An intuitive introduction to statistics

Péter Zoltán Csurcsia
pcsurcsi@vub.ac.be
Preface

• This catch up material is intended for BRUFACE master students only.
• The focus is here on the practical aspects and therefore some notions are not explained with the adequate scientific punctuality.
Statistics vs probability theorem

- **Probability theorem** deals with predicting the likelihood of future events, while **statistics** involves the analysis of past events.

- **Probability theorem** is theoretical which studies the consequences of mathematical definitions.

- **Statistics** is applied which tries to make sense of observations in the real world.

- **Statistics** is discrete while probability theorem can be discrete and continuous.
In statistics we are dealing with random samples from a larger set. The random samples are forming a sequence, which have a certain characteristics called distribution. The larger set is the population. The observed variable is the random/stochastic variable.

Example (to be continued):
Some politicians are $x=[1.68 \ 1.85 \ 1.80 \ 1.63 \ 1.65 \ 1.65]$m tall. Note: Of course there are many more politicians but these are the randomly chosen ones...

Source: BBC UK
Population/set and their interpretation

• In order to completely describe the population, all samples from the population would be needed.
• This is in practice not possible, so we use random samples (from the population).
• But it is still too complicated to interpret, and therefore we use metrics (variables) to show the most important properties of the represented population.

• Such metrics are, for instance: Mean, Standard deviation (dispersion), Median, Mode, etc.
Basic variables (1)

• (Sample) Mean/ Average/ Expected value (probability theorem)

\[ \mathbb{E}\{X\} = \mu \Rightarrow \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \]

where \( x \) is the random sequence, \( N \) is the number of samples.

• Sample variance (later in details discussed):

\[ \text{var}\{X\} \Rightarrow \sigma^2 = s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{x} - x_i)^2 \]

**Hint:** it tells us how spread are the values around the mean value. The population variance is the same expression but it is normalized by \( N \) instead of \( N-1 \).
Basic variables (2)

• (Sample) standard deviation: square root of the variance

\[
\text{std} = s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\bar{x} - x_i)^2}
\]

Hint: it tells us that with a very large likelihood a randomly chosen sample is within the \( \bar{x} \pm s \) interval.

• Motivation why these variables are not enough: in a company the average salary is 1 million euro. Does it mean that an average guy/most guys could earn one million?
Basic variables (3)

• Median: the middle of the sequence. It separates the lower half, and the higher half of the distribution.

Calculation: place the samples in an increasing order, chose the middle element (if there are odd samples then average the two in the middle)

• Mode (modus): the most frequent element. If a distribution has more modes then it is multimodal distribution.
Basic variables – example

• \( x = [1.68 \ 1.85 \ 1.80 \ 1.63 \ 1.65 \ 1.65] \text{m} \)

What is the mean/variance/std/med/mod?

Solutions

\[
\bar{x} = \frac{(1.68 + 1.85 + 1.80 + 1.63 + 1.65 + 1.65) \times 6}{6} = 1.71
\]

\[
s^2 = \frac{\sum_{i=1}^{6} (1.71 - x_i)^2}{5} = 0.0084
\]

\[s = \sqrt{s^2} = 0.00919\]

\( x' = [1.63 \ 1.65 \ 1.65 \ 1.68 \ 1.80 \ 1.85] \text{m} \)

Med(\( x \)) = (1.65 + 1.68)/2 = 1.665  \quad \text{Mod}(\text{x}) = 1.65 \text{ (2x)}
Basic variables – example solved in Matlab

```matlab
>> x=[1.68 1.85 1.80 1.63 1.65 1.65];
>> mean(x)
ans = 1.7100
>> var(x)
ans = 0.0084
>> std(x)
ans = 0.0919
>> mode(x)
ans = 1.6500
>> median(x)
ans = 1.6650
```
A histogram is a graphical representation (a bar diagram) of the random data. It provides the (empirical) distribution.

- X-axis: variable quantity (e.g. height in meter)
- Y-axis: the frequency (how many times does a thing occur in the related interval).

```matlab
>>x=[1.68 1.85 1.80 1.63 1.65 1.65];
>>hist(x)
```
Probability Density Function (PDF)

- Graph, gives the likelihood (y axis) at certain elements (x axis)

- Properties:
  - Continuous,
  - Max is limited to 1 (1=100%),
  - Min is limited to 0 (0=0%),
  - Integrate (sum) is 1
  - Typically denoted by $f(x)$
  - Similar to histogram (see later on)
Probability Mass Function (PMF)

• The discrete PDF is the PMF
• Known as empirical PDF as well, or discrete PDF
• Properties are the same as the properties of PDF

Examples:
• PMF of a flipping a coin
  \[ f(x=0) : \text{head} \]
  \[ f(x=1) : \text{tail} \]
• DIY: PMF of throwing a die
Conversion from histogram to PDF/PMF

- From histogram: normalize (divide) with the number of samples (if needed, rescale the x-axis, and use more bins)
- To check: sum of the normalized histogram should give exactly 1

Conversion example

- marks=[10 9 15 6 19 16 12 11 10 18 2 11 20];
- Plot the histogram and the distribution function
- Considering this as a distribution. What is the likelihood to pass this exam (mark is min 10), or the have a very good mark (mark is min 18)?
Conversion example: solution in Matlab

```matlab
>> marks=[10 9 15 6 19 16 12 11 10 18 2 11 20];
>> hist(marks); %general plot, insufficient resolution
>> hist(marks,1:20); %good plot, good resolution
>> pdf_marks=hist(marks,1:20)/length(marks);
>>> sum(pdf_marks) %to check if the pdf is correct
ans = 1.0000
>> plot(1:20,pdf_marks); xlabel('mark');ylabel('probability');

>> sum(pdf_marks(10:20)) %likelihood to pass the exam
ans = 0.7692

>> sum(pdf_marks(18:20)) %likelihood to very good
ans = 0.2308
```
Cumulative Distribution Function (CDF)

• It is the integrate of PDF. It gives the likelihood that a random variable will take a value less than or equal to $x$.
  • Monotonic increasing,
  • Min is 0,
  • Max is 1
  • Typically denoted by $F(x)$ (because PDF is $f(x)$)

Example – CDF conversion

• marks=[10 9 15 6 19 16 12 11 10 18 2 11 20];
• Plot the CDF
• Considering this statistics as a distribution. What is the likelihood to pass or to get a very good mark?
Example – CDF conversion solution in Matlab

```matlab
>> marks=[10 9 15 6 19 16 12 11 10 18 2 11 20];
>> pdf_marks=hist(marks,1:20)/length(marks);
>> cdf_marks=cumsum(pdf_marks); %conversion to CDF
>> plot(1:20,cdf_marks);xlabel('marks');ylabel('probablity');

>> 1-cdf_marks(9) % likelihood to pass
ans =
   0.7692
>> 1-cdf_marks(17) % likelihood to get very good
ans =
   0.2308
```
Normal (Gaussian) distribution

One of the most important distributions.

PDF

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

Standard normal distribution: Special case when the expected value is zero and the variance is one i.e. $\mathcal{N}(0,1)$
Standard deviation and the likelihood

- $1 \sigma$, $2 \sigma$, $3 \sigma$ rule
- In general $2\sigma$ (confidence) interval is enough
Two important theorems

• Law of large numbers:
The average of the results obtained from a large number experiment approaches the expected value, and it will tend to come closer and closer as more experiments are performed.

• Central limit theorem:
If sufficiently large number of experiments is performed on independent random variables (each with a finite mean and variance) then they will converge to Gaussian distribution, regardless of the underlying distribution.
Some distributions

Source:
We need to measure the dependency between two random variables ($X$ and $Y$), or in other word we need to know how they change together.

To measure this, for instance, a (population or sample) correlation coefficient ($r$) can be used.

Correlation and independence
Correlation examples

• An interesting example from US
Correlation examples

• But be very careful because...

![Graph showing the correlation between Divorce rate in Maine and Per capita consumption of margarine.](image)

**Divorce rate in Maine correlates with Per capita consumption of margarine**

Correlation: 99.26% (r=0.992558)

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture
Correlation and covariance

• To compute the correlation coefficient, first we need to extend the variance to covariance which measures the variance between two random variables

Population covariance – probability theorem
• \( \sigma_{XY}^2 = \text{COV}(X,Y) = \mathbb{E}\{(X - \mathbb{E}\{X\})\ (Y - \mathbb{E}\{Y\})\} \)

Sample covariance – statistics
• \( s_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{x} - x_i)(\bar{y} - y_i) \)

Of course, we get back the the variance if we compute it for only one variable, i.e. \( s = s_{XX} \)
Correlation coefficient and independence

- The (Pearson) sample correlation coefficient is the covariance normalized by the variances (in order to get a number between -1 and 1) and computed as:

\[ r_{XY} = \frac{S_{XY}}{S_X S_Y} = \cdots = \frac{\sum_{i=1}^{N} (\bar{x} - x_i)(\bar{y} - y_i)}{\sqrt{\sum_{i=1}^{N} (\bar{x} - x_i)^2} \sqrt{\sum_{i=1}^{N} (\bar{y} - y_i)^2}} \]

- Important that if two variables are independent then they are uncorrelated \( r_{XY} = 0 \) but the opposite statement is not true: uncorrelated does not (necessarily) mean independency.
Covariance matrix

• The covariance matrix is a 2×2, symmetric, positive semidefinite matrix where the (co)variances of the (cross-)variables are computed.

\[
\begin{bmatrix}
\text{VAR}(X) & \text{COV}(Y,X) \\
\text{COV}(X,Y) & \text{VAR}(Y)
\end{bmatrix}
\]

• In the same way the correlation matrix can be computed 😊

• Autocovariance (matrix): the covariance is computed element wise within the samples of a random variable (so the matrix size is NxN)
System identification and model estimation

- **Task**: guess what the unknown physical system is
- **Goal**: a (mathematical) model using (designing) the excitation signal and the measurement(s)

Modeling techniques:
- **Black box**: we don’t know anything about the system
- **Grey box**: we know something in advance
Parametric vs nonparametric modeling

• It is a representation

\[ x(t) = x_0 + vt + a \frac{t^2}{2} \]

SYSTEM

nonparametric models

parametric

impulse response function

time [sec]

amplitude [V]
Properties of an estimator (1)

• **Bias**: how much the difference is between the expected value of the estimator and the true (theoretical) value of the observed phenomenon
  If there is no bias, it is called unbiased.

• **Variance**: how much the values vary around the true value

• **MSE** (Mean Square Error) = $\text{bias}^2 + \text{variance}$
Properties of an estimator (2)

- **Consistency**: a good estimator gives the results closer and closer to the true value as the number of samples grows
- **Efficiency**: unbiased estimator with the smallest variance (which reaches the Cramér–Rao lowerbound)
- Minimal: minimum number of parameters to describe an observation
- Robustness
- Asymptotic normality
Estimators in general

- Least Squares
- Gauss-Markov
- Maximum Likelihood
- Bayes

- a priori knowledge
- complexity
Estimator: Least Squares (Maximum Likelihood)

Goal: minimize the sum of squared errors
Example: total household expenses

Parametric model

95€ + 155€/person
Estimator: Maximum A Posteriori (Bayes)

- We have prior information
- Balance between the measured and prior value

Measured: 221V ± 10V
Prior: 230V

Weighting: 50% measured + 50% prior

Estimated value: 225.5V
This example is solved by LS in matrix form.
The parameter to estimate is $\theta (R)$.
The observation matrix is $K(I)$ and the system output is $Y (V)$.

\[
\hat{\theta}_{LS} = \hat{R}_{LS} = \left[ K^T K \right]^{-1} K^T Y
\]
Inverse can be problematic...

\[
\text{>> } \text{inv}(X^*X)*X^*Y
\]
ans =
993.3333

\[
\text{\textbackslash also solves the equation but with help of QR decomposition and it is better conditioned}
\]
\[
\text{>> } K\backslash Y
\]
ans =
993.3333

\[
\text{The analytical way}
\hat{R}_{LS} = \frac{\sum KY}{\sum K^2} = \frac{\sum VI}{\sum I^2}
\]
\[
\text{>> } (X^*Y)/(X^*X)
\]
ans =
993.3333
„It is not because things are difficult that we do not dare,
   it is because we do not dare that they are difficult.”

Seneca
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of figures</td>
<td>3</td>
</tr>
<tr>
<td>List of tables</td>
<td>4</td>
</tr>
<tr>
<td>1 Introduction: probability versus statistics</td>
<td>7</td>
</tr>
<tr>
<td>1.1 What is probability?</td>
<td>7</td>
</tr>
<tr>
<td>1.2 What is statistics?</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Why a course about 'probability and statistics'?</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Goals of this course</td>
<td>9</td>
</tr>
<tr>
<td>2 Descriptive statistics</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Presentation of raw data</td>
<td>14</td>
</tr>
<tr>
<td>2.3 Histogram</td>
<td>15</td>
</tr>
<tr>
<td>2.4 The empirical distribution function</td>
<td>16</td>
</tr>
<tr>
<td>2.5 Measures of location: mean, median, mode</td>
<td>17</td>
</tr>
<tr>
<td>2.6 Measures of spread</td>
<td>18</td>
</tr>
<tr>
<td>2.7 MATLAB®-instructions</td>
<td>19</td>
</tr>
<tr>
<td>3 Describing random events</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>22</td>
</tr>
<tr>
<td>3.2 Random event</td>
<td>23</td>
</tr>
<tr>
<td>3.3 The axiomatic definition of a probability</td>
<td>24</td>
</tr>
<tr>
<td>3.4 Conditional probability</td>
<td>27</td>
</tr>
<tr>
<td>3.5 Independent events</td>
<td>29</td>
</tr>
<tr>
<td>3.6 Rule of Bayes</td>
<td>31</td>
</tr>
<tr>
<td>4 Stochastic variables and their probability distribution</td>
<td>34</td>
</tr>
<tr>
<td>4.1 Stochastic variable</td>
<td>34</td>
</tr>
<tr>
<td>4.2 Distribution function</td>
<td>36</td>
</tr>
<tr>
<td>4.2.1 Introduction and definition</td>
<td>36</td>
</tr>
<tr>
<td>4.2.2 Properties of a distribution function</td>
<td>37</td>
</tr>
<tr>
<td>4.2.3 Discrete and continuous distribution functions</td>
<td>38</td>
</tr>
<tr>
<td>4.2.4 Nonlinear transformation of a continuous stochastic variable</td>
<td>40</td>
</tr>
</tbody>
</table>
9.5.1 The chi-square distribution ................................ 94
9.5.2 Study of the sample variance ................................ 97
9.5.3 Confidence intervals for $S_n^2$ ............................... 98

9.6 Study of $\frac{\bar{X} - \mu}{S_n}$ ......................................... 99
9.6.1 Student’s $t$-distribution ..................................... 100
9.6.2 Confidence intervals ........................................... 102

10 Hypothesis testing .................................................. 104
10.1 Introduction ......................................................... 104
10.2 Hypothesis and significance testing: the basic ideas ...... 106
10.2.1 Formulation of hypotheses .................................... 106
10.2.1.1 Null hypothesis ............................................. 106
10.2.1.2 The alternative hypothesis ............................... 107
10.2.2 The test statistic ................................................. 108
10.2.3 Statistical significance and exceedance probability ... 108
10.2.4 Procedure ........................................................... 110

10.3 Specific testing .................................................... 111
10.3.1 The $T$-test: $H_0: \mu = \mu_0$ ............................... 111
10.3.2 The $\chi^2$-test: $H_0: \sigma = \sigma_0$ .......................... 113
10.3.3 Comparing mean values or the two-group $t$-test ...... 114
10.3.4 The $F$-test, comparing variances ......................... 114

10.4 Errors in hypothesis testing .................................... 115

Estimation of models starting from measurements ............... 117

Bibliography ............................................................. 129
List of Figures

2.1 Histogram of the students' marks. ........................................ 16
2.2 The empirical distribution function for the students' marks for
the chemistry exam. .......................................................... 17

4.1 The distribution function of a die. ..................................... 36
4.2 Example of a bar chart (left) and the corresponding distribution
function (right). .............................................................. 39
4.3 Example of a continuous distribution. ................................. 40

5.1 Example of independent (left) and dependent (right) stochastic
quantities. ........................................................................ 48

6.1 Example of a median. ....................................................... 53

7.1 Stochastic variables with different correlations. .................... 66

8.1 The binomial distribution for \( n = 10 \). ............................... 73
8.2 Example of a hypergeometric distribution. .......................... 76
8.3 Example of a geometric distribution. ................................... 77
8.4 Example of a Poisson probability density function: \( \lambda = 0.1, 1, 10 
and \( t = 1 \). .................................................................. 80
8.5 Example of the course of the probability density function of the
exponential distribution. ...................................................... 82

9.1 Example of the convergence of the distribution of the sum of
independent random variables to a normal distribution. .......... 87
9.2 The standard normal distribution \( \mathcal{N}(0, 1) \) ....................... 88
9.3 Example of a confidence interval, situated symmetrically around 0. 90
9.4 The probability density function of \( X_n \) for \( n = 1, 10, 100, 1000 \). 93
9.5 The probability density function of \( \chi^2_n \) for \( n = 2, 4, 8, 16 \). 95
9.6 95% confidence interval for the estimate of the spread from \( n 
measurements (\sigma = 1). ....................................................... 99
9.7 Example of the \( t_n \)-distribution for \( n = 2, 8, 32 \). As a reference,
the \( \mathcal{N}(0, 1) \) (in ...) is given. ............................................ 101
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Marks of 105 students for the chemistry exam, not sorted.</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Marks of 105 students for the chemistry exam, sorted.</td>
<td>14</td>
</tr>
<tr>
<td>2.3</td>
<td>Representation of the results, divided in classes.</td>
<td>15</td>
</tr>
<tr>
<td>3.1</td>
<td>Probability of a transmission line without failure.</td>
<td>30</td>
</tr>
<tr>
<td>9.1</td>
<td>Uncertainty on the mean value if the spread is exactly known.</td>
<td>92</td>
</tr>
<tr>
<td>9.2</td>
<td>95% confidence interval for $\sigma = 1$.</td>
<td>98</td>
</tr>
<tr>
<td>10.1</td>
<td>The two kinds of errors when testing hypotheses.</td>
<td>115</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction: probability versus Statistics
Chapter 1

Introduction: probability versus statistics

In this chapter we will briefly describe the two main topics of this course: probability on the one hand, statistics on the other hand. Next, the objectives of this course will be discussed.

1.1 What is probability?

Probability is a mathematical discipline in which one establishes an abstract probability-theoretical model to compute and study probabilities. Thereby, one starts with several basic axioms from which one derives some conclusions through deductive reasoning ¹.

¹Deduction: a process of reasoning of developing specific predictions (hypotheses) from general principles. This type of reasoning moves from the general to the particular.
CHAPTER 1. INTRODUCTION: PROBABILITY VERSUS STATISTICS

Examples:

- What is the chance you throw a six with a die?

- What is the chance that 997 of the 1000 CD's are good, when the probability of a bad CD is 0.1%?

- How large should the capacity of a telephone exchange be so that 99% of the telephone calls can be forwarded immediately?

- How much longer one needs to measure so that the uncertainty on the result becomes 10 times smaller?

1.2 What is statistics?

In statistics, inductive reasoning\(^2\) is used to verify if a certain probability-theoretical model is applicable to a given set of observations (measurements). This knowledge can then be used to make predictions.

Examples:

- I have thrown a coin a 100 times of which 60 times 'heads'. Is this coin 'fair' (in other words is the probability to throw 'heads' equal to 0.5)?

- Are the diagnosed leukaemia patients in the neighbourhood of the nuclear plant in Doel due to random causes or is the risk of this disease statistically significantly higher in this region?

- What is the quality of a communication channel? Can the amount of transmission errors be predicted by using a limited number of measurements?

- How many measurements do I need to perform to reduce the stochastic errors on the mean value to 1%?

\(^2\)Induction (Collins dictionary): a process of reasoning by which a general conclusion is drawn from a set of premises, based mainly on experience or experimental evidence.
• How big is the risk that a machinery breakdown will occur?

• How can we determine if 2 professors are equally severe or mild during an examination? How many students do they have to examine together to answer this question?

1.3 Why a course about ‘probability and statistics’?

In everyday life, and in your future career as an engineer or scientist, you get overwhelmed with data. No more than we can stop using words, we can stop using data. Like words on a page are meaningless for an illiterate – or confusing for someone with a lack of education – data do not speak for themselves, they need to be handled by someone who knows how to interpret them. Just like an author who chooses his words to get convincing arguments or an inconsistent story, data can be compulsive or misleading, or just irrelevant. Being numerically educated, having the capability to understand numerical arguments, is very important for your further education and career. It is essential to be able to express yourself in a numerical way – being an author instead of a reader. A course on statistics will help you to process critically your data. It learns you to produce data that give clear answers to important questions. To do so, you need solid methods to be able to draw reliable conclusions on the basis of data.

1.4 Goals of this course

Statistical analysis requires a thorough knowledge of probability. Indeed, confidence intervals, hypothesis tests, verification tests, independency tests, etc. cannot be made without the use of probability. In this course, both aspects (probability and statistics) will be studied. In the first part of the course, mainly theoretical models will be studied (theory of probability), while in the second
part we will examine how these models can be combined with observations and how we can extract from this knowledge the probability of a given event.
Chapter 2

Descriptive statistics
Chapter 2

Descriptive statistics

In this chapter we will study how to rearrange a large amount of raw data in order to get a better view on the information that is hidden in the data. Next, we will introduce a number of important quantities (measures) that characterize the distribution of the data with a few numbers.

The following concepts will be introduced:

- Data sample, raw data,
- histogram, division in classes, class width,
- empirical distribution function,
- mean, median, mode,
- empirical variance, standard deviation, median absolute deviation.

2.1 Introduction

In the descriptive statistics, one often needs to process a large amount of raw data. These data can be obtained by sampling (a test on a limited subset that one considers as a representation of the total population), or they can be obtained by collecting data of a well-defined set (e.g. the results of the students
in the first bachelor of engineering for the chemistry exam). Because such large amounts of numbers are not well-arranged, we want to order them in a certain way and represent the results in a graphical way (e.g. a histogram) on the one hand, and on the other hand we can put a number of important characteristics of these data together in a limited amount of measures, like their mean or median and their spread.

The first step, however, is to define the population and the data sample.

Definition 2.1.1: Population
The whole group of objects or persons of which we want information, is called the population.

Definition 2.1.2: Data sample
A data sample is a part of the population that is actually used to obtain information.

Note that the population is defined in terms of what we want to study. If we want to make a conclusion about all students of all universities in Belgium, this group of students is our population. The data sample is the part of the students that were interrogated and from which we make a general conclusion. It is obvious that collecting a data sample needs to be done very precisely. A data sample procedure that is badly designed can lead to spurious results: e.g. if one systematically selects only the 'good' students.

Usually, one is not interested in individual measurements, but wishes to extract some global characteristics, e.g. what is the most common mark? Therefore we wish to present the raw data in a more efficient way. This is shown in Section 2.2 by an example.
2.2 Presentation of raw data

The results of 105 students for the chemistry exam are shown in Table 2.1.

Table 2.1: Marks of 105 students for the chemistry exam, not sorted.

<table>
<thead>
<tr>
<th>12.0</th>
<th>14.0</th>
<th>13.0</th>
<th>18.5</th>
<th>14.0</th>
<th>10.5</th>
<th>17.0</th>
<th>10.0</th>
<th>15.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.0</td>
<td>14.5</td>
<td>14.5</td>
<td>12.0</td>
<td>15.0</td>
<td>13.5</td>
<td>13.0</td>
<td>13.5</td>
<td>15.5</td>
</tr>
<tr>
<td>14.5</td>
<td>13.5</td>
<td>11.0</td>
<td>13.5</td>
<td>12.0</td>
<td>17.5</td>
<td>14.5</td>
<td>11.5</td>
<td>12.0</td>
</tr>
<tr>
<td>12.0</td>
<td>12.0</td>
<td>12.5</td>
<td>13.0</td>
<td>12.0</td>
<td>11.5</td>
<td>13.5</td>
<td>14.0</td>
<td>10.0</td>
</tr>
<tr>
<td>9.0</td>
<td>10.5</td>
<td>14.5</td>
<td>12.0</td>
<td>15.0</td>
<td>13.0</td>
<td>9.0</td>
<td>12.5</td>
<td>11.5</td>
</tr>
<tr>
<td>12.5</td>
<td>11.5</td>
<td>13.5</td>
<td>10.5</td>
<td>10.5</td>
<td>11.5</td>
<td>10.0</td>
<td>15.5</td>
<td>14.5</td>
</tr>
<tr>
<td>7.0</td>
<td>16.5</td>
<td>14.5</td>
<td>7.5</td>
<td>16.0</td>
<td>14.0</td>
<td>17.0</td>
<td>9.0</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>14.0</td>
<td>11.0</td>
<td>13.0</td>
<td>14.5</td>
<td>14.0</td>
<td>11.5</td>
<td>12.0</td>
<td>16.5</td>
</tr>
<tr>
<td>15.5</td>
<td>17.0</td>
<td>12.5</td>
<td>14.5</td>
<td>10.5</td>
<td>16.0</td>
<td>17.0</td>
<td>15.5</td>
<td>14.5</td>
</tr>
<tr>
<td>13.5</td>
<td>16.0</td>
<td>9.5</td>
<td>13.0</td>
<td>16.5</td>
<td>12.5</td>
<td>14.0</td>
<td>11.5</td>
<td>16.5</td>
</tr>
<tr>
<td>13.5</td>
<td>13.5</td>
<td>15.0</td>
<td>11.0</td>
<td>13.5</td>
<td>16.0</td>
<td>9.0</td>
<td>14.0</td>
<td>12.5</td>
</tr>
<tr>
<td>13.5</td>
<td>15.0</td>
<td>13.0</td>
<td>11.0</td>
<td>10.5</td>
<td>17.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table of raw data does not contain much information. It's hard to find the most common mark, the position of the extremes... The same data, now sorted by magnitude, given in Table 2.2, contain much more information.

We can see immediately that all measurements are contained in the interval [7.0, 18.5]. The values around 13 and 14 appear most frequently.

Table 2.2: Marks of 105 students for the chemistry exam, sorted.

<table>
<thead>
<tr>
<th>7.0</th>
<th>7.5</th>
<th>9.0</th>
<th>9.0</th>
<th>9.0</th>
<th>9.0</th>
<th>9.0</th>
<th>9.0</th>
<th>9.5</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>10.0</td>
<td>10.5</td>
<td>10.5</td>
<td>10.5</td>
<td>10.5</td>
<td>10.5</td>
<td>10.5</td>
<td>10.5</td>
<td>11.0</td>
</tr>
<tr>
<td>11.0</td>
<td>11.0</td>
<td>11.0</td>
<td>11.0</td>
<td>11.5</td>
<td>11.5</td>
<td>11.5</td>
<td>11.5</td>
<td>11.5</td>
<td>11.5</td>
</tr>
<tr>
<td>11.5</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td>12.0</td>
<td>12.5</td>
<td>12.0</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
</tr>
<tr>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
</tr>
<tr>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>14.0</td>
<td>14.0</td>
</tr>
<tr>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>14.5</td>
<td>14.5</td>
</tr>
<tr>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
</tr>
<tr>
<td>15.0</td>
<td>15.0</td>
<td>15.0</td>
<td>15.5</td>
<td>15.5</td>
<td>15.5</td>
<td>15.5</td>
<td>15.5</td>
<td>16.0</td>
<td>16.0</td>
</tr>
<tr>
<td>16.0</td>
<td>16.0</td>
<td>16.0</td>
<td>16.5</td>
<td>16.5</td>
<td>16.5</td>
<td>16.5</td>
<td>17.0</td>
<td>17.0</td>
<td>17.0</td>
</tr>
<tr>
<td>17.0</td>
<td>17.0</td>
<td>17.0</td>
<td>17.5</td>
<td>17.5</td>
<td>17.5</td>
<td>18.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For an even better overview, the data can be divided in classes (mostly 5 till 30). To do this we choose a class width, e.g. 1, we divide the relevant interval in half-open subintervals with this selected width, e.g. [13.0, 14.0], etc., and we count the frequencies (the number of times that a measurement is contained in
CHAPTER 2. DESCRIPTIVE STATISTICS

a certain interval). The raw data were processed in this way in Table 2.3. It is obvious that this representation admits a much clearer view on the distribution of the results.

Table 2.3: Representation of the results, divided in classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Lower limit ≤</th>
<th>Upper limit &lt;</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>13</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>14</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>15</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>16</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>17</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>18</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>19</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

2.3 Histogram

Instead of using tables, the split in classes can be represented graphically in a histogram. We draw a rectangle in each subinterval with an area that is proportional to the frequency of the corresponding class. In Figure 2.1, a histogram is drawn for 2 different class widths. This Figure shows that the shape of the histogram strongly depends on the choice of the class width. In practice, the class width needs to be chosen in function of the number of measurements: the more measurements, the smaller the class width can be chosen to get a better picture of the distribution of the measurements. If the number of measurements is too small for a given number of classes, the picture of the distribution becomes
strongly disturbed because of random fluctuations.

![Histograms of students' marks](image)

Figure 2.1: Histogram of the students' marks. Left: class width = 1; Right: class width = 0.5. Remark: the MATLAB™-definition differs from the one used in this textbook.

## 2.4 The empirical distribution function

**Definition 2.4.1: Empirical distribution function**

The empirical distribution function $F_n(x)$ for $n$ measurements is

$$F_n(x) = \frac{\# \{x_i \leq x\}}{n} \quad (2.1)$$

A graphical representation of this function immediately shows the distribution of the data (amount of small values, amount of extremely large values, ...). In combination with a histogram, this function gives a lot of insight. The empirical distribution function will be supplemented with the 'theoretical' distribution later on (see Section 4.2).

In Figure 2.2, the empirical distribution function is shown for the data in Table 2.1. This function is a staircase function with a discontinuity in the points $x_i$. For continuous quantities (e.g. the weight of 2 month old pigs; the daily
precipitation measured at the KMI (the Belgian Royal Meteorological Institute); the voltage of the mains) this distribution function will converge to a continuous function.

![Graph](image.png)

Figure 2.2: The empirical distribution function for the students' marks for the chemistry exam.

2.5 Measures of location: mean, median, mode

Sometimes we want to reduce the data even more than we did using a histogram or an empirical distribution function. Therefore 2 measures will be used: one for the location and one for the spread. The mean, median and mode are commonly used measures for location; for the spread the standard deviation and sometimes the median absolute deviation are used.
CHAPTER 2. DESCRIPTIVE STATISTICS

Definition 2.5.1: Mean
The mean (sample mean) of the data \( \{x_i\}, i = 1, \ldots, n \) is

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{2.2} \]

Definition 2.5.2: Median
Consider the sorted data \( \{x_1 \leq x_2 \leq \ldots \leq x_n\} \). The median is the 'middle' observation if \( n \) is odd, and is the mean of the two middle observations if \( n \) is even:

\[ \text{med} = x_{\frac{n+1}{2}} \text{ if } n \text{ is odd, and } \text{med} = \frac{x_y + x_{y+1}}{2} \text{ if } n \text{ is even}. \tag{2.3} \]

Definition 2.5.3: Mode
The mode of a data sample is the observation that appears most often in the data set.

Note that the mode does not need to be unambiguous; different observations can appear equally.

In this example the mean is \( \bar{x} = 13.19 \), the median \( \text{med} = 13.5 \) and the mode is 13.5.

The mean is the easiest measure to calculate but is very sensitive to outliers. The median, on the other hand, is much more robust, but the data need to be sorted first (with the use of computers this is not a problem anymore).

Anecdote: Most people in Brussels have more than the average number of legs.

2.6 Measures of spread

Besides the location, it is important to know how much the measurements can differ from this location, in other words, how strongly the measurements are grouped. For this purpose we use the standard deviation and sometimes the median absolute deviation.
CHAPTER 2. DESCRIPTIVE STATISTICS

Definition 2.6.1: Standard deviation $s$, variance $s^2$
$s = \sqrt{s^2}$ with $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Definition 2.6.2: Median absolute deviation (MAD)
$\text{MAD} = \text{median} \{|x_i - \text{med}|, i = 1, \ldots, n\}$.

Note that half of the observations is between $\text{med} - \text{MAD}$ and $\text{med} + \text{MAD}$.
In the example: $s = 2.36$, $\text{MAD} = 1.89$.

2.7 MATLAB™-instructions

Below, some MATLAB™-instructions are described shortly. Full information can be found in MATLAB™ by typing help INSTRUCTION.

hist Histogram.

- $N = \text{hist}(Y)$ bins the elements of $Y$ into 10 equally spaced containers and returns the number of elements in each container. If $Y$ is a matrix, \text{hist} works down the columns. $N = \text{hist}(Y,M)$, where $M$ is a scalar, uses $M$ bins.

mean Average or mean value.

- For vectors, mean($X$) is the mean value of the elements in $X$.
- For matrices, mean($X$) is a row vector containing the mean value of each column. For N-D arrays, \text{MEAN($X$)} is the mean value of the elements along the first non-singleton dimension of $X$.
- mean($X$,$\text{DIM}$) takes the mean along the dimension $\text{DIM}$ of $X$.
- Example: If $X = \begin{bmatrix} 0 & 1 & 2; & 3 & 4 & 5 \end{bmatrix}$ then mean($X$,1) is $[1.5 \ 2.5 \ 3.5]$ and mean($X$,2) is $[1; \ 4]$.
std Standard deviation.

- For vectors, \(\text{std}(X)\) returns the standard deviation. \(\text{std}(X)\) normalizes by \((N-1)\) where \(N\) is the sequence length.
- For matrices, \(\text{std}(X, \text{FLAG}, \text{DIM})\) takes the standard deviation along the dimension \(\text{DIM}\) of \(X\). When \(\text{FLAG}=0\), \(\text{std}\) normalizes by \((N-1)\); otherwise \(\text{std}\) normalizes by \(N\).

median Median value.

- For vectors, \(\text{median}(X)\) is the median value of the elements in \(X\).
- For matrices, \(\text{median}(X)\) is a row vector containing the median value of each column.

mad Mean absolute deviation.

- \(Y = \text{mad}(X)\) calculates the mean absolute deviation (MAD) of \(X\).
- For matrix \(X\), \(\text{mad}\) returns a row vector containing the MAD of each column.
Chapter 3

Describing random events
Chapter 3

Describing random events

In this chapter we learn how to describe a random event. For this end, a probability model will be introduced and next the individual components of this probability model will be studied in more detail. The following concepts will be introduced:

- random event,
- probability model,
- sample space,
- probability: axiomatic definition and basic properties,
- conditional probability,
- (in)dependent events,
- rule of Bayes.

3.1 Introduction

It is a fact of life that some things are a coincidence. The result of flipping a coin, the time interval between the emission of two particles by a radioactive source, the gender of a baby, the result of these repeated measurements are all
unpredictable. Often, an intuitive concept of probability can be put down to these 'events':

- We know that the chance to throw a $\spadesuit$ or $\heartsuit$ with a die is smaller than the chance to throw a $\diamondsuit$, $\clubsuit$ or $\heartsuit$.

- If we step into a car, we assume that the chance of getting into a car accident is very small.

- We can decide at first glance if there is a chance of rain, and whether or not we need to take along an umbrella.

However, this intuitive experience with probabilities must be formalized before it can be used as a basic concept in mathematics. To this end, we need to introduce a probability model. A probability model for a random event exists of a sample space $\Omega$, that contains all possible outcomes, and a quantity $P$, that assigns a number $P(A)$ as a probability to each event $A$. All these concepts will be introduced below.

### 3.2 Random event

**Definition 3.2.1: random events**

A phenomenon is called a random event if all outcomes are uncertain, but nonetheless a regular distribution of the outcomes exists for a large number of iterations.

**Example:** if we flip a coin once, we cannot predict the result: heads or tails. However, if we flip the coin 10000 times the result appears to be about 5000 times heads and 5000 times tails. Despite the fact that we cannot predict the individual results, we can certainly describe the behaviour of a large number of experiments.

A random event can be described by its possible outcomes.

**Definition 3.2.2: Sample space**

$\Omega$ is the set of outcomes or elementary events (and is called the sample space).
CHAPTER 3. DESCRIBING RANDOM EVENTS

Example: if we roll a die this sample space is: \( \Omega = \{\, \Box, \, \bigcirc, \, \bigtriangleup, \, \bigtriangledown, \, \hbar, \, \} \).

These elementary events are not the only things we are interested in. Often, we want to study more complicated situations, where some elementary events are combined.

**Definition 3.2.3: Event**

An event \( A \) is a subset of \( \Omega \).

Next, we introduce a collection of all possible events:

**Axiom 3.2.1: Collection of all possible events \( \mathcal{A} \)**

1. \( \emptyset \) and \( \Omega \) are events: \( \emptyset \in \mathcal{A} \) and \( \Omega \in \mathcal{A} \)

2. If \( A \) is an event, then its complement is also an event:

\[
A \in \mathcal{A} \Rightarrow A^c = \Omega \setminus A \in \mathcal{A}, \tag{3.1}
\]

3. If \( A \) and \( B \) are events, then \( A \cup B \) is also an event:

\[
A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A} \tag{3.2}
\]

**Example 3.2.1:**

- The event 'throwing a \( \bigcirc \)' is \( A = \{ \bigcirc \} \).
- The event 'an even cast' is \( A = \{ \Box, \bigcirc, \hbar \} \).

### 3.3 The axiomatic definition of a probability

In order to completely characterize a random event, we do not only need to know all possible outcomes (the sample space \( \Omega \)), but also the probability of each of these outcomes.
CHAPTER 3. DESCRIBING RANDOM EVENTS

Definition 3.3.1:
For each event \( A \), a probability (function) \( P(A) \) exists that expresses the chance of appearance of this event.

Example: For a 'fair' die, our intuition tells us that \( P(A = \{2\}) = \frac{1}{6} \). In this course we define the probability concept by using 3 axioms. All other properties can be derived from this concept.

Probability function \( P \) of \( A \)

Axiom 3.3.1:
\[ 0 \leq P(A) \leq 1 \] for all \( A \in \mathcal{A} \)

Axiom 3.3.2:
\[ P(\emptyset) = 0 \] and \( P(\Omega) = 1 \)

Axiom 3.3.3:
\[ A, B \in \mathcal{A} \text{ and } A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B) \]

Note that \( A \cap B = \emptyset \) means that the events \( A \) and \( B \) cannot occur simultaneously.

Example 3.3.1:
If we roll a die, the sample space is \( \Omega = \{\text{1, 2, 3, 4, 5, 6}\} \) and the probability of an elementary event is 1/6. Prove that this example satisfies all axioms above.

Note that by using the axioms 3.3.1-3.3.3 one can find that the probability of an elementary event must be 1/6 for a 'fair' die, which is in agreement with our intuition from the example in the introduction of this section.

Remark: If \( \Omega \) contains an infinite amount of elements, all axioms above need to be extended. This can be done without problems for countable sets (e.g. the set of natural numbers), but must be done with care for continuous sets (e.g. the set of real numbers).

From the axioms above, we get the following properties:
CHAPTER 3. DESCRIBING RANDOM EVENTS

Property 3.3.1:
If $A, B \in A$, then $A \cap B \in A$

Property 3.3.2: Sum rule
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Property 3.3.3: Complement rule
$P(A) + P(A^c) = 1$

Probability originates from the study of games of chance. Rolling a die, distributing cards, and turning a roulette wheel are all examples of intentional randomization that look like a random data sample. However, the field of application reaches much further than this. Precise, recurrent measurements do not seem to be identical, part of the result is determined by coincidence and can be defined similarly as we did for a random event. However, probability has a large range of applications: describing highway traffic, data transmission in a telecommunication network or computer chip, genetic structure of individuals or populations, the energy state of elementary particles, the spread of an epidemic, analysis of the risks in financial markets and insurances, the choice of the height of a dike, etc.
CHAPTER 3. DESCRIBING RANDOM EVENTS

3.4 Conditional probability

Example 3.4.1:

1. What is the chance that a student of high school's senior year wants to become an engineer?

2. What is the chance that a student of high school's senior year wants to become an engineer if we know the student is a girl?

3. What is the chance that a student of high school's senior year wants to become an engineer if we know the student is a boy?

In all three cases we wish to know the chance that someone wants to become an engineer, but the prior knowledge we have is different for the 3 questions. This has an influence on the answer.

Such questions are formalized with the concept of conditional probability. In conditional probability, one searches for the probability of an event $A$, if one knows that event $B$ occurred:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (3.3)$$

One reads the expression $P(A|B)$ as follows: the probability of $A$ given $B$.

If we consider the conditional probability $P(A|B)$, we limit the set of events that can contain $A$ to the set $B$. Since $P(B|B) = 1$ should be valid, we must normalize all probabilities with $P(B)$ which explains the presence of the denominator in (3.3).

The use of conditional probabilities allows us to split very complex problems into a series of (more) simple problems.

Example 3.4.2:

We roll two dice. What is the probability that the sum of the two dice is even if one of the two dice shows a $\square$?
CHAPTER 3. DESCRIBING RANDOM EVENTS

Solution

1. The total number of possible outcomes when rolling 2 dice $6 \times 6 = 36$.
   Define the following events:
   $A$: sum of the 2 dice is even
   $B$: at least 1 of the dies is a $\lozenge$.

2. What is the probability of $B$?
   
   \[
   P(\text{one of the dies is a } \lozenge) = \frac{11}{36}
   \]

   \[
   P \left( \left\{ \begin{array}{c}
   1 \lozenge, 2 \lozenge, 3 \lozenge, 4 \lozenge, 5 \lozenge, 6 \lozenge
g, 2g, 3g, 4g, 5g, 6g
\end{array} \right. \right)
   \]

3. What is probability of the event $A \cap B$?

   \[P(A \cap B) = P \left( \left\{ \begin{array}{c}
   1 \lozenge, 2 \lozenge, 3 \lozenge, 4 \lozenge, 5 \lozenge, 6 \lozenge
g, 2g, 3g, 4g, 5g, 6g
\end{array} \right. \right) = \frac{5}{36}\]

4. Then we find immediately:

   \[P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{5/36}{11/36} = \frac{5}{11}\]

5. Remark: \[P(A) = \frac{3 \times 3 + 3 \times 3}{36} = \frac{1}{2}\]
3.5 Independent events

In the previous example we have seen that the answer to the question 'What is the probability that the sum of two dies is even' depends on the fact that one knows in advance that one of the two dies shows a 6. So the probability of event A depends on event B. If this is not the case, we talk about an independent event. So, if

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \text{ (this is what we claim),} \quad (3.4)$$

then we can find the result: $P(A \cap B) = P(A)P(B)$ immediately. We will use this as a formal definition.

**Definition 3.5.1: Independent events**

A and B are independent $\iff P(A \cap B) = P(A)P(B)$.

This definition can be extended for N events.

**Definition 3.5.2: Independent events**

The events $A_1, A_2, \ldots, A_N$ are independent if

$$P(A_{v_1} | A_{v_2}A_{v_3} \ldots A_{v_N}) = P(A_{v_1}) \quad (3.5)$$

For all possible choices of the indices $v_i$ for $i = 1, \ldots, N$ (with all $v_i$ different!).

This gives rise to the following property:

**Property 3.5.1: Product rule**

If the events $A_1, A_2, \ldots, A_N$ are independent then:

$$P(A_1 \cap A_2 \cap \ldots \cap A_N) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_N) \quad (3.6)$$

If one can state that events are independent, one can easily reduce the complexity of the problem because one splits a complex problem into a set of simplified problems.
Example 3.5.1:
Consider a transatlantic telephone cable that contains repeaters (restore and amplify the signal; these are not used anymore in optical fibers) at regular distances, e.g. 1000 repeaters. The line is only working if none of these repeaters is broken. Assuming that the probability of a failure over a period of 10 years is 0.001, what is the probability that this line will not fail for 10 years?

To calculate the answer, it is easier to consider the complementary event: the probability of 1 repeater to work correctly for 10 years is 0.999. If we assume that the failure of one repeater is independent of the failure of the other repeaters (independent events), we can use (3.6):

\[
P(\text{no failure along the line}) = P(\text{all repeaters are working}) = (P(\text{1 repeater is working}))^{1000}
\]

This results in the following reliability as a function of the probability of a failure: as given in Table 3.1. This clearly shows that for complex systems with a large number of individual components, one needs to make high reliability demands upon each individual component in order to have a completely reliable system.

<table>
<thead>
<tr>
<th>probability of 1 broken repeater</th>
<th>probability of a repeater without failure</th>
<th>probability of a line without failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.99</td>
<td>0.000043</td>
</tr>
<tr>
<td>0.001</td>
<td>0.999</td>
<td>0.37</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.9999</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 3.1: Probability of a transmission line without failure.
3.6 Rule of Bayes

In some cases one wishes to revert the conditional probabilities, in other words, can we find $P(B|A)$ knowing $P(A|B)$? The answer to this question is given by the rule of Bayes.

Definition 3.6.1: Rule of Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$  \hspace{1cm} \text{(3.9)}

This result follows directly from the equations:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

The above rule can be refined by noticing that

$$P(B) = P(B \cap \Omega)$$  \hspace{1cm} \text{(3.10)}
$$= P(B \cap (A \cup A^c))$$  \hspace{1cm} \text{(3.11)}
$$= P(B \cap A) + P(B \cap A^c)$$  \hspace{1cm} \text{(3.12)}
$$= P(B|A)P(A) + P(B|A^c)P(A^c)$$  \hspace{1cm} \text{(3.13)}

The last equation follows directly from point 3 of Axiom 3.2.1.

Finally, substitution in (3.9) results in:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$  \hspace{1cm} \text{(3.14)}

Example 3.6.1: Research in population

Because of the growing possibilities of medical diagnostic techniques, we often have to address the question if a global medical screening of the population should be made, e.g. in cervical cancer, in seropositivity, ..., is useful, cost effective and/or socially acceptable returns repeatedly.

Consider a test, characterized by:

31
CHAPTER 3. DESCRIBING RANDOM EVENTS

- \( P(\text{positive|ill}) = 0.999 \): this is a correct test
- \( P(\text{positive|not ill}) = 0.01 \): this is a false alarm

If one knows from previous screenings that 0.1% of the population has the disease, one can wonder which percentage of the positive tests is a false alarm, in other words, how many people will be made anxious unnecessarily with respect to the number of correct diagnosis.

Answer:

1. Define:
   - event \( A \): not infected,
   - event \( B \): positive.

2. The probability that one is not infected, while the test is still positive:

   \[
P(B|A) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}
   = \frac{0.01 \times (1 - 0.001)}{0.01 \times (1 - 0.001) + 0.999 \times 0.001} \approx 0.91
   \]

3. Conclusion: the majority of the positive tests when considering a full screening is a false alarm.

Check for yourself the probability that someone has the disease while the test is negative (the disease is not detected).
Chapter 4

Stochastic variables and their probability distribution
Chapter 4

Stochastic variables and their probability distribution

In this chapter we will describe a random event by a variable. Next we will describe the behaviour of this variable by its distribution function.

The following concepts are introduced:

- stochastic variable (discrete/continuous)
- probability distribution function and probability density function
- transformation of variables

4.1 Stochastic variable

The sample space \( \Omega \) does not need to consist of numbers. If we throw 4 coins, we can describe the outcome as a row of H's and T's ('heads' and 'tails'), e.g. HTTH. However, in statistics we are more interested in numerical outcomes, like the number of H's in 4 casts. It is convenient to use a short notation: \( X \) represents the number of H's. If the outcome is HTTH, then \( X = 2 \). If the outcome is TTTT, then the value of \( X \) changes to \( X = 1 \). If we throw 4 coins, the possible values of \( X \) are: 0, 1, 2, 3, 4. We call \( X \) a stochastic variable (or
random variable) because its values are unpredictable when repeatedly throwing
the coins. When the number of experiments is increased, the average behaviour
(e.g. the element 0 of $X$) will be more predictable.

A stochastic variable is a variable whose value is a numerical outcome at-
tached to a random event.

**Definition 4.1.1: Stochastic variable**
The real function $X$ from $\Omega$ to $\mathbb{R}$ is a stochastic variable if the image of $X$ is
compatible with the structure of the set $A$ of subsets in $\Omega$:

1. For all real numbers $a$: $\{ \omega \in \Omega | X(\omega) \leq a \} \in A$

2. The probabilities, defined on the elements of $A$, are projected along:

$$P(X \leq a) = P(\{ \omega \in \Omega | X(\omega) \leq a \})$$

Usually, we are more interested in the value $X(\omega)$ than in the elements $\omega$ of
the underlying set $\Omega$. If I want to sell shoes in this country, the distribution of
the shoe size is the only thing I need to know about the inhabitants in order to
order the right amount of the different sizes; so I need to know something about
the shoe sizes $-X(\omega)$ for each inhabitant $\omega \in \Omega$. 

35
4.2 Distribution function

4.2.1 Introduction and definition

One of the possibilities to describe a stochastic variable is the distribution function that indicates how the probability mass is distributed.

Definition 4.2.1: Distribution function

If \( X \) is a stochastic variable, then the function \( F_X \), with

\[
F_X(a) = P(X \leq a)
\]  \hspace{1cm} (4.1)

is called the distribution function of \( X \) (also called the cumulative distribution function).

Note that this function can be discrete or continuous.

Example 4.2.1: \( F_X \) is discrete

\[
\Omega = \{1 \text{ dot, 2 dots, ..., 6 dots}\},
\]

\[
X(k \text{ dots}) = k \text{ for } k = 1, 2, \ldots, 6
\]  \hspace{1cm} (4.2)

\[
P(X < 1) = 0
\]

\[
P(X \leq k) = \frac{k}{6} \quad \forall k \in \{1, 2, \ldots, 6\}
\]

\[
P(X > 6) = 0
\]  \hspace{1cm} (4.3)

This is shown graphically in Figure 4.1.

![Figure 4.1: The distribution function of a die.](image-url)
4.2.2 Properties of a distribution function

The properties of the distribution function are given below.

Property 4.2.1:
$0 \leq F_X(a) \leq 1$: all probabilities are between zero and one.

Property 4.2.2:
The distribution function is monotonically nondecreasing:

\[ a \leq b \Rightarrow F_X(a) \leq F_X(b) \quad (4.5) \]

Property 4.2.3:
$F_X$ is right-continuous:

\[ \lim_{\epsilon \to 0} F_X(a + \epsilon) = F_X(a) \quad (4.6) \]

When approaching from the left there can be a discontinuity.

Property 4.2.4:

\[ \lim_{x \to -\infty} F_X(x) = 0 \quad \lim_{x \to \infty} F_X(x) = 1 \quad (4.7) \]

Property 4.2.5:
The probability of a half-open interval is given by:

\[ P(a < X \leq b) = P(x \leq b) - P(x \leq a) \]
\[ = F_X(b) - F_X(a) \quad (4.8) \]

\[ P(X > a) = 1 - P(X < a) \]
\[ = 1 - F_X(a) \quad (4.9) \]
CHAPTER 4. STOCHASTIC VARIABLES AND THEIR PROBABILITY DISTRIBUTION

Property 4.2.6:
If we perform an affine transformation \( Y = aX + b \) on a stochastic variable \( X \), then the distribution function of \( Y \) is:

- if \( a > 0 \):
  \[
  F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y - b}{a}\right) = F_X\left(\frac{y - b}{a}\right)
  \]

- if \( a < 0 \):
  \[
  F_Y(y) = P\left(X \geq \frac{y - b}{a}\right) = 1 - P\left(X < \frac{y - b}{a}\right) = 1 - F_X\left(\frac{y - b}{a}\right) + P\left(X = \frac{y - b}{a}\right)
  \]

4.2.3 Discrete and continuous distribution functions

For simplicity of the discussions in the rest of this book, we consider the discrete and continuous distribution functions as 2 separate cases.

Definition 4.2.2:
A stochastic variable \( X \) is discrete if \( X \) can only adopt a finite number, or an countable infinite number of different values. In other words: \( X \in \{ x_i | i = 1, 2, \ldots \} \)
so that:

\[
\begin{align*}
P(X = x_i) &= p_i, \\
\sum_{i=1}^{\infty} p_i &= 1
\end{align*}
\]

The probability distribution of such a distribution can be represented by a bar chart; in each point \( x_i \) a bar of length \( p_i \) is raised. Then the distribution function \( F_X \) is a piecewise, constant function with discontinuities at the points \( x_i \) of magnitude \( p_i \). An example is given in Figure 4.2.
Definition 4.2.3:
A stochastic variable $X$ is continuous, if the distribution function $F_X$ is differentiable at each point (except, possibly, in a finite number of points).

\[ f_X(x) = \frac{dF_X(x)}{dx} \quad \text{(4.18)} \]

or

\[ F_X(x) = \int_{-\infty}^{x} f_X(t) dt \quad \text{(4.19)} \]

$f_X(x)$ is called a probability density function (pdf).

Remark: Since $F_X$ is monotonically nondecreasing: $f_X(x) \geq 0$, $\forall x$. Since the total probability mass is equal to 1, the tails of $f_x$ need to go to 0.

An example of a continuous distribution is given in Figure 4.3.
CHAPTER 4. STOCHASTIC VARIABLES AND THEIR PROBABILITY DISTRIBUTION

Figure 4.3: Example of a continuous distribution.
These plots were created using the function disttool in MATLAB™.

4.2.4 Nonlinear transformation of a continuous stochastic variable

Consider a continuous stochastic variable $X$ with probability density function $f_X(x)$. Next we transform this variable through a static nonlinear function:

$$y = g(x)$$ (4.20)

The probability density function of $Y$ becomes

$$f_Y(y) = \left| \frac{\partial y}{\partial x} \right| f_X(x)$$ (4.21)

Example 4.2.2:
Consider $y = x^3$, then

$$f_Y(y) = \frac{1}{3x^2} f_X(x) \quad \text{with } x = \sqrt[3]{y}$$ (4.22)

or

$$f_Y(y) = \frac{1}{3 \sqrt[3]{y^2}} f_X(\sqrt[3]{y})$$ (4.23)

If multiple solutions exist for $y = g(x)$, one needs to add the corresponding probability densities.
CHAPTER 4. STOCHASTIC VARIABLES AND THEIR PROBABILITY DISTRIBUTION

Example 4.2.3:
Consider \( y = x^2 \). Then the solutions are \( x_1 = -\sqrt{y} \) and \( x_2 = \sqrt{y} \), with

\[
\left| \frac{\partial y}{\partial x} \right| = |2x| = |2\sqrt{y}|
\]

(4.24)

so that

\[
f_Y(y) = \frac{fx(-\sqrt{y}) + fx(\sqrt{y})}{|2\sqrt{y}|}.
\]

(4.25)

These relations can be generalized to (non) linear dynamical systems, but this is outside the scope of this course.
Chapter 5

Probability vectors and independent stochastic variables
Chapter 5

Probability vectors and independent stochastic variables

In this chapter, multidimensional stochastic variables are introduced. Next we will learn to understand the difference between dependent and independent stochastic variables.

Introduced concepts:

- multidimensional stochastic variables
- two-dimensional distributions and probability density functions
- marginal distributions
- (in)dependent stochastic variables

5.1 Multidimensional stochastic variables

In many cases, one can assign more than one real value to an element of the sample space $\Omega$. 
CHAPTER 5. PROBABILITY VECTORS AND INDEPENDENT STOCHASTIC VARIABLES

Example:

- During an analysis of Belgian inhabitants one is interested in the length as well as in the weight of each inhabitant.

- To model a vibrating mechanical structure, one measures multiple quantities at the same time: the strength, the acceleration.

- To measure an impedance, the current and the voltage are measured.

Definition 5.1.1: $n$-dimensional stochastic variables

If we consider $n$ properties at the same time, we have a vector function $Z : \Omega \to \mathbb{R}^n$. If the components $X_1, X_2, \ldots, X_n$ of this vector function are stochastic variables, we call $Z$ an $n$-dimensional stochastic variable, or probability vector.

For simplicity, we limit ourselves to the case where $n = 2$.

5.2 Two-dimensional distribution functions

Definition 5.2.1: Two-dimensional distribution functions

The distribution function $F_Z$ of a two-dimensional probability vector $Z = (X, Y)$ is

$$F_Z(a, b) = P(X \leq a \text{ and } Y \leq b)$$  \hspace{1cm} (5.1)

$$= P(\omega \in \Omega : X(\omega) \leq a \text{ and } Y(\omega) \leq b)$$ \hspace{1cm} (5.2)

As for a one-dimensional distribution, we can consider discrete and continuous distributions. For continuous distributions, however, there is an extra requirement: all second ($n^{th}$ if $n$ dimensions) mixed partial derivatives of this distribution function must be continuous.
CHAPTER 5. PROBABILITY VECTORS AND INDEPENDENT STOCHASTIC VARIABLES

5.3 Two-dimensional density functions

As for the one-dimensional distribution, we can define a probability density function for the two-dimensional distributions. This is a measure for how the probability mass is distributed over the plane. For discrete distributions this is given by giving the probability of the individual points on the plane (that constitute a countable set): \( P(Z = z_i) = p_i \). For all other points the probability is 0.

Definition 5.3.1: Two-dimensional density function for continuous distributions

The probability density function \( f_Z(x, y) \) is given by:

\[
 f_Z(x, y) = \frac{\partial^2 F_Z(x, y)}{\partial x \partial y} \tag{5.3}
\]

Remarks:

- If \( f_Z \) is known, we can find the distribution function via integration:

\[
 F_Z(x, y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} f_Z(u, v) dv \tag{5.4}
\]

- The probability of \( Z \in A \) for each measurable subset \( A \subset \mathbb{R} \):

\[
 P(Z \in A) = \int \int_A f_Z(x, y) dx \, dy. \tag{5.5}
\]

5.4 Marginal distributions

Consider the probability vector \( Z = (X, Y) \) with known distribution function \( F_Z \) and probability density function \( f_Z \). Can we establish from this knowledge the distribution function and the probability density function of \( X \) for any value of \( Y \). In other words, can we eliminate \( Y \) in the description of \( X \). The answer is yes, and is given by the marginal distributions.
CHAPTER 5. PROBABILITY VECTORS AND INDEPENDENT STOCHASTIC VARIABLES

Definition 5.4.1: Marginal probability distributions

\[ F_X(a) = P(X \leq a) \quad (5.6) \]
\[ = P(X \leq a \text{ and } Y < \infty) \quad (5.7) \]
\[ = \lim_{y \to \infty} F_Z(a, y). \quad (5.8) \]

For continuous distributions:

\[ F_X(x) = \int_{-\infty}^{x} du \int_{-\infty}^{\infty} f_Z(u, v)dv. \quad (5.9) \]

Definition 5.4.2: Marginal probability density function for a continuous distribution

\[ f_X(x) = \frac{dF_X(x)}{dx} = \int_{-\infty}^{\infty} f_Z(x, v)dv. \quad (5.10) \]

5.5 Independent stochastic variables

The concept of 'independent events' (Section 3.5) lead to 'independent stochastic variables'.

Definition 5.5.1: Independent stochastic variables

Two stochastic variables \(X\) and \(Y\) are called independent if the events

\[(a_1 < X \leq b_1) \text{ and } (a_2 < Y \leq b_2) \quad (5.11)\]

are independent for all \(a_1, b_1 \in \mathbb{R}\), or, equivalently (see property 3.6), if

\[ P((a_1 < X \leq b_1) \cap (a_2 < Y \leq b_2)) = P(a_1 < X \leq b_1) \cdot P(a_2 < Y \leq b_2) \quad (5.12) \]
CHAPTER 5. PROBABILITY VECTORS AND INDEPENDENT
STOCHASTIC VARIABLES

Property 5.5.1: The distribution function of independent variables
The components of a two-dimensional probability vector $Z = (X, Y)$ are independent if and only if the distribution function of $Z$ is the product of the marginal distribution functions:

$$F_Z(a, b) = F_X(a)F_Y(b). \quad (5.13)$$

For continuous distributions this is also valid for the probability density functions:

$$f_Z(x, y) = f_X(x)f_Y(y). \quad (5.14)$$

Exercise 5.5.1:
If $X$ and $Y$ are two independent stochastic variables with a continuous distribution, then the probability density function of $X + Y$ is the convolution of the densities of $X$ and $Y$:

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(v)f_Y(u-v)\,dv = f_X * f_Y = f_Y * f_X. \quad (5.15)$$

This property will be studied during the exercises.
Example 5.5.1:
In Figure 5.1 the graphical representation of the (in)dependent stochastic variables is given. If the variables are independent, we get a cloud of points (shape of an ellipse with the principle axes parallel to the axes). The knowledge of one variable does not tell us anything about the value of the other variable. For dependent variables we have a sloping ellipse (for linear dependencies) or even a more complicated relation.

Figure 5.1: Example of independent (left) and dependent (right) stochastic quantities.
Chapter 6

Measures of location, spread and shape
Chapter 6

Measures of location, spread and shape

In this chapter we will show how a distribution can be characterized with a small number of 'easily' measurable quantities. The following concepts are introduced:

- expected value, median, mode,
- variance,
- higher order moments.

6.1 Introduction

In chapter 2 (Descriptive statistics), several measures of location and spread, that were calculated using a finite number of realizations of a stochastic quantity, were already introduced. Here, these quantities are defined again, but now based on the probability distribution and the probability density function. Next we will introduce the measures of shape. These give us information about the obliquity and the compactness of the distribution.
CHAPTER 6. MEASURES OF LOCATION, SPREAD AND SHAPE

6.2 Measures of location

The first series of measures that we will redefine are mean (here called expected value), median and mode.

6.2.1 Expected value $E\{X\}$

Definition 6.2.1: Expected value

The expected value is

$$E\{X\} = \int_{-\infty}^{\infty} xf_X(x) \, dx.$$  \hspace{1cm} (6.1)

For continuous distributions one can write

$$E\{X\} = \int_{-\infty}^{\infty} x f_X(x) \, dx.$$  \hspace{1cm} (6.2)

For discrete distributions this becomes

$$E\{X\} = \sum_{i=1}^{\infty} x_i p_i.$$  \hspace{1cm} (6.3)

Example 6.2.1: Lottery

There are 1000 lottery tickets that each cost 1 euro. When you draw the winning ticket you get a jackpot of 400 euros. Besides this, there are 5 consolation prizes of 20 euros each. All the others get nothing. What is the average profit?

Solution: $\Omega$ is the set of 1000 lottery tickets. $X(\omega)$ is the profit that one can make when buying one ticket:

$$X(\omega) = \begin{cases} 400 - 1 \text{ euro} & \text{winning ticket,} \\ 20 - 1 \text{ euro} & \text{consolation prize,} \\ -1 \text{ euro} & \text{other tickets} \end{cases}. \hspace{1cm} (6.4)$$
CHAPTER 6. MEASURES OF LOCATION, SPREAD AND SHAPE

As a result

\[ E\{X\} = (400 - 1) \frac{1}{1000} + (20 - 1) \frac{5}{1000} - 1 \frac{994}{1000} \]  \hspace{1cm} (6.5)

\[ = -0.5 \text{ euro}. \]  \hspace{1cm} (6.6)

In plenty of cases, one is interested in the expected value of a transformed quantity. Consider the function \( g : \mathbb{R} \to \mathbb{R} \), then we can define a new stochastic variable \( g(X) \). The expected value is then

\[ E\{g(x)\} = \int_{-\infty}^{\infty} g(x)f_x(x)dx, \]  \hspace{1cm} (6.7)

for continuous distributions. Similar definitions can be given in other situations.

Example 6.2.2:

Consider \( \theta \), equally distributed over \([0, 2\pi]\). \( (f_{\theta}(x) = \frac{1}{2\pi} \text{ if } 0 \leq \theta < 2\pi \text{ and } 0 \text{ elsewhere}) \). What is the average value of \( \sin \theta \)?

Solution: \( E\{\sin (\theta)\} = \int_{-\infty}^{\infty} \sin(x)f_{\theta}(x)dx = \int_{0}^{2\pi} \frac{\sin x}{2\pi} dx = 0. \)

6.2.1.1 Properties of the expected value

Considering the linearity of the integration, the expected value has a lot of useful properties:

Property 6.2.1:

\( E\{aX + b\} = aE\{X\} + b \)

Property 6.2.2:

\( E\{X + Y\} = E\{X\} + E\{Y\} \)

Property 6.2.3:

\( |E\{X\}| \leq E\{|X|\} \)

Property 6.2.4:

For independent variables \( X \) and \( Y \): \( E\{XY\} = E\{X\} \cdot E\{Y\} \)
6.2.2 The median

The median of a distribution is (roughly) a point on the x-axis that is situated in such a way, that on both sides of this point an equal probability mass is found. Special care needs to be taken when this point does not exist ($F_X$ is discontinuous and jumps over the point 0.5), or when the point is not unambiguous (the median becomes an interval).

Example 6.2.3:

In the following figure, the median is drawn for a normal distribution and a chi-square distribution (see later).

![Diagram](attachment:image.png)

(a) A symmetrical distribution (the normal distribution).

(b) A nonsymmetrical distribution (the chi-square distribution).

Figure 6.1: Example of a median.
Both examples were generated in MATLAB via the instruction `disttool`.
6.2.3 The mode

The mode is the value that appears most often in a set of data.

- When $X$ is discrete, the mode is the $x_j$ for which $p_j = P(X = x_j)$ reaches a maximum.

- When $X$ is continuous, $\text{mod}(X)$ is the point where $f_X$ reaches its absolute maximum.

The mode is not necessarily unambiguous! $f_X$ can have more than one absolute maximum and a discrete distribution can reach its maximum in more than one point (e.g. a fair die).

6.3 Measures of spread

6.3.1 Definitions

The main measure of spread is the variance.

Definition 6.3.1: Variance

The variance of $X$ is:

$$\text{Var} \{X\} = E \{(X - E\{X\})^2\}. \quad (6.8)$$

For example, for a continuous and a discrete distribution this becomes:

$$\text{Var} \{X\} = \int_{-\infty}^{\infty} (x - E\{X\})^2 f_X(x)dx \quad (6.9)$$

and

$$\text{Var} \{X\} = \sum_{k=0}^{\infty} p_k (x_k - E\{X\})^2. \quad (6.10)$$

Note that the variance only exists if the corresponding integral or infinite sum does not diverge.
CHAPTER 6. MEASURES OF LOCATION, SPREAD AND SHAPE

The standard deviation of \(X\) is \(\sigma_X = \sqrt{\text{Var}(X)}\). The variance is mostly written as \(\sigma^2_X\). Note that \(\sigma_X\) and \(X\) have the same dimensions.

The variance indicates how close \(X\) stays to its expected value. The smaller the variance, the less measurements are spread around their average value.

One can easily see that \(\text{Var}(X) = E\left\{ (X - E\{X\})^2 \right\} = E\{X^2\} - E\{X\}^2\). However, this is particularly important for theoretical insights. From a numerical point of view, it is not recommended to use this last expression because it is possible that one has to calculate the difference between 2 large, but approximately equal numbers. This results in large rounding errors.

6.3.2 Properties

The following properties can be shown:

Property 6.3.1:
\(\text{Var}\{aX + b\} = a^2 \text{Var}\{X\}\)

Property 6.3.2:
For independent variables \(X\) and \(Y\): \(\text{Var}\{X + Y\} = \text{Var}\{X\} + \text{Var}\{Y\}\)

Property 6.3.3:
\(\text{Var}\{X \cdot Y\} = \text{Var}\{X\} \cdot \text{Var}\{Y\} + E\{X\}^2 \text{Var}\{Y\} + \text{Var}\{X\} E\{Y\}^2\)

The variance is a measure for the magnitude of the range where we expect the largest part of the probability mass to be. This is formulated in the following result:

Theorem 6.3.1: Chebyshev’s inequality

If \(X\) is a stochastic variable with mean \(\alpha_1\) and variance \(\sigma^2\), then the following inequality is valid for each \(\lambda > 0\):

\[
P(|X - \alpha_1| \geq \lambda) \leq \frac{\sigma^2}{\lambda^2}.
\] (6.11)
CHAPTER 6. MEASURES OF LOCATION, SPREAD AND SHAPE

Proof. We only prove the case in which $X$ is continuously distributed.

\[ \sigma^2 = \int_{-\infty}^{\infty} (x - \alpha_1)^2 f_X(x)dx \]  \hspace{1cm} (6.12)

\[ \geq \int_{-\infty}^{\alpha_1-\lambda} (x - \alpha_1)^2 f_X(x)dx + \int_{\alpha_1+\lambda}^{\infty} (x - \alpha_1)^2 f_X(x)dx \]  \hspace{1cm} (6.13)

For $x \leq \alpha_1 - \lambda$ and $x \geq \alpha_1 + \lambda$ applies $|x - \alpha_1|^2 \geq \lambda^2$. Furthermore $f_X(x)$ is never negative. So:

\[ \sigma^2 \geq \int_{-\infty}^{\alpha_1-\lambda} \lambda^2 f_X(x)dx + \int_{\alpha_1+\lambda}^{\infty} \lambda^2 f_X(x)dx \]  \hspace{1cm} (6.14)

\[ = \lambda^2 P(X \leq \alpha_1 - \lambda) + \lambda^2 P(X \geq \alpha_1 + \lambda) \]  \hspace{1cm} (6.15)

\[ = \lambda^2 P(|X - \alpha_1| \geq \lambda) \]  \hspace{1cm} (6.16)

Remarks

- Mostly the Chebyshev inequality is very conservative, in other words, for most of the distributions the limit could be much sharper. In this case one can make much better estimations if one has priory knowledge about the true distribution. In general, however, one cannot make the limit sharper because distributions exist for which the equality is valid.

- In the inequality one assumes that the mean value and the variance are perfectly known. If these quantities need to be found via measurements (in other words, if one needs to estimate the mean value and the variance), the lower limit needs to be adjusted. In that case, the limit becomes even less sharp. The complete expression can be found e.g. in the book of Kendall\&Stuart, part 1 of The Advanced Theory of Statistics (present in library).
CHAPTER 6. MEASURES OF LOCATION, SPREAD AND SHAPE

6.4 Measures of shape

6.4.1 Introduction and definition

Besides the first (mean value) and second (variance) order moment, one can also introduce higher order moments. These will give more detailed information about the shape of the distribution. They are also present during stochastic calculations, e.g. the calculation of the variance of the variance results in fourth order moments. One can also calculate these higher order moments around the mean value (central moments) or around 0 (raw moments, also called crude moments).

Definition 6.4.1: Central and raw higher order moments

For all \( k \in \mathbb{N} \) we define a raw moment \( \alpha_k \) and a central moment \( \mu_k \) of order \( k \) by

\[
\alpha_k (X) = \mathbb{E} \{ X^k \} \\
\mu_k (X) = \mathbb{E} \{ (X - \mathbb{E} \{ X \})^k \}
\]

(6.17) \hspace{1cm} (6.18)

Note that:

\[
\alpha_1 (X) = \mathbb{E} \{ X \} \\
\mu_1 (X) = 0 \\
\mu_2 (X) = \text{Var} \{ X \} = \alpha_2 (X) - \alpha_1 (X)^2 \\
\mu_3 (X) = \alpha_3 (X) - 3\alpha_1 (X)\alpha_2 (X) + 2\alpha_1 (X)^3
\]

(6.19)
6.4.2 The skewness

The 3rd central order moment \( \mu_3(X) \) describes the skewness of the distribution.

For a symmetrical distribution \( f_X(-x) = f_X(x) \) this is zero (prove as an exercise).

If the distribution has a long wide tail to the left then \( \mu_3(X) < 0 \), for a long wide tail to the right applies \( \mu_3(X) > 0 \).

One often uses a normalized measure: the skewness coefficient

\[
\gamma_1(X) = \frac{\mu_3(X)}{\sigma_X^3}.
\] (6.20)

6.4.3 The kurtosis

In the 4th order moments, the contribution of the tails \( \left| \frac{x - \mu(X)}{\sigma_X} \right| > 1 \) is larger than at the lower order moments (because of \( .^4 \)). So, \( \mu_4(X) \) is a measure for the 'thickness' of the tails. Again, one can make the quantity dimensionless (by normalizing):

\[
\gamma_2(X) = \frac{\mu_4(X)}{\sigma_X^4} - 3.\] (6.21)

Often, 3 is subtracted because one uses the normal distribution (see later) as a reference. Therefore \( \gamma_2(X) = 0 \). If \( \gamma_2(X) > 0 \) the distribution has thicker tails than the normal distribution (the probability mass is not centered).
Chapter 7

Covariance and correlation coefficient
Chapter 7

Covariance and correlation coefficient

In this chapter we will introduce the covariance matrix. This matrix describes the second order moments of multivariable stochastic quantities. The covariance matrix is a very important factor when calculating uncertainties.

Introduced concepts:

- covariance matrix
- correlated and uncorrelated quantities
- calculation of the spread of a function of random variables
7.1 The covariance matrix

Consider two real stochastic variables $X, Y$. We can define the following second order moments:

- the variance of $X$

\[ \sigma_X^2 = \mathbb{E} \left\{ (X - \mathbb{E}\{X\})^2 \right\} = \mathbb{E} \{ (X - \mathbb{E}\{X\})(X - \mathbb{E}\{X\}) \} \]  

(7.1)

- the variance of $Y$

\[ \sigma_Y^2 = \mathbb{E} \left\{ (Y - \mathbb{E}\{Y\})^2 \right\} = \mathbb{E} \{ (Y - \mathbb{E}\{Y\})(Y - \mathbb{E}\{Y\}) \} \]  

(7.2)

- we can also define the covariance $\text{Cov}\{X,Y\} = \sigma_{XY}^2$:

\[ \sigma_{XY}^2 = \mathbb{E} \{ (X - \mathbb{E}\{X\})(Y - \mathbb{E}\{Y\}) \} \]  

(7.3)

All these quantities can be put into the covariance matrix:

\[ C(X,Y) = \begin{pmatrix}
\sigma_X^2 & \sigma_{XY}^2 \\
\sigma_{XY}^2 & \sigma_Y^2
\end{pmatrix} \]  

(7.4)

This is called the covariance matrix of $X$ and $Y$. 

61
7.2 Interpretation and use of the covariance matrix

The covariance matrix plays an important role in the analysis of uncertainties. Consider a continuous, differential function \( z = g(x, y) \) in the point \((x_0, y_0)\). We wish to analyze the behaviour of this function around zero \( z_0 = g(x_0, y_0) \) and for this end we will use the approximation:

\[
z = z_0 + \delta z = g(x_0 + \delta x, y_0 + \delta y) \approx g(x_0, y_0) + \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}
\]

(7.3)

where we take the derivatives in the point \((x_0, y_0)\). Define \( J = \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} \), then we have

\[
\delta z = \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = J \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}.
\]

(7.6)

Suppose that \( \text{E} \{ \delta x \} = 0, \text{E} \{ \delta y \} = 0 \), within the approximation errors that is

\[
\text{E} \{ \delta z \} = \text{E} \left\{ J \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \right\} = J \text{E} \left\{ \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \right\} = 0,
\]

(7.7)

and the variance is approximated by

\[
\sigma_z^2 = \text{E} \{ \delta z^2 \} = \text{E} \{ \delta_x \delta_x^T \}
\]

(7.8)

\[
= \text{E} \left\{ J \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \begin{bmatrix} \delta x & \delta y \end{bmatrix} J^T \right\}
\]

(7.9)

\[
= J \text{E} \left\{ \begin{bmatrix} \delta_x^2 & \delta_x \delta_y \\ \delta_x \delta_y & \delta_y^2 \end{bmatrix} \right\} J^T = JCJ^T
\]

(7.10)

This can be written as

\[
\sigma_z^2 = \left( \frac{\partial g}{\partial x} \right)^2 \sigma_x^2 + \left( \frac{\partial g}{\partial y} \right)^2 \sigma_y^2 + 2 \frac{\partial g}{\partial x} \frac{\partial g}{\partial y} \sigma_{xy}. \tag{7.11}
\]
CHAPTER 7. COVARIANCE AND CORRELATION COEFFICIENT

This proves that the variance of \( z \) is not only determined by \( \sigma_x^2, \sigma_y^2 \), but also by \( \sigma_{XY}^2 \).

Example 7.2.1: Special case \( g(x, y) = x + y \)

Consider two sound sources \( x, y \) that both contribute to the total sound in a given point in space. What is the power (this is proportional to the variance of the noise) that we observe in this point?

From (7.11) with \( g = x + y \) one can find immediately

\[
\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{XY}^2. \tag{7.12}
\]

We consider 3 different situations:

- The sources \( x, y \) are uncorrelated: \( \sigma_{XY}^2 = 0 \) (see next section).
  - In this case we have: \( \sigma_z^2 = \sigma_x^2 + \sigma_y^2 \).
  - So we need to add the powers. If \( \sigma_x = \sigma_y = 1 \), then \( \sigma_z = \sqrt{2} \).

- Anti sound: \( x = -y \).
  - In this case: \( \sigma_{XY}^2 = -\sigma_x^2 = -\sigma_y^2 \) and \( \sigma_z^2 = 0 \).
  - The sources extinguish each other.

- Coherent sound: \( x = ky \), with e.g. \( k = 1 \).
  - In this case: \( \sigma_{XY}^2 = \sigma_x^2 = \sigma_y^2 \)
  - \( \sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{XY}^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_X^2 = 4\sigma_x^2, \) or \( \sigma_z = 2\sigma_x \).
  - In that case the sound doubles.
CHAPTER 7. COVARIANCE AND CORRELATION COEFFICIENT

7.3 The correlation coefficient

The normalized covariance is given by the correlation coefficient $\rho$:

$$\rho = \frac{\sigma_{XY}}{\sigma_x \sigma_y}.$$  \hspace{1cm} (7.13)

Note that this is a dimensionless quantity.

**Theorem 7.3.1:**

The correlation coefficient $\rho$ is limited: $-1 \leq \rho \leq 1$

**Proof.** Consider the following semi definite, positive, quadratic form in $a$ and $b$:

$$\forall a, b \in \mathbb{R} :$$

$$0 \leq E \left\{ (a(X - \mu_X) + b(Y - \mu_Y))^2 \right\} = a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2ab$$

$$= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}.$$  \hspace{1cm} (7.14)

The determinant of this quadratic form should be positive:

$$\sigma_X^2 \sigma_Y^2 - (\sigma_{XY})^2 \geq 0$$  \hspace{1cm} (7.16)

or

$$\sigma_X^2 \sigma_Y^2 \geq \sigma_{XY}^2$$  \hspace{1cm} (7.17)

The situation $\rho = 0$ corresponds to uncorrelated quantities.
CHAPTER 7. COVARIANCE AND CORRELATION COEFFICIENT

Definition 7.3.1: Uncorrelated quantities
Two stochastic variables $X$ and $Y$ are uncorrelated if $\sigma_{XY}^2 = 0$.

Note that in general, the uncorrelated quantities are not independent. The only exceptions are quantities with a normal distribution (see Chapter 9).

Example 7.3.1: Uncorrelated quantities that are not independent
Consider $y = x^2$, where $x$ has a symmetric probability density function with a mean value $E\{x\} = 0$. Then one can find immediately that

\[
\sigma_{XY}^2 = E\{(y - E\{y\})(x - E\{x\})\} \tag{7.18}
\]

\[
= E\{(x^2 - \alpha x^2)\} \tag{7.19}
\]

\[
= E\{x^3\} - \alpha x^2 E\{x\} = 0. \tag{7.20}
\]

So the quantities are uncorrelated. On the other hand, it is clear that they are completely dependent, the knowledge of $x$ admit to perfectly calculate $y$!

Example 7.3.2: Correlated random variables
In the following Figure 7.1, a $x,y$-plot of 2 random variables with different correlations is made. The spread of $x$ and $y$ is chosen to be equal to 1 for all plots.
Figure 7.1: Stochastic variables with different correlations.
Chapter 8

Discrete distributions
Chapter 8

Discrete distributions

In this chapter, a number of discrete distributions are introduced. We emphasize the origin of those distributions, sketch the problems they originate from, and show their mutual interconnection.

Introduced concepts:

- Bernoulli-experiment
- binomial distribution
- hypergeometric distribution
- geometric distribution
- Poisson distribution
- exponential distribution

8.1 Introduction

In the next two chapters we will briefly discuss a number of probability density functions. This is important for several reasons. On the one hand this gives a clearer view on the behaviour of stochastic variables, on the other hand the knowledge is directly applicable to a large number of practical problems. We
CHAPTER 8. DISCRETE DISTRIBUTIONS

split the distributions in 2 groups: the discrete distributions and the continuous distributions (these are discussed in the following chapter).

These 2 chapters are certainly not complete. There exist a huge range of distributions. These can be found mostly in physical problems that are studied and described in detail in the literature. Considering the limited time of this course, we cannot handle all of them. It is however very important to know that standard solutions exist for a large number of problems. Therefore, every time you need to describe a stochastic behaviour, we highly recommend you to verify thoroughly if a certain distribution is described in the literature that can solve your problem because this can save you months or even years of time. For this reason, we will pay a lot of attention to the physical interpretation of these distributions while studying them: what kind of problems do they describe? This will always be done by some examples. Next we will formalize these examples with the help of some axioms that define precisely when the distributions are valid. At last, the main properties of the distributions will be given.

In this chapter we will study 2 families of discrete distributions. Each of these contain a number of distributions that are strongly connected to each other.

The first family describes what happens if we take samples from a population (the binomial distribution, the hypergeometric distribution and the geometric distribution). The second family (the Poisson distribution and the exponential distribution) is very often used in ‘wait’ problems (e.g. the length of the queue at a counter, radioactive decay, etc). For both families, we will start by giving the link between the distributions, next we will discuss the individual distributions in more detail.
CHAPTER 8. DISCRETE DISTRIBUTIONS

8.2 The binomial, hypergeometric and geometric distribution

8.2.1 Introduction

The first family of distributions that we will study verifies what happens if we take a data sample in 'qualitative' populations. In this case there are often 2 complementary categories of which one wants to know more: defect ↔ good, sufficient ↔ inadequate, success ↔ failure. A sample of such a population is a Bernoulli-experiment, and this is obviously described by a sample space with only 2 elements \( \Omega = \{0, 1\} \). When 1 appears with a probability \( p \), then the alternative 0 appears with a probability \( q = 1 - p \). The question is: what happens if we take several samples? There are 2 possibilities:

- The population from which we sample is infinitely big. In this case we can assume that the population does not change by taking a sample. As a consequence, it is reasonable to accept that the successive samples are independent (if the sample is taken blindly). This will lead to the binomial distribution.

- The population from which we sample is small. In these circumstances the previous assumption is not valid anymore: taking a sample will affect the population and consequently we cannot assume anymore that the successive samples are independent. This will lead to the hypergeometric distribution. It is clear that the latter will converge to the binomial distribution if the population grows with respect to the number of samples.

Another question is: how many samples do I have to take before I will make a first successful experiment? E.g. how many light bulbs do I have to test before I can find one that is broken? More in general, what is the distance between two successful events in a binomial process? The answer is given by the geometric distribution.
CHAPTER 8. DISCRETE DISTRIBUTIONS

Some problems that can be solved with such distributions are described below.

Binomial distribution:

- Quality control: what fraction of products is good? How many products do I have to test to achieve a given accuracy? What fraction of a battery production can deliver a current of 100 mA for at least 10 hours?

- Production planning: the final product consists of a combination of several basic choices e.g. is an option present or not (radio, airco, airbags,...)?

- Example: a complex electronic system is built with a certain amount of backup units to increase its reliability. How many backup units need to be foreseen?

Hypergeometric distribution:

- The applications of the hypergeometric distribution are similar to those of the binomial distribution, with one big difference: the population and the variation of the population when taking 1 (n) sample(s) is not negligible. Knowledge of the properties of the hypergeometric distribution allows us to decide under what conditions (how large does the population need to be?) one can use the simpler binomial distribution.

- Example: a limited number of large transformers is sent. Some of them are thoroughly tested before all of them can be sent.

Geometric distribution:

- By using the geometric distribution one can check how many samples need to be taken before an event can occur in a series of Bernoulli experiments with a certain probability (see explanation above). Such questions frequently appear in reliability analysis.
CHAPTER 8. DISCRETE DISTRIBUTIONS

- What is the lifespan of a solar panel of a satellite, if there is a probability
  \( p \) of an impact of one micrometeorite a day?

These 3 distributions are formalized and discussed in more detail below.

8.2.2 A Bernoulli-experiment

Consider an experiment with:

- a binary property, formally represented by a sample space \( \Omega = \{0, 1\} \).
- \( P(X = 1) = p \)
- \( P(X = 0) = q = 1 - p \)

Then we say that \( X \) has a Bernoulli distribution with a chance of 'success' \( p \), and is written as:

\[
X \sim \mathcal{B}(1, p). \tag{8.1}
\]

Exercise 8.2.1:
Calculate the mean value and the variance of \( X \).

Solution:

\[
E(X) = p \quad \text{and} \quad \sigma_X^2 = p(1 - p)
\]

Mostly, the importance of a Bernoulli-experiment is the fact that it can be considered as a base for more complex problems, that, as discussed before, give rise to the binomial and the hypergeometric distribution.

8.2.3 The binomial distribution

Consider \( n \) independent iterations of a Bernoulli-experiment (e.g. throw a coin \( n \) times). What is the probability of \( k \) successful experiments \( (X = 1) \)?

This can also be described by considering the sum of the outcomes \( X_i \):

\[
Y = X_1 + X_2 + \ldots + X_n \tag{8.2}
\]
CHAPTER 8. DISCRETE DISTRIBUTIONS

Definition 8.2.1: Binomial distribution

Consider \( n \) independent Bernoulli experiments \( X_i \sim \mathcal{B}(1,p) \).

Then, the sum \( Y = X_1 + X_2 + \ldots + X_n \) has a binomial distribution: \( Y \sim \mathcal{B}(n,p) \).

Example 8.2.1: Binomial distribution: \( Y \sim \mathcal{B}(10,p) \)

In Figure 8.1, the course of the binomial distribution is drawn for different values of \( p \) and \( n = 10 \). Note that this is a discrete distribution, the lines between the points are drawn to separate the different graphs.

![Figure 8.1: The binomial distribution for \( n = 10 \). The probabilities were calculated by using the instruction binopdf in MATLAB™.](image)

73
CHAPTER 8. DISCRETE DISTRIBUTIONS

Properties of a binomial distribution A number of the following properties can be found in Figure 8.1.

Property 8.2.1:

\[ P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k} \text{ with } k = 0, 1, 2, \ldots, \text{n and } \binom{n}{k} = \frac{n!}{k!(n-k)!} \]

Property 8.2.2:

\[
\begin{align*}
E\{Y\} &= E\{X_1 + X_2 + \ldots + X_n\} \\
       &= E\{X_1\} + E\{X_2\} + \ldots + E\{X_n\} \\
       &= np
\end{align*}
\]

Property 8.2.3:

\[
\begin{align*}
\text{Var}\{Y\} &= \text{Var}\{X_1 + X_2 + \ldots + X_n\} \\
                 &= \text{Var}\{X_1\} + \text{Var}\{X_2\} + \ldots + \text{Var}\{X_n\} \\
                 &= npq
\end{align*}
\]

Property 8.2.4: Skewness

\[ \gamma_1 = \frac{q - p}{\sqrt{npq}} \]

Property 8.2.5: Kurtosis

\[ \gamma_2 = \frac{1 - 6pq}{npq} \]

Properties 8.2.2 and 8.2.3 directly follow from the properties of independent variables. Note that the skewness and the kurtosis go to 0 when \( n \to \infty \).
CHAPTER 8. DISCRETE DISTRIBUTIONS

Exercise 8.2.2:
One wishes to gauge the popularity of the MOGAB party via a data sample. How many people need to be interrogated for their political opinion in a survey, if one wishes the spread of the final result to be certainly smaller than 1%?

8.2.4 The hypergeometric distribution

Consider a population with $N$ elements (note that we do not only consider a finite population, $N$ is also explicitly given). This population contains $P$ elements for which $X = 1$, and $N - P$ elements for which $X = 0$ (so we have again a binary characteristic to study). What is the probability of $k$ successful experiments ($X = 1$) when we take samples without replacement?

This can be described by considering the sum of the outcomes $X_i$:

$$Y = X_1 + X_2 + \ldots + X_n$$  \hspace{1cm} (8.3)

Definition 8.2.2: Hypergeometric distribution

Consider $n$ Bernoulli experiments $X_i \sim B(1, p)$ without replacement from the population with $N$ elements. Then the sum $Y = X_1 + X_2 + \ldots + X_n$ has a hypergeometric distribution: $Y \sim \mathcal{H}(N, p, n)$, with $p = \frac{P}{N}$.
CHAPTER 8. DISCRETE DISTRIBUTIONS

Properties of the hypergeometric distribution

Property 8.2.6:

\[ P(Y = j) = \binom{j}{r} \binom{N-p}{n-j} \frac{N}{\binom{N}{n}} \]

Property 8.2.7:

\[ E\{Y\} = np \]

Property 8.2.8:

\[ \sigma_Y^2 = npe \frac{N-n}{N-1} \]

Example 8.2.2:

Hypergeometric distribution: \( N = 40, 20 \) samples without replacement, \( k \) number of successful samples

![Graph](image)

Figure 8.2: Example of a hypergeometric distribution. The corresponding binomial distribution was plotted as a reference (dotted line). Calculated via the MATLAB\textsuperscript{™}-instruction: `hygepdf(k,N,K,20)`, with \( K = p \cdot N \) = total number of successful samples in the population.
8.2.5 The geometric distribution

The geometric distribution gives the probability of a first success with sample \( k + 1 \) after \( Y = k \) times of failure in a binomial process. Since this means that all previous samples must be a failure, we can immediately see that

\[
P(X_1 = 0, \ldots, X_k = 0, X_{k+1} = 1) = pq^k.
\]

(8.4)

The average value \( E(Y) = \frac{q}{p} \), the variance \( \sigma_Y^2 = \frac{q}{p^2} \).

Example 8.2.3: Example of a geometric distribution

![Graph showing geometric distribution for different values of p](image)

Figure 8.3: Example of a geometric distribution. Generated via the MATLAB\textsuperscript{\textregistered}-function: geopdf(k,p).
CHAPTER 8. DISCRETE DISTRIBUTIONS

8.3 The exponential and Poisson distribution

8.3.1 Introduction

How many cash desks need to be opened in a supermarket so that the queues do not become too long? How many phone calls need to be handled simultaneously (capacity of a telephone exchange)? How big is the probability that a telephone line is operational without interruption for a month? How many radioactive particles do we count per minute near radioactive waste?

All these questions are described by the same underlying probability density function: the Poisson distribution.

Associated to the previous questions one can verify the behaviour of the time interval between two events (e.g. entry of a client, start of a phone call, time between the detection of 2 radioactive particles). These processes are described by the exponential distribution.

One can consider the Poisson distribution and the exponential distribution as limit situations of respectively the binomial and the geometric distribution. These transitions are out of the scope of this course.

Instead of speaking of events (a 0 or a 1), one speaks of 'incidents' with these distributions. An incident is an event with $X = 1$. One speaks of a Poisson incident flow.

8.3.2 The Poisson distribution

With the Poisson incident flow one checks if an incident occurs during a certain time interval. One divides the time in equal subintervals and lets the width $\Delta t$ of these intervals go to 0. One speaks of a Poisson incident flow if the following 3 hypotheses are satisfied.
CHAPTER 8. DISCRETE DISTRIBUTIONS

Basic assumptions for a Poisson process

Assumption 8.3.1:
The number of incidents during two non-overlapping time intervals are stochastically independent.

Assumption 8.3.2:
The probability of 1 incident in a time interval of width $\Delta t$ is proportional to $\Delta t$ if $\Delta t$ goes to 0:

$$P(\text{1 incident in interval of width } \Delta t) = \lambda \Delta t + o(\Delta t) \quad (8.5)$$

with

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

with the speed $\lambda$ independent of the time (same for all intervals).

Assumption 8.3.3:
The probability of multiple incidents in a time interval of width $\Delta t$ is negligible if $\Delta t$ goes to 0:

$$P(\text{multiple incidents in 1 interval of width } \Delta t) = o(\Delta t) \quad (8.6)$$

with

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

These assumptions are valid in very good approximation in several processes and phenomena.

Definition 8.3.1: Poisson incident flow
Consider an incident flow that satisfies the Assumptions 8.3.1, 8.3.2 and 8.3.2.
Such a process is called a Poisson incident flow and is described by the following probability density function:

$$P_n = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad (8.7)$$
CHAPTER 8. DISCRETE DISTRIBUTIONS

with $p_n$ the probability of $n$ incidents in a time interval of width $t$. $\lambda$ is defined in (8.5).

In Figure 8.4 several examples of a Poisson distribution are shown.

Figure 8.4: Example of a Poisson probability density function: $\lambda = 0.1, 1, 10$ and $t = 1$.

Properties of a Poisson distribution

Property 8.3.1:

$E \{ Y \} = \lambda$

Property 8.3.2:

$\text{Var} \{ Y \} = \lambda$

The MATLAB instruction for the Poisson distribution is `poisspdf(X, Lambda)`.

$X$ is a natural number, $\text{Lambda}=\lambda t$. 

80
CHAPTER 8. DISCRETE DISTRIBUTIONS

8.3.3 The exponential distribution

The distance between 2 successive incidents of a Poisson incident flow is exponentially distributed.

Definition 8.3.2: Exponential distribution

The distance \( t \) between 2 successive incidents of a Poisson process is exponentially distributed with parameter \( \lambda \). The probability density function is

\[
f_T(t) = \begin{cases} 
\lambda e^{-\lambda t} & \forall t > 0 \\
0 & \forall t \leq 0 
\end{cases} \quad (8.8)
\]

And the distribution function is

\[
F_T(t) = \begin{cases} 
1 - e^{-\lambda t} & \forall t > 0 \\
0 & \forall t \leq 0 
\end{cases} \quad (8.9)
\]

The exponential distribution has a remarkable property. If one starts measuring at an arbitrary instant, then the probability distribution of the time until the next incident is independent of what happened before.

Property 8.3.3: Forgetfulness of the exponential distribution

At the instant \( t \) the period of time until the next incident is independent of the period of time between the occurrence of the previous incident and the start of the considered time interval. For \( s > 0, t > 0 \) is valid:

\[
P(T > s + t | T > t) = \frac{P(T > s + t \text{ en } T > t)}{P(T > t)} \quad (8.10)
\]

\[
= \frac{P(T > s + t)}{P(T > t)} = \frac{1 - P(T \leq s + t)}{1 - P(T \leq t)} \quad (8.11)
\]

\[
= \frac{1 - (1 - e^{-\lambda(s+t)})}{1 - (1 - e^{-\lambda t})} = \frac{e^{-\lambda s} e^{-\lambda t}}{e^{-\lambda t}} \quad (8.12)
\]

\[
= e^{-\lambda s} = 1 - P(T \leq s) \quad (8.13)
\]

\[
= P(T > s). \quad (8.14)
\]
CHAPTER 8. DISCRETE DISTRIBUTIONS

Properties of an exponential distribution

Property 8.3.4:
\[ E\{Y\} = \frac{1}{\lambda} \]

Property 8.3.5:
\[ \text{Var}\{Y\} = \frac{1}{\lambda^2} \]

The MATLAB\textsuperscript{TM}-instruction for the exponential distribution is \texttt{exppdf(t, \mu = \frac{1}{\lambda})}.

Example 8.3.1: The course of the exponential distribution

![Graph of the exponential distribution]

Figure 8.5: Example of the course of the probability density function of the exponential distribution.
Generated with the MATLAB\textsuperscript{TM}-instruction: \texttt{exppdf(t, 1/\lambda)}.

Remark: One can generalize all these results to the waiting time of the \( k \text{th} \) incident. This waiting time is described by the Gamma distribution (out of the scope of this course).
Chapter 9

Continuous distributions
Chapter 9

Continuous distributions

In this chapter a number of continuous distributions are studied. The normal
distribution $\mathcal{N}(\mu, \sigma)$ is the most important one here. This plays an important
role when establishing uncertainty intervals. In practice, the mean value $\mu$ and
the spread $\sigma$ need to be replaced by measured values. In order to understand
the behaviour of uncertainty limits under these circumstances, one needs to in-
troduce 2 additional distributions: the chi-square distribution and the Student’s t-distribution.

Introduced concepts:

- central limit theorem,
- normal distribution,
- establishing uncertainty intervals,
- behaviour of the sample mean and variance,
- the $\chi^2$-distribution and the Student’s t-distribution.
CHAPTER 9 CONTINUOUS DISTRIBUTIONS

9.1 Introduction

In this chapter we will study a number of continuous distributions. The normal distribution \( N(\mu, \sigma) \) is the most important one here. All other distributions that we will consider are introduced starting from and in function of the use of the normal distribution. The reason for this is an important theorem: the central limit theorem. Essentially, this theorem tells us that the sum of a large number of stochastic variables is asymptotically described by a normal distribution.

In practice, for an engineer, random varying quantities occur in many processes/systems, that are combined in an unknown way before they become visible at the output. Because of the central limit theorem, the disturbances are often normally distributed.

This admits, starting from a normal distribution, to give uncertainty intervals for measurements. These give an idea about how reliable a result is with respect to stochastic measurement errors. Note that the systematic measurement errors can be eliminated via calibrations with known elements! In order to generate these uncertainty intervals one needs to know the parameters of the normal distribution. The one-dimensional normal distribution is fully characterized by its mean value and its variance. These are two quantities that can easily be determined from measurements (see also Chapter 2).

However, when replacing the true values of the mean and the variance by the measured values, a new problem occurs.

- What is the quality of these measured quantities?

- What is the influence on the uncertainty intervals if one replaces a true value by a measured one?

To answer these questions, one needs to consider a number of additional distributions: the chi-square distribution and the Student’s \( t \)-distribution.

This immediately sets the structure of this chapter.
9.2 The central limit theorem

We quote first a limited form of the central limit theorem.

**Theorem 9.2.1: Central limit theorem**

If $X_1, X_2, \ldots, X_n$ are independent stochastic variables with probability distributions for which the expected value and the variance exist and are uniformly bounded,

\[
\begin{align*}
\mu_j &= \mathbb{E} \{ X_j \} \text{ with } |\mu_j| \leq M \\
\sigma_j &= \text{Var} \{ X_j \} \text{ with } \sigma_j \leq V
\end{align*}
\forall j \in \mathbb{N}.
\tag{9.1}
\]

and if $Y_n$ is their sum with expected value $\bar{\mu}_n$ and variance $\bar{\sigma}_n^2$:

\[
\begin{align*}
Y_n &= \sum_{j=1}^{n} X_j \\
\bar{\mu}_n &= \sum_{j=1}^{n} \mu_j \\
\bar{\sigma}_n^2 &= \sum_{j=1}^{n} \sigma_j^2
\end{align*}
\tag{9.2-4}
\]

then $Z_n = \frac{Y_n - \bar{\mu}_n}{\bar{\sigma}_n}$ converges, and the limit is standard-normally distributed:

\[
\lim_{n \to \infty} Z_n = W, \text{ with } W \sim \mathcal{N}(0,1).
\tag{9.5}
\]

The exact meaning of $\mathcal{N}(0,1)$ will be explained later on.

**Generalities:** There exist central limit theorems where one replaces the strict condition of independency by a less strict condition, where samples 'in each others neighbourhood' may be dependent.
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

Example 9.2.1: The central limit theorem
In Figure 9.1, the distribution for \( X_1, X_1 + X_2 \) and \( X_1 + X_2 + X_3 \) is shown, where the random variables \( X_i \) are independent and uniformly distributed \(^1\). This actual distribution is compared to the normal distribution (dashed line) \( N(0,1) \) that is reached in the limit.

![Figure 9.1: Example of the convergence of the distribution of the sum of independent random variables to a normal distribution.](image)

9.3 The normal distribution

As mentioned in the introduction, the normal distribution plays a central role in probability and in statistics, and this in theory as well as in practice.

Definition 9.3.1: Normal distribution
We say that a stochastic variable \( X \) is normally distributed with mean \( \mu \) and spread \( \sigma \), \( X \sim N(\mu, \sigma) \) if \( \frac{X-\mu}{\sigma} \sim N(0,1) \), where the probability density function of \( N(0,1) \) is given by \( \varphi(x) \), with

\[
\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad (9.6)
\]

and the distribution function \( \Phi(x) \) by

\[
\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt. \quad (9.7)
\]

\(^1\)The uniform distribution is a continuous distribution where the mass is equally distributed over a finite interval: \( P(X) = \frac{1}{b - a} \) if \( x \in [a, b] \) and \( 0 \) if \( x \notin [a, b] \).
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

The integral in \( \Phi \) is not exactly expressible in terms of elementary functions and the value of \( \Phi(x) \) for a given \( x \) will need to be calculated via numerical integration, searched via tables, or functions in numerical packets. For this, one can use the error function (ERF):

\[
\text{ERF}(X) = \frac{2}{\sqrt{\pi}} \int_0^X e^{-t^2} dt
\]

so that

\[
\Phi(x) = \frac{1}{2} \left( 1 + \text{ERF} \left( \frac{x}{\sqrt{2}} \right) \right)
\]

In MATLAB\textsuperscript{TM}, one can use the function \texttt{erf}(x).

The shape of the probability density function and of the distribution function is given in Figure 9.2.

![Probability Density and Distribution Functions](image)

(a) the probability density function. (b) the probability distribution function.

Figure 9.2: The standard normal distribution \( \mathcal{N}(0, 1) \).

Note: the probability density function of the standard normal distribution \( \mathcal{N}(\mu, \sigma) \) is given by:

\[
\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.
\]

(9.10)

9.3.1 Properties of the normal distribution \( \mathcal{N}(0, 1) \)

Property 9.3.1:

\( \mathbb{E}\{X\} = 0 \)

Property 9.3.2:
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

\[ \text{Var}\{X\} = 1 \]

**Property 9.3.3:**

\( \gamma_1 = 0 \)

**Property 9.3.4:**

\( \gamma_2 = 0 \quad \mu_4 = \mathbb{E}\{X^4\} = 3 \)

**Property 9.3.5:**

Consider 2 independent normally distributed stochastic variables \( X \sim \mathcal{N}(\mu_1, \sigma_1) \) and \( Y \sim \mathcal{N}(\mu_2, \sigma_2) \), then \( X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}) \).

This last property is also valid for dependent variables, but in that case the variance needs to be calculated via (7.12).

### 9.3.2 Establishing uncertainty intervals

A measurement result or a prediction only gets a meaning if one can also indicate how reliable the result is. E.g. everyone can predict the rate of the stock market tomorrow, but only a few can do this with a high reliability.

There are different possibilities to specify an uncertainty, e.g. via percentiles (course of the distribution function where one looks at the value of \( x \), so that for example \( P(X \leq x) = 40\% \)). Mostly, one uses the standard deviation as a measure for the uncertainty. Assume that one estimates e.g. the mean \( \mu \) via the arithmetic mean \( \bar{x} \). The problem then is that the probability \( P(\bar{x} - 2\sigma \leq \mu \leq \bar{x} + 2\sigma) \) depends on the distribution of \( \bar{x} \), in other words, one cannot say anything about the reliability. In practice however, one can usually assume, on the basis of the central limit theorem, that the measurement/prediction (\( \bar{x} \) in the example) is normally distributed. Note that exceptions exist to this rule (which we will discuss later on) where one knows from the beginning that another distribution should be used. In the remaining of this section we will limit ourselves to the normal distribution.

The question is: what is the uncertainty interval for a quantity \( x \), if we know...
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

that \( x \sim \mathcal{N}(\mu, \sigma) \)? Or: find \( a \) so that:

\[
P \left( \mu \in [\mu - a\sigma, \mu + a\sigma] \right) = 1 - \alpha.
\] (9.11)

Here, \( \alpha \) is the probability that \( \mu \) lays outside the considered interval. Such an interval is called an uncertainty interval with reliability \( 1 - \alpha \) (e.g. the 95% confidence interval).

The answer to this question can be found in tables of the normal distribution. To this end, we first need to do a normalization by replacing (9.11) by

\[
P \left( \frac{x - \mu}{\sigma} \in [-a, a] \right) = 1 - \alpha.
\] (9.12)

This allows us to use only 1 table, e.g. based on the standard table of the erf function.

Note that the confidence interval is chosen symmetrically around the mean value.

Figure 9.3: Example of a confidence interval, situated symmetrically around 0.
9.3.3 Establishing and using uncertainty intervals in practical situations

The use of the uncertainty interval, as given in (9.12), is less simple than it looks like: one needs to know the true value \( \sigma \) of the distribution of \( x \). In practice, one needs to obtain this true value via measurements, and often one only has a limited number of measurements available. The first question one needs to study is: What is the influence of replacing the true values of \( \sigma \) by its measured value?

Another frequent question is: Is a measured quantity \( x \) significantly different from 0? In other words, does \( x_0 = 0 \) belong to the confidence intervals in (9.12)? To this end, one needs to know the statistic \( \bar{x} \), e.g. \( x \) is the average of a number of measurements and \( \sigma \) is the measured spread.

These questions are treated in the following sections.

1. First we will discuss the distribution of the sample mean:

\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\] (9.13)

2. Next we will study the behaviour of the empirical sample variance:

\[
S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.
\] (9.14)

This will give rise to the chi-square distribution.

3. At last we will study the ratio \( \bar{X}_n/S_n \). This will give rise to the Student's \( t \)-distribution.

9.4 Study of the sample mean

We consider \( n \) independent variables \( X_i \sim N(\mu, \sigma) \), \( i = 1, \ldots, n \) and calculate the arithmetic mean (9.13). What is the distribution of \( \bar{X}_n \)?
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

The answer follows directly from property 9.3.5

\[ nX_n \sim N(n\mu, \sigma^2/n) \tag{9.15} \]

and as a consequence, the distribution of \( X_n \) is given by (9.16):

**Property 9.4.1: Distribution of the sample mean \( X_n \)**

\[ X_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{9.16} \]

A first consequence is that \( E\{X_n\} = \mu \).

A second consequence is that the standard deviation \( \sigma_{X_n} \) decreases with \( \frac{1}{\sqrt{n}} \). The more measurements, the more reliable the result. However, the gain grows very slowly. 100 measurements reduce \( \sigma_{X_n} \) with a factor 10, while 10000 measurements only reduce the uncertainty with a factor 100.

If the standard deviation \( \sigma \) is a priori exactly known, then it is very easy to give the uncertainty interval for \( X_n \) with the help of the probability distribution function \( \Phi \) (see (9.7)):

\[ P\left( X_n - \mu \leq \frac{r\sigma}{\sqrt{n}} \right) = P\left( \frac{X_n - \mu}{\sigma/\sqrt{n}} \leq r \right) = \Phi(r), \text{ with } \frac{X_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{9.17} \]

A more symmetric result is:

\[ P\left( |X_n - \mu| \leq \frac{r\sigma}{\sqrt{n}} \right) = \Phi(r) - \Phi(-r) = 2\Phi(r) - 1, \text{ (} r > 0 \text{).} \tag{9.18} \]

In Table 9.1 a number of typical values are given.

**Table 9.1: Uncertainty on the mean value if the spread is exactly known.**

| \( r \) | \( P\left( |X_n - \mu| \leq \frac{r\sigma}{\sqrt{n}} \right) \) |
|--------|----------------------------------|
| 1      | 0.68                             |
| 1.645  | 0.9                              |
| 2      | 0.95                             |
| 3      | 0.997                            |

92
If the standard deviation \( \sigma \) is not exactly known, but is also taken from the same data set, then the problem becomes much more complicated. This is discussed later on in Section 9.6.2.

The probability density function of \( \bar{X}_n \) is plotted in Figure 9.4 for \( n = 1, 10, 100, 1000 \).

![Figure 9.4: The probability density function of \( \bar{X}_n \) for \( n = 1, 10, 100, 1000 \).]
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

9.5 Study of the sample variance

A candidate for measuring the sample variance is given by \(2\)

\[
s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

(9.19)

and thus also depends on \(\bar{X}_n\). This strongly complicates the study. That is why we will first consider a simplified problem: what is the behaviour of

\[
\sum_{i=1}^{n} X_i^2 \text{ with } X_i \sim \mathcal{N}(0,1) ?
\]

(9.20)

This will give rise to the chi-square distribution.

9.5.1 The chi-square distribution

First we will formally define the chi-square distribution, next we will discuss an application and finally, in the next section, we will use it to characterize the sample variance.

Definition 9.5.1: The chi-square distribution: \(\chi^2_n\)

If \(Z_i \sim \mathcal{N}(0,1), i = 1, \ldots, n\) is independently distributed, then the sum of squares \(X\),

\[
X = Z_1^2 + Z_2^2 + \ldots + Z_n^2
\]

(9.21)

has a chi-square distribution with \(n\) degrees of freedom and we write:

\[
X \sim \chi^2_n.
\]

(9.22)

Properties of the \(\chi^2_n\)-distribution

Property 9.5.1:

\[E\{X\} = n\]

\(^2\text{Attention: we will see later on that one mostly uses } s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \text{ in practice.}\)
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

Property 9.5.2:
\[ \text{Var}\{X\} = 2n \]

Property 9.5.3:
If \( X \sim \chi^2_n \), \( Y \sim \chi^2_m \) are independent, then \( X + Y \sim \chi^2_{n+m} \).

Property 9.5.4:
For large \( n \) is valid: \( \chi^2_n \approx N(n, \sqrt{2n}) \).

In Figure 9.5 the probability density function is shown for a number of values of \( n \). One clearly sees that the distribution is asymmetrical, but that it converges to a normal distribution if \( n \) increases.

![Figure 9.5: The probability density function of \( \chi^2_n \) for \( n = 2, 4, 8, 16 \).](image)

The \( \chi^2 \) in \( X \) for \( V \) degrees of freedom can be calculated via the MATLAB\textsuperscript{®} instruction: `chi2pdf(X,V).

Example 9.5.1: Measurement of the RMS-value of a noise signal
In many cases the RMS-value\(^3\) of a noise signal has to be estimated. One of

\(^3\text{RMS: root mean square}\)
the possibilities is to sample the signal \( n \) times and then to calculate

\[
V_{\text{RMS}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}.
\] (9.23)

As was mentioned before, a lot of noise signals are normally distributed, in our case e.g. \( x_i \sim \mathcal{N}(0, V_{\text{RMS}_0}) \). Furthermore, if we assume that the samples \( x_i \) are not taken too fast, then we can consider them as independent. Under these conditions, it is obvious that

\[
V_{\text{RMS}}^2 \sim \frac{V_{\text{RMS}_0}^2}{n} \chi_n^2.
\] (9.24)

The scale factor is due to the fact that the \( x_i \) have a spread \( V_{\text{RMS}} \) instead of a spread 1. From this result we can directly determine the quality of the measurement:

- \( E \{ V_{\text{RMS}}^2 \} = \frac{V_{\text{RMS}_0}^2}{n} \ E \{ \chi_n^2 \} = V_{\text{RMS}_0}^2 \).

Since, per definition, the RMS-value is given by \( E \{ x^2 \} = V_{\text{RMS}_0}^2 \), it follows immediately that the expected value of the measurement converges to the true value.

- The uncertainty of a measurement follows directly from the properties of \( \chi_n^2 \):

\[
\sigma_{V_{\text{RMS}}} = \frac{V_{\text{RMS}_0} \sqrt{2/n}}{\sqrt{n}} = \sqrt{\frac{2}{n}} V_{\text{RMS}_0}.
\] (9.25)

- Thus, the uncertainty \( \sigma_{V_{\text{RMS}}} \) decreases with \( \frac{1}{\sqrt{n}} \). From these results we can design an experiment: how long do we have to measure to guarantee a certain accuracy?

This result is often used in more advanced measurements where one needs to measure the transfer function of vibrating structures (e.g. bridges, aircraft wings), electrical systems (e.g. amplifiers, satellite connection), acoustic systems (e.g. speakers, microphones), ...
9.5.2 Study of the sample variance

Now, we will study again the behaviour of

\[ s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \quad (9.26) \]

where we will now consider the additional term \( \overline{X}_n \), in comparison with Section 9.5.1.

A detailed calculation shows that

\[ E \{ n s_n^2 \} = E \left\{ \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \right\} = (n - 1) \sigma^2. \quad (9.27) \]

It appears that \( n s_n^2 \) behaves like a \( \chi_{n-1}^2 \), and this is indeed the case (prove: see lecture for detailed calculation). The reason for this is exactly the presence of the estimated value \( \overline{X}_n \) instead of the true value \( \mu \), an additional parameter is estimated by which 1 degree of freedom disappears (without prove: the number of degrees of freedom is the number of real measurements – the number of estimated real parameters). If one does not estimate the mean value but uses the true value, then one will see that \( E \left\{ \sum_{i=1}^{n} (X_i - \mu)^2 \right\} = n \sigma^2 \) (this follows directly from the properties of the normal distribution!), and \( \sum_{i=1}^{n} (X_i - \mu)^2 \) has indeed \( n \) degrees of freedom.

Conclusions:

1. A good estimator for a sample variance is

\[ S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \quad (9.28) \]

2. The quantity

\[ \frac{(n-1) S_n^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (9.29) \]

3. \( S_n^2 \) and \( \overline{X}_n \) are independent (without prove).
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

9.5.3 Confidence intervals for $S_n^2$

From (9.29) one can again establish an uncertainty interval for $S_n$, when assuming that the variance is exactly known. If we exclude at both sides of the distribution a probability mass $\alpha/2$, then we have a total mass of

$$ P \left( \chi^2_{n-1, \frac{\alpha}{2}} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{n-1, 1-\frac{\alpha}{2}} \right) = 1 - \alpha \tag{9.30} $$

for the remaining interval. From this we can of course establish an uncertainty interval for $\sigma$ starting from the measurement $S_n$:

$$ P \left( \frac{(n-1)S_n^2}{\chi^2_{n-1, 1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{n-1, \frac{\alpha}{2}}} \right) = 1 - \alpha. \tag{9.31} $$

Note that this interval is not symmetrical around $\sigma^2$ because the $\chi^2$ itself is not symmetrical either. In Table 9.2 the 95% confidence interval is given for $\sigma$ in function of the number of measurements $n$. To this end, the value of $\sigma$ was normalized to 1. The table was calculated from (9.31) by taking the square root (since we consider $\sigma$).

Table 9.2: 95% confidence interval for $\sigma = 1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>lower limit</th>
<th>upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>does not exist</td>
<td>does not exist</td>
</tr>
<tr>
<td>2</td>
<td>0.45</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>3.73</td>
</tr>
<tr>
<td>8</td>
<td>0.66</td>
<td>2.04</td>
</tr>
<tr>
<td>16</td>
<td>0.74</td>
<td>1.55</td>
</tr>
<tr>
<td>32</td>
<td>0.80</td>
<td>1.33</td>
</tr>
<tr>
<td>64</td>
<td>0.85</td>
<td>1.21</td>
</tr>
<tr>
<td>128</td>
<td>0.89</td>
<td>1.14</td>
</tr>
</tbody>
</table>

A graphical representation is given in Figure 9.6.

This calculation was performed with the help of the MATLAB™-instruction:

- `chi2inv(\alpha, n)` for the calculation of the inverse of $\chi^2_{n-1, \frac{\alpha}{2}}$. 

98
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

![Graph showing confidence intervals](image)

Figure 9.6: 95% confidence interval for the estimate of the spread from \( n \) measurements \((\sigma = 1)\).

### 9.6 Study of \( \frac{\bar{X}_n - \mu}{S_n} \)

We consider again the full question. We wish to calculate the mean value from \( n \) measurements. We assume that these are normally distributed, but we do not know the true value of the mean \((\mu)\), neither the true value of the spread \((\sigma)\).

The expressions for \( \bar{X}_n \) and \( S_n \) remain unchanged:

\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{9.32}
\]

\[
S_n^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{9.33}
\]

Thus the question is limited to generating uncertainty bounds for \( \bar{X}_n \). To this end we use the normalized quantity:

\[
T = \frac{\bar{X}_n - \mu}{S_n} \tag{9.34}
\]
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

The distribution of $T$ is studied in detail (by W. Gosset who published under the alias: Student). First we will study this distribution, next we will use these results to generate the wanted uncertainty intervals.

9.6.1 Student's $t$-distribution

Definition 9.6.1: Student's $t$-distribution ($t_n$)

If $X$ has a chi-square distribution with $n$ degrees of freedom and if $Z$ is standard normally distributed and independent of $X$, then the quotient $T'$

$$T = \frac{Z}{\sqrt{X/n}}, Z \sim \mathcal{N}(0,1), \text{ and } X \sim \chi^2_n$$

(9.35)

has a $t$-distribution with $n$ degrees of freedom, and we write: $T \sim t_n$.

Properties of the $t_n$-distribution

Property 9.6.1:

The probability distribution of $T$ is symmetrical.

- This follows from the fact that $Z$ is independent of $X$ and is symmetrically distributed.

Property 9.6.2:

The probability density function is given by (you don't have to learn this by heart):

$$f_{t_n}(x) = C_n \left(1 + \frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)}$$

(9.36)

with $C_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}$ and with $\Gamma(\cdot)$ the gamma function.

Property 9.6.3:

Only the first $n-1$ moments exist (the higher order moments do not converge).

One can see this as follows: $x^k f_{t_n}(x) \approx C_n x^{k-n-1}$ for large $|x|$. If $k > n-1$ we get the integral of $1/x^m$ at the tails, with $m < 2$.  

100
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

Property 9.6.4:
\[ E\{T\} = 0 \text{ if } n \geq 2; \text{ it does not exist for } n = 1. \]

Property 9.6.5:
\[ \text{Var}\{T\} = \frac{n}{n-2} \text{ if } n \geq 3; \text{ it does not exist for } n = 1 \text{ or } n = 2. \]

Property 9.6.6:
For large values of \( n \), the following approximation is valid: \( t_n \approx \mathcal{N}(0,1) \).

Following MATLAB®-instructions exist for the \( t \)-distribution:

- \texttt{tpdf}(x,n): admits to determine the probability distribution in \( x \) for \( n \) degrees of freedom.

- \texttt{x=inv}(P,n): determines the inverse of the cumulative \( t \)-distribution for \( n \) degrees of freedom. One can generate confidence intervals with this function.

In Figure 9.7, the distribution for a number of degrees of freedom is given. As a reference, the normal distribution is again given in the same figure.

Figure 9.7: Example of the \( t_n \)-distribution for \( n = 2, 8, 32 \). As a reference, the \( \mathcal{N}(0,1) \) (in ...) is given.
CHAPTER 9. CONTINUOUS DISTRIBUTIONS

9.6.2 Confidence intervals

With the help of the previous results, one can directly calculate the confidence interval for $\mu$.

Note that the stochastic $T = \frac{\bar{X}_n - \mu}{\bar{S}_n / \sqrt{n}}$ has a Student's $t$-distribution with $n - 1$ degrees of freedom.

- Indeed:

$$
T = \frac{\bar{X}_n - \mu}{\bar{S}_n / \sqrt{n}} = \frac{\bar{X}_n - \mu \sigma / \sqrt{n}}{S_n / \sqrt{n}} = \frac{Z}{\sqrt{n-1}},
$$

(9.37)

- with $Z \sim N(0, 1)$ and $Y = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

As in Section 9.5 we have

$$
P \left( \left| \frac{\bar{X}_n - \mu}{\bar{S}_n / \sqrt{n}} \right| \leq t_{n-1,1-\frac{\alpha}{2}} \right) = 1 - \alpha,
$$

(9.38)

or

$$
P \left( \bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \leq \mu \leq \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \right) = 1 - \alpha.
$$

(9.39)
Chapter 10

Hypothesis testing
Chapter 10

Hypothesis testing

In this chapter we will study some techniques to verify if there are statistically relevant indications to determine whether a 'hypothesis' is right or wrong.

Introduced concepts:

- formulation of hypotheses
- hypothesis and significance testing
- statistical significance and exceedance probability
- $T$-test, $\chi^2$-test, $F$-test
- errors in hypothesis testing

10.1 Introduction

In the previous chapters we have learned how to estimate a characteristic value of a stochastic quantity (e.g. the mean value) via a data sample, using a finite (limited) number of observations. Since this estimation varies from experiment to experiment, we also introduced uncertainty intervals. Such a description is informative, but does not give an answer to all questions. A further treatment of the results is necessary. We wish to know if a certain 'hypothesis' gives a plausi-
CHAPTER 10. HYPOTHESIS TESTING

ble explanation for the observed data, or maybe if there are other explanations that are more reasonable.

Example 10.1.1:
An inspector wishes to know if a baker indeed sells bread of 800 grams each? Or do these bread not weigh enough?

Answering this question is not simple. The weight of a bread varies from bread to bread. So, it is not practical to demand that each bread has a weight of exactly 800 grams. On the other hand, we expect that the average weight is 800 grams, and the variation of the weight is not too large.

If the inspector wants to test these weakened demands, he still has a problem. Assume that he measures the weight of 10 bread and finds a measured mean value ($\bar{X}_{10}$) of 780 grams and a standard deviation ($s_{10}$) of 15 grams. Can he decide from this experiment that there is a problem. Or can he formalize as a hypothesis that the baker indeed makes bread of 800 grams each? Again the answer to this question is not quite simple. Or one can conclude either that the observations are useless (the inspector accidentally picked out the 'lightweight' bread), or that the hypothesis that the bread weigh 800 grams is wrong.

The above described example is strongly simplified in order to clearly outline the problem. However, similar questions exist for a large range of problems: e.g. quality control (does a manufacturer achieve the claimed specifications), scientific research (is a proposed theory correct, or are there enough experimental data to reject it?), ...

In all these problems we can never make a fully correct statement, we can only make statements that contain a certain probability of truth. A test does not prove a statement; the user will need to decide at a certain point if a certain hypothesis is too unlikely to be true on the basis of the observations he has at its disposal.

By formalizing the question it seemed to be possible to establish a general approximation for such problems. To this end, one can make the difference

105
CHAPTER 10. HYPOTHESIS TESTING

between the 'null hypothesis' (in the example: each bread weighs 800 grams) and the 'alternative hypothesis' (the bread do not weigh 800 grams). Next, one will verify if the available data admits to replace the null hypothesis by the alternative hypothesis. For this end, the required probabilities need to be defined (significance, exceedance probability).

In the next sections these aspects are introduced step by step. At first we will keep a general reasoning and we will focus on the philosophy behind the method. Therefore the probability distribution of a test statistic is not yet specified in the next section. Next we will consider/mention a number of specific cases for which the probability distribution is indeed specified, and which gives rise to the T-test, chi-square-test, F-test, etc.

10.2 Hypothesis and significance testing: the basic ideas

10.2.1 Formulation of hypotheses

10.2.1.1 Null hypothesis

The first step in hypothesis testing is the formulation of the statement against which we try to find evidence, called the null hypothesis. Usually the null hypothesis is a statement of the form 'no effect' or 'no difference'.

Definition 10.2.1: Null hypothesis

The statement that is tested in a significance test, is called the null hypothesis. The significance test is designed to determine the evidence against the null hypothesis. Notation: \( H_0 \).

Example: \( H_0: \mu_{\text{brood}} = 800 \text{ g} \). Note that hypotheses always refer to some population or model, not to a specific result.
10.2.1.2 The alternative hypothesis

The alternative hypothesis represents the effect that we put in the place of the null hypothesis. There are several possibilities here, and the user needs to decide which one is the most appropriate choice. In the example of the bread, this can be: the bread does not weigh 800 grams; the bread is too light weighted; the bread is too heavy.

It is obvious that a bread that is too heavy is not a problem in most cases. If, however, one wants a cable with a given diameter, then the tolerances need to be respected in both directions (too large, too small). A CD may not be too thin (otherwise it breaks too easily), and not too thick because then the production is too expensive.

Too large or too small (both are considered together) is called a two-sided hypothesis. Or too large, or too small (one is only interested in one of both situations) is called a one-sided hypothesis.

**Definition 10.2.2: Alternative hypothesis**

The alternative hypothesis is a statement of which one wishes to verify if it is correct, instead of the null hypothesis $H_0$. Notation: $H_1$. 
10.2.2 The test statistic

The test is based on a quantity that estimates a parameter that is present in a hypothesis. If $H_0$ is true, we expect that the estimator accepts a value in the neighbourhood of the parameter value, specified in $H_0$. Values of the estimator that are distant from the parameter value, specified in $H_0$, constitute a prove against $H_0$. This quantity is called the test statistic.

Definition 10.2.3: Test statistic

A test statistic measures the consensus between the null hypothesis and the data. It is a stochastic variable of which the probability distribution needs to be known.

Next, we will test the probability of significance of $H_0$ via probability.

10.2.3 Statistical significance and exceedance probability

A significance test measures the strength of the prove against the null hypothesis in terms of probabilities. If the observed outcome, under the assumption that the null hypothesis is true, is unlikely, and on the contrary, has a larger probability if the alternative hypothesis is true, then this outcome is an indication against $H_0$ to the credit of $H_1$. The less probable the outcome, the stronger the indication that $H_0$ is false.

In practice it is of course not easy to say what is acceptable and what is not. E.g. the Supreme Court of the United States said that 'two out of three standard deviations' is their criterion to reject $H_0$ for a normal distribution. However, because not all testing quantities are normally distributed, a more general criterion is needed. To this end we will determine the probability of getting an outcome that is as extreme as or even more extreme than the actual, observed outcome. 'Extreme' means: far from what we would expect if $H_0$ is true. In order to calculate this exceedance probability, one needs to know the sample distribution of the test statistic.

Definition 10.2.4: Exceedance probability
CHAPTER 10. HYPOTHESIS TESTING

The probability, calculated under the assumption that $H_0$ is true, that the test statistic $h$ will adopt a value that is as extreme as or even more extreme than the actual, observed outcome $h_n$ (from $n$ observations), is called the exceedance probability of a test.

E.g. for a two-sided test this becomes:

$$P(|h| > h_n) \mid H_0,$$

(10.1)

and for a one-sided test:

$$P(h > h_n) \mid H_0 \text{ or } P(h < h_n) \mid H_0.$$

The smaller the exceedance probability, the stronger the prove, delivered by the data, against $H_0$.

Thanks to the use of computers and statistical packets, the exceedance probability is usually easy to calculate. In the past, one needed to use tables. For that reason, the use of a somewhat simpler concept to handle became popular: the statistical significance.

**Definition 10.2.5: Statistical significance**

If the exceedance probability is smaller or equal to $\alpha$, we say that the data are statistically significant at level $\alpha$.

Typical values for $\alpha$ are 5% and 1%.
CHAPTER 10  HYPOTHESIS TESTING

Remark: 'statistically significant' does not mean important. The original meaning of the word significant is 'meaning something'. In statistics, the term is only used to indicate that the strength of the prove against the null hypothesis has reached the by $\alpha$ supposed standard. The more data one has at his/her disposal, the smaller the deviation that one can determine as 'significant'. Whether these significant deviations are important is another question. It is e.g. irrelevant if a bread weighs 800.000 grams or 800.001 grams, even if one can measure this in a significant way.

10.2.4 Procedure

Below we will describe the whole procedure step by step. In the next sections we will apply these steps to some specific situations.

1. Formulate the null hypothesis $H_0$ and the alternative hypothesis $H_1$. The test is designed to measure the strength of the evidence against $H_0$. $H_1$ is the statement that we will accept if the evidence allows one to reject $H_0$.

2. Specify the level of significance $\alpha$. This expresses how much prove against $H_0$ we will consider as crucial.

3. Calculate the value of the test statistic $h_n$ on which the test is based. This is a quantity that measures how close the data and $H_0$ are.

4. Find the exceedance probability for the observed data. This is the probability, calculated under the assumption that $H_0$ is true, that the outcome of the test statistic is at least as strong against $H_0$ as with the observed data. If the exceedance probability is smaller or equal to $\alpha$, the test result is significant at level $\alpha$.  

110
CHAPTER 10. HYPOTHESIS TESTING

10.3 Specific testing

Hypothesis testing is used very intensively in a large range of applications. Finally, however, in lots of cases the test turns out to convert to one of the following standard problems:

1. Is the mean value equal to (smaller, bigger than) a predefined value, e.g. $0$?

2. Is the variance equal to (smaller, bigger than) a predefined value?

3. Is the variance of one quantity equal to (smaller, bigger than) the variance of another one?

4. Is the mean value of one quantity equal to (smaller, bigger than) the mean value of another one?

These standard problems are studied in detail in the literature. Thereby one mostly assumes that the underlying stochastic variables are normally distributed so that one can explicitly establish the probability distribution of the test statistic $h_n$. A number of cases are described below.

10.3.1 The $T$-test: $H_0: \mu = \mu_0$

Two-sided test. We want to test the hypothesis that the mean value $\mu$ of a normally distributed quantity $X$ is equal to $\mu_0$, starting from $n$ independent measurements $\{x_1, x_2, \ldots, x_n\}$. As alternative hypothesis we say that $\mu \neq \mu_0$.

This leads to the following formal problem:

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu \neq \mu_0 \tag{10.3}$$

In order to answer this question we consider the following test statistic:

$$T = \frac{\overline{X}_n - \mu_0}{S_n / (\sqrt{n})} \sim t_{n-1}, \quad |H_0|, \tag{10.4}$$
CHAPTER 10. HYPOTHESIS TESTING

with

\[ X_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]  

and

\[ S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - X_n)^2. \]  

(10.5)

Note that \( T \) is precisely the quantity that was studied in Section 9.6 where it was explained that this quantity has a Student’s t-distribution with \( n - 1 \) degrees of freedom (for this reason the test statistic was represented in (10.4) by \( T \) instead of \( h \)). So the distribution of the test statistic is known and the exceedance probability and the significance level can be calculated immediately.

If we choose a significance level \( \alpha \), then:

\[ P \left( |T| \leq t_{n-1,1-\alpha/2} \right) = 1 - \alpha \quad \text{if } H_0. \]  

(10.6)

One needs to put \( \alpha/2 \) into \( t_{n-1,1-\alpha/2} \) because we consider a two-sided test \( (\mu \neq \mu_0) \) and the Student’s t-distribution is symmetrical. The admissible region (the interval in which the test statistic needs be located) is \( [-t_{n-1,1-\alpha/2}, t_{n-1,1-\alpha/2}] \).

Alternatively, one could also have calculated the exceedance probability for the observed value of \( T \) and the given number of measurements \( n \), e.g. via the MATLAB™-routine: \texttt{tinv} (see also Section 9.6).

**One-sided test** Instead of verifying if \( H_1 : \mu \neq \mu_0 \), one can also verify if \( H_1 : \mu > \mu_0 \) (or \( H_1 : \mu < \mu_0 \)). This gives rise to a one-sided test where one verifies if the test statistic is bigger (smaller) than the significance level. Formally this becomes:

\[ H_0 : \mu \leq \mu_0 \quad \text{and} \quad H_1 : \mu > \mu_0 \]  

(10.7)

and the test becomes

\[ P \left( T \leq t_{n-1,1-\alpha} \right) = 1 - \alpha \quad \text{if } H_0. \]  

(10.8)

Note that one can no longer consider the absolute value and that one uses \( \alpha \) in \( t_{n-1,1-\alpha} \). This is because the admissible region for the one-sided test is \( (-\infty, +t_{n-1,1-\alpha}] \).
10.3.2 The $\chi^2$-test: $H_0 : \sigma = \sigma_0$

In this test one verifies whether the spread of a distribution is equal to a certain value. The reasoning is the same as for the $t$-test but one uses the following test statistic:

$$ Y = \frac{(n-1) S_n^2}{\sigma_0^2} \sim \chi^2_{n-1} \mid H_0 \quad (10.9) $$

that has a $\chi^2_{n-1}$ distribution (see Section 9.5).

This results in the following tests:

Two-sided test

$$ \begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases} \quad (10.10) $$

with admissible region

$$ [\chi_{n-1,\alpha/2}, \chi_{n-1,1-\alpha/2}] \quad (10.11) $$

One-sided test

$$ \begin{cases} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases} \quad (10.12) $$

with admissible region

$$ [0, \chi_{n-1,1-\alpha}] \quad (10.13) $$

One-sided test

$$ \begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{cases} \quad (10.14) $$

with admissible region

$$ [\chi_{n-1,\alpha}, \infty) \quad (10.15) $$
10.3.3 Comparing mean values or the two-group t-test

Consider two independent data samples \( \{X_1, X_2, \ldots, X_m\} \) and \( \{Y_1, Y_2, \ldots, Y_n\} \), with \( X_i \sim \mathcal{N}(\mu_1, \sigma_1) \) and \( Y_i \sim \mathcal{N}(\mu_2, \sigma_2) \). Next, we wish to know if \( H_0 : \mu_1 = \mu_2 \)?

This problem can be solved very easily if \( \sigma_1 = \sigma_2 \) (see e.g. the course of Inleiding in de waarschijnlijkheidsrekening en de statistiek van Prof. Caenepeel and De Groen). In the general case, one can establish an approximate test that is valid if \( m \) and \( n \) are large. Use in this case

\[
Z = \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \sim \mathcal{N}(0, 1) \quad |H_0 \quad (10.16)
\]

as test statistic and apply the methods that were explained for this end.

10.3.4 The \( F \)-test, comparing variances

(this section is not a part of the material for the examination)

If one wishes to verify whether the variances of two independent data samples are equal, one can use an \( F \)-test. Here, one considers as test statistic:

\[
F = \frac{S_1^2}{S_2^2} \sim F_{m-1, n-1} \quad |H_0, \text{ with } H_0 : \sigma_1^2 = \sigma_2^2 \quad (10.17)
\]

The \( F \)-distribution is exhaustively documented in standard works. It is also available in statistical packets.
10.4 Errors in hypothesis testing

Significant tests focus on $H_0$, the null hypothesis. However, if a decision is needed, there is no reason to suggest $H_0$. There are simply two hypotheses, and we need to accept one and reject the other. We hope that our decision will be correct, but sometimes it will not. There can occur 2 classes of errors:

1. Errors of the first kind: we reject $H_0$, while $H_0$ is in fact true.

2. Errors of the second kind: we accept $H_0$, while $H_0$ is in fact false.

Globally, we have four possible situations, characterized by conditional probabilities as is given in Table 10.1.

Table 10.1: The two kinds of errors when testing hypotheses.

<table>
<thead>
<tr>
<th>$H_0$ is true</th>
<th>$H_1$ is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>reject $H_0$</td>
<td>error of the first kind</td>
</tr>
<tr>
<td>accept $H_0$</td>
<td>correct decision</td>
</tr>
<tr>
<td></td>
<td>correct decision</td>
</tr>
<tr>
<td></td>
<td>error of the second kind</td>
</tr>
</tbody>
</table>

A significant test with level $\alpha$ has a probability $\alpha$ of an error of the first kind. The probability of an error of the second kind depends on the true value of the tested parameter. The detailed study of these errors is out of the scope of this course.
Chapter 11

Estimation of models starting from measurements
Why do you need identification methods?

A simple experiment

Multiple measurements lead to conflicting results.

How to combine all this information?

Why do you need identification methods
Measurement of a resistance

2 sets of measurements

Group A  Group B
3 different estimators

\[ \hat{R}_{SA}(N) = \frac{1}{N} \sum_{k=1}^{N} u(k) \]

\[ \hat{R}_{LS}(N) = \frac{1}{N} \sum_{k=1}^{N} \frac{u(k)i(k)}{\sum_{k=1}^{N} i(k)^2} \]

\[ \hat{R}_{EV}(N) = \frac{1}{N} \sum_{k=1}^{N} \frac{u(k)}{\sum_{k=1}^{N} i(k)} \]

and their results

Remarks

- variations decrease as function of \( N \), except for \( \hat{R}_{LS} \)
- the asymptotic values are different
- \( \hat{R}_{SA} \) behaves 'strange'
Repeating the experiments.

Group A

Group B

Observed pdf of $\hat{R}(N)$ for both groups, from the left to the right $N = 10, 100, \text{ and } 1000$

- the distributions become more concentrated around their limit value
- $R_{SA}$ behaves 'strange' for group A

Repeating the experiments

Group

Group B

Standard deviation of $\hat{R}(N)$ for the different estimators, and comparison with $1/\sqrt{N}$: full dotted line: $R_{SA}$, dotted line: $R_{LS}$, full line: $R_{EV}$, dashed line $1/\sqrt{N}$.

- the standard deviation decrease in $\sqrt{N}$
- the uncertainty also depends on the estimator
Strange behaviour of $\hat{R}_{SA}$ for group A

Group A

Group B

Histogram of the current measurements.

- The current takes negative values for group A

- The estimators tend to a normal distribution although the noise behaviour is completely different

Simplified analysis

Why do the asymptotic values depend on the estimator?

Can we explain the behaviour of the variance?

Why does the $\hat{R}_{SA}$ estimator behave strange for group A?

More information is needed to answer these questions

- noise model of the measurements

\[ i(k) = i_0 + n_i(k) \quad u(k) = u_0 + n_u(k) \]

- Assumption: $n_i(k)$ and $n_u(k)$ are mutually independent zero mean iid (independent and identically distributed) random variables with a symmetric distribution and with variance $\sigma_i^2$ and $\sigma_u^2$. 
Statistical tools

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} x(k) = 0
\]

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} x(k)^2 = \sigma_x^2
\]

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} x(k)y(k) = 0
\]

Asymptotic value of \( \hat{R}_{LS} \)

\[
\lim_{N \to \infty} \hat{R}_{LS}(N) = \lim_{N \to \infty} \left( \frac{\sum_{k=1}^{N} u(k)\hat{u}(k)}{\sum_{k=1}^{N} \hat{u}^2(k)} \right)
\]

\[
= \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} (u_0 + n_u(k))(i_0 + n_i(k))
\]

\[
= \frac{1}{N} \sum_{k=1}^{N} (i_0 + n_i(k))^2
\]

Or

\[
\lim_{N \to \infty} \hat{R}_{LS}(N) = \frac{u_0i_0 + \frac{u_0}{N} \sum_{k=1}^{N} n_u(k)(i_0 + n_i(k)) + \frac{i_0}{N} \sum_{k=1}^{N} n_i(k) + \frac{1}{N} \sum_{k=1}^{N} n_u(k)n_i(k)}{i_0^2 + \frac{1}{N} \sum_{k=1}^{N} n_i^2(k) + \frac{2}{N} \sum_{k=1}^{N} n_i(k)n_i(k)}
\]

And finally

\[
\lim_{N \to \infty} \hat{R}_{LS}(N) = \frac{u_0^2}{i_0^2 + \sigma_i^2} = \frac{R_0}{1 + \sigma_i^2/i_0^2}
\]

It converges to the wrong value!!!
Asymptotic value of $\hat{R}_{EV}$

$$\lim_{N \to \infty} \hat{R}_{EV}(N) = \lim_{N \to \infty} \left( \frac{\sum_{k=1}^{N} u(k)}{\sum_{k=1}^{N} i(k)} \right)$$

$$= \lim_{N \to \infty} \frac{\frac{1}{N} \sum_{k=1}^{N} (u_0 + n_u(k))}{\frac{1}{N} \sum_{k=1}^{N} (i_0 + n_i(k))}$$

$$= \left\{ \begin{array}{ll}
& \frac{u_0 + \frac{1}{N} \sum_{k=1}^{N} n_u(k)}{i_0 + \frac{1}{N} \sum_{k=1}^{N} n_i(k)} \\
& \lim_{N \to \infty}
\end{array} \right.$$

$$= R_0$$

It converges to the exact value!!!

Asymptotic value of $\hat{R}_{SA}$

$$\hat{R}_{SA}(N) = \frac{1}{N} \sum_{k=0}^{N} \frac{u(k)}{i(k)} = \frac{1}{N} \sum_{k=0}^{N} \frac{u_0 + n_u(k)}{i_0 + n_i(k)} = \frac{1}{N} \sum_{k=0}^{N} \frac{1 + n_u(k) / u_0}{1 + n_i(k) / i_0}$$

The series expansion exist only for small noise distortions

$$\frac{1}{1 + x} = \sum_{l=0}^{\infty} (-1)^l x^l$$ for $|x| < 1$

**Group A:** A detailed analysis shows that the expected value does not exist for the data of group A.

The estimator does not converge.

**Group B:** For group B the series converges and

$$\lim_{N \to \infty} \hat{R}_{SA}(N) = R_0 \left( 1 + \frac{\sigma_i^2}{i_0^2} \right)$$

The estimator converges to the wrong value!!
Variance expressions

First order approximation

\[ \sigma_{R_{ls}}^2(N) = \sigma_{R_{ev}}^2(N) = \sigma_{R_{sa}}^2(N) = \frac{R_0^2 \sigma_u^2}{N} + \frac{\sigma_i^2}{u_0^2} \]

- variance decreases in \(1/N\)

- variance increases with the noise

- for low noise levels, all estimators have the same uncertainty

--- Experiment design

Cost function interpretation

The previous estimates match the model \(u = Ri\) as good as possible on the data.

A criterion to express the goodness of the fit is needed ----> Cost function interpretation.

\[ R_{SA}(N) \]

\[ V_{SA}(R) = \frac{1}{N} \sum_{k=1}^{N} (R(k) - R)^2. \]

\[ R_{LS}(N) \]

\[ V_{LS}(R) = \frac{1}{N} \sum_{k=1}^{N} (u(k) - Ri(k))^2 \]

\[ R_{EV}(N) \]

\[ V_{EV}(R, i_0, u_0) = \frac{1}{N} \left( \sum_{k=1}^{N} (u(k) - u_0)^2 + \sum_{k=1}^{N} (i(k) - i_0)^2 \right) \]

subject to \(u_0 = Ri_0\)
Conclusion

- A simple problem

- Many solutions

- How to select a good estimator?

- Can we know the properties in advance?

----> need for a general framework !!

Characterizing estimators

Location properties: are the parameters concentrated around the 'exact value'?

Dispersion properties: is the uncertainty small or large?
Location properties

unbiased and consistent estimators

Unbiased estimates: the mean value equals the exact value

Definition
An estimator $\hat{\theta}$ of the parameters $\theta_0$ is unbiased if $E\{\hat{\theta}\} = \theta_0$, for all true parameters $\theta_0$. Otherwise it is a biased estimator.

Asymptotic unbiased estimates: unbiased for $N \rightarrow \infty$

The sample mean

\[ \hat{u}(N) = \frac{1}{N} \sum_{k=1}^{N} u(k) \]

Unbiased?

\[ E\{\hat{u}(N)\} = \frac{1}{N} \sum_{k=1}^{N} E\{u(k)\} = \frac{1}{N} \sum_{k=1}^{N} \theta_0 = \theta_0 \]

The sample variance

\[ \hat{\sigma}_u(N) = \frac{1}{N} \sum_{k=1}^{N} (u(k) - \hat{u}(N))^2 \]

Unbiased?

\[ E\{\hat{\sigma}_u(N)\} = \frac{N-1}{N} \sigma_u^2 \]

Alternative expression

\[ \frac{1}{N-1} \sum_{k=1}^{N} (u(k) - \hat{u}(N))^2 \]
Consistent estimates: the probability mass gets concentrated around the exact value

$$\lim_{N \to \infty} \text{Prob}(|\hat{\theta}(N) - \theta_0| > \delta > 0) = 0$$

Example

$$\text{plim}_{N \to \infty} \hat{R}_{EV}(N) = \frac{\text{plim} \left( \frac{1}{N} \sum_{k=1}^{N} u(k) \right)}{\text{plim} \left( \frac{1}{N} \sum_{k=1}^{N} i(k) \right)}$$

$$= \frac{\mu_0}{i_0} = R_0$$
Dispersion properties

efficient estimators

- Mostly the covariance matrix is used, however alternatives like percentiles exist.

- For a given data set, there exists a minimum bound on the covariance matrix: the Cramér-Rao lower bound.

\[ CR(\theta) = Fi^{-1}(\theta_0) \]

with

\[ Fi(\theta_0) = E \left( \frac{\partial}{\partial \theta}(Z|\theta)^T \frac{\partial}{\partial \theta}(Z|\theta) \right) = -E \left\{ \frac{\partial^2}{\partial \theta^2} l(Z|\theta) \right\}. \]

The derivatives are calculated in \( \theta = \theta_0 \)

The likelihood function (ter informatie, geen examenstof)

1) Consider the measurements \( Z \in \mathbb{R}^N \)
2) \( Z \) is generated by a hypothetical, exact model with parameters \( \theta_0 \)
3) \( Z \) is disturbed by noise \( \rightarrow \) stochastic variables
4) Consider the probability density function \( f(Z|\theta_0) \) with

\[ \int_{Z \in \mathbb{R}} f(Z|\theta_0) dZ = 1. \]

5) Interpret this relation conversely, viz:

how likely is it that a specific set of measurements \( Z = Z_m \) are generated by a system with parameters \( \theta \)?

In other words, we consider now a given set of measurements and view the model parameters as the free variables:

\[ L(Z_m|\theta) = f(Z = Z_m|\theta), \]

with \( \theta \) the free variables.

\( L(Z_m|\theta) \) is called the likelihood function.
Basic steps in identification

1) collect the information: experiment setup

2) select a model
   parametric > nonparametric models
   white < black box models
   linear < nonlinear models
   linear-in-the-parameters < nonlinear-in-the-parameters

   \[ e = y - (a_1 u + a_2 u^2), \quad e(\omega) = Y(\omega) - \frac{\alpha_0 + \alpha_1 j\omega}{\beta_0 + \beta_1 j\omega} U(\omega) \]

3) match the model to the data
   select a cost function
   --> LS, WLS, MLE, Bayes estimation

4) validation
   does the model explain the data?
   can it deal with new data?

Remark: this scheme is not only valid for the classical identification theory. It also applies to neural nets, fuzzy logic, ...

A statistical framework: choice of the cost functions

\[ y_0 = G(u, \theta_0), \quad y = y_0 + n_y, \quad e = y - G(u, \theta_0) \]

Least squares estimation

\[ V_{\text{LS}}(\theta) = \frac{1}{N} \sum_{k=1}^{N} e^2(k, \theta) \]

Weighted least squares estimation

\[ V_{\text{WLS}}(\theta) = \frac{1}{N} e(\theta)^T W e(\theta) \]

Maximum likelihood estimation

\[ f(y | \theta_0) = f_n(y - G(u, \theta_0)) \]

\[ \theta_{\text{ML}} = \arg \max_{\theta} f(y | \theta) \]
Bibliography

In the library there are several good basic books available about the subjects that were treated in this course. I have explicitly used the following books in order to establish these course notes.

  - This work is available at the vubtiek but is only available in Dutch. The notation in this course is completely similar to the one used in this more detailed course. Students who want to study some aspects more in-depth, can find more information in this course.
  - http://homepages.vub.ac.be/~scaenepe/homepage.html#kans

  - This work gives more insight in the use of statistics, without using a forced mathematical approach. Unfortunately, it is only available in Dutch.


