Probability and Statistics I – 2019/20

These notes are a summary of what is lectured. They do not contain all examples or explanations and are NOT a substitute for your own notes. The notes are based on a previous version of the module by Hugo Maruri-Aguilar and on the textbook mentioned below.

Course web page: Full details of the module (lecture notes, model solutions, contact details, office hours, video recordings, etc.) are available on the course web page:

https://qmplus.qmul.ac.uk/course/view.php?id=10517

If you have an urgent query contact me by email: w.just@qmul.ac.uk

Final exam: The final exam will count with 80% towards your module mark.

- **The final exam will be OPEN SOURCE. All questions count.**
- **PERMITTED:**
  - **Any printed, handwritten, photocopied etc. material.**
  - **The use of electronic devices such as calculators, laptops, mobile phones etc. to access the Internet (Google, Wikipedia,...) and to run software (such as R).**
- **PROHIBITED:**
  - **To use electronic devices for the purpose of communication (e.g. Email, Twitter, Instagram, Facebook, mobile phone calls, etc.).**
  - **To share material with other students.**

Marked exam script (feedback): You can obtain a copy of your MARKED FINAL EXAM SCRIPT, at about 3 weeks after the final exam. If you wish to obtain a copy of your marked final exam script send me a request by email, by April 10th 2020 at the very latest: w.just@qmul.ac.uk.

Use your COLLEGE EMAIL ADDRESS and put your STUDENT ID in the subject field.

Mid term test: We will run a multiple choice mid term test in week 7 which counts with 20% towards your module mark.

Tutorials: The lecture notes contain the problem sheets as well, see the appendix. Most (but not all) of the problems require the use of software, hence we run the tutorials in computer labs.

Solutions to courseworks will be made available on the course web page at the end of the relevant week.

There is no need to submit any coursework. If you want to receive personalised feedback on your work speak to me.

It is vital that you attend the computer labs and that you do the coursework questions, as otherwise you won't acquire the necessary skills to pass the module.

Tutorials will start in week 2.

There are no tutorials or lectures in week 7 (reading week).

Literature: Parts of the lecture notes have been based on the textbook:

Michael Sullivan
Statistics: Informed Decisions Using Data
Pearson

Contents

§0 Introduction

By and large Statistics is data analysis using predominantly tools and concepts from Probability.

a) Illustrative examples

The ideas behind and the challenges in Statistics can be suitably illustrated by a couple of examples.

Example 0.1 (The german tank problem) *The tanks produced by Germany in WWII had equipment with a consecutive serial number, i.e., tanks were essentially labelled by integers* $1, 2, \dots, N$*. The Allies knew about this practise and they were interested to get a good estimate about the number* N *of tanks produced. They had serial numbers from captured or destroyed tanks, i.e., they had a random sample of size* $n < N$ *of serial numbers (say for instance* $124, 3, 65, 122, 48$*). Given such an information what would be a suitable estimate for* N*. In the example above* N *is certainly larger than* 124*, and it is unlikely that* N *is very large, say* $20,000$ *(as it would have been unlikely to capture only tanks with very low serial numbers). What would be the best estimate for* N*, or an expected value, or even a complete probability distribution? The Allied solved this highly nontrivial problem surprisingly well.*

Example 0.2 (Simpon's paradox) *Two treatments for a disease, say drug A and drug B, are tested on a group of 390 patients. Drug A is given to 160 patients of whom 100 are men and 60 are women. 20 of these men and 40 of these women recover. Drug B is given to 230 patients of whom 210 are men and 20 are women. 50 of these men and 15 of these women recover. With drug A* $20/100 = 20\%$ *of the men recover while with drug B* $50/210 = 23.8\ldots\%$ *of the men recover. Hence drug B is more effective for men. As for women with drug A* $40/60 = 66.6\ldots\%$ *of women recover while with drug B* $15/20 = 75\%$ *of women recover. Hence drug B is more effective for women. However if we ignore the gender then with drug A* $60/160 = 37.5\%$ *of patients recover while with drug B only*

$65/230 = 28.2\ldots\%$ of the patients recover. Hence drug A seems to be more effective. This paradox and the inconclusive results is a consequence of the unfortunate design of the study.

Example 0.3 (Polling) *On June 27th 2019 ukpollingreport.co.uk reports some latest results on UK voting intention:*

"I am a little cautious of the value of voting intention polls at this point, we can expect the appointment of a new Prime Minister to have a significant impact on political support, so voting intention polls right now seem a trifle redundant. However, for what they are worth there have been two new polls this week so far.

YouGov for the Times had topline figures of CON 22%(+2), LAB 20%(nc), LDEM 19%(-2), BREX 22%(-1), GRN 10%(+1). Fieldwork was Monday to Tuesday, and changes are from mid-June.

Ipsos MORI's monthly political monitor in the Standard had topline figures of CON 26%(+1), LAB 24%(-3). LDEM 22%(+7), BREX 12%(-4), GRN 8%(-1). Fieldwork was over the weekend, and changes are from last month."

How accurate are these figures, in fact how is polling done? Polling happens in many areas of social life, for instance, pricing of adverts in TV programmes, introduction of a new product, etc.. Accurate polls have an enormous economical and social value. But how can we estimate or improve their accuracy?

b) Basic notions

Like any other subject, Statistics uses a common language to formulate problems and solutions. However, these notions are often not as rigorously defined as, say, a pure mathematician would prefer.

The entire group of objects to be studied is called the *population.* A population may be finite or infinite, a population may be real or hypothetical. A *sample* is a subset of the population that is being studied.

Different types of collecting data: There are different methods to obtain data for a statistical analysis:

- A *survey* is the collection of data from a sample of the population. If data are obtained from the whole population the study is called a *census*.
- In an *observational study* researchers observe the behaviour of individuals without trying to influence the outcome of the study.
- In a *designed experiment* researchers apply some treatment to the units under investigation and measure the response.

Example 0.4 *400 people are asked whether they prefer Coke or Pepsi. Such a study is a survey.*

The UK BioBank has recruited 500,000 volunteers and follows the medical records until they die, recording which diseases they get. This is an observational study.

Sixty patients with carpal tunnel syndrome are divided into two groups. One group is treated weekly with acupuncture and an exercise regimen. The other group is treated weekly with the same exercise regimen only. After one year, both groups are questioned about their level of pain. This is a designed experiment.

As already mentioned these concepts have to be taken with a grain of salt, as we do not speak about precisely defined objects (like in real life).

Different types of data: Studies may express their measurements in terms of (real) numbers (say the ambient temperature) or categories (say: satisfied/dissatisfied). The former type is called quantitative data, the latter type qualitative data. While some textbooks distinguish between variables and data (the variable is the abstract concept similar to a function f, the data is the outcome similar to the value $f(x)$ in the range) I won't make this distinction here.

- *(Quantitative) Continuous* variables/data are variables which are given in terms of real numbers, such as the blood pressure.

- *(Quantitative) Discrete* variables/data are variables which are given by integers (or any other countable number set) such as attendance numbers in lectures.
- *(Qualitative) Categorical* variables/data (sometimes also called *factors*) are variables which are expressed in terms of categories (such as, e.g., eye colour) where the categories have no sense of order.
- *(Qualitative) Ordinal* variables/data are variables which are expressed in terms of categories (for instance see the QMUL module evaluation: Definitely agree, Mostly agree, Neutral, Mostly disagree, Definitely disagree) where the categories have an obvious order.

Again this notion suggest a level of rigour which is often not met in applications. Consider e.g. the delay times of busses. If the delay time is given, say, in seconds then the data is discrete, but if you allow to express delay times at an arbitrary precision (say 5 minutes 23 seconds 438 milliseconds 128 microseconds 745 nanoseconds ...) then the data is continuous. Also the notion of categorical and ordinal variables may be slightly blurred. Suppose you record the colour of cars (blue, red, black, yellow, ...). You may say that there is no obvious order (i.e. the data is categorical). However it may be possible to impose a natural order by arranging colours along the visible spectrum.

Example 0.5 *The 2018/19 module evaluation questionnaire in MTH4107 showed the following result for the question "My mathematical background has prepared me well for this module": Definitely agree (12), Mostly agree (37), Neutral (23), Mostly disagree (6), Definitely disagree (1). The variable/data is ordinal (note that the type of variable/data is determined by the type of answers to the question, not by the frequencies/student numbers).*

c) Software

We will use the software R to perform the required calculations in this module. R is an open source statistics software package which is freely available (see e.g. https://www.r-project.org/ if you want to install the software on your computer). R is a command line software. The graphics interface Rstudio facilitates the use of R. R and Rstudio are installed on the QMUL teaching network.

General help how to use R is available, e.g., at https://www.statmethods.net/r-tutorial/index.html or by using Google. More details can be found, e.g., in https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

You can use R as well on your mobile via web pages without the need of installing the software, for instance

- https://rdrr.io/snippets/
- https://rextester.com/l/r_online_compiler
- https://www.tutorialspoint.com/execute_r_online.php
- http://makemeanalyst.com/run-your-r-code/

Check some of those pages with your mobile as you may find it convenient to use R in exams (and mobiles will be permitted in exams for using web pages).

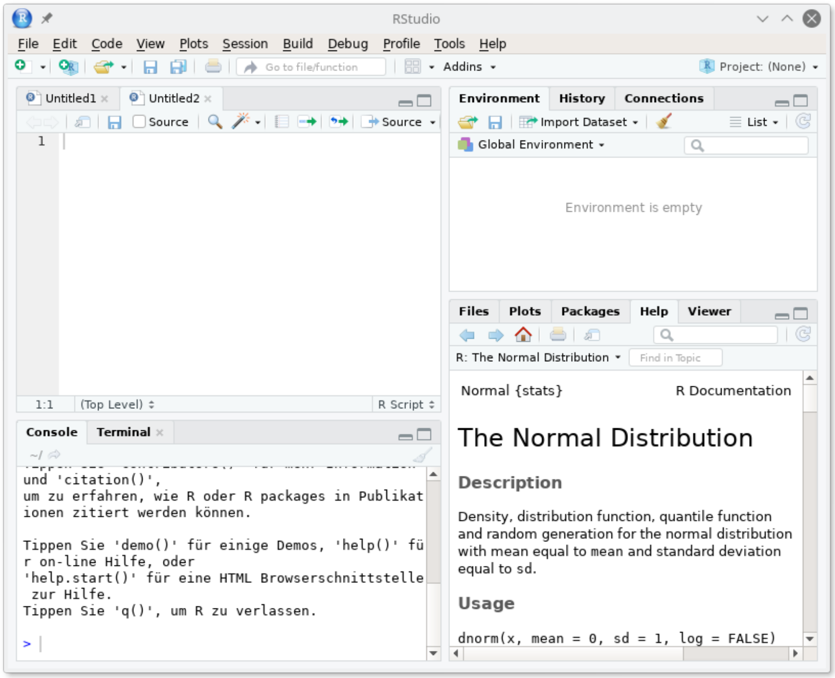
If you start Rstudio a window will open which consists of three main panels (see figure 0.1)

R can do elementary computations

```
> 2+3*6; 1/7; 1/2-1/3
[1] 20
[1] 0.1428571
[1] 0.1666667
```

One may enter multiple commands separated by ;

R knows plenty of analytic functions and constants



```
und 'citation()',
um zu erfahren, wie R oder R packages in Publikat
ionen zitiert werden können.

Tippen Sie 'demo()' für einige Demos, 'help()' fü
r on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle
 zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

>
```

Normal {stats} R Documentation

The Normal Distribution

Description

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
```

Figure 0.1: Main window in Rstudio. The bottom left panel is the workspace for entering commands at the prompt >.

```
> sqrt(2); sin(2); exp(-1.5); cos(pi)
[1] 1.414214
[1] 0.9092974
[1] 0.2231302
[1] -1
```

Values can be “stored” in variables by assigning numbers to symbols using <-

```
> a<-2.5; b<- -2.2
```

The variables then appear in your environment panel (top right) and can be used for computations

```
> a+b
[1] 0.3
```

It is often useful to have numbers in a list/vector, e.g., -1,0,1,2,3. There are (at least) two ways to generate such a list. You can simply use `-1:3` to produce a list of integers

```
> -1:3
[1] -1  0  1  2  3
```

or you can use the sequence command `seq(-1,3,1)` which is more flexible. The third argument is the increment between successive entries

```
> seq(-1,3,1)
[1] -1  0  1  2  3
```

With the sequence command you can generate lists where the entries are non integer and the increments are any real number, e.g.

```
> seq(0.2,2.2,0.5)
[1] 0.2 0.7 1.2 1.7 2.2
```

You can also type in a list by hand, e.g. 20,0.5,12,-2.2, by using the `c` command to obtain the correct data format

```
> c(20,0.5,12,-2.2)
[1] 20.0  0.5 12.0 -2.2
```

You can assign lists to variables as usual, e.g., assign the list $\pi/2, \pi, 3\pi/2, 2\pi 5\pi/2$ to the variable x

```
> x<-seq(pi/2,5*pi/2,pi/2)
> x
[1] 1.570796 3.141593 4.712389 6.283185 7.853982
```

If one types in the variable name the content of the variable is displayed.

If one applies a function to a list, say sin, then the function is applied to each entry. For instance

```
> sin(x)
[1]  1.000000e+00  1.224647e-16 -1.000000e+00 -2.449294e-16
[5]  1.000000e+00
```

Note that R does computations in double precision (up to 16 digits) hence there occur rounding errors of the size of 10^{-16}. The second and the fourth entry of the previous list are actually zero.

You will have noticed that R display funny labels `[1]` and `[5]`. The labels label the elements of the output, i.e. the first element of the list has label `[1]`, the second label `[2]`, and the last label `[5]`. Labels are only displayed if a new output line starts. We can access elements in a list by using these labels, e.g. `x[3]` stands for the third element of the list `x`

```
> x[3]
[1] 4.712389
```

(Note the `[1]` in the output is the label for the single output). Since `sin(x)` is here also a list (see the output above) we can similarly access its elements, e.g.

```
> sin(x)[4]
[1] -2.449294e-16
```

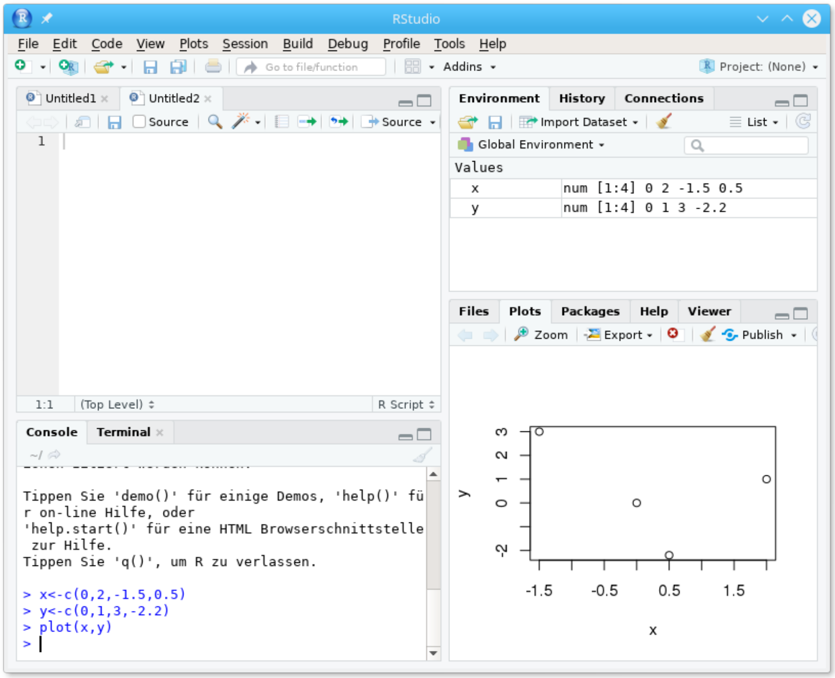
(Note carefully the sequence of the brackets). For the computer geeks: `sin(x[4])` will give the same result, but it is processed differently. While `sin(x)[4]` just looks up the fourth element in the list `sin(x)` (which is computationally fast) the command `sin(x[4])` looks up the fourth element in the list `x` and then computes the sin of this number (which is computationally time consuming). But we are not running a programming module here, so do not worry about efficiency. One can as well replace elements in a list by different numbers using assignments. If we want to change the third element in the list x to zero then just assign 0 to `x[3]`

```
> x
[1] 1.570796 3.141593 4.712389 6.283185 7.853982
> x[3]<-0
> x
[1] 1.570796 3.141593 0.000000 6.283185 7.853982
>
```

R offers plenty of high level commands to perform rather diverse tasks and we will use and study them during the course of this module. One of those is the command for plotting data (and there are different ways doing that). In its simplest form the command `plot` aims at plotting points in the plane. The command expects the x and y coordinates to appear in two lists. Hence for plotting the points $(0,0),(2,1),(-1.5,3),(0.5,-2.2)$ assign the x and y coordinates to two lists and then plot y versus x using the `plot` command

```
> x<-c(0,2,-1.5,0.5)
> y<-c(0,1,3,-2.2)
> plot(x,y)
```

The plot will appear in the lower right panel of Rstudio, see figure 0.2. You can connect the points by line segments using the option `type`.



```
Values
x     num [1:4] 0 2 -1.5 0.5
y     num [1:4] 0 1 3 -2.2
```

```
Tippen Sie 'demo()' für einige Demos, 'help()' fü
r on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle
 zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

> x<-c(0,2,-1.5,0.5)
> y<-c(0,1,3,-2.2)
> plot(x,y)
> 
```

Figure 0.2: Plotting points in the plane. See as well the list of variables in the environment, top right panel.

```
> plot(x,y,type="l")
```

As you see the new plot command will replace the previous plot.

There are different ways to plot functions but none of them is completely satisfactory. Suppose you want to plot the function $x^2\sin(3x)$ in the interval $[-5,5]$. One way doing that is to generate x and y values in lists. First generate a sufficiently dense list of x

values where you want to evaluate the function, and then produce the corresponding list of y values. Finally plot y versus x

```
> x<-seq(-5,5,0.1)
> y<-x^2*sin(3*x)
> plot(x,y,type="l")
```

Now suppose we want to include the graph of the function x^2 in our plot as well. Generate the corresponding y values and assign them to another list (say `z`). If we would use the plot command the previous plot would disappear. To add an element to the existing plot use the command `lines`.

```
> z<-x^2
> lines(x,z,type="l")
```

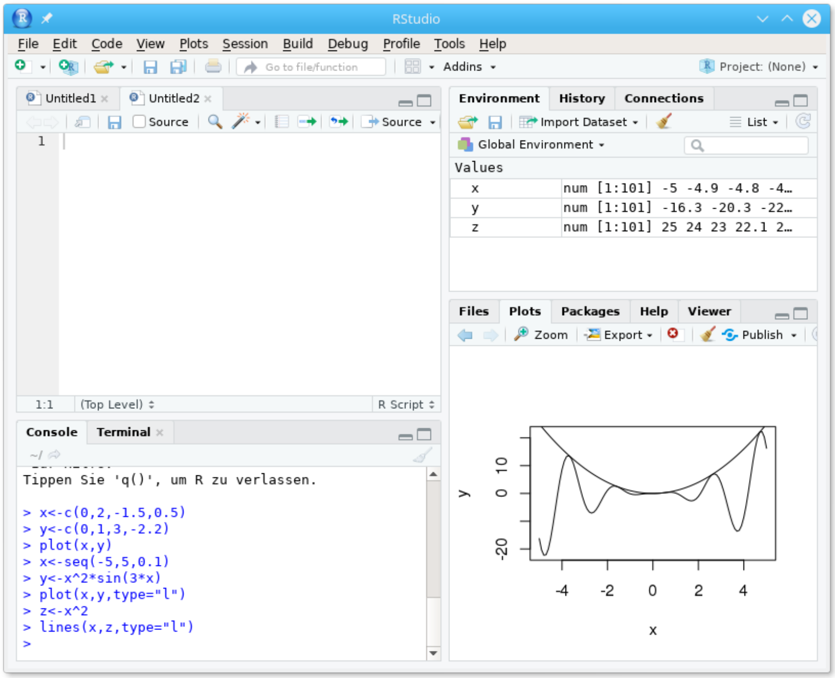
The result is shown in figure 0.3

R offers as well online help on each command. If you want to learn more about, e.g., the `plot` command use `?`, i.e.,

```
> ?plot
```

The lower right panel will then display the online help page which you can study carefully to learn about various options of the command. The online help works similarly for other R commands.

The horizontal menu bars of Rstudio allow you to scroll through the plots you have produced, to save plots in pdf format, and to save your entire session.



```
Values
  x            num [1:101] -5 -4.9 -4.8 -4…
  y            num [1:101] -16.3 -20.3 -22…
  z            num [1:101] 25 24 23 22.1 2…
```

```
Tippen Sie 'q()', um R zu verlassen.

> x<-c(0,2,-1.5,0.5)
> y<-c(0,1,3,-2.2)
> plot(x,y)
> x<-seq(-5,5,0.1)
> y<-x^2*sin(3*x)
> plot(x,y,type="l")
> z<-x^2
> lines(x,z,type="l")
>
```

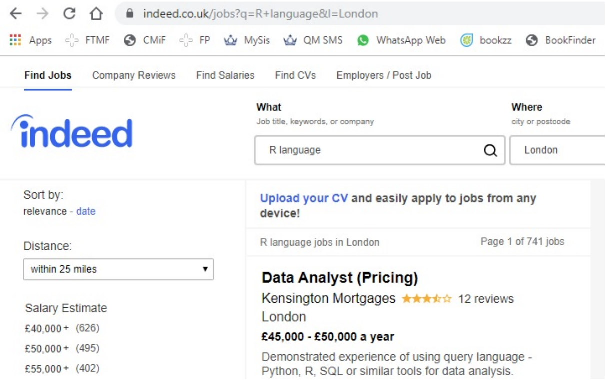
Figure 0.3: Rstudio window when plotting the graphs of $y = x^2 \sin(x)$ and $z = x^2$ as described in the text.

Rstudio has plenty of predefined data sets and they can be listed with `data()`. The commands `View(EuStockMarkets)` or `head(EuStockMarkets)` allow you to display some details of a particular data set (the one which appears as the argument). Data are organised in a table/matrix, i.e., you can access say the third entry in the second column by `EuStockMarkets[3,2]`. The third column is obtained by leaving the row index blank, i.e. `EuStockMarkets[,3]`. For instance `plot(EuStockMarkets[,4],EuStockMarkets[,1])` plots the FTSE vs. the DAX.

d) R and the world

Before we conclude the introduction let me briefly comment on a general non mathematical issue. A mathematics degree will provide you with all the skills to continue your career either in industry or in academia, and this module aims to contribute to this success. For real world problems (i.e. everything beyond problem sheets) Statistics (and any other mathematical subject!) requires computations to be done by electronic means. There are various different programming tools and the one we introduce in this module, R, is widely recognised in the commercial and the academic world.

- R knowledge is highly valued in the job market (see e.g. figure 0.4). If you like R make sure you invest time learning more.



What
Job title, keywords, or company
R language

Where
city or postcode
London

Sort by:
relevance - date

Distance:
within 25 miles

Salary Estimate
£40,000+ (626)
£50,000+ (495)
£55,000+ (402)

Upload your CV and easily apply to jobs from any device!

R language jobs in London Page 1 of 741 jobs

Data Analyst (Pricing)
Kensington Mortgages 12 reviews
London
£45,000 - £50,000 a year
Demonstrated experience of using query language - Python, R, SQL or similar tools for data analysis.

Figure 0.4: Screenshot of a job portal, https://www.indeed.co.uk/Job-Portal-jobs, indicating the salary ranges for applicants with background in Statistics and Programming.

- There are many online resources where you can learn more R. You just need to check Google or Youtube. The one flagged in figure 0.5 teaches you how to install R and R studio as well as the basics of R. Not all tutorials you will find are equally suited so make sure you check a few until you find one that is right for you.

Figure 0.5: Screenshot of an online R tutorial, https://www.youtube.com/watch?v=UYclmg1_KLk.

- Statistics is the essential ingredient of techniques that are currently revolutionising industry: Artificial Intelligence, Machine Learning, Data Science, Data Analytics,. . . . The more you invest in this module the better placed you will be to participate in these exciting new developments.
- Make sure that you check the careers pages of any companies mentioned by the school/college (see figure 0.6 for examples) and similar ones as they might offer exciting internship opportunities that could boost your employability.

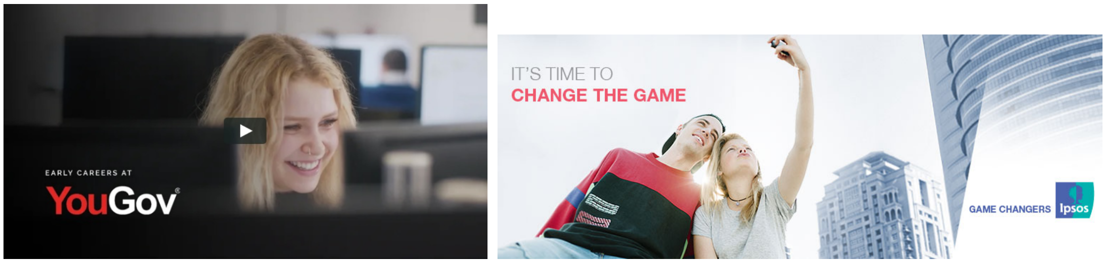


Figure 0.6: Left: Screenshot from YouGov's early careers page https://jobs.yougov.com/early-careers . Right: Screenshot from Ipsos Mori's career page, https://www.ipsos.com/ipsos-mori/en-uk/graduates .

But you do not have to write your job applications this afternoon. I hope you will first enjoy (at least parts of) the current module.

Progress Check:

1. What kinds of methods do you know to collect data?
2. Explain the difference between qualitative and quantitative data.
3. Give examples for discrete data, continuous data, ordinal data, and categorical data.
4. Make yourself familiar with using R on your mobile, your tablet, your laptop, or any other electronic device.
5. Do you know how to implement (assign) a list in R?
6. Are you able to plot a simple function like $\ln(x)$ in R?
7. Do you know how to get information about the datasets which are implemented in R?

§1 Discrete Random Variables

Random variables are functions defined on the sample space $\mathcal{S}$. Discrete random variables take discrete values, here we assume they take integer values (the most frequent case).

Definition 1.1 (Discrete random variable) *A* discrete random variable *is a function from $\mathcal{S}$ to $\mathbb{Z}$.*

We just recall a few important fundamental concepts.

a) Basic quantities

A random variable is completely determined by its probability mass function (pmf).

Definition 1.2 (Probability mass function) *The* probability mass function *(pmf) of a random variable X is the function which given k has output $\mathbb{P}(X = k)$*

$$k \mapsto \mathbb{P}(X = k) \, .$$

Expectation and variance tell us something about the typical values and the uncertainty related with the random variable.

Definition 1.3 (Expectation, variance, and standard deviation) *If X is a discrete random variable which takes values k then the* expectation *of X (or the expected value of X), called $E(X)$, is defined by*

$$E(X) = \sum_k k\mathbb{P}(X = k) \, .$$

The variance *of X, called $Var(X)$, is defined by*

$$Var(X) = E(X^2) - (E(X))^2 = E((X - E(X))^2)$$

while the standard deviation *σ_X is defined by*

$$\sigma_X = \sqrt{Var(X)} \, .$$

Example 1.1 *We toss a fair coin twice and count the number of heads with the random variable X. The sample space reads $\{tt, th, ht, hh\}$ and X takes values $0, 1, 2$. Obviously the pmf reads*

k	0	1	2
$P(X = k)$	$1/4$	$1/2$	$1/4$

Expectation, variance, and standard deviation are given by

$$\begin{aligned} E(X) &= 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1 \\ Var(X) &= (0-1)^2 \times \frac{1}{4} + (1-1)^2 \times \frac{1}{2} + (2-1)^2 \times \frac{1}{4} = \frac{1}{2} \\ \sigma_X &= \frac{1}{\sqrt{2}}\,. \end{aligned}$$

b) Special discrete random variables

Bernoulli(p) distribution: Consider a random variable X which takes values 0 and 1 with pmf

k	0	1
$P(X = k)$	$1-p$	p

which we call the Bernoulli(p) distribution. We write $X \sim$ Bernoulli(p). Examples of the distribution are shown in figure 1.1.

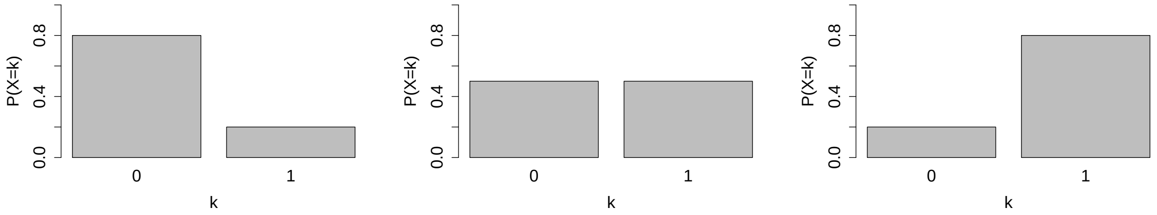


Figure 1.1: Bernoulli(p) distribution as a bar plot for $p = 0.2$ (left), $p = 0.5$ (middle) and $p = 0.8$ (right).

Expectation, variance, and standard deviation:

$$\mathrm{E}(X) = p, \quad \mathrm{Var}(X) = p(1-p), \quad \sigma_X = \sqrt{p(1-p)}\,.$$

Example 1.2 *Toss a biased (or fair) coin once and let X count the number of heads seen. Then $X \sim$ Bernoulli(p) where p measures the bias of the coin (actually: $p - 1/2$ quantifies the bias).*

Binomial distribution: Denote by X_ℓ with $\ell = 1, \ldots, n$ n independent Bernoulli(p) random variables and consider the sum $X = X_1 + \ldots + X_n$. The pmf of X is given by the Binomial(n, p) distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}\,.$$

We say $X \sim \mathrm{Bin}(n, p)$.

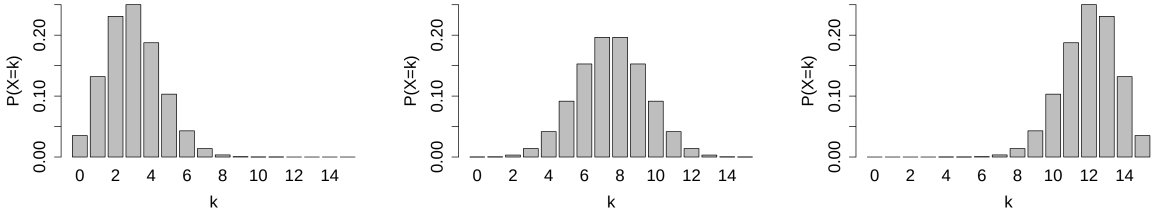


Figure 1.2: Binomial(n, p) distribution as a bar plot for $n = 15$ and $p = 0.2$ (left), $p = 0.5$ (middle) and $p = 0.8$ (right).

Expectation, variance, and standard deviation:

$$\mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p), \quad \sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{np(1-p)}\,.$$

Example 1.3 *If a (biased) coin with probability p coming up "head" is tossed n times, and if X denotes the random variable counting the numbers of heads then $X \sim$ Bin(n, p).*

Geometric distribution: Suppose we make an unlimited number of independent Bernoulli(p) trials. Let T be the number of trials up to and including the first success. The pmf of T is given by

$$P(T = k) = (1-p)^{k-1} p\,.$$

We say T has the *geometric distribution*, $T \sim \mathrm{Geom}(p)$.

Expectation, variance, and standard deviation:

$$\mathrm{E}(T) = \frac{1}{p}, \quad \mathrm{Var}(T) = \frac{1-p}{p^2}, \quad \sigma_T = \frac{\sqrt{1-p}}{p}\,.$$

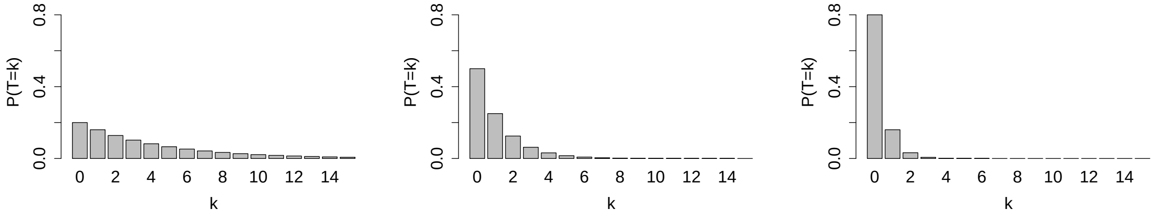


Figure 1.3: Geometric(p) distribution as a bar plot for $p = 0.2$ (left), $p = 0.5$ (middle) and $p = 0.8$ (right).

Example 1.4 *A biased coin has probability* p *to show head. We toss the biased coin until we see head for the first time. Let* T *denote the random variable counting the number of coin tosses.* $T \sim$ *Geom*(p) *and* $\mathbb{P}(T = k)$ *is the probability that we perform exactly* k *coin tosses.*

Hypergeometric distribution: A bag contains n balls, m white balls and $n-m$ black balls. You pick ℓ balls at random without replacement. Let X denote the number of white balls you pick. Then the pmf of X reads

$$\mathbb{P}(X = k) = \frac{\binom{m}{k}\binom{n-m}{\ell-k}}{\binom{n}{\ell}} .$$

We say X obeys the *Hypergeometric*(n, m, ℓ) distribution $X \sim \mathrm{Hg}(n, m, \ell)$.

Remark: We pick ℓ balls from n balls. With unordered sampling without replacement the size of the sample space is $\binom{n}{\ell}$ (the denominator of the expression). The event $X = k$ contains outcomes where we pick k white balls and $\ell - k$ black balls. The white balls we pick from the subset of m white balls, giving $\binom{m}{k}$ possibilities, whereas the black balls we pick from the subset of $n - m$ black balls, giving $\binom{n-m}{\ell-k}$ possibilities (giving rise to the numerator).

Expectation, variance, and standard deviation:

$$\mathrm{E}(X) = \ell\frac{m}{n}, \quad \mathrm{Var}(X) = \ell\frac{m}{n}\frac{n-m}{n}\frac{n-\ell}{n-1}, \quad \sigma_X = \frac{\sqrt{\ell m(n-m)(n-\ell)}}{n\sqrt{n-1}} .$$

Remark:

- The actual computation of the expectation and the variance can be accomplished

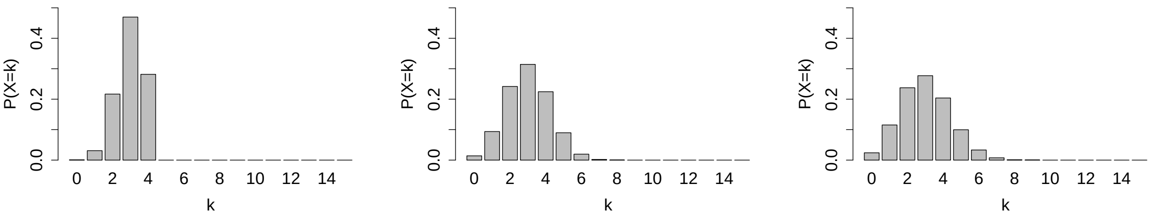


Figure 1.4: Hypergeometric(n, m, ℓ) distribution as a bar plot for $\ell = 15$ and: $n = 20$, $m = 4$ (left), $n = 40$, $m = 8$ (middle) and $n = 80$, $m = 16$ (right). Obviously the pmf is zero if $k > m$.

using factorial moments (see e.g. Introduction to Probability). But the computation is far from trivial.

- If n and m are large the issue of picking with or without replacement becomes less important and the probability to pick a white ball is approximately m/n for each pick. Hence the Hypergeometric distribution $\mathrm{Hg}(n, m, \ell)$ is well approximated by the Binomial distribution $\mathrm{Bin}(m/n, \ell)$ for n and m large (compare figure 1.2, left panel, with figure 1.4, right panel – and problem 2).

Negative binomial distribution: Consider a sequence of independent Bernoulli(p) trials. Given a fixed integer r denote by T the random variable which counts the number of failures when we have seen in total r successes (for the first time). An outcome of the event $T = k$ consists of a sequence of r successes and k failures, $fsff\ldots\ldots sfss$. Each such outcome (each such simple event) has probability $p^r(1-p)^k$. Since the last Bernoulli trial has to be a success and all the other trials can come in any order, there are $\binom{r+k-1}{r-1}$ possibilities to chose the $r-1$ successes among the possible $r+k-1$ instances. Hence the pmf of T is given by

$$\mathbb{P}(T = k) = \binom{k+r-1}{r-1} p^r (1-p)^k \,.$$

We say X obeys the *NegativeBinomial*(r, p) distribution, $X \sim \mathrm{NB}(r, p)$.

Expectation, variance, and standard deviation:

$$\mathrm{E}(T) = \frac{(1-p)r}{p}, \quad \mathrm{Var}(T) = \frac{(1-p)r}{p^2}, \quad \sigma_T = \frac{\sqrt{(1-p)r}}{p}$$

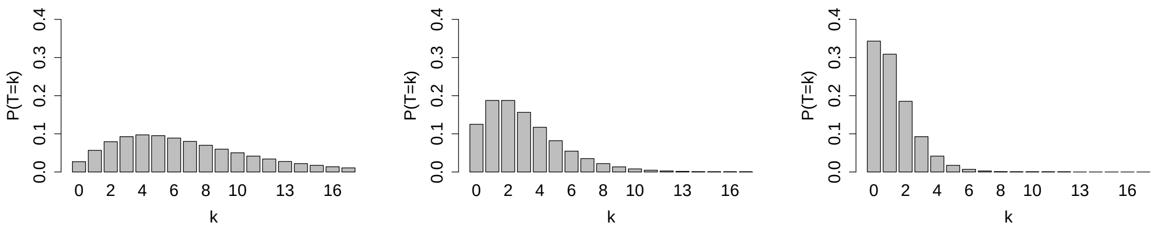


Figure 1.5: NegativeBinomial(r, p) distribution as a bar plot for $r = 3$ and: $p = 0.3$ (left), $p = 0.5$ (middle) and $p = 0.7$ (right).

Remark:

- Some parts of the literature may use a slightly different definition of the negative binomial distribution The most common variations are where the random variable T is counting different things. The definition supplied here is consistent with the implementation in R.
- The computation of the expectation and the variance is possible but slightly tedious.

Example 1.5 *We roll a die repeatedly until we have seen a "6" three times (for the first time). Compute the probability that we roll the die 10 times.*

Denote by $p = 1/6$ the probability that we roll a six. Each roll is a Bernoulli(p) trial. If T denotes the number of failures (the number of times we do not roll a six) then T obeys the NegativeBinomial(r, p) *distribution with $r = 3$ and $p = 1/6$. The total number of rolls is $T + 3$, hence the probability to roll the die 10 times is given by*

$$\mathbb{P}(T = 7) = \binom{7+3-1}{3-1} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^7 .$$

The numerical value can be easily computed in R

```
> dnbinom(7, size=3, prob=1/6)
[1] 0.04651361
```

Hence the probability for exactly 10 rolls is at about 4.7%.

Uniform distribution: Assume you roll a fair die and X denotes the number shown. All the events $X = k$ with $k = 1, \ldots, 6$ have the same probability $P(X = k) = 1/6$. We call such a pmf a uniform distribution. In general: assume X takes values $m, m + 1, \ldots, m + n$ and all such events $X = k$ are equally likely, $P(X = k) = 1/(n + 1)$, $k = m, m + 1, \ldots, m + n$. The corresponding pmf

$$\mathbb{P}(X = k) = \begin{cases} 1/(n+1) & \text{if } m \leq k \leq n + m \\ 0 & \text{else} \end{cases}$$

is called (discrete) uniform distribution, $X \sim \mathrm{U}(m, m + n)$.

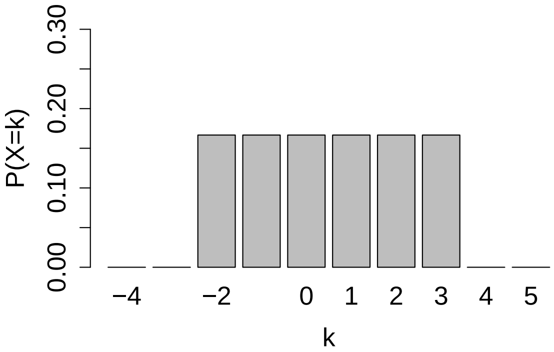


Figure 1.6: Discrete uniform distribution $\mathrm{U}(m, m + n)$ for $m = -2$ and $n = 5$.

Expectation, variance, and standard deviation:

$$\mathrm{E}(X) = m + \frac{n}{2}, \quad \mathrm{Var}(X) = \frac{n(n+2)}{12}, \quad \sigma_X = \frac{\sqrt{n(n+2)}}{2\sqrt{3}}\,.$$

Remark: The computation of expectation and variance is fairly straightforward.

Poisson distribution: Consider a random variable X which takes values $k = 0, 1, 2, 3, \ldots$. Denote by $\lambda > 0$ a fixed positive real number. Define the pmf of the random variable X by

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}\,.$$

We say that X has the Poisson(λ) distribution, in short $X \sim$ Poisson(λ).

Expectation, variance, and standard deviation:

$$\mathrm{E}(X) = \lambda, \quad \mathrm{Var}(X) = \lambda, \quad \sigma_X = \sqrt{\lambda}\,.$$

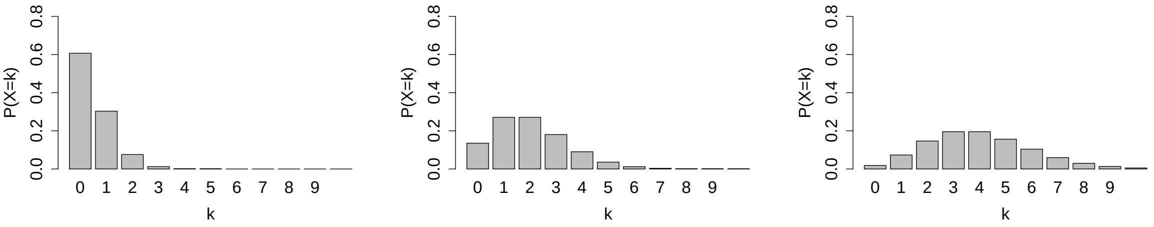


Figure 1.7: Poisson distribution for $\lambda = 0.5$ (left), $\lambda = 2$ (middle) and $\lambda = 4$ (right).

Example 1.6 *The first real application of the Poisson distribution was published by Ladislaus von Bortkiewicz in 1898 ("The Law of Small Numbers") where he recorded the number of Prussian cavalry soldiers kicked to death by their horses per year (in 10 cavalry corps over 20 years). The actual historical data are (frequencies and relative frequencies)*

No. deaths per year	*0*	*1*	*2*	*3*	*4*
frequency	*109*	*65*	*22*	*3*	*1*
relative freq.	*0.545*	*0.325*	*0.110*	*0.015*	*0.005*

It is easy to check that the mean number of deaths is 0.61*. The recorded data agree surprisingly well with the Poisson distribution with* $\lambda = 0.61$*.*

```
> dpois(0:4, 0.61)
[1] 0.543350869 0.331444030 0.101090429 0.020555054 0.003134646
```

The Poisson distribution captures cases where in a large population – here the Prussian army – things happen with small probability – here soldiers being killed by horsekicks (or more formally: a large number of Bernoulli trials where each trial has a small probability).

Summary:

	range	E(X)	Var(X)	R command (pmf)
$X \sim \mathrm{Ber}(p)$	$0, 1$	p	$p(1-p)$	-
$X \sim \mathrm{Bin}(n, p)$	$0, 1, \ldots, n$	np	$np(1-p)$	`dbinom(k,n,p)`
$X \sim \mathrm{Geom}(p)$	$1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	`dgeom(k,p)`
$X \sim \mathrm{Hg}(n, m, \ell)$	$(0, 1, \ldots, \ell)$	$\ell\frac{m}{n}$	$\ell\frac{m}{n}\frac{n-m}{n}\frac{n-\ell}{n-1}$	`dhyper(k,m,n-m,l)`
$X \sim \mathrm{NB}(r, p)$	$0, 1, 2, \ldots$	$\frac{(1-p)r}{p}$	$\frac{(1-p)r}{p^2}$	`dnbinom(k,r,p)`
$X \sim \mathrm{U}(m, m+n)$	$m, \ldots, n+m$	$m + \frac{n}{2}$	$\frac{n(n+2)}{12}$	-
$X \sim \mathrm{Poisson}(\lambda)$	$0, 1, 2, \ldots$	λ	λ	`dpois(k,lambda)`

Remark: Note that pmfs have a bell type shape if a large number of independent incidents is involved, see for instance figures 1.2 (middle), 1.4 (right), or 1.7 (right). We will explore this important issue in section 3

c) Advanced concepts

Cumulative distribution function (cdf): So far we have focused on the probability that a random variable X takes a certain value, $\mathbb{P}(X = k)$, the so called pmf. Now we consider the probability that a random variable does not exceed a certain value, $\mathbb{P}(X \leq k)$.

Example 1.7 *You roll five fair die. Let X denote the largest number seen. To compute the probability that the largest number seen is four, i.e, $\mathbb{P}(X = 4)$, consider the two events $X \leq 4$ and $X \leq 3$. Obviously the event $X = 4$ contains all the outcomes which are contained in $X \leq 4$ but not in $X \leq 3$, hence $\mathbb{P}(X = 4) = \mathbb{P}(X \leq 4) - \mathbb{P}(X \leq 3)$. The expressions of the right hand side are easily evaluated as $\mathbb{P}(X \leq 4) = (4/6)^5$ and $\mathbb{P}(X \leq 3) = (3/6)^5$.*

The example illustrates the usefulness of the so called cumulative distribution function.

Definition 1.4 (Cumulative distribution function) *The* cumulative distribution function *(cdf) of a random variable X is the function which given t has output $\mathbb{P}(X \leq t)$*

$$t \mapsto \mathbb{P}(X \leq t)\,.$$

Remark:

- Unlike the pmf (see definition 1.2) the cdf is a function which is defined for any real value t. For instance, in example 1.7 $P(X \leq 4) = (4/6)^5$ and $\mathbb{P}(X \leq 4.38) = (4/6)^5$ as well. The cdf is a piecewise constant function, see figures 1.8 and 1.9.

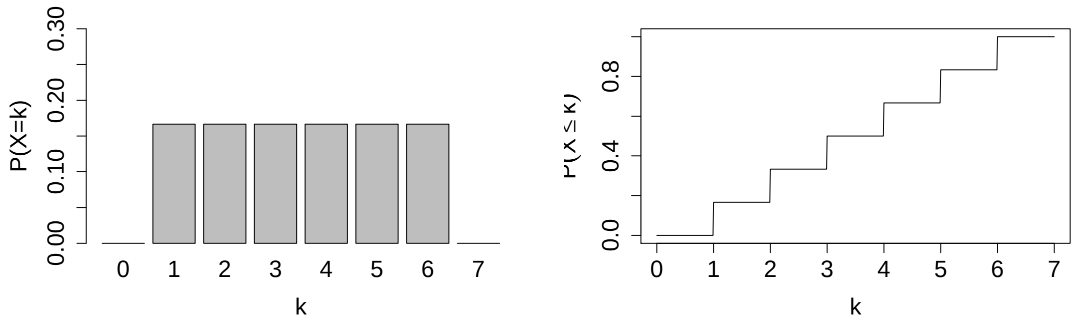


Figure 1.8: Bar plot of the pmf (left) and graph of the cdf (right) of the (discrete) uniform distribution $\mathrm{U}(1,6)$, i.e., pmf and cdf of the distribution for rolling a fair die.

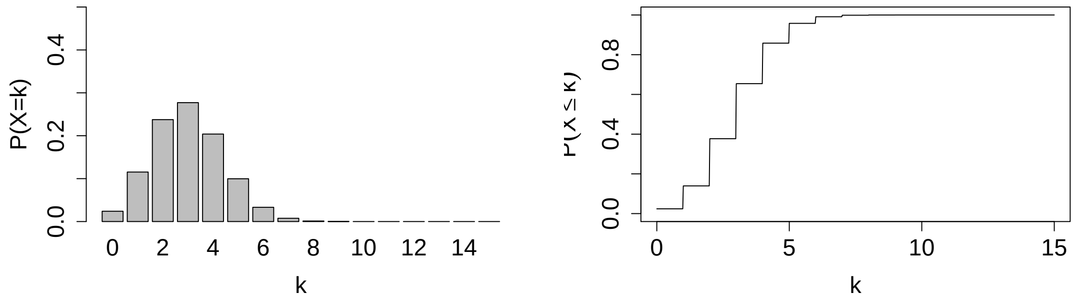


Figure 1.9: Bar plot of the pmf (left) and graph of the cdf (right) of the Hypergeometric distribution with $m = 16$, $n = 80$ and $\ell = 15$, see figure 1.4 (right)

- If X is a random variable which takes values $0, 1, 2, 3, \ldots$ then

$$P(X \leq t) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \ldots + \mathbb{P}(X = \lfloor t \rfloor)$$

  where $\lfloor t \rfloor$ denotes the largest integer which is smaller or equal to t, e.g. $\lfloor 2.8 \rfloor = 2$.

- Obviously the pmf and the cdf are related by

$$P(X = k) = \mathbb{P}(X \leq k) - \mathbb{P}(X \leq k - 1)$$

and more generally for any integers $k \leq \ell$

$$P(k \leq X \leq \ell) = \mathbb{P}(X \leq k) - \mathbb{P}(X \leq \ell - 1)\,.$$

While pmfs are computed in R using the prefix `d` (see above) the cdfs of various random variables are easily computed in R with a prefix `p`.

Example 1.8 *A bag contains 20 red balls and 30 blue balls. You pick 15 balls at random without replacement. Let X denote the number of blue balls you pick. Compute the probability that you pick between (and including) 5 and 10 blue balls, i.e. compute $\mathbb{P}(5 \leq X \leq 10)$.*

Obviously X is a hypergeometric random variable with $n = 50$ (total number of balls), $m = 30$ (number of blue balls) and $\ell = 15$ (number of balls we pick), i.e. $X \sim Hg(50, 30, 15)$. The required probability can be expressed in terms of the cumulative distribution

$$\mathbb{P}(5 \leq X \leq 10) = \mathbb{P}(X \leq 10) - \mathbb{P}(X \leq 4)\,.$$

The expressions can be computed in R using the command `phyper` *for the cdf of the hypergeometric distribution*

```
> phyper(10,30,20,15)-phyper(4,30,20,15)
[1] 0.8248473
```

Moment generating function (mgf): Expectations of powers, such as $E(X^2)$ or $E(X^5)$ (so called moments of the pmf) can be efficiently computed though a moment generating function.

Definition 1.5 (Moment generating function) *Let X be a discrete random variable which takes integer values. The* moment generating function *(mgf) of X is the function which given t has output $E(e^{tX})$*

$$t \mapsto M_X(t) = E(e^{tX})\,.$$

Remark:

- If for instance X takes values $0, 1, 2, 3, \ldots$ the mgf reads

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_k e^{tk}\mathbb{P}(X=k) = \mathbb{P}(X=0) + e^t\mathbb{P}(X=1) + e^{2t}\mathbb{P}(X=2) + \ldots$$

- At $t = 0$ the mgf evaluates as

$$M_X(0) = E(e^{tX})\big|_{t=0} = \sum_k e^0\mathbb{P}(X=k) = 1\,.$$

  Since the first derivative of e^{kt} with respect to t is given by ke^{kt} we obtain for the first derivative at $t = 0$

$$M_X'(0) = \left.\frac{d\mathrm{E}(e^{tX})}{dt}\right|_{t=0} = \sum_k ke^0\mathbb{P}(X=k) = E(X)\,.$$

  In the same way for higher order derivatives

$$M_X^{(n)}(0) = \left.\frac{d^n\mathrm{E}(e^{tX})}{dt^n}\right|_{t=0} = \sum_k k^n e^0\mathbb{P}(X=k) = E(X^n)\,.$$

  Hence the mgf $M_X(t)$ allows to compute moments of any order.

Example 1.9 *Assume the random variable X obeys the Binomial distribution with parameters n and p, $X \sim Bin(n,p)$. Using the so called Binomial theorem the mgf of the Binomial distribution reads*

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{n} e^{tk}\mathbb{P}(X=k) = \sum_{k=0}^{n}\binom{n}{k}\left(e^t\right)^k p^k(1-p)^{n-k} = \left(1-p+pe^t\right)^n\,.$$

Clearly $M_X(0) = 1$ and the derivative

$$M_X'(t) = np(1-p+pe^t)^{n-1}e^t, \quad M_X''(t) = n(n-1)p^2(1-p+pe^t)^{n-2}e^{2t} + np(1-p+pe^t)^{n-1}e^t$$

gives $M_X'(0) = E(X) = np$, $M_X''(0) = E(X^2) = n(n-1)p^2 + np$ with $Var(X) = E(X^2) - (E(X))^2 = np - np^2$ (as required, see section 1b)).

Progress Check:

1. What is a discrete random variable
2. Explain what is meant by a pmf.
3. What is a cdf and why is this concept useful? What is a mgf and why is this concept useful?
4. Give an example of an experiment/random variable which is described by the binomial distribution, by the geometric distribution, by the hypergeometric distribution, and by the negative binomial distribution.
5. Which kind of experiments/set up are described by the Poisson distribution.
6. State the pdf of the Bernoulli(p) distribution and calculate the cdf and the mgf.
7. How do the pmfs of the uniform, of the geometric, and of the binomial distribution look like. In which way do they differ (e.g. in terms of range and shape)?

§2 Continuous Random Variables

So far we have considered random variables which take integer values (e.g. rolling a die and recording the number). Expressions such as $\mathbb{P}(X = 2)$ were perfectly meaningful (defining the pmf).

a) Probability densities

Let us now consider continuous random variables, that means random variables which can take any real value. Examples can be waiting times at bus stops (as the waiting time does not have to be an integer multiple of seconds) or diameters of footballs. In that case an expression like $\mathbb{P}(X = 2)$ is meaningless (or to be precise: it is zero) as the probability that you wait exactly two seconds is zero (it can easily happen that you wait 3 milliseconds longer, or 5 microseconds shorter). Probabilities make only sense if you specify an interval, say $\mathbb{P}(1.99 < X \leq 2.01)$ for the probability to wait between 1.99 seconds and 2.01 seconds. Of course such intervals can be as short or as long as you wish. Since probabilities of disjoint events are additive, e.g. $\mathbb{P}(1.99 < X \leq 2) + \mathbb{P}(2 < X \leq 2.01) = \mathbb{P}(1.99 < X \leq 2.01)$ the expression can be written as an integral containing the so called probability density function f_X

Definition 2.1 (Probability density function) *The* probability density function *(pdf) f_X of a continuous random variable X is defined by*

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t)dt$$

for any values $a \leq b$.

Remark:

- The area under the graph of the pdf f_X over an interval $[a, b]$ represents the probability $\mathbb{P}(a \leq X \leq b)$ of observing the random variable X in that interval (see figure 2.1).

- Since the probability of a continuous random variable X taking a single value is zero, e.g., $\mathbb{P}(X=a)=0$ the following probabilities are equivalent

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a < X \le b) = \mathbb{P}(a \le X < b) = \mathbb{P}(a < X < b)\,.$$

- Positivity of probabilities implies that $f_X(x) \ge 0$. Since X takes a value between $-\infty$ and $+\infty$ the condition $\mathbb{P}(-\infty < X < \infty) = 1$ implies the normalisation condition

$$1 = \int_{-\infty}^{\infty} f_X(t)dt\,.$$

Example 2.1 *Denote by X the waiting time at a bus stop. Assume that the waiting time is uniformly distributed between 3 and 5 minutes. That means the pdf is constant in the interval $[3,5]$ and zero elsewhere*

$$f_X(x) = \begin{cases} c & \textit{if} \quad 3 \le x \le 5 \\ 0 & \textit{else} \end{cases}\,.$$

Normalisation gives the value of the constant c

$$1 = \int_{-\infty}^{\infty} f_X(t)dt = \int_3^5 cdt = 2c \quad \Rightarrow \quad c = \frac{1}{2}\,.$$

The probability that we wait between 4 and 4 1/2 minutes is given by

$$\mathbb{P}(4 < X < 4.5) = \int_4^{4.5} f_X(t)dt = \int_4^{4.5} \frac{1}{2}dt = \frac{1}{4}\,.$$

Expectation and variance are defined in the usual way (see definition 1.3) with sums being replaced by integrals.

Definition 2.2 (Expectation and variance) *If X is a continuous random variable with pdf $f_X(x)$ the* expectation *of X (or the expected value of X), called $E(X)$, is defined by*

$$E(X) = \int_{-\infty}^{\infty} t f_X(t)dt\,.$$

The variance *of X, called $Var(X)$, is defined by*

$$Var(X) = E(X^2) - (E(X))^2 = E((X-E(X))^2) = \int_{-\infty}^{\infty} (t-E(X))^2 f_X(t)dt\,.$$

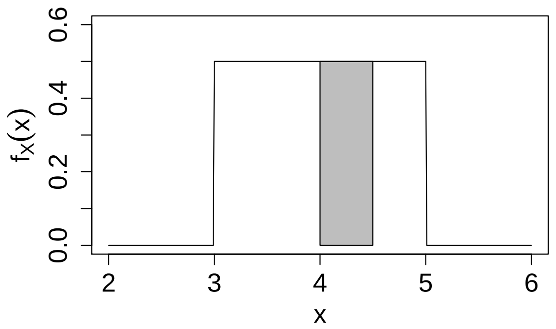


Figure 2.1: Uniform pdf of bus waiting times. The shaded area indicates the probability $\mathbb{P}(4 < X < 4.5)$.

Remark: One often uses the symbol $\mu_X = \mathrm{E}(X)$ (or simply $\mu = \mathrm{E}(X)$) for the expectation and $\sigma_X^2 = \mathrm{Var}(X)$ (or simply $\sigma^2 = \mathrm{Var}(X)$) for the variance of a random variable. Then $\sigma_X = \sqrt{\mathrm{Var}(X)}$ denotes the so called standard deviation.

Example 2.2 *Consider the bus waiting time of example 2.1. The expectation is given by*

$$\mu = E(X) = \int_{-\infty}^{\infty} t f_X(t) dt = \int_3^5 t \frac{1}{2} dt = 4$$

as expected. The variance follows as

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (t-\mu)^2 f_X(t) dt = \int_3^5 (t-4)^2 \frac{1}{2} dt = \int_{-1}^{1} \frac{s^2}{2} ds = \frac{1}{3}\,.$$

For the cumulative distribution function of a continuous random variable we can directly use the definition 1.4, i.e., the cdf is given by

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(t) dt\,.$$

Therefore the derivative of the cdf gives the pdf

$$\frac{dF_X(x)}{dx} = f_X(x)\,.$$

Since the pdf is non-negative, $f_X(x) \ge 0$, the slope of the cdf is non-negative as well, i.e., $F_X(x)$ is a monotonic increasing function.

In addition (see as well the fundamental theorem of Calculus)

$$\mathbb{P}(a \le X \le b) = \int_a^b f_X(t) dt = \int_{-\infty}^{b} f_X(t) dt - \int_{-\infty}^{a} f_X(t) dt = F_X(b) - F_X(a)\,.$$

Example 2.3 *The cdf of the bus waiting times in example 2.1 follows directly by integration*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & \textit{if} \quad x < 3 \\ x/2 & \textit{if} \quad 3 \leq x \leq 5 \\ 1 & \textit{if} \quad x > 5 \end{cases}$$

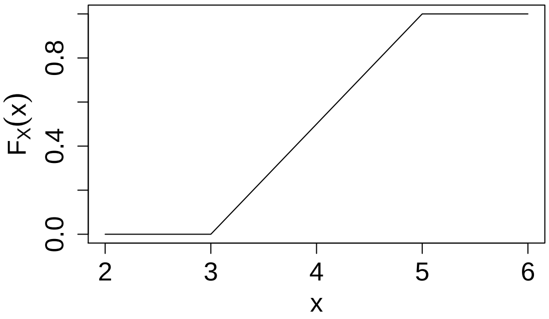


Figure 2.2: Cdf $F_X(x)$ of bus waiting times in example 2.1, see figure 2.1 for the pdf $f_X(x)$. Clearly the derivative of the cdf gives the pdf, $F'_X(x) = f_X(x)$.

b) Special continuous random variables

Uniform distribution: The pdf of a continuous random variable which is uniformly distributed in an interval $[a, b]$ is given by (see example 2.1)

$$f_X(x) = \begin{cases} 1/(b-a) & \text{if} \quad a \leq x \leq b \\ 0 & \text{else} \end{cases} .$$

We say that X is uniformly distributed in the interval $[a, b]$, $X \sim \mathrm{U}(a, b)$.

Expectation and variance follow by straightforward integration (see example 2.2)

$$\mu = \mathrm{E}(X) = \int_a^b \frac{t}{b-a} dt = \frac{a+b}{2}$$

and

$$\sigma^2 = \mathrm{Var}(X) = \int_a^b \frac{t^2}{b-a} dt - \frac{(a+b)^2}{4} = \frac{b^2 + ab + a^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} .$$

The cumulative distribution function follows easily by integration (see example 2.3)

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & \text{if} \quad x < a \\ x/(b-a) & \text{if} \quad a \le x \le b \\ 1 & \text{if} \quad x > b \end{cases} .$$

The pdf and the cdf can be evaluated in R with the commands `dunif(x,min=a,max=b)` and `punif(x,min=a,max=b)` respectively.

Exponential distribution: Consider a continuous random variable X which takes non-negative values with pdf

$$f_X(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \lambda e^{-\lambda x} & \text{if} \quad x \ge 0 \end{cases} .$$

We say X is exponentially distributed with decay rate λ, $X \sim \text{Exp}(\lambda)$.

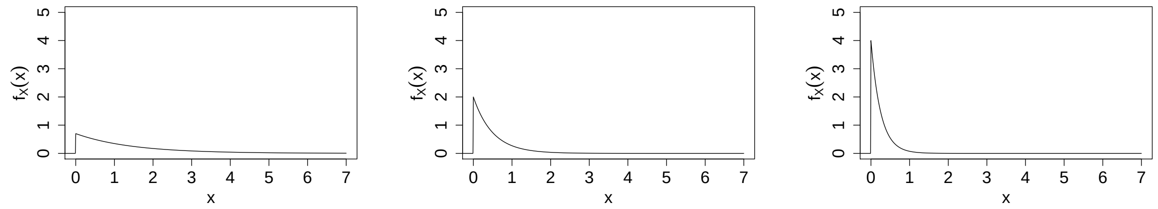


Figure 2.3: Pdf $f_X(x)$ of an exponential random variable, $X \sim \text{Exp}(\lambda)$, with rate $\lambda = 0.7$ (left), $\lambda = 2$ (middle), $\lambda = 4$ (right).

Example 2.4 *The times between two successive radioactive decays is exponentially distributed.*

Expectation and variance follow by straightforward integration

$$\mu = \text{E}(X) = \int_0^{\infty} t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$$

and

$$\sigma^2 = \text{Var}(X) = \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt - \frac{1}{\lambda^2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

In a similar way we derive the cdf. Obviously $F_X(x) = 0$ if $x < 0$ (since the pdf vanishes in this case) and

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, \quad (x \geq 0)\,.$$

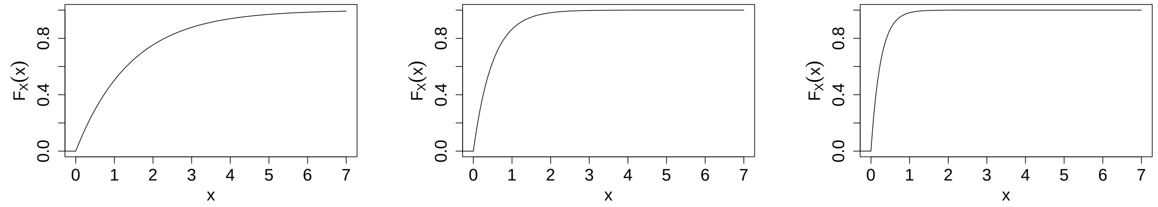


Figure 2.4: Cdf $F_X(x)$ of an exponential random variable, $X \sim \text{Exp}(\lambda)$, with rate $\lambda = 0.7$ (left), $\lambda = 2$ (middle), $\lambda = 4$ (right). See figure 2.3 for the corresponding pdfs $f_X(x)$. The pdf is the derivative of the cdf $F'_X(x) = f_X(x)$.

The pdf $f_X(x)$ of the exponential distribution can be computed in R by `dexp(x,rate=lambda)` while the cdf $F_X(x)$ in R is given by `pexp(x,rate=lambda)`

Example 2.5 *Suppose the waiting time for a bus is exponentially distributed $X \sim Exp(0.4)$ when measured in minutes. The expected waiting time is*

$$E(X) = 1/\lambda = 1/0.4 = 2.5 \quad (minutes)\,.$$

The probability that one waits for more than 5 minutes is

$$\mathbb{P}(X > 5) = 1 - \mathbb{P}(X \leq 5) = 1 - F_X(5)$$

The probability can be evaluated with a calculator or in R

```
> 1-pexp(5,rate=0.4)
[1] 0.1353353
```

c) Normal distribution

A continuous random variable X with pdf

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

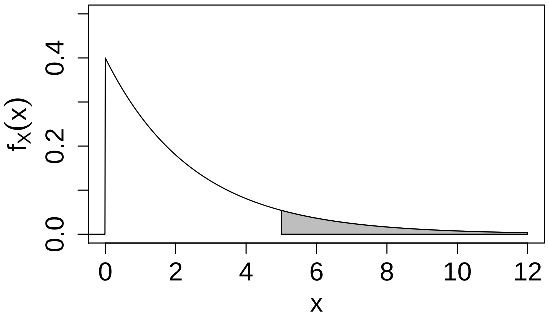


Figure 2.5: Pdf $f_X(x)$ of exponentially distributed bus waiting times, $X \sim \text{Exp}(0.4)$. The shaded area stands for the probability $\mathbb{P}(X > 5)$.

is said to be *normally distributed*.

The graph of the pdf is bell shaped with a maximum at μ and a width approximately given by the standard deviation σ.

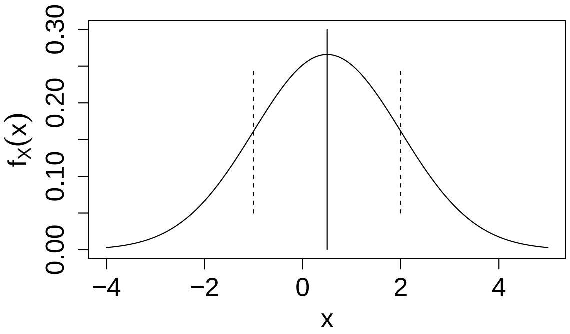


Figure 2.6: Pdf of a normal random variable with $\mu = 0.5$ and $\sigma = 1.5$. The vertical solid line at $x = \mu = 0.5$ indicates the value of μ whereas the two vertical dashed lines at $x = \mu \pm \sigma$ illustrate the meaning of the parameter σ.

Normalisation: Using the substitution $z = (x - \mu)/\sigma$ one can confirm that the pdf is normalised

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1\,.$$

The final integral $\int_{-\infty}^{\infty} e^{-z^2/2}dz = \sqrt{2\pi}$ cannot be computed by elementary methods.

Expectation: Using the same substitution $z = (x - \mu)/\sigma$ we obtain

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-\infty}^{\infty} (z\sigma + \mu) \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \sigma \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz + \mu = \mu \, .$$

Variance: Using again the substitution $z = (x - \mu)/\sigma$ we obtain

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx = \int_{-\infty}^{\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = \sigma^2$$

where the penultimate step has been done by integration by parts.

We say the random variable is a normal random variable with expectation μ and variance σ^2, $X \sim \mathrm{N}(\mu, \sigma^2)$. The normal pdf is determined by expectation and variance.

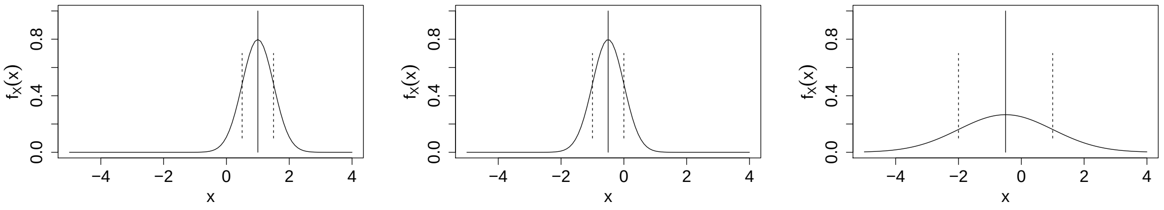


Figure 2.7: Pdf of a normal random variable with: $\mu = 1.0$ and $\sigma = 0.5$ (left), $\mu = -0.5$ and $\sigma = 0.5$ (middle), $\mu = -0.5$ and $\sigma = 1.5$ (right). The vertical solid line at $x = \mu$ indicates the value of the expectation whereas the two vertical dashed lines at $x = \mu \pm \sigma$ illustrate the meaning of the standard deviation σ.

Standardisation: In the previous calculations we have used the substitution

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

If we use this substitution to define a new random variable Z then this random variable has expectation $\mathrm{E}(Z) = (\mathrm{E}(X) - \mu)/\sigma = 0$ and variance $\mathrm{Var}(Z) = \mathrm{Var}\ (X)/\sigma^2 = 1$, i.e., $Z \sim \mathrm{N}(1, 0)$. We call Z a *standard normal random variable* and all computations can be reduced to the standard normal pdf. Since $X < a$, $X - \mu < a - \mu$ and $(X - \mu)/\sigma < (a - \mu)/\sigma$ denote the same events we have

$$\mathbb{P}(a < X \leq b) = \mathbb{P}((a - \mu)/\sigma < Z \leq (b - \mu)/\sigma) = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \, .$$

In what follows we will therefore consider the simpler case of a standard normal random variable with $\mu = 0$ and $\sigma = 1$.

Cumulative density function: The cdf of the standard normal random variable $Z \sim N(1,0)$ is given by

$$F_Z(z) = \mathbb{P}(Z \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \,.$$

The integral cannot be solved by elementary methods, it defines a new analytic function Φ, the so called *probability integral.* The graph of the function has a characteristic S shape (see figure 2.8) with $\Phi(-\infty) = 0$, $\Phi(\infty) = 1$ and $\Phi(-z) = 1 - \Phi(z)$. The function can be computed in R with the command `pnorm(x)` (if no additional arguments are given R assumes the expectation to be $\mu = 0$ and the variance to be $\sigma^2 = 1$).

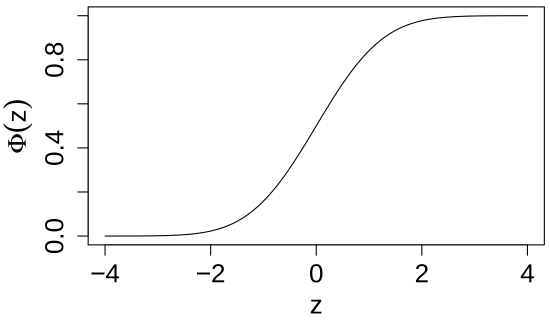


Figure 2.8: Probability integral (cdf of a standard normal random variable, $Z \sim \mathrm{N}(0,1)$).

Example 2.6 *Assume X is a normal random variable with expectation μ and variance σ^2, $X \sim N(\mu, \sigma^2)$. Compute the probability that X differs from its expectation by more than one standard deviation, i.e., compute $\mathbb{P}(|X - \mu| > \sigma)$.*

Since the events $|X - \mu| > \sigma$ and $|(X - \mu)/\sigma| > 1$ are identical standardisation gives

$$\mathbb{P}(|X - \mu| > \sigma) = \mathbb{P}(|Z| > 1)$$

with Z being a standard normal random variable. Since

$$\mathbb{P}(|Z| > 1) = \mathbb{P}(Z > 1) + \mathbb{P}(Z < -1) = 1 - \mathbb{P}(Z < 1) + \mathbb{P}(Z < -1)$$

we have

$$\mathbb{P}(|X - \mu| > \sigma) = \mathbb{P}(|Z| > 1) = 1 - \Phi(1) + \Phi(-1)$$

The value can be easily computed in R

```
> 1-pnorm(1)+pnorm(-1)
[1] 0.3173105
```

Hence a normal random variable differs from its expectation by more than one standard deviation with probability 31.7% *(see figure 2.9).*

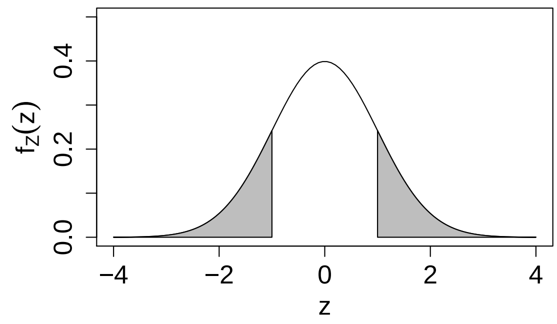


Figure 2.9: Pdf of a standard normal random variable $Z \sim \mathrm{N}(0,1)$ with expectation $\mu = 0$ and variance $\sigma^2 = 1$. The shading indicates the region where the random variable differs from its expectation by more than one standard deviation, i.e., it represents the probability $\mathbb{P}(|Z| > 1)$.

Similarly, the probability that a normal random variable differs from its mean by more than two standard deviations is given by

$$\mathbb{P}(|X - \mu| > 2\sigma) = \mathbb{P}(|Z| > 2) = 1 - \Phi(2) + \Phi(-2)$$

```
> 1-pnorm(2)+pnorm(-2)
[1] 0.04550026
```

i.e. by 4.6% *(see figure 2.10). Hence a normal random variable* $X \sim N(\mu, \sigma^2)$ *takes a value in the interval* $[\mu - 2\sigma, \mu + 2\sigma]$ *with 95.4% probability.*

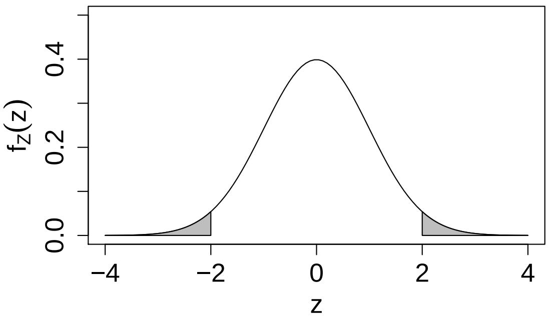


Figure 2.10: Pdf of a standard normal random variable $Z \sim \mathrm{N}(0, 1)$ with expectation $\mu = 0$ and variance $\sigma^2 = 1$. The shading indicates the region where the random variable differs from its expectation by more than two standard deviations, i.e., it represents the probability $\mathbb{P}(|Z| > 2)$.

z-scores/quartiles: Let $Z \sim \mathrm{N}(0, 1)$ denote a standard normal random variable. So far we computed probabilities that Z exceeds a threshold, $\mathbb{P}(Z > a)$. Now consider the reverse problem. Given a probability α let us compute the threshold z_α such that

$$\mathbb{P}(Z > z_\alpha) = \alpha$$

i.e. that Z exceeds the threshold with probability α, see figure 2.11.

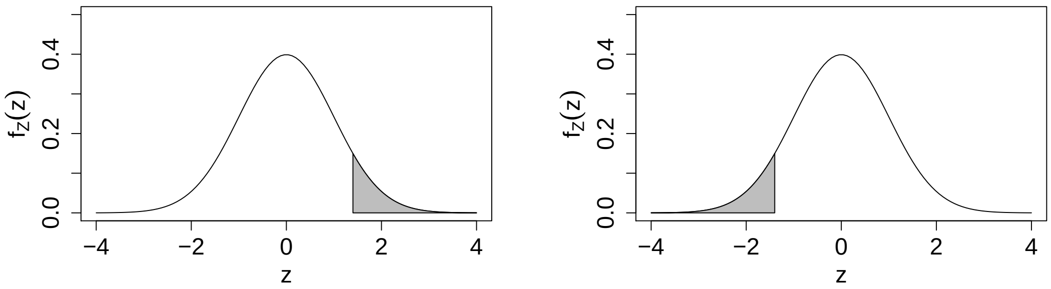


Figure 2.11: Pdf of a standard normal random variable $Z \sim \mathrm{N}(0, 1)$ and graphical illustration of the z-score z_α. Left: If α denotes the size of the shaded area then the left boundary of the shaded area is at $z = z_\alpha$. Right: If α denotes the size of the shaded area then the right boundary of the shaded area is at $z = -z_\alpha$.

To compute z_α observe that

$$\alpha = \mathbb{P}(Z > z_\alpha) = 1 - \mathbb{P}(Z \leq z_\alpha) = 1 - \Phi(z_\alpha)$$

We solve

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

for z_α using the inverse function Φ^{-1}

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

The inverse probability integral is available in R by the command `qnorm(x)`.

The value z_α is sometimes called a standard normal quantile. the value z_α is the threshold such that $\mathbb{P}(Z \leq z_\alpha) = 1 - \alpha$. Hence:

- $\alpha = 1/2$: $P(Z \leq z_{1/2}) = 1/2$. $z_{1/2}$ is called the median (see section 4**??**)
- $\alpha = 1/4$: $P(Z \leq z_{1/4}) = 3/4$. $z_{1/4}$ is called the upper quartile (see section 4**??**)
- $\alpha = 3/4$: $P(Z \leq z_{3/4}) = 1/4$. $z_{3/4}$ is called the lower quartile (see section refsec4b)

of the standard normal random variable Z.

Example 2.7 *Consider a normal random variable X with expectation μ and variance σ^2. Compute the threshold a so that X exceeds its expectation with probability 1%, i.e.. compute C such that $\mathbb{P}(X - \mu > C) = 0.01$. Standardisation tells us that*

$$0.01 = \mathbb{P}(X - \mu > C) = \mathbb{P}(Z > C/\sigma) = \mathbb{P}(Z > z_{0.01})$$

so that $C/\sigma = z_{0.01}$ with $z_{0.01} = \Phi^{-1}(1 - 0.01) = \Phi^{-1}(0.99)$.

```
> qnorm(0.99)
[1] 2.326348
```

Hence $C = \sigma 2.32\ldots$ and X exceeds its expectation by 2.3 standard deviations with probability 1%.

Since the pdf of Z is a symmetric function the area under the graph in the interval $[z_\alpha, \infty)$ is identical to the area under the graph in the interval $(-\infty, -z_\alpha]$, see figure 2.11. Hence we have shown

Lemma 2.1 (z-score) *Denote by Z a standard normal random variable, $Z \sim N(0,1)$. Then*

$$\alpha = \mathbb{P}(Z > z_\alpha) = \mathbb{P}(Z < -z_\alpha)$$

and

$$\mathbb{P}(|Z| > z_\alpha) = \mathbb{P}(Z > z_\alpha) + \mathbb{P}(Z < -z_\alpha) = 2\mathbb{P}(Z > z_\alpha) = 2\alpha$$

where

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

Example 2.8 *Consider a normal random variable X with expectation μ and variance σ^2. Compute the threshold b so that X deviates from its expectation with probability 5%, i.e.. compute b such that $\mathbb{P}(|X-\mu| > b) = 0.05$. Standardisation tells us that*

$$0.05 = \mathbb{P}(|X-\mu| > b|) = \mathbb{P}(|Z| > a/\sigma) = \mathbb{P}(|Z| > z_{0.025})$$

so that $b/\sigma = z_{0.025}$ with $z_{0.025} = \Phi^{-1}(1-0.025) = \Phi^{-1}(0.975)$.

```
> qnorm(0.975)
[1] 1.959964
```

Hence $b = \sigma 1.95\ldots$ and X deviates from its expectation by 1.96 standard deviations with probability 5%. That is in fact consistent with the result of example 2.6.

Moment generating function: The mgf, definition 1.5, of a normal random variable with expectation μ and variance σ^2 can be computed explicitly

Lemma 2.2 *Denote by X a normal random variable with expectation μ and variance σ^2. The mgf is given by*

$$M_X(t) = E(e^{tX}) = e^{t\mu + t^2\sigma^2/2}\,.$$

Proof: Using the standardisation $X = \mu + \sigma Z$ with the standard normal random variable $Z \sim \mathrm{N}(0,1)$ we have

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}(e^{t(\sigma Z + \mu)}) = e^{t\mu}\mathrm{E}(e^{t\sigma Z})\,.$$

With the definition 2.2 and the pdf of a standard normal random variable the expectation evaluates as

$$\begin{aligned}
\mathrm{E}(e^{t\sigma Z}) &= \int_{-\infty}^{\infty} e^{t\sigma z}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(z-t\sigma)^2/2}e^{\sigma^2t^2/2}dz \\
&= e^{\sigma^2t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du = e^{\sigma^2t^2/2}\,.
\end{aligned}$$

Hence the mgf of the random variable $X \sim \mathrm{N}(\mu, \sigma^2)$ reads

$$M_X(t) = e^{t\mu + t^2\sigma^2/2}\,.$$

□.

Obviously $M_X(0) = 1$. Furthermore $M'_X(t) = (\mu + t\sigma^2)e^{t\mu+t^2\sigma^2/2}$ and $M'_X(0) = \mu = \mathrm{E}(X)$ gives the expectation. Finally $M''_X(0) = \sigma^2 + \mu^2 = \mathrm{E}(X^2)$ gives the second moment in accordance with the variance.

Progress Check:

1. Explain the difference between a discrete and a continuous random variable.

2. What is meant by a pdf? How are pdfs related with probabilities?

3. What is the meaning of the area under the graph of a pdf.

4. State the expectation and the standard deviation of the exponential distribution.

5. Explain the meaning of the parameters which are contained in the normal distribution.

6. What does the term probability integral mean? Why is this concept useful?

7. What does the notion z-score mean. How is this concept related with probabilities?

§3 Central Limit Theorem

Example 3.1 *You toss a coin which has probability $p = 0.3$ coming up head $n = 50$ times. X_k, $k = 1, \ldots, n$ is the Bernoulli(p) random variable which gives 1 if the kth toss shows head (and zero if the coin shows tail). Hence the sum $X = X_1 + X_2 + \ldots + X_n$ is the random variable which counts the number of heads seen. We know that X obeys the Binomial(n, p) distribution (see example 1.3), that the expectation is given by $E(X) = np$ and the variance is given by $Var(X) = np(1-p)$. If n is large (say if $np(1-p)$ is larger than 10) the pmf of X is very well approximated by the pdf of a normal random variable with $\mu = np$ and $\sigma^2 = np(1-p)$, see figure 3.1.*

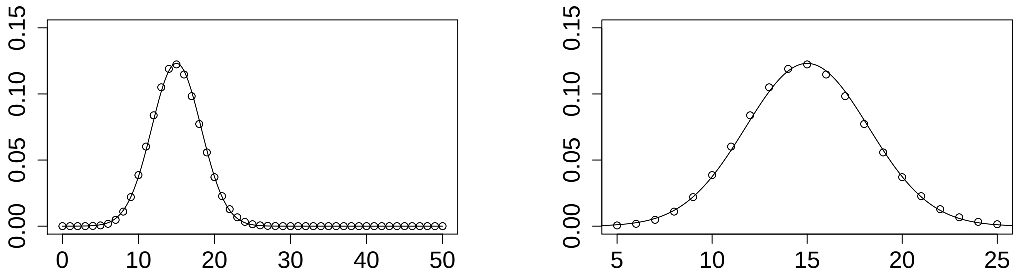


Figure 3.1: Symbols: Binomial distribution with $n = 50$ and $p = 0.3$. Line: Normal distribution with expectation $\mu = np = 15$ and variance $\sigma^2 = np(1-p) = 10.5$. Left: Distributions over the full range. Right: same data on a finer scale.

The exact probability that we see k heads is given by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \,.$$

Let Y denote the continuous normal random variable Y with expectation $\mu = np$ and variance $\sigma^2 = np(1-p)$, $Y \sim N(np, np(1-p))$. The exact probability $\mathbb{P}(X = k)$ is very well approximated by the probability of Y taking values in the symmetric interval $[k - 1/2, k + 1/2]$, i.e.,

$$\mathbb{P}(X = k) \approx \mathbb{P}(k - 1/2 \le Y \le k + 1/2) = \int_{k-1/2}^{k+1/2} \frac{1}{\sqrt{2\pi}} e^{-(t-\mu)^2/(2\sigma^2)} dt$$

To judge the quality of the approximation we compute the exact value (for $n = 50$, $p = 0.3$, $k = 16$) with R

```
> n<-50
> p<-0.3
> k<-16
> dbinom(k,size=n,prob=p)
[1] 0.1147002
```

To compute the approximation we first use standardisation to reduce the quantity to the probability integral

$$\begin{aligned}\mathbb{P}(k-1/2 \leq Y \leq k+1/2) &= \mathbb{P}(Y \leq k+1/2) - \mathbb{P}(Y \leq k-1/2)\\ &= \mathbb{P}\left(\frac{Y-\mu}{\sigma} < \frac{k+1/2-\mu}{\sigma}\right) - \mathbb{P}\left(\frac{Y-\mu}{\sigma} < \frac{k-1/2-\mu}{\sigma}\right)\\ &= \Phi\left(\frac{k+1/2-\mu}{\sigma}\right) - \Phi\left(\frac{k-1/2-\mu}{\sigma}\right)\end{aligned}$$

The numerical value can now be easily computed using R (`n`*,* `p` *and* `k` *have been already assigned in the R commands above)*

```
> mu<-n*p
> sigma<-sqrt(n*p*(1-p))
> pnorm((k+1/2-mu)/sigma)-pnorm((k-1/2-mu)/sigma)
[1] 0.1169709
```

The approximation reproduces the first two digits.

There is no need to approximate the binomial distribution by a normal distribution. But the example tells us that sums of random variables can be well approximated by a normal random variable. Such a statement is in fact valid in general, and that is the reason why normal random variables appear frequently in mathematics and in applications.

Proposition 3.1 (Central limit theorem) *Let X_1, X_2, ..., X_n be n independent identically distributed random variables with expectation $E(X_k) = \mu$ and variance $Var(X_k) =$*

σ^2. Denote by $X = X_1+X_2+\ldots+X_n$ the sum of these random variables and by Y a normal (continuous) random variable with expectation $E(Y) = n\mu$ and variance $Var(Y) = n\sigma^2$. Then the pmf of X is well approximated by the pdf of Y for large values of n

$$\mathbb{P}(a \leq X \leq b) \approx \mathbb{P}(a - 1/2 \leq Y \leq b + 1/2)\,.$$

The proposition has been stated in an informal way, i.e., we have not defined what we mean by "well approximated", and what we mean by "large values of n". Writing down the proposition in rigorous terms would require a machinery way beyond first year mathematics (and that means a proof is even more beyond first year mathematics). Nevertheless the proposition is key to almost all approaches in mathematical statistics.

Example 3.2 *We roll a fair die 120 times and record the sum of all rolls. X_k denotes the number shown in the kth roll. The sum of all rolls is the random variable $X = X_1 + X_2 + \ldots + X_n$. Compute the probability that the sum of all rolls is between (and including) 410 and 430, i.e., compute $\mathbb{P}(410 \leq X \leq 430)$.*

Here $n = 120$, $\mu = E(X_k) = 7/2$ and $\sigma^2 = Var(X_k) = 35/12$. We may apply the normal approximation of the central limit theorem since n is reasonably large. Denote by Y a normal continuous random variable with expectation $\mu_Y = E(Y) = n\mu = 420$ and variance $\sigma_Y^2 = Var(Y) = n\sigma^2 = 350$. Then

$$\mathbb{P}(410 \leq X \leq 430) \approx \mathbb{P}(410{-}1/2 \leq Y \leq 430{+}1/2) = \mathbb{P}(Y \leq 430{+}1/2) - \mathbb{P}(Y \leq 410{-}1/2)\,.$$

Using standardisation (with $\mu_Y = 420$ and $\sigma_Y = \sqrt{350}$)

$$\begin{aligned} & P(Y \leq 430 + 1/2) - \mathbb{P}(Y \leq 410 - 1/2) \\ = \; & \mathbb{P}\left(\frac{Y - 420}{\sqrt{350}} \leq \frac{430 + 1/2 - 420}{\sqrt{350}}\right) - \mathbb{P}\left(\frac{Y - 420}{\sqrt{350}} \leq \frac{410 - 1/2 - 420}{\sqrt{350}}\right) \\ = \; & \Phi\left(\frac{10 + 1/2}{\sqrt{350}}\right) - \Phi\left(\frac{-10 - 1/2}{\sqrt{350}}\right) \end{aligned}$$

the numerical result can be easily obtained using R

```
> pnorm((10+1/2)/sqrt(350))- pnorm((-10-1/2)/sqrt(350))
[1] 0.4253719
```

That means with 42.5% *chance the sum of the rolls is between 410 and 430.*

Furthermore within the normal approximation example 2.6 tells us that with 95.4% *chance the random variable X (i.e. the sum) takes values in the interval* $[\mu_X - 2\sigma_X, \mu_X + 2\sigma_X]$ *(here: $\mu_X = E(X) = nE(X_k) = 420$, $\sigma_X^2 = Var(X) = n\,Var(X_k) = 350$), i.e. in the interval* $[382, 458]$.

The central limit theorem guarantees normal behaviour for a large number of terms in the sum. If the individual terms are normal random variables as well, then the sum is a normal random variable even for a small number of terms.

Proposition 3.2 *Let X_1, X_2, …, X_n be n independent normal random variables, $X_k \sim N(\mu, \sigma^2)$ with expectation $E(X_k) = \mu$ and variance $Var(X_k) = \sigma^2$. Denote by $X = X_1 + X_2 + \ldots + X_n$ the sum of these random variables. Then X is a normal random variable, $X \sim N(n\mu, n\sigma^2)$ with expectation $E(X) = n\mu$ and variance $Var(X) = n\sigma^2$.*

Proof: The mgf of X reads

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}(e^{t(X_1+X_2+\ldots+X_n)}) = \mathrm{E}(e^{tX_1}e^{tX_2}\ldots e^{tX_n}) = \mathrm{E}(e^{tX_1})\mathrm{E}(e^{tX_2})\ldots\mathrm{E}(e^{tX_n})$$

where we have used the independence of X_k to write the expectation of the product as the product of the expectations. Each factor is the mgf of a normal random variable, so that lemma 2.2 tells us that

$$M(t) = \left(e^{t\mu+t^2\sigma^2/2}\right)^n = e^{tn\mu+t^2n\sigma^2/2}\,.$$

This result is the mgf of a normal random variable with expectation $n\mu$ and variance $n\sigma^2$. □.

Progress Check:

1. Explain in lay terms what is meant by the central limit theorem.

2. How can the central limit theorem be used to evaluate probabilities of sums of random variables?

3. Consider the binomial distribution. Under which conditions can the binomial distribution be approximated by a normal distribution.

4. Let Z_1 and Z_2 be two standard normal random variables. State the pdf of the sum, $Z_1 + Z_2$.

5. You roll a die 3000 times. How can you compute the probability that the sum of all rolls exceeds 15000 using the central limit theorem.

6. You roll a die 3 times. Explain why you cannot use the central limit theorem to compute the probability that the sum of all rolls exceeds 15.

7. A bag contains 500 gold and 500 copper coins. You pick 700 coins without replacement. Let X_k count whether the kth pick has resulted in a gold coin (i.e. $X_k = 1$ if the kth pick has been a gold coin and $X_k = 0$ otherwise) so that $X = X_1 + X_2 + \ldots + X_{700}$ is the total number of gold coins you have picked. Explain why one cannot use the central limit theorem to compute the pdf of X.

§4 Exploratory Data Analysis

We want to use the previous abstract concepts to obtain information from measured data sets. A first step is just the visualisation of some statistical aspects of data sets.

a) Basic concepts

Qualitative data:

Example 4.1 *Data for the 2018 final exam in MTH4107 lists for each of the 207 students the grade* $(x_1, x_2, \ldots, x_{207})$*. This type of data is ordinal (as grades have a natural order: A,B,C,D,E,F). To display a data summary compute frequencies, i.e. list for each grade the number of students: A grade (124 students), grade B (37 students), grade C (23 students), grade D (8 students), grade E (4 students), grade F (11 students).*

For qualitative data (ordinal or categorical) one can lists for each category the number of occurrences. A *frequency distribution* lists for each category the number of occurrences (as above). The *relative frequency* is the proportion of observations within a category, i.e., the frequency divided by the sum of all frequencies. A relative frequency distribution lists for each category the relative frequency.

Example 4.2 *In example 4.1 the sum of all frequencies (total number of students) is 207. The relative frequency distribution is: grade A: 124/207, grade B: 37/207, grade C: 23/207, grade D: 8/207, grade E: 4/207, grade F: 11/207.*

Frequency distributions may be visualised by *bar charts* where one plots for each category the frequency or relative frequency by a bar of corresponding height. R easily allows to produce bar charts with the command `barplot` and two arguments, containing the data and the names of the categories (see figure 4.1).

```
> x<-c(124,37,23,8,4,11)
> grade<-c("A","B","C","D","E","F")
> barplot(x,names.arg=grade)
```

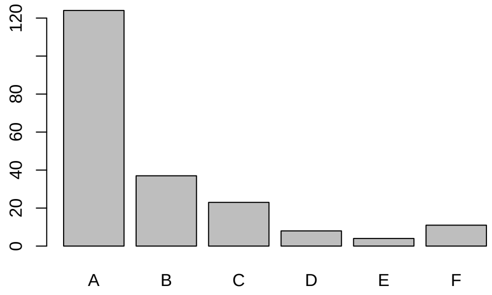


Figure 4.1: Bar chart of the grades in the MTH4107 final exam in 2018.

For a relative frequency distribution just compute the sum of all frequencies by `sum(x)` and divide each entry of `x` by the sum using `x/sum(x)`.

Quantitative discrete data:

Example 4.3 *The marks of the students failing MTH4107 in 2018 were: 26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34. A frequency distribution can be obtained as before by using e.g. the marks/data as categories, i.e., mark 23 (1 student), mark 25 (1 student), mark 32 (2 students), mark 33 (1 student), mark 34 (1 student), mark 35 (2 students), mark 38 (1 student), mark 39 (1 student).*

A bar chart can be directly obtained from the data set using the command `hist`. The option `breaks` tells R how to group the data, here we chose break points at half integer marks $(22.5, 23.5, \ldots, 38.5, 39.5)$.

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> hist(x,breaks=seq(22.5,39.5,1))
```

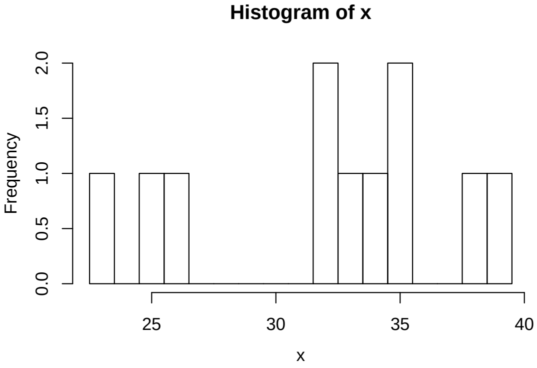


Figure 4.2: Histogram of the fails in the MTH4107 final exam in 2018.

Bar charts/histograms (for quantitative data or ordinal data) can be characterised as "symmetric", "unimodal" (one maximum), "multimodal" (more than one maximum), "skewed right" (tilted towards right), "skewed left" (tilted towards left).

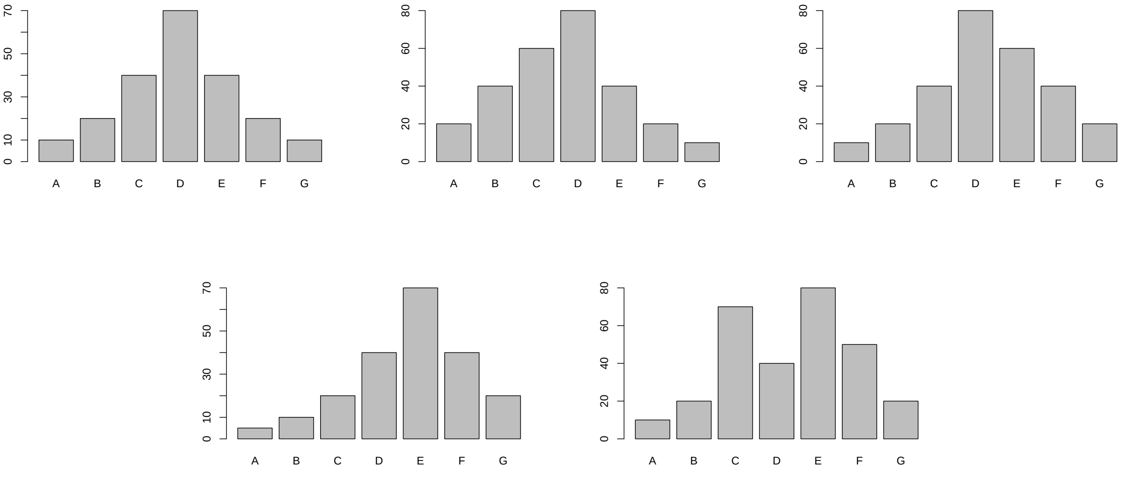


Figure 4.3: Top (from left to right): bar chart of a symmetric, a skewed-left and a skewed-right distribution. Bottom (from left to right): bar chart of a unimodal and of a bimodal/multimodal distribution.

b) Measures for centrality and dispersion

Consider quantitative datasets. The number of entries (observations) is denoted by n. Example 4.3 is a quantitative dataset with $n = 11$.

Mean and median: Denote by $(x_1, x_2, \ldots, x_n)$ the data set. The so called sample mean is defined by

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Assume the data points $(x_1, x_2, \ldots, x_n)$ are sorted in increasing order (i.e., $x_k \leq x_{k+1}$), a so called *order statistics*. If n is odd the entry in the middle of the sequence has index $(n+1)/2$ and we call the data entry $x_{(n+1)/2} = Q_2$ the median. If n is even there are "two middle entries" at $n/2$ and $n/2+1$ and we call the average $(x_{n/2} + x_{n/2+1})/2 = Q_2$ the median.

Example 4.4 *Consider the data set of example 4.3. The order statistics of this data set reads 23, 25, 26, 32, 32, 33, 34, 35, 35, 38, 39 with* $n = 11$*. Hence entry* $(11+1)/2 = 6$ *is the middle entry and* $x_6 = 33$ *is the median. A simple calculation shows that the sample mean is* $\bar{x} = 32$*. Mean and median can be easily evaluated using R*

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> mean(x); median(x)
[1] 32
[1] 33
```

Remark:

- Mean and median quantify the average value of the data set, e.g., the value where the bar chart has its central part (see figure 4.2).
- If the bar chart is symmetric we have $\bar{x} = Q_2$, if the bar chart is skewed left we have $\bar{x} < Q_2$, and if the bar chart is skewed right we have $\bar{x} > Q_2$.

Variance and quartiles: Given a data set $(x_1, x_2, \ldots, x_n)$ the sample mean locates the central part of the data set, and the numbers $x_k - \bar{x}$ somehow characterise the spread of the data set. The sample mean resembles an expectation value if we assume all entries in the data set to be equally likely (i.e. having probability $1/n$ to occur). Hence it is tempting to define a measure for the spread along the lines of a variance of a random variable as

$$\sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 .$$

This quantity is called the *population variance* (if the size n of the sample is the entire population size). For a pretty subtle reason it is better to consider the so called *sample variance* defined by

$$s_x^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

since the population variance is a biased estimator (to be discussed in section 5**??**). The R command `var(x)` computes the sample variance of the data set. As discussed previously the *sample standard deviation* s_x (or simply s) gives a measure for the spread/dispersion of the bar chart. The standard deviation can be computed using `sd(x)`.

Suppose we have sorted the data in increasing order $(x_1, x_2, \ldots, x_n)$. The median Q_2 divides the data set in two parts of equal size, a lower part and an upper part, e.g. $(x_1, \ldots, x_{(n+1)/2-1})$ and $(x_{(n+1)/2+1}, \ldots x_n)$ if n is odd, or $(x_1, \ldots, x_{n/2})$ and $(x_{n/2+1}, \ldots x_n)$ if n is even. The median of the lower part is called the *lower quartile* Q_1 and the median of the upper part is called the *upper quartile* Q_3. Hence the quartiles Q_1, Q_2, Q_3 split the data set in four parts of equal size.

Example 4.5 *Consider the order statistics of the data set of example 4.3, (23, 25, 26, 32, 32, 33, 34, 35, 35, 38, 39). The median is* $Q_2 = 33$*. The lower half of the sorted data set is (23, 25, 26, 32, 32) and its median gives the lower quartile* $Q_1 = 26$ *of the full data set. The upper half of the data set is (34, 35, 35, 38, 39) and its median is the upper quartile* $Q_3 = 35$ *of the original data set.*

Quartiles can be directly computed in R using the command `quantile`*. The option* `probs=0.25` *selects the lower quartile, the option* `probs=0.75` *selects the upper quartile. There are at least nine (!) different non equivalent ways to define quartiles, the option* `type=2` *selects the definition presented here.*

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> c(quantile(x,probs=0.25, type=2), quantile(x,probs=0.75, type=2))
25% 75%
 26  35
```

Median and quartiles split the bar chart of relative frequencies in regions where the bar chart has 25% *weight, see figure 4.4*

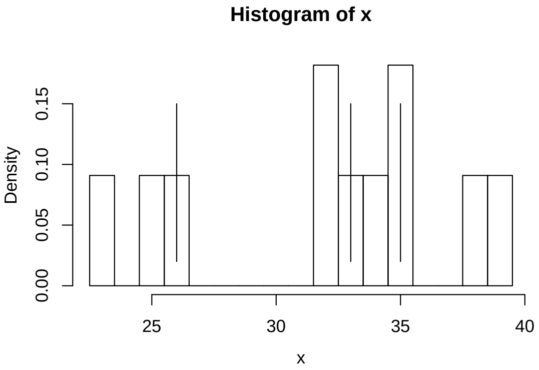


Figure 4.4: Relative frequencies of the fails in the MTH4107 final exam in 2018, see figure 4.2. The vertical lines are (from left to right) lower quartile Q_1, median Q_2 and upper quartile Q_3.

The *interquartile range* (IQR) is defined by

$$IQR = Q_3 - Q_1 \, .$$

It gives a measure for the spread/dispersion of the data set/bar chart. `IQR(x,type=2)` computes the IQR.

Example 4.6 *Consider the data set of example 4.3. Sample variance, sample standard deviation, and IQR are given by*

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> var(x)
[1] 27.4
> sd(x); IQR(x,type=2)
[1] 5.234501
[1] 9
```

As expected the IQR is at about twice the standard deviation.

Five-number summary and boxplots: Denote by $(x_1, \ldots, x_n)$ a quantitative data set. Denote by m_x the smallest number in the data set (the minimum) and by M_x the largest number in the data set (the maximum). The five numbers (the so called *five-number summary*) m_x, Q_1, Q_2, Q_3, M_x give a brief quantitative characterisation of the data set (in terms of range, central part, and spread).

Example 4.7 *Consider the data set of example 4.3. The five-number summary can be computed with the R command* `summary` *(which returns as well the mean, the option* `quantile.type` *specifies the version of quantiles used)*

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> summary(x,quantile.type=2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     23      26      33      32      35      39 
```

The five-number summary can be visualised in a so called boxplot. The boxplot consists of a rectangle positioned above the interval $[Q_1, Q_3]$. The median is shown by a line through the box positioned at Q_2. Two lines outside the box extend from Q_3 to the maximum of the data set, and from Q_1 to the minimum of the data set. A boxplot for the data set of example 4.3 is shown in figure 4.5

A boxplot can be produced in R with the command `boxplot(x)` The command uses a different definition of the quartiles, hence the position of Q_1 and Q_3 may differ slightly from the previous calculation. A boxplot may contain as well particular data points, which are considered to be outliers (see the R documentation for details).

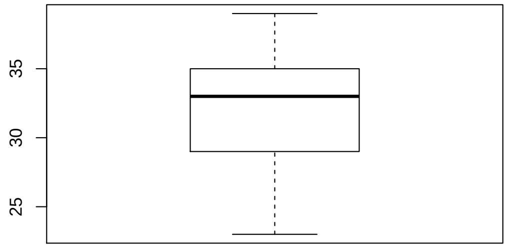


Figure 4.5: Boxplot of the data set of example 4.3, with median $Q_2 = 33$ indicated by the horizontal line through the box, upper quartile $Q_3 = 33$ indicated by the upper bound of the box, lower quartile $Q_1 = 29$ indicated by the lower bound of the box, and the two whiskers extending to minimum 23 and maximum 39 of the data set.

c) Correlation and scatter diagrams

So far we have considered quantitative data of the type $(x_1, x_2, \ldots, x_n)$, i.e., a single quantity x_k measured for each item k, so called *univariate data*. Let us now consider the case that we measure (at least) two quantities x_k and y_k for each item, so called *multivariate data* $([x_1, y_1], [x_2, y_2], \ldots, [x_n, y_n])$.

Example 4.8 *The raw exam marks of 29 students in the 2019 Calculus I exam and the Introduction to Probability exam are listed in a table:*

Calc I	*62*	*84*	*18*	*64*	*88*	*69*	*66*	*69*	*73*	*84*	*95*	*82*	*68*	*91*	*65*	*100*	*83*	*95*	*91*	*82*	*44*	*83*	*45*	*82*	*93*	*74*	*73*	*91*	*14*
Int. Prob	*84*	*73*	*47*	*74*	*64*	*86*	*88*	*76*	*76*	*79*	*98*	*85*	*72*	*81*	*93*	*91*	*90*	*89*	*95*	*77*	*91*	*75*	*62*	*90*	*61*	*77*	*69*	*65*	*49*

For each student there are two mark entries, i.e., we have bivariate/multivariate data. For each module (each component) one can perform the univariate data analysis described in the previous section. However, the interesting aspect is whether there are correlations between the marks in both modules.

Scatter plots: To visualise any correlations within a multivariate data set ($[x_1, y_1]$, $[x_2, y_2], \ldots, [x_n, y_n]$) produce a plot where you show x_k versus y_k, a so called *scatter plot.* If (an extreme case) y_k would be a function of x_k, i.e., $y_k = f(x_k)$ the scatter plot would just show the graph of the function. If points are scattered in the plane it means that there is no such simple relationship.

Example 4.9 *Represent the multivariate data of example 4.8 in a scatter plot, i.e., plot Probability marks y_k vs. Calculus marks x_k. In R assign marks to lists and plot lists against each other.*

```
x<-c(62,84,18,64,88,69,66,69,73,84,95,82,68,91,65,
     100,83,95,91,82,44,83,45,82,93,74,73,91,14)
y<-c(84,73,47,74,64,86,88,76,76,79,98,85,72,81,93,
     91,90,89,95,77,91,75,62,90,61,77,69,65,49)
plot(x,y,xlab="Calc I",ylab="Int. Prob.")
```

The scatter plot indicates that there is some minor positive correlation between the two module marks as a higher mark in Calculus I has a tendency to have a higher mark in Introduction to Probability. But the correlation does not look strong.

Correlation coefficient: To quantify a (linear) correlation one can resort to the idea of covariance and correlation coefficients. If you recall the definition of the covariance of two random variables $\mathrm{Cov}(X, Y) = \mathrm{E}(X - E(X)(Y - E(Y))$ and if you compare the definition of variance $\mathrm{Var}(X)$ and sample variance s_x^2 it is sensible to introduce the *sample covariance*

$$s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

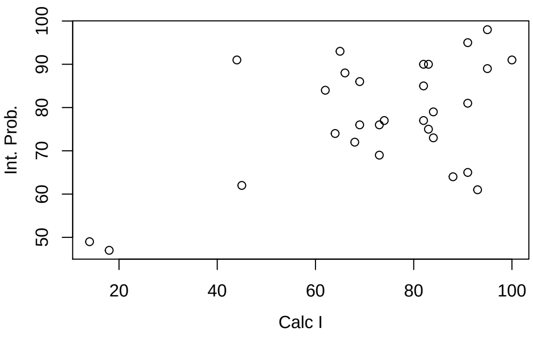


Figure 4.6: Scatter plot of the 2019 final exam marks of Calculus I and Introduction to Probability for 29 students.

as a measure for the correlation between the two data sets. Like the correlation coefficient $\mathrm{corr}(X,Y) = Cov(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$ one divides by the sample variances to introduce the *sample linear correlation coefficient*

$$r = \frac{s_{xy}}{s_x s_y} \, .$$

The sample correlation coefficient has the properties

- $-1 \leq r \leq 1$
- If the two data are linearly related with positive slope $a > 0$, $y_k = ax_k + b$, then $r = 1$
- If the two data are linearly related with positive slope $a < 0$, $y_k = ax_k + b$, then $r = -1$

Hence r quantifies the degree by which data are linearly correlated. The proof of these statements is fairly straightforward but requires some familiarity with manipulating sums like, e.g., the Cauchy Schwarz inequality (see as well the proofs for analogous properties of the correlation coefficient).

Example 4.10 *Consider the multivariate data of example 4.8. The sample correlation coefficient can be computed with the R command* `cor`*.*

```
x<-c(62,84,18,64,88,69,66,69,73,84,95,82,68,91,65,
     100,83,95,91,82,44,83,45,82,93,74,73,91,14)
y<-c(84,73,47,74,64,86,88,76,76,79,98,85,72,81,93,
     91,90,89,95,77,91,75,62,90,61,77,69,65,49)
> cor(x,y)
[1] 0.5281328
```

The correlation coefficient is positive (indicating that a higher mark in Calculus I points towards a higher mark in Introduction to probability, and vice versa) but there is no linear relation between both marks as the correlation coefficient differs substantially from 1.

Remark:

- A positive or negative correlation does not mean any causality (e.g. that one feature causes the other). Correlations may be mitigated by other factors or they may be spurious (i.a. accidental).
- A certain value of the correlation coefficient does not mean that the value is significant (i.e. relevant). We will discuss those issues in section 6 (these kind of questions are actually a key aspect in Statistics).

Progress Check:

1. Explain what is meant by a frequency distribution and by a relative frequency distribution.
2. Which measures for centrality do you know? What do these measures tell you?
3. Which measures for dispersion do you know? What do these measures tell you?
4. Which information is contained in a boxplot?
5. The QMUL student evaluation questionnaires contain questions which can be answered by "definitely agree", "mostly agree", "neutral, "mostly disagree", "definitely disagree". Which of the following quantities make sense to summarise the answers: the mean, the median, the standard deviation, the interquartile range?
6. How do you produce a scatter plot, and which information does the plot contain?
7. Which information is contained in the correlation coefficient? What does a positive/negative value mean, and how is this reflected in the scatter plot?

§5 Sampling Distributions

Example 5.1 *We want to obtain the average height of the British population. We could do a census, i.e., recording the height of each individual and then compute the arithmetic mean, which gives the exact average height (the so called population mean). But such a procedure is impractical or at least far to expensive. Instead we could do a survey, i.e., we take a random sample of say* $n = 10,000$ *people, record their heights* $(x_1, x_2, \ldots, x_n)$ *and compute the sample mean* $\bar{x}$. *But how accurate is such a sample mean? If we do another survey we obtain a different data set* $(x_1, x_2, \ldots, x_n)$ *and a different sample mean. In fact each entry* x_k *can be viewed as the outcome of a random variable* X_k *which gives the value of the* k*th data entry. The sample mean* $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ *becomes a random variable as well. The random variables* X_1,X_2,…,X_n *are independent since we take a random sample of the population and their expectation* $\mu = E(X_k)$ *is the exact population mean (the number we want to estimate). So the research question is: what can we say about the distribution of* $\bar{X}$ *and how likely is it that its values differ much from* μ.

In general: if we have collected a random sample of size n, then the actual data set $(x_1, x_2, \ldots, x_n)$ may be viewed as a realisation of n (mutually) independent random variables $(X_1, X_2, \ldots, X_n)$ which have identical distribution. In fact, the random variables are independent if we sample with replacement (which is unlikely in most surveys) or if we sample from an infinite population. If we sample without replacement then the size N of the population needs to be much larger than the sample size n to guarantee independence. See for instance the hypergeometric distribution (see figure 1.4) which covers picking balls without replacement and the Binomial distribution (see figure 1.2) which covers picking balls with replacement.

a) Estimators for mean and proportion

Denote by $X_1, X_2, \ldots, X_n$ n (mutually) independent identically distributed random variables with expectation $\mu = \mathrm{E}(X)$ and variance $\mathrm{Var}(X_k) = \sigma^2$. Denote by

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$$

the sample mean. Since we use the sample mean to estimate μ the sample mean is called an *estimator* for the mean μ. Given a particular sample with values $x_1, x_2, \ldots, x_n$ the numerical value of the sample mean $\bar{x}$ is called a *point estimate* for the expectation μ.

Here we will first study the properties of the sample mean, assuming that μ and σ are given. This knowledge will then be used in section 6 to estimate parameters of a distribution from a sample.

Definition 5.1 (Statistic) *A real function $g(X_1, X_2, \ldots, X_n)$ of random variables X_1, X_2, …, X_n is called a* statistic.

Obviously the sample mean is a statistic.

Proposition 5.1 *Let X_1, X_2,…, X_n be n mutually independent random variables having the same expectation and variance, $\mu = E(X_k)$ and $\sigma^2 = Var(X_k)$ for $k = 1, \ldots, n$. Then the expectation $\mu_{\bar{X}}$ and the variance $\sigma^2_{\bar{X}}$ of the sample mean $\bar{X}$ are given by*

a) $\mu_{\bar{X}} = E(\bar{X}) = \mu$

b) $\sigma^2_{\bar{X}} = Var(\bar{X}) = \dfrac{\sigma^2}{n}$

Proof: The proof is a simple application of the linearity of expectation and independence

a) Linearity of expectation tells us that

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\mathrm{E}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + \mathrm{E}(X_n)\right) = \frac{1}{n}n\mu = \mu$$

b) Mutual independence of the random variables tells us that

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

□.

Remark:

- The pdf of the random variable $\bar{X}$ is called the sampling distribution (of the sample mean)
- The variance of $\bar{X}$, $\sigma_{\bar{X}} = \sigma^2/n$, becomes smaller when the sample size n increases.
- The central limit theorem (see proposition 3.1) tells us that regardless of the shape of the underlying population the sampling distribution of $\bar{X}$ becomes approximately normal, $\bar{X} \sim N(\mu, \sigma^2/n)$ for large n.

Example 5.2 *The exact population mean of the final exam marks in Introduction to Probability in 2019 has been $\mu = 70.21$ with variance $\sigma^2 = 380.27$ (by computing the averages among all students). You ask a random sample of 10 students who took the exam for their final exam mark and you compute the sample mean. Compute the probability that the sample mean is 5 or more marks below the actual mean μ.*

If X_k is the random variable which gives the mark of the kth student in your sample the sample mean is

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

with $n = 10$. We are supposed to compute the probability $\mathbb{P}(\bar{X} \leq 70.21 - 5)$.

Proposition 3.1 tells us that $\bar{X}$ is an approximate normal random variable with mean and variance given by proposition 5.1, $\bar{X} = N(\mu, \sigma^2/n)$ with $\mu = 70.21$, $\sigma^2 = 380.27$, and $n = 10$. This statement has to be taken with a grain of salt as the sample size n is fairly small. Standardisation of the random normal variable with $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$ gives

$$\mathbb{P}(\bar{X} \leq 65.21) = \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq \frac{-5}{\sqrt{\sigma^2/n}}\right) = \Phi\left(\frac{-5}{\sqrt{\sigma^2/n}}\right) .$$

The numerical value can be easily evaluated in R

```
> pnorm(-5/sqrt(38.027))
[1] 0.2087348
```

Hence at about 20.9% of the random samples (that means at about one in five) underestimate the mean mark by 5 or more marks.

Example 5.3 *Consider the setup of example 5.2. How large must the size n of the random sample of students be so that less than 5% of the random samples have a sample mean which is 5 or more marks below the actual mean (i.e. that less than 5% of the random samples underestimate the actual mean mark).*

In this case we have to compute n such that

$$\mathbb{P}(\bar{X} \leq 65.21) < 0.05\,.$$

If we use standardisation for the right hand side we obtain (see example 5.2)

$$\Phi\left(\frac{-5}{\sqrt{\sigma^2/n}}\right) < 0.05\,.$$

Since the probability integral is a monotonic increasing function (i.e. application of the inverse function preserves inequalities) application of Φ^{-1} gives

$$\frac{-5}{\sqrt{\sigma^2/n}} < \Phi^{-1}(0.05)$$

which can be easily solved for n (observe that $\Phi^{-1}(0.05)$ is negative, see e.g. figure 2.8 !)

$$n > \frac{(-\Phi^{-1}(0.05))^2\sigma^2}{5^2}\,.$$

The numerical value can be easily computed in R

```
> (-qnorm(0.05))^2*380.27/5^2
[1] 41.15348
```

You would have to ask more than 41 randomly sampled students (and all of them would have to be willing to tell you their mark). Again some critical thinking should be exercised. The required value of n is fairly large compared to the population size, and the assumption of independence may be violated.

So far we based our analysis on the mean μ and the variance σ^2 of each observation X_k to evaluate properties of the sample mean $\bar{X}$. The situation becomes simpler if we consider the special case to estimate a proportion of a population.

Example 5.4 *The true proportion p of home owners in the UK can be obtained by a census (i.e. ask each individual whether they own a property). Alternatively one may estimate the proportion by a survey, i.e., from a random sample $(x_1, \ldots, x_n)$ of size n. One may record home owners by a digit $x_k = 1$ and non home owners by a digit $x_k = 0$. Hence home ownership becomes a Bernoulli(p) random variable X_k (taking values 0 and 1). The estimate of the proportion p is the number of home owners (i.e. the number of ones) divided by the sample size n, i.e., the sample proportion $\bar{X} = (X_1 + \ldots + X_n)/n$ gives an estimator for the proportion. Since E(X_ℓ) $= p$ and Var(X_ℓ) $= p(1-p)$ proposition 5.1 tells us that E($\bar{X}$) $= p$ and Var($\bar{X}$) $= p(1-p)/n$.*

In general: If $X_1, \ldots, X_n$ are Bernoulli(p) random variables telling us whether an individual belongs to a group in the population then the sample proportion $\bar{X}$ is an unbiased estimator for the proportion of members of the group and

$$\mathrm{E}(\bar{X}) = p, \quad \mathrm{Var}(\bar{X}) = \frac{p(1-p)}{n} \,.$$

Example 5.5 *In the 2019 Calculus I final exam 41.4% of students got an A grade. You ask a random sample of 15 students (who took that exam) whether they obtained an A. Compute the probability that more than 50% of students in your sample obtained an A grade.*

Denote by $X_1, \ldots, X_n$ the $n = 15$ Bernoulli(p) random variables with $p = 0.414$. The variables count whether a student in the sample got an A (i.e. $X_k = 1$ if student k got an A, and $X_k = 0$ otherwise). We have E(X_ℓ) $= p$ and Var(X_ℓ) $= p(1-p)$. The sample proportion $\bar{X}$ counts the fraction of students in the sample which got an A grade. We need to compute $\mathbb{P}(\bar{X} > 0.5)$. Proposition 5.1 tells us that $\mu_{\bar{X}} = p$ and Var($\bar{X}$) $= p(1-p)/n$ with $p = 0.414$ and $n = 15$.

Within the normal approximation standardisation tells us that

$$\mathbb{P}(\bar{X} > 0.5) = 1 - \mathbb{P}(\bar{X} \le 0.5) = 1 - \mathbb{P}(\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \le \frac{0.5 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = 1 - \Phi\left(\frac{0.5 - p}{\sqrt{p(1-p)/n}}\right) .$$

Evaluation in R is straightforward

```
> 1-pnorm((0.5-0.414)/sqrt(0.414*(1-0.414)/15))
[1] 0.249447
```

That means for at about 25% (i.e a quarter) of the random samples one obtains a sample proportion with more than 50% A grades.

Remark:

- Independence condition: To simplify the analysis we assume that the random variables $X_1, \dots, X_n$ in a random sample of size n are independent. If we sample randomly from an infinite population or from a finite population with replacement then independence is guaranteed. If we sample from a finite population of size N without replacement independence is a good approximation if the sample size is at most 5% of the population size, $n \leq 0.05N$.
- Normality condition: If each individual is modelled by a normal random variable $X_k \sim \mathrm{N}(\mu, \sigma^2)$ then the sample mean $\bar{X}$ is a normal random variable as well, by proposition 3.2, for any sample size n. If X_k are Bernoulli(p) random variables the sample mean $\bar{X}$ is well approximated by a normal random variable if the sample size is sufficiently large, $np(1-p) > 10$. For general random variables X_k the sample mean is often well approximated by a normal random variable if $n > 20$.

b) Bias

Definition 5.2 (Unbiased estimator) *Consider n identically distributed independent random variables $X_1, \dots, X_n$ (a sample of size n) where the distribution of X_k contains a parameter θ. The statistic $g(X_1, \dots, X_n)$ is called an unbiased estimator for θ is*

$$E(g(X_1, \dots, X_n)) = \theta \, .$$

Remark:

- Being an unbiased estimator means that "on average" the estimator gives the correct parameter value.

- The previous subsection has shown that the sample mean $\bar{X}$ is an unbiased estimator for the mean $\mu = \mathrm{E}(X_\ell)$. If we consider Bernoulli(p) random variables then the sample proportion $\bar{X}$ is an unbiased estimator for the sample proportion $p = \mathrm{E}(X_\ell)$.

Sample variance: Consider a sample of size n, i.e., n identically distributed independent random variables $X_1, \ldots, X_n$ with $\mathrm{Var}(X_\ell) = \sigma^2$. The naive way to estimate the variance is the statistic

$$S_P^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \left(\bar{X}\right)^2$$

This estimator is, however, biased.

Lemma 5.1 *Denote by $X_1, \ldots, X_n$ n identically distributed, independent random variables with $E(X_\ell) = \mu$ and $Var(X_\ell) = \sigma^2$. Then*

$$E(S_P^2) = \frac{n-1}{n}\sigma^2 \,.$$

That means S_P^2 is biased and it slightly underestimates the variance.

Proof: Using linearity of expectation

$$\mathrm{E}\left(S_P^2\right) = \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}(X_k^2) - \mathrm{E}(\bar{X}^2)$$

Using the definition of variance we have

$$\mathrm{E}(X_k^2) = \mathrm{Var}(X_k) + (\mathrm{E}(X_k))^2 = \sigma^2 + \mu^2$$

and (using proposition 5.1 as well)

$$\mathrm{E}(\bar{X}^2) = \mathrm{Var}(\bar{X}) + (E(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Therefore

$$\mathrm{E}\left(S_P^2\right) = \frac{1}{n}n(\sigma^2+\mu^2) - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2$$

$\square$.

To obtain an unbiased estimator we multiply S_P^2 with $n/(n-1)$ resulting in the sample variance (see section 4**??**)

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2 .$$

c) Mean square error

Consider a sample of size n, $(X_1, \ldots, X_n)$. The sample mean is an unbiased estimator for the mean μ, but there are other unbiased estimators. For instance X_1 (just the first data entry, ignoring all the other $n-1$ entries) is an unbiased estimator since by definition $\mathrm{E}(X_1) = \mu$. Intuitively the sample mean is the better estimator, but why?

The size of the variance of the estimator tells us by how much a typical value differs from the mean. By definition $\mathrm{Var}(X_1) = \sigma^2$, but by proposition 5.1 $\mathrm{Var}(\bar{X}) = \sigma^2/n$, i.e., the variance decreases with the sample size and becomes small for large samples. Hence the estimate becomes more accurate. Therefore the sample mean is the much better estimator.

Definition 5.3 (Mean square eror) *Consider n independent identically distributed random variables $X_1, \ldots, X_n$ (a sample of size n). Denote by θ parameter of the distribution of X_k, and by $g(X_1, \ldots, X_n)$ a statistic which is an estimator for the parameter θ. The* mean square error *(mse) of the estimator is defined by*

$$MSE_g = E((g(X_1, \ldots, X_n) - \theta)^2) .$$

Remark: If the estimator is unbiased, $\theta = \mathrm{E}(g(X_1, \ldots, X_n))$, the mean square error is given by

$$MSE_g = \mathrm{E}((g(X_1, \ldots, X_n) - \mathrm{E}(g(X_1, \ldots, X_n)))^2)) = \mathrm{Var}(g(X_1, \ldots, X_n)) .$$

d) Simulation of random variables

We have discussed various pmfs and pdfs in sections 1b), 2??, 2??, for instance the Binomial distribution. One may consider such a random variable as a hypothetical (or theoretical) population. Any n numbers $(x_1, \ldots, x_n)$ drawn from such a distribution can be considered as a sample of size n. These so called pseudorandom numbers can be generated numerically (but the actual algorithms are way beyond first year UG mathematics). Numerically generated finite samples can be used to model systems or to check and confirm theoretical results.

Example 5.6 *A bag contains* $n = 50$ *balls 20 red balls and* $m = 30$ *blue balls. You pick* $\ell = 15$ *balls at random, see example 1.8, and we count the number of blue balls we pick. Let* X *denote the number of blue balls you pick. The random variable obeys the hypergeometric distribution* $X \sim Hg(50, 30, 15)$*. If we perform the experiment repeatedly, say* $n = 200$ *we obtain a data set of 200 numbers* $(x_1, \ldots, x_n)$*. Instead of doing the actual 200 experiments we can generate such a data set in R using the command* `rhyper`

```
> x<-rhyper(200,30,20,15)
> x[10:20]
 [1]  9  8 10 10 13 13 10  8  9  9  6
```

(the second command `x[10:20]` *just displays the part from* x_{10} *to* x_{20}*). We can compute the sample mean and the sample variance*

```
> mean(x); var(x)
[1] 9.085
[1] 2.701281
```

The values slightly differ from the theoretical prediction, see section 1b), $E(X) = \ell m/n = 9$ *and* $Var(X) = 18/7$ *as we base our computation on a particular finite sample. We may as well plot the relative frequencies of the data set*

```
> hist(x,freq=FALSE,breaks=seq(-0.5,15.5,1))
```

The option `freq=FALSE` *tells R to compute the relative frequencies, and the sequence* `breaks=seq(-0.5,14.5,1)` *specifies the grouping of the data. We may as well add the exact theoretical result by points*

```
> lines(0:15,dhyper(0:15,30,20,15),type="p")
```

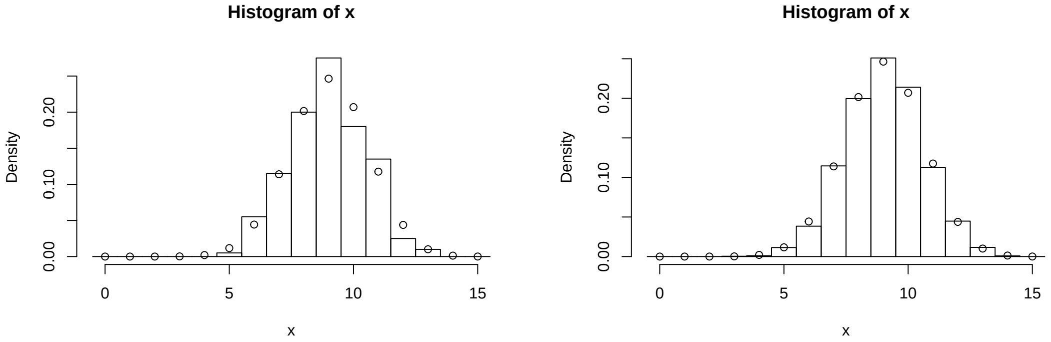


Figure 5.1: Relative frequencies as a histogram obtained from a simulation of the Hypergeometric distribution Hg$(50, 30, 15)$. The symbols are the exact values of the hypergeometric distribution. Left: Sample size 200, right: sample size 5000.

Agreement becomes better for larger sample size (see figure 5.1).

Continuous random variables may be treated the same way.

Example 5.7 *Assume the waiting times (in units of minutes) at a bus stop are exponentially distributed with decay rate* $\lambda = 0.4$*,* $T \sim \mathit{Exp}(0.4)$*. We can simulate a sample of 500 waiting times*

```
> x<-rexp(500,0.4)
> x[120:130]
 [1] 2.33918482 0.22072580 0.09658925 0.97787379
 [5] 2.66325354 0.49060909 2.14893974 5.94831907
 [9] 1.48453527 3.74587232 1.11054274
```

Sample mean and sample variance of the data set are

```
> mean(x);var(x)
[1] 2.354746
[1] 6.160399
```

and they agree reasonable well with the exact theoretical results $E(T) = 1/\lambda = 2.5$ *and* $Var(T) = 1/\lambda^2 = 6.25$. *The histogram of the dataset and the comparison with the theoretical curve are shown in figure 5.2 (We chose the grouping of the histogram sensibly, not too small, not too large, here a bin size of 0.5. We also chose the resolution for the line reasonable fine, here 0.1).*

```
> hist(x,freq=FALSE,breaks=seq(0,25,0.5))
> lines(seq(0,25,0.1),dexp(seq(0,25,0.1),0.4),type="l")
```

We can even perform rather complex tasks, e.g., simulating the sample distribution of the sample mean.

Example 5.8 *Consider a random variable which obeys the Poisson distribution with parameter* $\lambda = 4$, $X \sim Poisson(\lambda)$. *Consider samples of size* $n = 50$. *A single sample* $(x_1, \ldots, x_n)$ *and its sample mean* $\bar{x}$ *can be easily computed*

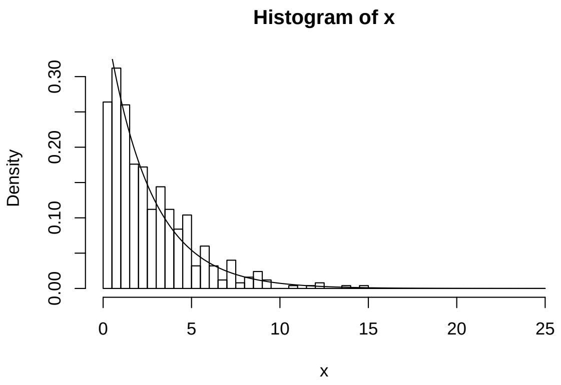


Figure 5.2: Relative frequencies as a histogram obtained from a simulation of the (continuous) exponential distribution Exp(0.4) with sample size 500. The line is the exact result of the exponential distribution.

```
> x<-rpois(50,lambda=4)
> mean(x)
[1] 3.98
```

We can compute the sample mean in one line as well

```
> mean(rpois(50,lambda=4))
[1] 3.64
```

Note that the numerical value differs as we have generated a different sample, the sample mean is a random variable! To plot the bar chart of the sample mean we need to generate a sufficiently large data set of sample means (i.e. we have to execute the above command repeatedly). That can be done with the R command `replicate` *which performs the command repeatedly, say 3 times*

```
> replicate(3,mean(rpois(n=50,lambda=4)))
[1] 3.76 4.03 4.34
```

The result are three sample means in list format. Lets us compute 1000 sample means and then produce a histogram

```
> xbar<-replicate(1000,mean(rpois(n=50,lambda=4)))
> hist(xbar,freq=FALSE,seq(3,5,0.1))
```

Proposition 5.1 tells us that the sample mean obeys approximately a normal distribution with mean $\mu_{\bar{X}} = E(X) = \lambda = 4$ *and variance* $\sigma^2_{\bar{X}} = \lambda/n = 2/25$. *We can add the graph of the pdf to our plot (with reasonably high resolution), see figure 5.3.*

```
> lines(seq(3,5,0.02),dnorm(seq(3,5,0.02),mean=4,sd=sqrt(2/25)))
```

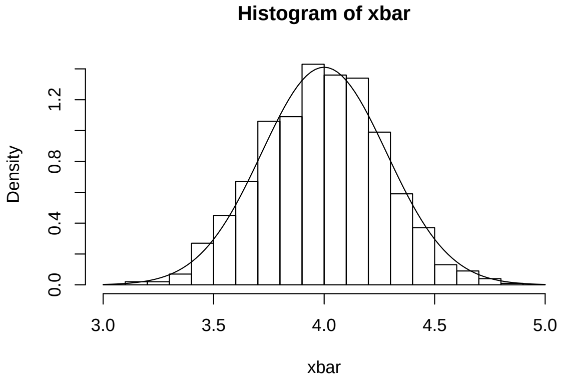


Figure 5.3: Histogram of the sample mean of $n = 50$ Poisson random variables with $\lambda = 4$. The histogram has been obtained from a data set of 1000 samples. The line is the normal approximation for the sampling distribution with $\mu_{\bar{X}} = \lambda = 4$ and $\sigma^2_{\bar{X}} = \lambda/n$.

As expected the sample mean is approximately a normal random variable.

Progress Check:

1. Explain the difference between a population and a sample.
2. What is meant by a statistic. Which statistic do you use for estimating the mean, and which do you use for estimating a proportion.
3. What is meant by a sampling distribution. Describe the circumstances under which the shape of the sampling distribution is approximately normal.
4. What is meant by an unbiased estimator?
5. Which statistic do you use for estimating the variance?
6. What does the mean square error tell you?
7. You toss a fair coin repeatedly until you have seen three heads (and then you stop tossing). Denote by X the number of tails you have seen. Which R command do you use to generate 20 outcomes of this experiment, i.e., a sample of size 20?

§6 Confidence Intervals

So far we have developed unbiased estimators for the mean, the variance, and the proportion of a population. We have as well studied their statistical properties in terms of the sampling distribution. How accurate are the numerical values obtained from the estimators (e.g. how well does the sample mean of a sample reflect the true mean of a population), i.e., how accurate are the point estimates. We will employ our knowledge of the sampling distribution to tackle this problem.

Example 6.1 *You ask a random sample of ten students for their 2018 final exam mark in MTH4107 Introduction to Probability: 63, 43, 41, 59, 49, 54, 74, 17, 67, 92. The sample mean is*

```
> x<-c(63, 43, 41, 59, 49, 54, 74, 17, 67, 92)
> mean(x)
[1] 55.9
```

The population variance of the final exam mark is 248.58. What can we say about the true expected mark in the module (i.e. the population mean)?

a) Interval estimate for the sample mean

Consider a sample of size n, $(X_1, \ldots, X_n)$, where X_k denote independent identically distributed random variables with expectation μ and variance σ^2. Assume for simplicity that we know the variance σ. The sample mean $\bar{X}$ is approximately a normal random variable with expectation $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$ for suitable sample size. Given a (small) probability α (say $\alpha = 0.05$) we want to find an interval $[\mu-\delta, \mu+\delta]$ around the expectation (i.e. we want to find δ) so that with (large) probability $1-\alpha$ the sample mean $\bar{x}$ is contained in that interval. That means the sample mean differs from the expectation by more than δ with (small) probability α. The threshold δ is determined by

$$\mathbb{P}(|\bar{X}-\mu|<\delta)=1-\alpha \quad \Rightarrow \quad \mathbb{P}(|\bar{X}-\mu|>\delta)=\alpha\,.$$

Using standardisation

$$\mathbb{P}\left(\frac{|\bar{X}-\mu|}{\sigma_{\bar{X}}}>\frac{\delta}{\sigma_{\bar{X}}}\right)=\mathbb{P}\left(|Z|>\frac{\delta}{\sigma_{\bar{X}}}\right)=\alpha\,.$$

The z-score, Lemma 2.1, tells us $\alpha=\mathbb{P}(|Z|>z_{\alpha/2})$, i.e.,

$$\frac{\delta}{\sigma_{\bar{X}}}=z_{\alpha/2}=\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \quad \Rightarrow \quad \delta=\sigma_{\bar{X}}z_{\alpha/2}\,.$$

Hence, the interval $[\mu-\sigma_{\bar{X}}z_{\alpha/2},\mu+\sigma_{\bar{X}}z_{\alpha/2}]$ contains the sample mean of the sample with probability $1-\alpha$.

In return that means a sample mean $\bar{x}$ results in so called *confidence interval*

$$[\bar{x}-\sigma_{\bar{X}}z_{\alpha/2},\bar{x}+\sigma_{\bar{X}}z_{\alpha/2}]=[\bar{x}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\bar{x}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}]$$

The fraction of $1-\alpha$ of such intervals contain the population mean μ. We call $1-\alpha$ the *level of confidence.* A level of confidence $1-\alpha$ means that the fraction $1-\alpha$ of all possible samples result in confidence intervals that contain the parameter μ.

Note: Confidence intervals do not mean that the parameter μ is contained in the interval with probability $1-\alpha$. It is the statement that the fraction $1-\alpha$ of all confidence intervals contain the parameter, hence it is a statement about the sample statistic.

Remark: In applications it is sensible to check the independence and the normality conditions, since we have based the theoretical analysis on the normal approximation of the sampling distribution.

Example 6.2 *Given the data from example 6.1 construct a* 95% *confidence interval for the expectation.*

Since $\sigma^2=248.58$ *and* $n=10$ *we have* $\sigma_{\bar{X}}^2=\sigma^2/n=24.858$.

For confidence level $1-\alpha=0.95$ *the z-score is* $z_{\alpha/2}=z_{0.025}=\Phi^{-1}(0.975)$

```
> qnorm(0.975)
[1] 1.959964
```

With $\bar{x} = 55.9$ *the confidence interval at confidence level* $1 - \alpha = 0.95$ *reads*

$$[\bar{x} - \sigma_{\bar{X}} z_{0.025}, \bar{x} + \sigma_{\bar{X}} z_{0.025}]$$

which gives

```
> barx<-55.9
> sigma<-sqrt(248.58)
> n<-10
> sigmaX<-sigma/sqrt(n)
> alpha<-0.05
> zsc<-qnorm(1-alpha/2)
> barx-sigmaX*zsc; barx+sigmaX*zsc
[1] 46.12805
[1] 65.67195
```

Hence assuming that our sample does not belong to the 5% ($\alpha = 0.05$*) of exceptional samples the correct mean mark of the module is contained in the interval* $[46, 66]$.

Remark: For simplicity we have assumed that the variance $\sigma^2 = \mathrm{Var}(X_k)$ is known. If σ^2 is a priori unknown we may take the sample variance s^2 as a proxy if the sample is sufficiently large. One should keep in mind that a more sophisticated theory would be required (beyond first year UG mathematics) to deal with this case in a consistent way.

Example 6.3 *Consider the marks sample of example 6.1. If the variance* σ^2 *would not be known we could use the sample variance* s^2 *instead, to compute a confidence interval*

```
> x<-c(63, 43, 41, 59, 49, 54, 74, 17, 67, 92)
> var(x)
[1] 418.5444
```

Using the approximation $\sigma^2_{\bar{X}} = s^2/n$ *the confidence interval at confidence level* $1-\alpha = 0.95$ *is then given by*

$$[\bar{x} - \sigma_{\bar{X}} z_{0.025}, \bar{x} + \sigma_{\bar{X}} z_{0.025}]$$

which can be easily evaluated in R

```
> x<-c(63, 43, 41, 59, 49, 54, 74, 17, 67, 92)
> n<-length(x)
> barx<-mean(x)
> sigmaX<-sqrt(var(x)/n)
> alpha<-0.05
> zsc<-qnorm(1-alpha/2)
> barx-sigmaX*zsc; barx+sigmaX*zsc
[1] 43.22001
[1] 68.57999
```

The 95% confidence interval differs noticeably from the previous calculation. The sample size is not large enough to justify the normality condition.

b) Interval estimate for the sample proportion

Consider a sample of size n, $(X_1, \ldots, X_n)$. If X_k are Bernoulli(p) random variables which signal membership of the individual to a group (i.e., $X_k = 1$ if individual k belongs to the group and $X_k = 0$ otherwise) then the proportion mean $\bar{X}$ is an unbiased estimator for the population proportion p. Hence we can use the same confidence interval for the sample proportion as well. Here we have $\sigma^2 = \mathrm{Var}(X_k) = p(1-p)$ and the variance of the sample proportion is $\sigma_{\bar{X}}^2 = \sigma^2/n = p(1-p)/n$. Therefore the confidence interval at confidence level $1-\alpha$ is given by

$$[\bar{x} - \sigma_{\bar{X}} z_{\alpha/2}, \bar{x} + \sigma_{\bar{X}} z_{\alpha/2}] = [\bar{x} - \sqrt{\frac{p(1-p)}{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \bar{x} + \sqrt{\frac{p(1-p)}{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)].$$

Since the exact proportion p is unknown we may replace p by the estimate $\bar{x}$ to evaluate the sample variance and we arrive at

$$[\bar{x} - \sigma_{\bar{X}} z_{\alpha/2}, \bar{x} + \sigma_{\bar{X}} z_{\alpha/2}] = \left[\bar{x} - \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} z_{\alpha/2}, \bar{x} + \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} z_{\alpha/2}\right]$$

for the confidence interval of the sample proportion at confidence level $1-\alpha$, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

Example 6.4 *You ask a random sample of 15 students for their marks in the 2017 final exam in Introduction to Probability:*

79,75,70,26,88,32,83,69,46,50,69,37,72,76,49

To compute the proportion of students getting an A grade convert the data set where 1/0 indicates an A grade/lower grade

1,1,1,0,1,0,1,0,0,0,0,0,1,1,0

The sample proportion estimates the proportion of A grades in this exam, i.e., the point estimate is given by

```
> x<-c(1,1,1,0,1,0,1,0,0,0,0,0,1,1,0)
> barx<-mean(x)
> barx
[1] 0.4666667
```

To construct a 95% confidence interval we need to estimate the variance σ^2 *to compute the sample variance* $\sigma_{\bar{X}}^2 = \sigma^2/n$. *Since we deal with Bernoulli(*p*) random variables the variance is Var(*X_k*)* $= p(1-p)$, *where we replace* p *by its estimate* $\bar{x}$

```
> sigma2<-barx*(1-barx)
> sigma2
[1] 0.2488889
```

The sample variance s^2 *would give a very similar estimate*

```
> var(x)
[1] 0.2666667
```

The variance of the sample mean $\sigma^2_{\bar{X}} = \sigma^2/n$ *then gives*

```
> sigma2X<-sigma2/length(x)
> sigma2X
[1] 0.01659259
```

For a 95% confidence interval we have confidence level $1 - \alpha = 0.95$*, i.e.,* $\alpha = 0.05$ *and the z-score* $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ *is given by*

```
> zsc<-qnorm(1-0.05/2)
> zsc
[1] 1.959964
```

Finally the 95% confidence interval $[\bar{x} - \sigma_{\bar{X}} z_{0.025}, \bar{x} + \sigma_{\bar{X}} z_{0.025}]$ *reads*

```
> barx-sqrt(sigma2X)*zsc; barx+sqrt(sigma2X)*zsc
[1] 0.2141993
[1] 0.719134
```

That means unless our sample of grades is among 5% of exceptional samples the actual proportion of A grades is contained in the interval $[0.214, 0.719]$.

To compute a 99% confidence interval we can keep the estimate $\bar{x}$ *and the sample variance* $\sigma^2_{\bar{X}}$. *We just need to adjust the z-score* $z_{\alpha/2}$ *since now* $1 - \alpha = 0.99$ *and* $\alpha = 0.01$

```
> zsc<-qnorm(1-0.01/2)
> zsc
[1] 2.575829
> barx-sqrt(sigma2X)*zsc; barx+sqrt(sigma2X)*zsc
[1] 0.1348683
[1] 0.798465
```

The 99% confidence interval $[0.135, 0.798]$ *is larger. Higher confidence requires that a larger number of samples result in intervals which contain the actual population proportion, i.e., higher confidence requires a larger confidence interval.*

Which sample size n *would be needed to obtain a 95% confidence interval* $[\bar{x} - \delta, \bar{x} + \delta]$ *of half width* $\delta = 0.1$*? Since the half width of the confidence interval is given by* $\sigma^2_{\bar{X}} z_{\alpha/2}$ *with confidence level* $1 - \alpha = 0.95$, *i.e.,* $\alpha = 0.05$ *the condition on the sample size reads*

$$\delta = \sigma^2_{\bar{X}} z_{\alpha/2} = \sqrt{\frac{p(1-p)}{n}} z_{\alpha/2}$$

which gives

$$n = \frac{p(1-p) z_{\alpha/2}}{\delta^2}$$

with $\alpha = 0.05$ *and* $\delta = 0.1$. *For (the unknown) value* p *we could use the estimate* $\bar{x}$. *Alternatively* $p(1-p) \leq 1/4$ *and we could use this value as an upper bound. Hence*

$$n \leq \frac{1}{4} \frac{\Phi^{-1}(1 - \alpha/2)}{\delta^2}$$

which gives

```
> 1/4*qnorm(1-0.05/2)/0.1^2
[1] 48.9991
```

Hence one needs the grades of 49 randomly sampled students to guarantee the 95% confidence interval $[0.467 - 0.1, 0.467 + 0.1]$ *for the proportion of students scoring an A grade. However: recall that our approach was based on the assumption of independence, i.e., that the sample size is small compared to the population size (i.e. the number of students sitting the exam). If such a condition is not valid, i.e., if independence cannot be guaranteed then one needs more sophisticated approaches.*

Progress Check:

1. Explain in general terms the difference between a point estimate and an interval estimate.

2. Which quantities do you need to know/evaluate to compute a 95% confidence interval for the sample mean?

3. What does a 95% level of confidence mean?

4. What does a 99% confidence interval for the sample proportion tell you?

5. How does the size of a confidence interval depend on the sample size?

6. Given a sample you compute a 95% and a 99% confidence interval for the sample proportion. Which interval is smaller? Explain why the interval is smaller.

7. Which conditions does a sample have to satisfy so that you can compute a confidence interval for the sample mean?

§7 Hypothesis Testing

a) Basic notions

Example 7.1 *We want to figure out whether a coin is fair. We toss the coin repeatedly, say 10000 times, and we count the number of heads seen. If this number is close to 5000 the coin has a good chance to be fair. Tossing a coin is captured by a Bernoulli(p) random variable (say, $X = 1$ meaning head, and $X = 0$ meaning tail). The statement the coin is fair is a statement about the bias p, i.e., about a parameter of the probability mass function, here $p = 1/2$. Such a statement is called a* hypothesis

Definition 7.1 (Hypothesis) *A* hypothesis *is a statement about a parameter θ of the pmf (or pdf) of a random variable X.*

Example 7.2 *By tossing the coin many times, say 10000 times, we will not be able to confirm whether the coin is fair ($p = 0.5$) just by counting the number of heads. We won't be able to exclude a tiny bias, say $p = 0.50000000001$ (e.g. adding a single atom to one side of the coin). Hence hypothesis testing does not mean to confirm the hypothesis, it means finding evidence against the hypothesis. If we find evidence against the hypothesis we reject ("nullify") the hypothesis. If we cannot find evidence we do not reject the hypothesis (and we may be tempted to accept the hypothesis for the time being).*

Definition 7.2 (Null hypothesis) *In hypothesis testing the central claim, the* null hypothesis *H_0 is a statement $\theta = \theta_0$ which we intend to find evidence against (to "nullify the hypothesis").*

Example 7.3 *We toss a coin. Our null hypothesis H_0 is to have a fair coin, $p = 1/2$. Our test procedure consists in tossing the coin 10000 times and counting the number of heads seen. If this number is close to 5000, say $X \in [4900, 5100]$ we do not reject H_0, otherwise we reject H_0. Denote by X the number of heads seen $X \sim Bin(10000, p)$. If we assume H_0 to be true then $X \sim Bin(10000, 1/2)$ and the probability that the test rejects*

H_0 *is*

$$\mathbb{P}(X < 4900 \textit{ or } X > 5100) = \mathbb{P}(X \leq 4899) + \mathbb{P}(X \geq 5101) = 1 - \mathbb{P}(X \leq 5100) + \mathbb{P}(X \leq 4899)$$

The numerical value can be easily evaluated in terms of the cdf of the Binomial$(10000, 1/2)$ *distribution*

```
> 1-pbinom(5100,size=10000,prob=1/2)+pbinom(4899,size=10000,prob=1/2)
[1] 0.0444258
```

That means only with a small probability of 4.4% the test rejects the null hypothesis assuming the hypothesis is actually true. Such an error is called type-I error and the corresponding probability α *is called the significance level.*

Definition 7.3 (Significance level) *Assume the null hypothesis* H_0 *is valid. We say a type-I error has occurred if the test procedure for* H_0 *rejects the null hypothesis. The probability of a type-I error occurring is called the* significance level α *of the test procedure.*

In general we test a null hypothesis against an alternative hypothesis, e.g., in example 7.3 $p = 1/2$ (the coin being fair) against $p \neq 1/2$ (the coin being biased).

Definition 7.4 (Alternative hypothesis) *The test procedure tests the null hypothesis* H_0 *against the so called* alternative hypothesis H_1. *The alternative hypothesis specifies under which conditions the null hypothesis should be rejected.*

The properties of the alternative hypothesis determine the type of test procedure. In example 7.3 with alternative hypothesis $p \neq 1/2$ we used a test to reject H_0 if $X > 5100$ or $X < 4900$ a so called two-sided test.

Example 7.4 *We toss a coin. We want to test the null hypothesis* H_0*, the coin being fair (*$p = 1/2$*) against the alternative hypothesis* H_1 *that the coin shows more tails than* *heads ($p < 1/2$). We use the test procedure: toss the coin 10000 times and reject H_0 if we see less than 4900 heads, i.e. we reject H_0 if $X < 4900$ (and we do not reject H_0 if $X \geq 4900$). This is a so called* one-sided test *(left-tailed test).*

To compute the significance level α assume H_0 is valid. The probability that the test rejects H_0 is given by

$$\alpha = \mathbb{P}(X < 4900) = \mathbb{P}(X \leq 4899)$$

and the value can be easily evaluated with the cdf of the Binomial$(10000, 1/2)$ distribution

```
> pbinom(4899,size=10000,prob=1/2)
[1] 0.0222129
```

That means if the number of heads seen is less than 4900, $X < 4900$, we reject H_0 (against H_1) with significance level 2.2%. If $X \geq 4900$ we do not reject H_0 (no matter how many tails we see !).

Remark:

- If we test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ we need a two-sided test.
- If we test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ we need a one-sided test (normally: a right-tailed test).
- If we test $H_0 : \theta = \theta_0$ again-ts $H_1 : \theta < \theta_0$ we need a one-sided test (normally: a left-tailed test).

Definition 7.3 states that a type-I error occurs if the test rejects the null hypothesis H_0 even though H_0 is valid. The complementary situation can occur as well, i.e., the test does not reject H_0 even though in some sense H_0 is false. That is called a type-II error.

Example 7.5 *Assume we use the two-sided test of example 7.3 to test our null hypothesis (i.e. to find evidence against H_0). Assume that $p = 0.48$ (i.e. H_0 is invalid). Then*

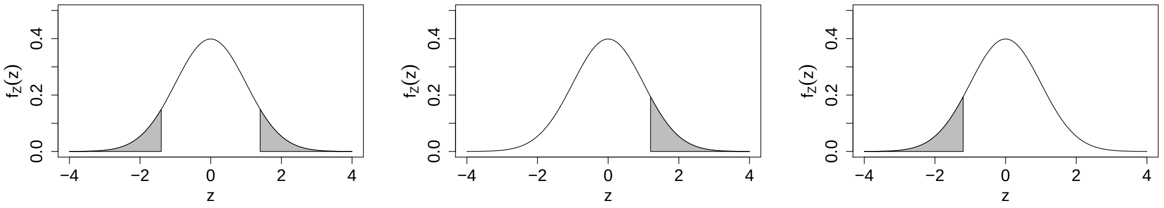


Figure 7.1: Illustration of a two-sided test (left), right-tailed test (middle), and left-tailed test (right). The graph depicts the pdf of the test statistic assuming that the null hypothesis is valid. The shaded regions indicate the rejection region with the size of the shaded area being the significance level α.

$X \sim Bin(10000, 0.48)$ and any outcome $X \in [4900, 5100]$ produces a type-II error, i.e. the test does not reject H_0. The probability for a type-II error is (assuming $p = 0.48$)

$$\mathbb{P}(4900 \le X \le 5100) = \mathbb{P}(X \le 5100) - \mathbb{P}(X \le 4899)$$

and can be easily evaluated in terms of the cdf of the $Binomial(10000, 0.48)$ distribution

```
> pbinom(5100,size=10000,prob=0.48)-pbinom(4899,size=10000,prob=0.48)
[1] 0.02322684
```

Hence there is a 2.3% chance for not rejecting a false hypothesis H_0 if $p = 0.48$.

Definition 7.5 (Power) *Assume the null hypothesis H_0 ($\theta = \theta_0$) is not valid and the parameter takes a value $\theta = \theta_1 \neq \theta_0$. We say that a type-II error has occurred if the test for H_0 does not reject H_0. The probability for a type-II error is denoted by β. $1 - \beta$ is called the power of the test (as with regards to the alternative $\theta = \theta_1$).*

In a nutshell type-I and and type-II errors are summarised in table 1.

Example 7.6 *Consider the setup of example 7.3. We change the test procedure: we toss the coin 10000 times and we do not reject H_0 if $X \in [4950, 5050]$.*

	H_0 is valid	H_0 is (somehow) invalid
Test does not reject H_0	correct conclusion	type-II error
Test rejects H_0	type-I error	correct conclusion

Table 1: Summary of the meaning of type I and type II errors.

The significance level of this test is (assuming H_0 is valid, i.e., $p = 1/2$)

$$\begin{aligned} \alpha &= \mathbb{P}(X < 4950 \textit{ or } X > 5050) = \mathbb{P}(X \leq 4949) + \mathbb{P}(X \geq 5051) \\ &= 1 - \mathbb{P}(X \leq 5051) + \mathbb{P}(X \leq 4949)\,. \end{aligned}$$

```
> 1-pbinom(5050,size=10000,prob=1/2)+pbinom(4949,size=10000,prob=1/2)
[1] 0.3124952
```

The significance level has increased $\alpha = 0.31$, but that means a higher probability for type-I errors (rejecting a valid null hypothesis) of 31.2%.

The probability for type-II errors as with regards to the alternative $p = 0.48$ (see example 7.5) is (assuming $p = 0.48$)

$$\beta = \mathbb{P}(4950 \leq X \leq 5050) = \mathbb{P}(X \leq 5050) - \mathbb{P}(X \leq 4949)$$

which evaluates as

```
> pbinom(5050,size=10000,prob=0.48)-pbinom(4949,size=10000,prob=0.48)
[1] 0.001387635
```

Hence the probability of type-II errors is tiny 0.14% and the power $1 - \beta = 0.9986\ldots$ has increased (as compared to 0.9767... in example 7.5). The test with higher power is better in “identifying” biased coins (only very few biased coins are not rejected by the test) but the probability of a type I error is massive (i.e. many fair coins will be rejected by the test with higher power).

Remark:

- Typically increasing α decreases β. That means increasing the significance level (increasing the likelihood of type-I errors) decreases the likelihood of type-II errors (increases the power of the test).
- Similarly decreasing α typically increases β. That means decreasing the significance level (decreasing the likelihood of type-I errors) increases the likelihood of type-II errors (decreases the power of the test).

But given this notation mumbo jumbo how do we design suitable test procedures?

b) Test for a population proportion

Example 7.7 *QMUL suggest a percentage of (at least)* 35% *A grades in typical modules. A randomly selected sample of 15 students who took the 2017 final exam in Introduction to Probability (see example 6.4 for a different sample) shows 3 A grades and 12 lower grades, that means only 20% of students in the sample have an A grade. Does that mean there is enough evidence to state that the final exam did not meet the required 35% of A grades?*

We sample from a population containing two types of objects/individuals, and we want to test the proportion p of those individuals. The sample of size n, $(X_1, \ldots, X_n)$, may be considered as a collection of independent identically distributed Bernoulli(p) random variables. We want to design a test for the null hypothesis H_0 that the proportion p has the value $p = p_0$.

We base the test on the estimator $\bar{X}$. Section 5**??** and proposition 5.1 tells us that $\bar{X}$ is an (approximate) normal random variable with mean $\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = p$ and variance $\sigma^2_{\bar{X}} = \mathrm{Var}(\bar{X}) = \mathrm{Var}(X_k)/n = p(1-p)/n$ if the sample has a suitable size (see the independence and the normality condition). Assume that the null hypothesis $p = p_0$ is valid. Then standardisation tells us that the statistic

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

is a standard normal random variable $Z \sim N(0,1)$. We reject the null hypothesis if $\bar{X}$ substantially deviates from p_0, i.e., if Z exceeds a threshold δ. The threshold will be determined by the significance level of the test.

Two-sided test: We test our null hypothesis $H_0 : p = p_0$ against the alternative $H_1 : p \neq p_0$. We reject the null hypothesis if $|Z| > \delta$. The significance level (the probability of a type-I error, i.e., the probability of rejection assuming H_0 is valid) is given by

$$\alpha = \mathbb{P}(|Z| > \delta) \, .$$

Using lemma 2.1 we have

$$\delta = z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) \, .$$

We reject the null hypothesis $p = p_0$ (in favour of the alternative $p \neq p_0$) at significance level α if the sample mean obeys $|Z| > \delta$, i.e., if

$$\left| \frac{\bar{x} - p_0}{\sqrt{p_0(1 - p_0)/n}} \right| > z_{\alpha/2} \, .$$

Example 7.8 *In 1994, 52% of parents of children in high school felt it was a serious problem that high school students were not being taught enough maths and science. A recent survey found that 256 of 800 parents of children in high school felt it was a serious problem that high school students were not being taught enough math and science. Do parents feel differently today than they did in 1994, if we apply a significance level of* $\alpha = 0.05$

Denote by p the proportion of parents who think there is not enough math and science taught at high school today. Our null hypothesis H_0 *is: the proportion of parents has not changed since 1994, i.e.* $p = p_0 = 0.52$*. The survey gives the estimate* $\bar{X} = 256/800 = 0.32$*. Does that indicate a change at significance level* $\alpha = 0.05$*, i.e., should we reject the null hypothesis? Our alternative hypothesis in this case reads* $p \neq p_0$*.*

Here our sample is small compared to the whole population and $np_0(1 - p_0) > 10$*, that means, our test conditions apply (i.e. the normality and the independence condition are* *satisfied, see section 5??). We check the inequality*

$$\left|\frac{\bar{x}-p_0}{\sqrt{p_0(1-p_0)/n}}\right| > z_{\alpha/2}$$

with $p_0 = 0.52$, $n = 800$, $\alpha = 0.05$, *and* $\bar{x} = 256/800$ *using R.*

Left-hand side: The standardised proportion mean is given by

```
> p0<-0.52
> xbar<-256/800
> n<-800
> (xbar-p0)/sqrt(p0*(1-p0)/n)
[1] -11.32277
```

Right-hand side: the z*-score reads*

```
> alpha<-0.05
> qnorm(1-alpha/2)
[1] 1.959964
```

Hence the inequality is valid and there is sufficient evidence to reject the null hypothesis at 5% significance level, i.e., the parents' perception has changed.

Left-tailed test: We test our null hypothesis $H_0 : p = p_0$ against the alternative $H_1 : p < p_0$. We reject the null hypothesis if Z is too small, $Z < -\delta$ (with $\delta > 0$). The significance level (the probability of a type-I error, i.e., the probability of rejection assuming H_0 is valid) is given by

$$\alpha = \mathbb{P}(Z < -\delta)\,.$$

Using lemma 2.1 we have

$$\delta = z_\alpha = \Phi^{-1}(1-\alpha)\,.$$

We reject the null hypothesis $p = p_0$ (in favour of the alternative $p < p_0$) at significance level α if the sample proportion obeys

$$\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha \, .$$

Example 7.9 *Consider the setting of example 7.7. QMUL suggest a percentage of (at least) 35% A grades in typical modules. Consider the 2017 final exam in Introduction to Probability (see example 6.4). A randomly selected sample of 15 students shows 3 A grades. Using this sample is there enough evidence to state that the final exam does not comply with the college suggestion of 35% A grades at a 10% significance level?*

Our sample has size $n = 15$ *with 3 A-grades, i.e.,* $\bar{x} = 3/15 = 0.2$. *We use the null hypothesis of a proportion of* $p = p_0 = 0.35$ *A-grades, with alternative hypothesis* $p < p_0$ *(as overperformance in modules is not an issue). To reject the null hypothesis we need to confirm the inequality*

$$\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha$$

with $n = 15$, $\bar{x} = 3/15$, $p_0 = 0.35$, *and* $\alpha = 0.1$.

Left-hand side: The standardised proportion mean is given by

```
> p0<-0.35
> xbar<-3/15
> n<-15
> (xbar-p0)/sqrt(p0*(1-p0)/n)
[1] -1.217997
```

Right-hand side: the (negative) z-score reads

```
> alpha<-0.1
> -qnorm(1-alpha)
[1] -1.281552
```

Hence the inequality is violated and we have not enough evidence to reject the null hypothesis, i.e, there is so far no data based evidence to assume that the final exam did not comply with the 35% A-grade constraint at a 10% significance level.

One has to consider that the sample size of $n = 15$ *is fair small and the normal approximation used in the analysis may not be valid. In fact* $np_0(1-p_0) > 10$ *does not hold, so that more sophisticated tools may be needed to perform a solid hypothesis test.*

Right-tailed test: We test our null hypothesis $H_0 : p = p_0$ against the alternative $H_1 : p > p_0$. We reject the null hypothesis if $Z > \delta$. The significance level (the probability of a type-I error, i.e., the probability of rejection assuming H_0 is valid) is given by

$$\alpha = \mathbb{P}(Z > \delta)\,.$$

Using lemma 2.1 we have

$$\delta = \Phi^{-1}(1-\alpha) = z_\alpha\,.$$

We reject the null hypothesis $p = p_0$ (in favour of the alternative $p > p_0$) at significance level α if the sample proportion obeys

$$\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha\,.$$

Example 7.10 *In April 2009, the Gallup organisation surveyed 676 adults aged 18 and older and found that 352 believed they would not have enough money to live comfortably in retirement. Does the sample evidence suggests that a majority of adults believe they will not have enough money in retirement if we use a significance level of 5%?*

The first (and major) problem is the formulation of the null and the alternative hypothesis. Denote by p the fraction of adults who believe they will not have enough money. The hypothesis $p = p_0 = 0.5$ *means that (only) half of the adults believe they have not enough money, while* $p > 0.5$ *means the majority is of that believe. Hence we can state the null hypothesis as* $H_0 : p = p_0 = 0.5$ *and we are looking for evidence against this hypothesis in favour of the alternative* $H_1 : p > p_0 = 0.5$.

We reject H_0 *(in favour of* H_1*) if the inequality*

$$\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha$$

is satisfied, where $p_0 = 0.5$, $n = 676$, $\bar{x} = 352/676$, *and* $\alpha = 0.05$.

Left-hand side: The standardised sample proportion is given by

```
> p0<-0.5
> xbar<-352/676
> n<-676
> (xbar-p0)/sqrt(p0*(1-p0)/n)
[1] 1.076923
```

Right-hand side: the z*-score reads*

```
> alpha<-0.05
> qnorm(1-alpha)
[1] 1.644854
```

Hence the inequality is not valid and there is not enough evidence to reject the hypothesis H_0 *at significance level 5% (i.e. we cannot reject that only a half, or less, of the population believes to have insufficient money), even though* $\bar{x} = 352/676 = 0.520\ldots$.

For this example the independence and the normality condition are both satisfied as the sample size is small compared to the population size, and $np_0(1-p_0) = 676 \cdot 0.5(1-0.5) > 10$.

c) Test for a population mean

Example 7.11 *QMUL modules have a target mean mark of 60. Consider the data of example 6.4 for the 2017 final exam in Introduction to Probability. The sample mean of* *these marks is* $\bar{x} = 61.4$. *Is there sufficient data based evidence to state that students in this module did better as compared to the college guidelines?*

We consider a sample of size n modelled by independent identically distributed random variables $X_1, \ldots, X_n$. Based on the sample mean $\bar{X}$ we want to design a test for the population mean $\mu = \mathrm{E}(X_k)$ with null hypothesis $\mu = \mu_0$. For simplicity we assume that the variance $\sigma^2 = \mathrm{Var}(X_k)$ is known.

Section 5**??** and proposition 5.1 tells us that $\bar{X}$ is a normal random variable with variance $\sigma_{\bar{X}} = \mathrm{Var}(X_k)/n = \sigma^2/n$. Assume that the null hypothesis $\mu = \mu_0$ is valid. Then standardisation tells us that the statistic

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

is a standard normal random variable, $Z \sim N(0,1)$. We reject the null hypothesis if $\bar{X}$ substantially differs from μ_0. Depending on the alternative we apply a two-sided or one-sided test.

Two-sided test: We test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$. We reject the null hypothesis if $\bar{X}$ differs substantially from μ_0, that means $|Z| > \delta$. The threshold δ is related with the significance level α (see above)

$$\alpha = \mathbb{P}(|Z| > \delta) \quad \Rightarrow \quad \delta = z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)\,.$$

We reject the null hypothesis $\mu = \mu_0$ (in favour of the alternative $\mu \neq \mu_0$) at significance level α if the sample mean obeys

$$\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}\,.$$

Example 7.12 *In 2002 the mean age of an inmate on death row in the US was 40.7 years with standard deviation 9.6 years. To check whether the mean age of a death row inmate has changed 32 death row inmates are randomly sampled and one finds a sample mean of 38.9 years. Assume for simplicity that the standard deviation has not changed over time. Does the sample mean support a change of mean age at a 5% significance level?*

We seek evidence against the null hypothesis $H_0 : \mu = \mu_0 = 40.7$ with alternative $H_1 : \mu \neq \mu_0$, i.e., we apply a two-sided test. The sample of size $n = 32$ tells us $\bar{x} = 38.9$ with $\sigma = 9.6$ (assumed to be unaffected). To reject the null hypothesis we need to confirm the inequality

$$\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

with significance level $\alpha = 0.05$.

Left-hand side: The standardised sample mean is given by

```
> mu0<-40.7
> xbar<-38.9
> n<-32
> sigma<-9.6
> (xbar-mu0)/(sigma/sqrt(n))
[1] -1.06066
```

Right-hand side: the z-score reads

```
> alpha<-0.05
> qnorm(1-alpha/2)
[1] 1.959964
```

The inequality is not valid, hence there is not enough evidence to reject the null hypothesis, i.e., the sample does not support any change of the mean since 2002.

The sample size is small compared to the population size (hence we can justify independence), and the sample size seems to be sufficiently large to justify normality.

Left-tailed test: We test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu < \mu_0$. We reject the null hypothesis if Z is too small, $Z < -\delta$ where the threshold δ is determined by the significance level (see previous subsection), $\delta = z_\alpha$.

We reject the null hypothesis $\mu = \mu_0$ (in favour of the alternative $\mu < \mu_0$) at significance level α if the sample mean obeys

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha \,.$$

Example 7.13 *In a study published in the American Journal of Psychiatry (May 2000) researches measured the effect of alcohol on the development of the hippocampal region in adolescents. The researchers randomly selected 12 adolescents with alcohol use disorder. They wanted to determine whether the hippocampal volumes in the alcoholic adolescents were less than the normal volume of 9.02 cm^3. The standard deviation of the volume is known to be $\sigma = 0.7cm^3$. An analysis of the sample data revealed that the hippocampal volume is approximately normal with $\bar{x} = 8.10cm^3$. Do the data support a reduction in volume at a significance level of $\alpha = 0.01$.*

Use the null hypothesis $\mu = \mu_0 = 9.02$ with alternative $\mu < \mu_0$. Apply the left-tailed test with $\bar{x} = 8.10$, $\sigma = 0.7$, $n = 12$ and $\alpha = 0.01$, i.e. check the inequality

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

to reject the null hypothesis.

Left-hand side: The standardised sample mean is given by

```
> mu0<-9.02
> xbar<-8.1
> n<-12
> sigma<-0.7
> (xbar-mu0)/(sigma/sqrt(n))
[1] -4.552819
```

Right-hand side: the z-score reads

```
> alpha<-0.01
> -qnorm(1-alpha)
[1] -2.326348
```

The inequality is valid, hence there is sufficient evidence to reject the null hypothesis $\mu = \mu_0$ *in favour of* $\mu < \mu_0$ *at significance level 1%. Observe however that the sample size* $n = 12$ *is fairly small and the usage of the normal approximation is questionable (see the independence and the normality conditions).*

Right-tailed test: We test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$. We reject the null hypothesis if Z is too large, $Z > \delta$ where the threshold δ is determined by the significance level (see previous subsection), $\delta = z_\alpha$.

We reject the null hypothesis $\mu = \mu_0$ (against the alternative $\mu > \mu_0$) at significance level α if the sample mean obeys

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \,.$$

Example 7.14 *QMUL modules have a target mean mark of 60. Consider the data of example 6.4 for the 2017 final exam in Introduction to Probability. Is there any evidence at 5% significance level that students in this module did better compared to the college target. The standard deviation of marks in this module is* $\sigma = 19.7$*.*

We test the null hypothesis $H_0 : \mu = \mu_0 = 60$ *against the alternative* $H_1 : \mu > \mu_0$*, i.e., we use a right tailed test. We have to check the inequality*

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

with $\mu_0 = 60$*,* $\sigma = 19.7$*,* $n = 15$*,* $\alpha = 0.05$ *and* $\bar{x}$ *computed from the data set of example 6.4*

Left-hand side: The standardised sample mean is given by

```
> x<-c(79,75,70,26,88,32,83,69,46,50,69,37,72,76,49)
> mu0<-60
> xbar<-mean(x)
> n<-15
> sigma<-19.7
> (xbar-mu0)/(sigma/sqrt(n))
[1] 0.2752374
```

Right-hand side: the z-score reads

```
> alpha<-0.05
> qnorm(1-alpha)
[1] 1.644854
```

The inequality is not satisfied, i.e., there is insufficient evidence to reject the null hypothesis $\mu = \mu_0 = 60$ *in favour of the alternative* $\mu > \mu_0$ *at significance level 5%.*

the sample size is at about 7% of the population size (at about 200 students were sitting the exam), i.e., the independence condition is almost satisfied. However the sample size is fairly small to comply with the normality condition, and hence more sophisticated tools may be needed.

d) P-values

P-values are an alternative to perform hypothesis testing. Given an sample of size n, the P-value, in a nutshell, is the probability to observe an outcome as extreme as the current observation, assuming the null hypothesis is true. Hence low P-values mean outcomes as extreme as our current samples are unlikely and we reject the null hypothesis.

Definition 7.6 (P-value) *Given an observation of a sample of size n and a null hypothesis H_0, the P*-value *of the observation is the probability to observe a sample statistic as extreme or more extreme as the observation under the assumption that the null hypothesis is true.*

At significance level α the null hypothesis is rejected if the P-value obeys $P < \alpha$.

Remark: Guideline for the interpretation of P-values

- $P > 0.1$: No evidence against the null hypothesis
- $0.05 < P < 0.1$: Weak evidence against the null hypothesis
- $0.01 < P < 0.05$: Moderate evidence against the null hypothesis
- $0.001 < P < 0.1$: Strong evidence against the null hypothesis
- $P < 0.001$: Overwhelming evidence against the null hypothesis

Test for a population proportion: Denote by $x_1, \ldots, x_n$ a given (fixed) observation of n independent identically distributed Bernoulli observables with estimator $\bar{x}$ for the sample proportion. Denote by $X_1, \ldots, X_n$ a general sample with estimator $\bar{X}$. Denote by $H_0 : p = p_0$ the null hypothesis and by $H_1 : p \neq p_0$, $H_1 : p > p_0$ and $H_1 : p < p_0$ three possible alternative hypothesis' corresponding to a two-sided, right-tailed, and left-tailed test. The P-values in theses cases are given by

- Two-sided test ($H_0 : p = p_0$, $H_1 : p \neq p_0$)

$$\begin{aligned} P &= \mathbb{P}(|\bar{X} - p_0| > |\bar{x} - p_0|) = \mathbb{P}\left(|Z| > \frac{|\bar{x} - p_0|}{\sqrt{p_0(1-p_0)/n}}\right) \\ &= 2\mathbb{P}\left(Z > \frac{|\bar{x} - p_0|}{\sqrt{p_0(1-p_0)/n}}\right) = 2\left(1 - \Phi\left(\frac{|\bar{x} - p_0|}{\sqrt{p_0(1-p_0)/n}}\right)\right) \end{aligned}$$

  where we use the standardisation $Z = (\bar{X} - p_0)/\sqrt{p_0(1-p_0)n}$.

- Right-tailed test ($H_0 : p = p_0$, $H_1 : p > p_0$)

$$P = \mathbb{P}(\bar{X} - p_0 > \bar{x} - p_0) = \mathbb{P}\left(Z > \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}}\right) = 1 - \Phi\left(\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

- Left-tailed test ($H_0 : p = p_0$, $H_1 : p < p_0$)

$$P = \mathbb{P}(\bar{X} - p_0 < \bar{x} - p_0) = \mathbb{P}\left(Z < \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}}\right) = \Phi\left(\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

Example 7.15 *According to the U.S. Census Bureau, in 2009, 2.1% of Americans worked at home. An economist believes that the percentage of Americans working at home has increased since then. They randomly select 150 working Americans and find that 6 of them work at home. Does a higher proportion of Americans work at home?*

Compare the null hypothesis H_0 that the percentage has not changed ($p = p_0 = 0.021$) with the alternative hypothesis H_1 that the percentage has increased ($p > p_0$). We apply a right-tailed test. The P-value of the survey is given by

$$P = 1 - \Phi\left(\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

with $\bar{x} = 6/150 = 0.04$, $n = 150$, and $p_0 = 0.021$.

```
> barx=0.04
> p0=0.021
> n=150
> 1-pnorm((barx-p0)/sqrt(p0*(1-p0)/n))
[1] 0.0523028
```

Here $P = 0.0523\ldots$ that means at about 5.2% of samples will give a sample proportion as high or higher as the one we have seen, assuming that the null hypothesis is valid.

$P > \alpha = 0.05$ and we have no evidence to reject the null hypothesis at a 5% significance level. However we reject the null hypothesis at a 10% significance level as $P < \alpha = 0.1$. There is weak evidence against the null hypothesis since $0.05 < P < 0.1$.

Test for a sample mean: Denote by $x_1, \ldots, x_n$ a given (fixed) observation of n independent identically distributed random variables with estimator $\bar{x}$ for the population mean. Denote by $X_1, \ldots, X_n$ a general sample with estimator $\bar{X}$. Assume that the variance of the random variables is known, $\sigma^2 = \mathrm{Var}(X_k)$. Denote by $H_0 : \mu = \mu_0$ the null hypothesis for the mean and by $H_1 : \mu \neq \mu_0$, $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$ three possible alternative hypothesis' corresponding to a two-sided, right-tailed, and left-tailed test. Again using standardisation the P-values in theses cases are given by

- Two- sided test ($H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$)

$$\begin{aligned} P &= \mathbb{P}(|\bar{X} - \mu_0| > |\bar{x} - \mu_0|) = \mathbb{P}\left(|Z| > \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right) \\ &= 2\mathbb{P}\left(Z > \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right) = 2\left(1 - \Phi\left(\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right)\right) . \end{aligned}$$

- Right-tailed test ($H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$)

$$P = \mathbb{P}(\bar{X} - \mu_0 > \bar{x} - \mu_0) = \mathbb{P}\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) .$$

- Left-tailed test ($H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$)

$$P = \mathbb{P}(\bar{X} - \mu_0 < \bar{x} - \mu_0) = \mathbb{P}\left(Z < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) .$$

Example 7.16 *Consider the setup of example 7.14. We test the null hypothesis* $H_0 : \mu = \mu_0 = 60$ *against the alternative* $H_1 : \mu > \mu_0$*, i.e., we use a right-tailed test. The P value of the survey is given by*

$$P = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

with $\mu_0 = 60$, $\sigma = 19.7$ *and* $\bar{x}$ *computed from the data set*

```
> x<-c(79,75,70,26,88,32,83,69,46,50,69,37,72,76,49)
> barx<-mean(x)
> mu0<-60
> sigma<-19.7
> n<-length(x)
> 1-pnorm((barx-mu0)/(sigma/sqrt(n)))
[1] 0.3915669
```

Here $P = 0.3915\ldots$ *which means that at about* 39.2% *of samples will give a mean as high or higher as the one we have seen (which means our value is rather typical).*

Since $P > \alpha = 0.05$ *there is insufficient evidence to reject the null hypothesis at a 5% significance level (as already seen in example 7.14). In fact there no evidence against the null hypothesis since* $P > 0.1$.

Progress Check:

1. What is a hypothesis? What is the main purpose of testing a hypothesis?
2. Explain what the notions null hypothesis and alternative hypothesis mean, and how they are related?
3. According to the U.S. Census Bureau, 10.5% of registered births in the United States in 2007 were to teenage mothers. A sociologist believes that this percentage has increased since then. Formulate the null and the alternative hypothesis.
4. For the example described in item 3 explain what it means that the test results in a type-I error, and what it means that the test results in a type-II error.
5. Explain under which conditions do you apply a two-sided test, a right-tailed test, or a left-tailed test.
6. Which statistic do you use to test for a population mean?
7. Explain the meaning of of the notion P-value. A test produces a P-value of $P = 0.15$. What do you conclude?

§8 Appendix

a) Problem sheet 1 / week 2

Problem 1: A bag contains 10 red balls and 20 blue balls. You pick 5 balls at random. Denote by X the number of red balls you pick.

a) Assume you pick balls without replacement. Compute the probabilities $\mathbb{P}(X = k)$ for all k values, $0, 1, \ldots, 5$. Which of these probabilities has the largest value (the so called *mode* of the pmf).

b) Produce a bar chart of the pmf using the R command `barplot`, and produce a plot of the pmf using the R command `plot` with the option `type` = ”`p`”.

c) Assume you pick balls with replacement. Compute the probabilities $\mathbb{P}(X = k)$ for all k values, $0, 1, \ldots, 5$. Which of these probabilities has the largest value (i.e. state the so-called *mode* of the pmf).

d) Compute the difference of the probabilities computed in part a) and c) (i.e. the difference of the probabilities when picking without and with replacement). Which of the probabilities is larger? Is there any pattern visible?

e) Produce a bar chart of the pmfs using the R command `barplot`. Your plot should contain both pmfs, that of part a) and that of part c) (use e.g. the R command `cbind(y1,y2) ~ x` within `barplot` and the option `beside=TRUE`). Produce as well a plot of both pmfs in a single figure using the R commands `plot` and `lines`.

Problem 2: Redo the entire problem 1 with a bag containing 1000 red balls and 2000 blue balls when you pick again 5 balls at random. In particular, comment in detail on your findings in part c) (i.e. on the difference between probabilities when picking balls with or without replacement).

b) Problem sheet 2 / week 3

Problem 3: We toss a fair coin repeatedly until we have seen in total 20 heads (and then we stop tossing the coin). Denote by X the total number of coin tosses.

a) Plot the pmf of X as a bar chart using the R command `barplot`. Plot the pmf of X using the R command `plot`.

b) Compute/state the expected number of coin tosses.

c) Compute the probability that you toss the coin more than 50 times or less than 30 times.

Problem 4: A library contains $N = 10,000$ books. $M = 100$ of those books are on Probability& Statistics, i.e., only a small proportion $p = M/N = 0.01 = 1\%$ of books is on this topic. A student prepares for an open book exam by making a random selection (with replacement/repetition) of $n = 150$ books from the library catalogue. Let X denote the number of selected books on Probability& Statistics.

a) Plot the pmf of X as a bar chart using the R command `barplot` in the range $[0, 10]$. Plot the pmf of X in the range $[0, 10]$ using the R command `plot`. State the mode of the pmf and the expectation of X.

b) Plot the Poisson distribution with $\lambda = np$ and the pmf of part a) in a single diagram (using `barplot` or using `plot`). Summarise your findings.

Problem 5: The number of hits to a Web site follows a Poisson distribution; hits occur at the rate of 1.4 per minute. Compute the probability that the number of hits between 7:30pm and 7:35pm is:

a) exactly seven.

b) fewer than seven.

c) at least seven.

c) Problem sheet 3 / week 4

Problem 6: Two bus lines, line A and line B, service a station. The passenger waiting times for both buses are independent and exponentially distributed. The passenger waiting time (in units of minutes) for bus A obeys $T_1 \sim \text{Exp}(0.2)$ and the passenger waiting time for bus B obeys $T_2 \sim \text{Exp}(0.25)$.

a) Compute the probability that you wait for bus A more than 3 minutes.

b) Compute the probability that you wait for bus B less than 4 minutes.

c) Using the results of problem 8 compute the probability that you wait for a bus (either bus A or bus B) less than 2 minutes.

Problem 7: The reading speed of sixth-grade students is approximately normal, with a mean speed of 125 words per minute and a standard deviation of 24 words per minute.

a) What is the probability that a randomly selected sixth-grade student reads less than 100 words per minute?

b) What is the probability that a randomly selected sixth-grade student reads more than 140 words per minute?

c) What is the probability that a randomly selected sixth-grade student reads between 110 and 130 words per minute?

d) Would it be unusual for a sixth-grader to read more than 200 words per minute? Why?

Problem 8: Denote by X_1 and X_2 two continuous random variables which are exponentially distributed, $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$. Assume X_1 and X_2 to be independent.

a) Compute the probability $\mathbb{P}(X_1 > a)$ (the so called complementary cdf).

b) Compute the probability $\mathbb{P}(X_2 > a)$ (the so called complementary cdf).

c) Denote by $Z = \min\{X_1, X_2\}$ the minimum of both variables. Express $\mathbb{P}(Z > a)$ as a joint probability and compute $\mathbb{P}(Z > a)$ using the results of part a) and b).

d) Using the result of part c) compute the cdf of the random variable Z and hence show that the pdf of Z is again an exponential distribution with parameter $\lambda_1 + \lambda_2$.

d) Problem sheet 4 / week 5

Problem 9: Suppose X is a random variable which obeys the Poisson distribution $X \sim \text{Poisson}(\lambda)$.

a) Suppose that $\lambda = 5$. Compute/state the expectation $\mathrm{E}(X)$ and the standard deviation σ_X of X. Plot the pmf of X and plot the normal approximation of the pmf (i.e. a normal distribution with expectation $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$) in a single diagram. Chose the range of the plot sensibly.

b) Suppose that $\lambda = 20$. Compute/state the expectation $\mathrm{E}(X)$ and the standard deviation σ_X of X. Plot the pmf of X and plot the normal approximation of the pmf (i.e. a normal distribution with expectation $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$) in a single diagram. Chose the range of the plot sensibly.

c) Suppose that $\lambda = 80$. Compute/state the expectation $\mathrm{E}(X)$ and the standard deviation σ_X of X. Plot the pmf of X and plot the normal approximation of the pmf (i.e. a normal distribution with expectation $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$) in a single diagram. Chose the range of the plot sensibly. Summarise your findings.

Problem 10: A bag contains 500 gold coins and 1000 silver coins. You pick 100 coins at random (without replacement). Let X denote the number of gold coins you pick.

a) Compute and plot the pmf of X. Your plot should contain as well the normal approximation of the pmf, i.e., a normal distribution with mean $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$.

b) Compute the difference between $P(X = k)$ and the normal approximation of this probability, i.e., the difference between $P(X = k)$ and the integral

$$\int_{k-1/2}^{k+1/2} \frac{1}{\sqrt{2\pi}} e^{-(t-\mu)^2/(2\sigma^2)} dt$$

for all k values. Plot the difference as a function of k. Estimate (or compute) the maximum value of the error.

Problem 11: A single bus line services a bus station with buses arriving/departing in a random way. Denote by T the time interval between the arrival of two buses. Assume that T obeys the geometric distribution with expected value $\mathrm{E}(T) = 4.2$ (time is measured in unit of minutes). Assume that the arrival of buses are independent events, i.e. that subsequent time intervals $T_1, T_2, \ldots$ are independent identically distributed random variables. Denote by X the time interval such that exactly 20 buses arrive (i.e. the time interval between the arrival of a bus and the arrival of the 20th bus).

a) Compute the expectation and the variance of X.

b) Using the normal approximation compute the probability that the arrival of 20 buses takes more than 2 hours.

c) Is it reasonable to assume that 20 buses arrive within 5 minutes? Give a reason.

e) Problem sheet 5 / week 6

Problem 12: The R datasets package contains the dataset `eurodist` which gives the road distances (in km) between 21 cities in Europe. The data are taken from a table in The Cambridge Encyclopaedia.

a) Compute the fivenum of intercity distances, e.g., using the R command `summary`.

b) Compute the measures for dispersion of the data set, i.e., compute the variance, the standard deviation and the IQR.

c) Plot a frequency distribution of the data using the R command `hist`. Try different groupings of the data.

d) Produce a boxplot and compare the diagram to your findings in part a).

e) Summarise your findings, e.g., comment on the average value of the data (in terms of mean, median, mode), the dispersion (in terms of standard deviation and IQR). Is the frequency distribution symmetric or skewed? Do the findings look reasonable?

Problem 13: The R datasets package contains the dataset `morely` which contains classical data of Michelson on measurements done in 1879 on the speed of light (the speed of light is at about 300,000 km/sec). The data consists of five experiments, each consisting of 20 consecutive ‘runs’. The response is the speed of light measurement, suitably coded (in km/sec, with 299,000 subtracted).

a) Compute the fivenum of the data, e.g., using the R command `summary`.

b) Compute the measures for dispersion of the data set, i.e., compute the variance, the standard deviation, and the IQR.

c) Plot a frequency distribution of the data using the R command `hist`. Try different groupings of the data.

d) Produce a boxplot and compare the diagram to your findings in part a).

e) Summarise your findings, e.g., comment on the average value of the data (in terms of mean, median, mode), the dispersion (in terms of standard deviation and IQR). Is the frequency distribution symmetric or skewed? Do the findings look reasonable?

Problem 14: The R datasets package contains the dataset `state.x77` which contains data related to the 50 states of the United States of America.

a) Familiarise yourself with the dataset by using, e.g., the R command `View`.

b) Plot a bar chart of the illiteracy (given in percent of the population, data from 1970) using the R command `barplot`.

c) Plot a bar chart of the murder and non-negligent manslaughter rate (given per 100,000 population, data from 1976) using the R command `barplot`.

d) Produce a scatter plot of the illiteracy vs. the murder rate.

e) Compute the correlation coefficient between the illiteracy and the murder rate. Comment on your findings.

f) Problem sheet 6 / week 8

Problem 15: We want to estimate the unknown proportion p of registered Scottish voters who agree that Scotland should be independent. Anne takes a random sample of 30 voters and records the number X who agree that Scotland should be independent. Ben takes another random sample of 20 voters and records the number Y who agree that Scotland should be independent. Assume that these two samples are independent of each other.

a) Anne proposes to estimate p by the estimator $A = X/30$. Find the bias and the mean square error of this estimator.

b) Ben proposes to estimate p by the estimator $B = Y/20$. Find the bias and the mean square error of this estimator.

c) Craig proposes to estimate p by the estimator $C = (X/30 + Y/20)/2$. Find the bias and the mean square error of this estimator.

d) Donald proposes to estimate p by the estimator $D = (X + Y)/50$. Find the bias and the mean square error of this estimator.

e) Which estimator do you think is the best?

Problem 16: The S&P 500 is a collection of 500 stocks of publicly traded companies. Using the data obtained from Yahoo! Finance, the monthly rates of return of the S&P 500 since 1950 are normally distributed. The mean rate of return is 0.7233% and the standard deviation for the rate of return is 4.135% .

a) What is the probability that a randomly selected month has a positive rate of return?

b) Treating the next 6 months as a simple random sample, what is the probability that the mean monthly rate of return will be positive?

c) Treating the next 12 months as a simple random sample, what is the probability that the mean monthly rate of return will be positive?

d) Treating the next 24 months as a simple random sample, what is the probability that the mean monthly rate of return will be positive?

e) Use the results of parts b)-d) to describe the likelihood of earning a positive rate of return on stocks as the investment time horizon increases.

g) Problem sheet 7 / week 9

Problem 17: Consider two independent exponentially distributed random variables, $X_1 \sim \text{Exp}(2.5)$ and $X_2 \sim \text{Exp}(4)$ (see problem 8).

a) Denote by $Z = \min\{X_1, X_2\}$ the minimum of the variables. Generate a sample of size 10,000 (use e.g. the R command `rexp(1,2.5)` and `rexp(1,4)` to generate single random values for the variables X_1 and X_2, use `min` to compute the value of Z, and the command `replicate` to generate the sample of size 10,000).

b) Use your sample to plot a histogram for the pdf of Z with a sensible grouping. Your diagram should include as well the exact analytic result of the pdf (see problem 8d).

c) Denote by $U = \min\{X_1, X_2\}$ the maximum of the variables. Generate a sample of size 10,000 and plot a histogram for the pdf of U with a sensible grouping. Comment on the shape of the pdf (or provide an exact analytic expression, if possible).

Problem 18: Consider the setup of problem 11. Denote by $\bar{T}$ the sample mean of 19 consecutive waiting time intervals (i.e. $19\bar{T}$ is the time interval for the arrival of 20 buses).

a) Compute the expectation and the variance of $\bar{T}$.

b) Generate a sample of size 10,000 and plot the pdf of $\bar{T}$ as a histogram with a sensible grouping of data. Include in your plot the normal approximation of the sampling distribution. Comment on the accuracy of the normal approximation.

Problem 19: The 2006 General Social Survey asked 26540 respondents "If ever married, how old were you when you first married?". The survey gave a sample mean of 22.15 yrs with standard deviation 4.885 yrs.

a) Use this information to construct a 99% confidence interval for the mean age of first marriage.

b) What does the confidence interval mean?

h) Problem sheet 8 / week 10

Problem 20: In a survey of 3611 adult Americans 18 years and older conducted in July 2010 by SmartRevenue, it was found that 542 have used their smartphone to make a purchase.

a) Obtain a point estimate for the population proportion of adult Americans 18 years and older who have used their smartphone to make a purchase.

b) Verify that the requirements for constructing a confidence interval about p are satisfied.

c) Construct a 95% confidence interval for the population proportion of adult Americans who have used their smartphone for purchase.

d) Interpret the interval.

Problem 21: Consider the sample of marks in the 2019 Calculus I final exam given in example 4.8 of the lecture notes.

a) Compute a point estimate for the population proportion of students scoring an A grade.

b) Compute a 90% confidence interval for the population proportion.

c) Interpret the interval.

Problem 22: For the following two setups formulate the null and the alternative hypothesis, explain what it would mean to make a type-I error, and explain what it would mean to make a type-II error.

a) According to the national Association of Home Builders, the mean price of an existing single-family home in 2009 was \$218,600. A real estate broker believes that because of the recent credit crunch, the mean price has decreased since then.

b) According to the *CTIA-The Wireless Association*, the mean monthly cell phone bill was \$41.47 in 2010. A researcher suspects that the mean monthly cell phone bill is different today.

i) Problem sheet 9 / week 11

Problem 23: In August 2003, 56% of employed adults in the United States reported that basic mathematical skills were critical or very important to their job. The supervisor of the job placement office at a year-4 college thinks this percentage has increased due to the increased use of technology in the workplace. He takes a random sample of 480 employed adults and finds that 297 of them feel basic mathematical skills are critical or very important to their job. Is this sufficient evidence to conclude that the percentage of employed adults who feel basic mathematical skills are critical or very important to their job has increased at the $\alpha = 0.05$ level of significance?

Problem 24: Suppose you wish to find out the answer to the age-old question, "Do Americans prefer Coke or Pepsi?" You conduct a blind taste test in which individuals are randomly asked to drink one of the colas first, followed by the other cola, and then asked to disclose which drink they prefer. Results of your taste test indicate that 53 of 100 individuals prefer Pepsi.

a) Conduct a hypothesis test H_0: $p = p_0$ versus $H_1 : p \neq p_0$ for $p_0 = 0.42, 0.43, 0.44, \ldots, 0.64$ at the $\alpha = 0.05$ level of significance. For which values of p_0 do you reject the null hypothesis? What do each of the values of p_0 represent?

b) Construct a 95% confidence interval for the proportion of individuals who prefer Pepsi.

c) Suppose you changed the level of significance in conducting the hypothesis test to $\alpha = 0.01$. What would happen to the range of values of p_0 for which the null hypothesis is not rejected? Why does that make sense?

j) Problem sheet 10 / week 12

Problem 25: A Fair Isaac Corporation (FICO) score is used by credit agencies (such as mortgage companies and banks) to assess the creditworthiness of individuals. Its value ranges from 300 to 850. An individual with FICO score over 750 is considered to be a quality credit risk. According to Fair Isaac Corporation, the mean FICO score is 703.5. A credit analyst wondered whether high-income individuals (incomes in excess of £100,000 per year) had higher credit scores. He obtained a random sample of 40 high-income individuals and found the sample mean credit score to be 714.2 with a standard deviation of 83.2. Conduct the appropriate test to determine if high-income individuals have higher FICO scores at the $\alpha = 0.05$ level of significance.

Problem 26: A math teacher claims that she has developed a review course that increases the scores of students on the math portion of the SAT exam. Based on data from the College Board, SAT scores are normally distributed with a mean 515. The teacher obtains a random sample of 1800 students, puts them through the review class, and finds that the mean SAT math score of the 1800 students is 519 with a standard deviation of 111.

a) Formulate the null and the alternative hypothesis.

b) Conduct the appropriate test and compute the P-value. Is a mean SAT math score of 519 significantly higher than 515, say at the $\alpha = 0.1$ level of significance.

c) Test the hypothesis with 400 students by computing the P-value. Assume the same sample statistics. Is a sample mean of 519 significantly more than 515, say at the $\alpha = 0.1$ level of significance. What do you conclude about the impact of large samples on the P-value?

Problem 27: The exponential probability density can be used to model waiting time in line or the lifetime of electronic components. Its density function is skewed right. Suppose the wait-time in a line can be modelled by the exponential distribution with expectation 5 minutes and 15 seconds.

a) Simulate a random sample of size $n = 100$, i.e., simulate a random sample of 100 individuals with average wait-time of 5 minutes and 15 seconds.

b) Consider the null hypothesis $H_0 : \mu = 5$ versus the alternative $H_1 : \mu \neq 5$. Use your sample and compute the P-value of the test. Interpret the P-value, i.e., is there any evidence against the null hypothesis, based on your sample.

k) R commands for figures

- Fig.1.1

```
> barplot(c(0.8,0.2) ~ c(0,1),xlab="k",ylab="P(X=k)",ylim=c(0,1),
    cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(c(0.5,0.5) ~ c(0,1),xlab="k",ylab="P(X=k)",ylim=c(0,1),
    cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(c(0.2,0.8) ~ c(0,1),xlab="k",ylab="P(X=k)",ylim=c(0,1),
    cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
```

- Fig.1.2

```
> barplot(dbinom(0:15,size=15,prob=0.2)~seq(0,15,1),xlab="k",
    ylab="P(X=k)",ylim=c(0,0.25),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
> barplot(dbinom(0:15,size=15,prob=0.5)~seq(0,15,1),xlab="k",
    ylab="P(X=k)",ylim=c(0,0.25),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
> barplot(dbinom(0:15,size=15,prob=0.5)~seq(0,15,1),xlab="k",
    ylab="P(X=k)",ylim=c(0,0.25),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
```

- Fig.1.3

```
> barplot(dgeom(0:15,prob=0.2)~seq(0,15,1),xlab="k",ylab="P(T=k)",
    ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(dgeom(0:15,prob=0.5)~seq(0,15,1),xlab="k",ylab="P(T=k)",
    ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(dgeom(0:15,prob=0.8)~seq(0,15,1),xlab="k",ylab="P(T=k)",
    ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
```

- Fig.1.4

```
> barplot(dhyper(0:15,4,20-4,15)~seq(0,15,1),xlab="k",ylab="P(X=k)",
    ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(dhyper(0:15,8,40-8,15)~seq(0,15,1),xlab="k",ylab="P(X=k)",
    ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(dhyper(0:15,16,80-16,15)~seq(0,15,1),xlab="k",ylab="P(X=k)",
    ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
```

- Fig.1.5

```
> barplot(dnbinom(0:17, size=3, prob=0.3) ~ seq(0,17,1),xlab="k",
    ylab="P(T=k)",ylim=c(0,0.4),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
> barplot(dnbinom(0:17, size=3, prob=0.5) ~ seq(0,17,1),xlab="k",
    ylab="P(T=k)",ylim=c(0,0.4),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
> barplot(dnbinom(0:17, size=3, prob=0.7) ~ seq(0,17,1),xlab="k",
    ylab="P(T=k)",ylim=c(0,0.4),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
```

- Fig.1.6

```
> barplot(dunif(-4:5,min=-2.5,max=3.5) ~ seq(-4,5,1),xlab="k",
    ylab="P(X=k)",ylim=c(0,0.3),cex.names=1.5, cex.axis=1.5,
    cex.lab=1.5)
```

- Fig.1.7

```
> barplot(dpois(0:10,0.5) ~ seq(0,10,1),xlab="k",ylab="P(X=k)",
    ylim=c(0,0.8),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(dpois(0:10,2) ~ seq(0,10,1),xlab="k",ylab="P(X=k)",
    ylim=c(0,0.8),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> barplot(dpois(0:10,4) ~ seq(0,10,1),xlab="k",ylab="P(X=k)",
    ylim=c(0,0.8),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
```

- Fig.1.8

```
> barplot(dunif(0:7,min=0.5,max=6.5) ~ seq(0,7,1),xlab="k",
  ylab="P(X=k)",ylim=c(0,0.3),cex.names=1.5, cex.axis=1.5,
  cex.lab=1.5)
> plot(c(0,0.99,1,1.99,2,2.99,3,3.99,4,4.99,5,5.99,6,7),
  punif(c(0,0,1,1,2,2,3,3,4,4,5,5,6,7),0,6),xlab="k",
  ylab=expression(P(X<=k)),xlim=c(0,7),ylim=c(0,1),t="l",
  cex.axis=1.5, cex.lab=1.5)
```

- Fig.1.9

```
> barplot(dhyper(0:15,16,80-16,15)~seq(0,15,1),xlab="k",ylab="P(X=k)",
  ylim=c(0,0.85),cex.names=1.5, cex.axis=1.5, cex.lab=1.5)
> plot(seq(0,15,0.02),phyper(seq(0,15,0.02),16,80-16,15),xlab="k",
  ylab=expression(P(X<=k)),xlim=c(0,15),ylim=c(0,1),t="l",
  cex.axis=1.5, cex.lab=1.5)
```

- Fig.2.1

```
> par(mgp=c(2.5,1,0))
> plot(seq(2,6,0.01),dunif(seq(2,6,0.01),3,5),type="l",xlab="x",
  ylab=expression(f[X](x)),xlim=c(2,6),ylim=c(0,0.6),cex.axis=1.5,
  cex.lab=1.5)
> polygon(c(4,4,4.5,4.5,4),c(0,0.5,0.5,0,0), col="grey")
```

- Fig.2.2

```
> par(mgp=c(2.5,1,0))
> plot(seq(2,6,0.01),punif(seq(2,6,0.01),3,5),type="l",xlab="x",
  ylab=expression(F[X](x)),xlim=c(2,6),ylim=c(0,1),cex.axis=1.5,
  cex.lab=1.5)
```

- Fig.2.3

```
> par(mgp=c(2.5,1,0))
> plot(seq(-1,7,0.01),dexp(seq(-1,7,0.01),0.7),type="l",xlab="x",
  ylab=expression(f[X](x)),xlim=c(0,7),ylim=c(0,5),cex.axis=1.5,
  cex.lab=1.5)
> plot(seq(-1,7,0.01),dexp(seq(-1,7,0.01),2),type="l",xlab="x",
  ylab=expression(f[X](x)),xlim=c(0,7),ylim=c(0,5),cex.axis=1.5,
  cex.lab=1.5)
> plot(seq(-1,7,0.01),dexp(seq(-1,7,0.01),4),type="l",xlab="x",
  ylab=expression(f[X](x)),xlim=c(0,7),ylim=c(0,5),cex.axis=1.5,
  cex.lab=1.5)
```

- Fig.2.4

```
> par(mgp=c(2.5,1,0))
> plot(seq(-1,7,0.01),pexp(seq(-1,7,0.01),0.7),type="l",xlab="x",
    ylab=expression(F[X](x)),xlim=c(0,7),ylim=c(0,1),cex.axis=1.5,
    cex.lab=1.5)
> plot(seq(-1,7,0.01),pexp(seq(-1,7,0.01),2),type="l",xlab="x",
    ylab=expression(F[X](x)),xlim=c(0,7),ylim=c(0,1),cex.axis=1.5,
    cex.lab=1.5)
> plot(seq(-1,7,0.01),pexp(seq(-1,7,0.01),4),type="l",xlab="x",
    ylab=expression(F[X](x)),xlim=c(0,7),ylim=c(0,1),cex.axis=1.5,
    cex.lab=1.5)
```

- Fig.2.5

```
> par(mgp=c(2.5,1,0))
> plot(seq(-1,12,0.01),dexp(seq(-1,12,0.01),0.4),type="l",xlab="x",
    ylab=expression(f[X](x)),xlim=c(0,12),ylim=c(0,0.5),cex.axis=1.5,
    cex.lab=1.5)
> xp<-c(5,seq(5,12,0.01),12,5)
> yp<-c(0,dexp(seq(5,12,0.01),0.4),0,0)
> polygon(xp,yp,col="grey")
```

- Fig.2.6

```
> mu<-0.5
> sd<-1.5
> x<-seq(-4,5,length=100)
> y<-dnorm(x,mu,sd)
> par(mgp=c(2.5,1,0))
> plot(x, y, type="l",xlab="x",ylab=expression(f[X](x)),xlim=c(-4,5),
    ylim=c(0,0.3),cex.axis=1.5, cex.lab=1.5)
> lines(c(mu,mu),c(0,0.3),type="l",lty="solid")
> lines(c(mu-sd,mu-sd),c(0.05,0.25),type="l",lty="dashed")
> lines(c(mu+sd,mu+sd),c(0.05,0.25),type="l",lty="dashed")
```

- Fig.2.7

```
> mu<-1.0; sd<-0.5; x<-seq(-5,4,length=100); y<-dnorm(x,mu,sd)
> plot(x, y, type="l",xlab="x",ylab=expression(f[X](x)),xlim=c(-5,4),
    ylim=c(0,1.0),cex.axis=1.5, cex.lab=1.5)
> lines(c(mu,mu),c(0,1.0),type="l",lty="solid")
> lines(c(mu-sd,mu-sd),c(0.1,0.7),type="l",lty="dashed")
> lines(c(mu+sd,mu+sd),c(0.1,0.7),type="l",lty="dashed")
> mu<--0.5; sd<-0.5; x<-seq(-5,4,length=100); y<-dnorm(x,mu,sd)
> plot(x, y, type="l",xlab="x",ylab=expression(f[X](x)),xlim=c(-5,4),
    ylim=c(0,1.0),cex.axis=1.5, cex.lab=1.5)
> lines(c(mu,mu),c(0,1.0),type="l",lty="solid")
> lines(c(mu-sd,mu-sd),c(0.1,0.7),type="l",lty="dashed")
> lines(c(mu+sd,mu+sd),c(0.1,0.7),type="l",lty="dashed")
> mu<--0.5; sd<-1.5; x<-seq(-5,4,length=100); y<-dnorm(x,mu,sd)
> plot(x, y, type="l",xlab="x",ylab=expression(f[X](x)),xlim=c(-5,4),
    ylim=c(0,1.0),cex.axis=1.5, cex.lab=1.5)
> lines(c(mu,mu),c(0,1.0),type="l",lty="solid")
> lines(c(mu-sd,mu-sd),c(0.1,0.7),type="l",lty="dashed")
> lines(c(mu+sd,mu+sd),c(0.1,0.7),type="l",lty="dashed")
```

- Fig.2.8

```
> plot(seq(-4,4,0.02),pnorm(seq(-4,4,0.02)),type="l",xlab="z",
    ylab=expression(Phi(z)),xlim=c(-4,4),ylim=c(0,1.0),cex.axis=1.5,
    cex.lab=1.5)
```

- Fig.2.9

```
> a<-1.0
> x<-seq(-4,4,length=100)
> y<-dnorm(x)
> x2<-seq(a,4,length=100)
> x1<-c(a,x2,4,a)
> y1<-c(0,dnorm(x2),0,0)
> par(mgp=c(2.5,1,0))
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(x1,y1, col="grey")
> polygon(-x1,y1, col="grey")
```

- Fig.2.10

```
> a<-2.0
> x<-seq(-4,4,length=100)
> y<-dnorm(x)
> x2<-seq(a,4,length=100)
> x1<-c(a,x2,4,a)
> y1<-c(0,dnorm(x2),0,0)
> par(mgp=c(2.5,1,0))
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(x1,y1, col="grey")
> polygon(-x1,y1, col="grey")
```

- Fig.2.11

```
> a<-1.4
> x<-seq(-4,4,length=100)
> y<-dnorm(x)
> x2<-seq(a,4,length=100)
> x1<-c(a,x2,4,a)
> y1<-c(0,dnorm(x2),0,0)
> par(mgp=c(2.5,1,0))
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(x1,y1, col="grey")
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(-x1,y1, col="grey")
```

- Fig.3.1

```
> n<-50
> p<-0.3
> mu<-n*p
> sigma<-sqrt(n*p*(1-p))
> par(mgp=c(2.5,1,0))
> plot(0:n,dbinom(0:n,n,p),,xlab=" ",ylab=" ",xlim=c(0,50),
    ylim=c(0,0.15),cex.axis=1.5, cex.lab=1.5)
> x<-seq(0,n,0.1)
> lines(x,dnorm(x,mu,sigma))
> plot(0:n,dbinom(0:n,n,p),,xlab=" ",ylab=" ",xlim=c(5,25),
    ylim=c(0,0.15),cex.axis=1.5, cex.lab=1.5)
> lines(x,dnorm(x,mu,sigma))
```

- Fig.4.1

```
> x<-c(124,37,23,8,4,11)
> grade<-c("A","B","C","D","E","F")
> barplot(x,names.arg=grade)
```

- Fig.4.2

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> hist(x,breaks=seq(22.5,39.5,1))
```

- Fig.4.3

```
> barplot(c(10,20,40,70,40,20,10),
    names.arg=c("A","B","C","D","E","F","G"))
> barplot(c(20,40,60,80,40,20,10),
    names.arg=c("A","B","C","D","E","F","G"))
> barplot(c(10,20,40,80,60,40,20),
    names.arg=c("A","B","C","D","E","F","G"))
> barplot(c(5,10,20,40,70,40,20),
    names.arg=c("A","B","C","D","E","F","G"))
> barplot(c(10,20,70,40,80,50,20),
    names.arg=c("A","B","C","D","E","F","G"))
```

- Fig.4.4

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> hist(x,breaks=seq(22.5,39.5,1),freq=FALSE)
> lines(c(26,26),c(0.02,0.15))
> lines(c(33,33),c(0.02,0.15))
> lines(c(35,35),c(0.02,0.15))
```

- Fig.4.5

```
> x<-c(26, 35, 23, 32, 35, 33, 39, 32, 25, 38, 34)
> boxplot(x)
```

- Fig.4.6

```
> x<-c(62,84,18,64,88,69,66,69,73,84,95,82,68,91,65,
       100,83,95,91,82,44,83,45,82,93,74,73,91,14)
> y<-c(84,73,47,74,64,86,88,76,76,79,98,85,72,81,93,
       91,90,89,95,77,91,75,62,90,61,77,69,65,49)
> plot(x,y,xlab="Calc I",ylab="Int. Prob.")
```

- Fig.5.1

```
> x<-rhyper(200,30,20,15)
> hist(x,freq=FALSE,breaks=seq(-0.5,15.5,1))
> lines(0:15,dhyper(0:15,30,20,15,type="p")
> x<-rhyper(5000,30,20,15)
> hist(x,freq=FALSE,breaks=seq(-0.5,15.5,1))
> lines(0:15,dhyper(0:15,30,20,15,type="p")
```

- Fig.5.2

```
> x<-rexp(500,0.4)
> hist=(x,freq=FALSE,breaks=seq(0,25,0.5)
> lines(seq(0,25,0.1),dexp(seq(0,25,0.1),0.4),type="l")
```

- Fig.5.3

```
> xbar<-replicate(1000,mean(rpois(n=50,lambda=4)))
> hist(xbar,freq=FALSE,seq(3,5,0.1))
> lines(seq(3,5,0.02),dnorm(seq(3,5,0.02),mean=4,sd=sqrt(2/25)))
```

- Fig.7.1

```
> a<-1.4
> x<-seq(-4,4,length=100); y<-dnorm(x)
> x2<-seq(a,4,length=100); x1<-c(a,x2,4,a); y1<-c(0,dnorm(x2),0,0)
> par(mgp=c(2.5,1,0))
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(x1,y1, col="grey")
> polygon(-x1,y1, col="grey")
> a<-1.2
> x2<-seq(a,4,length=100); x1<-c(a,x2,4,a); y1<-c(0,dnorm(x2),0,0)
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(x1,y1, col="grey")
> plot(x, y, type="l",xlab="z",ylab=expression(f[Z](z)),xlim=c(-4,4),
    ylim=c(0,0.5),cex.axis=1.5, cex.lab=1.5)
> polygon(-x1,y1, col="grey")
```

l) What is Statistics?

A large part of Statistics deals with so called statistical inference, i.e., to determine properties and parameters of a (large) population from (small) samples. The methods used and sometimes the accuracy of the estimate may depend on how one interprets probabilities. There are two main strands of Statisticians. So called frequentists consider probabilities to be relative frequencies, i.e., one repeats an experiment a large number of times and the probability of an event is the number of occurrences of the event divided by the number of repetitions (in the limit of an infinite number of repetitions). Within a so called Bayesian view probabilities are an intrinsic (abstract) property of an event, and the values of probabilities can even come from belief or logical support.

We will illustrate these issues by the German tank problem (see example 0.1). In abstract terms we pick a sample of size m, $\{x_1, x_2, \ldots, x_m\}$, from the set of integers (the population) $\{1, 2, \ldots, N\}$. The task is to estimate the unknown population size N from the sample. For instance, if we have the sample $\{23, 2, 97, 128\}$ ($m = 4$) what can we say about N (certainly $N \geq 128$)? For convenience (and with a slight abuse) of notation we denote by x_m the largest value in our sample, the maximum of the set. The interpretation of probabilities will have an impact on how to approach the problem to estimate N.

Since we want to estimate the parameter N we could resort to the approaches outlined in sections 5-7. The sample $X_1, X_2, \ldots, X_m$ may be considered as a collection of independent random variables, if we assume N to be large (compared to m), the so called independence condition. Since X_ℓ are uniformly distributed we have (see section 1b)) $\mathrm{E}(X_\ell) = (N+1)/2$ and $\mathrm{Var}(X_\ell) = (N^2-1)/12$. Using the sample mean $\bar{X}$ it follows that $\mathrm{E}(\bar{X}) = (N+1)/2$. Hence $Y = 2\bar{X} - 1$ is an unbiased estimator for the population size since $\mathrm{E}(Y) = N$. The mean square error of this estimator is then simply given by $MSE_Y = \mathrm{Var}(Y) = 4\mathrm{Var}(\bar{X}) = (N^2-1)/(3m)$. Intuitively it should be possible to come up with a better estimator. For instance, the sample mean does not even guarantee that the estimate exceeds the maximum of the sample, e.g., for the sample $\{23, 2, 97, 128\}$ the sample mean is $\bar{x} = 62.5$ and the estimate gives $Y = 124$ which is of course too small. Certainly the maximum of the sample, X_m, would give a better estimate. In addition, it may be that the independence and the normality condition (which require N and m to be large) are not met in applications. Hence, we are going to deal with a couple of issues. How to perform parameter estimates and hypothesis testing in small populations, how to deal with different statistics, and to explore the impact of a frequentist or Bayesian view on the inference problem.

Population size estimate (frequentist view): The parameter N is considered as an unknown but otherwise fixed numerical value. Consider a given sample of size m

$$\{x_1, x_2, \ldots, x_m\} = \{x_1, \ldots, x_{m-1}\} \cup \{x_m\}$$

where for convenience x_m denotes the maximal value. We will use the statistics X_m to estimate the parameter N. Key to the analysis is the sampling distribution, i.e., the pmf $\mathbb{P}(X_m = \ell)$. We cannot rely on our analysis of section 5 as we are dealing with a different statistics and we cannot assume N or m to be large. Hence the pmf has to be computed from scratch.

Our sample space contains $\binom{N}{m}$ samples (the number of m element subsets chosen from N elements). The event $X_m = \ell$ contains sets with $m-1$ elements chosen from $\{1, \ldots, \ell-1\}$, hence the event has size $\binom{\ell-1}{m-1}$. Thus the pmf reads

$$\mathbb{P}(X_m = \ell) = \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}}, \quad (m \le \ell \le N) \quad N \text{ fixed}\,.$$

Expectation and variance can be computed as an elaborate exercise in binomial coefficients (see the end of the paragraph) and result in

$$\mathrm{E}(X_m) = \frac{m(N+1)}{m+1}, \quad \mathrm{Var}(X_m) = \frac{m(N+1)(N-m)}{(m+2)(m+1)^2}\,.$$

The expectation tells us that

$$Y = \frac{m+1}{m} X_m - 1$$

is an unbiased estimator for N (as $\mathrm{E}(Y) = N$) with mean square error

$$MSE_Y = \mathrm{Var}(Y) = \left(\frac{m+1}{m}\right)^2 \mathrm{Var}(X_m) = \frac{(N+1)(N-m)}{m(m+2)}\,.$$

This mean square error is substantially smaller than the mean square error based on the sample mean. Hence the estimate of this paragraph gives better values.

Let us conclude with the explicit calculation of expectation and variance. Using the identity

$$\binom{k}{m} = \binom{k+1}{m+1} - \binom{k}{m+1}$$

we conclude that

$$\sum_{k=m}^{N} \binom{k}{m} = \sum_{k=m}^{N} \left(\binom{k+1}{m+1} - \binom{k}{m+1} \right) = \binom{N+1}{m+1}$$

which shows that the pmf $\mathbb{P}(X_m = \ell)$ is normalised. For the expectation observe that

$$\binom{\ell}{m} = \frac{\ell}{m} \binom{\ell-1}{m-1}$$

so that

$$\mathrm{E}(X_m) = \frac{\sum_{\ell=m}^{N} \ell \binom{\ell-1}{m-1}}{\binom{N}{m}} = \frac{m \sum_{\ell=m}^{N} \binom{\ell}{m}}{\binom{N}{m}} = \frac{m \binom{N+1}{m+1}}{\binom{N}{m}} = \frac{m}{m+1}(N+1)\,.$$

To compute the variance consider the identity

$$\binom{\ell+1}{m+1} = \frac{\ell(\ell+1)}{m(m+1)} \binom{\ell-1}{m-1}$$

and evaluate the factorial moment

$$\begin{aligned} \mathrm{E}(X_m(X_m+1)) &= \frac{\sum_{\ell=m}^{N} \ell(\ell+1) \binom{\ell-1}{m-1}}{\binom{N}{m}} = \frac{m(m+1) \sum_{\ell=m}^{N} \binom{\ell+1}{m+1}}{\binom{N}{m}} \\ &= \frac{m(m+1) \binom{N+2}{m+2}}{\binom{N}{m}} = \frac{m}{m+2}(N+1)(N+2)\,. \end{aligned}$$

The variance then follows by a straightforward (but slightly lengthy) calculation

$$\mathrm{Var}(X_m) = \mathrm{E}(X_m(X_m+1)) - \mathrm{E}(X_m) - (\mathrm{E}(X_m))^2 = \frac{m}{m+2} \frac{(N+1)(N-m)}{(m+1)^2}\,.$$

Population size estimate (Bayesian view): Consider N to be a random variable with pmf $P(N = n)$ (a so called "prior pmf"). Assume that the maximal value in our sample has value $X_m = \ell$. We are interested in the conditional pmf $P(N = n | X_m = \ell)$ (the so called "posterior pmf") as this pmf tells us something about the population size taking the knowledge from the sample into account. The posterior can be written in terms of the prior using Bayes theorem

$$P(N = n | X_m = \ell) = \frac{\mathbb{P}(X_m = \ell | N = n) \mathbb{P}(N = n)}{\mathbb{P}(X_m = \ell)} = \frac{\mathbb{P}(X_m = \ell | N = n) \mathbb{P}(N = n)}{\sum_{\nu=1}^{\infty} \mathbb{P}(X_m = \ell | N = \nu) \mathbb{P}(N = \nu)}\,.$$

The remaining condition probability $\mathbb{P}(X_m = \ell | N = n)$ has been already worked out in the previous paragraph (as the condition $N = n$ fixes N)

$$\mathbb{P}(X_m = \ell | N = n) = \begin{cases} \frac{\binom{\ell-1}{m-1}}{\binom{n}{m}} & \text{if} \quad m \le \ell \le n \\ 0 & \text{if} \quad \ell < m \text{ or } \ell > n \end{cases}\,.$$

We still need an expression for the prior and that expression is “pulled from the air” (e.g. an educated guess, or something similar). If the Bayesian approach is meaningful the computed posterior pmf will not depend substantially on the chosen prior. In the present case we assume that N is uniformly distributed among a large set of numbers $1 \leq N \leq \Omega$ (to express the lack of knowledge about N), i.e. we chose the prior

$$\mathbb{P}(N) = \begin{cases} 1/\Omega & \text{if} \quad 1 \leq n \leq \Omega \\ 0 & \text{if} \quad n > \Omega \end{cases} .$$

Other sensible priors will give similar final results, but with the current prior all computations can be done analytically. The posterior pmf then reads

$$P(N = n | X_m = \ell) = \frac{\mathbb{P}(X_m = \ell | N = n)}{\sum_{\nu=\ell}^{\Omega} \mathbb{P}(X_m = \ell | N = \nu)} = \frac{\binom{n}{m}^{-1}}{\sum_{\nu=\ell}^{\Omega} \binom{\nu}{m}^{-1}} .$$

The remaining sum in the denominator can be evaluated again as an elaborate exercise in binomial coefficients (see the end of the paragraph). In fact, as the series in the denominator converges, all sufficiently large values of Ω will result in the same expression indicating that the impact of the prior is minimal

$$P(N = n | X_m = \ell) = \frac{m-1}{\ell} \frac{\binom{\ell}{m}}{\binom{n}{m}} .$$

To estimate the population size we can take e.g. the (conditional) expectation, while the variance will give an indication of the mean square error (the explicit calculation will be given at the end of this paragraph)

$$\mathrm{E}(N | X_m = \ell) = \frac{m-1}{m-2}(\ell - 1), \quad \mathrm{Var}(N | X_m = \ell) = \frac{(m-1)(\ell-1)(\ell-m-1)}{(m-3)(m-2)^2} .$$

The estimate and the mean square error differ slightly from the frequentist approach but they are of the same order, e.g. for the sample $\{23, 2, 97, 128\}$ the frequentist approach gives the estimate 159 while the Bayesian estimate reads 190.5.

We conclude with the explicit calculation of the conditional pmf and the related expectation and variance. From the identity

$$\begin{aligned} \frac{1}{\binom{n}{m}} - \frac{1}{\binom{n+1}{m}} &= \frac{m!(n-m)!}{n!} - \frac{m!(n+1-m)!}{(n+1)!} = \frac{m!}{(n+1)!}(n-m)!(n+1-(n+1-m)) \\ &= \frac{m}{m+1} \frac{(m+1)!(n-m)!}{(n+1)!} = \frac{m}{m+1} \frac{1}{\binom{n+1}{m+1}} \end{aligned}$$

we obtain

$$\binom{n}{m}^{-1} = \frac{m}{m-1} \left(\binom{n-1}{m-1}^{-1} - \binom{n}{m-1}^{-1} \right) .$$

Hence the denominator of the posterior pmf evaluates as

$$\begin{aligned}\sum_{\nu=\ell}^{\Omega}\binom{\nu}{m}^{-1} &= \frac{m}{m-1}\sum_{\nu=\ell}^{\Omega}\left(\binom{\nu-1}{m-1}^{-1} - \binom{\nu}{m-1}^{-1}\right)\\ &= \frac{m}{m-1}\left(\binom{\ell-1}{m-1}^{-1} - \binom{\Omega-1}{m-1}^{-1}\right).\end{aligned}$$

The series converges, i.e., the limit $\Omega \to \infty$ exists as the last term vanishes in this limit. That means the posterior pmf does not depend much on the prior for sufficiently large values of Ω. In that limit the posterior pmf reads

$$P(N=n|X_m=\ell) = \frac{\binom{n}{m}^{-1}}{\frac{m}{m-1}\binom{\ell-1}{m-1}^{-1}} = \frac{m-1}{\ell}\frac{\binom{\ell}{m}}{\binom{n}{m}}\,.$$

The computation of the (conditional) expectation follows the lines of the previous identities

$$\begin{aligned}\mathrm{E}(N|X_m=\ell) &= \frac{m-1}{\ell}\binom{\ell}{m}\sum_{n=\ell}^{\infty}\frac{n}{\binom{n}{m}} = \frac{m-1}{\ell}\binom{\ell}{m}\sum_{n=\ell}^{\infty}\frac{m}{\binom{n-1}{m-1}}\\ &= \frac{m-1}{\ell}\binom{\ell}{m}m\frac{m-1}{m-2}\binom{\ell-2}{m-2}^{-1} = \frac{m-1}{m-2}(\ell-1)\,.\end{aligned}$$

To evaluate the variance one first computes the factorial moment

$$\begin{aligned}\mathrm{E}(N(N-1)|X_m=\ell) &= \frac{m-1}{\ell}\binom{\ell}{m}\sum_{n=\ell}^{\infty}\frac{n(n-1)}{\binom{n}{m}} = \frac{m-1}{\ell}\binom{\ell}{m}\sum_{n=\ell}^{\infty}\frac{m(m-1)}{\binom{n-2}{m-2}}\\ &= \frac{m-1}{\ell}\binom{\ell}{m}m(m-1)\frac{m-2}{m-3}\binom{\ell-3}{m-3}^{-1} = \frac{m-1}{m-3}(\ell-1)(\ell-2)\,.\end{aligned}$$

The variance then follows as

$$\mathrm{Var}(N|X_m=\ell) = \mathrm{E}(N(N-1)|X_m=\ell)+\mathrm{E}(N|X_m=\ell)-(\mathrm{E}(N|X_m=\ell))^2 = \frac{(m-1)(\ell-1)(\ell-m+1)}{(m-3)(m-2)^2}\,.$$

Hypothesis test design: We can formulate the German tank problem as a hypothesis test. Want to test the null hypothesis that the population has a given size N_0 (say $N_0 = 150$), $H_0 : N = N_0$. It could be quite severe if we underestimate N while overestimation may not be such a big issue (at least if you think about tanks in a battle). Hence it may be sensible to test the null hypothesis against the alternative $H_1 : N > N_0$. We will base our test on the statistic X_m, i.e. on the largest value in our sample of size m. To find evidence against H_0 (and in favour of H_1) we use the criterion $X_m > \delta$ (large values for X_m are evidence that the actual population size N exceeds N_0) where the integer value δ is the threshold we need to determine.

If we take δ to be N_0, i.e. the test criterion $X_m > N_0$ (to reject H_0) then any sample which is rejected certainly comes from a population with $N > N_0$. Hence this threshold avoids any type-I errors. The significance level (assuming H_0 to be valid) $\alpha = \mathbb{P}(X_m > N_0) = 0$ vanishes. However this test may produce massive type-II errors, as samples coming from a population which violate H_0 (in favour of H_1), i.e. coming from a population with $N > N_0$ may pass the test (if the maximal value of the sample does not exceed N_0). Hence the test has no real power. To design a test with more power, i.e., with smaller probability for type-II errors, requires small values of δ. But if the threshold δ is too small then the test $X_m > \delta$ will reject many samples even if H_0 is valid (i.e. the type-I error increases). Hence the design of the test consists in balancing type-I and type-II errors.

Significance level: Assume that the null hypothesis $H_0 : N = N_0$ is valid. We test against the alternative $H_1 : N > N_0$ with the criterion $X_m > \delta$. The sampling distribution of the test statistics, i.e., the pmf of X_m is given by (see above)

$$\mathbb{P}(X_m = \ell) = \frac{\binom{\ell-1}{m-1}}{\binom{N_0}{m}}, \quad (m \leq \ell \leq N_0)\,.$$

The significance level of the test $X_m > \delta$ is defined by

$$\begin{aligned} \alpha &= \mathbb{P}(X_m > \delta) = \sum_{\ell=\delta+1}^{N_0} \mathbb{P}(X_m = \ell) = \frac{\sum_{\ell=\delta+1}^{N_0} \binom{\ell-1}{m-1}}{\binom{N_0}{m}} = \frac{\sum_{\ell=\delta+1}^{N_0} \left(\binom{\ell}{m} - \binom{\ell-1}{m}\right)}{\binom{N_0}{m}} \\ &= \frac{\binom{N_0}{m} - \binom{\delta}{m}}{\binom{N_0}{m}} = 1 - \frac{\binom{\delta}{m}}{\binom{N_0}{m}}\,. \end{aligned}$$

As usual we obtain a relation between the significance level α and the threshold δ (the sample size m is given by the test design, and N_0 is given by the null hypothesis). The formula cannot be solved easily for δ but the dependence can e.g. be conveniently given in a table (say for $m = 4$ and $N_0 = 150$):

δ	140	141	142	143	144	145	146	147	148	149	150
α	0.243	0.221	0.199	0.176	0.152	0.128	0.103	0.078	0.053	0.027	0

The significance level is a decreasing function of the threshold. For instance, the test $X_m > \delta$ (for a sample of size $m = 4$) with $\delta = 146$ would reject the null hypothesis $N = 150$ (in favour of $N > 150$) with significance level 10.3% .

Power: To discuss the power of the test we need to specify the alternative hypothesis. Assume the test should be able to distinguish between $N = N_0$ (the null hypothesis) and

$H_1 : N = N_1$ (say for instance $N_1 = 1.5N_0$ if the test should be able to detect an excess of 50%). Assume H_1 to be valid, i.e. $N = N_1$. The sampling distribution of the statistics X_m reads

$$\mathbb{P}(X_m = \ell) = \frac{\binom{\ell-1}{m-1}}{\binom{N_1}{m}}, \quad (m \leq \ell \leq N_1)\,.$$

The probability of a type-II error (i.e. the probability that the test $X_m > \delta$ does not reject the null hypothesis, even though the null hypothesis is not valid) is

$$\beta = \mathbb{P}(X_m \leq \delta) = 1 - \mathbb{P}(X_m > \delta) = 1 - \left(1 - \frac{\binom{\delta}{m}}{\binom{N_1}{m}}\right) = \frac{\binom{\delta}{m}}{\binom{N_1}{m}}$$

if we use the result for the cumulative distribution function of the previous paragraph. The power of the test is given by $1 - \beta$. Again the threshold δ determines the power and the dependence can be easily illustrated in a table (say for $m = 4$, $N_0 = 150$ and $N_1 = 1.5N_0 = 225$):

δ	140	141	142	143	144	145	146	147	148	149	150
β	0.147	0.152	0.156	0.161	0.165	0.170	0.175	0.180	0.186	0.190	0.195
$1 - \beta$	0.853	0.848	0.844	0.839	0.835	0.830	0.825	0.820	0.814	0.810	0.805

The power is a decreasing function of the threshold, i.e. the test becomes less powerful if the threshold becomes larger. Comparing both tables (the probability of type-I errors, α, and the probability of type-II errors, β) the threshold $\delta = 144$ seems to be a good compromise to minimise both types of errors. So with the sample $\{23, 2, 97, 128\}$ the test $X_m > \delta = 144$ does not reject the null hypothesis $N = N_0 = 150$ at level of significance 15.2%. If one aims at better test results (i.e. smaller type-I and smaller type-II errors) one needs samples of larger size.