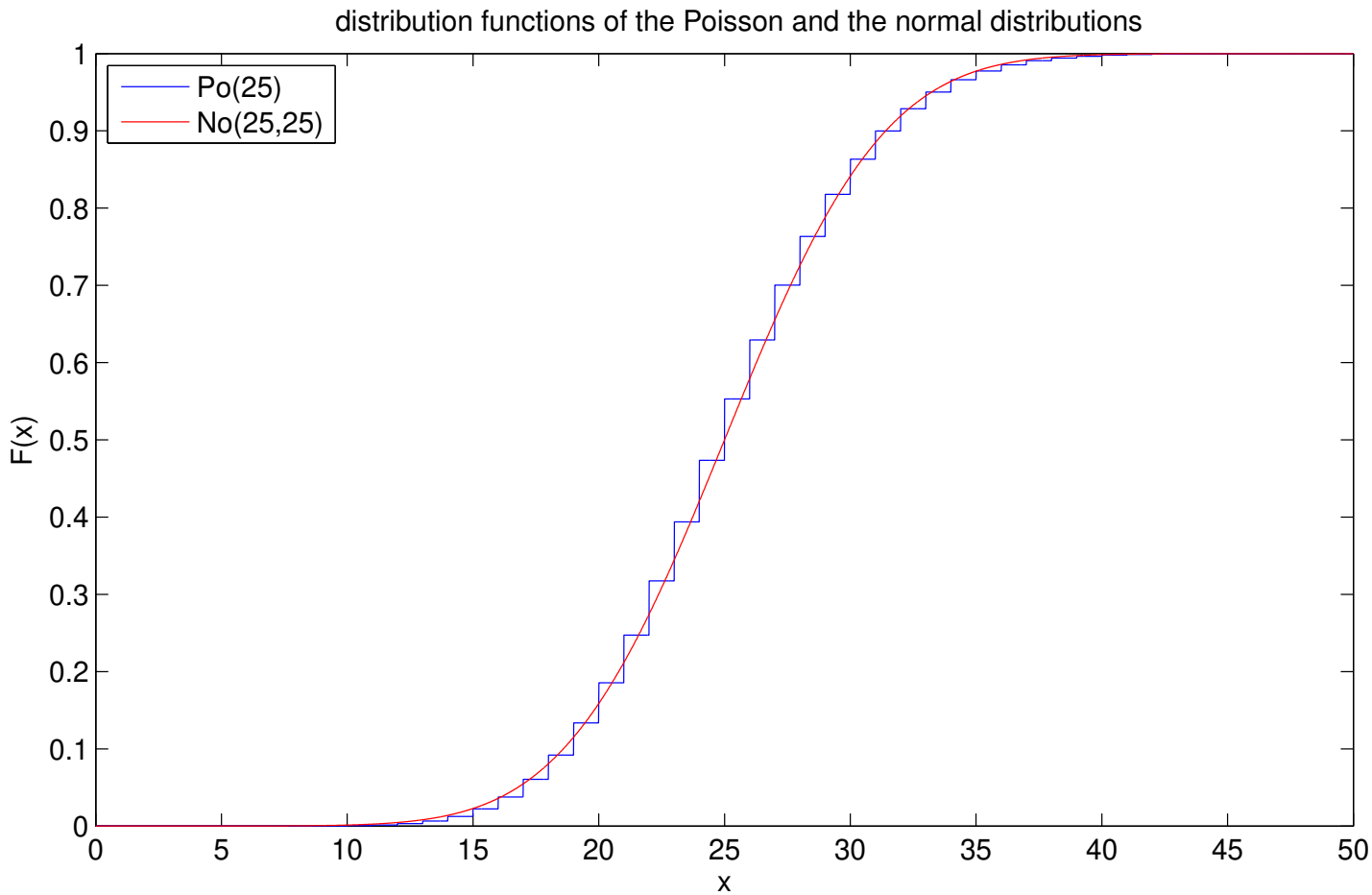
$$I_\lambda = \frac{n}{\lambda}$$

The Poisson Process

The Poisson distribution for large λ

- A random variable distributed according to $\mathrm{Po}(\lambda)$ can be represented as a sum of λ $P(1)$-distributed random variables.
- Therefore, according to the central limit theorem, the Poisson distribution for $\lambda \to \infty$ must tend toward a normal distribution.
- The following figure shows the distribution function of the Poisson distribution $\mathrm{Po}(\lambda)$ with $\lambda = 25$, and the distribution function of the normal distribution $\mathrm{No}(\mu, \sigma^2)$ with $\mu = \lambda = 25$ and $\sigma^2 = \lambda = 25$.

The Poisson Process



Subsection: estimation of the mean

estimation of the mean

Theorem

Let $X_1, \ldots, X_n$ be a sample from the Poisson distribution $\mathrm{Po}(\lambda)$ and $\overline{X}$ be the sample mean. Then holds:

- $\overline{X}$ is an unbiased and consistent estimator of λ.
- $\overline{X}$ is efficient.
- $\overline{X}$ is asymptotically normally distributed with mean $\mu = \lambda$ and variance $\sigma^2 = \lambda/n$.

Section 15: Data from Bernoulli and Drawing Experiments

Subsection: Repeating Bernoulli Experiments

Repeating Bernoulli Experiments

The alternative or Bernoulli distribution $\mathrm{Al}(p)$

- A Bernoulli experiment or alternative experiment has two possible outcomes, "success" and "failure" respectively.
- The random variable X assigns the value 1 to success and 0 to failure.
- If p is the probability of success, the density $f_{\mathrm{Al}}(x;p)$ is as follows:

$$f_{\mathrm{Al}}(0;p) = 1-p, \;\; f_{\mathrm{Al}}(1;p) = p$$

or

$$f_{\mathrm{Al}}(x;p) = p^x(1-p)^{1-x}, \quad x \in \{0,1\}$$

Repeating Bernoulli Experiments

- If the alternative test is performed n times independently, there are 2^n possible outcomes, namely all sequences of the form $\boldsymbol{e} = (i_1, \ldots, i_n), \quad i_j \in \{0, 1\}$.
- The discrete random variable Y maps the sequence $\boldsymbol{e}$ to the frequency of 1:

$$Y(\boldsymbol{e}) = \sum_{j=1}^{n} i_j$$

- The range of values of Y is the set $\{0, 1, \ldots, n\}$. Mapped to the number k $(0 \leq k \leq n)$ are all sequences where 1 occurs exactly k times. There are C_k^n such sequences, and each has probability $p^k(1-p)^{n-k}$.

Repeating Bernoulli Experiments

The binomial distribution $\mathrm{Bi}(n, p)$

- Therefore, the density f of Y is:

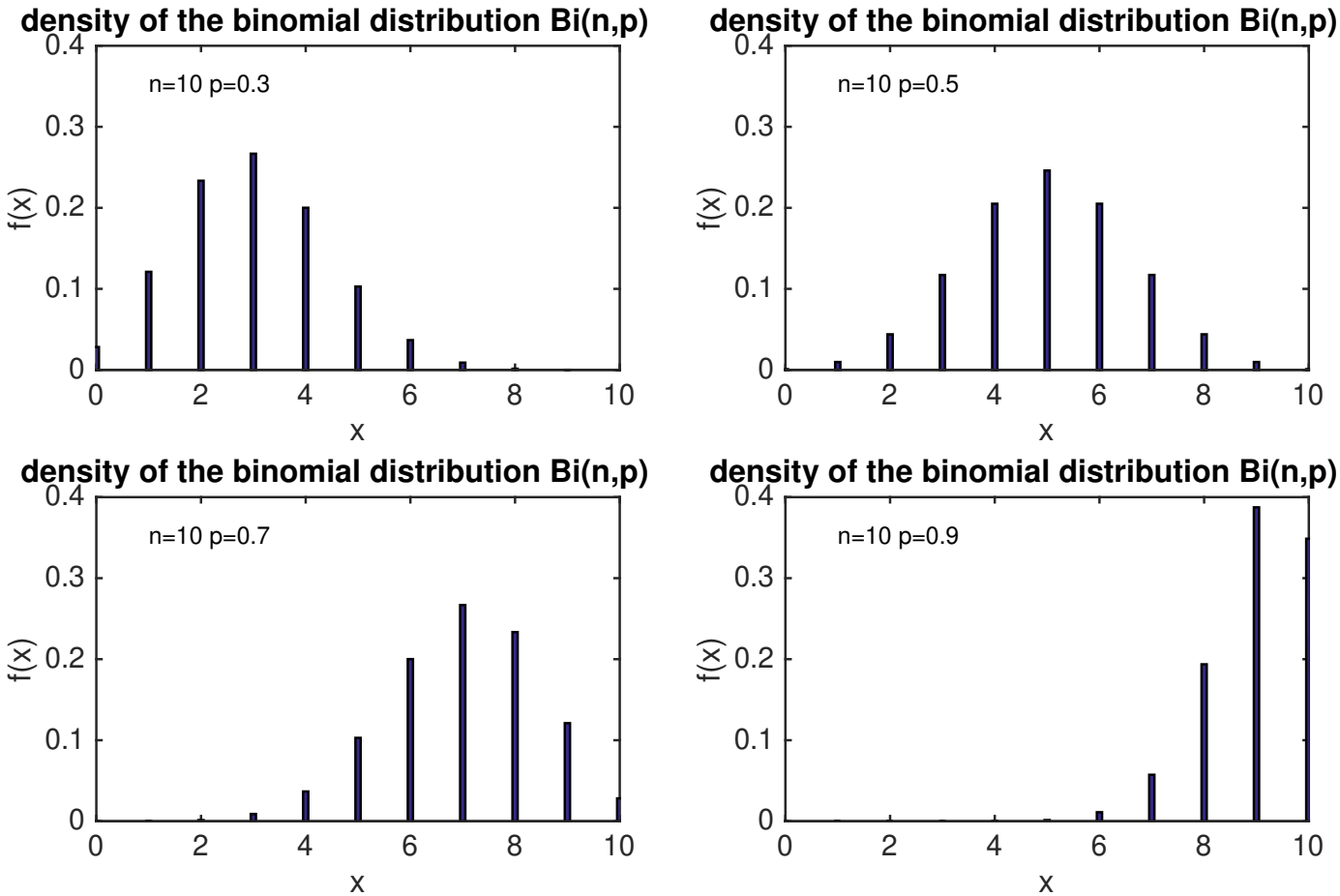
$$f_{\mathrm{Bi}}(k; p) = \binom{n}{k} p^k (1-p)^{n-k}, \ 0 \leq k \leq n$$

- The distribution of Y is called **binomial distribution** $\mathrm{Bi}(n, p)$ with parameters n and p.
- It holds:

$$\sum_{k=0}^{n} f_{\mathrm{Bi}}(k; p) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$$

This is just the binomial theorem.

Repeating Bernoulli Experiments



Repeating Bernoulli Experiments

- The mode M is equal to

$$M = \begin{cases} \lfloor (n+1)p \rfloor, & \text{if } p = 0 \text{ or } (n+1)p \notin \mathbb{N} \\ (n+1)p \text{ and} & \\ (n+1)p - 1, & \text{wenn } (n+1)p \in \{1, \ldots, n\} \\ 1, & \text{if } p = 1 \end{cases}$$

- The distribution function can be expressed by the distribution function of the beta distribution $\mathrm{Be}(x; n-k, k+1)$:

$$F_{\mathrm{Bi}}(k; n, p) = F_{\mathrm{Be}}(1-p; n-k, k+1)$$

Repeating Bernoulli Experiments

Exkurs: The beta distribution $\mathrm{Be}(a, b)$

- The density of the beta distribution is:

$$f_{\mathrm{Be}}(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \cdot I_{[0,1]}(x)$$

- Its distribution function is the regularized incomplete beta function:

$$F_{\mathrm{Be}}(x; a, b) = \int_0^x \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)} \, \mathrm{d}t$$

- The mode M is at $x = (a-1)/(a+b-2)$ when $a, b > 1$.

Repeating Bernoulli Experiments

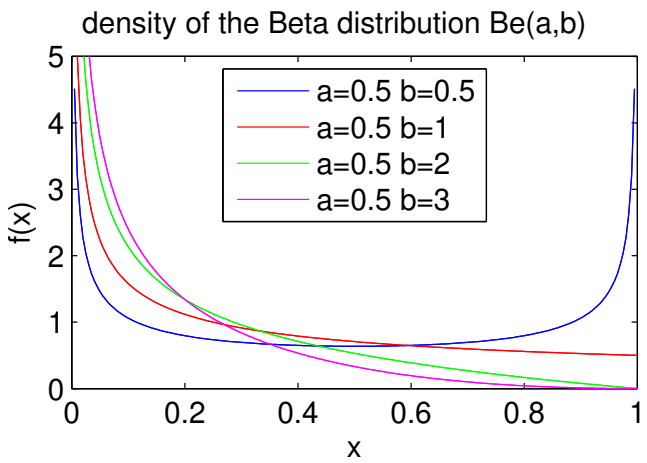
The moments of the $\mathrm{Be}(a, b)$ distribution

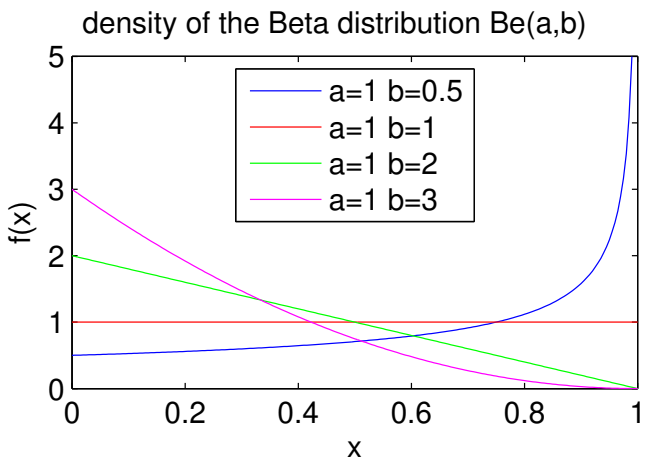
let $X \sim \mathrm{Be}(a, b)$. Then holds:

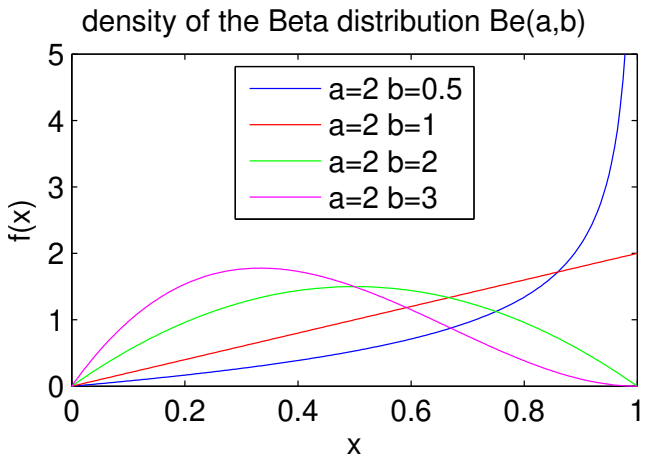
- $\mathsf{E}[X] = \dfrac{a}{a+b}$
- $\mathsf{var}[X] = \dfrac{ab}{(a+b)^2(a+b+1)}$
- $\gamma[X] = \dfrac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$
- $\kappa[X] = \dfrac{6(a-b)^2 + 3ab(a+b+2)}{ab(a+b+2)(a+b+3)}$

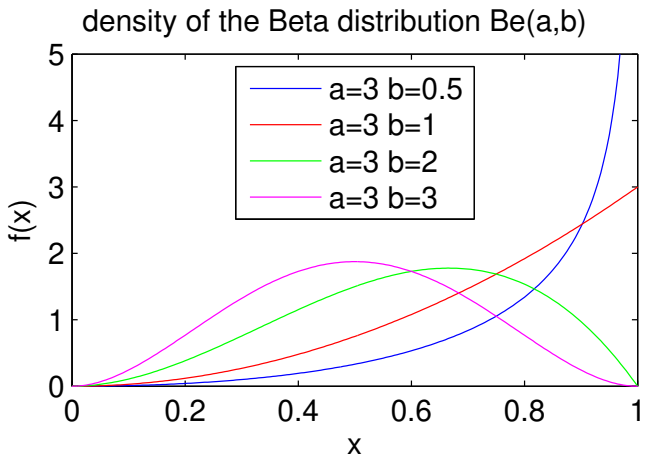
- The α quantile of the $\mathrm{Be}(a, b)$ distribution is denoted by $\beta_{\alpha;a,b}$.

Repeating Bernoulli Experiments









Repeating Bernoulli Experiments

The moments and the CF of the binomial distribution

Let $X \sim \mathrm{Bi}(n,p)$. Then holds:

- $\mathsf{E}[X] = np$
- $\mathsf{var}[X] = np(1-p)$
- $\gamma[X] = \dfrac{1-2p}{\sqrt{np(1-p)}}$
- $\kappa[X] = 3 - \dfrac{6}{n} + \dfrac{1}{np(1-p)}$
- $\varphi_X(t) = \left(1-p+pe^{\mathrm{i}t}\right)^n$

Fisher information of a binomial distributed observation

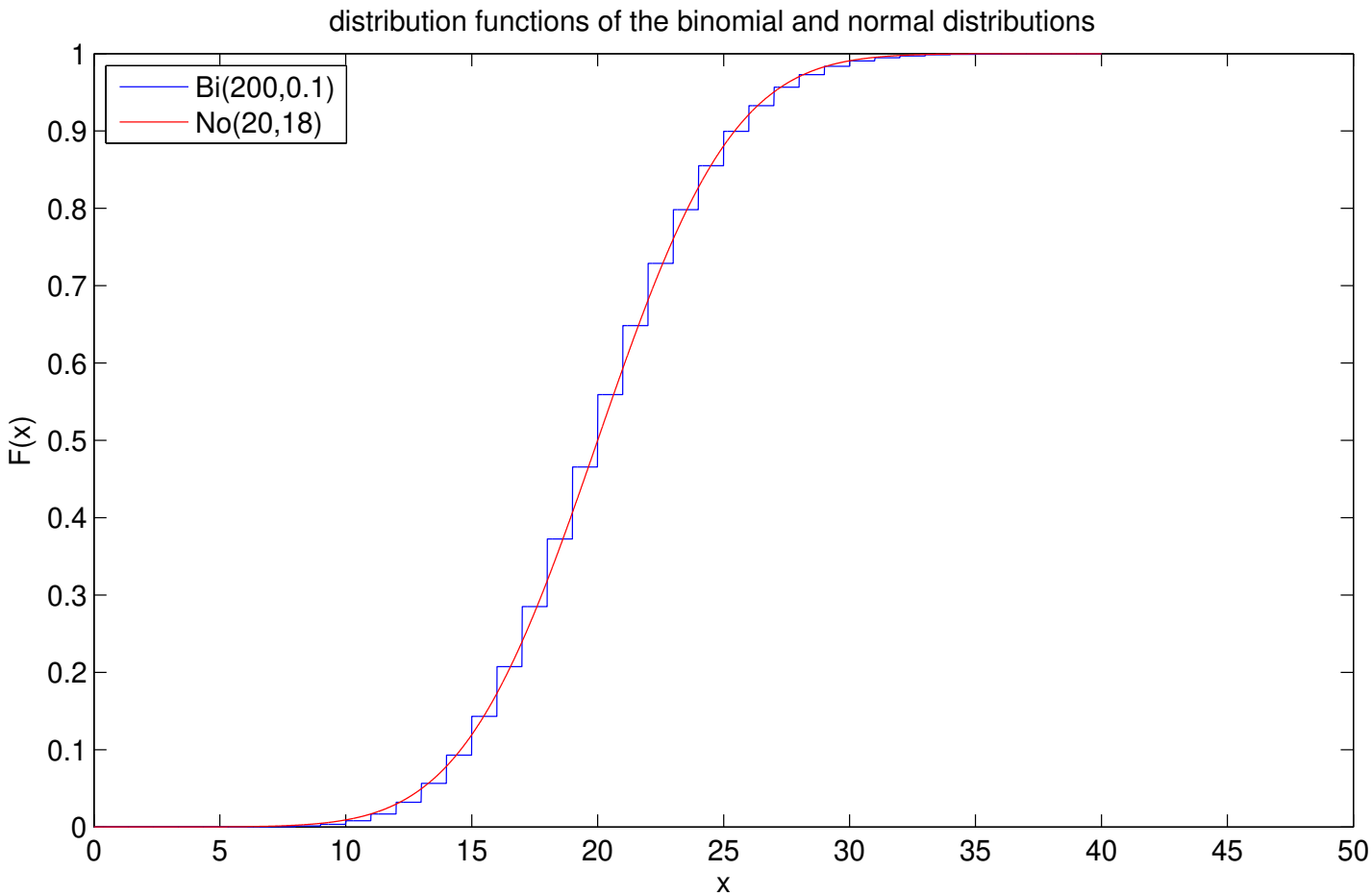
Let X be an observation from the binomial distribution $\mathrm{Bi}(n,p)$. Then holds:

$$I_p = \frac{n}{p(1-p)}$$

The Binomial Distribution for Large n

- A random variable distributed according to $\mathrm{Bi}(n, p)$ can be represented as a sum of n Bernoulli distributed random variables.
- Therefore, according to the central limit theorem, the binomial distribution for $n \to \infty$ and fixed p must tend toward a normal distribution.
- The following figure shows the distribution function of the binomial distribution $\mathrm{Bi}(n, p)$ with $n = 200$ and $p = 0.1$, and the distribution function of the normal distribution $\mathrm{No}(\mu, \sigma^2)$ with $\mu = np = 20$ and $\sigma^2 = np(1-p) = 18$.

Repeating Bernoulli Experiments



Repeating Bernoulli Experiments

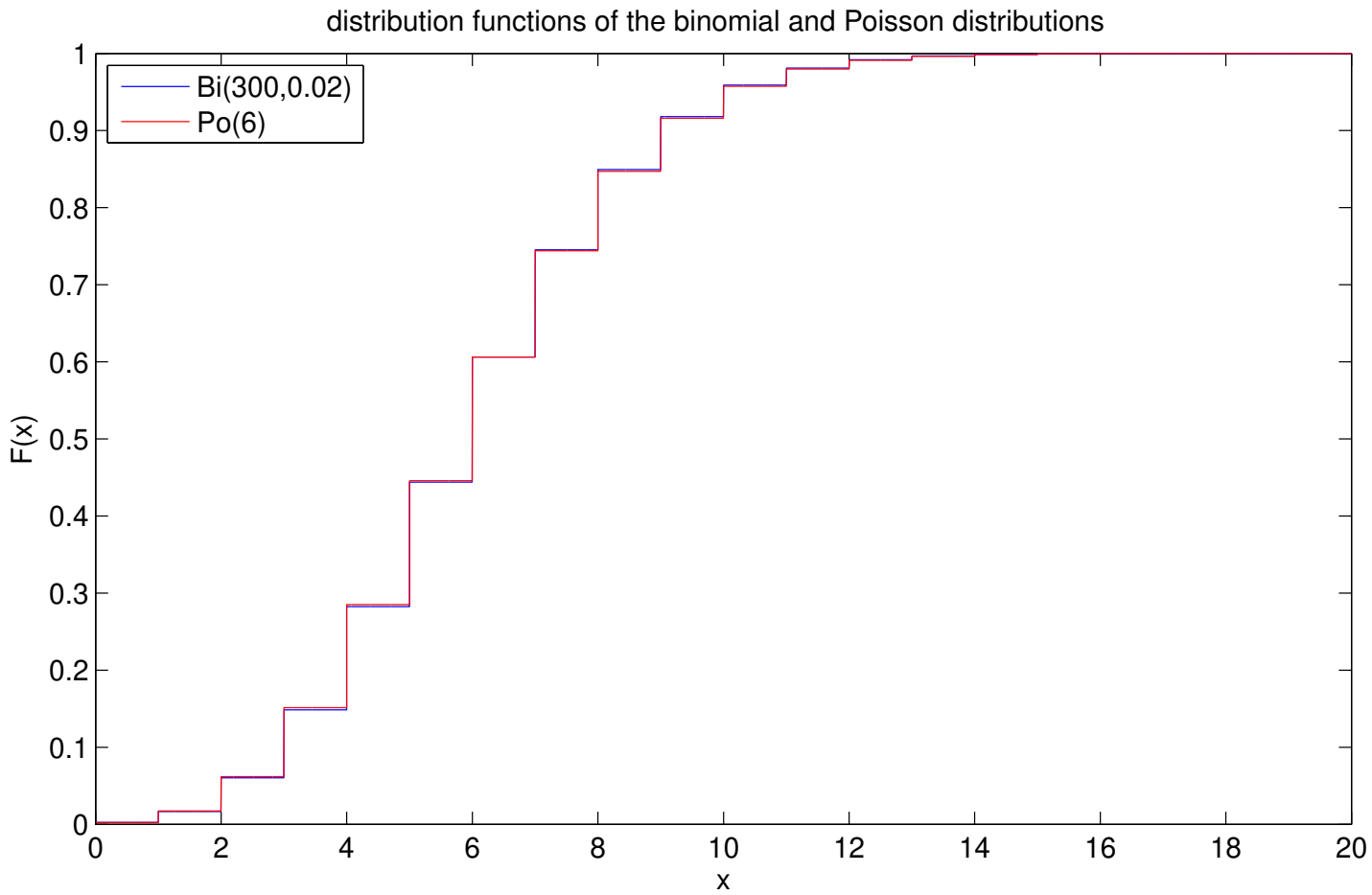
The binomial distribution for large n and $p = \lambda/n$

- In the limit $n \to \infty$ and $p = \lambda/n \to 0$ the binomial distribution $\mathrm{Bi}(n, p)$ tends to the Poisson distribution $\mathrm{Po}(\lambda)$:

$$\lim_{n\to\infty} \binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{n-k} = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

- Thus, if the number of trials n is increased and the probability of success is decreased in inverse proportion, the number of successes is asymptotically Poisson distributed.

Repeating Bernoulli Experiments



Subsection: estimation of probability of success

estimation of probability of success

Theorem

Let k be an observation from the binomial distribution $\mathrm{Bi}(n, p)$. Then holds:

- $\hat{p} = k/n$ is an unbiased estimator of p.
- $\hat{p}$ is efficient.
- $\hat{p}$ is asymptotically normally distributed with mean $\mu = p$ and variance $\sigma^2 = p(1-p)/n$.

Subsection: Drawing Experiments

Drawing Experiments

Drawing with Placing Back

- Population of N objects, of which M objects are **feature bearers**, thus have a certain property E.
- n objects are drawn, each object having the same probability of being drawn.
- Each object drawn is immediately put back.
- The number of feature bearers drawn is a random variable X.
- X is then **binomially distributed** according to $\mathrm{Bi}(n, M/N)$.
- If N and M are much larger than n, X is binomially distributed to a good approximation even if drawn objects are **not** put back.
- An example of this is a survey in which the sample size n is much smaller than N and M.

Drawing Experiments

Pulling without backspacing

- Population of N objects, of which M objects are **feature bearer**, thus have a certain property E.
- n objects are drawn, each object having the same probability of being drawn.
- Once drawn, objects are **not** put back.
- The number of feature bearers drawn is a random variable X.
- The distribution of X is called a **hypergeometric distribution** $\mathrm{Hy}(N, M, n)$.

Drawing Experiments

The hypergeometric distribution $\mathrm{Hy}(N, M, n)$

- Its density is:

$$f_{\mathrm{Hy}}(m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}, \quad 0 \leq m \leq \min(n, M)$$

- The mode D is equal to

$$D = \left\lfloor \frac{(n+1)(M+1)}{N+2} \right\rfloor$$

Drawing Experiments

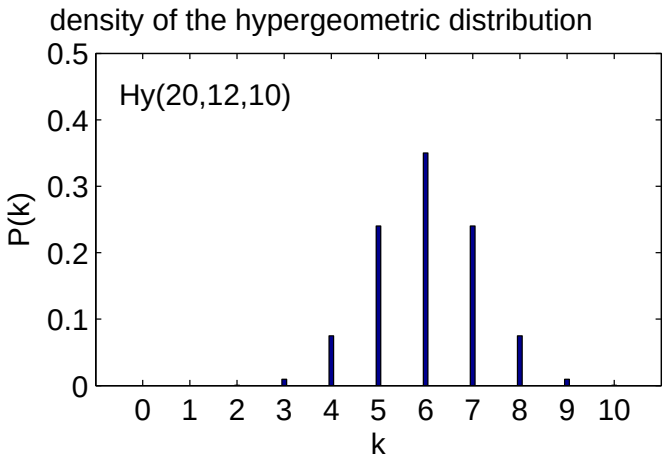
The moments of the hypergeometric distribution

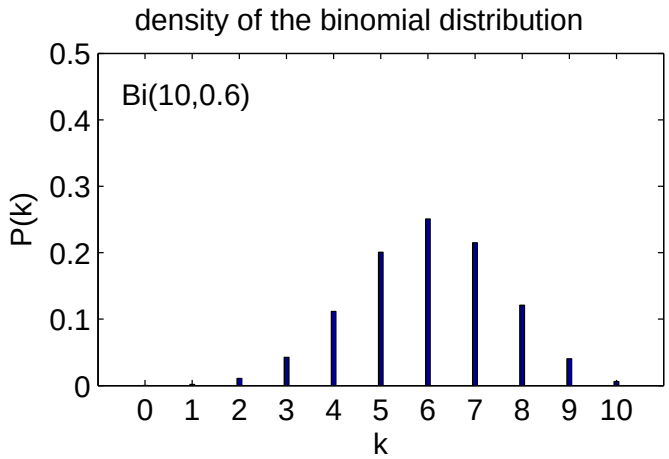
Let $X \sim \mathrm{Hy}(N, M, n)$ and $p = M/N$. Then holds:

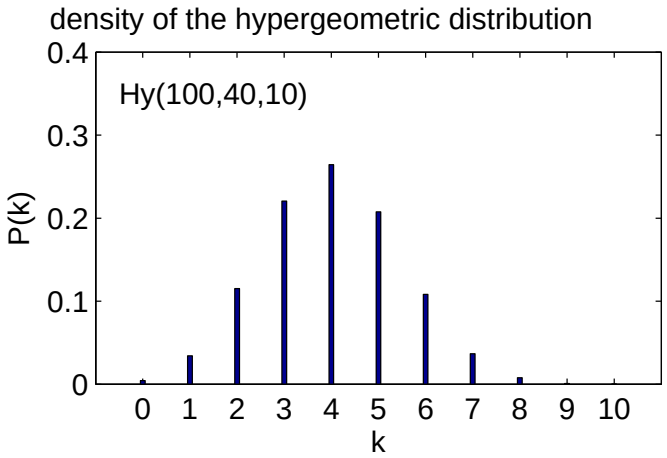
- $\mathrm{E}[X] = \frac{nM}{N} = np$
- $\mathrm{var}[X] = \frac{nM}{N}\frac{N-M}{N}\frac{N-n}{N-1} = np(1-p)\frac{N-n}{N-1}$

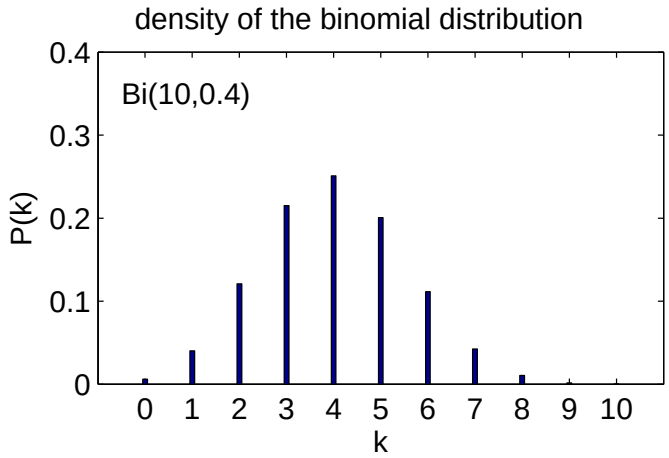
- The term $(N-n)(N-1)$ in the variance is called **finity correction**. It is close to 1 if $N \gg n$.
- If $N \gg n$ and $M \gg n$, the hypergeometric distribution can be approximated by the binomial distribution because then the drawing of the sample changes the composition of the population only insignificantly.

Drawing Experiments









Drawing Experiments

Estimating the number of feature bearers

Theorem

Let k be an observation from the hypergeometric distribution $\mathrm{Hy}(N, M, n)$ with known N but unknown M. Then holds:

- $\hat{M} = \frac{kN}{n}$ is an unbiased estimator of M.
- $\mathrm{var}[\hat{M}] = M \frac{N-M}{n} \frac{N-n}{N-1}$

Drawing Experiments

Estimating the size of the population

Theorem

Let k be an observation from the hypergeometric distribution $\mathrm{Hy}(N, M, n)$ with known M but unknown N. Then holds:

- $\hat{N} = \frac{nM}{k}$ is a biased estimator of N.
- $1/\hat{N} = \frac{k}{nM}$ is an unbiased estimator of $1/N$.
- With error propagation we get:

$$\mathrm{E}[\hat{N}] \approx N + \frac{N-M}{M}\frac{N-n}{n}$$

$$\mathrm{var}[\hat{N}] \approx N\,\frac{N-M}{M}\frac{N-n}{n}$$

Section 16: Maximum Likelihood Estimator

Maximum Likelihood Estimator

Definition (ML Estimator)

- Let $X_1, \ldots, X_n$ be a sample with joint density $g(x_1, \ldots, x_n|\vartheta)$. The function

$$L(\vartheta|X_1, \ldots, X_n) = g(X_1, \ldots, X_n|\vartheta)$$

  is called the **likelihood function** of the sample.
- The **plausible** or **maximum likelihoodestimator** $\hat{\vartheta}$ is that value of ϑ that maximizes the likelihood function of the sample.

- Often, instead of the likelihood function, its logarithm, the log-likelihood function $\ell(\vartheta) = \ln L(\vartheta)$ is maximized.
- In simple cases the maximization succeeds analytically, otherwise it must be maximized numerically.

Maximum Likelihood Estimator

- The ML estimator is invariant under (differentiable) transformations of the parameter: if $\hat{\vartheta}$ is the ML estimator of ϑ, then $h(\hat{\vartheta})$ is the ML estimator of $h(\vartheta)$.

Example (ML estimator of a Bernoulli parameter)

Let $X_1, \ldots, X_n$ be a sample from the Bernoulli distribution $\mathrm{Al}(p)$. The joint density is:

$$g(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Therefore, the log-likelihood function is:

$$\ell(p) = \sum_{i=1}^{n} X_i \ln p + \left(n - \sum_{i=1}^{n} X_i\right) \ln(1-p)$$

Maximum Likelihood Estimator

Example (Continuation)

Deriving to p results in:

$$\frac{\partial \ell(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right)$$

Zeroing the derivative and solving for p yields:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

The ML estimator is unbiased and efficient.

Maximum Likelihood Estimator

Example (ML estimator of a Poisson parameter)

Let $X_1, \ldots, X_n$ be a sample from the Poisson distribution $\mathrm{Po}(\lambda)$. The joint density is:

$$g(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Therefore, the log-likelihood function is:

$$\ell(\lambda) = \sum_{i=1}^{n} [X_i \ln \lambda - \lambda - \ln(x_i!)]$$

Deriving to λ gives:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n$$

Maximum Likelihood Estimator

Example (Continuation)

Zeroing the derivative and solving for λ gives:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

. The ML estimator is unbiased and efficient.

Maximum Likelihood Estimator

Example (ML estimator of a mean lifetime)

Let $X_1, \ldots, X_n$ be a sample from the exponential distribution $\mathrm{Ex}(\tau)$. The joint density is:

$$g(x_1, \ldots, x_n|\tau) = \prod_{i=1}^{n} \frac{e^{-x_i/\tau}}{\tau}$$

Therefore, the log-likelihood function is:

$$\ell(\tau) = \sum_{i=1}^{n} [-\ln \tau - \frac{1}{\tau} X_i]$$

Deriving to τ gives:

$$\frac{\partial \ell(\tau)}{\partial \tau} = -\frac{n}{\tau} + \frac{1}{\tau^2} \sum_{i=1}^{n} X_i$$

Maximum Likelihood Estimator

Example (Continuation)

Zeroing the derivative and solving for τ yields:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

. The ML estimator is unbiased and efficient.

Maximum Likelihood Estimator

Example (ML estimator of the parameters of a normal distribution)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$. The joint density is:

$$g(x_1, \ldots, x_n | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\,\sigma^2}\right]$$

Therefore, the log-likelihood function is:

$$\ell(\mu, \sigma^2 | \boldsymbol{X}) = \sum_{i=1}^{n} \left[-\ln\sqrt{2\pi} - \frac{1}{2}\ln\sigma^2 - \frac{(X_i - \mu)^2}{2\,\sigma^2}\right]$$

Deriving by μ and σ^2 yields:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma^2}, \quad \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} \left[-\frac{1}{2\,\sigma^2} + \frac{(X_i - \mu)^2}{2\,\sigma^4}\right]$$

Maximum Likelihood Estimator

Example (Continuation)

Zeroing the derivatives and solving for μ and σ^2 yields:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

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

The ML estimator of μ is unbiased and efficient. The ML estimator of σ^2 is asymptotically unbiased and asymptotically efficient.
The ML estimator of σ is equal to.

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}}S$$

It is also asymptotically unbiased and asymptotically efficient.

Maximum Likelihood Estimator

- The normalized likelihood function can be interpreted as an a posteriori distribution of the estimated parameter.
- For large n, the variance of the likelihood estimate $\hat{\vartheta}$ can therefore be read from the second central moment of the normalized likelihood function.
- If the estimated parameter ϑ is the mean of a normal distribution, this procedure is exact for any n:

$$L(\vartheta) = \frac{1}{\sigma^n \sqrt{2\,\pi}^n} \exp\left[-\frac{n}{2\,\sigma^2}\left((\hat{\vartheta}-\vartheta)^2 + \frac{1}{n}\sum(x_i-\hat{\vartheta})^2\right)\right]$$

- If $L(\vartheta)$ is normalized, the "'density"' of a normal distribution with mean $\hat{\vartheta}$ and variance $\frac{\sigma^2}{n}$, i.e., just the variance of the estimate $\hat{\vartheta} = \frac{1}{n}\sum x_i$.

Maximum Likelihood Estimator

Example (Estimation of the parameter a of a gamma distribution)

The sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ consists of $n = 200$ values drawn independently from a $\mathrm{Ga}(a, 1)$ distribution:

$$f(x_i|a) = \frac{x_i^{a-1} e^{-x_i}}{\Gamma(a)}, \quad i = 1, \ldots, n$$

The (unknown) true value of a is $a_\mathrm{w} = 2$. The log-likelihood function is

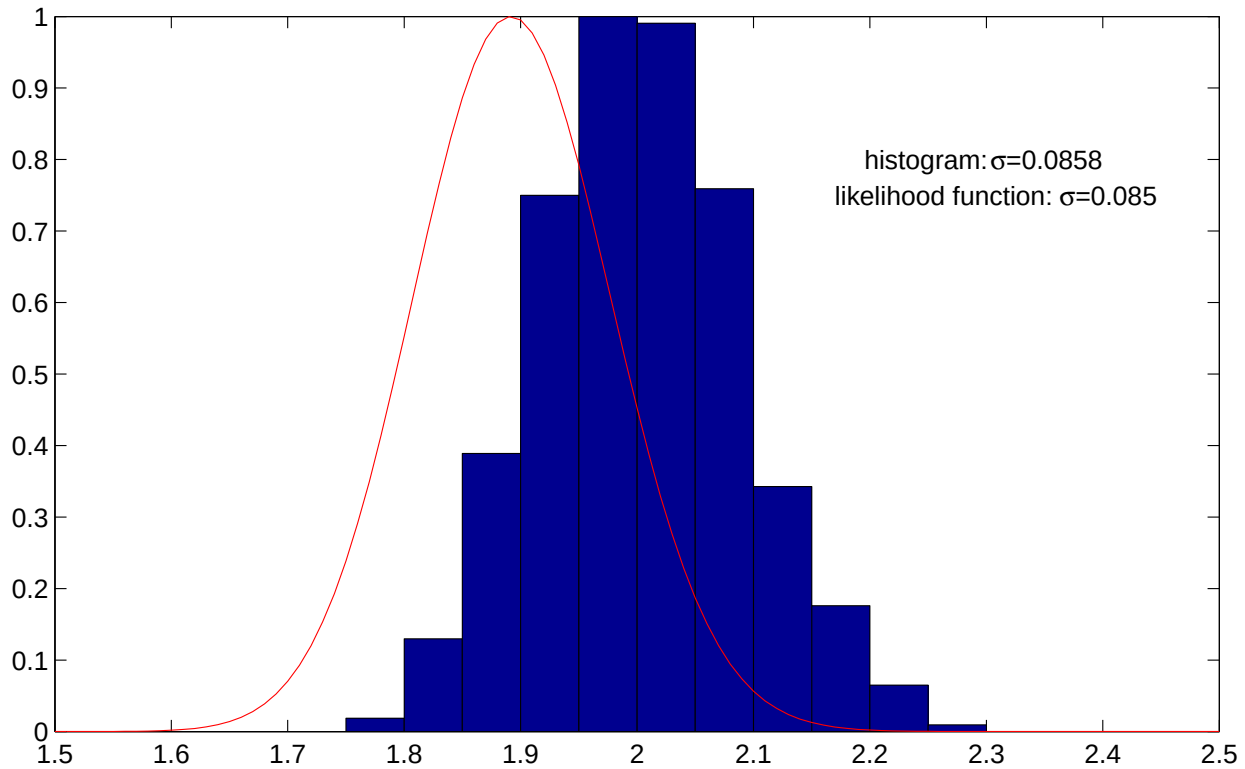
$$\ln L(a|\boldsymbol{X}) = \sum_{i=1}^{n} \ln f(X_i|a) = (a-1) \sum_{i=1}^{n} \ln X_i - \sum_{i=1}^{n} X_i - n \ln \Gamma(a)$$

Maximum Likelihood Estimator

Example (Continuation)

Numerical maximization of $\ln L(a)$ gives the maximum likelihood estimate $\hat{a}$. The experiment is repeated N times and the estimates of each experiment $(\hat{a}^{(k)}, k = 1, \ldots, N)$ are histogrammed. Comparison of the individual (normalized) likelihood function with the histogram $(N = 500)$ shows good agreement between the standard deviations.

Example (Continuation)



Maximum Likelihood Estimator

Example (Continuation)

With known b, the Fisher information of the sample with respect to a is equal to

$$I_a = n \frac{\mathrm{d}^2 \ln \Gamma(a)}{\mathrm{d}a^2} \approx 128.99$$

This corresponds to a standard deviation $\sigma = 0.088$. Thus, the ML estimator is practically efficient.

Because of $\mathrm{E}[\overline{X}] = a$, $\overline{X}$ is a biased estimator of a for all n. However, it has a larger standard deviation than the ML estimator:

$$\sigma[\overline{X}] = \sqrt{\frac{a}{n}} = 0.1$$

On the other hand, it is much easier to calculate.

Maximum Likelihood Estimator

- The ML estimator has the following important property:

Theorem

If the first two derivatives of $L(\vartheta)$ exist, the information $I_g(\vartheta)$ exists for all ϑ, and if $E\left[(\ln L)'\right] = 0$, then the likelihood estimate $\hat{\vartheta}$ is asymptotically normally distributed with mean ϑ and variance $1/I_g(\vartheta)$. $\hat{\vartheta}$ is therefore asymptotically unbiased and asymptotically efficient.

- From this follows immediately the next property:

Theorem

The likelihood estimator $\hat{\vartheta}$ is consistent (under the same assumptions).

Maximum Likelihood Estimator

Example (ML estimation of the location parameter of a Cauchy distribution)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from the Cauchy distribution $\mathrm{t}(1)$ with location parameter μ. The joint density is:

$$g(x_1, \ldots, x_n|\mu) = \prod_{i=1}^{n} \frac{1}{\pi[1 + (x_i - \mu)^2]}$$

Therefore, the log-likelihood function is:

$$\ell(\mu|\boldsymbol{X}) = -n \ln \pi - \sum_{i=1}^{n} \ln[1 + (X_i - \mu)^2]$$

The maximum $\hat{\mu}$ of $\ell(\mu|\boldsymbol{X})$ must be found numerically.

Maximum Likelihood Estimator

Example (Continuation)

It can be shown that the Fisher information of the sample is equal to

$$I_\mu = \frac{n}{2}$$

. Therefore, for large samples, the variance of the ML estimator $\hat{\mu}$ must be approximately equal to $2/n$.

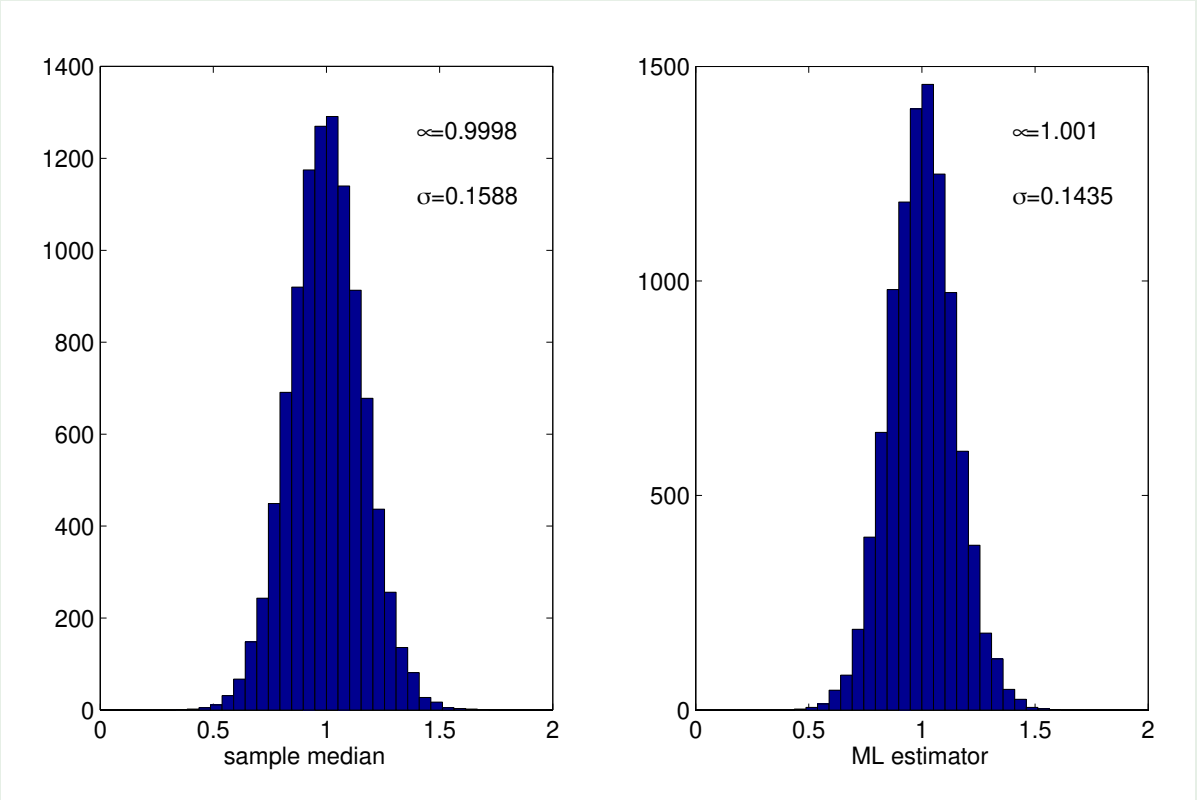
The sample median $\tilde{X}$ is also a consistent estimator for μ. Its variance is asymptotically equal to $\pi^2/(4n) \approx 2.47/n$. Thus, it is about 23 percent larger than the variance of the ML estimator.

The sample mean $\overline{X}$, on the other hand, is **no** consistent estimator for μ. Indeed, one can show that $\overline{X}$ has the same distribution as a single observation.

Maximum Likelihood Estimator

Example (Continuation)

Simulation of 10000 samples of size $n = 100$:

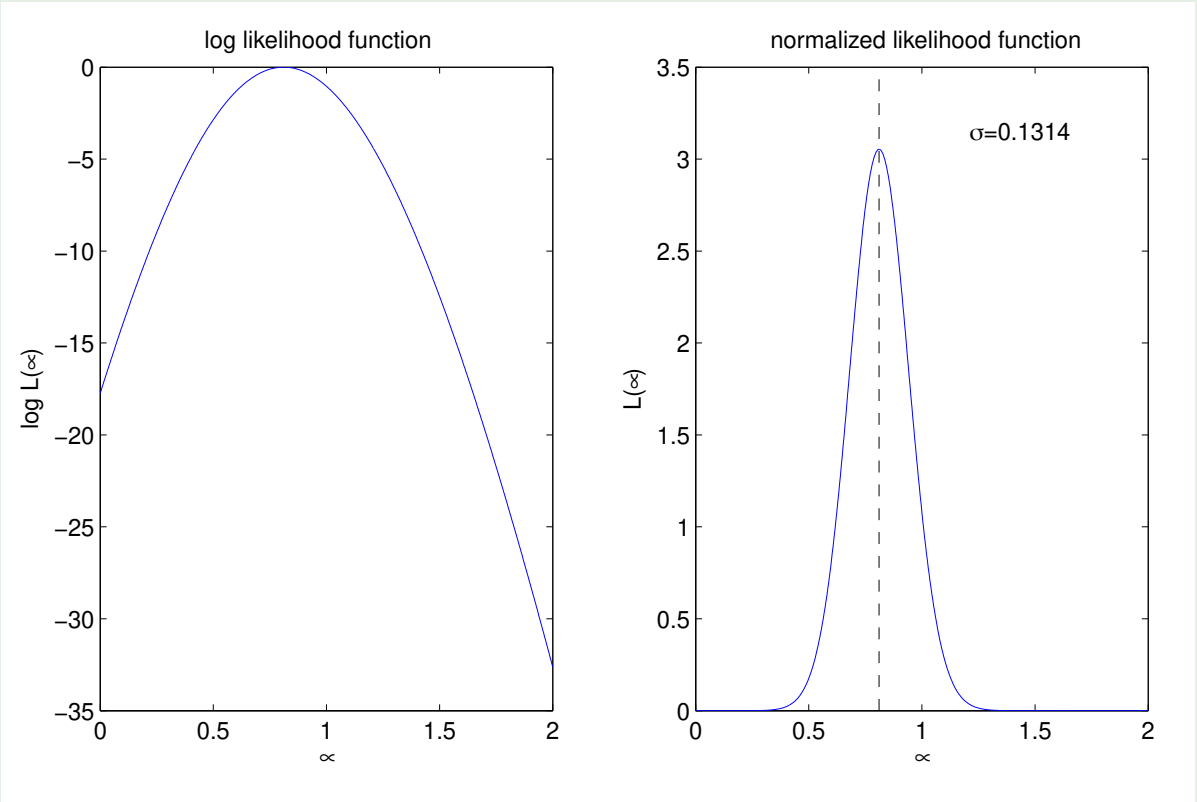


The correlation between $\tilde{x}$ and $\hat{\mu}$ is about 90%.

Maximum Likelihood Estimator

Example (Continuation)

The standard deviation of the ML estimator can again be approximated from the normalized likelihood function of a sample:



Maximum Likelihood Estimator

Example (ML estimation of the upper limit of a uniform distribution)

Let $X_1, \ldots, X_n$ be a sample from the uniform distribution $\mathrm{Un}(0, b)$ with upper limit b. The joint density is:

$$g(x_1, \ldots, x_n | b) = \frac{1}{b^n}, 0 \leq x_1, \ldots, x_n \leq b$$

Therefore, the largest value of likelihood function is at.

$$\hat{b} = \max_i X_i$$

Since there is a marginal maximum, the usual asymptotic properties do not apply.

Maximum Likelihood Estimator

Example (Continuation)

The density of $\hat{b} = \max_i X_i$ is:

$$f(x) = \frac{n x^{n-1}}{b^n}$$

From this, expectation and variance can be calculated:

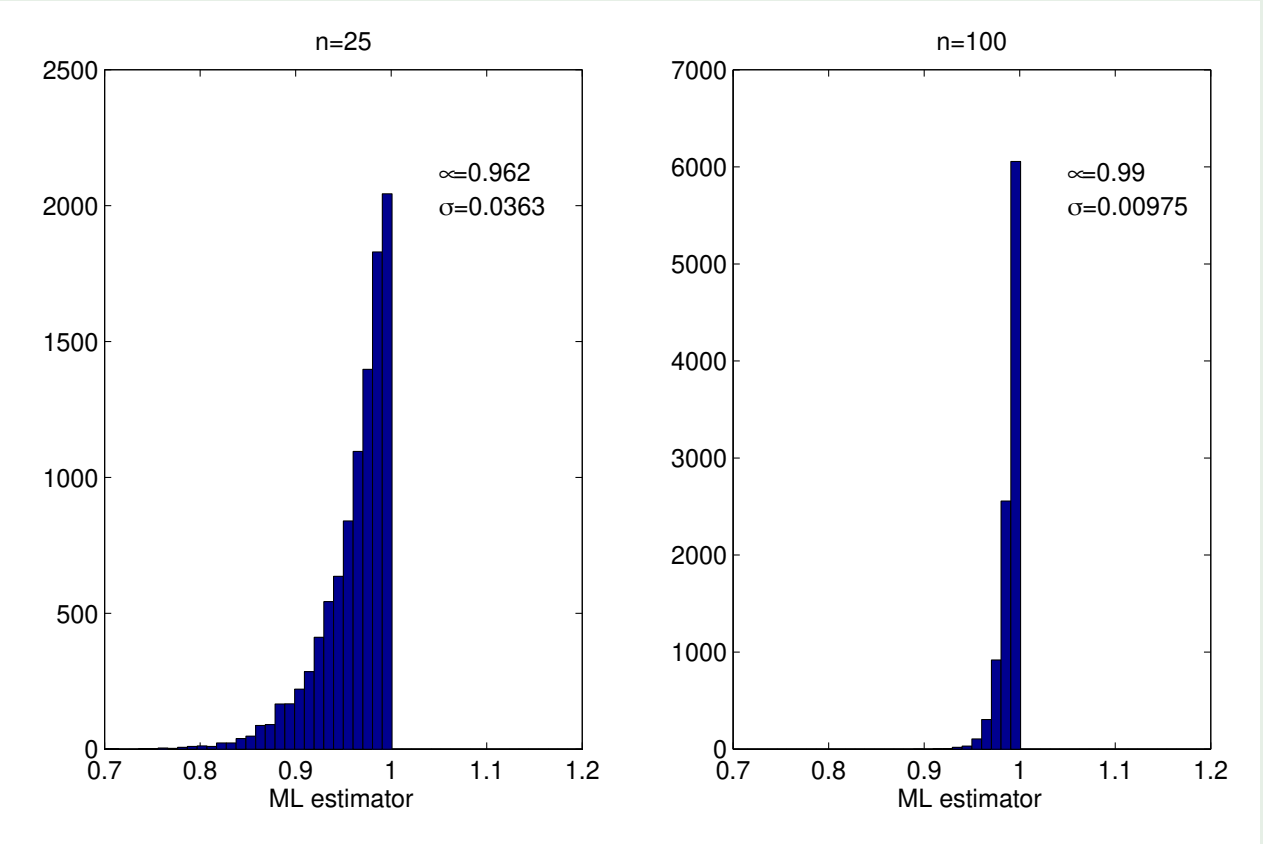
$$\mathrm{E}[\hat{b}] = \frac{bn}{n+1}, \quad \mathrm{var}[\hat{b}] = \frac{b^2 n}{(n+2)(n+1)^2}$$

The estimator is asymptotically unbiased, but the variance approaches zero, as does $1/n^2$! The estimator is also not asymptotically normally distributed.

Maximum Likelihood Estimator

Example (Continuation)

Simulation of 10000 samples ($b = 1$) of size $n = 25$ or $n = 100$:



Maximum Likelihood Estimator

- If the likelihood function is (approximately) normal, the log-likelihood function is (approximately) parabolic.
- In this case, error intervals of the ML estimator can be read from the log-likelihood function.

Width of the log-likelihood function

Let $\ell(\vartheta)$ be the log-likelihood function of the parameter ϑ, further $\hat{\vartheta}$ be the ML estimator and $\sigma[\hat{\vartheta}]$ its standard deviation. Then, approximately:

$$\ell(\hat{\vartheta} - k\sigma[\hat{\vartheta}]) = \ell(\hat{\vartheta} + k\sigma[\hat{\vartheta}]) = \ell(\hat{\vartheta}) - k^2/2$$

Maximum Likelihood Estimator

- If two parameters $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)$ are estimated simultaneously, the log-likelihood function $\ell(\boldsymbol{\vartheta})$ is asymptotically usually an elliptic paraboloid.
- The contour lines of $\ell(\boldsymbol{\vartheta})$ are then ellipses.
- If a contour line is to have the probability content $1 - \alpha$, then its height z is equal to

$$z = \ell(\hat{\boldsymbol{\vartheta}}) - \frac{1}{2}\chi^2_{1-\alpha;2} = \ell(\hat{\boldsymbol{\vartheta}}) + \ln(\alpha)$$

- For example, the ellipse with $1 - \alpha = 0.95$ is the contour line to the height $z = \ell(\hat{\boldsymbol{\vartheta}}) - 2.996$.
- The covariance matrix of the ML estimator $\hat{\boldsymbol{\vartheta}}$ can be approximated by the inverse negative Hessian matrix at the maximum.

Maximum Likelihood Estimator

Example

Mean μ and standard deviation σ of a normal distribution are estimated from a sample of size $n = 500$. The true values are $\mu = 5, \sigma = 1.5$, the estimated values are $\hat{\mu} = 4.980, \hat{\sigma} = 1.529$. The inverse negative Hessian matrix in the maximum of log-likelihood function is equal to.

$$\mathbf{V} = \begin{pmatrix} 0.0047 & 0 \\ 0 & 0.0023 \end{pmatrix}$$

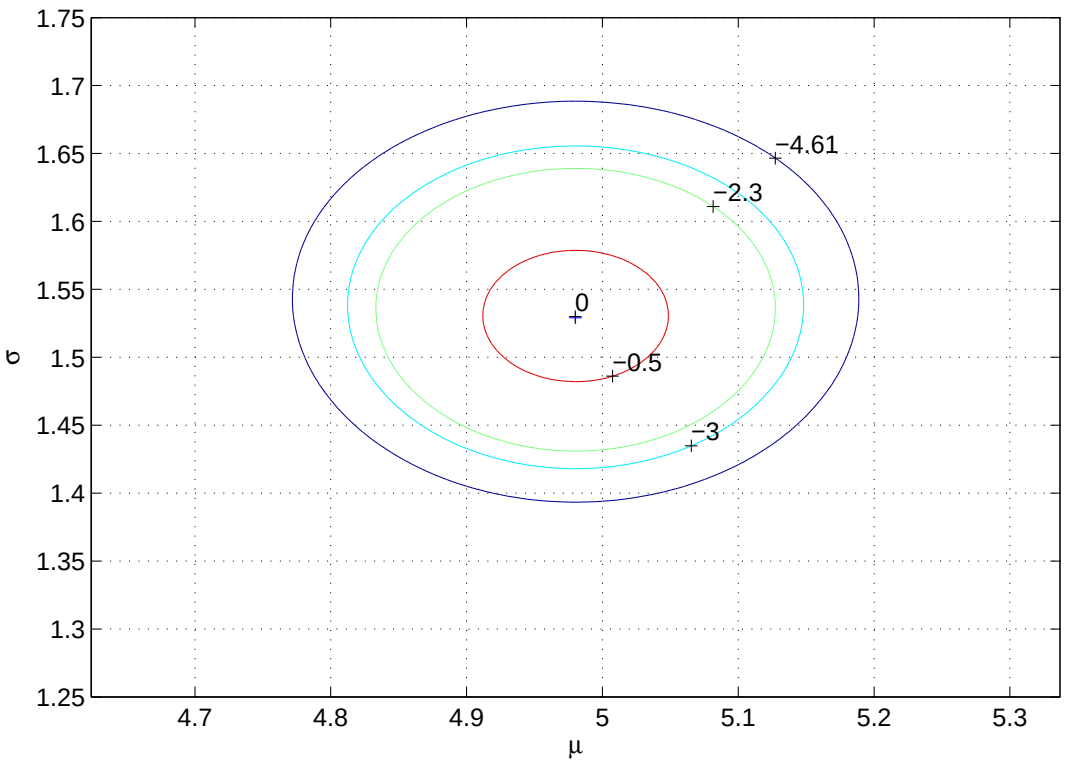
This corresponds to a standard error of $\sigma[\hat{\mu}] = 0.0684$ and $\sigma[\hat{\sigma}] = 0.0483$. The inverse Fisher information matrix is equal to

$$\mathbf{I}_{\mu,\sigma} = \begin{pmatrix} 0.0045 & 0 \; 0 & 0.0022 \end{pmatrix}$$

Maximum Likelihood Estimator

Example (Continuation)

Contour lines of log-likelihood function with probability contents $0.3935, 0.9, 0.95, 0.99$.



Maximum Likelihood Estimator

Numerical Calculation of the Hessian matrix

- Let $\ell(\vartheta_1, \vartheta_2)$ be the log-likelihood function, with maximum in $(\hat{\vartheta}_1, \hat{\vartheta}_2)$.
- Furthermore let $\varepsilon_1, \varepsilon_2 > 0$ and

$$\mathbf{L} = \begin{pmatrix} \ell(\hat{\vartheta}_1 - \varepsilon_1, \hat{\vartheta}_2 - \varepsilon_2) & \ell(\hat{\vartheta}_1, \hat{\vartheta}_2 - \varepsilon_2) & \ell(\hat{\vartheta}_1 + \varepsilon_1, \hat{\vartheta}_2 - \varepsilon_2) \\ \ell(\hat{\vartheta}_1 - \varepsilon_1, \hat{\vartheta}_2) & \ell(\hat{\vartheta}_1, \hat{\vartheta}_2) & \ell(\hat{\vartheta}_1 + \varepsilon_1, \hat{\vartheta}_2) \\ \ell(\hat{\vartheta}_1 - \varepsilon_1, \hat{\vartheta}_2 + \varepsilon_2) & \ell(\hat{\vartheta}_1, \hat{\vartheta}_2 + \varepsilon_2) & \ell(\hat{\vartheta}_1 + \varepsilon_1, \hat{\vartheta}_2 + \varepsilon_2) \end{pmatrix}$$

- The Hessian matrix is then equal to

$$\mathbf{H} = \begin{pmatrix} \dfrac{\mathbf{L}_{21} + \mathbf{L}_{23} - 2\,\mathbf{L}_{22}}{\varepsilon_1^2} & \dfrac{\mathbf{L}_{11} + \mathbf{L}_{33} - \mathbf{L}_{13} - \mathbf{L}_{31}}{4\,\varepsilon_1 \varepsilon_2} \\ \dfrac{\mathbf{L}_{11} + \mathbf{L}_{33} - \mathbf{L}_{13} - \mathbf{L}_{31}}{4\,\varepsilon_1 \varepsilon_2} & \dfrac{\mathbf{L}_{12} + \mathbf{L}_{32} - 2\,\mathbf{L}_{22}}{\varepsilon_2^2} \end{pmatrix}$$

Section 17: Mixed Distributions and the EM Algorithm

Mixed Distributions and the EM Algorithm

- Experimental observations often come from different distributions, e.g. signal and background.
- Such data can be described by a **mixed distribution**.

Definition (Mixed Distribution)

A **mixing distribution** with k components is a distribution whose density has the following form:

$$f(x) = \sum_{j=1}^{k} w_j f_j(x), \quad w_j \geq 0, \quad \sum_{j=1}^{k} w_j = 1$$

. The w_j are nonnegative and are called the **weights** of the components.

- The components f_j are typically normal distributions or other simple distributions.

Mixed Distributions and the EM Algorithm

- The l-th moment μ'_l around 0 of the mixture distribution is the mixture of the corresponding moments μ'_{lj} around 0 of the components:

$$\mu'_l = \sum_{j=1}^{k} w_j \mu'_{lj}$$

expectation and variance of a mixture distribution

Let a mixture distribution consist of k components with means μ_j and variances σ_j^2. Then the mean μ and variance σ^2 of the mixture distribution are given by:

$$\mu = \sum_{j=1}^{k} w_j \mu_j, \qquad \sigma^2 = \sum_{j=1}^{k} w_j \left(\mu_j^2 + \sigma_j^2\right) - \mu^2$$

Example

- A mixture distribution of k normally distributed components has $3k-1$ parameters, viz.

$$\boldsymbol{\vartheta} = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, w_1, \ldots, w_{k-1})$$

- A mixture distribution of k exponentially distributed components has $2k-1$ parameters, viz.

$$\boldsymbol{\vartheta} = (\tau_1, \ldots, \tau_k, w_1, \ldots, w_{k-1})$$

Mixed Distributions and the EM Algorithm

- The estimation of the parameters can be done with the maximum likelihoodmethod.
- In the case of a mixture of normal distributions, the log-likelihood function of a sample $\boldsymbol{X}$ of size n is:

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} w_j \cdot \varphi(x_i; \mu_j, \sigma_j^2) \right)$$

- The log-likelihood function must be maximized numerically, subject to the constraint $\sum_{j=1}^{k} w_j = 1$.
- The log-likelihood function may have several local maxima. In this case, repeated maximization with different initial values is recommended.

Mixed Distributions and the EM Algorithm

- One method for maximizing the log-likelihood function is the **EM algorithm**.
- The EM algorithm is iterative. The choice of initial values can influence which local maximum is reached.
- In each iteration, Bayes' theorem is used to calculate the a posteriori probabilities p_{ij} that observation i comes from component j.
- The means and variances of the components are then estimated by weighted sample means and weighted sample variances, respectively.
- These two steps are iterated until the estimate stabilizes.
- The log-likelihood function cannot get smaller in any iteration; however, reaching the global maximum is **not** guaranteed.

EM algorithm for a mixture of normal distributions

1. Choice of initial values for mixture parameters $\boldsymbol{\vartheta} = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, w_1, \ldots, w_k)$
2. Calculation of p_{ij} and p_j:

$$p_{ij} = \frac{w_j \, \varphi(X_i; \mu_j, \sigma_j^2)}{\sum_{l=1}^{k} w_l \; \varphi(X_i; \mu_l, \sigma_l^2)}, \quad p_j = \sum_{i=1}^{n} p_{ij}$$

3. Estimation of weights and moments:

$$w_j = \frac{p_j}{n}, \; \mu_j = \frac{\sum_{i=1}^{n} p_{ij} X_i}{p_j}, \; \sigma_j^2 = \frac{\sum_{i=1}^{n} p_{ij} (X_i - \mu_j)^2}{p_j}$$

4. Repeat steps 2 and 3 until convergence.

Example

Data set 2 is a mixture of 500 values from $\mathrm{No}(5, 1)$ and 100 values from $\mathrm{No}(8, 9)$. The EM algorithm yields the following estimates of the mixture parameters:

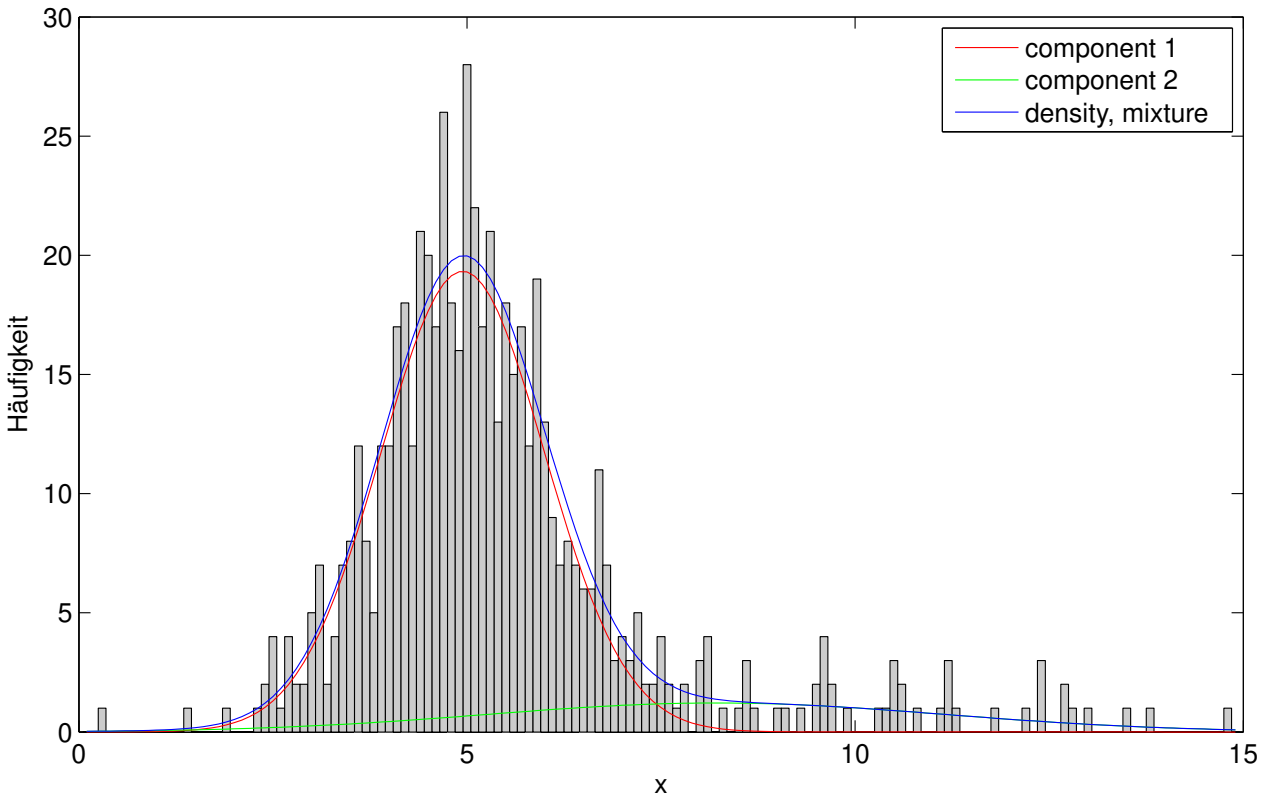
$$\mu_1 = 4.946, \sigma_1 = 1.053, w_1 = 0.850$$
$$\mu_2 = 8.206, \sigma_2 = 2.944, w_2 = 0.150$$

Maximizing the log-likelihood function with the simplex algorithm of Nelder and Mead yields almost identical values:

$$\mu_1 = 4.946, \sigma_1 = 1.052, w_1 = 0.850$$
$$\mu_2 = 8.195, \sigma_2 = 2.946, w_2 = 0.150$$

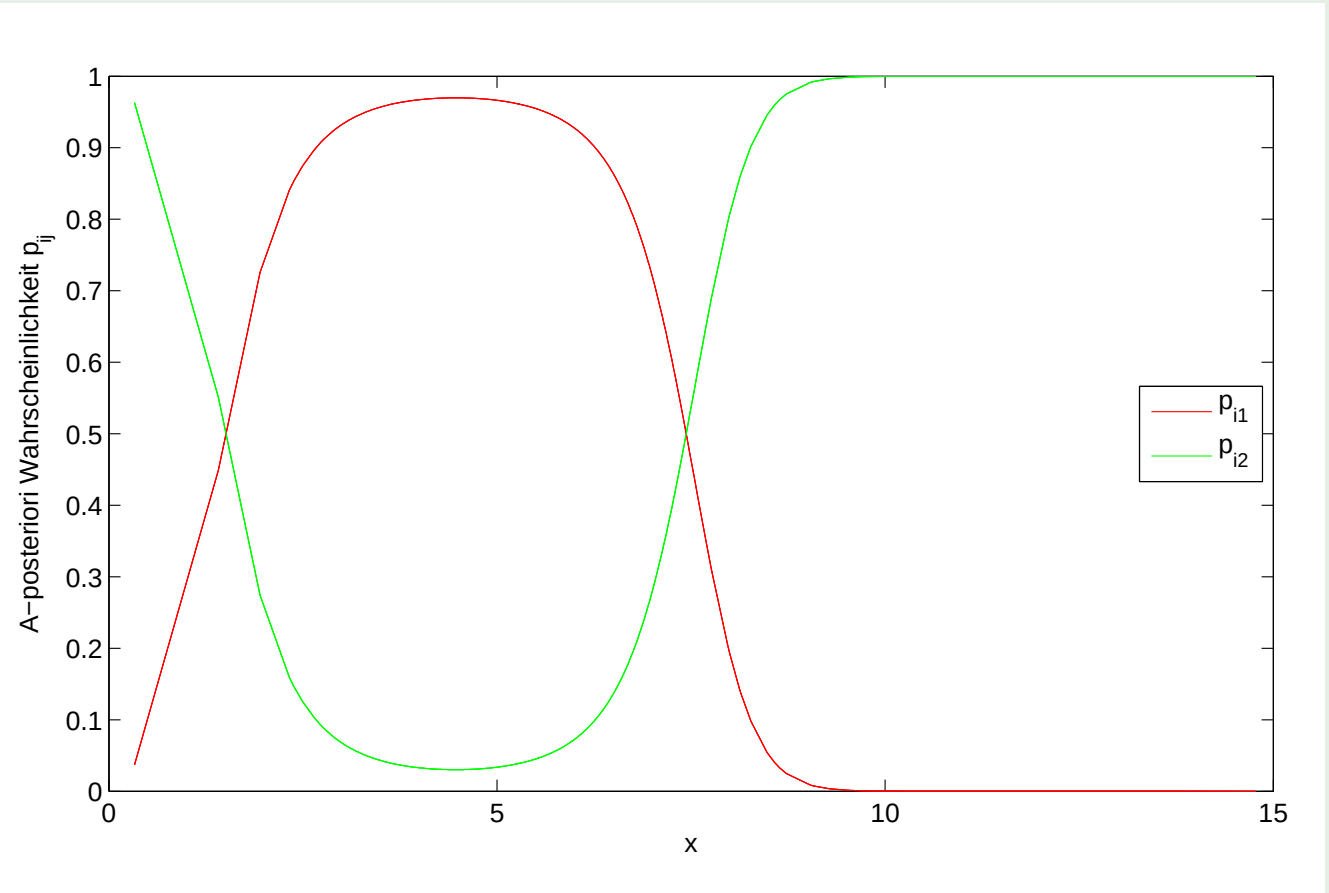
Example (Continuation)

Densities of components and mixture density:



Example (Continuation)

A-posteriori probabilities p_{ij}:



Part V

Confidence Intervals

Overview Part 5

Section 18: Basic Terminology

Basic Terminology

- Besides the estimated value itself, its dispersion around the true value is also of interest.
- We want to determine an interval from a sample that contains the true value with a certain probability.

Definition (Confidence Interval)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from the distribution F with unknown parameter ϑ and $0 < \alpha < 1$. An interval with boundaries $G_1 = g_1(\boldsymbol{X})$ and $G_2 = g_2(\boldsymbol{X})$ is called an **confidence interval** with certainty $1 - \alpha$ if holds:

$$W(G1 \leq G2) = 1$$
$$W(G1 \leq \vartheta \leq G2) \geq 1 - \alpha$$

Such an interval is called a $(1 - \alpha)$ confidence interval, in short.

Basic Terminology

- For each value of certainty $1-\alpha$ there are many different confidence intervals.
- If F is continuous, there are infinitely many confidence intervals with certainty $1-\alpha$.
- If F is discrete, the confidence is usually greater than $1-\alpha$.
- A **symmetric** confidence interval exists if it is true that:

$$W(\vartheta \leq G_1) = W(\vartheta \geq G_2)$$

- A **one-sided** confidence interval exists if it is true that:

lower tail: $W(\vartheta \leq G_2) \geq 1-\alpha$ or

upper tail: $W(G_1 \leq \vartheta) \geq 1-\alpha$

Section 19: General Construction According to Neyman

General Construction According to Neyman

- Let $Y = h(\boldsymbol{X})$ be a sampling function. The distribution G of Y then also depends on the unknown parameter ϑ.
- For each value of ϑ, we determine an **prediction interval** $[y_1(\vartheta), y_2(\vartheta)]$ of level $1 - \alpha$:

$$W(y_1(\vartheta) \leq Y \leq y_2(\vartheta)) \geq 1 - \alpha$$

- If the observation is equal to $Y = y_0$, the confidence interval $[G_1(Y), G_2(Y)]$ is given by:

$$G1 = \min_{\vartheta}\{\vartheta | y_1(\vartheta) \leq y_0 \leq y_2(\vartheta)\}$$

$$G2 = \max_{\vartheta}\{\vartheta | y_1(\vartheta) \leq y_0 \leq y_2(\vartheta)\}$$

General Construction According to Neyman

Example

Let $\boldsymbol{X}$ be a sample of $\mathrm{No}(\mu, \sigma^2)$ with unknown variance σ^2. Then $(n-1)S^2/\sigma^2$ χ^2-distributed with $n-1$ degrees of freedom. Therefore, it holds:

$$W\left(\frac{\sigma^2\chi^2_{\alpha/2;n-1}}{n-1} \leq S^2 \leq \frac{\sigma^2\chi^2_{1-\alpha/2;n-1}}{n-1}\right) = 1-\alpha$$

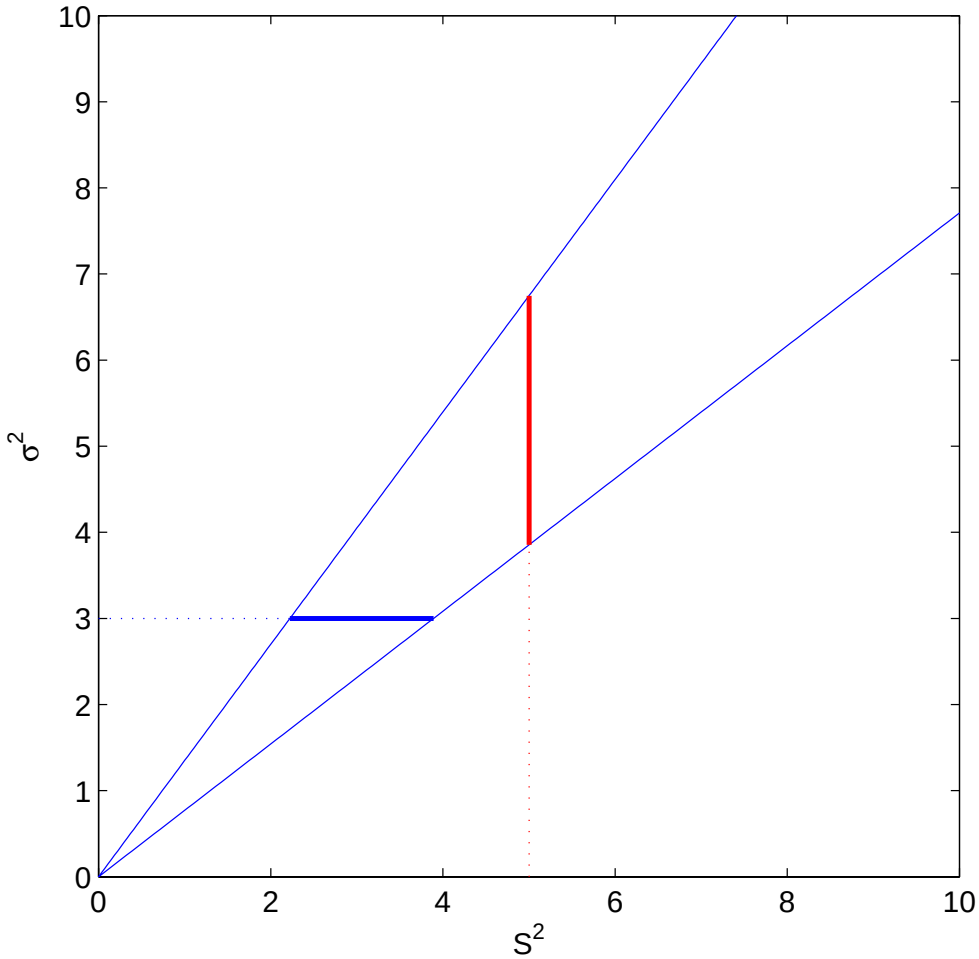
The expression in the parenthesis can be transformed to:

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}$$

From this follows

$$G_1(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}, \quad G_2(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}$$

General Construction According to Neyman



Blue: prediction interval for $\sigma^2 = 3$; red: confidence interval for $S^2 = 5$.

Section 20: Normally Distributed Data

Subsection: Mean

18 Basic Terminology

19 General Construction According to Neyman

20 Normally Distributed Data

- Mean
- Variance
- Difference between Two Mean Values

21 Exponentially Distributed Data

22 Poisson Distributed Data

23 Binomially Distributed Data

24 Data from Other Distributions

Mean

- Let $\boldsymbol{X}$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$.
- $\overline{X}$ is normally distributed according to $\mathrm{No}(\mu, \sigma^2/n)$.
- If σ^2 is known, the standard score is.

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

standard normally distributed. From

$$W(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

then follows

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

Mean

Symmetric AI for the Mean, Known Variance

$$G_1(\boldsymbol{X}) = \overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}, \quad G_2(\boldsymbol{X}) = \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}$$

- The lower and upper tail confidence intervals are constructed, analogously.

Lower tail CI for the mean, known variance

$$G_1(\boldsymbol{X}) = -\infty, \quad G_2(\boldsymbol{X}) = \overline{X} + z_{1-\alpha}\sigma/\sqrt{n}$$

Upper tail AI for the mean, known variance

$$G_1(\boldsymbol{X}) = \overline{X} - z_{1-\alpha}\sigma/\sqrt{n}, \quad G_2(\boldsymbol{X}) = \infty$$

Mean

quantiles of the standard normal distribution

α	$z_{1-\alpha/2}$	$z_{1-\alpha}$
0.001	3.29	3.09
0.002	3.09	2.88
0.005	2.81	2.58
0.01	2.58	2.33
0.02	2.33	2.05
0.05	1.96	1.64
0.1	1.64	1.28

Mean

- If σ^2 is unknown, σ^2 is estimated by the sample variance S^2. The standard score

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

is then t-distributed with $n-1$ degrees of freedom. From

$$W(-t_{1-\alpha/2;n-1} \leq T \leq t_{1-\alpha/2;n-1}) = 1 - \alpha$$

follows

Symmetric CI for the Mean, Unknown Variance

$$G_1(\boldsymbol{x}) = \overline{X} - t_{1-\alpha/2;n-1} S/\sqrt{n},\ G_2(\boldsymbol{X}) = \overline{X} + t_{1-\alpha/2;n-1} S/\sqrt{n}$$

Mean

Example

A sample of size $n = 50$ from the standard normal distribution has sample mean $\overline{X} = 0.0540$ and sample variance $S^2 = 1.0987$. If the variance is assumed to be known, the symmetric 95%-confidence interval for μ is:

$$G_1 = 0.0540 - 1.96/\sqrt{50} = -0.2232$$
$$G_2 = 0.0540 + 1.96/\sqrt{50} = 0.3312$$

If the variance is assumed to be unknown, the symmetric 95%-confidence interval for μ is:

$$G_1 = 0.0540 - 2.01 \cdot 1.0482/\sqrt{50} = -0.2439$$
$$G_2 = 0.0540 + 2.01 \cdot 1.0482/\sqrt{50} = 0.3519$$

MATLAB: `make_KI_normal`

Subsection: Variance

Variance

- Let $X_1, \ldots, X_n$ be a sample from the normal distribution $\mathrm{No}(\mu, \sigma^2)$.
- $(n-1)S^2/\sigma^2$ is χ^2-distributed with $n-1$ degrees of freedom. From

$$W\left(\chi^2_{\alpha/2;n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2;n-1}\right) = 1-\alpha$$

follows

Symmetric CI for the Variance

$$G_1(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}, \quad G_2(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}$$

Variance

Lower Tail CI for the Variance

$$G_1(\boldsymbol{X}) = 0, \quad G_2(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{\alpha;n-1}}$$

right-sided AI for variance

$$G_1(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{1-\alpha;n-1}}, \quad G_2(\boldsymbol{X}) = \infty$$

Variance

Example

A sample of size $n = 50$ from the normal distribution $\mathrm{No}(0, 4)$ has sample variance $S^2 = 4.3949$. The symmetric 95%-confidence interval for σ^2 is:

$$\begin{aligned} G_1 &= 49 \cdot 4.3949/70.2224 = 3.0667 \\ G_2 &= 49 \cdot 4.3949/31.5549 = 6.8246 \end{aligned}$$

If the quantiles of the χ^2 distribution $\chi^2(n-1)$ are replaced by the quantiles of the normal distribution $\mathrm{No}(n-1, 2(n-1))$, the confidence intervalis:

$$\begin{aligned} G_1 &= 49 \cdot 4.3949/68.4027 = 3.1483 \\ G_2 &= 49 \cdot 4.3949/29.5973 = 7.2760 \end{aligned}$$

MATLAB: `make_KI_normal_variance.m`

Subsection: Difference between Two Mean Values

Difference between Two Mean Values

- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent samples from the normal distributions $\mathrm{No}(\mu_x, \sigma_x^2)$ respectively, $\mathrm{No}(\mu_y, \sigma_y^2)$.
- We are looking for a confidence interval for $\mu_x - \mu_y$. The difference $D = \overline{X} - \overline{Y}$ is normally distributed according to $\mathrm{No}(\mu_x - \mu_y, \sigma^2)$, with $\sigma_D^2 = \sigma_x^2/n + \sigma_y^2/m$.
- If the variances are known, the standard score of D is standard normally distributed. From

$$W\left(-z_{1-\alpha/2} \leq \frac{D - (\mu_x - \mu_y)}{\sigma_D} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

follows

Symmetric CI for the Difference between Two Mean Values

$$G_1(\boldsymbol{X}) = D - z_{1-\alpha/2}\sigma_D, \quad G_2(\boldsymbol{X}) = D + z_{1-\alpha/2}\sigma_D$$

Difference between Two Mean Values

- If the variances are unknown and equal, is

$$S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

  χ^2-distributed with $k = n + m - 2$ degrees of freedom.

- The standard score

$$T = \frac{D - (\mu_x - \mu_y)}{S_D}$$

  with $S_D = S\sqrt{1/n + 1/m}$ is therefore t-distributed with $k = n + m - 2$ degrees of freedom.

Difference between Two Mean Values

- Off

$$W\left(-t_{1-\alpha/2;k} \leq T \leq t_{1-\alpha/2;k}\right) = 1-\alpha$$

follows

Symmetric CI for the Difference between Two Means, Unknown Variance

$$G_1(\boldsymbol{X}) = D - t_{1-\alpha/2;k} S_D, \quad G_2(\boldsymbol{X}) = D + t_{1-\alpha/2;k} S_D$$

with $k = n + m - 2$.

Difference between Two Mean Values

Example

A sample of $\mathrm{No}(2, 4)$ of size $n = 50$ has sample mean $\overline{X} = 2.1080$ and sample variance $S_x^2 = 4.3949$; a second sample of $\mathrm{No}(1, 4)$ of size $m = 25$ has sample mean $\overline{X} = 1.6692$ and sample variance $S_x^2 = 5.2220$. If the variances are assumed to be known, the 95%-confidence interval for $\mu_x - \mu_y$ is:

$$\begin{aligned} G_1 &= 0.4388 - 1.96 \cdot 0.4899 = -0.5213 \\ G_2 &= 0.4388 + 1.96 \cdot 0.4899 = 1.3990 \end{aligned}$$

Difference between Two Mean Values

Example (Continuation)

If the variances are assumed to be unknown, $S^2 = 4.6668$ and $S_D = 0.5292$. The 95%=confidence interval for $\mu_x - \mu_y$ is then:

$$G_1 = 0.4388 - 1.993 \cdot 0.5292 = -0.6158$$
$$G_2 = 0.4388 + 1.993 \cdot 0.5292 = 1.4935$$

MATLAB: `make_KI_normal_difference.m`

Section 21: Exponentially Distributed Data

Exponentially Distributed Data

- Let $\boldsymbol{X}$ be a sample from the exponential distribution $\mathrm{Ex}(\tau)$.
- Then the sample mean $\overline{X}$ is gamma distributed according to $\mathrm{Ga}(n, \tau/n)$ and has the following density:

$$f(x) = \frac{x^{n-1}}{(\tau/n)^n \Gamma(n)} \exp\left(-\frac{x}{\tau/n}\right)$$

- Therefore, for each τ holds:

$$W\left(\gamma_{\alpha/2;n,\tau/n} \leq \overline{X} \leq \gamma_{1-\alpha/2;n,\tau/n}\right) = 1 - \alpha$$

Exponentially Distributed Data

- From this follows

$$W\left(\gamma_{\alpha/2;n,1/n} \leq \frac{\overline{X}}{\tau} \leq \gamma_{1-\alpha/2;n,1/n}\right) = 1-\alpha$$

and

$$W\left(\frac{\overline{X}}{\gamma_{1-\alpha/2;n,1/n}} \leq \tau \leq \frac{\overline{X}}{\gamma_{\alpha/2;n,1/n}}\right) = 1-\alpha$$

- Thus:

Symmetric CI for the Mean

$$G_1(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{1-\alpha/2;n,1/n}}, \quad G_2(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{\alpha/2;n,1/n}}$$

Exponentially Distributed Data

Lower tail CI for the Mean Value

$$G_1(\boldsymbol{X}) = 0, \quad G_2(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{\alpha;n,1/n}}$$

Upper tail CI for the Mean Value

$$G_1(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{1-\alpha;n,1/n}}, \quad G_2(\boldsymbol{X}) = \infty$$

Section 22: Poisson Distributed Data

Poisson Distributed Data

- Let K be an observation from the Poisson distribution $\mathrm{Po}(\lambda)$.
- The quantiles of the Poisson distribution can be calculated using the distribution function of the gamma distribution.
- For the practical calculation of the confidence interval, therefore, the **quantiles of the gamma distribution** can be used.

Symmetric CI for the Mean

$$G_1(K) = \gamma_{\alpha/2;K,1}, \quad G_2(K) = \gamma_{1-\alpha/2;K+1,1}$$

Poisson Distributed Data

- If n observations $K_1, \ldots, K_n$ are available, then $L = \sum K_i$ is Poisson distributed with mean $n\lambda$. The symmetric confidence interval for λ is then:

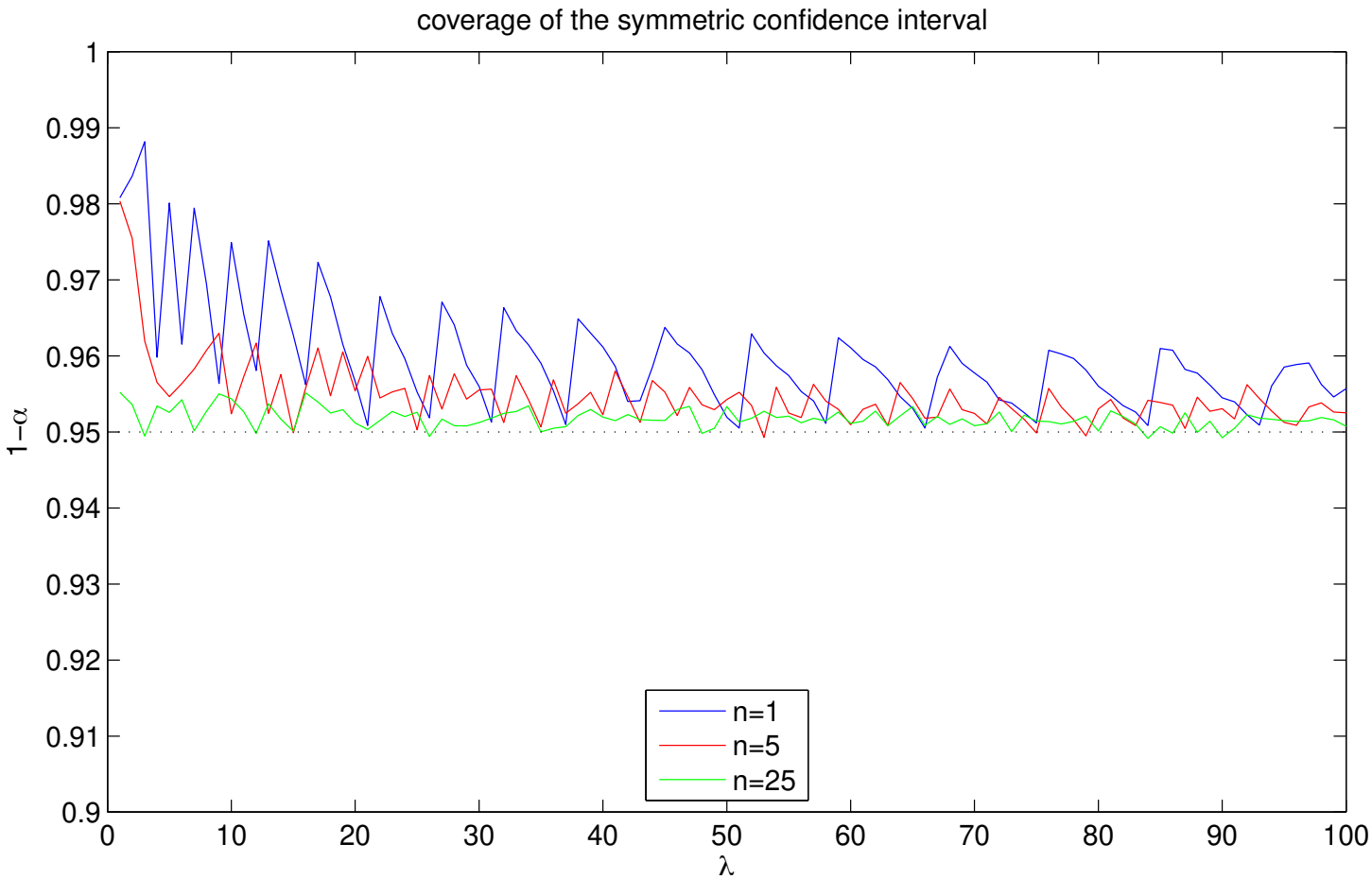
Symmetric CI for the Mean

$$G_1(L) = \gamma_{\alpha/2;L,1/n}, \quad G_2(L) = \gamma_{1-\alpha/2;L+1,1/n}$$

- This interval is **conservative** in the sense that the certainty is practically always greater than $1 - \alpha$.

MATLAB: `make_KI_Poisson_symmetric`

Poisson Distributed Data



Poisson Distributed Data

- The one-sided confidence intervals are determined in an analogous way.

Lower tail CI for the Mean Value

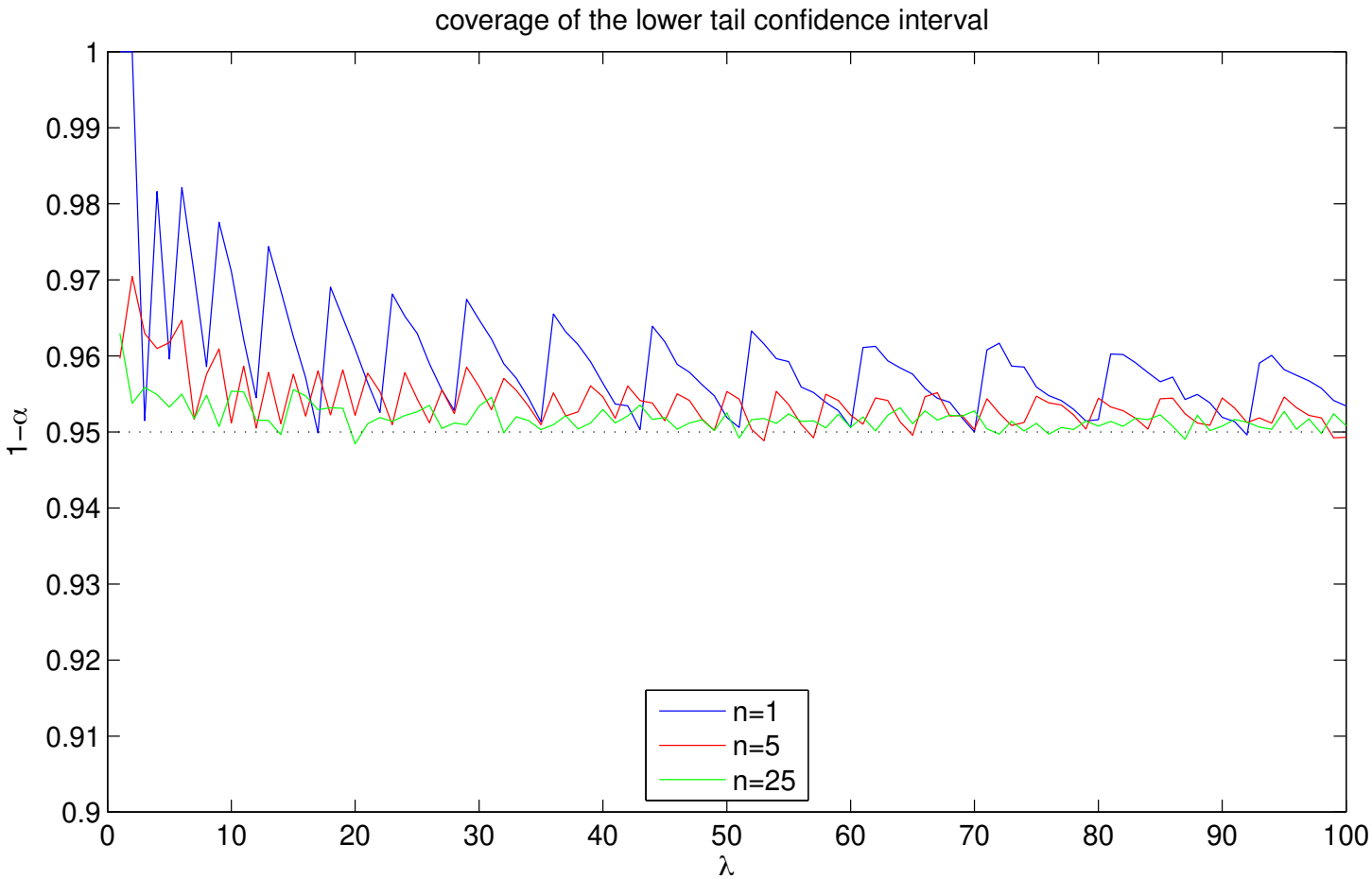
$$G_1(L) = 0, \quad G_2(L) = \gamma_{1-\alpha;L+1,1/n}$$

Upper tail CI for the Mean Value

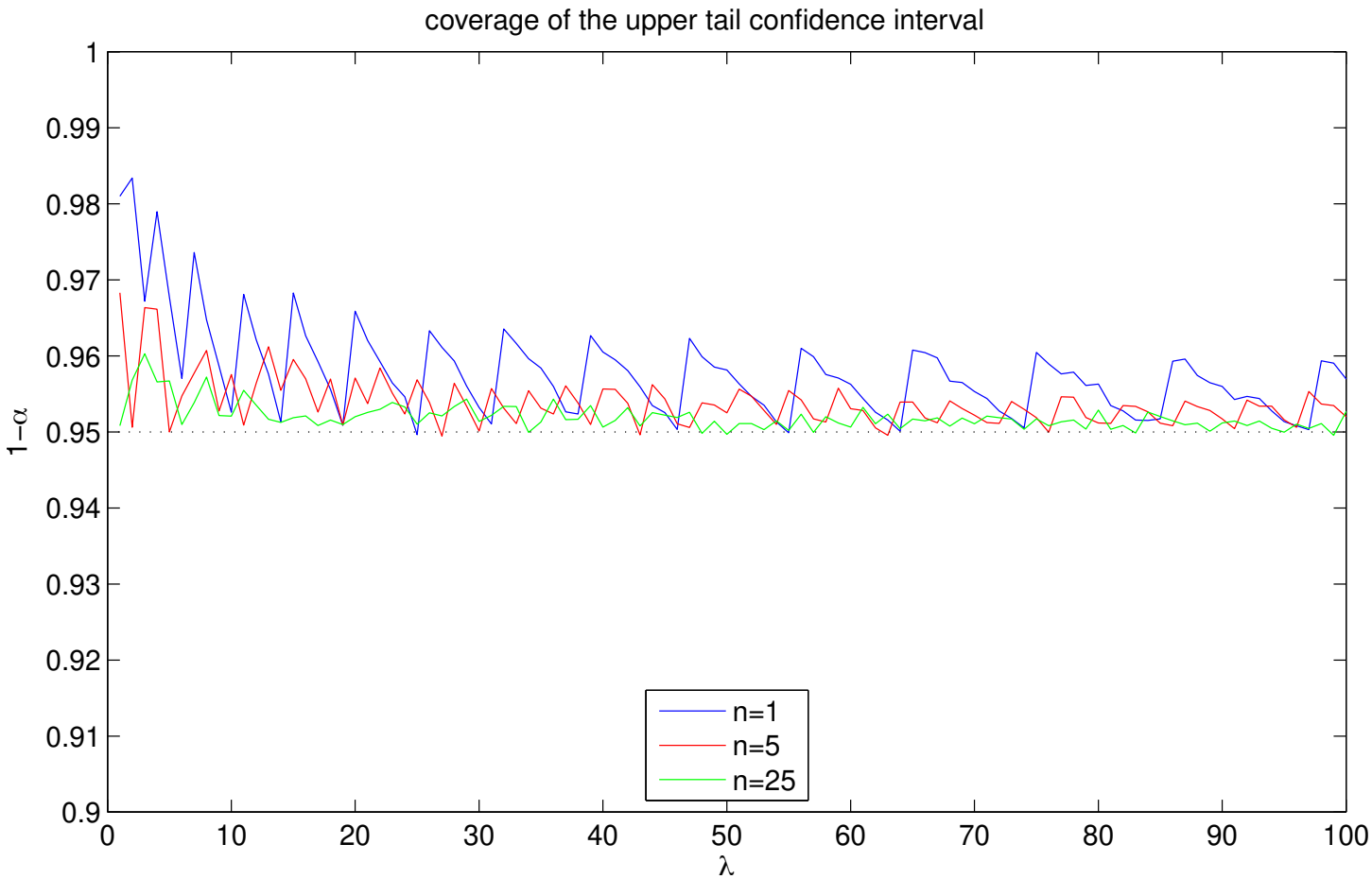
$$G_1(L) = \gamma_{\alpha;L,1/n}, \quad G_2(L) = \infty$$

- MATLAB: `make_KI_Poisson_left`
- MATLAB: `make_KI_Poisson_right`

Poisson Distributed Data



Poisson Distributed Data



Section 23: Binomially Distributed Data

Binomially Distributed Data

- Let K be an observation from the binomial distribution $\mathrm{Bi}(n, p)$.
- The quantiles of the binomial distribution can be calculated using the distribution function of the beta distribution.
- For the practical calculation of the confidence interval the **quantiles of the beta distribution** can be used.

Symmetric AI according to Clopper and Pearson

$$G_1(K) = \beta_{\alpha/2;K,n-K+1}, \quad G_2(K) = \beta_{1-\alpha/2;K+1,n-K}$$

- Special cases: for $K = 0$, $G_1(0) = 0$, for $K = n$, $G_2(n) = 1$.
- This interval is **conservative** in the sense that the certainty is practically always greater than $1 - \alpha$.

Binomially Distributed Data

- The one-sided confidence intervals are determined in an analogous way.

Lower tail AI according to Clopper and Pearson

$$G_1(K) = 0, \quad G_2(K) = \beta_{1-\alpha;K+1,n-K}$$

- If $K = 0$, the left-hand confidence interval extends from 0 to $\beta_{1-\alpha;1,n} = 1 - \sqrt[n]{\alpha}$.

Upper tail AI according to Clopper and Pearson

$$G_1(K) = \beta_{\alpha;K,n-K+1}, \quad G_2(K) = 1$$

- If $K = n$, the right-hand confidence interval extends from $\beta_{\alpha;n,1} = \sqrt[n]{\alpha}$ to 1.

Binomially Distributed Data

- For sufficiently large n, $\hat{p} = K/n$ is approximately normally distributed according to $\mathrm{No}(p, p(1-p)/n)$.
- The standard score

$$Z = \frac{\hat{p} - p}{\sigma[p]}$$

  is then approximately standard normally distributed.
- Off

$$W(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

  follows

$$W(\hat{p} - z_{1-\alpha/2}\sigma[p] \leq p \leq \hat{p} + z_{1-\alpha/2}\sigma[p]) = 1 - \alpha$$

Binomially Distributed Data

- Since p is not known, $\sigma[p]$ must be approximated.
- **bootstrap method**: p is approximated by $\hat{p}$:

$$\sigma[p] \approx \sigma[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Symmetric bootstrapped CI

$$G_1(K) = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$G_2(K) = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Binomially Distributed Data

- **Robust method**: p is chosen so that $\sigma[p]$ is maximal, i.e. $p = 0.5$ and

$$\sigma[p] \approx \frac{1}{2\sqrt{n}}$$

Symmetric AI using the Robust Method

$$G_1(K) = \hat{p} - z_{1-\alpha/2}\frac{1}{2\sqrt{n}}$$

$$G_2(K) = \hat{p} + z_{1-\alpha/2}\frac{1}{2\sqrt{n}}$$

Binomially Distributed Data

- procedure according to **Agresti–Coull**: We set

$$c = z_{1-\alpha/2}, K' = K + c^2/2, n' = n + c^2, p' = K'/n'$$

and

$$\sigma[p'] = \sqrt{\frac{p'(1-p')}{n'}}$$

Symmetric AI according to Agresti-Coull

$$G_1(K) = p' - z_{1-\alpha/2}\sqrt{\frac{p'(1-p')}{n}}$$

$$G_2(K) = p' + z_{1-\alpha/2}\sqrt{\frac{p'(1-p')}{n}}$$

Binomially Distributed Data

Example

Angabe: In a survey of $n = 400$ people, $k = 157$ people say they know product X. We are looking for a 95% confidence interval for the level of familiarity p.

Clopper-Pearson:

$$G_1(k) = \beta_{0.025;157,244} = 0.3443$$
$$G_2(k) = \beta_{0.975;158,243} = 0.4423$$

approximation by normal distribution:

It is $\hat{p} = 0.3925$ and $z_{0.975} = 1.96$. Using the bootstrap procedure, $\sigma[\hat{p}] = 0.0244$ is obtained. The limits of the confidence interval are therefore

$$G_1 = 0.3925 - 1.96 \cdot 0.0244 = 0.3446$$
$$G_2 = 0.3925 + 1.96 \cdot 0.0244 = 0.4404$$

Binomially Distributed Data

Example (Continuation)

Using the robust procedure, we get $\sigma[\hat{p}] = 0.025$ and the bounds

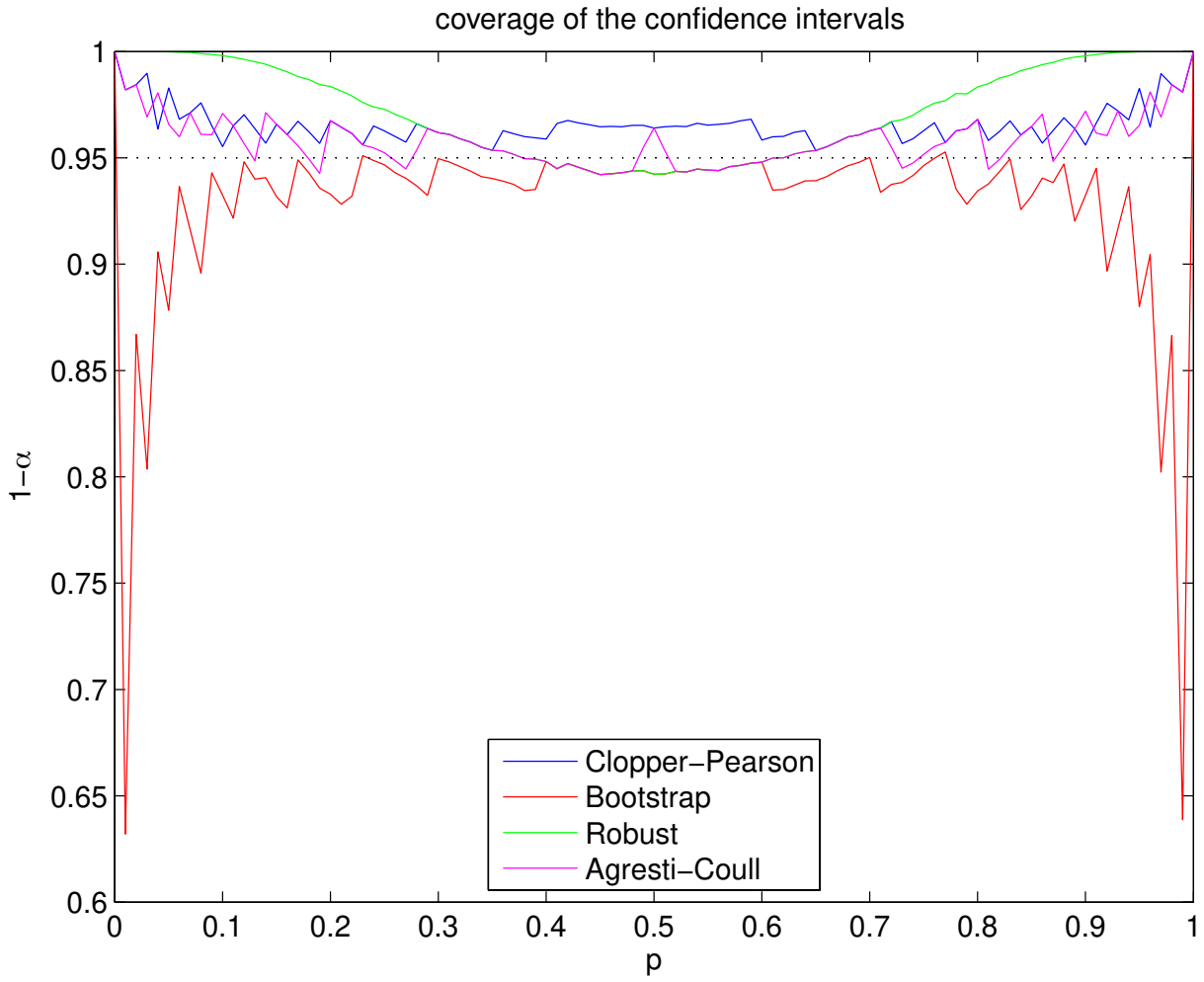
$$G_1 = 0.3925 - 1.96 \cdot 0.025 = 0.3435$$
$$G_2 = 0.3925 + 1.96 \cdot 0.025 = 0.4415$$

The robust interval is only marginally longer than the bootstrap interval.
With the Agresti-Coull correction, we get $\hat{p} = 0.3936$. The limits of the confidence interval are then

$$G_1 = 0.3936 - 1.96 \cdot 0.0244 = 0.3457$$
$$G_2 = 0.3936 + 1.96 \cdot 0.0244 = 0.4414$$

MATLAB: `make_KI_binomial`

Binomially Distributed Data



Section 24: Data from Other Distributions

Data from Other Distributions

- Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from distribution F with mean μ and variance σ^2.
- Given the central limit theorem, the standard score Z of the sample mean is:

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

for large samples approximately normally distributed. Thus, **approximate**

$$W(\overline{X} - z_{1-\alpha/2} S/\sqrt{n} \leq \mu \leq \overline{X} + z_{1-\alpha/2} S/\sqrt{n}) \approx 1 - \alpha$$

Approximate AI for the mean

$$G_1(\boldsymbol{X}) = \overline{X} - z_{1-\alpha/2} S/\sqrt{n}, \quad G_2(\boldsymbol{X}) = \overline{X} + z_{1-\alpha/2} S/\sqrt{n}$$

Data from Other Distributions

Example

For exponentially distributed samples of size n, the following table gives the confidence of 95%-confidence interval approximated by normal distribution, estimated from $N = 20000$ samples:

n	25	50	100	200	400
$1-\alpha$	0.9112	0.9289	0.9408	0.9473	0.9476

MATLAB: `make_KI_exponential`

Part VI

Testing Hypotheses

Overview Part 6

Section 25: Introduction

Introduction

- We observe a sample $X_1, \ldots, X_n$ from a distribution F.
- A test is to determine whether the observations are compatible with some assumption about F.
- The assumption is called the **null hypothesis** H_0.
- If the form of F is specified except for one or more parameters, the test is called **parametric**.
- If the form of F is not specified, the test is called **nonparametric** or **parameter free**.
- The test decides whether the sample is consistent with the hypothesis, not whether the hypothesis is correct!
- If the sample is consistent with the hypothesis, the hypothesis may can still deviate from the truth, even if only to a limited extent.

General Procedure

- A test size (test statistic) T is calculated from the sample.
- The range of values of T is divided, depending on H_0, into a **rejection range** (critical range) C and a **acceptance range** C'.
- The acceptance range is usually a **forecast interval** for T.
- If the value of T falls within the rejection range, H_0 is discarded.
- Otherwise, H_0 is retained for the time being.
- However, this is not a confirmation of H_0. It simply means that the data are consistent with the hypothesis.

One-sided and Two-sided Tests

- If the acceptance range is the symmetric prediction interval for T, the test is called **two-sided**. The critical range then breaks down into two subintervals.
- If the acceptance range is an interval of the form $T \leq c$ or $T \geq c$, the test is called **one-sided**. The critical region is then an interval of the form $T > c$ or $T < c$.

Tests and Confidence Intervals

- In many cases, the acceptance range is a confidence interval for the parameter under test.
- The null hypothesis is rejected if the hypothesized value falls outside the confidence interval, because it is then not sufficiently plausible.

The p Value

- The test can alternatively be performed using the p value $P(T)$.
- The p-value indicates how likely it is to observe at least the value T or at most the value T, assuming the null hypothesis.
- Two-sided test: If $F_0(x)$ is the distribution function of T under the null hypothesis, then the p value is equal to

$$P(T) = 2\min(F_0(T), 1 - F_0(T))$$

- One-sided test: if $F_0(x)$ is the distribution function of T under the null hypothesis, then the p value is equal to

$$P(T) = F_0(T) \text{ resp. } p = 1 - F_0(T)$$

- The null hypothesis is rejected if $P(T) < \alpha$.

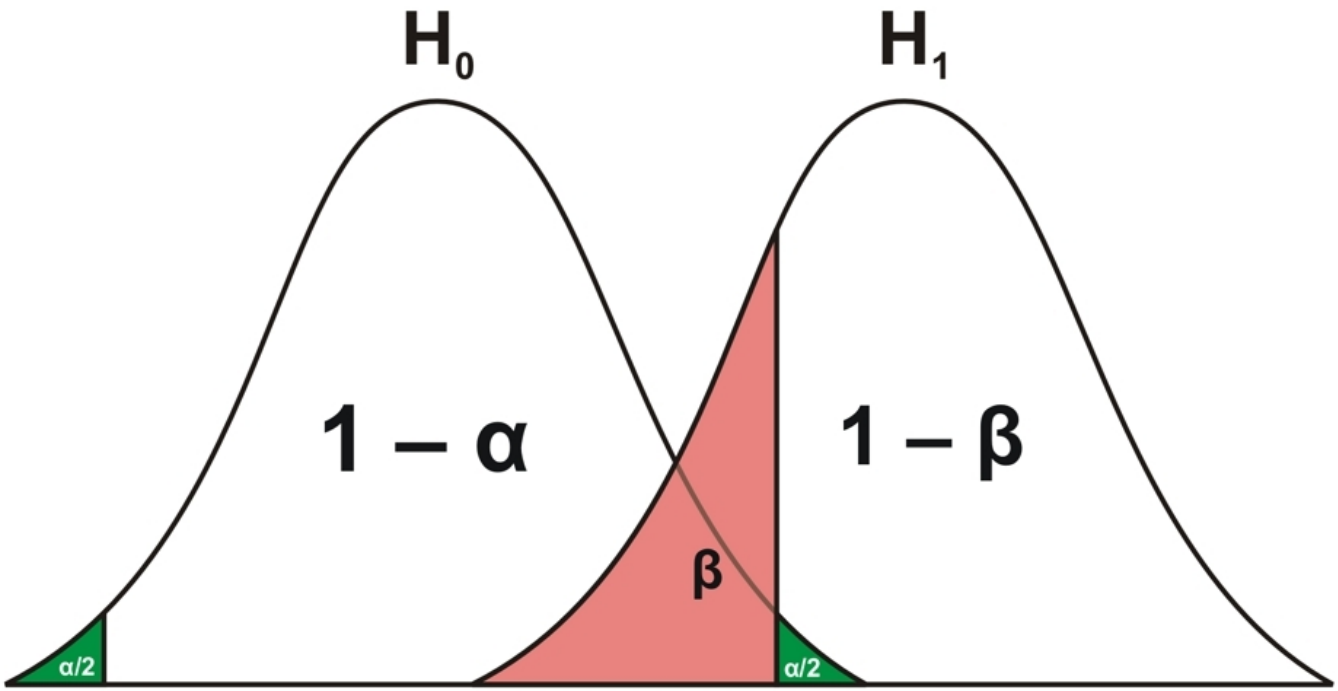
Significance and Quality

- Two types of errors are possible in any test procedure.
  - **error of 1st kind**: The hypothesis H_0 is rejected although it is true.
  - **error 2nd kind**: The hypothesis H_0 is retained, although it is not true.
- The distribution of T assuming H_0 is determined.
- The rejection range is determined such that the probability of a 1st kind error is at most equal to a value α.
- α is called the **significance level** of the test. Common values are $\alpha = 0.05, 0.01, 0.005$.

Introduction

- If the rejection range is defined, the probability $\beta(H_1)$ of a failure 2.2. („alternative hypothesis") H_1 the probability $\beta(H_1)$ of a 2nd kind error can be calculated.
- $1 - \beta(H_1)$ is called the **goodness** of the test for H_1.
- The goodness should never be smaller than α.
- If the goodness is never less than α, the test is called **unbiased**.
- One goal of test theory is to construct unbiased tests with maximum goodness (UMPU).

Introduction



Section 26: Parametric Testing

Subsection: Basics

Basics

- We consider a sample $X_1, \ldots, X_n$ from a distribution F specified down to one or more parameters.
- Tests of hypotheses about F are called **parametric**.
- A null hypothesis H_0 can be thought of as a subset of the parameter space Θ.
- The test decides whether the sample is consistent with the hypothesis.
- Before application, clarify whether the assumed parametric form is plausible.

Basics

- First, the test statistic T and the significance level α are chosen.
- Then the critical region C is set so that

$$W(T \in C | \vartheta \in H_0) \leq \alpha$$

- To a null hypothesis H_0, a **counterhypothesis** H_1 can be formulated.
- H_1 can also be conceived as a subset of the parameter space Θ.
- If the significance level α is fixed, for each $\vartheta \in H_1$ the goodness can be calculated:

$$1 - \beta(\vartheta) = W(T \in C | \vartheta \in H_1)$$

- $1 - \beta(\vartheta)$ is called the **quality function** of the test.

Example with Exponential Distribution

- $X_1, \ldots, X_n$ is an exponentially distributed sample of $\mathrm{Ex}(\tau)$.
- The hypothesis $H_0 : \tau = \tau_0$ is to be tested using the sample.
- As test statistic T we choose the sample mean: $T = \overline{X}$.
- Assuming H_0, T has the following density:

$$f(t) = \frac{t^{n-1}}{(\tau_0/n)^n \Gamma(n)} \exp\left(-\frac{t}{\tau_0/n}\right)$$

- T is thus distributed according to $\mathrm{Ga}(n, \tau_0/n)$.
- The symmetric prediction interval $[y_1(\tau_0), y_2(\tau_0)]$ for T to the level $1-\alpha$ is obtained by:

$$y_1(\tau_0) = \gamma_{\alpha/2;n,\tau_0/n}, \quad y_2(\tau_0) = \gamma_{1-\alpha/2;n,\tau_0/n}$$

Basics

- Therefore, the discard area with significance level α is the set

$$C = [0, y_1(\tau_0)] \cup [y_2(\tau_0), \infty[$$

- So H_0 is rejected if T is "far away" from the hypothetical value τ_0.
- The quality function for a value τ is given by:

$$1 - \beta(\tau) = W(T \in C) = G(y_1(\tau)) + 1 - G(y_2(\tau))$$

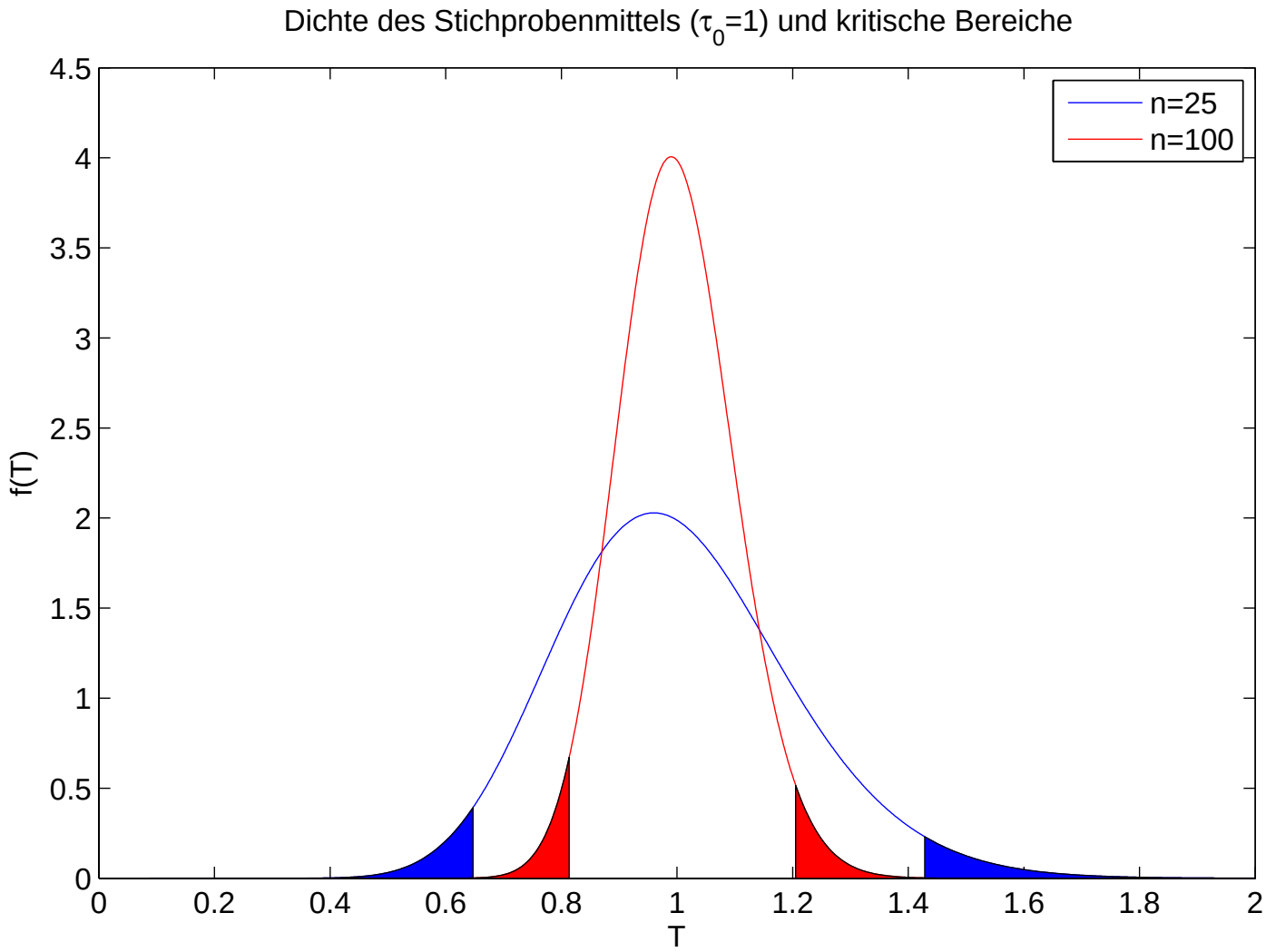
where G is the distribution function of the $\mathrm{Ga}(n, \tau/n)$ distribution.

- The test is not unbiased because, for example, for $\tau_0 = 1$ and $n = 25$.

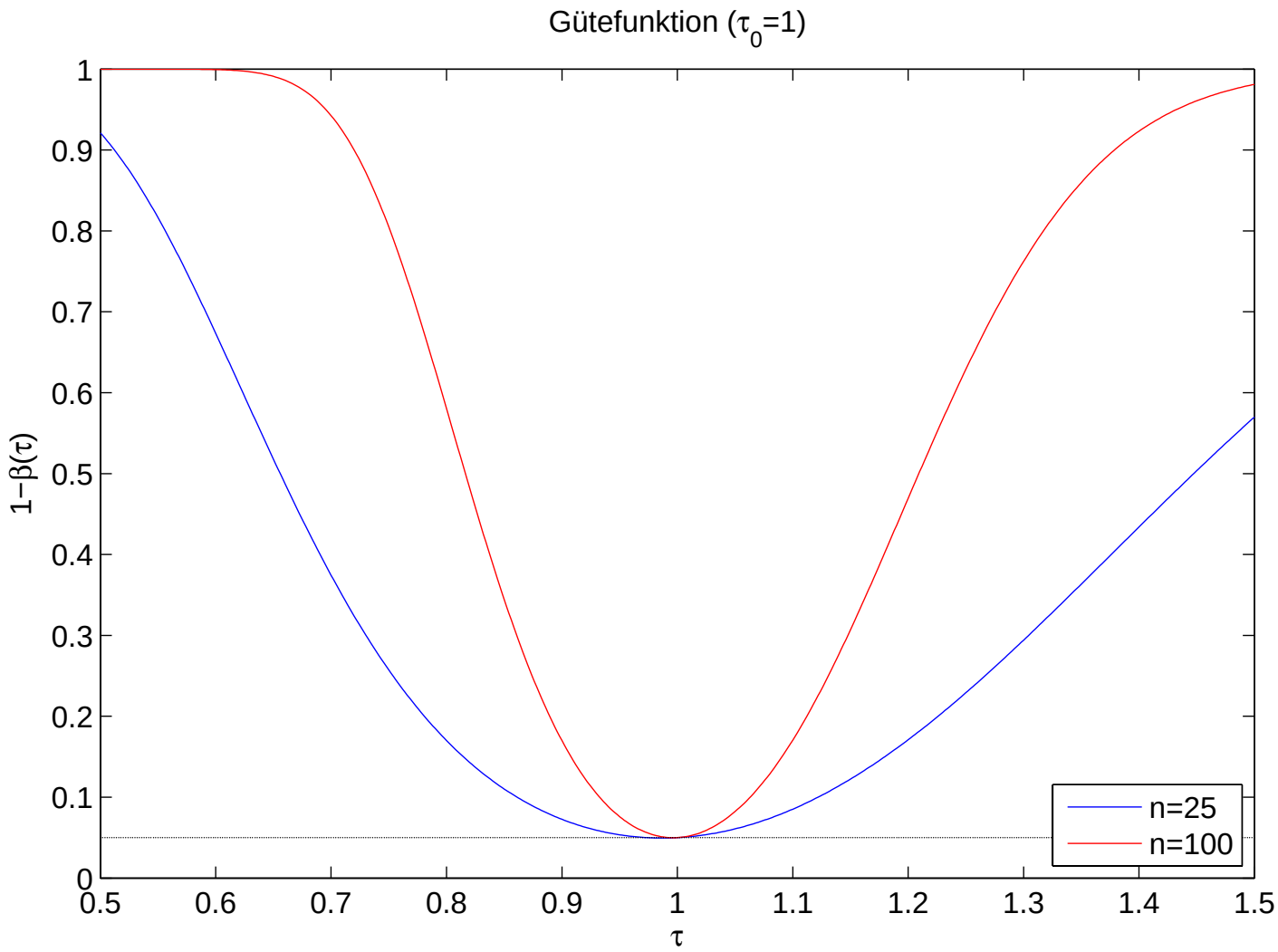
$$1 - \beta(0.986) = 0.0495 < \alpha$$

MATLAB: `make_test_exponential_mean.m`

Basics


Dichte des Stichprobenmittels (τ_0=1) und kritische Bereiche
n=25
n=100
f(T)
T
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

Basics


Gütefunktion (τ_0=1)
1−β(τ)
τ
n=25
n=100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5

Subsection: Tests for Normally Distributed Data

Tests for Normally Distributed Data

Expected Value with Known Variance

- $X_1, \ldots, X_n$ is a normally distributed sample of $\mathrm{No}(\mu, \sigma^2)$ with known σ^2.
- The hypothesis $H_0 : \mu = \mu_0$ is to be tested on the basis of the sample against the Counterhypothesis $H_1 : \mu \neq \mu_0$ is to be tested.
- As test statistic T we choose the standard score of the sample mean:

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$$

- Assuming H_0, $T \sim \mathrm{No}(0, 1)$.
- H_0 is rejected if T is not in a prediction interval of level $1 - \alpha$ of the standard normal distribution.

Tests for Normally Distributed Data

Two-sided Test

- The hypothesis H_0 is rejected if.

$$|T| = \frac{\sqrt{n}\left|\overline{X} - \mu_0\right|}{\sigma} > z_{1-\alpha/2}$$

- The quality function for a value μ is given by:

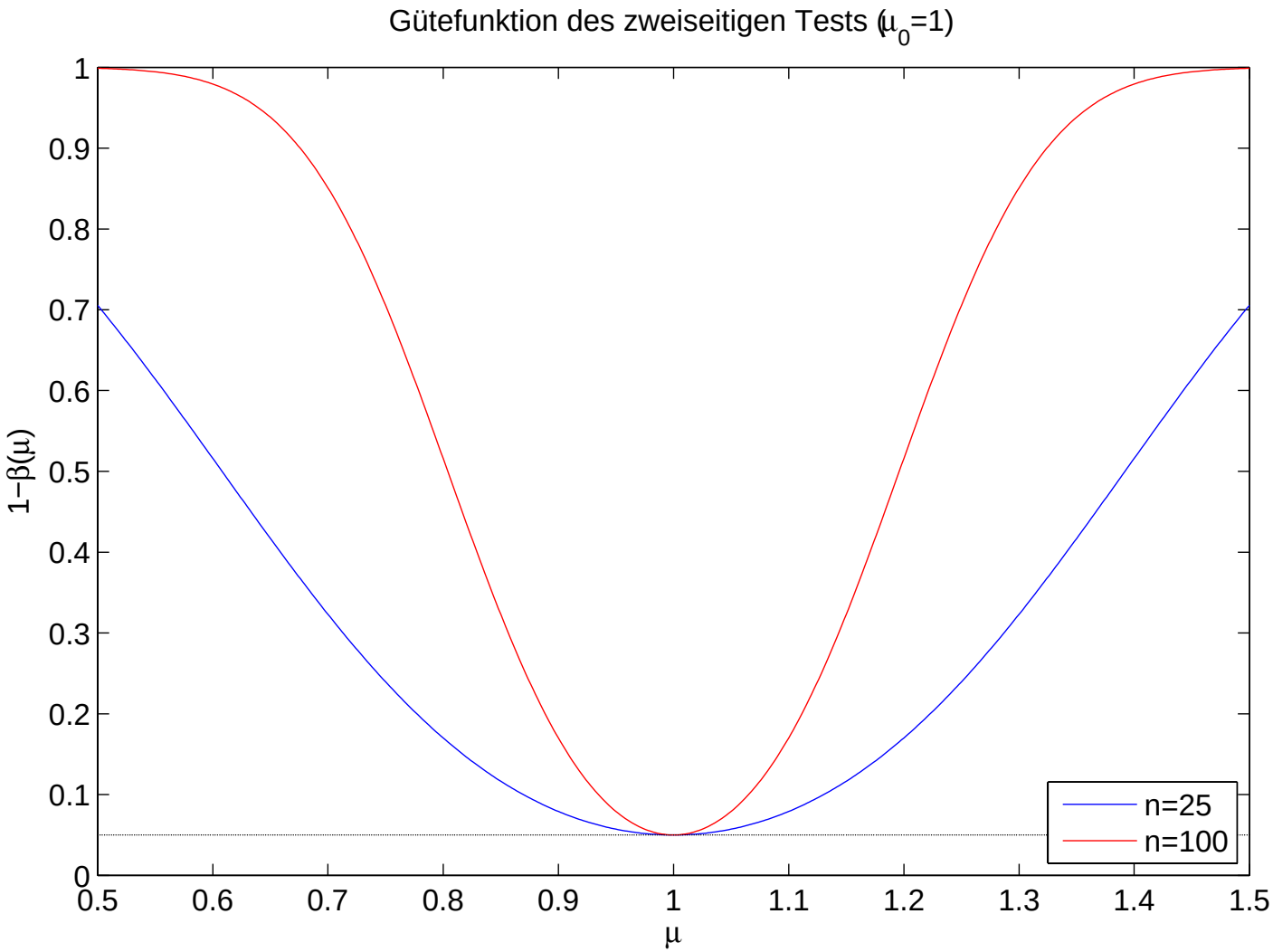
$$1 - \beta(\mu) = W(T \in C) = G(z_{\alpha/2}) + 1 - G(z_{(1-\alpha)/2})$$

where G is the distribution function of $\mathrm{No}(\sqrt{n}(\mu - \mu_0)/\sigma, 1)$-distribution.

- The test is unbiased.

MATLAB: `make_test_normal_mean.m`

Tests for Normally Distributed Data



Tests for Normally Distributed Data

One-sided Test

- The hypothesis $H_0 : \_mu \leq \mu_0$ is to be tested with the test statistic T against the counterhypothesis $H_1 : \mu > \mu_0$ is to be tested.
- H_0 is rejected if T is "'‘too large'"'.
- A rejection region with significance level α is the quantity

$$C = [z_{1-\alpha}, \infty[$$

- Thus, the hypothesis H_0 is rejected if.

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{\sigma} > z_{1-\alpha}$$

Tests for Normally Distributed Data

- The quality function for a value $\mu > \mu_0$ is given by:

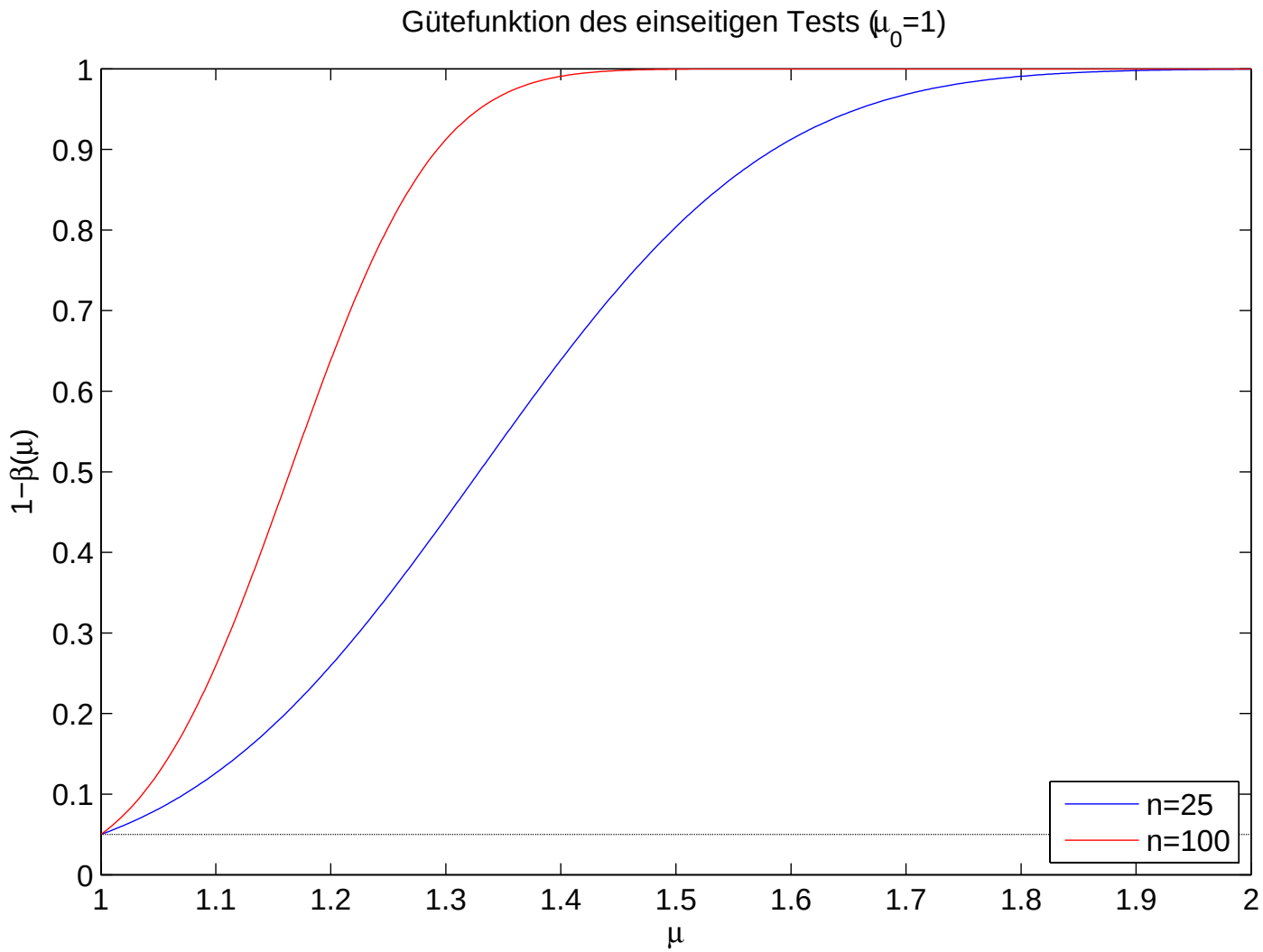
$$1 - \beta(\mu) = W(T \in C) = 1 - G(z_{1-\alpha})$$

where G is the distribution function of the $\mathrm{No}(\sqrt{n}(\mu - \mu_0)/\sigma, 1)$-distribution.

- Analogously, the test with $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$.

MATLAB: `make_test_normal_mean.m`

Tests for Normally Distributed Data



Tests for Normally Distributed Data

Expected Value with Unknown Variance: t-Test

- $X_1, \ldots, X_n$ is a normally distributed sample of $\mathrm{No}(\mu, \sigma^2)$ with unknown σ^2.
- The hypothesis $H_0 : \mu = \mu_0$ is to be tested against the sample against the Counterhypothesis $H_1 : \mu \neq \mu_0$ is to be tested.
- As test statistic T we choose the standard score of the sample mean, using the sample variance S^2:

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$$

- Assuming H_0, $T \sim \mathrm{t}(n-1)$.

Tests for Normally Distributed Data

- H_0 is rejected if T is not in a prediction interval of level $1-\alpha$ of the t distribution with $n-1$ degrees of freedom.
- A rejection region with significance level α is the quantity

$$C =]-\infty, t_{\alpha/2;n-1}] \cup [t_{1-\alpha/2;n-1}, \infty[$$

  where $t_{p;n}$ is the quantile of the t distribution with $n-1$ degrees of freedom at level p.
- Thus, the hypothesis H_0 is rejected if.

$$|T| = \frac{\sqrt{n}\,|\overline{X} - \mu_0|}{S} > t_{1-\alpha/2;n-1}$$

Tests for Normally Distributed Data

- The quality function for a value μ is given by:

$$1 - \beta(\tau) = W(T \in C) = G(z_{\alpha/2}) + 1 - G(z_{(1-\alpha)/2})$$

where G is the distribution function of the noncentral $\mathrm{t}(n-1, \delta)$ distribution with

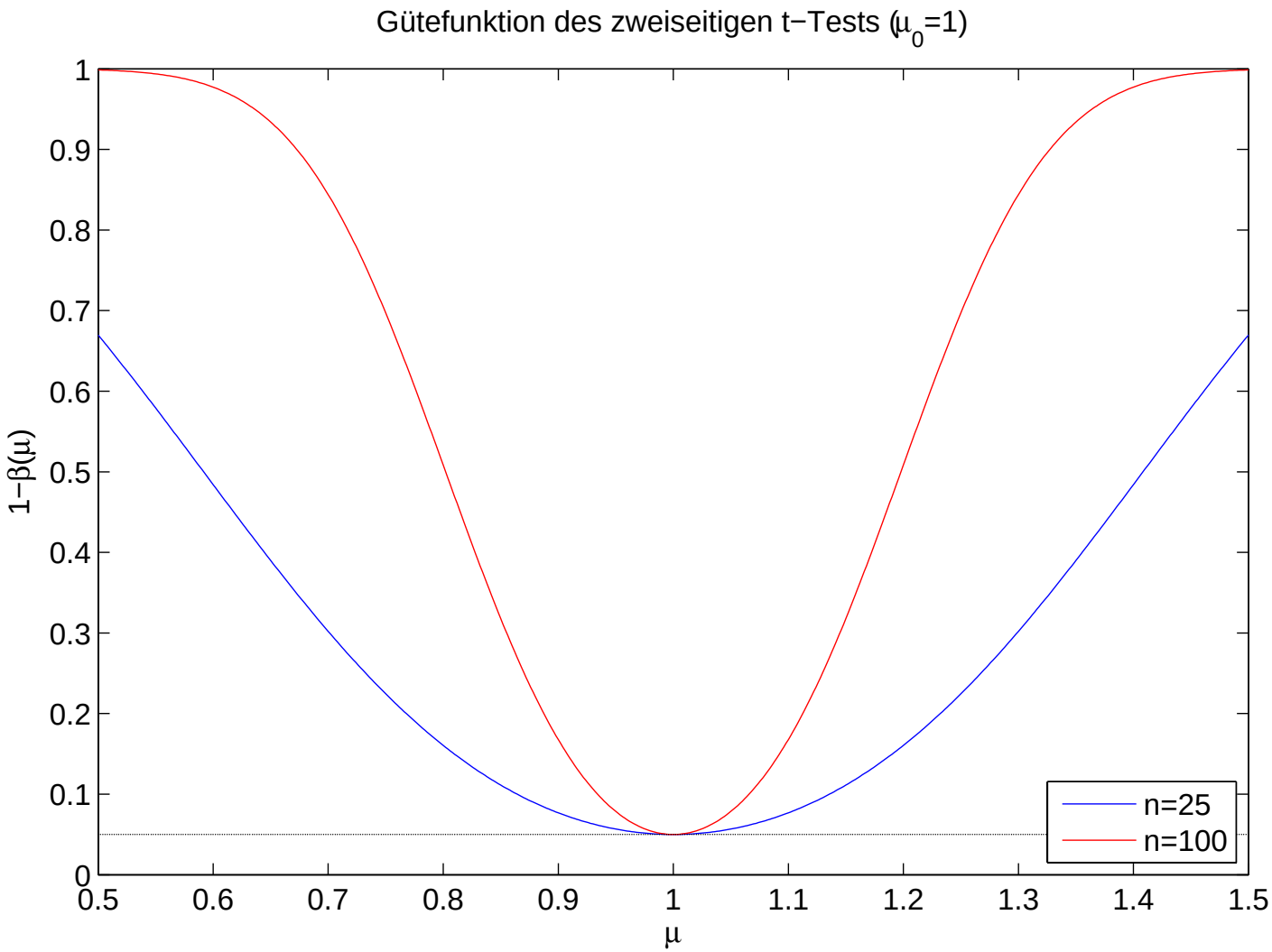
$$\delta = \sqrt{n}(\mu - \mu_0)/\sigma$$

is.

- The test is unbiased.

MATLAB: `make_test_normal_mean.m`

Tests for Normally Distributed Data



Tests for Normally Distributed Data

Equality of Two Expected Values

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two independent normally distributed samples of $\mathrm{No}(\mu_x, \sigma_x^2)$ and. $\mathrm{No}(\mu_y, \sigma_y^2)$.
- The hypothesis $H_0 : \mu_x = \mu_y$ is supposed to be tested against the counter hypothesis $H_1 : \mu_x \neq \mu_y$ to be tested.
- If the **variances are known**, we choose as test statistic T the difference of the sample means:

$$T = \overline{X} - \overline{Y}$$

- Assuming H_0, $T \sim \mathrm{No}(0, \sigma_x^2/n + \sigma_y^2/m)$.

Tests for Normally Distributed Data

- The default score

$$Z = \frac{T}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

  is then standard normally distributed.
- Thus, the hypothesis H_0 is rejected if.

$$|Z| > z_{1-\alpha/2}$$

  or

$$\frac{|\overline{X} - \overline{Y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} > z_{1-\alpha/2}$$

Tests for Normally Distributed Data

- If the **variances are unknown and equal**, the variance can be estimated from the combined ('pooled') sample:

$$S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

- Assuming H_0 is.

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2(1/n + 1/m)}}$$

t-distributed with $n + m - 2$ degrees of freedom.

- Thus, the hypothesis H_0 is rejected if.

$$|T| > t_{1-\alpha/2;n+m-2}$$

where $t_{1-\alpha/2;n+m-2}$ is the quantile of the t distribution with $n + m - 2$ degrees of freedom.

Tests for Normally Distributed Data

t Test for Paired Samples

- Paired samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ occur when the same quantity is measured twice for each observed object, before and after a given intervention.
- The effect of the intervention is described by the differences $W_i = Y_i - X_i, i = 1, \ldots, n$.
- We assume that $W_1, \ldots, W_n$ is normally distributed with mean μ_w and unknown variance σ_w^2.
- The hypothesis $H_0 : \mu_w = 0$ (no effect of the intervention) is to be tested against the sample against the counter-hypothesis $H_1 : \mu_w \neq 0$ to be tested.
- This is done with the t test for single samples.

Tests for Normally Distributed Data

Test of Variance

- $X_1, \ldots, X_n$ is a normally distributed sample with unknown expected value μ and unknown variance σ^2.
- The hypothesis $H_0 : \sigma^2 = \sigma_0^2$ is to be tested on the basis of the sample against the counter hypothesis $H_1 : \sigma^2 \neq \sigma_0^2$ is to be tested against the sample.
- As test statistic T we choose:

$$T = \frac{(n-1)S^2}{\sigma_0^2}$$

- Assuming H_0, T is χ^2-distributed with $n-1$ degrees of freedom.

Tests for Normally Distributed Data

- Thus, the hypothesis H_0 is rejected if.

$$T < \chi^2_{\alpha/2;n-1} \quad \text{oder} \quad T > \chi^2_{1-\alpha/2;n-1}$$

where $\chi^2_{p;k}$ is the p quantile of the χ^2 distribution with k degrees of freedom.

- The quality function for a value σ^2 is given by:

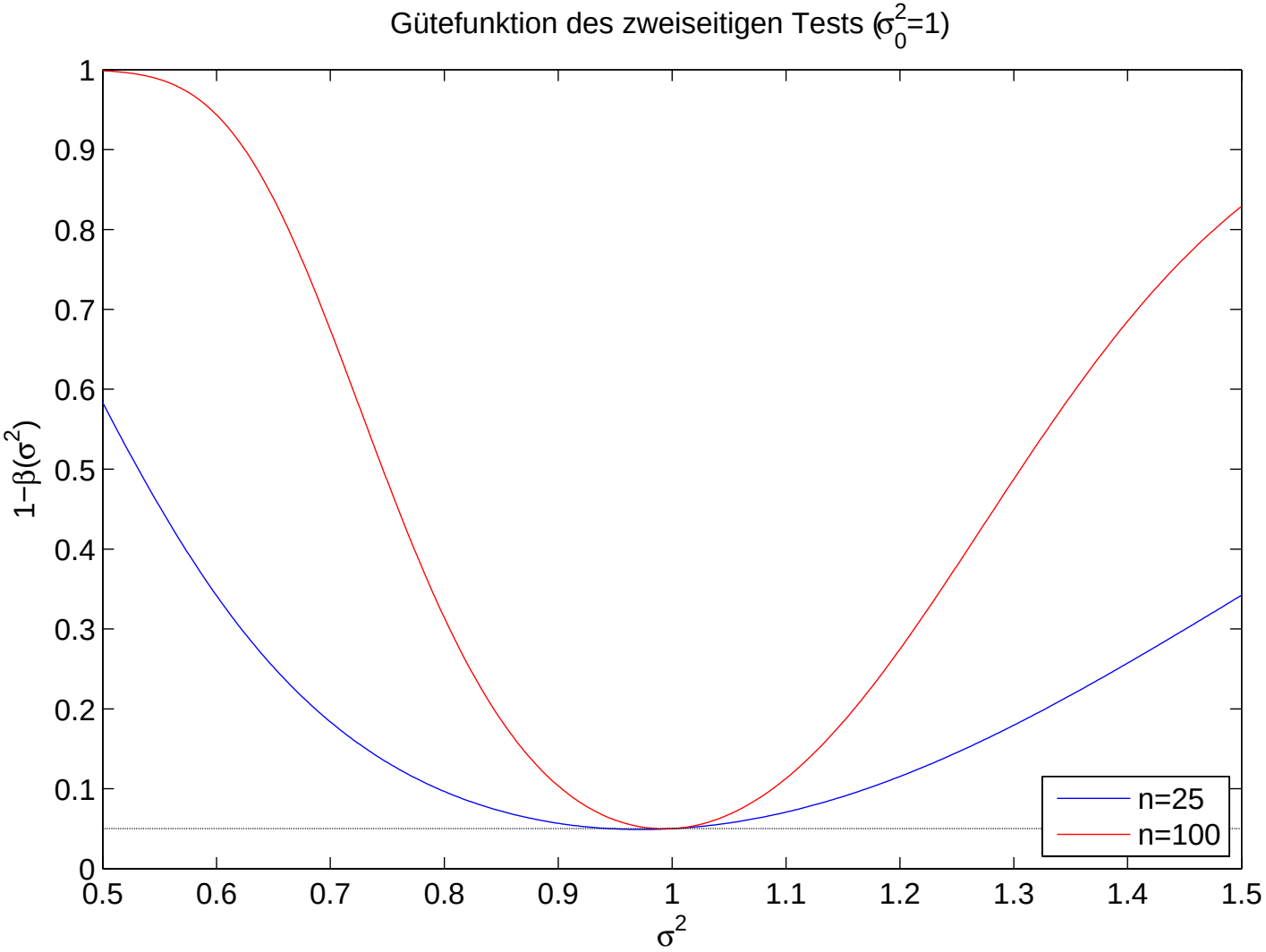
$$1 - \beta(\sigma^2) = G(\sigma_0^2/\sigma^2 \cdot \chi^2_{\alpha/2;n-1}) + 1 - G(\sigma_0^2/\sigma^2 \cdot \chi^2_{1-\alpha/2;n-1})$$

where G is the distribution function of the $\chi^2(n-1)$-distribution.

- The test is not unbiased.

MATLAB: `make_test_normal_variance.m`

Tests for Normally Distributed Data



Subsection: Tests for Poisson Distributed Data

Tests for Poisson Distributed Data

Test for Two-sided Expected Value

- Let $X_1, \ldots, X_n$ be a Poisson distributed sample of $\mathrm{Po}(\lambda)$.
- The hypothesis $H_0 : \lambda = \lambda_0$ is to be tested against the sample against the counter hypothesis $H_1 : \lambda \neq \lambda_0$ is to be tested.
- As test statistic T we choose the sample sum:

$$T = \sum_{i=1}^{n} X_i$$

- T is Poisson distributed according to $\mathrm{Po}(n\lambda)$.
- H_0 is rejected if T is "'‘too small"' or "'‘too large"', i.e. if

$$\sum_{k=0}^{T} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha/2 \text{ or } \sum_{k=T}^{\infty} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha/2$$

Tests for Poisson Distributed Data

- If the distribution function of the Poisson distribution is expressed by the distribution function of the gamma distribution, the following critical region is obtained:

$$n\lambda_0 < \gamma_{\alpha/2;T,1} \quad \text{oder} \quad n\lambda_0 > \gamma_{1-\alpha/2;T+1,1}$$

- H_0 is thus rejected if $n\lambda_0$ is not in the confidence interval based on observation T.

Tests for Poisson Distributed Data

One-sided Test for Expected Value

- The hypothesis $H_0 : \lambda \leq \lambda_0$ is rejected if T is "'too large"' and thus the p value is too small:

$$P(T) = \sum_{k=T}^{\infty} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha \quad \text{oder} \quad n\lambda_0 < \gamma_{\alpha;T,1}$$

- The hypothesis $H_0 : \lambda \geq \lambda_0$ is rejected if T is "'too small"' and thus the p value is too small:

$$P(T) = \sum_{k=0}^{T} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha \quad \text{oder} \quad n\lambda_0 > \gamma_{1-\alpha;T+1,1}$$

Tests for Poisson Distributed Data

Example

A manufacturer aims to produce no more than 25 defective components per day on average in a factory. A sample of 5 days results in 28,34,32,38 and 22 defective components. Has the manufacturer achieved his goal?

The test size is equal to $T = 154$. So it is valid:

$$P(T) = \sum_{k=T}^{\infty} \frac{(125)^k e^{-125}}{k!} = 0.0067 < 0.01$$

$$\gamma_{0.01;154,1} = 126.61 < 154$$

Thus, the hypothesis can be refuted at a significance level of 1 percent.

MATLAB: `make_test_poisson_mean.m`

Tests for Poisson Distributed Data

Approximation by Normal Distribution

- If n is sufficiently large, the distribution of T can be approximated by a normal distribution $\mathrm{No}(n\lambda, n\lambda)$.
- H_0 is rejected if the standard score.

$$Z = \frac{T - n\lambda_0}{\sqrt{n\lambda_0}}$$

is not in a prediction interval of level $1 - \alpha$ of the standard normal distribution.

Example

Given the last example, the approximation yields:

$$Z = 2.5938 > z_{0.99} = 2.3263$$

Thus, the hypothesis can be rejected at a significance level of 1 percent.

Subsection: Tests for Binomially Distributed Data

Tests for Binomially Distributed Data

Bipartite Test for Parameter p

- k is an observation from the binomial distribution $\mathrm{Bi}(n, p)$.
- The hypothesis $H_0 : p = p_0$ is to be tested by the observation against the counter hypothesis $H_1 : p \neq p_0$.
- H_0 is rejected if k assuming H_0 is not in the symmetric prediction interval $[y_1(p_0), y_2(p_0)]$, i.e., is "'too small"' or "'too large"'.

Tests for Binomially Distributed Data

- This is the case if either

$$\sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i} = F_{\mathrm{Be}}(p_0; k, n-k+1) < \alpha/2$$

or

$$\sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} = F_{\mathrm{Be}}(1-p_0; n-k, k+1) < \alpha/2$$

holds where $F_{\mathrm{Be}}(x; a, b)$ is the distribution function of $\mathrm{Be}(a, b)$.

Tests for Binomially Distributed Data

One-sided Test for Parameter p

- The hypothesis $H_0 : p \leq p_0$ is to be tested on the basis of the observation k against the counter hypothesis $H_1 : p > p_0$.
- H_0 is rejected if k is "'‘too large"' and thus the p value is too small:

$$P(k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} = B(p_0; k, n-k+1) < \alpha$$

- The hypothesis $H_0 : p \geq p_0$ is rejected if k is "'‘too small"' and thus the p value is too small:

$$P(k) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i} = B(1-p_0; n-k, k+1) < \alpha$$

Tests for Binomially Distributed Data

Example

A manufacturer claims that no more than 2 percent of a certain component is defective. In a sample of size 300, 9 pieces are defective. Can the manufacturer's claim be refuted?
It holds:

$$P(k) = \sum_{i=9}^{300} \binom{300}{i} 0.02^i 0.98^{300-i} = 0.1507$$

So the manufacturer's claim cannot be refuted at a 5 percent significance level.

MATLAB: `make_test_binomial.m`

Tests for Binomially Distributed Data

Approximation by Normal Distribution

- If n is sufficiently large, the distribution of k can be approximated by a normal distribution $\mathrm{No}(np, np(1-p))$.
- H_0 is rejected if the standard score is

$$Z = \frac{k - np_0}{\sqrt{np(1-p_0)}}$$

  does not lie within a prediction interval of level $1-\alpha$ of the standard normal distribution.
- Two-sided test: H_0 is rejected if.

$$Z < z_{\alpha/2} \text{ oder } Z > z_{1-\alpha/2}$$

- One-sided test: H_0 is rejected if.

$$Z < z_{\alpha} \text{ or } Z > z_{1-\alpha}$$

Tests for Binomially Distributed Data

Example

Given the last example, the approximation yields:

$$Z = 1.2372 < z_{0.95} = 1.6449$$

Thus, the hypothesis cannot be rejected.

MATLAB: `make_test_binomial.m`

Subsection: Likelihood-Ratio Tests

Likelihood-Ratio Tests

- A hypothesis is called **simple** if it has no free parameters, i.e. if the statistical model is completely given.
- If two **simple** hypotheses $H_0 : \theta = \theta_0$, and $H_1 : \theta = \theta_1$, a test can be constructed from the ratio of the likelihoods. likelihoods can be constructed.

$$\Lambda(x) := \frac{L(\theta_0|x)}{L(\theta_1|x)}$$

- The null hypothesis H_0 is rejected if the test statistic is below of a certain value $\Lambda(X) \leq \eta$ falls.
- Here η is chosen such that $\alpha = P(\Lambda(X) \leq \eta|H_0)$. As usual, the significance level (,,test size") α is again freely chosen. Common values are $\alpha = 0.05, 0.1$ or 0.01.

Likelihood-Ratio Tests

Example (Likelihood Ratio Test)

Let $(X_1, \ldots, X_n)$ be a sample of a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with fixed μ. We test for the *variance* of the distribution, with $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 = \sigma_1^2, \sigma_1^2 > \sigma_0^2$.

- The likelihood ratio is

$$\Lambda(X) = \frac{L(\sigma_0^2|X)}{L(\sigma_1^2|X)} = \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{-\frac{n}{2}} \exp\left\{-\frac{\sigma_0^{-2} - \sigma_1^{-2}}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

- $\sum_{i=1}^{n}(x_i - \mu)^2$ is the only term of the likelihood ratio that depends on the data, therefore H_0 is rejected when the sum is extreme enough under the null hypothesis. Since we required that $\sigma_1^2 > \sigma_0^2$, we can restrict ourselves to test to test whether the sum is too large (one-sided test).

Likelihood-Ratio Tests

Neyman-Pearson lemma

The likelihood-ratio test is under certain conditions a test with maximum goodness at a given significance level α.

- A hypothesis that is not *simple* is called **composite**.
- Often the counterhypothesis H_1 is constructed as the complement of the Null hypothesis constructed, i.e. $H_0 : \theta \in \theta_0, H_1 : \theta \in \Theta_0^C$.
- In this case we speak of a **nested model**. In this case the Neyman-Pearson lemma holds for:

$$\Lambda(x) = \frac{\sup\{L(\theta|x) : \theta \in \Theta_0\}}{\sup\{L(\theta|x) : \theta \in \Theta\}}$$

Likelihood-Ratio Tests

In addition, for nested models, the:

Wilks' Theorem

For nested models, the test statistic $T = -2 \ln \Lambda(X)$ converges under H_0 asymptotically converges to a $\chi^2(n)$ distribution with n degrees of freedom, where n is given as the difference of the dimensionalities of Θ and Θ_0: $n = \dim(\Theta) - \dim(\Theta_0)$.

- Wilks' theorem can be used to specify η: $\eta = \chi^2_{\alpha;n}$, where α is again the significance level of the test and n is the number of degrees of freedom of the test.

Likelihood-Ratio Tests

Example (Likelihood Ratio Test for P_λ)

We construct a test for the intensity parameter λ of a Poisson distribution, $H_0 : \lambda = \lambda_0, H_1 : \lambda \neq \lambda_0$. Let $(X_1, \ldots, X_n)$ be a sample of a Poisson distributed random variable.

- The supremum of the counter hypothesis H_1 (the ,,Maximum Likelihood") is reached at $\lambda = \overline{X}$.
- From this follows for the likelihood ratio $\Lambda(X)$:

$$\Lambda(X) = e^{n(\overline{X}-\lambda_0)} \prod_{i=1}^{n} \left(\frac{\lambda_0}{\overline{X}} \right)^{X_i}$$

- The test statistic $T = -2 \ln \Lambda(X)$:

$$T = -2n \left[\overline{X} - \lambda_0 + \overline{X} \ln \left(\frac{\lambda_0}{\overline{X}} \right) \right]$$

Likelihood-Ratio Tests

Example (likelihood ratio test for P_λ, continued)

- The test statistic $T = -2 \ln \Lambda(X)$:

$$T = 2n \left[\lambda_0 - \overline{X} + \overline{X} \ln \left(\frac{\overline{X}}{\lambda_0} \right) \right]$$

- According to Wilks' theorem T is asymptotically χ^2-distributed with one degree of freedom. The 95% quantile of $\chi^2(1)$ is 3.84. The critical region of the test at a significance level of 5% is therefore

$$C = [3.84, \infty)$$

.

Section 27: Non-parametric Tests

Non-parametric Tests

- A test that checks the hypothesis whether the data can come from a certain distribution is called a **fitting test**.
- The distribution may be completely determined or determined except for unknown parameters.
- A goodness-of-fit test may precede a parametric test to check its applicability.

Subsection: The Chisquare Test

The Chisquare Test

The Chisquare Test for Discrete Data

- The sample $X_1, \ldots, X_n$ comes from a discrete distribution with range of values $\{1, \ldots, k\}$.
- We test the hypothesis H_0 that the density f has values $f(j) = p_j, j = 1, \ldots, k$:

$$H_0 : W(X_i = j) = p_j, j = 1, \ldots, k$$

vs.

$$H_1 : W(X_i = j) \neq p_j, \text{for one } j$$

- Let Y_j be the number of observations equal to j.
- Under the null hypothesis, $Y_1, \ldots, Y_k$ is multinomially distributed according to $\mathrm{Mu}(n, p_1, \ldots, p_k)$ and $\mathrm{E}[Y_j] = np_j$.

The Chisquare Test

Excursion: the Multinomial Distribution $\mathrm{Mu}(n, p_1, \ldots, p_d)$

- The alternative experiment can be generalized to allow not only two, but d elementary events $e_1, \ldots, e_d$, to which are assigned the probabilities $p_1, \ldots, p_d$, which are only

$$\sum_{i=1}^{d} p_i = 1$$

  must satisfy.
- Performing the **generalized alternative experiment** n times, the elementary events are the sequences of the form:

$$(e_{i_1}, \ldots, e_{i_n}),\ 1 \leq i_j \leq d$$

The Chisquare Test

- If the n partial trials are independent, then:

$$W(e_{i_1}, \ldots, e_{i_n}) = \prod_{j=1}^{n} W(e_{i_j}) = \prod_{j=1}^{n} p_{i_j} = \prod_{i=1}^{d} p_i^{n_i}$$

  Where n_i is the number of occurrences of e_i. Therefore, the sum of n_i is n.

- The d-dimensional random variable $\boldsymbol{X} = (X_1, \ldots, X_d)$ maps the sequence $(e_{i_1}, \ldots, e_{i_n})$ to the vector $(n_1, \ldots, n_d)$. In this case $n!/(n_1! \cdots n_d!)$ sequences are mapped onto the same vector.
- Therefore, the density of $\boldsymbol{X}$ is:

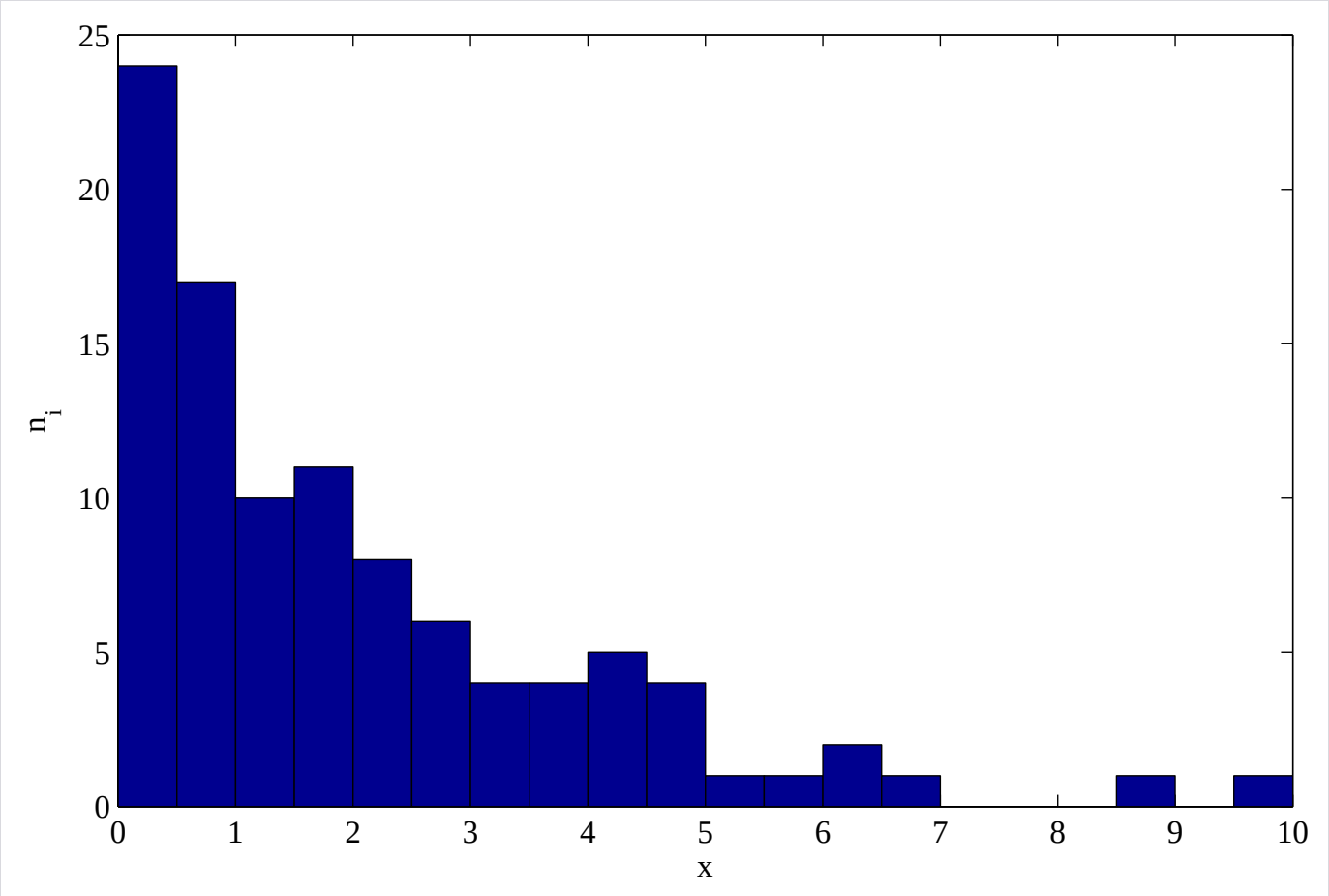
$$f_{\boldsymbol{X}}(n_1, \ldots, n_d) = \frac{n!}{n_1! \ldots n_d!} \prod_{i=1}^{d} p_i^{n_i}, \ \sum_{i=1}^{d} n_i = n, \ \sum_{i=1}^{d} p_i = 1$$

The Chisquare Test

- The distribution of $\boldsymbol{X}$ is called **multinomial distribution** with parameters n and $p_1, \ldots, p_d$: $W_{\boldsymbol{X}} = \mathrm{Mu}(n, p_1, \ldots, p_d)$
- The classic example of a multinomial distributed random vector is the histogram (grouped frequency distribution), which is used to graphically represent the (absolute) **experimental frequency**.
- X_i is the number of times the random variable R, the experimental outcome, falls into group i.
- Let the probability that R falls into group i be equal to p_i.
- If n outcomes are filled into the histogram, the group contents $(X_1, \ldots, X_d)$ are multinomially distributed according to $\mathrm{Mu}(n, p_1, \ldots, p_d)$.

The Chisquare Test

- A histogram



The Chisquare Test

The moments of the $\mathrm{Mu}(n, p_1, \ldots, p_d)$ distribution

Let $\boldsymbol{X} = (X_1, \ldots, X_d) \sim \mathrm{Mu}(n, p_1, \ldots, p_d)$. Then holds:

- $\mathsf{E}[X_i] = np_i$
- $\mathsf{var}[X_i] = np_i(1 - p_i)$
- $\mathsf{cov}[X_i, X_j] = -np_ip_j$

Definition (covariance)

The **covariance** of two random variables X and Y is defined by:

$$\mathsf{cov}[X, Y] = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]$$

If $\mathsf{cov}[X, Y] = 0$, X and Y are called **uncorrelated**.

The Chisquare Test

Properties of Covariance

- $\mathrm{cov}[aX + b, cY + d] = ac \cdot \mathrm{cov}[X, Y]$
- $\mathrm{var}[X + Y] = \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}[X, Y]$
- $\mathrm{cov}[X, X] = \mathrm{var}[X]$

Definition (correlation)

The **correlation coefficient** $\rho[X, Y]$ is defined by:

$$\rho[X, Y] = \frac{\mathrm{cov}[X, Y]}{\sqrt{\mathrm{var}[X]\mathrm{var}[Y]}}$$

The Chisquare Test

properties of correlation coefficient

- $-1 \leq \rho[X, Y] \leq 1$
- $\rho[aX + b, cY + d] = \rho[X, Y]$
- $\rho[X, X] = 1$

The Chisquare Test

Definition (covariance matrix)

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector. If all variances and covariances exist, they are combined in the **covariance matrix** $\mathbf{C} = \mathrm{Cov}[\boldsymbol{X}]$:

$$\mathbf{C}_{ij} = \mathrm{cov}[X_i, X_j]$$

Similarly, the correlations are summarized in the **correlation matrix** $\mathbf{R} = \mathrm{Cor}[X, Y]$:

$$\mathbf{R}_{ij} = \rho[X_i, X_j]$$

- The covariance matrix is always **symmetric** and **positive definite**. All eigenvalues are real and positive.
- The correlation matrix is also **symmetric** and **positive definite**. All diagonal elements are equal to 1.

The Chisquare Test

- The test variable compares the observed frequencies Y_j with their expected values:

$$T = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}$$

- The null hypothesis is rejected if T is large.
- The critical region can be determined according to the following result.

Theorem

Assuming the null hypothesis, the random variable T is asymptotic, i.e. for $n \to \infty$, χ^2-distributed with $k-1$ degrees of freedom.

The Chisquare Test

- If the test is to have significance level α, H_0 is rejected if.

$$T \geq \chi^2_{1-\alpha;k-1}$$

where $\chi^2_{1-\alpha;k}$ is the $(1-\alpha)$ quantile of the χ^2 distribution with $k-1$ degrees of freedom.

- The reason that T has only $k-1$ degrees of freedom is the linear relationship between the Y_j:

$$\sum_{j=1}^{k} Y_j = n$$

- As a rule of thumb, n should be large enough that $np_j > 5, j = 1, \ldots, k$.
- If this is not satisfied, the rejection range should be determined by simulation.

The Chisquare Test

Example

We use a sample of size 50 to test whether a die is symmetric, that is, whether the number of eyes X has the following distribution:

$$W(X = 1) = \ldots = W(X = 6) = \frac{1}{6}$$

A simulation of $N = 100000$ samples yields:

$$\overline{T} = 5.000, \quad S_T^2 = 9.789$$

The 0.95 quantile of the χ^2 distribution with five degrees of freedom is $\chi^2_{0.95;5} = 11.07$, and

$$W(T \geq 11.07) = 0.048$$

MATLAB: `make_chi2test_cube.m`

The Chisquare Test

The Chisquare Test for Continuous Data

- The sample $X_1, \ldots, X_n$ comes from a continuous distribution F.
- We test the hypothesis $H_0 : F(x) = F_0(x)$.
- To do this, we divide the range of values of X into k groups $G_1, \ldots, G_k$.
- Let Y_j be the number of observations in group G_j.
- Under the null hypothesis, $Y_1, \ldots, Y_k$ is multinomially distributed according to $\mathrm{Mu}(n, p_1, \ldots, p_k)$ and $\mathrm{E}[Y_j] = np_j$, with.

$$p_j = W(X \in G_j | H_0)$$

- The test continues as in the discrete case.

The Chisquare Test

Unknown Parameters

- The null hypothesis need not be fully specified. We consider the case where the p_j still depend on unknown parameters $\boldsymbol{\vartheta}$:

$$W(X \in G_j) = p_j(\boldsymbol{\vartheta})$$

- The statistic T is now a function of the unknown parameters:

$$T(\boldsymbol{\vartheta}) = \sum_{j=1}^{k} \frac{(Y_j - np_j(\boldsymbol{\vartheta}))^2}{np_j(\boldsymbol{\vartheta})}$$

- First, the parameters are estimated, by ML estimation or minimization of T:

$$\tilde{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}} T(\boldsymbol{\vartheta})$$

The Chisquare Test

- The critical area can be determined according to the following result.

Theorem

If m parameters are estimated from the sample, $T(\tilde{\boldsymbol{\vartheta}})$ is asymptotically χ^2-distributed with $k-1-m$ degrees of freedom.

- If the test is to have significance level α, H_0 is rejected if.

$$T \geq \chi^2_{1-\alpha;k-1-m}$$

where $\chi^2_{1-\alpha;k-1-m}$ is the $(1-\alpha)$ quantile of the χ^2 distribution with $k-1-m$ degrees of freedom.

The Chisquare Test

Example

Angabe: The number of occupational accidents was surveyed in a large company over a period of 30 weeks. The following values resulted:

$$\boldsymbol{X} = \{8, 0, 0, 1, 3, 4, 0, 2, 12, 5, 1, 8, 0, 2, 0, \\ 1, 9, 3, 4, 5, 3, 3, 4, 7, 4, 0, 1, 2, 1, 2\}$$

To test the hypothesis that the observations are Poisson distributed according to $\mathrm{Po}(\lambda)$.

Lösung: The observations are divided into five groups:

group	1	2	3	4	5
X	0	1	2–3	4–5	> 5

The frequencies of the groups are:

$$Y_1 = 6, Y_2 = 5, Y_3 = 8, Y_4 = 6, Y_5 = 5$$

The Chisquare Test

Example (Continuation)

The estimated value for λ is the sample mean:

$$\tilde{\lambda} = 3.1667$$

The expected values of Y_j assuming $H_0 = \mathrm{Po}(\tilde{\lambda})$ are:

j	1	2	3	4	5
$\mathsf{E}[Y_1]$	1.2643	4.0037	13.0304	8.6522	3.0493

The test size T is equal to $T = 21.99$. The 99% quantile of the χ^2 distribution with three degrees of freedom is equal to $\chi^2_{0.99;3} = 11.35$. Thus, the hypothesis that the observations are Poisson distributed must be rejected.

MATLAB: `make_chi2test_poisson.m`

Subsection: The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov Test

One Sample

- The sample $X_1, \ldots, X_n$ is from the continuous distribution with distribution function F.
- We test the hypothesis $H_0 : F(x) = F_0(x)$.
- The test statistic D_n is the maximum absolute deviation of the empirical distribution function $F_n(x)$ of the sample from the hypothesized distribution function $F_0(x)$:

$$D_n = \max_x |F_n(x) - F_0(x)|$$

- For samples from F_0, the distribution function of $\sqrt{n}D$ for $n \to \infty$ tends to:

$$K(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}$$

The Kolmogorov-Smirnov Test

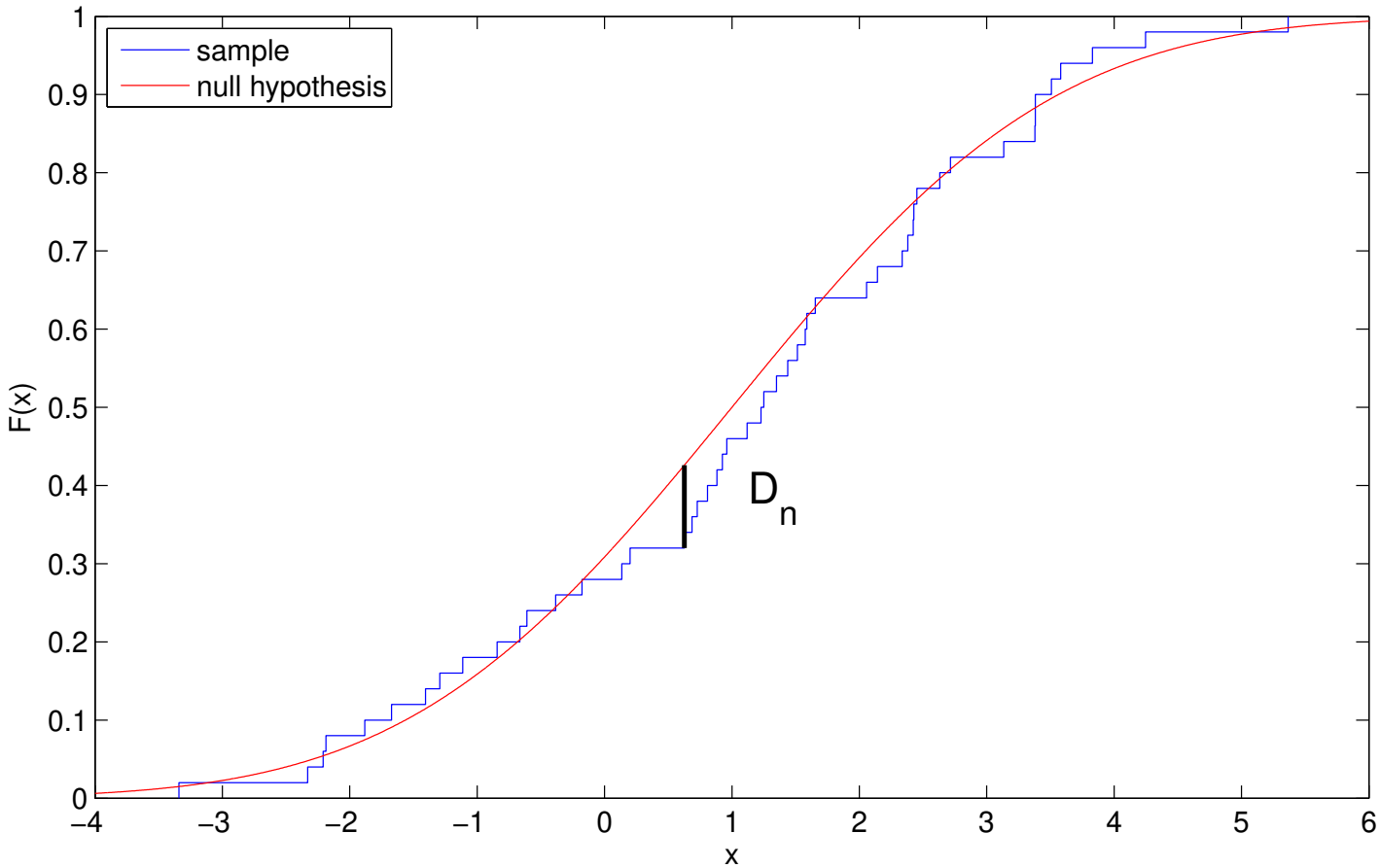
- Quantiles $K_{1-\alpha}$ can be calculated from the asymptotic distribution function.
- The null hypothesis is rejected if

$$\sqrt{n} D_n > K_{1-\alpha}$$

- If parameters of F_0 are estimated before the test, the quantiles are no longer valid.
- In this case, the rejection range must be determined by simulation.

MATLAB: `kstest`

The Kolmogorov-Smirnov Test



D_n, the test statistic of the Kolmogorov-Smirnov test.

The Kolmogorov-Smirnov Test

Two Samples

- We test whether two samples of size n or m are from the same distribution F.
- The test size is the maximum absolute difference of the empirical distribution functions:

$$D_{n,m} = \max_x \left| F_n^1(x) - F_m^2(x) \right|$$

- The null hypothesis is rejected if.

$$\sqrt{\frac{nm}{n+m}} D_{n,m} > K_{1-\alpha}$$

MATLAB: `kstest2`

Part VII

Regression and Linear Models

Overview Part 7

Section 28: Introduction

Introduction

- Regression analysis examines the dependence of observations on various variables.
- **influence variable** (independent variable) $\boldsymbol{x} = (x_1, \ldots, x_r)$.
- **outcome variable** (dependent variable) Y.
- **regression model**:

$$Y = f(\boldsymbol{\beta}, \boldsymbol{x}) + \varepsilon$$

with **regression coefficients** $\boldsymbol{\beta}$ and **error term** ε.

- The objective is to **estimate** $\boldsymbol{\beta}$ using observations $Y_1, \ldots, Y_n,\ n \geq r$.
- One influence variable: single regression; Multiple influence variables: multiple (multiple) regression.

Introduction

- Each observation Y_i has an error term ε_i.
- The error terms need not all have the same distribution, nor need they be independent.
- It is often assumed that the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ has a **multivariate normal distribution**.
- However, other distributions of $\boldsymbol{\varepsilon}$ are also conceivable.

Section 29: Multidimensional Random Variable

Subsection: Basic Terminology

Basic Terminology

Definition (Multidimensional Random Variable)

A mapping $\boldsymbol{X}$:

$$\omega \in \Omega \mapsto \boldsymbol{x} = \boldsymbol{X}(\omega) \in \mathbb{R}^d$$

that assigns a real vector $\boldsymbol{x} \in \mathbb{R}^d$ to each element ω of the sample space Ω is called a d-dimensional random variable.

- Each component of a d-dimensional random variable is itself a random variable.
- Each component can be discrete or continuous.

Basic Terminology

Definition (distribution function)

If $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a d-dimensional random variable, then its distribution function $F_{\boldsymbol{X}}$ is given by

$$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = W(X_1 \leq x_1 \cap \ldots \cap X_d \leq x_d)$$

defined.

Definition (density function)

If $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a d-dimensional discrete random variable, then its density function $f_{\boldsymbol{X}}$ is given by

$$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = W(X_1 = x_1 \cap \ldots \cap X_d = x_d)$$

defined.

Basic Terminology

Example

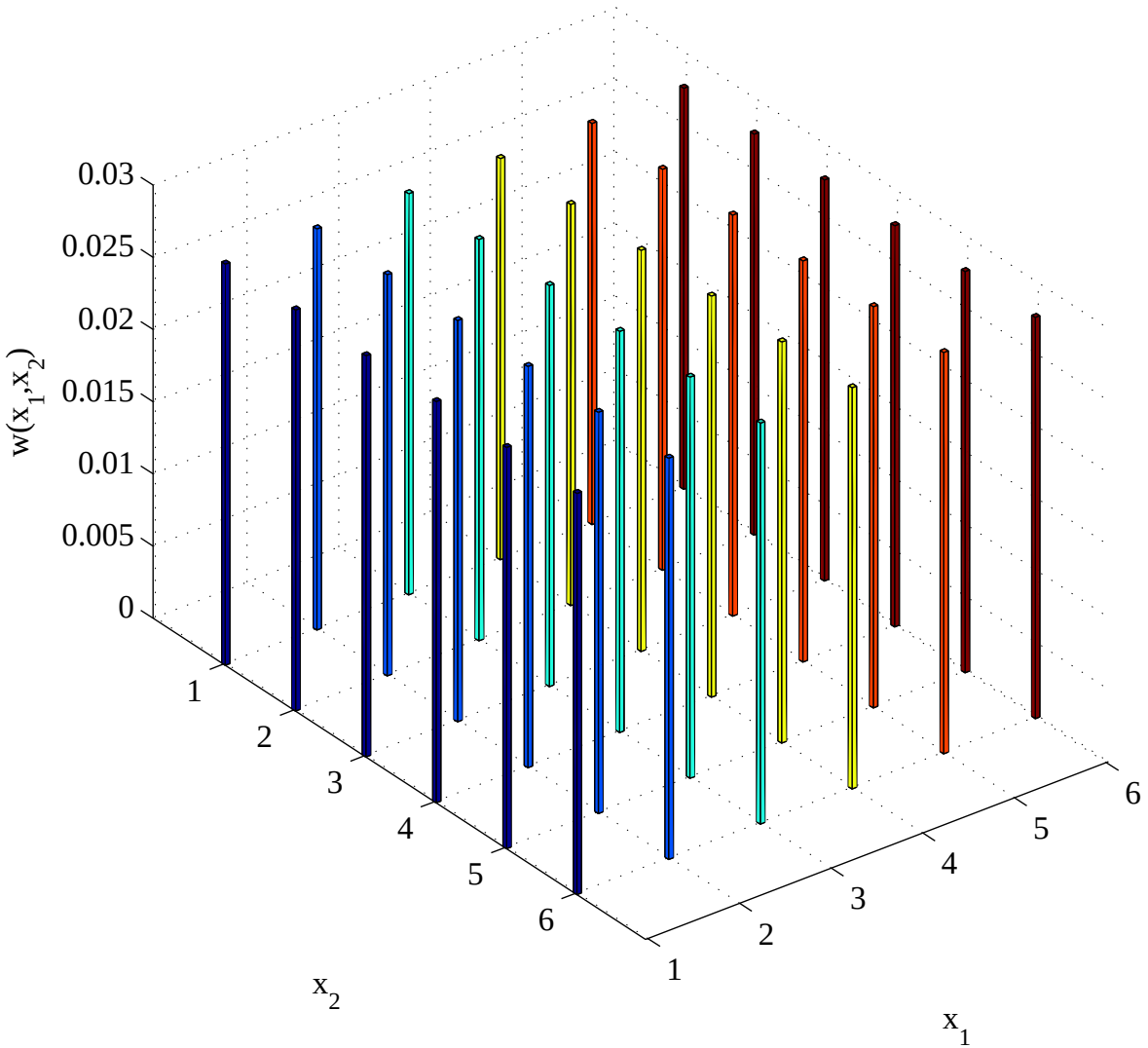
The bivariate (two-dimensional) random variable $\boldsymbol{X} = (X_1, X_2)$ assigns the numbers of eyes (i, j) to the outcome (e_i, e_j) of the roll with two dice. If all outcomes are equally probable, $W_{\boldsymbol{X}}$ is given by:

$$W_{\boldsymbol{X}}\left(\{(i,j)\}\right) = \frac{1}{36}$$

The density $f_{\boldsymbol{X}}$ is:

$$f_{\boldsymbol{X}}(x_1, x_2) = . \begin{cases} \frac{1}{36}, & x_1 \in \{1, \ldots, 6\} \cap x_2 \in \{1, \ldots, 6\} \\ 0, & \text{other} \end{cases}$$

Basic Terminology



Example (Continuation)

Therefore, the distribution function $F_{\boldsymbol{X}}$ is:

$$F_{\boldsymbol{X}}(x_1, x_2) = W(X_1 \leq x_1 \cap X_2 \leq x_2) = \sum_{i \leq x_1 \cap j \leq x_2} f(i,j)$$

For example, $F_{\boldsymbol{X}}(3,4) = \sum_{i \leq 3 \cap j \leq 4} \frac{1}{36} = \frac{12}{36} = \frac{1}{3}$.

- Because of the countability of the elementary events, they can also be uniquely mapped into $\mathbb{R}$ by a univariate random variable Y, e.g.:

  $$Y : (e_i, e_j) \longrightarrow 6i + j - 6$$

  The set of values of Y are the natural numbers between 1 and 36, and W_Y is given by:

  $$W_Y\big(\{k\}\big) = \frac{1}{36}, \ 1 \leq k \leq 36$$

Basic Terminology

Definition (density function)

If $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a d-dimensional continuous random variable, then its density function $f_{\boldsymbol{X}}$ is given by

$$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = \frac{\partial^d F_{\boldsymbol{X}}}{\partial x_1 \ldots \partial x_d}$$

defined.

Definition (moments)

- Expectation: $\mathrm{E}[\boldsymbol{X}] = (\mathrm{E}[X_1] \cdots \mathrm{E}[X_d])$
- Covariance matrix: $\mathrm{Cov}[\boldsymbol{X}]_{ij} = \mathrm{cov}[X_i, X_j]$

Subsection: Marginal Distributions and Conditional Distributions

Marginal Distributions and Conditional Distributions

- If X_1 and X_2 are two (discrete or continuous) univariate random variables, then $\boldsymbol{X} = (X_1, X_2)$ is a bivariate random variable. The distribution (distribution function, density) of $\boldsymbol{X}$ is also called the **common distribution** (distribution function, density) of X_1 and X_2.
- The following problem now arises: can we calculate the distribution of X_1 or X_2 from the joint distribution?

Marginal Distributions and Conditional Distributions

- Let F be the distribution function and f be the density of continuous random variables $\boldsymbol{X} = (X_1, X_2)$. Then the distribution function F_1 of X_1 is given by:

$$F_1(x_1) = W(X_1 \leq x_1) = W(X_1 \leq x_1 \cap -\infty < X_2 < \infty) = \\ = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2\, \mathrm{d}x_1$$

- It follows:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2$$

is the density of X_1.

Marginal Distributions and Conditional Distributions

Definition (marginal distribution)

Let $\boldsymbol{X} = (X_1, X_2)$ be a bivariate continuous random variable with distribution function F and density f. The distribution of X_1 is called the **marginal distribution** of X_1 with respect to $\boldsymbol{X}$. Its density f_1 is:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_2.$$

Analogously, if $\boldsymbol{X} = (X_1, X_2)$ is discrete with density f, then the density f_1 of the marginal distribution of X_1 with respect to $\boldsymbol{X}$ is given by:

$$f_1(k_1) = \sum_{k_2} f(k_1, k_2)$$

Marginal Distributions and Conditional Distributions

- So the distributions of X_1 and X_2 can be calculated from the joint distribution of X_1 and X_2.
- The reverse process is generally not possible because the joint distribution also contains information about correlations (coupling) between X_1 and X_2.
- Let X_1 and X_2 be two discrete random variables with joint density $f(k_1, k_2)$ and marginal densities $f_1(k_1)$ and $f_2(k_2)$. Then the conditional probability of the event $X_1 = k_1$ under the condition $X_2 = k_2$ is given by:

$$W(X_1 = k_1 | X_2 = k_2) = \frac{W(X_1 = k_1 \cap X_2 = k_2)}{W(X_2 = k_2)} = \frac{f(k_1, k_2)}{f_2(k_2)}$$

Marginal Distributions and Conditional Distributions

Definition (Conditional Density)

Let $\boldsymbol{X} = (X_1, X_2)$ be a bivariate discrete random variable with density $f(k_1, k_2)$ and marginal distribution densities $f_1(k_1)$ and $f_2(k_2)$, respectively. The function $f(k_1|k_2)$, defined by:

$$f(k_1|k_2) = \frac{f(k_1, k_2)}{f_2(k_2)}$$

is called the density

of X_1 **conditioned by** X_2**.**

- For fixed k_2, the conditional density is the density of a distribution, the **conditional distribution** of X_1 given by $X_2 = k_2$.

Marginal Distributions and Conditional Distributions

- If $\boldsymbol{X} = (X_1, X_2)$ is continuous, then analogously $f(x_1|x_2)$ is defined by:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \; (f_2(x_2) \neq 0)$$

- $f(x_1|x_2)$ is for fixed x_2 the density of a distribution, the distribution of X_1 conditioned by $X_2 = x_2$.
- That $f(x_1|x_2)$ is indeed a density can be easily verified:

$$\int_{-\infty}^{\infty} f(x_1|x_2)\,\mathrm{d}x_1 = \int_{-\infty}^{\infty} \frac{f(x_1, x_2)}{f_2(x_2)}\,\mathrm{d}x_1 = \frac{f_2(x_2)}{f_2(x_2)} = 1$$

and analogously for discrete $\boldsymbol{X}$.

Marginal Distributions and Conditional Distributions

Conditional Density Properties

It is true:

$$f(x_1, x_2) = f(x_1|x_2) \cdot f_2(x_2)$$

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1|x_2) \cdot f_2(x_2)\,\mathrm{d}x_2$$

and analogously for discrete densities.

Definition (independence of random variables)

If the (unconditional) density of the marginal distribution of X_1 is equal to the density conditioned by X_2, then X_1 and X_2 are called **independent**.

$$X_1 \text{ and } X_2 \text{ independent} \iff f(x_1|x_2) = f_1(x_1).$$

Marginal Distributions and Conditional Distributions

- For independent random variables X_1 and X_2 holds:

$$f(x_1|x_2) = f_1(x_1) \iff f(x_2|x_1) = f_2(x_2)$$
$$\iff f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

and analogously for discrete $\boldsymbol{X}$.

Properties of Independent Random Variables

- $f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$
- $\mathsf{E}[X_1 X_2] = \mathsf{E}[X_1] \cdot \mathsf{E}[X_2]$
- $\mathsf{cov}[X_1, X_2] = \rho[X_1, X_2] = 0$

- Uncorrelated random variable are not necessarily also independent!

Marginal Distributions and Conditional Distributions

- If $\boldsymbol{X} = (X_1, \ldots, X_d),\ d > 2$, then the definitions of the marginal distribution, conditional densities and independence must be generalized accordingly.
- The density $f_{i_1,\ldots,i_m}$ of the marginal distribution of $X_{i_1}, \ldots, X_{i_m}$ is obtained by integrating or summing over all other variables.
- The density of X_i conditioned by X_j is given by:

$$f(x_i|x_j) = \frac{f_{i,j}(x_i, x_j)}{f_j(x_j)}$$

where $f_{i,j}(x_i, x_j)$ is the marginal distribution density of X_i, X_j.

- $X_{i_1}, \ldots, X_{i_k}$ are called independent if the density of their marginal distribution is the product of the densities of the marginal distributions of each X_{i_j}.

Marginal Distributions and Conditional Distributions

Example (The Acceptance or Detection Probability)

Let X be a random variable with density $f(x)$. If X takes the value x, then there is a probability $a(x)$ that x is actually observed. One now defines a random variable I that is 1 if x is observed and 0 otherwise. Then I is Bernoulli distributed under the condition $X = x$ according to $\mathrm{Al}(a(x))$:

$$W(I = 1|X = x) = a(x)$$
$$W(I = 0|X = x) = 1 - a(x)$$

Therefore, the joint density of X and I is:

$$f(x, 1) = a(x)f(x)$$
$$f(x, 0) = [1 - a(x)]f(x)$$

Example (Continuation)

Because the experimenter can only work with observed quantities, he restricts his population to the detected events, i.e. he needs the density of X under the condition that X is observed:

$$f_A(x) = f(x|I=1) = \frac{f(x,1)}{f_2(1)} = \frac{a(x)f(x)}{\int a(x)f(x)\,\mathrm{d}x}$$

As a concrete example, consider the measurement of a lifetime. Let the measurement start at $t_{\min}$ and end at $t_{\max}$. Then $a(t)$ has the following shape:

$$a(t) = \begin{cases} 0, & \text{for } t \leq t_{\min} \\ 1, & \text{for } t_{\min} < t \leq t_{\max} \\ 0, & \text{for } t > t_{\max} \end{cases}$$

Marginal Distributions and Conditional Distributions

Example (Continuation)

For the measured probability density:

$$f_A(t) = \begin{cases} 0, \ t \leq t_{\min} \\ \dfrac{\frac{1}{\tau}\exp(-t/\tau)}{\exp(-t_{\min}/\tau) - \exp(-t_{\max}/\tau)}, \ t_{\min} \leq t < t_{\max} \\ 0, \ t > t_{\max} \end{cases}$$

The denominator $[\exp(-t_{\min}/\tau) - \exp(-t_{\max}/\tau)]$ corrects for those particles that decay before $t_{\min}$ or after $t_{\max}$.
The detection probability $a(t)$ can also have a much more complicated dependence on t. For example, whether or not a decay can be observed at t may depend on the nature and geometric configuration of the decay products.

Subsection: The Multivariate Normal Distribution

The Multivariate Normal Distribution

The multivariate normal distribution $\mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$

- Its density is:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\mathbf{V}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Moments

Let $\boldsymbol{X} = (X_1, \ldots, X_d) \sim \mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$. Then holds:

- $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$
- $\mathrm{Cov}[\boldsymbol{X}] = \mathbf{V}$
- $\mathbf{V}^{-1}$ is also symmetric and positive definite, and is called a weight or information matrix.

The Multivariate Normal Distribution

Linear Transformations

Let $\boldsymbol{X} \sim \mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{H}$ be a $m \times d$ matrix. Then $\boldsymbol{Y} = \mathbf{H}\boldsymbol{X} \sim \mathrm{No}(\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\mathbf{V}\mathbf{H}^{\mathrm{T}})$.

Marginal Distributions

Each marginal distribution of a normal distribution is again a normal distribution. The mean and matrix of the marginal distribution are obtained by deleting the columns and rows of the remaining variables.

Conditional Distributions

Each conditional distribution of a normal distribution is again a normal distribution.

The Multivariate Normal Distribution

- If $\boldsymbol{X} \sim \mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$, then $\mathbf{V}$ can be brought to diagonal form as a positive definite symmetric matrix using an orthogonal transformation (rotation):

$$\mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}} = \mathbf{D}^2$$

- All diagonal elements of $\mathbf{D}^2$ are positive. The random variable $\boldsymbol{Z} = \mathbf{D}\mathbf{U}(\boldsymbol{X} - \boldsymbol{\mu})$ is then multivariate standard normal distributed, i.e.:

$$\mathrm{E}[\boldsymbol{Z}] = \mathbf{0}, \quad \mathrm{Cov}[\boldsymbol{Z}] = \mathbf{I}$$

The rotation $\mathbf{U}$ is called **major axis transformation**.

The Multivariate Normal Distribution

The bivariate normal distribution

- For $d = 2$ and $\boldsymbol{\mu} = \mathbf{0}$, the density can be written as follows:

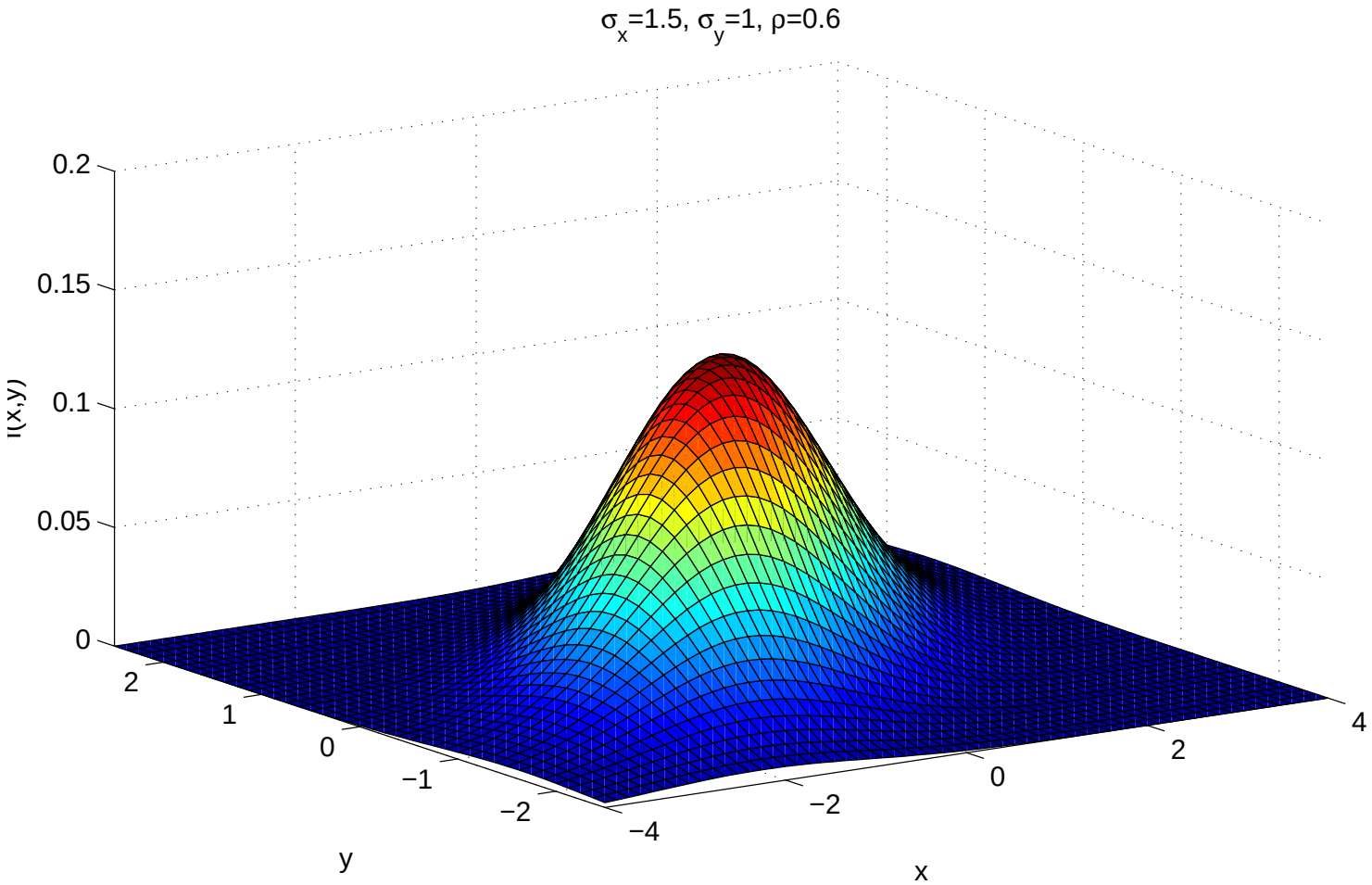
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2\,\rho\,x\,y}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)\right]$$

- $\rho = \sigma_{xy}/(\sigma_x\sigma_y)$ is the correlation coefficient. If X and Y are uncorrelated, i.e. $\rho = 0$, it follows:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right] = f_1(x)\cdot f_2(y)$$

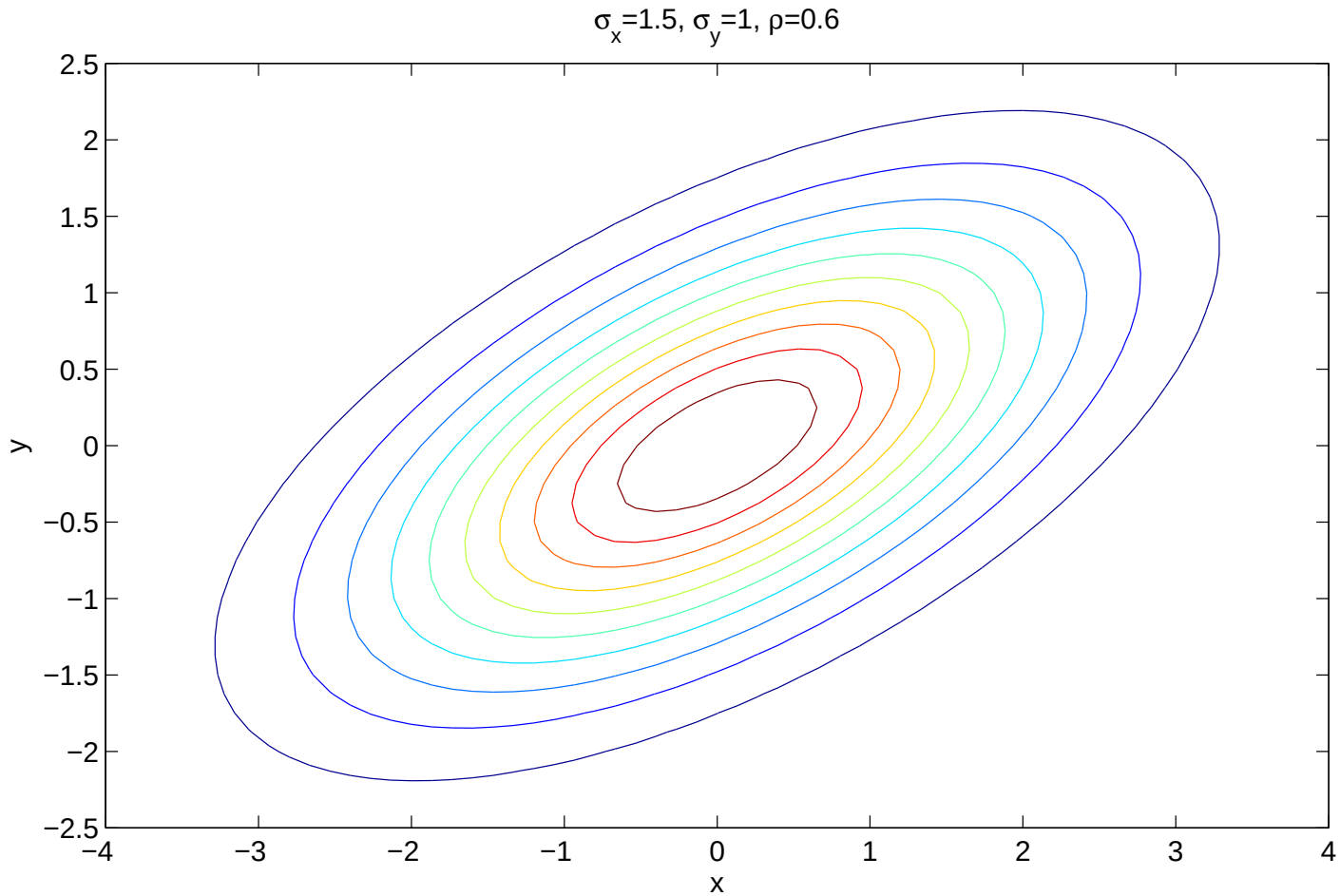
- Two uncorrelated normally distributed random variables with common normal distribution are therefore **independent**.

The Multivariate Normal Distribution



density function of a bivariate normal distribution

The Multivariate Normal Distribution



contour lines of a bivariate normal distribution

The Multivariate Normal Distribution

- The **conditional density** $f(y|x)$ is given by

$$f(y|x) = \frac{f(x,y)}{f(x)} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,\sigma_y^2\,(1-\rho^2)}\left(y - \frac{\rho\,y\,\sigma_y}{\sigma_x}\right)^2\right]$$

- $Y|X = x$ is therefore a normally distributed random variable with expectation

$$\mathrm{E}[Y|X] = \rho\,x\,\sigma_y/\sigma_x$$

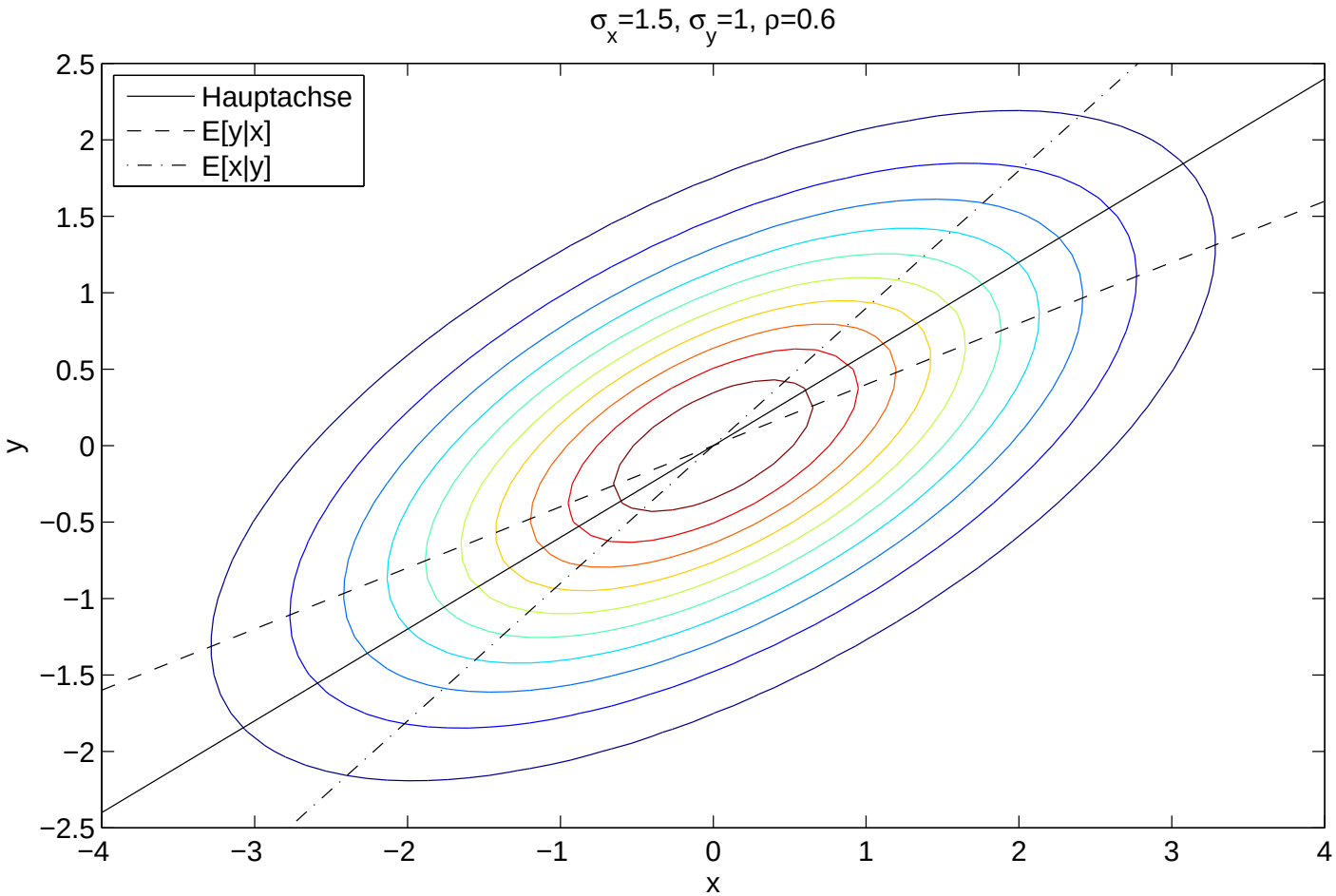
- $\mathrm{E}[Y|X]$ is called the **conditional expectation** or **regression of** y **on** x.

The Multivariate Normal Distribution

- Depending on the sign of ρ, the conditional expectation of Y falls or grows as X grows.
- If $\rho = 1$, X and Y are proportional: $Y = X\,\sigma_y/\sigma_x$.
- The contour lines of the density function are **ellipses**.
- The **major axis transformation** is that rotation which brings the ellipses into axis-parallel position.
- It depends in the case $d = 2$ only on ρ. If $\rho = 0$, X and Y are already independent, and the rotation angle is equal to 0. If $\rho \neq 0$, the rotation matrix U is equal to

$$U = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \text{ with } \varphi = -\frac{1}{2}\,\mathrm{arccot}\,\frac{\sigma_y^2 - \sigma_x^2}{2\rho\sigma_x\sigma_y}$$

The Multivariate Normal Distribution



Principal axis and regression lines of a bivariate normal distribution

Subsection: Multivariate Error Propagation

Multivariate Error Propagation

Affine Transformations

- Let $\boldsymbol{X}$ be a random variable with density $f(\boldsymbol{x})$ and $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X} + \boldsymbol{b}$ with regular $\mathbf{A}$.
- The density $g(\boldsymbol{y})$ of $\boldsymbol{Y}$ is then equal to

$$g(\boldsymbol{y}) = \frac{1}{|\mathbf{A}|} f\left(\mathbf{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})\right)$$

- Furthermore, for any $\mathbf{A}$:

$$\mathrm{E}[\boldsymbol{Y}] = \mathbf{A} \cdot \mathrm{E}[\boldsymbol{X}] + \boldsymbol{b}$$
$$\mathrm{Cov}[\boldsymbol{Y}] = \mathbf{A} \cdot \mathrm{Cov}[\boldsymbol{X}] \cdot \mathbf{A}^{\mathrm{T}}$$

Multivariate Error Propagation

Nonlinear Transformations

- Let $\boldsymbol{X}$ be a random variable with density $f(\boldsymbol{x})$ and $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$.
- If $\boldsymbol{h}(\boldsymbol{x})$ is bijective, the density $g(\boldsymbol{y})$ of $\boldsymbol{Y}$ is equal to

$$g(\boldsymbol{y}) = \left| \frac{\partial \boldsymbol{h}^{-1}}{\partial \boldsymbol{y}} \right| f\left(\boldsymbol{h}^{-1}(\boldsymbol{y})\right)$$

- The expectation and variance of $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$ can be calculated **approximately** using the Taylor expansion of $\boldsymbol{h}(\boldsymbol{x})$, even if $\boldsymbol{h}(\boldsymbol{x})$ is not bijective.

Multivariate Error Propagation

- With the development point $\boldsymbol{x}_0$ holds in linear approximation

$$\boldsymbol{h}(\boldsymbol{x}) \approx \boldsymbol{h}(\boldsymbol{x}_0) + \mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0), \quad \mathbf{H} = \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

- With the choice $\boldsymbol{x}_0 = \mathrm{E}[\boldsymbol{X}]$ follows.

Theorem

$$\mathrm{E}[\boldsymbol{h}(\boldsymbol{X})] \approx \boldsymbol{h}(\mathrm{E}[\boldsymbol{X}])$$

$$\mathrm{Cov}[\boldsymbol{h}(\boldsymbol{X})] \approx \mathbf{H} \cdot \mathrm{Cov}[X] \cdot \mathbf{H}^{\mathrm{T}} \quad \text{(Linear Error Propagation)}$$

Section 30: Simple Regression

Subsection: Linear Regression

Linear Regression

- The simplest regression model is a straight line:

$$Y = \alpha + \beta x + \varepsilon, \quad \mathrm{E}[\varepsilon] = 0, \ \mathrm{var}[\varepsilon] = \sigma^2$$

- Now let $Y_1, \ldots, Y_n$ be the results for the values $x_1, \ldots, x_n$ of the influence variable x.
- The estimation of α and β can be done using the principle of least squares error.
- The following objective function is minimized:

$$SS = \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2$$

- gradient of SS:

$$\frac{\partial SS}{\partial \alpha} = -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i), \ \frac{\partial SS}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (Y_i - \alpha - \beta x_i)$$

Linear Regression

- Zeroing the gradient gives the **normal equations**:

$$\sum_{i=1}^{n} Y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i Y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$$

- The estimated regression coefficients are:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\bar{x}$$

Linear Regression

- The following applies

$$\mathsf{E}[\hat{\alpha}] = \alpha, \quad \mathsf{E}[\hat{\beta}] = \beta$$

- The variance of the error term is estimated with expectation confidence by:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2$$

with

$$r_i = Y_i - \hat{Y}_i, \quad \hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$$

- The r_i are called the **residuals** of the regression.

Linear Regression

- The covariance matrix of the estimated regression coefficients is obtained by linear error propagation:

$$\mathrm{Cov}[\hat{\alpha}, \hat{\beta}] = \sigma^2 \begin{pmatrix} \frac{\sum x_i^2}{n\left(\sum x_i^2 - n\bar{x}^2\right)} & -\frac{\sum x_i}{n\left(\sum x_i^2 - n\bar{x}^2\right)} \\ -\frac{\sum x_i}{n\left(\sum x_i^2 - n\bar{x}^2\right)} & \frac{1}{\sum x_i^2 - n\bar{x}^2} \end{pmatrix}$$

Linear Regression

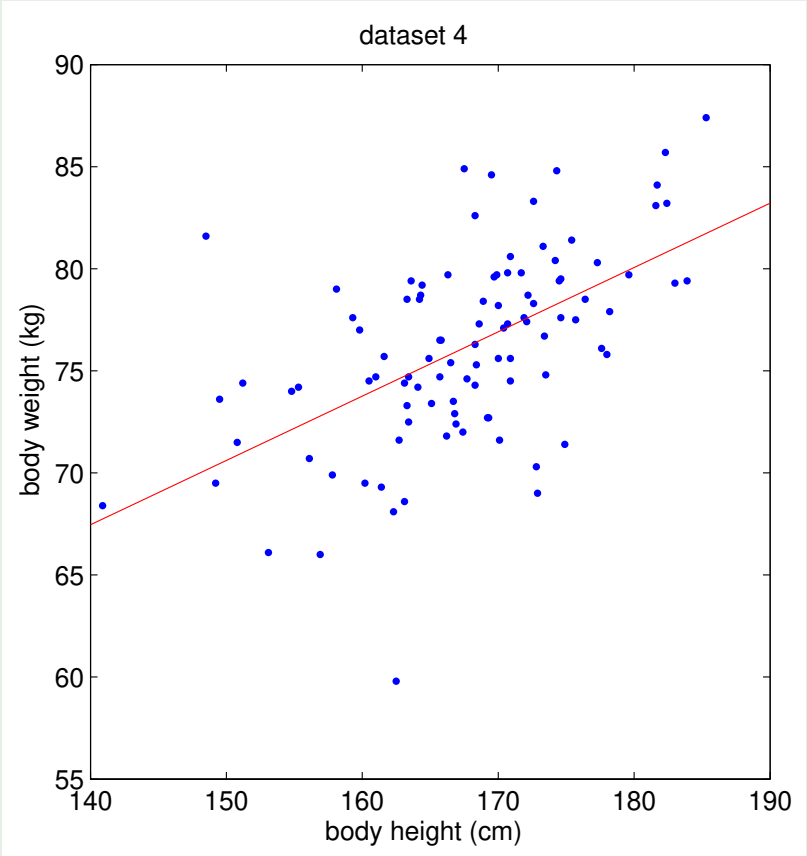
Example

Data set 4:

$$\bar{x} = 167.60 \qquad r_{xy} = 0.5562$$
$$\bar{y} = 76.16 \qquad \hat{a} = 23.37$$
$$s_x = 8.348 \qquad \hat{b} = 0.3150$$
$$s_y = 4.727$$

MATLAB: `make_dataset4`

Example (Continuation)



scatter plot with regression line

Linear Regression

- The scatter of the values Y_i has different causes in the regression model. causes.
- On the one hand, there are systematic differences due to different values of x.
- On top of that there is the random scattering of the data.

Explained scatter $$SS^* = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = r_{xy}^2 n s_Y^2$$

Residual scatter $$SS_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (1 - r_{xy}^2) n s_Y^2$$

Total dispersion $$SS_T = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = n s_Y^2$$

Linear Regression

Streuungszerlegung

$$SS_T = SS^* + SS_R$$

- The goodness of the regression line can be determined by the **measure of determination**:

Measure of determination of the regression

$$B = \frac{SS^*}{SS_T} = r_{xy}^2$$

- It indicates what proportion of the total scatter can be explained by the correlation of x and Y.

Subsection: Tests, Confidence and Prediction Intervals

Tests, Confidence and Prediction Intervals

- If $\beta = 0$, the result does not depend on the influence variables at all.
- A test of the null hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ is based on the following theorem.

Theorem

If ε is normally distributed, then.

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}}, \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}}$$

t-distributed with $n - 2$ degrees of freedom, where.

$$\hat{\sigma}^2_{\hat{\alpha}} = \frac{\hat{\sigma}^2 \sum x_i^2}{n\left(\sum x_i^2 - n\bar{x}^2\right)}, \quad \hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}$$

Tests, Confidence and Prediction Intervals

- The null hypothesis $H_0 : \beta = 0$ is rejected, if the test statistic

$$T = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

is relatively small or relatively large, i.e. if

$$\frac{|\hat{\beta}|}{\hat{\sigma}_{\hat{\beta}}} > t_{1-\alpha/2;n-2}$$

where $t_{p;n-2}$ is the p quantile of the t distribution with $n-2$ degrees of freedom.

- An analogous test can be performed for the null hypothesis $H_0 : \alpha = 0$ can be performed.

Tests, Confidence and Prediction Intervals

Symmetric Confidence Intervals

$$\hat{\alpha} \pm \hat{\sigma}_{\hat{\alpha}} \cdot t_{1-\alpha/2;n-2}, \quad \hat{\beta} \pm \hat{\sigma}_{\hat{\beta}} \cdot t_{1-\alpha/2;n-2}$$

- For $n \gtrsim 50$, the quantiles of the t distribution can be replaced by quantiles of the standard normal distribution.
- We now want to predict the outcome $Y_0 = Y(x_0)$ for a given value x_0 of the influence variable x.
- The expected value of Y_0 is

$$\mathsf{E}[Y_0] = \hat{\alpha} + \hat{\beta} x_0$$

- The variance of $\mathsf{E}[Y_0]$ is obtained by error propagation:

$$\mathsf{var}[\mathsf{E}[Y_0]] = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2} \right]$$

Tests, Confidence and Prediction Intervals

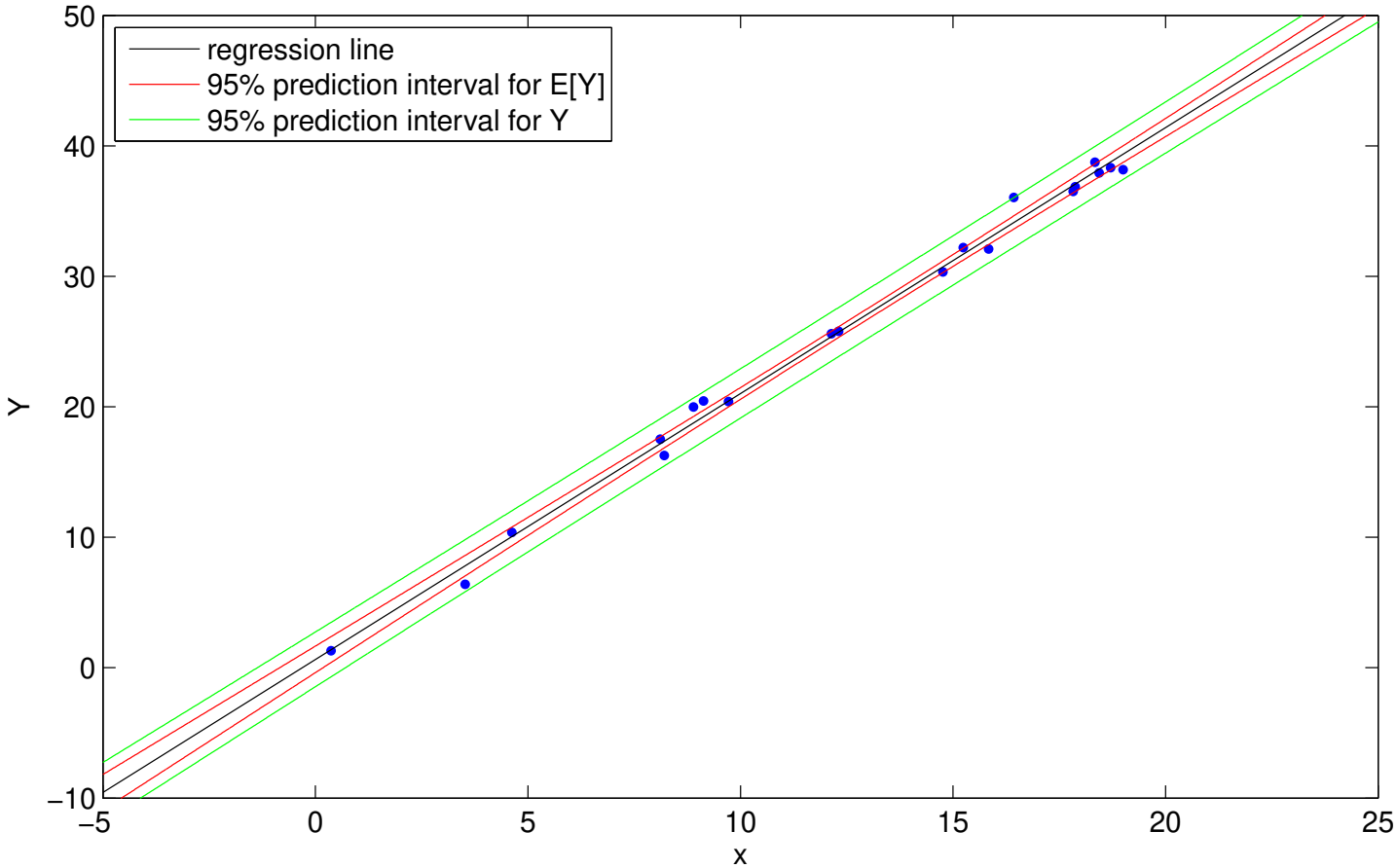
- Since Y_0 scatters around its expected value with variance σ^2, we get:

$$\text{var}[Y_0] = \sigma^2 \left[\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2} \right]$$

- The symmetric forecast interval for Y_0 with certainty α is therefore the same:

$$\hat{\alpha} + \hat{\beta} x_0 \pm t_{1-\alpha/2;n-2} \hat{\sigma} \sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}}$$

Tests, Confidence and Prediction Intervals



prediction bands for $\mathrm{E}[Y]$ and Y

Tests, Confidence and Prediction Intervals

- The adequacy of the model can be checked by examining the **studentized residuals** (residual errors).
- The residual r_k has the variance

$$\text{var}[r_k] = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_k - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}\right]$$

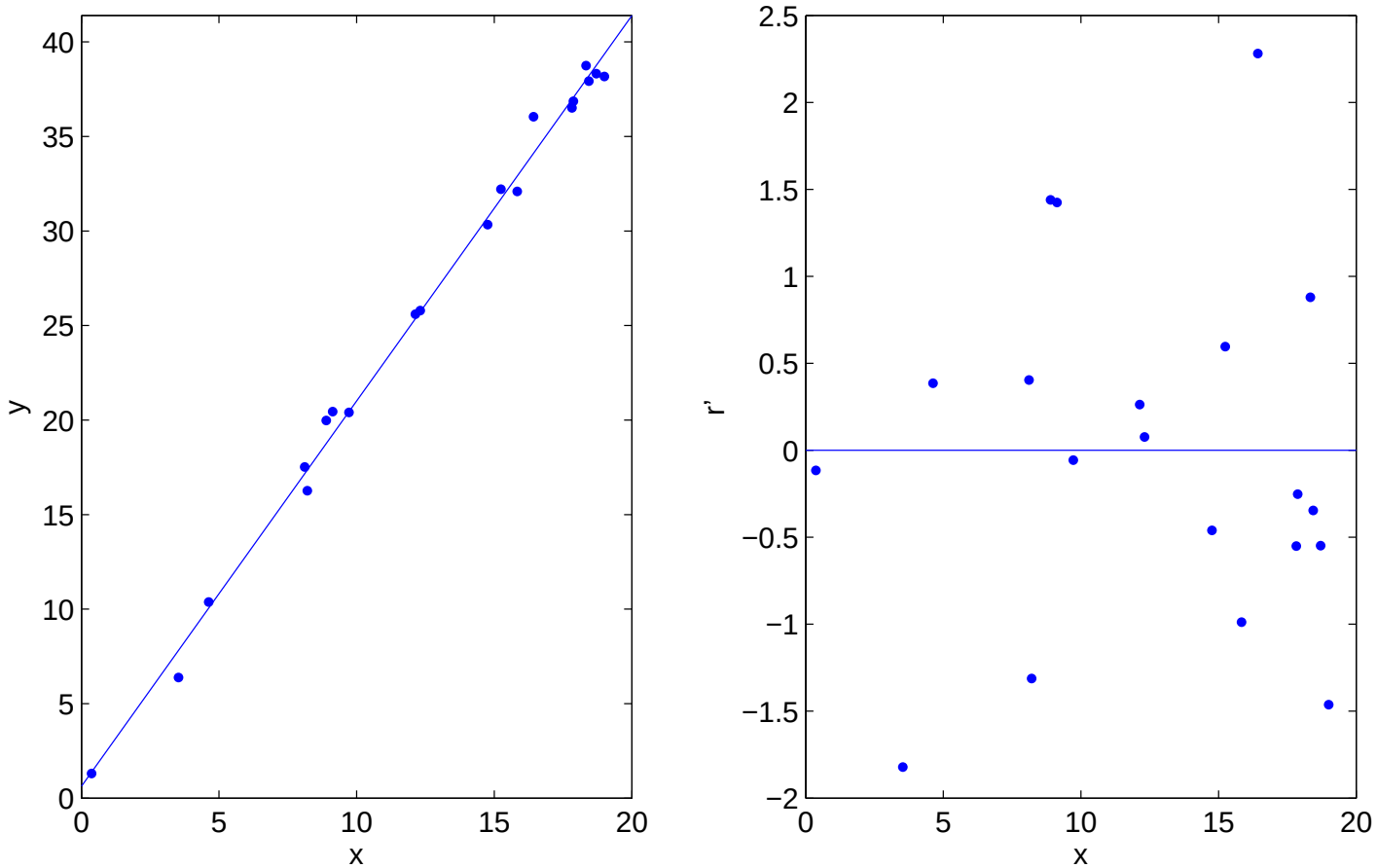
- The studentized residual is then.

$$r'_k = \frac{r_k}{\hat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_k-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}}$$

- It has expectation 0 and variance 1.

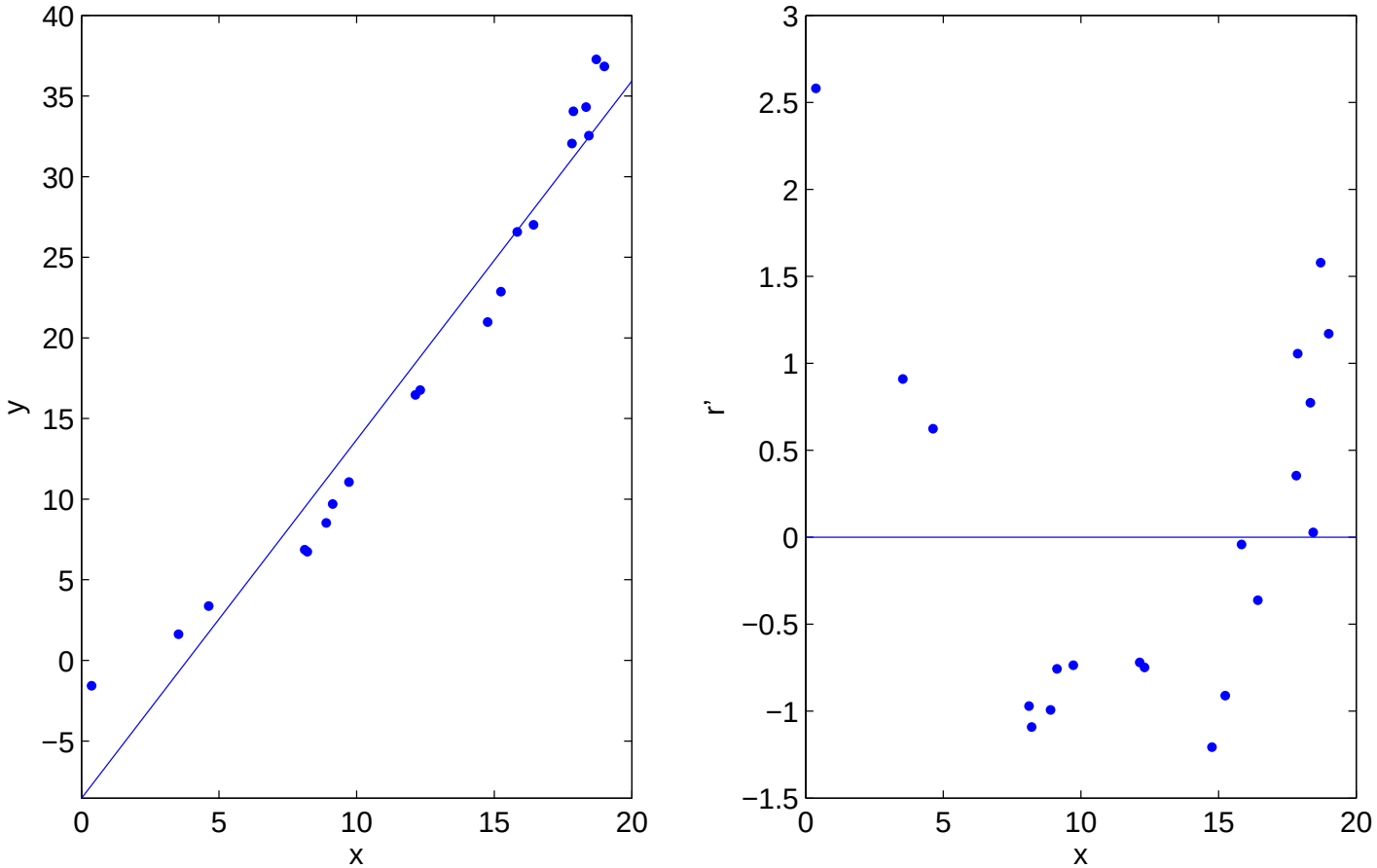
MATLAB: `make_regression_diagnostics`

Tests, Confidence and Prediction Intervals



regression line and studentized residuals.

Tests, Confidence and Prediction Intervals



regression line and studentized residuals.

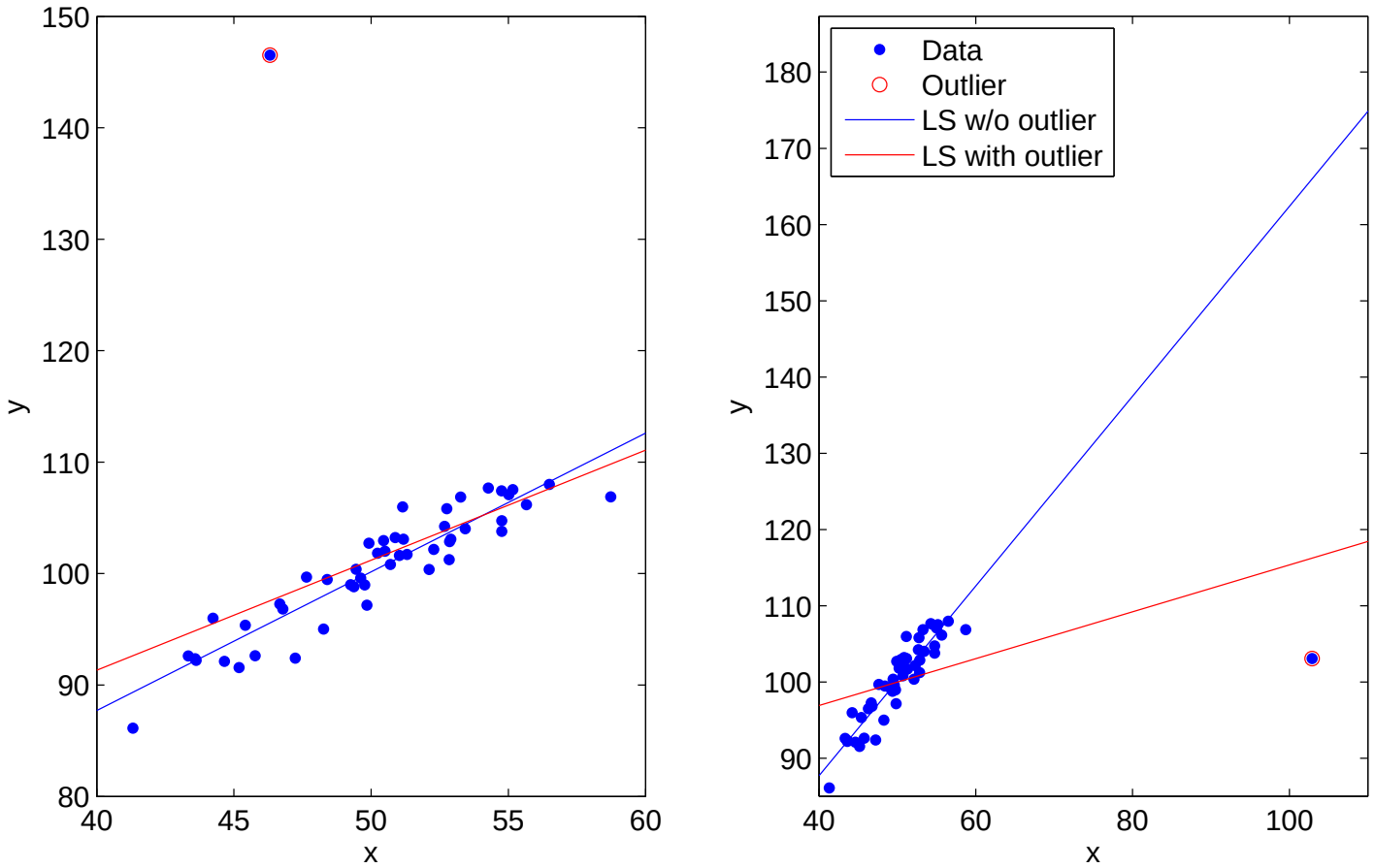
Subsection: Robust Regression

Robust Regression

- As an LS estimator, the regression line is not robust, i.e. sensitive to outliers.
- Single outliers in the outcome variable y usually result in a slight distortion of the regression line.
- Single outliers in the influence variable x, so-called leverage points, can cause catastrophic distortions. have.

MATLAB: `make_regression_outliers`

Robust Regression



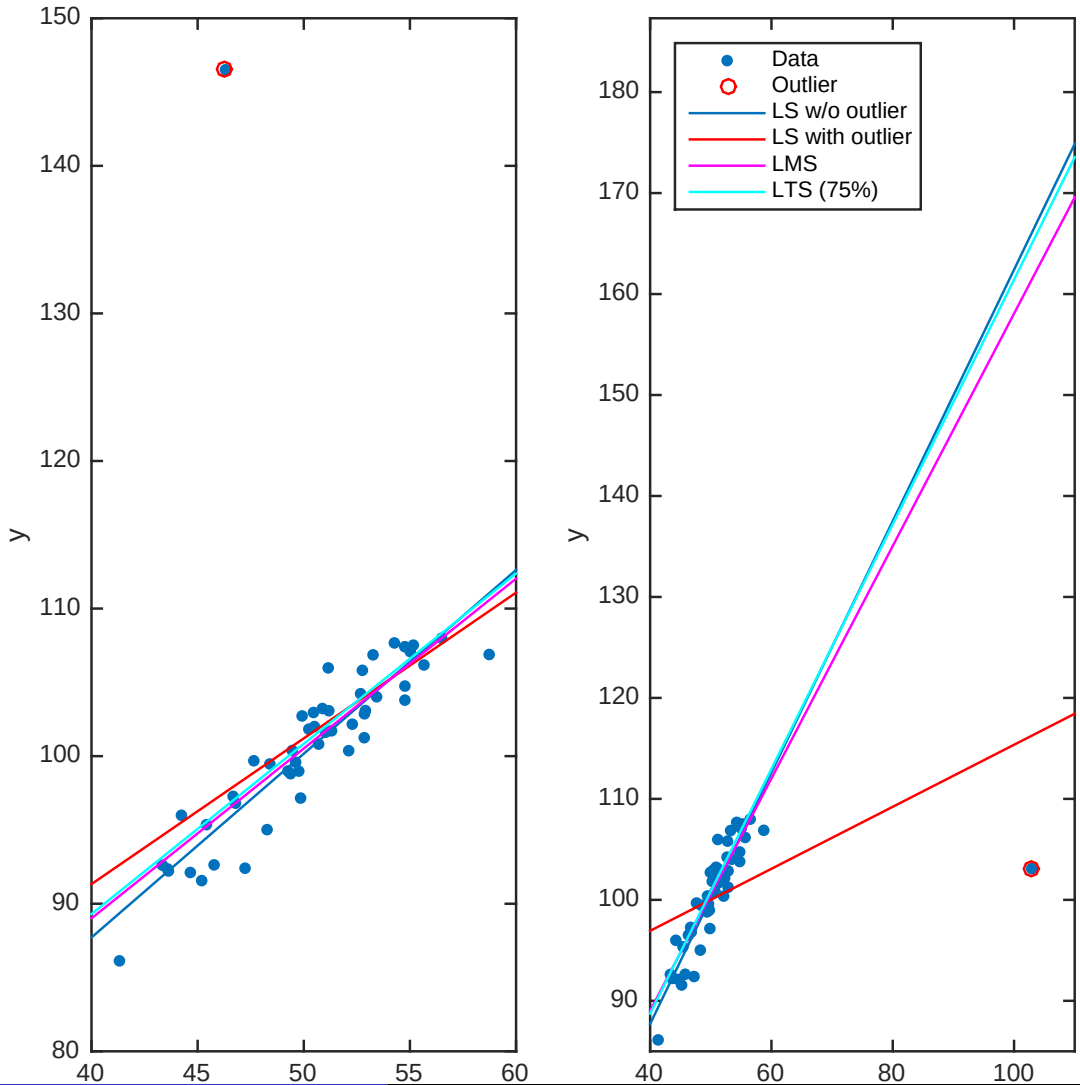
linear regression with outliers

Robust Regression

- **LMS (Least Median of Squares)**: Instead of the sum of the error squares, the **median** of the error squares is minimized.
- "Exact fit property": The LMS straight line passes through two data points.
- Calculation combinatorial.
- **LTS (Least Trimmed Squares)**: The sum of a fixed number $h \leq n$ of error squares is minimized.
- Calculation iterative (FAST-LTS).
- Both methods go back to P. Rousseeuw.

MATLAB: `make_robust_regression`

Robust Regression



Subsection: Polynomial Regression

Polynomial Regression

- If the relationship between x and Y is not approximately linear, you can try to fit a polynomial.
- The model is then:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + \varepsilon, \quad \mathsf{E}[\varepsilon] = 0, \ \mathsf{var}[\varepsilon] = \sigma^2$$

- Let again $Y_1, \ldots, Y_n$ be the results for the values $x_1, \ldots, x_n$ of the influence variables x.
- In matrix vector notation:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathsf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathsf{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}$$

with

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^r \\ 1 & x_2 & x_2^2 & \cdots & x_2^r \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_n & x_n^2 & \cdots & x_n^r \end{pmatrix}$$

Polynomial Regression

- The following target function is minimized:

$$SS = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- gradient of SS:

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Zeroing the gradient gives the **normal equations**:

$$\mathbf{X}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

- The solution is:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$$

Polynomial Regression

- $\hat{\beta}$ is an expectation-trusted estimator of β.
- The variance of the error term is estimated expectation-trusted by:

$$\hat{\sigma}^2 = \frac{1}{n-r-1} \sum_{i=1}^{n} r_i^2$$

with the vector of residuals

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \quad \hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

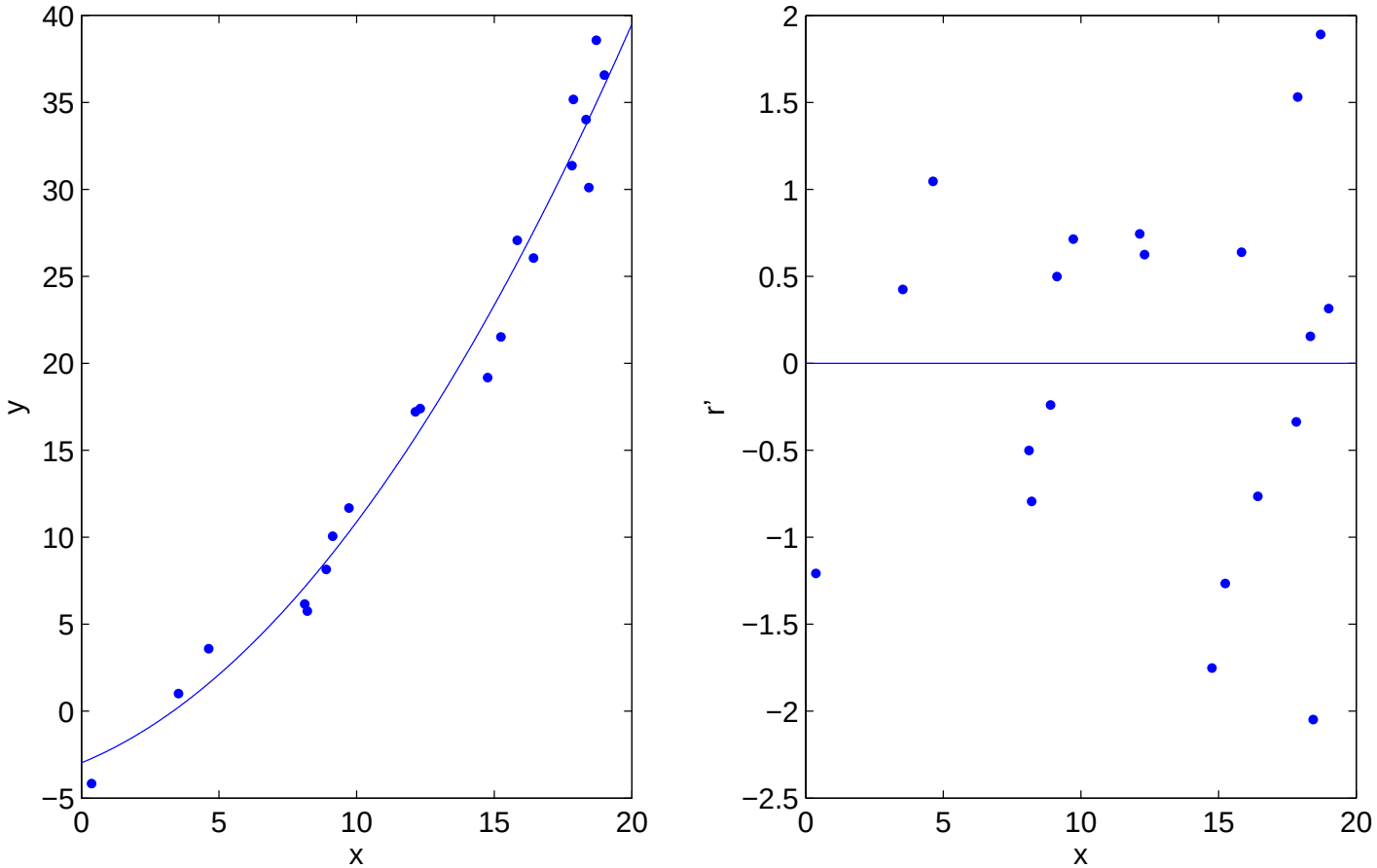
- covariance matrix of estimated regression coefficients:

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

- covariance matrix of residuals $\boldsymbol{r}$:

$$\mathrm{Cov}[\boldsymbol{r}] = \sigma^2 \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]$$

Polynomial Regression



regression parabola and studentized residuals.

Section 31: Multiple Regression

Subsection: The Linear Model

The Linear Model

- If the outcome Y depends on several influencing variables, the simplest linear regression model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \varepsilon, \quad \mathsf{E}[\varepsilon] = 0, \ \mathsf{var}[\varepsilon] = \sigma^2$$

- Let again $Y_1, \ldots, Y_n$ be the results for n values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the influence variables $\boldsymbol{x} = (x_1, \ldots, x_r)$.
- In matrix vector notation:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathsf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathsf{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}$$

with

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,r} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,r} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,r} \end{pmatrix}$$

Subsection: Estimation, Tests, and Prediction Intervals

Estimation, Tests, and Prediction Intervals

- The following objective function is minimized:

$$SS = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- gradient of SS:

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Zeroing the gradient gives the **normal equations**:

$$\mathbf{X}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

- The solution is:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$$

Estimation, Tests, and Prediction Intervals

- $\hat{\boldsymbol{\beta}}$ is an expectation-trusted estimator of $\boldsymbol{\beta}$.
- The variance of the error term is estimated expectation-trusted by:

$$\hat{\sigma}^2 = \frac{1}{n-r-1}\sum_{i=1}^{n} r_i^2$$

with

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \quad \hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

- covariance matrix of estimated regression coefficients:

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

- covariance matrix of residuals $\boldsymbol{r}$:

$$\mathrm{Cov}[\boldsymbol{r}] = \sigma^2 \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]$$

Estimation, Tests, and Prediction Intervals

- If $\beta_k = 0$, the result does not depend at all on the influence variables x_k.
- A test of the null hypothesis $H_0 : \beta_k = 0$ against $H_1 : \beta_k \neq 0$ is based on the following theorem.

Theorem

If ε is normally distributed, then

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$$

t-distributed with $n - r - 1$ degrees of freedom, where $\hat{\sigma}^2_{\hat{\beta}_k}$ is the k-th diagonal element of the estimated covariance matrix.

$$\hat{\sigma}^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

is.

Estimation, Tests, and Prediction Intervals

- The null hypothesis $H_0 : \beta_k = 0$ is rejected, if the test statistic

$$T = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}$$

is relatively small or relatively large, i.e. if

$$\frac{|\hat{\beta}_k|}{\hat{\sigma}_{\hat{\beta}_k}} > t_{1-\alpha/2;n-r-1}$$

- The symmetric confidence interval for β_k with 95% confidence is:

$$\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} \cdot t_{1-\alpha/2;n-r-1}$$

Estimation, Tests, and Prediction Intervals

- We now want to predict the outcome $Y_0 = Y(\boldsymbol{x}_0)$ for a given value $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{0r})$ of the influence variable.
- We extend $\boldsymbol{x}_0$ by the value 1: $\boldsymbol{x}_+ = (1, x_{01}, \ldots, x_{0r})$.
- The expected value of Y_0 is then

$$\mathsf{E}[Y_0] = \boldsymbol{x}_+ \cdot \hat{\boldsymbol{\beta}}$$

- The variance of $\mathsf{E}[Y_0]$ is obtained by error propagation:

$$\mathsf{var}[\mathsf{E}[Y_0]] = \sigma^2 \boldsymbol{x}_+ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \boldsymbol{x}_+{}^{\mathrm{T}}$$

Estimation, Tests, and Prediction Intervals

- Since Y_0 scatters around its expected value with variance σ^2, we get:

$$\text{var}[Y_0] = \sigma^2 \left[1 + \boldsymbol{x}_+ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \boldsymbol{x}_+^{\mathrm{T}}\right]$$

- The symmetric prediction interval for Y_0 with certainty α is therefore the same:

$$\boldsymbol{x}_+ \cdot \hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2;n-k-1} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_+ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \boldsymbol{x}_+^{\mathrm{T}}}$$

Subsection: Weighted Regression

Weighted Regression

- In the general case, the error term can have any covariance matrix:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{Cov}[\boldsymbol{\varepsilon}] = \mathbf{V}$$

- If $\mathbf{V}$ is known, the objective function is:

$$SS = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}), \quad \mathbf{G} = \mathbf{V}^{-1}$$

- gradient of SS:

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Zeroing the gradient gives the **normal equations**:

$$\mathbf{X}^{\mathrm{T}}\mathbf{G}\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\boldsymbol{\beta}$$

- The solution is:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{G}\boldsymbol{Y}$$

Weighted Regression

- Covariance matrix of the estimated regression coefficients:

$$\text{Cov}[\hat{\boldsymbol{\beta}}] = \left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\right)^{-1}$$

- covariance matrix of residuals $\boldsymbol{r}$:

$$\text{Cov}[\boldsymbol{r}] = \mathbf{V} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$$

- Tests and prediction intervals can be modified accordingly.

Subsection: Nonlinear Regression

Nonlinear Regression

- In practice, the dependence of the results on the regression coefficients is often nonlinear:

$$\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \quad \mathrm{Cov}[\boldsymbol{\varepsilon}] = \mathbf{V}$$

- If $\mathbf{V}$ is known, the objective function is:

$$SS = [\boldsymbol{Y} - \boldsymbol{h}(\boldsymbol{\beta})]^{\mathrm{T}} \mathbf{G} [\boldsymbol{Y} - \boldsymbol{h}(\boldsymbol{\beta})], \quad \mathbf{G} = \mathbf{V}^{-1}$$

- SS can be minimized using the Gauss-Newton method.
- For this $\boldsymbol{h}$ is linearized at a point $\boldsymbol{\beta}_0$:

$$\boldsymbol{h}(\boldsymbol{\beta}) \approx \boldsymbol{h}(\boldsymbol{\beta}_0) + \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \boldsymbol{c} + \mathbf{H}\boldsymbol{\beta}, \quad \mathbf{H} = \left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0}$$

Nonlinear Regression

- The $\boldsymbol{\beta}$ estimate is:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \boldsymbol{c})$$

- $\boldsymbol{h}$ is linearized again at the point $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}$.
- The procedure is iterated until the estimate does not change significantly.
- Many other methods for minimizing SS are available.

MATLAB: `fminsearch,fminunc`

Section 32: Data Reconciliation

Data Reconciliation

Linear Constraints

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector of observations of the n unknown quantities $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$:

$$\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}, \mathrm{Cov}[\boldsymbol{\epsilon}] = \mathbf{V},$$

with known $\mathbf{V}$. Moreover, let it be known that the quantities $\boldsymbol{\beta}$ m must satisfy independent linear constraints, which r unknown Parameters $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_r)^{\mathrm{T}}$ include:

$$\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} = \boldsymbol{c}$$

Where $\mathbf{A}$ is a matrix of dimension $m \times n$, $\mathbf{B}$ is a matrix of dimension $m \times r$, and $\boldsymbol{c}$ a $m \times 1$ dimensional column vector.

Data Reconciliation

Example (Continuation)

$\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ are estimated according to the principle of least squares of error estimated. The objective function is

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{G} (\boldsymbol{Y} - \boldsymbol{\beta}), \mathbf{G} = \mathbf{V}^{-1}$$

under the condition:

$$\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} = \boldsymbol{c}.$$

The problem is determinate if $r \leq m$ and inconsistent if $m \leq n + r$. The minimization of S can be done using m Lagrangians $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathrm{T}}$. The extended objective function is

$$L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \boldsymbol{\lambda}) = (\boldsymbol{Y} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{G} (\boldsymbol{Y} - \boldsymbol{\beta}) + 2\boldsymbol{\lambda}^{\mathrm{T}} (\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} - \boldsymbol{c})$$

with gradients

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\mathbf{G}(\boldsymbol{Y} - \boldsymbol{\beta}) + 2\mathbf{A}^{\mathrm{T}}\boldsymbol{\lambda}, \frac{\partial L}{\partial \boldsymbol{\vartheta}} = 2\mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda}, \frac{\partial L}{\partial \boldsymbol{\lambda}} = 2(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} - \boldsymbol{c})$$

Data Reconciliation

Example (Continuation)

Zeroing the gradient leads to the following system of linear equations:

$$\mathbf{G}\boldsymbol{\beta} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{G}\boldsymbol{Y}$$

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{0}$$

$$\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} = \boldsymbol{c}$$

With the additional assumption $m \leq n$ and the labels $\mathbf{G_A} = .(\mathbf{AVA}^{\mathrm{T}})^{-1}, \mathbf{V_B} = (\mathbf{B}^{\mathrm{T}}\mathbf{G_A}\mathbf{B})^{-1}$ we obtain the following closed-form solution:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{Y} + \mathbf{VA}^{\mathrm{T}}\mathbf{G_A}[\mathbf{I} - \mathbf{B}\mathbf{V_B}^{\mathrm{T}}\mathbf{G_A}](\boldsymbol{c} - \mathbf{A}\boldsymbol{Y})$$

$$\hat{\boldsymbol{\vartheta}} = \mathbf{V_B}\mathbf{B}^{\mathrm{T}}\mathbf{G_A}(\boldsymbol{c} - \mathbf{A}\boldsymbol{Y})$$

The joint covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\vartheta}}$ is:

$$\mathrm{Cov}\left[\begin{pmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\vartheta}}\end{pmatrix}\right] = \begin{pmatrix}\mathbf{V} + \mathbf{VA}^{\mathrm{T}}\mathbf{G_A}(\mathbf{BV_B}\mathbf{B}^{\mathrm{T}}\mathbf{G_A} - \mathbf{I})\mathbf{AV} & -\mathbf{VA}^{\mathrm{T}}\mathbf{G_A}\mathbf{BV_B} \\ -\mathbf{V_B}\mathbf{V}^{\mathrm{T}}\mathbf{G_A}\mathbf{AV} & \mathbf{V_B}\end{pmatrix}$$

Data Reconciliation

Example (Continuation)

If the constraints do not contain unknown parameters, the following applies:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{Y} + \mathbf{V}\mathbf{A}^{\mathrm{T}}\mathbf{G}_{\mathbf{A}}(\boldsymbol{c} - \mathbf{A}\boldsymbol{Y}), \mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \mathbf{V} - \mathbf{V}\mathbf{A}^{\mathrm{T}}\mathbf{G}_{\mathbf{A}}\mathbf{A}\mathbf{V}$$

The χ^2 statistic of the balance is defined by:

$$\chi^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \hat{\boldsymbol{\beta}})$$

If $\boldsymbol{Y}$ is normally distributed with covariance matrix $\mathbf{V} = \mathbf{G}^{-1}$, then χ^2 χ^2-distributed with $m - r$ degrees of freedom.

Data Reconciliation

Example (balancing angles in triangle)

In a triangle, the angles $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are measured. Let the values in degrees be $\boldsymbol{Y} = (34.26, 86.07, 59.52)$, the measurement errors $\sigma_1 = 0.1, \sigma_2 = 0.08, \sigma_3 = 0.12$. The measured angles can be compensated with the condition $\beta_1 + \beta_2 + \beta_3 = 180$ can be compensated.

- It is $\mathbf{A} = (1, 1, 1), c = 180$, and $\mathbf{V} = \mathrm{diag}(.1^2, .08^2, .12^2)$.
- Furthermore, $\mathbf{G_A} = (\mathbf{AVA}^{\mathrm{T}})^{-1} = 32.468$.
- Thus $\hat{\boldsymbol{\beta}} = \boldsymbol{Y} + \mathbf{VA}^{\mathrm{T}}\mathbf{G_A}(c - \mathbf{A}\boldsymbol{Y}) = (34.31, 86.10, 59.59)$
- $\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \mathbf{V} - \mathbf{VA}^{\mathrm{T}}\mathbf{G_A}\mathbf{AV} = 10^{-2} \cdot \begin{pmatrix} .68 & -.21 & -.47 \\ -.21 & .51 & -.30 \\ -.47 & -.30 & .77 \end{pmatrix}$.

  All estimated values are negatively correlated.
- $\chi^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\beta}})^T \mathbf{G} (\boldsymbol{Y} - \hat{\boldsymbol{\beta}}) = 0.846, p = 0.36$. The errors seem to be realistic.

Data Reconciliation

Nonlinear Constraints

If the constraints are nonlinear, they can be written in the following general form:

$$\boldsymbol{h}(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = \mathbf{0}$$

First, $\boldsymbol{h}$ is linearized at appropriate point $(\boldsymbol{\beta}_0, \boldsymbol{\vartheta}_0)$:

$$\boldsymbol{h}(\boldsymbol{\beta}, \boldsymbol{\vartheta}) \approx \boldsymbol{h}(\boldsymbol{\beta}_0, \boldsymbol{\vartheta}_0) + \mathbf{A}_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \mathbf{B}_0(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) = \mathbf{A}_0\boldsymbol{\beta} + \mathbf{B}_0\boldsymbol{\vartheta} - \boldsymbol{c}_0$$

with

$$\mathbf{A}_0 = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta}}, \mathbf{B}_0 = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\vartheta}}, \boldsymbol{c}_0 = \mathbf{A}_0\boldsymbol{\beta}_0 + \mathbf{B}_0\boldsymbol{\vartheta}_0 - \boldsymbol{h}(\boldsymbol{\beta}_0, \boldsymbol{\vartheta}_0).$$

The estimation follows as in the linear case. Then $\boldsymbol{h}$ is estimated at the Place $\boldsymbol{\beta}_1, \boldsymbol{\vartheta}_1 = \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\vartheta}}$ is re-linearized. It is repeated until the constraints are satisfied exactly enough.

Part VIII

Bayes Statistics

Overview Part 8

Section 33: Introduction and Basic Terminology

Introduction and Basic Terminology

- In Bayesian statistics, probabilities are interpreted as rational estimates of facts.
- In addition to data, unknown parameters of distributions are also considered as **random variables**.
- Prior information about parameters is summarized in a **a priori distribution**.
- The information in the data leads to improved information about the parameters by applying Bayes' theorem, which is expressed by the **a-posteriori distribution**.
- The a-posteriori distribution can be used for estimating parameters, calculating confidence intervals, testing, forecasting, model selection, etc.

Introduction and Basic Terminology

- By choosing the a priori distribution, a subjective component is introduced into the data analysis.
- This can have a noticeable impact on the results, especially when the data is very small.
- However, there are methods for constructing a priori densities that minimize this influence.
- If there is a lot of data, the influence of the a priori distribution becomes negligible.
- In this case, Bayesian analysis gives approximately the same results as classical "frequentist" methods.
- However, more information is contained in the a posteriori distribution than in classical point or interval estimators.

Introduction and Basic Terminology

- The **sample space** $\mathcal{Y}$ is the set of all possible samples $\boldsymbol{y}$.
- The **parameter space** Θ is the set of all possible values of the parameter ϑ.
- The **a priori distribution** $\pi(\vartheta)$ describes our estimate of whether a particular value ϑ describes the true distribution of the data.
- The **likelihood function** $p(\boldsymbol{y}|\vartheta)$ describes the probability that the sample $\boldsymbol{y}$ will be observed if ϑ is true.
- If $\boldsymbol{y}$ is observed, our estimate of ϑ can be updated by using Bayes' theorem to compute the **a-posteriori distribution**:

$$p(\vartheta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\vartheta)\pi(\vartheta)}{\int_\Theta p(\boldsymbol{y}|\vartheta)\pi(\vartheta)\,\mathrm{d}\vartheta}$$

Introduction and Basic Terminology

- The a-posteriori distribution describes our estimate of ϑ in light of the data $\boldsymbol{y}$.
- If new data are added, the a-posteriori distribution can be used as the new a-priori distribution.
- If the integral in the denominator cannot be calculated analytically, one must resort to Monte Carlo methods.
- The **posterior mean)** or the **posterior mode)** of $p(\vartheta|\boldsymbol{y})$ are often used as point estimators of ϑ, occasionally also the **posterior median)**.

Section 34: A-priori Distributions

A-priori Distributions

- The choice of a-priori distribution can have a relatively large effect on the result for small samples.
- We distinguish **informative** and **non-informative** a-priori distributions.
- An informative a-priori distribution describes actual information about the unknown parameter ϑ. This can be subjective or objective.
- A non-informative a-priori distribution describes the absence of such information.
- In any case, the sensitivity of the result with respect to the a-priori distribution should be investigated.

A-priori Distributions

- The a-priori distribution need not be normalized because the normalization factor truncates away.
- If the a-priori distribution can be normalized, it is called **proper**, otherwise it is called **improper**.
- Also an impropere a-priori distribution can lead to a proper a-posteriori distribution.
- The choice of a-priori distribution is also often influenced by purely computational considerations.
- For some forms of the likelihood function, there are a-priori distributions that produce a-posteriori distributions from the same distribution family.
- Such a-priori distributions are called **conjugate** a-priori distributions.
- In some cases, a non-informative a-priori distribution is also a conjugate a-priori distribution.

A-priori Distributions

- There are several suggestions for choosing a non-informative a-priori distribution:
  - principle of **maximum entropy**
  - invariance under parameter transformations: **Jeffrey's prior**
  - Minimum influence of a-priori distribution on a-posteriori distribution: **Reference prior**
- For one-dimensional ϑ, Jeffrey's prior and Reference prior are mostly identical; this is no longer true in the multidimensional case.

A-priori Distributions

Continuous distributions with maximum entropy

- $\mathrm{E}[X]$ and $\mathrm{var}[X]$ are given: Normal distribution
- $X \geq 0$, $\mathrm{E}[\ln X]$ and $\mathrm{var}[\ln X]$ given: Lognormal distribution
- $X \geq 0$, $\mathrm{E}[X]$ and $\mathrm{E}[\ln X]$ given: Gamma distribution
- $X \geq 0$, $\mathrm{E}[X]$ given: Exponential distribution
- $X \in [a, b]$: Uniform distribution on $[a, b]$

Discrete distributions with maximum entropy

- $X \in \{1, \ldots, n\}$: Uniform distribution on $\{1, \ldots, n\}$
- $X \in \mathbb{N}$, $\mathrm{E}[X]$ given: Geometric distribution

A-priori Distributions

- **Jeffrey's prior** is constructed such that the a-priori distribution is invariant under a transformation of the parameter ϑ.
- Let $\tau = h(\vartheta)$ be the transformed parameter and $\pi_{\mathrm{J}}(\vartheta)$ be Jeffrey's prior in ϑ.
- Then the transformed a-priori distribution in τ is equal to

$$\pi(\tau) = \pi_{\mathrm{J}}(\vartheta) \left| \frac{\mathrm{d}\vartheta}{\mathrm{d}\tau} \right|$$

- The transformed Fisher information is equal to.

$$I(\tau) = I(\vartheta) \left(\frac{\mathrm{d}\vartheta}{\mathrm{d}\tau} \right)^2$$

- Jeffrey's prior is therefore chosen to be:

$$\pi_{\mathrm{J}}(\vartheta) \propto \sqrt{I(\vartheta)}$$

Section 35: Binomially Distributed Data

Binomially Distributed Data

- We consider a Bernoulli experiment repeated n times with probability of success ϑ.
- The number Y of successes is then binomially distributed according to $\mathrm{Bi}(n, \vartheta)$.
- We want to obtain a statement about the success probability ϑ from an observation y.
- To do this, we need an a priori distribution of ϑ.
- The maximum entropy principle yields the uniform distribution $\mathrm{Un}(0, 1) = \mathrm{Be}(1, 1)$ as a a-priori distribution:

$$\pi(\vartheta) = I_{[0,1]}(\vartheta)$$

- likelihood function is equal to.

$$p(y|\vartheta) = \binom{n}{y} \vartheta^y (1-\vartheta)^{n-y}$$

Binomially Distributed Data

- Bayes' theorem provides the a posteriori density:

$$p(\vartheta|y) \propto \vartheta^y (1-\vartheta)^{n-y} I_{[0,1]}(\vartheta)$$

- Since the a posteriori density is proportional to the density of the beta distribution $\mathrm{Be}(y+1, n-y+1)$, it must be identical to it.
- The expected value of the a-posteriori distribution is equal to

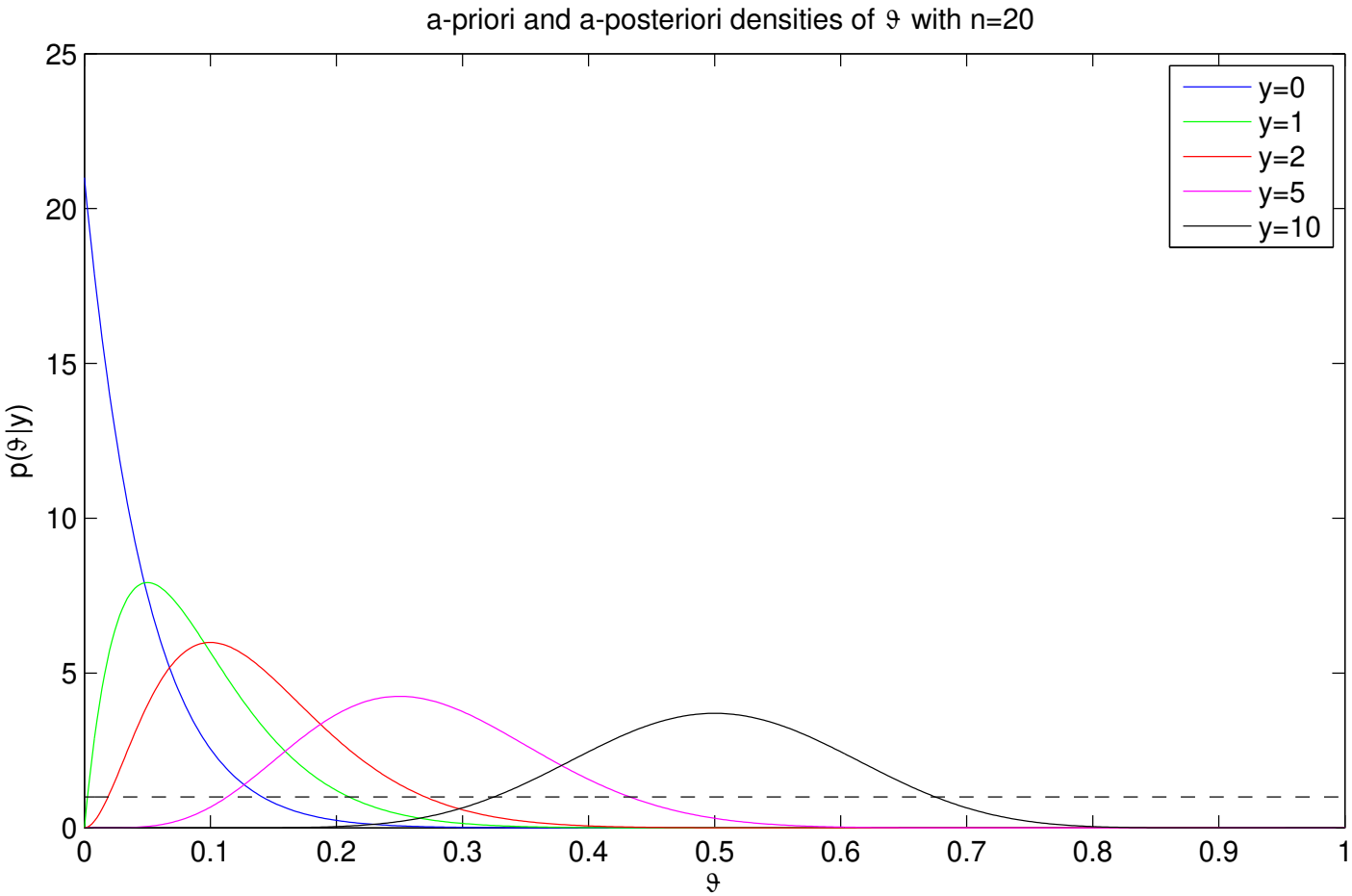
$$\mathsf{E}[\vartheta|y] = \frac{y+1}{n+2}$$

- The mode of the a-posteriori distribution is equal to.

$$\hat{\vartheta} = \frac{y}{n}$$

and thus the maximum likelihood estimator of ϑ.

Binomially Distributed Data



Binomially Distributed Data

- The Fisher information of an observation is equal to

$$I(\vartheta) = \frac{1}{\vartheta(1-\vartheta)}$$

- Jeffrey's prior is thus $\mathrm{Be}(0.5, 0.5)$.
- Bayes' theorem again provides the a posteriori density:

$$p(\vartheta|y) \propto \vartheta^{y-0.5}(1-\vartheta)^{n-y-0.5} I_{[0,1]}(\vartheta)$$

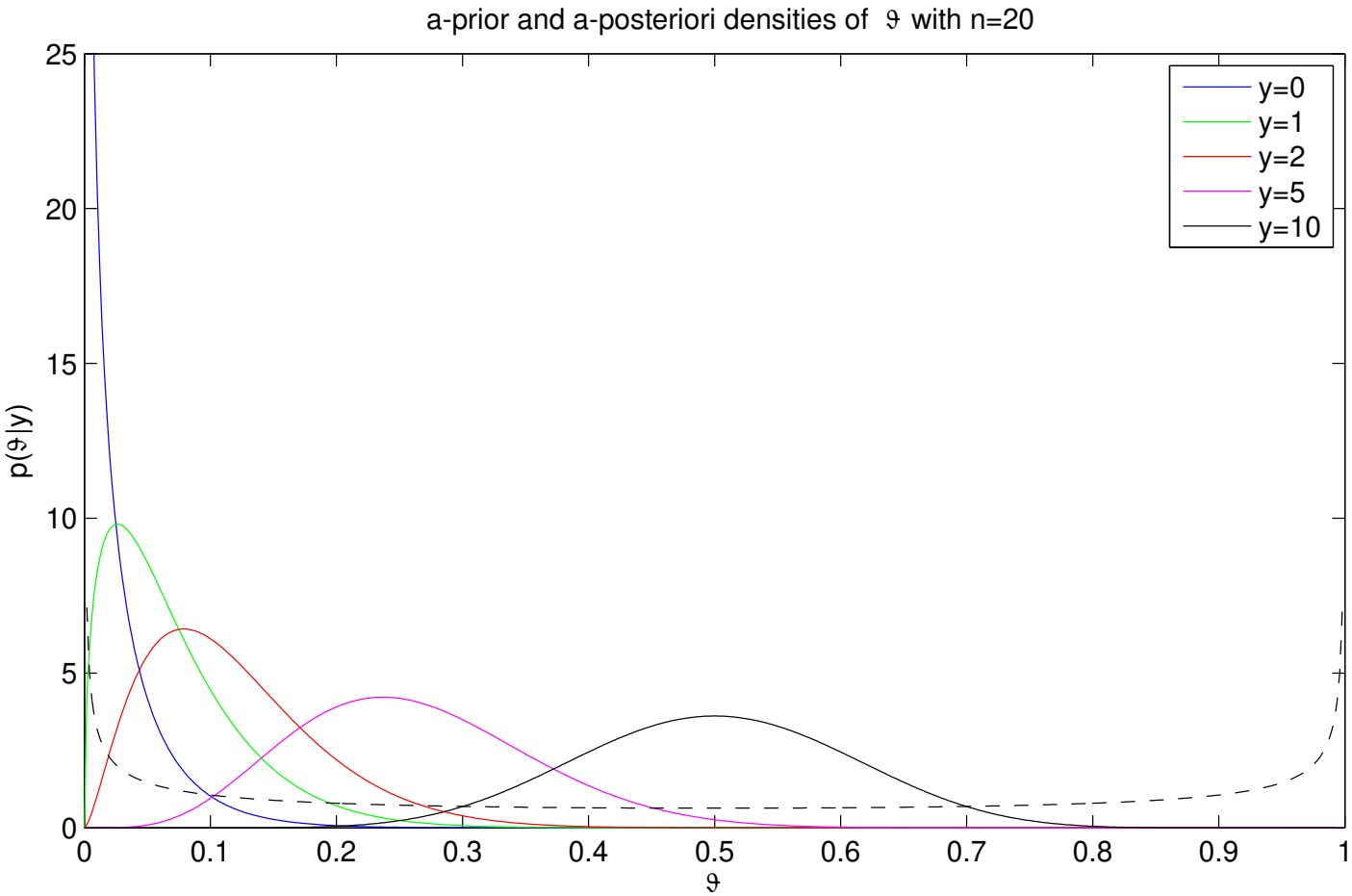
- Since the a posteriori density is proportional to the density of the beta distribution $\mathrm{Be}(y+0.5, n-y+0.5)$, it must be identical to it.
- The expected value of the a-posteriori distribution is equal to

$$\mathsf{E}[\vartheta|y] = \frac{y+0.5}{n+1}$$

- The mode of the a-posteriori distribution is equal to.

$$\hat{\vartheta} = \frac{y-0.5}{n-1}$$

Binomially Distributed Data



Binomially Distributed Data

- If prior information about ϑ is available, it can be included by a suitable a-priori density.
- The computationally simplest treatment results if the a priori distribution is also a beta distribution.
- Let

$$\pi(\vartheta) = \frac{\vartheta^{a-1}(1-\vartheta)^{b-1}}{B(a,b)}$$

- Then the a-posteriori density is equal to.

$$\begin{aligned} p(\vartheta|y) &\propto \vartheta^{y}(1-\vartheta)^{n-y}\vartheta^{a-1}(1-\vartheta)^{b-1} \\ &= \vartheta^{y+a-1}(1-\vartheta)^{n-y+b-1} \end{aligned}$$

- The a posteriori distribution is again a beta distribution, namely $\mathrm{Be}(y+a, n-y+b)$.

Binomially Distributed Data

- Obviously, an a-priori beta distribution with a binomial likelihood function again gives an a-posteriori beta distribution.
- Therefore, the beta distribution is the **conjugate** a-priori distribution to the binomial distribution.
- The expected value of the a-posteriori distribution is equal to

$$\mathsf{E}[\vartheta|y] = \frac{a+y}{a+b+n}$$

- The mode of the a-posteriori distribution is equal to.

$$\hat{\vartheta} = \frac{a+y-1}{a+b+n-2}$$

Binomially Distributed Data

- The expected value of the a posteriori distribution can be written as a weighted average of a priori information and data:

$$\begin{aligned}\mathsf{E}[\vartheta|y] =& \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\ =& \frac{a+b}{a+b+n} \times \text{a-priori expectation} \\ +& \frac{n}{a+b+n} \times \text{average of data}\end{aligned}$$

- a and b can be interpet as "a-priori data":

a	number of successes a-priori
b	number of failures a-priori
$a+b$	number of attempts a-priori

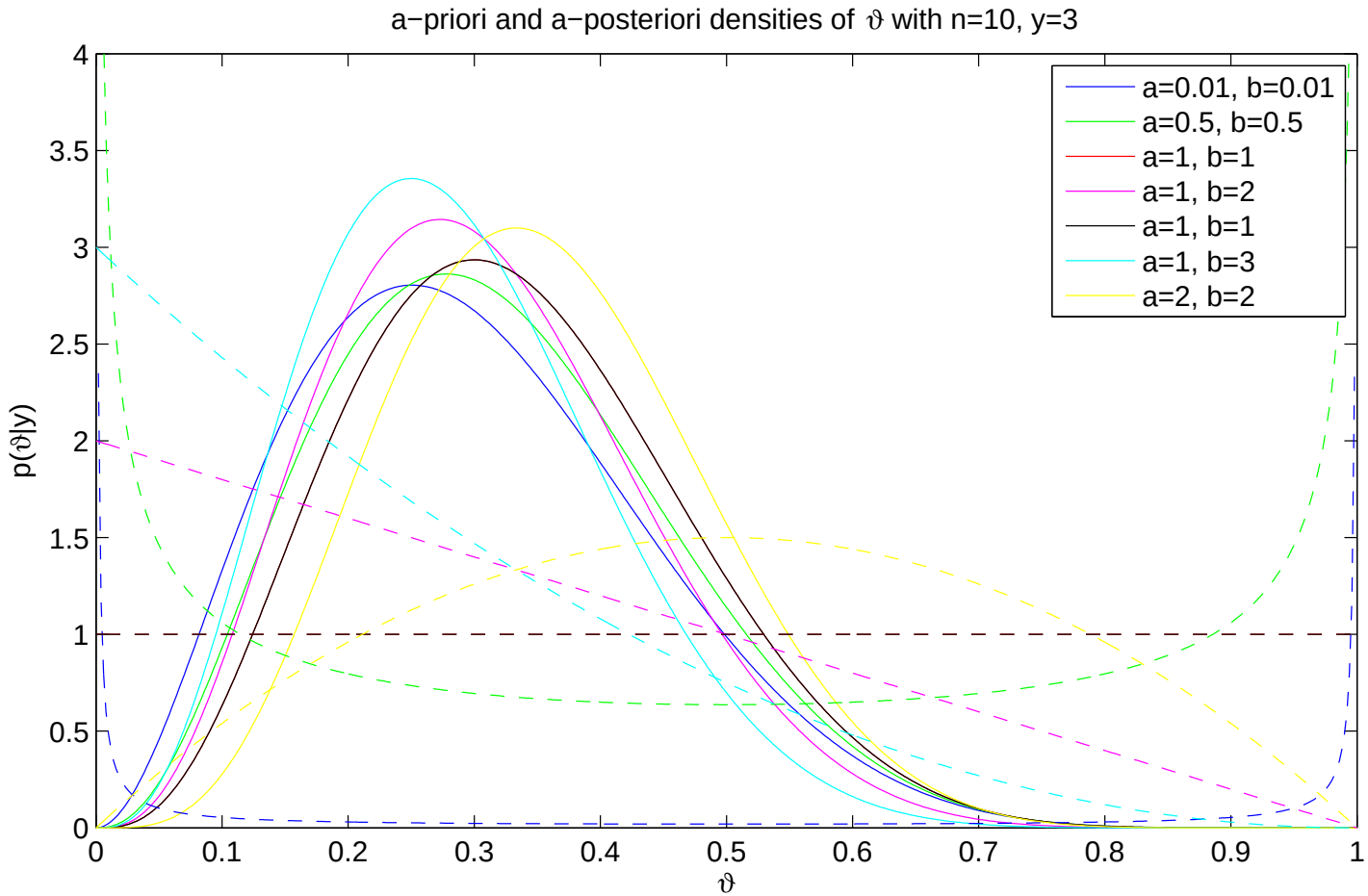
Binomially Distributed Data

- If the number of trials a-priori is set equal to 0, we get **Haldane's prior**:

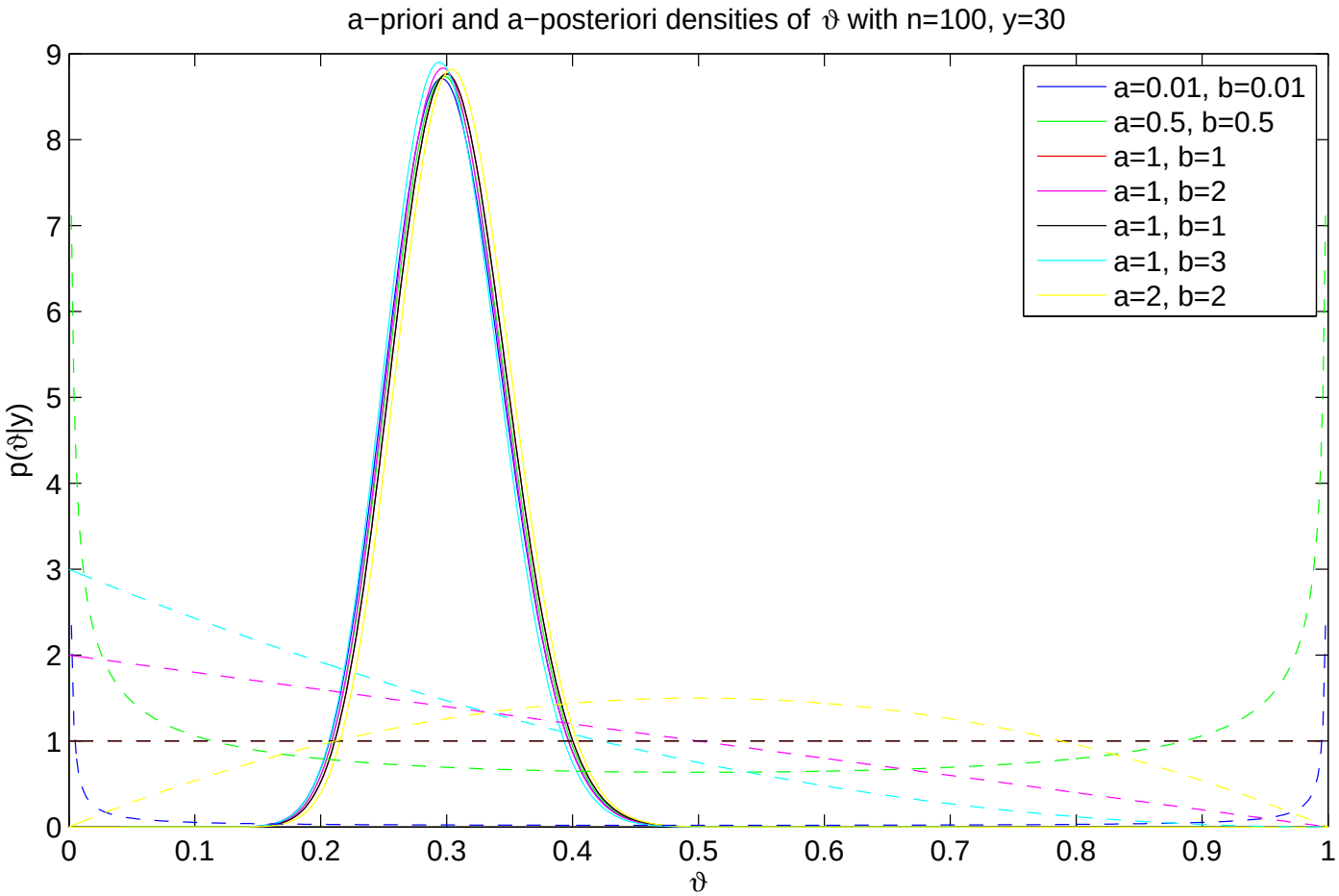
$$\pi_{\mathrm{H}} = \frac{1}{\vartheta(1-\vartheta)}$$

- Haldane's prior can be interpreted as $\mathrm{Be}(0,0)$, but is improper.
- The a posteriori mean is then equal to the ML estimator.
- Paradox: The a-priori distribution without any prior information gives the values $\vartheta = 0$ and $\vartheta = 1$ the highest probability!

Binomially Distributed Data



Binomially Distributed Data



Binomially Distributed Data

- We now want to construct subranges of the parameter space Θ that contain the true value of ϑ with high confidence $1 - \alpha$.
- Such a range is called a **trust range**. It is usually an interval (confidence interval).
- The simplest construction of a confidence interval $[\vartheta_1(y), \vartheta_2(y)]$ uses the quantiles of the a posteriori distribution.
- The symmetric confidence interval is:

$$\vartheta_1(y) = q_{\alpha/2}, \quad \vartheta_1(y) = q_{1-\alpha/2}$$

where q_p is the p-quantile of the a posteriori distribution $p(\vartheta|y)$.

Binomially Distributed Data

Example

Let $y = 4$ be the number of successes in $n = 20$ independent alternative trials with probability of success ϑ. With the uniform distribution as an a priori distribution, the a posteriori distribution of ϑ is a $\mathrm{Be}(5, 17)$ distribution. The symmetric confidence interval with $1 - \alpha = 0.95$ then has the limits

$$\vartheta_1(y) = \beta_{0.025;5,17} = 0.0822$$
$$\vartheta_2(y) = \beta_{0.975;5,17} = 0.4191$$

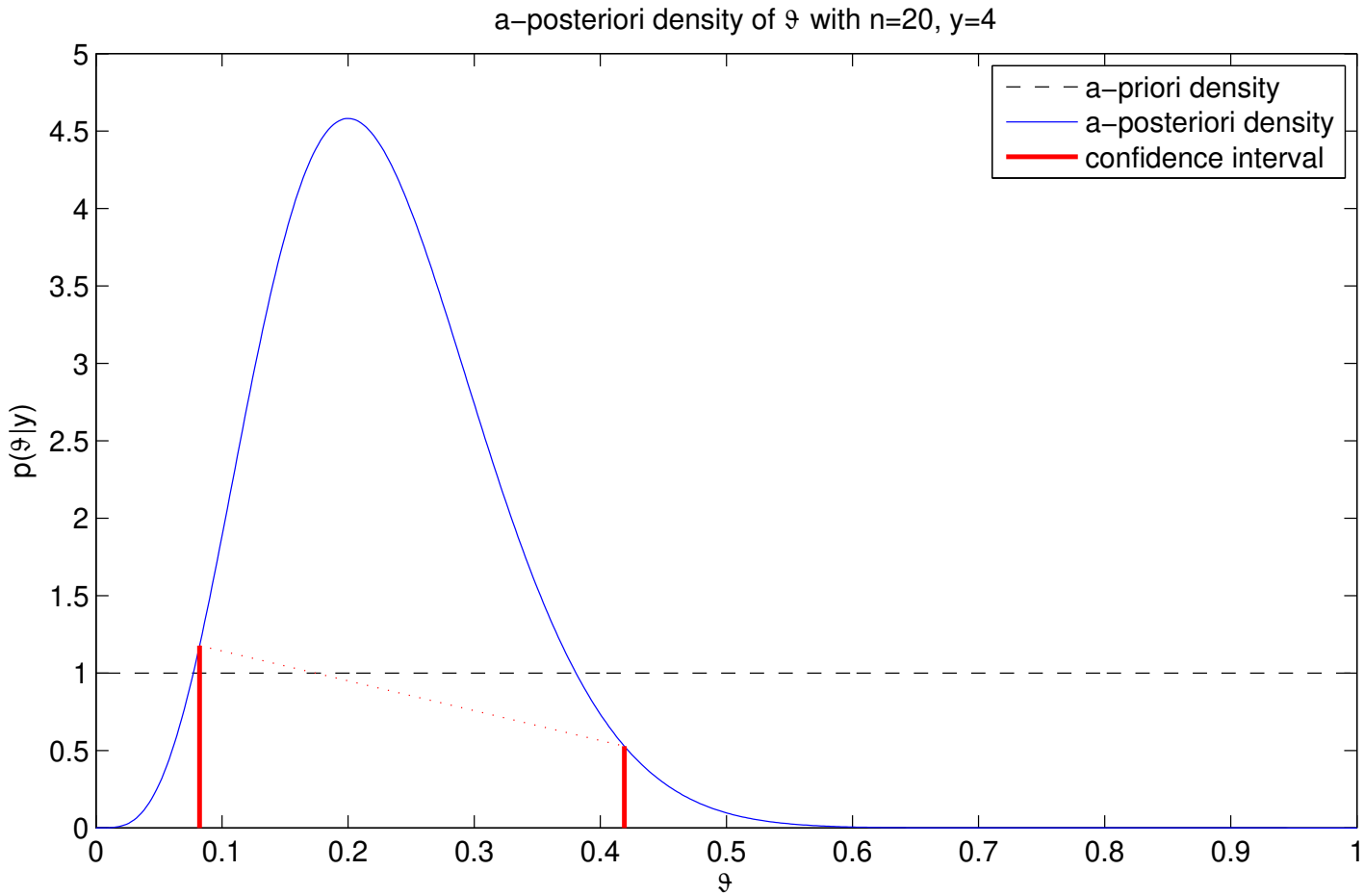
The expected value of the a posteriori distribution is equal to

$$\mathrm{E}[\vartheta|y] = \frac{5}{22} = 0.2273$$

The mode of the a posteriori distribution is the same

$$\hat{\vartheta} = \frac{4}{20} = 0.2$$

Binomially Distributed Data



Binomially Distributed Data

Example (Continuation)

Using Jeffrey's prior, the a posteriori distribution of ϑ is a $\mathrm{Be}(4.5, 16.5)$ distribution. The symmetric confidence interval with $1 - \alpha = 0.95$ then has the limits

$$\vartheta_1(y) = \beta_{0.025;4.5,16.5} = 0.0715$$
$$\vartheta_2(y) = \beta_{0.975;4.5,16.5} = 0.4082$$

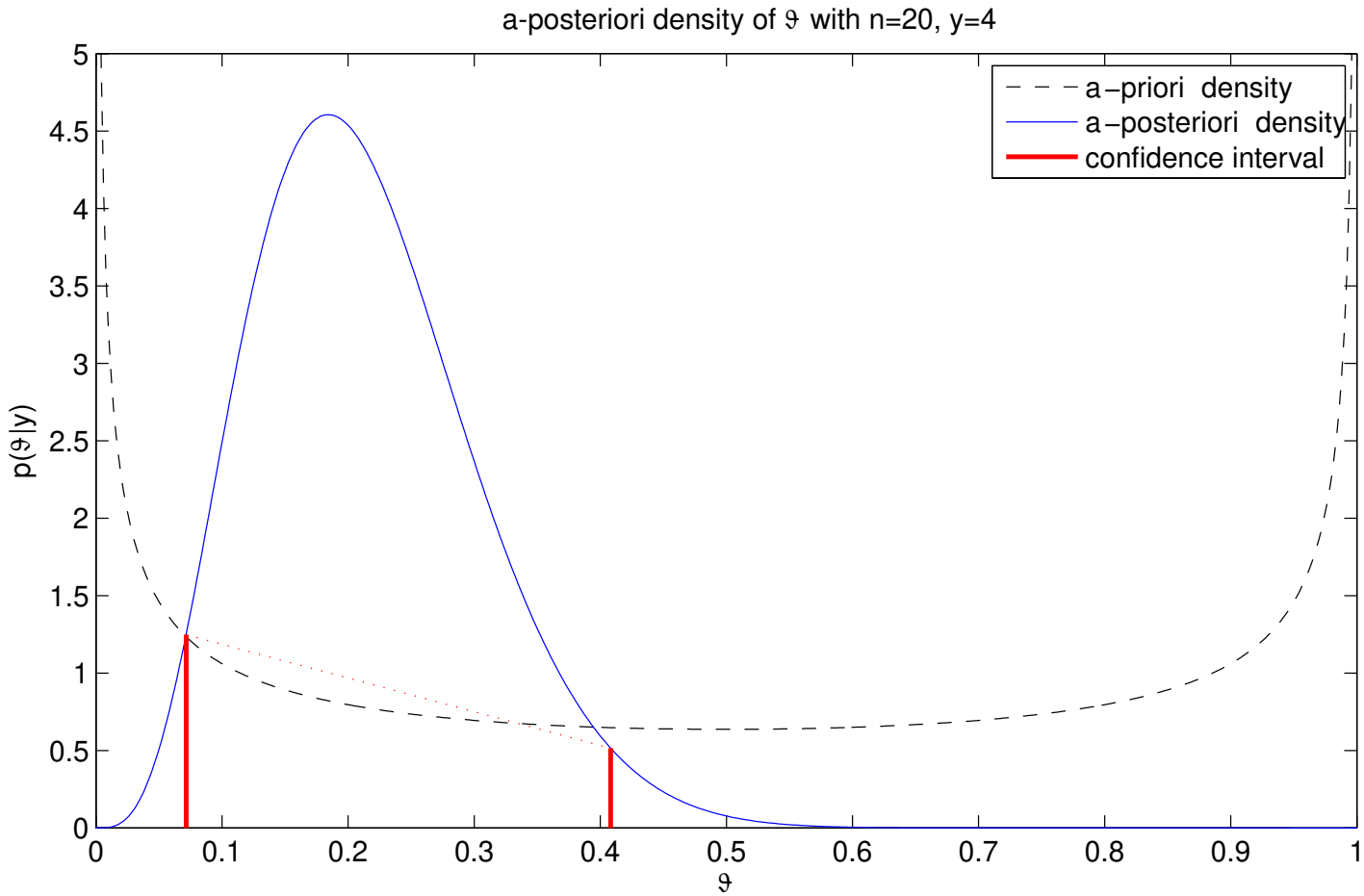
The expected value of the a posteriori distribution is equal to

$$\mathrm{E}[\vartheta|y] = \frac{4.5}{21} = 0.2143$$

The mode of the a posteriori distribution is the same

$$\hat{\vartheta} = \frac{3.5}{19.5} = 0.1795$$

Binomially Distributed Data



Binomially Distributed Data

- The symmetric confidence interval contains values of ϑ that have lower a-posteriori probability than some points outside the interval.
- A range in which all values of ϑ have higher a-posteriori probability than all values outside the range is called a high-posterior density range, or **HPD** range for short.
- If the a-posteriori density is unimodal, the HPD range is an HPD interval.
- In this case, the ϑ_1, ϑ_2 bounds of the HPD interval are obtained as the solution of the system of equations:

$$p(\vartheta_2|y) - p(\vartheta_1|y) = 0$$
$$P(\vartheta_2|y) - P(\vartheta_1|y) = 1 - \alpha$$

Here $P(\vartheta|y)$ is the a posteriori distribution function.

Binomially Distributed Data

- The system of equations must be solved numerically.

Example (Continuation)

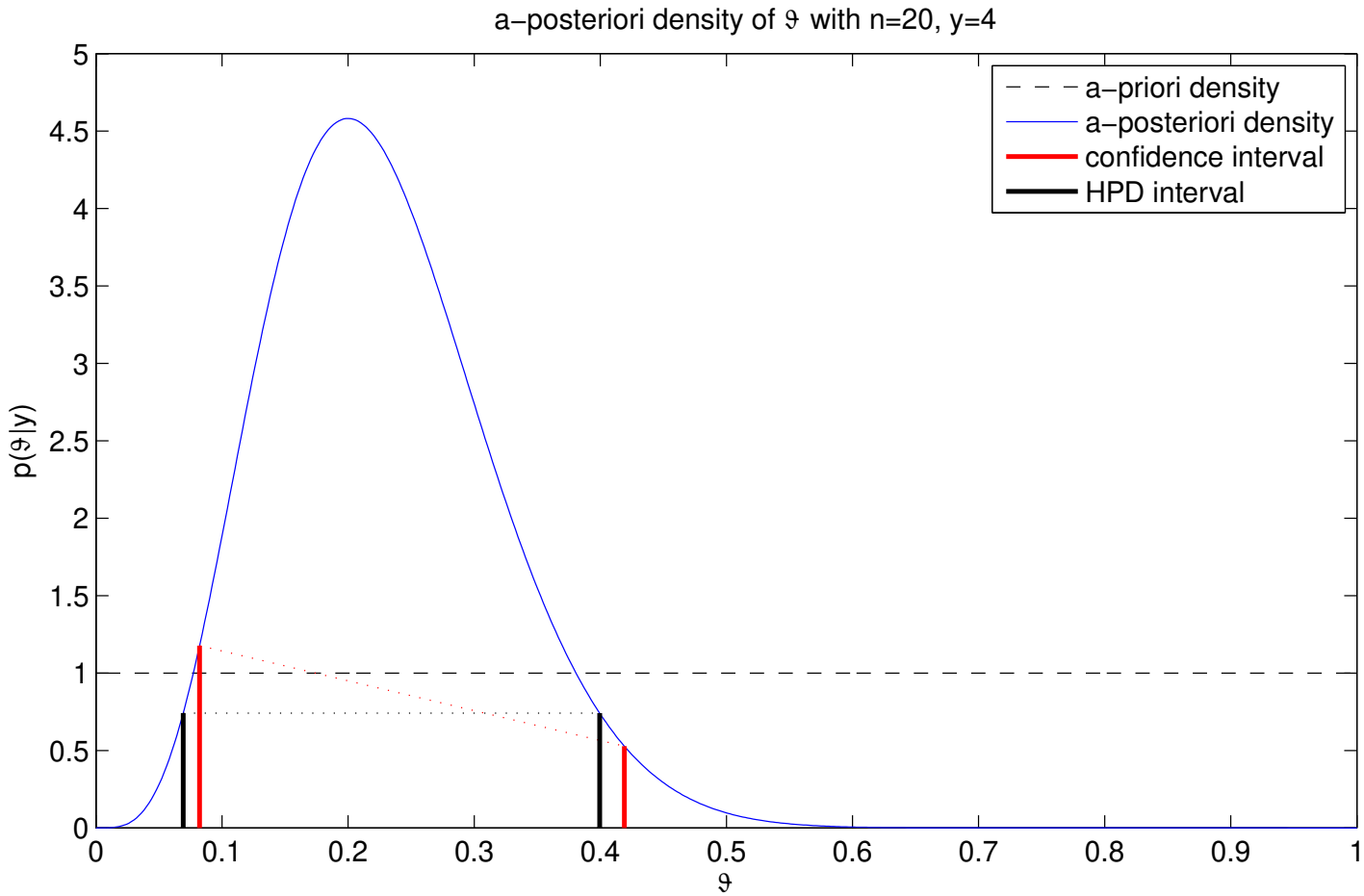
The HPD interval with $1 - \alpha = 0.95$ has the limits

$$\vartheta_1(y) = 0.06921, \quad \vartheta_1(y) = 0.3995$$

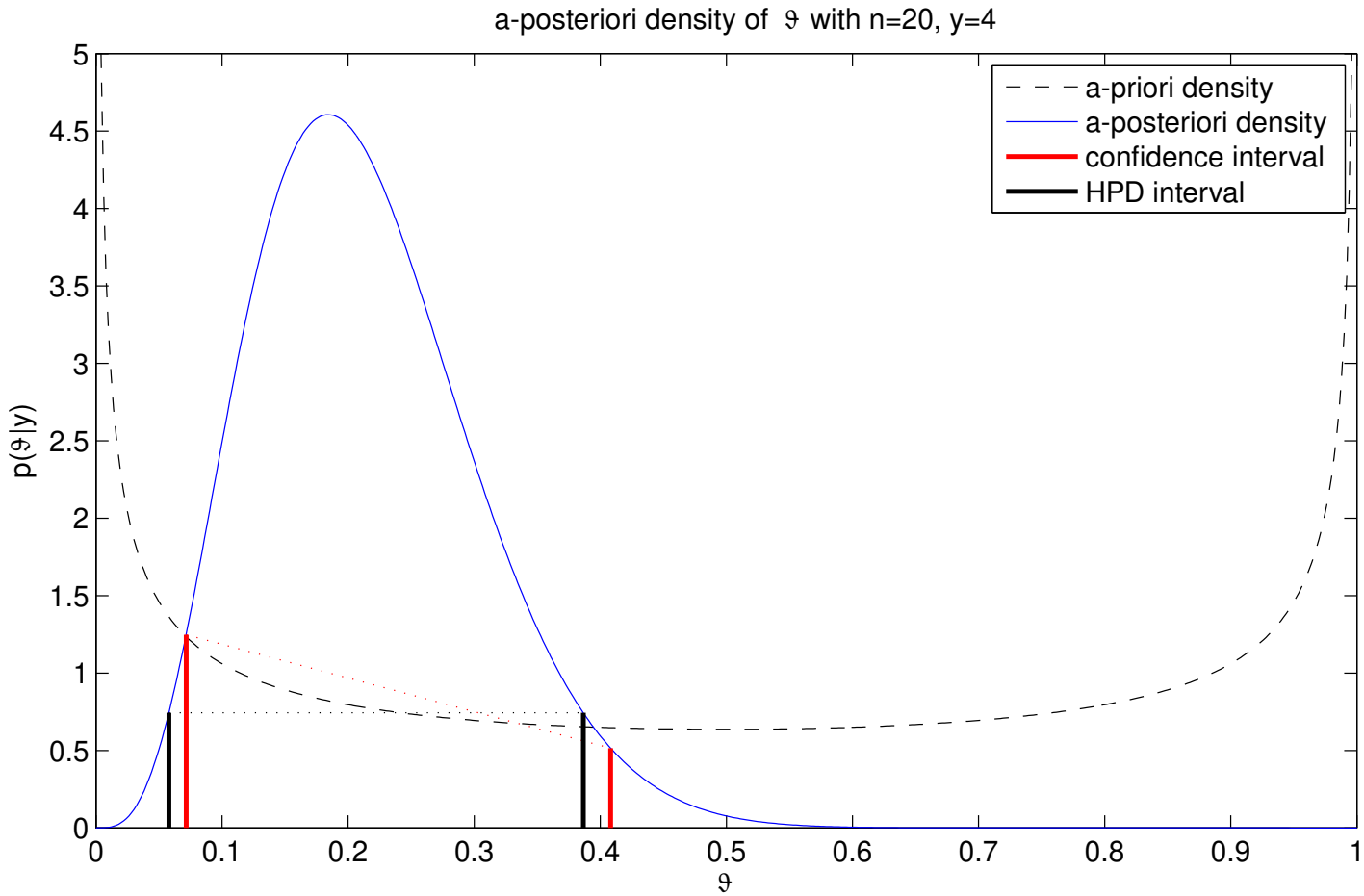
With a length of 0.3303, it is shorter than the symmetric confidence interval, which has a length of 0.3369.

MATLAB: `make_posterior_binomial`

Binomially Distributed Data



Binomially Distributed Data



Binomially Distributed Data

Example

An alternative trial is repeated $n = 20$ times and does not yield a single success ($k = 0$). What can be said about the probability of success ϑ?

- With a priori density $\pi(\vartheta) = 1$, the a posteriori distribution is $\mathrm{Be}(1, 21)$. The mode is equal to 0, and the expected value is equal to 0.0455. The HPD interval with $1 - \alpha = 0.95$ is equal to $[0, 0.1329]$.
- With Jeffrey's prior, the a posteriori distribution is $\mathrm{Be}(0.5, 20.5)$. The mode is equal to 0, and the expected value is equal to 0.0238. The HPD interval with $1 - \alpha = 0.95$ is equal to $[0, 0.0905]$.
- If it is known that ϑ is rather small, e.g. $\mathrm{Be}(0.5, 5)$ can be chosen as a priori distribution. The a posteriori distribution is then $\mathrm{Be}(0.5, 25)$. The mode is equal to 0, the expected value is equal to 0.0196, and the HPD interval is equal to $[0, 0.0747]$.

Binomially Distributed Data

Example (Continuation)

- The likelihood estimator of ϑ is equal to 0.
- The one-sided Clopper-Pearson confidence interval is equal to $[0, 0.1391]$.
- The approximation by normal distribution is only useful with the correction according to Agresti-Coull, otherwise the confidence interval shrinks to zero. The estimate is $\hat{\vartheta} = 0.0833$, the symmetric confidence interval is $[-0.0378, 0.2045]$, the left-hand confidence interval is $[0, 0.1850]$.

Section 36: Poisson Distributed Data

Poisson Distributed Data

- Observing a Poisson process yields the values $\boldsymbol{y} = y_1, \ldots, y_n$.
- We want to obtain an estimate of the intensity λ of the process from the data.
- The likelihood function is:

$$p(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \propto \lambda^{\sum y_i} e^{-n\lambda}$$

- It depends on the data only via $s = \sum y_i$.
- As non-informative a-priori distributions can be considered:
  - The improper density $\pi(\lambda) = 1$
  - Jeffrey's prior $\pi_{\mathrm{J}}(\lambda) = \lambda^{-1/2}$, also improper

Poisson Distributed Data

- With $\pi(\lambda) = 1$, the a posteriori density is proportional to the likelihood function:

$$p(\lambda|s) \propto \lambda^s e^{-n\lambda}$$

- Since the a posteriori density is proportional to the density of the gamma distribution $\mathrm{Ga}(s+1, 1/n)$, it must be identical to it:

$$p(\lambda|s) = \frac{\lambda^s e^{-n\lambda}}{n^{-(s+1)}\Gamma(s+1)}$$

- The expected value of the a posteriori distribution is equal to.

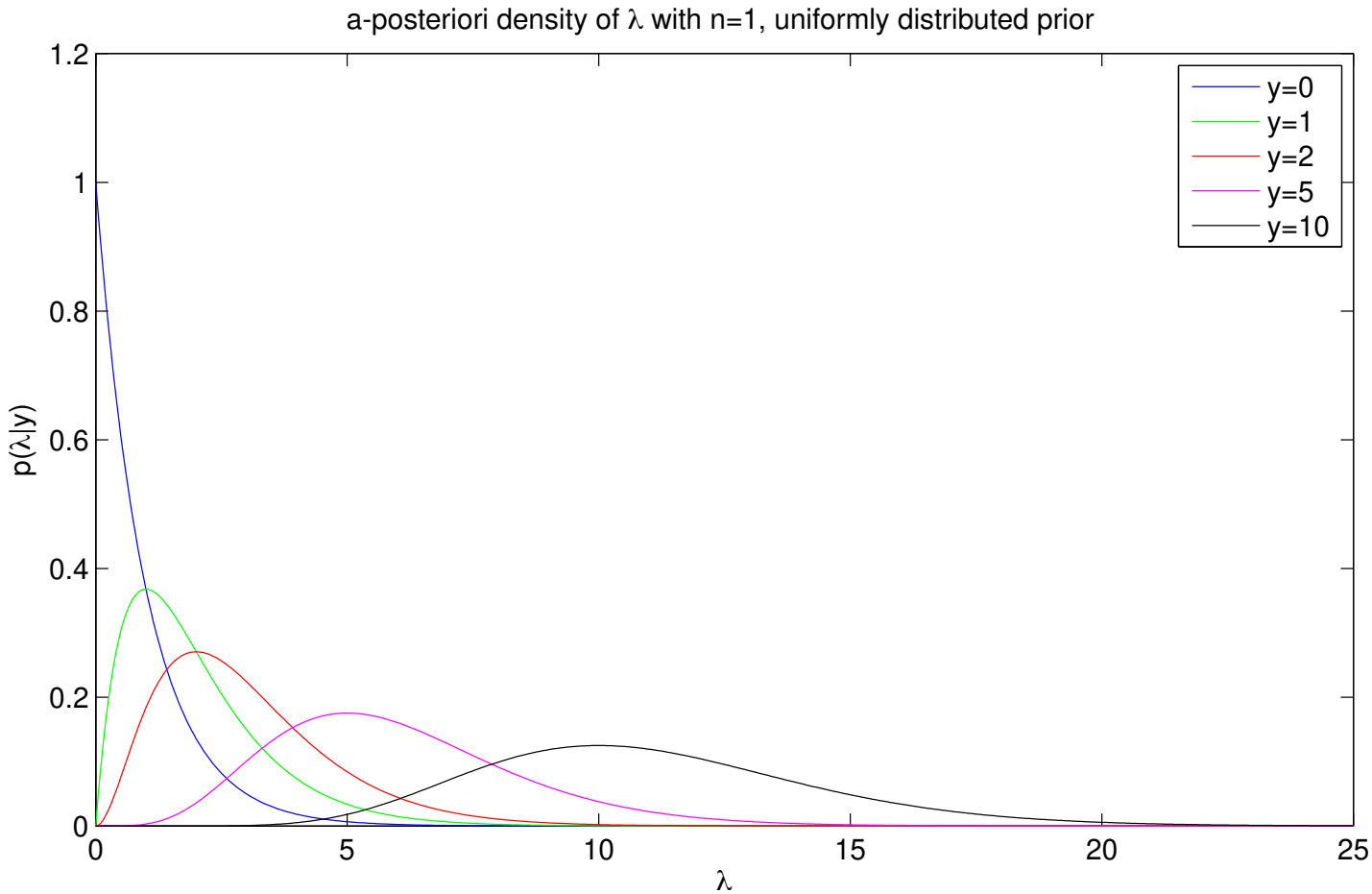
$$\mathrm{E}[\lambda|s] = \frac{s+1}{n}$$

- The mode of the a posteriori distribution is equal to.

$$\hat{\lambda} = \frac{s}{n}$$

and thus the maximum likelihood estimator of λ.

Poisson Distributed Data



Poisson Distributed Data

- With $\pi(\lambda) = \lambda^{-1/2}$ the a posteriori density is:

$$p(\lambda|s) \propto \lambda^{s-0.5} e^{-n\lambda}$$

- Since the a-posteriori density is proportional to the density of the gamma distribution $\mathrm{Ga}(s+0.5, 1/n)$, it must be identical to it:

$$p(\lambda|s) = \frac{\lambda^{s-0.5} e^{-n\lambda}}{n^{-(s+0.5)}\Gamma(s+0.5)}$$

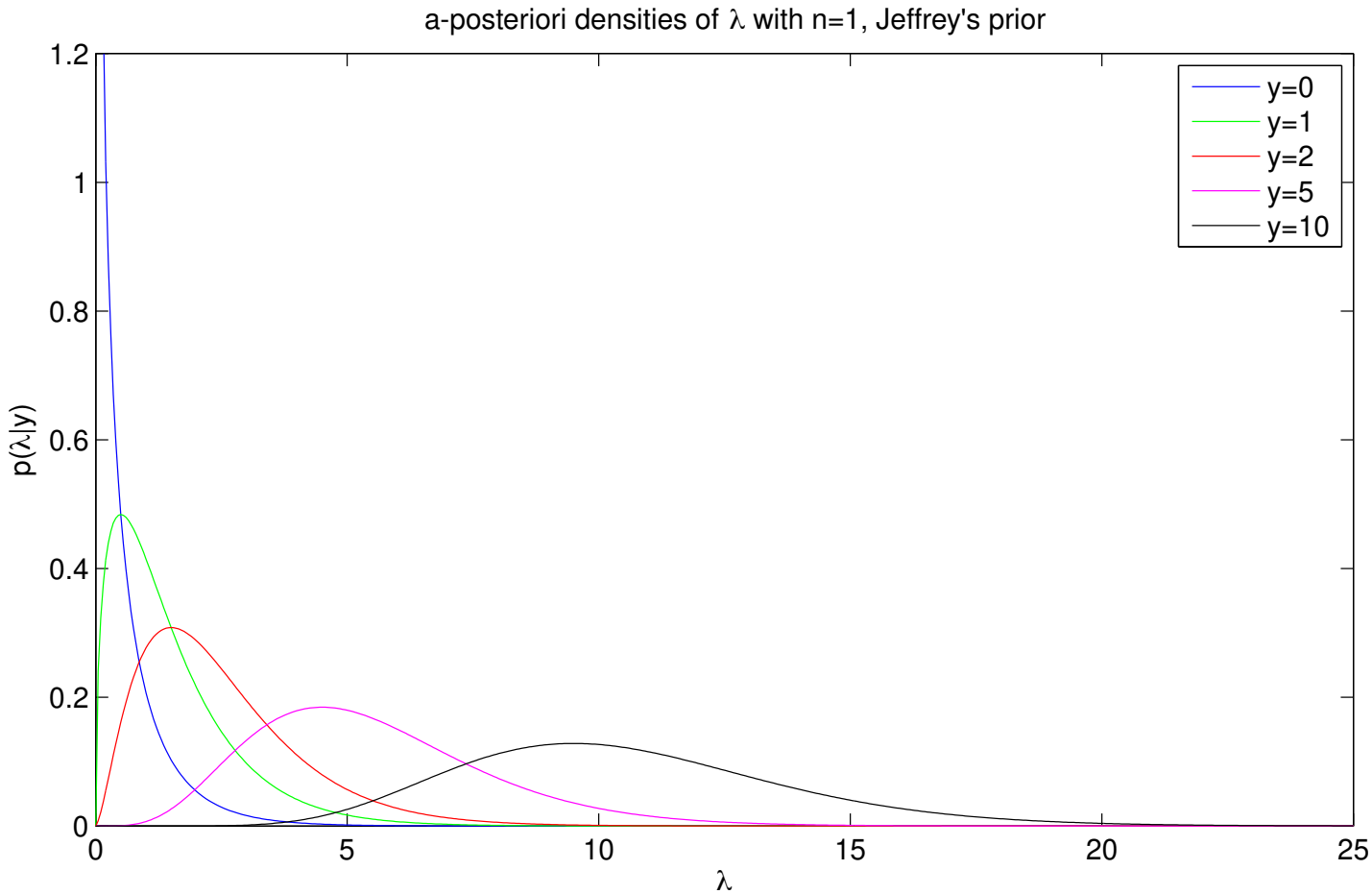
- The expected value of the a-posteriori distribution is equal to.

$$\mathsf{E}[\lambda|s] = \frac{s+0.5}{n}$$

- The mode of the a posteriori distribution is equal to.

$$\hat{\lambda} = \frac{s-0.5}{n}$$

Poisson Distributed Data



Poisson Distributed Data

- If prior information about λ is available, it can be included by a suitable a-priori density.
- The computationally simplest treatment results when the a priori distribution is a gamma distribution.
- Let

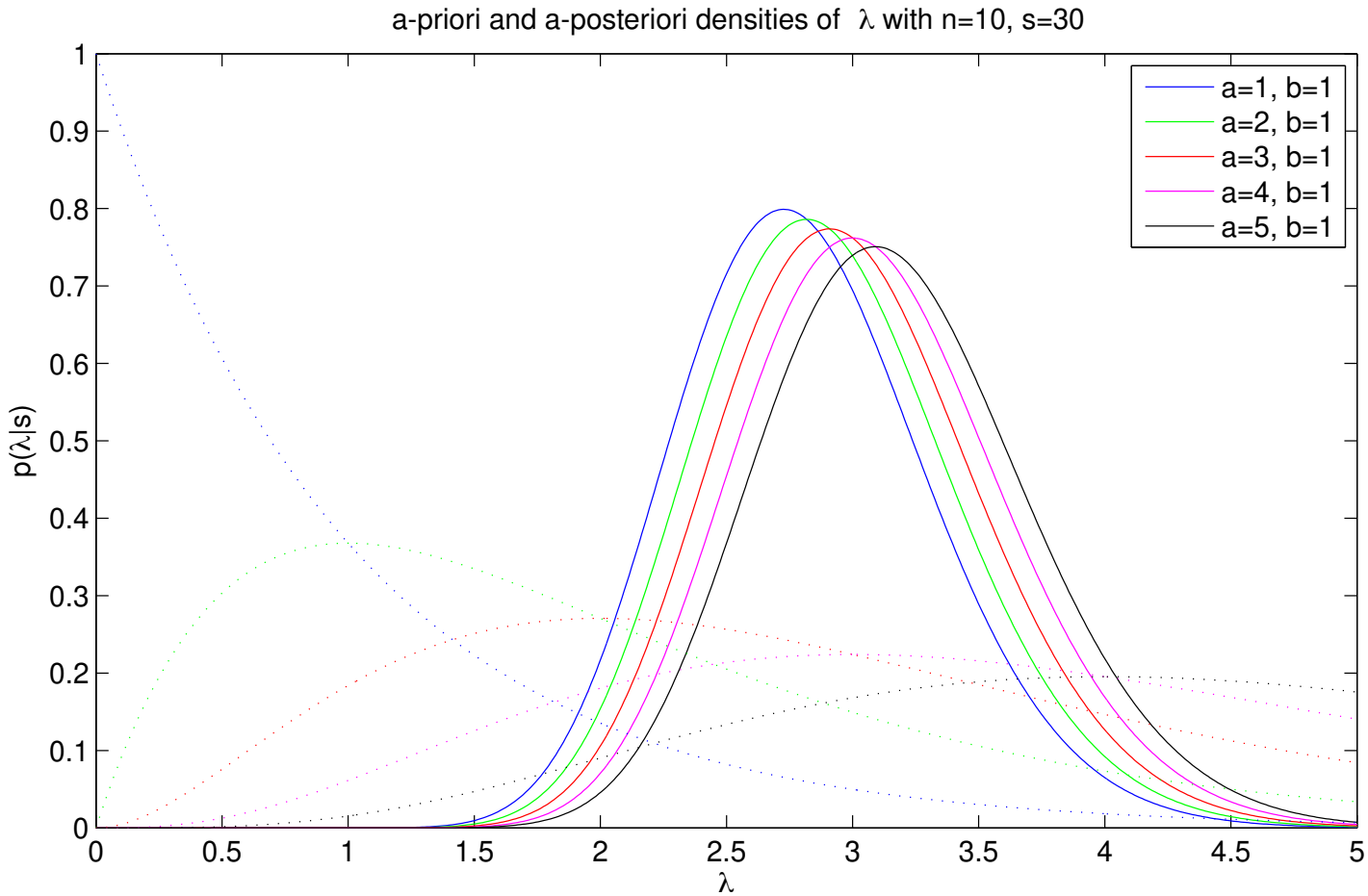
$$\pi(\lambda) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$$

  This is the density of the gamma distribution $\mathrm{Ga}(a, 1/b)$.
- Then the a-posteriori density is equal to

$$\begin{aligned} p(\lambda|s) &\propto \lambda^s e^{-n\lambda} \lambda^{a-1} e^{-b\lambda} \\ &= \lambda^{s+a-1} e^{-(b+n)\lambda} \end{aligned}$$

- The a posteriori distribution is the gamma distribution $\mathrm{Ga}(a+s, 1/(b+n))$. Thus, the gamma distribution is conjugate to the Poisson distribution.

Poisson Distributed Data



Poisson Distributed Data

- The expected value of the a-posteriori distribution is equal to

$$\mathsf{E}[\lambda|s] = \frac{a+s}{b+n}$$

- The mode of the a-posteriori distribution is equal to.

$$\hat{\lambda} = \frac{a+s-1}{b+n}$$

- The expected value of the a-posteriori distribution can be written as a weighted average of a-priori information and data:

$$\begin{aligned}\mathsf{E}[\lambda|s] = \frac{a+s}{b+n} &= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{s}{n} \\ &= \frac{b}{b+n} \times \text{a-priori expectation} \\ &+ \frac{n}{b+n} \times \text{average of data}\end{aligned}$$

Poisson Distributed Data

- The a-priori parameters a and b can be interpetted as follows:

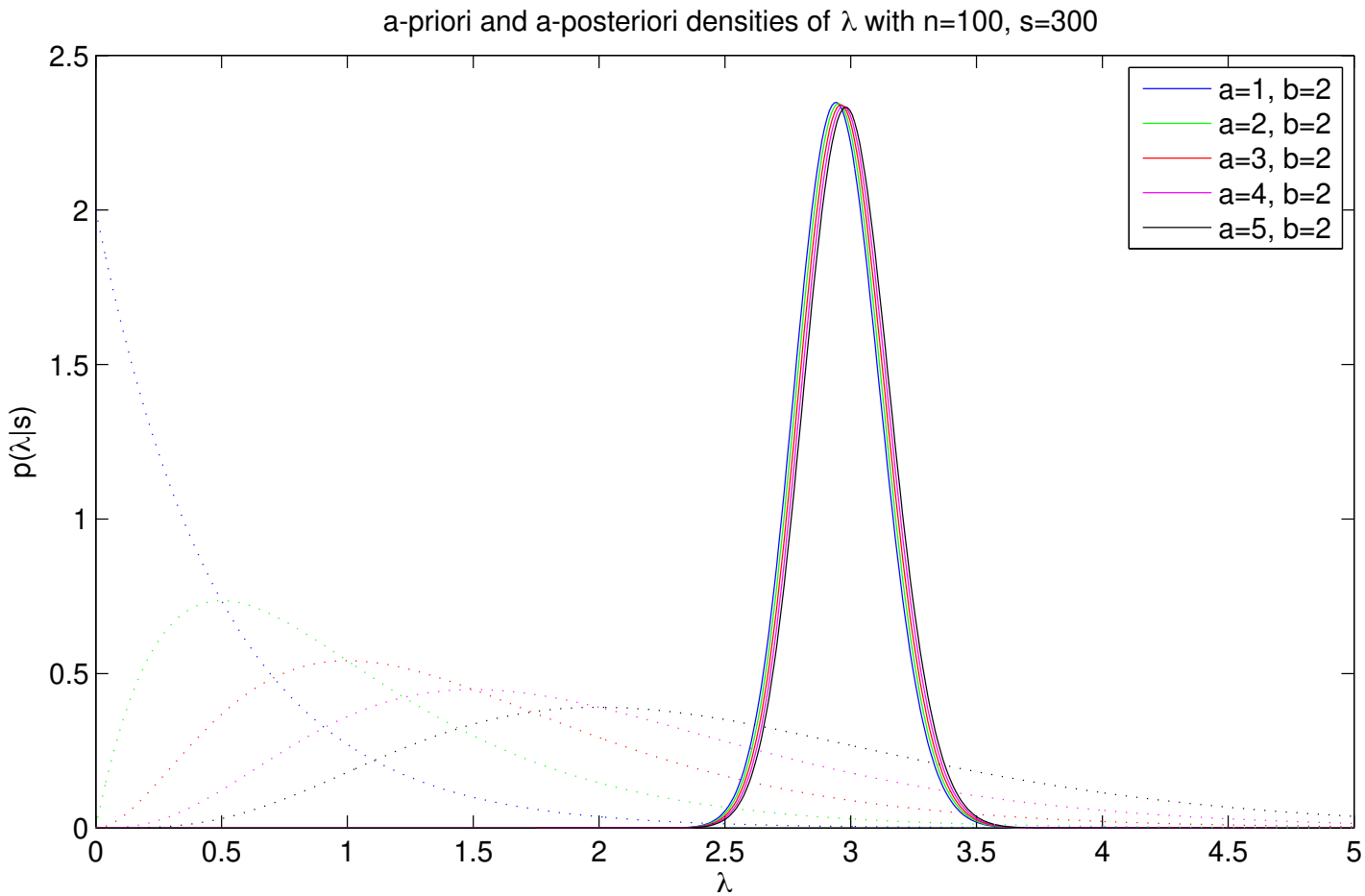
  a sum of data a-priori

  b number of data a-priori

- If $b \ll n$, dominate the data:

$$\mathsf{E}[\lambda|s] \approx \frac{s}{n}$$

Poisson Distributed Data



Poisson Distributed Data

- Confidence intervals $[\lambda_1(s), \lambda_2(s)]$ can be easily constructed using the quantiles of the a posteriori gamma distribution.
- The symmetric confidence interval is equal to

$$\left[\gamma_{\alpha/2;a+s,1/(b+n)}, \gamma_{1-\alpha/2;a+s,1/(b+n)}\right]$$

- The one-sided confidence intervals are.

$$[0, \gamma_{\alpha;a+s,1/(b+n)}] \quad \text{bzw.} \quad [\gamma_{\alpha;a+s,1/(b+n)}, \infty]$$

- HPD ranges can be determined using numerical methods.

Poisson Distributed Data

Example

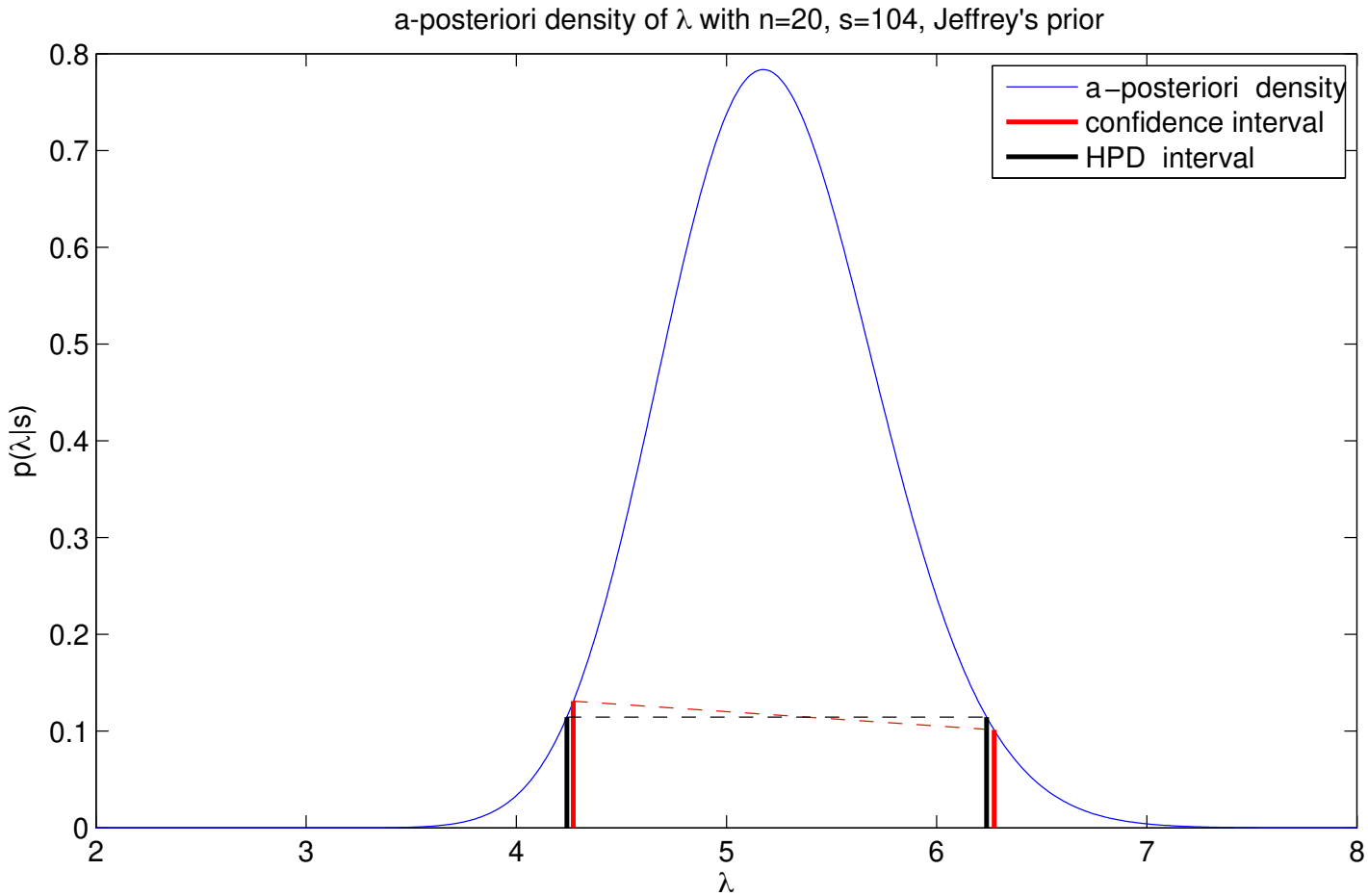
You measure the background radiation in a laboratory over a period of 20 seconds. The count values are

$$6, 2, 6, 1, 6, 8, 5, 3, 8, 4, 2, 5, 7, 8, 5, 4, 7, 9, 4, 4$$

Their sum is equal to $s = 104$. Using Jeffrey's prior, the a posteriori distribution is the gamma distribution $\mathrm{Ga}(104.5, 0.05)$. Its expectation is 5.225, and its mode is 5.1750. The symmetric confidence interval with $1 - \alpha = 0.95$ is $[4.2714, 6.2733]$, the HPD interval is $[4.2403, 6.2379]$. Since the distribution is almost symmetric, the two intervals are practically the same length.

MATLAB: `make_posterior_poisson`

Poisson Distributed Data



Poisson Distributed Data

Example

An observation from a Poisson distribution has the value $k = 0$. What can be said about the mean λ?

- With a priori density $\pi(\lambda) = 1$, the a posteriori distribution is $\mathrm{Ga}(1, 1)$. The mode is 0, and the mean is 1. The HPD interval with $1 - \alpha = 0.95$ is $[0, 2.9957]$.
- With Jeffrey's prior, the a posteriori distribution is $\mathrm{Ga}(0.5, 1)$. The mode is 0, and the mean is 0.5. The HPD interval with $1 - \alpha = 0.95$ is $[0, 1.9207]$.
- If it is known that λ is significantly smaller than 1, for example, $\mathrm{Ga}(0.5, 0.5)$ can be chosen as the a priori distribution. The a-posteriori distribution is then $\mathrm{Ga}(0.5, 0.6667)$. The mode is 0, the mean is 0.3333, the HPD interval is $[0, 1.2805]$.
- The likelihood estimator of λ is 0.
- The left-hand confidence interval is $[0, 2.9957]$, so it is identical to the HPD interval with improper a priori density.

Section 37: Normally Distributed Data

Normally Distributed Data

- We consider a normally distributed sample $\boldsymbol{y} = (y_1, \ldots, y_n)$.
- We want to gain an estimate of the mean and variance of the distribution from which the data are drawn.
- We first assume that the variance σ^2 is known.
- Then the likelihood function is:

$$p(\boldsymbol{y}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \mu)^2}{2\,\sigma^2}\right]$$

- In the absence of any prior information about the actual value of μ, we choose the improper a priori density $\pi(\mu) = 1$, which is also Jeffrey's prior.

Normally Distributed Data

- Then the a-posteriori density is equal to

$$p(\mu|\boldsymbol{y},\sigma^2) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i-\mu)^2}{2\,\sigma^2}\right]$$
$$\propto \exp\left[-\frac{n(\mu-\bar{y})^2}{2\,\sigma^2}\right]$$

- Since the a posteriori density is proportional to the density of the normal distribution $\mathrm{No}(\bar{y},\sigma^2/n)$, it must be identical to it.
- The expected value of the a-posteriori distribution is equal to

$$\mathrm{E}[\mu|\boldsymbol{y}] = \bar{y}$$

  and thus the maximum likelihood estimator of μ.
- The mode is also equal to $\bar{y}$.

Normally Distributed Data

- If prior information about μ is available, it can be included by a suitable a-priori density.
- The computationally simplest treatment results if the a priori distribution is also a normal distribution.
- Let

$$\pi(\mu|\sigma^2) = \frac{1}{\sqrt{2\pi}\tau_0} \exp\left[-\frac{(\mu-\mu_0)^2}{2\,\tau_0^2}\right]$$

- Then the a-posteriori density is equal to.

$$p(\mu|\boldsymbol{y},\sigma^2) \propto \exp\left[-\frac{n(\mu-\bar{y})^2}{2\,\sigma^2}\right] \exp\left[-\frac{(\mu-\mu_0)^2}{2\,\tau_0^2}\right]$$
$$\propto \exp\left[-\frac{a(\mu-b/a)^2}{2}\right]$$

with

Normally Distributed Data

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad \text{und} \quad b = \frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}$$

- The a posteriori distribution is therefore the normal distribution $\mathrm{No}(\mu_n, \tau_n^2)$ with

$$\mu_n = \frac{b}{a} = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

- The mean μ_n is the weighted average of the prior information μ_0 and the mean of the data $\bar{y}$, where the weights are given by the **precision**, i.e. the inverse variance.

Normally Distributed Data

- The precision $1/\tau_n^2$ is the sum of the precision of the prior information μ_0 and the precision of the mean value of the data $\bar{y}$.
- The smaller the precision of the prior information is, i.e. the larger τ_0^2 is, the smaller is the influence of the prior information on the a posteriori distribution.

Normally Distributed Data

- If the variance σ^2 is unknown, we need a joint a priori density for μ and σ^2.
- The calculation is simpler if we use the precision $\zeta = 1/\sigma^2$ instead of the variance σ^2.
- A possible choice is the impropere a-priori density

$$\pi(\mu, \zeta) = \frac{1}{\zeta}$$

- Then:

$$p(\mu, \zeta | \boldsymbol{y}) \propto \frac{1}{\zeta} \prod_{i=1}^{n} \sqrt{\zeta} \exp\left[-\frac{\zeta}{2} \sum_{i=1}^{n} (y_i - \mu)^2\right]$$

$$\propto \zeta^{n/2-1} \exp\left[-\frac{\zeta}{2} \sum_{i=1}^{n} (y_i - \mu)^2\right]$$

Normally Distributed Data

- integration over μ gives the a posteriori density of ζ:

$$p(\zeta|\boldsymbol{y}) \propto \int \zeta^{n/2-1} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \mathrm{d}\mu$$

$$\propto \zeta^{(n-3)/2} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

- ζ is therefore a-posteriori gamma distributed according to $\mathrm{Ga}((n-1)/2, 2/\sum(y_i-\bar{y})^2)$. Because of

$$\sum(y_i-\bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

the distribution depends only on $s_1 = \sum y_i$ and $s_2 = \sum y_i^2$ resp. depends only on the sample mean $\bar{y}$ and the sample variance S^2.

Normally Distributed Data

- The a posteriori density of σ^2 is obtained by the transformation $\sigma^2 = 1/\zeta$:

$$g(\sigma^2) = \frac{p(1/\sigma^2|\boldsymbol{y})}{\sigma^4}$$

- The distribution of σ^2 is an inverse gamma distribution.

The Inverse Gamma Distribution $\mathrm{IG}(a, b)$

- The density of the inverse gamma distribution is:

$$f_{\mathrm{IG}}(x; a, b) = \frac{(1/x)^{a+1} e^{-1/(xb)}}{b^a \Gamma(a)} \cdot I_{[0,\infty)}(x)$$

- Its distribution function is:

$$F_{\mathrm{IG}}(x; a, b) = 1 - F_{\mathrm{Ga}}(1/x; a, b)$$

- The mean is $1/(b(a-1))$ when $a > 1$; the mode is $m = 1/(b(a+1))$.

Normally Distributed Data

- We obtain the a posteriori distribution of μ conditioned by ζ from

$$
\begin{aligned}
p(\mu|\zeta,\boldsymbol{y}) &= \frac{p(\mu,\zeta|\boldsymbol{y})}{p(\zeta|\boldsymbol{y})} \\
&\propto \zeta^{0.5}\exp\left\{-\frac{\zeta}{2}\left[\sum(y_i-\mu)^2-\sum(y_i-\bar{y})^2\right]\right\} \\
&\propto \zeta^{0.5}\exp\left[-\frac{n\zeta}{2}(\mu-\bar{y})^2\right] \\
&= \frac{1}{\sigma}\exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right]
\end{aligned}
$$

- This is the normal distribution $\mathrm{No}(\bar{y},\sigma^2/n)$.

Normally Distributed Data

- Finally, we obtain the a posteriori distribution of μ by integrating over ζ:

$$p(\mu|\boldsymbol{y}) \propto \int \zeta^{n/2-1} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \mathrm{d}\zeta$$

$$\propto \frac{1}{\left[\sum(\mu-y_i)^2\right]^{n/2}}$$

- It follows by transformation that the default score is.

$$t = \frac{\mu - \bar{y}}{S/\sqrt{n}}$$

a-posteriori t-distributed with $n-1$ degrees of freedom is.

Normally Distributed Data

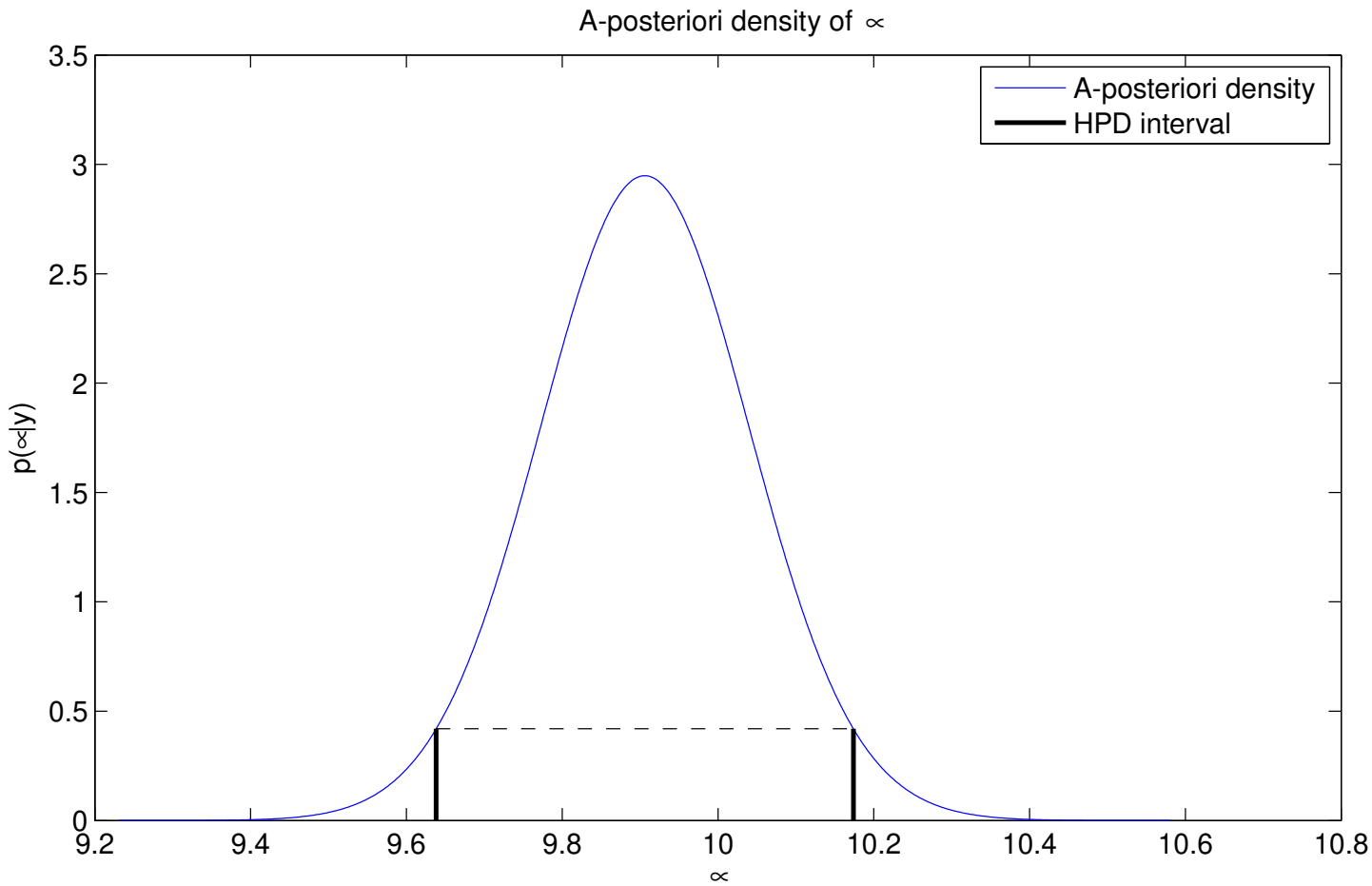
- If one wants an actual a priori density for ζ, the gamma distribution is a good choice, since it is the conjugate distribution.
- The a posteriori distribution $p(\zeta|\boldsymbol{y})$ is then again a gamma distribution.

Example

From a normally distributed data set of size $n = 100$ from $\mathrm{No}(10, \sqrt{2})$, $\bar{y} = 9.906$ and $S = 1.34981$ are given. The a posteriori density of μ is a $\mathrm{t}(n-1)$ distribution scaled by $S/\sqrt{n}$ and shifted by $\bar{y}$. The HPD interval with $1 - \alpha = 0.95$ is equal to $[9.6382, 10.1739]$.

MATLAB: `make_posterior_normal`

Normally Distributed Data



Normally Distributed Data

Example (Continuation)

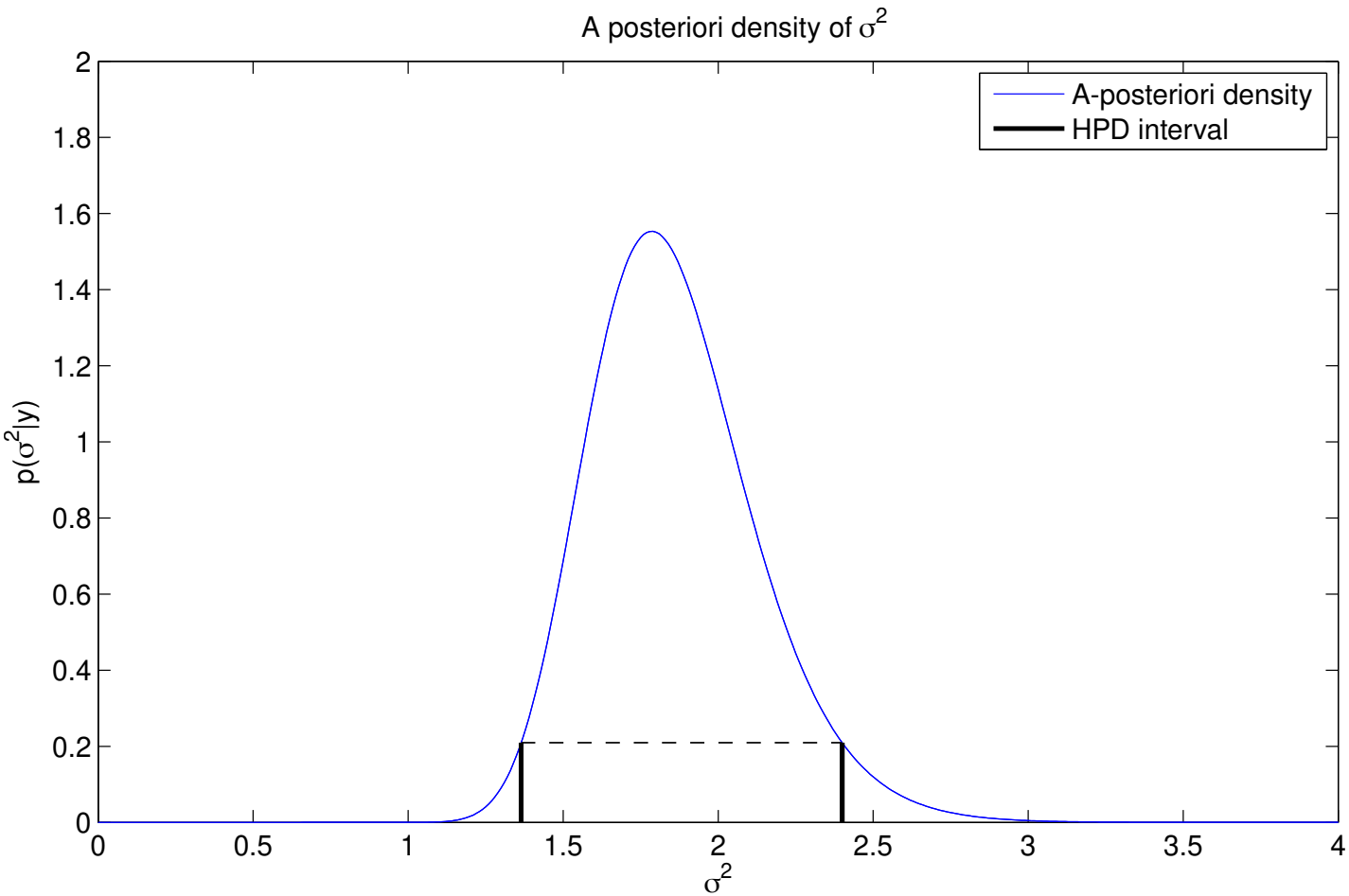
The a posteriori density $p(\sigma^2|\boldsymbol{y})$ of the variance σ^2 is the inverse gamma distribution

$$\mathrm{IG}((n-1)/2, 2/(S^2(n-1))) = \mathrm{IG}(49.5, 0.0110879)$$

The HPD interval with $1-\alpha = 0.95$ is equal to $[1.3645, 2.4002]$.

MATLAB: `make_posterior_normal`

Normally Distributed Data



Section 38: Exponentially Distributed Data

Exponentially Distributed Data

- We consider an exponentially distributed sample $\boldsymbol{y} = (y_1, \ldots, y_n),\ n > 1$.
- We want to obtain an estimate of the mean τ of the exponential distribution from which the data are drawn.
- The likelihood function is:

$$p(\boldsymbol{y}|\tau) = \prod_{i=1}^{n} \frac{1}{\tau} \exp\left(-\frac{y_i}{\tau}\right) = \frac{1}{\tau^n} \exp\left(-\frac{\sum y_i}{\tau}\right)$$

- The Fisher information is equal to $I_\tau = n/\tau^2$; Jeffrey's prior is therefore $\pi_{\mathrm{J}}(\tau) = \tau^{-1}$ and hence improper.
- The a-posteriori distribution is then proportional to

$$p(\tau|\boldsymbol{y}) \propto \frac{1}{\tau^{n+1}} \exp\left(-\frac{\sum y_i}{\tau}\right) = \frac{1}{\tau^{n+1}} \exp\left(-\frac{s}{\tau}\right)$$

Exponentially Distributed Data

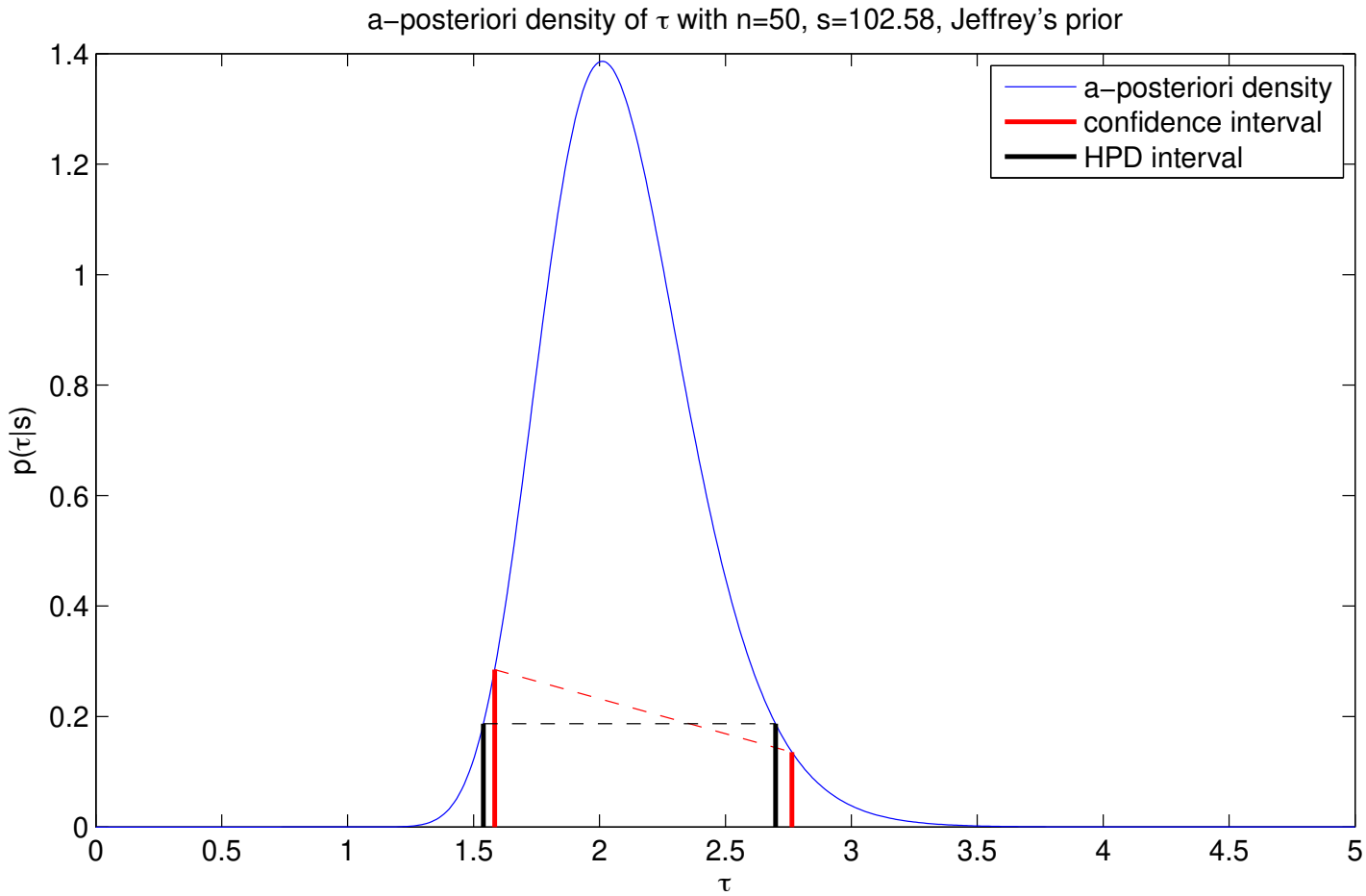
- The a-posteriori distribution is therefore the inverse gamma distribution $\mathrm{IG}(n, 1/s)$.
- The mean is $s/(n-1)$, the mode is $s/(n+1)$.

Example

There is an exponentially distributed sample of size $n = 50$. The sum of the data is $s = 102.58$. The a-posteriori mean is 2.0935, the a-posteriori mode is 2.0114. The HPD interval with $1 - \alpha = 0.95$ is $[1.5389, 2.6990]$.

MATLAB: `make_posterior_exponential`

Exponentially Distributed Data



Exponentially Distributed Data

- The likelihood function can also be parameterized with the mean rate λ:

$$p(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda y_i) = \lambda^n \exp\left(-\lambda \sum y_i\right) = \lambda^n \exp(-\lambda s)$$

- Jeffrey's prior is $\pi_{\mathrm{J}}(\lambda) = \lambda^{-1}$.
- The a-posteriori distribution is then proportional to

$$p(\lambda|\boldsymbol{y}) \propto \lambda^{n-1} \exp(-\lambda s)$$

- The a-posteriori distribution is therefore the gamma distribution $\mathrm{Ga}(n, 1/s)$ with mean n/s and mode $(n-1)/s$. The mean is equal to the maximum likelihood estimator of λ.
- The gamma distribution is also the conjugate a-priori distribution.

Section 39: Markov Chain Monte Carlo

Markov Chain Monte Carlo

- The sequence $X_0, X_1, \ldots$ of discrete random variables is called a **discrete Markov chain** if the **Markov property** holds:

$$W(X_{n+1} = j | X_1 = i_1, \ldots, X_n = i_n) = W(X_{n+1} = j | X_n = i_n)$$

- The transition probabilities are calculated in a **transition matrix** $p_{ij}^{(n)}$, with:

$$p_{ij}^{(n)} = W(X_{n+1} = j | X_n = i)$$

$p_{ij}^{(n)}$ is an **stochastic matrix**: the row sums are 1.

- In a **time homogeneous** Markov chain, $p_{ij}^{(n)} = p_{ij}$ holds for all n.
- Let $\boldsymbol{\pi}^{(0)}$ be the density of X_0. Then the density of X_n is.

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)}\, p_{ij}^{(n)}$$

Markov Chain Monte Carlo

- Under certain circumstances (the chain must be irreducible, all states must be positively recurrent), a temporally homogeneous Markov chain has a **stationary distribution** $\boldsymbol{\pi}$, with

$$\boldsymbol{\pi} = \boldsymbol{\pi}\, p_{ij}$$

- Obviously, $\boldsymbol{\pi}$ is a left eigenvector of p_{ij} with the eigenvalue $\lambda = 1$.

Example

Simple weather model. States: sunny=1, rainy=2. Transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

Example (Continuation)

Evolution over time, starting from a sunny day:

n	$W(1)$	$W(2)$
0	1.0	0.0
1	0.9	0.1
2	0.85	0.15
3	0.825	0.175
4	0.813	0.188
5	0.806	0.194
...	...	...

Example (Continuation)

Time evolution, starting with a rainy day:

n	$W(1)$	$W(2)$
0	0.0	1.0
1	0.4	0.6
2	0.6	0.4
3	0.7	0.3
4	0.75	0.25
5	0.775	0.225
...	...	...

Stationary distribution:

$$\boldsymbol{\pi} = \begin{pmatrix} 0.8 & 0.2 \end{pmatrix}$$

Markov Chain Monte Carlo

- A discrete Markov chain is called **reversible** if the condition of **detailed balance** holds, i.e. if there is a distribution p_{ij} for which holds:

$$\pi_i p_{ij}^{(n)} = \pi_j p_{ji}^{(n)}$$

- In that case p_{ij} is a stationary distribution of the chain.
- For **continuous** random variables X_i, the transition matrix is replaced by a **transition kernel** $q(x, y)$ with:

$$\int q(x, y)\, \mathrm{d}y = 1$$

- If at time n the state of the chain is equal to x, then the State y at time $n + 1$ is given by $q(x, y)$.

Markov Chain Monte Carlo

- **Markov Chain Monte Carlo (MCMC)** proceeds in reverse: given a target distribution $\pi(x)$, we look for a transition kernel $q(x, y)$ that takes $\pi(x)$ as a stationary distribution.
- In general, it is not possible to find such a kernel explicitly. find such a kernel. However, a practical solution is the **Metropolis–Hastings** algorithm.
- Assume that for certain x, y holds:

$$\pi(x)q(x, y) > \pi(y)q(y, x)$$

  Then the chain evolves from x to y more often than from y to x.
- To correct this, the frequency of evolutions from x to y is reduced by introducing a **acceptance probability** $\alpha(x, y) < 1$ is introduced.

Markov Chain Monte Carlo

- From the detailed balance condition we get

$$\alpha(x,y) = \min\left[1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right]$$

- $\alpha(x,y)$ can be calculated **without knowledge of the normalization constant of** $\pi(x)$!

The Metropolis–Hastings Algorithm

- Generate a starting value x_0.
- For i=1,. . . ,N:
  - Generate a default value y from $q(x_{i-1}, y)$ and u from $\mathrm{Un}(0,1)$.
  - If $u \leq \alpha(x_{i-1}, y)$, set $x_i = y$, otherwise $x_i = x_{i-1}$.
- Discard the first n links of the chain (,,burn-in").

Markov Chain Monte Carlo

- The M–H algorithm can be used to draw from densities with unknown normalization, such as posteriori densities.
- If $q(x, y) = q(y, x)$, then the acceptance probability is

$$\alpha(x, y) = \min\left[1, \frac{\pi(y)}{\pi(x)}\right]$$

- If $\pi(y) > \pi(x)$, then assume the development step; otherwise take the step with probability $\pi(y)/\pi(x)$. This is the basis of **optimization by "simulated annealing"'**.
- If $q(x, y) = g(y - x)$, then the chain is a **random walk**, since. $y = x + z$, z describes the stochastic noise with a distribution $g(z)$. If $g(x)$ is symmetric, the probability of acceptance reduces again as above.

Markov Chain Monte Carlo

- In case $q(x, y) = p(y)$ does not depend on x, the algorithm is also called **Independence Sampler**, and $p(y)$ is called the **Suggestion Density**. The acceptance probability is

$$\alpha(x, y) = \min\left[1, \frac{\pi(y)p(x)}{\pi(x)p(y)}\right]$$

Literature

1. S. Chib and E. Greenberg, Understanding the Metropolis–Hastings Algorithm. The American Statistician 49/4, 1995.
2. L. Tierney, Markov Chains for Exploring Posterior Distributions. The Annals of Statistics 22/4, 1994.
3. W.K. Hastings, Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57/1, 1970.
4. O. Cencic and R. Frühwirth, Data reconciliation with non-normal distributions I: linear constraints. Submitted to Computers & Chemical Engineering.