MAT14303 - Basic Statistics

Lecture Notes

dr. Sabine K. Schnabel, and dr.*ing.* Maikel P.H. Verouden

2026-01-22

Preface

The **Basic Statistics** course (MAT14303) provides an introduction to statistics. **Basic Statistics** deals with the subject matter covered by A-level Mathematics in Dutch pre-university education, as also taught in **MAT15303 Statistics 1** at Wageningen University & Research, as well as the subject matter covered in **MAT15403 Statistics 2**. **Basic Statistics** familiarizes students with notions such as population and sample, simple probability theory, probability distributions, expectation and variance, hypothesis testing, and simple linear regression.

The **Basic Statistics** course prepares students for the continuation courses: Advanced Statistics (AS, MAT20306), Advanced Statistics for Nutritionists (ASN, MAT24306), or Data Analysis for Biosystems Engineering (DABE, MAT26806). These continuation courses deal in particular with linear models: multiple linear regression, analysis of variance, and analysis of covariance. Also they also discuss the analysis of binary data, contingency tables, and the use of methods based on ranks.

Students, who are interested in statistics, can subsequently take the courses "Data Analysis for Plant and Animal Breeding" (ABG30806), "Statistics for Data Scientists" (MAT32806), "Data Science for Plant Breeding and Genetics" (MAT33306) and/or "Bayesian Data Analysis" (MAT34806) as a follow-up. All these courses are presented in English and build upon AS, ASN, DABE. For example in ABG30806, logistic and log linear models (for binary and count data), variance components models (for dependent data, both balanced and unbalanced), general inference techniques (maximum likelihood estimation, Wald test, likelihood ratio test), posterior Bayesian inference using Markov Chain Monte Carlo, and applications in genetics and epidemiology are discussed.

The aforementioned courses are organized by or in cooperation with Biometris. Being a business unit within Wageningen Plant Research and fully integrated with the Mathematical and Statistical Methods Group, Biometris () is a part of Wageningen University & Research and is also responsible for a number of short courses for PhD students and participates in many other courses. Biometris is a partner in the M.Sc. programme Statistics and Data Science at Leiden University. Staff members from Biometris are actively involved in research, mainly in statistical genetics, systems biology and ecology, food health and safety, omics, and big data. We have, however, an interest in other topics as well.

Contents

General Introduction

Structure of these Lecture Notes

These are the Lecture Notes to the **Basic Statistics** course MAT14303. The teaching methods used are tutorials and computer practicals. At the tutorials, a lecturer will explain the subject matter. During and after each tutorial, students are expected to do the exercises as indicated in these Lecture Notes. Feedback on these exercises is provided in Brightspace. At the computer practicals, statistical calculations are carried out with the help of a computer. These Lecture Notes provide an overview of the subjects covered during the tutorials and computer practicals. They also tell you how to prepare for each session.

The information provided for each tutorial gives an overview about the recommended reading as well as the main concepts covered. For some tutorials examples are listed. The exercises for after the tutorials are also mentioned. The recommended reading refers to *An Introduction to Statistical Methods and Data Analysis* by R. Lyman Ott and Michael Longnecker (hereafter abbreviated as 'O&L'). This is the same textbook as used in the **Statistics 1**, **Statistics 2** and the continuation course **Advanced Statistics** and its variations.

The O&L page numbers, mentioned in these Lecture Notes, refer to the 7^{th} edition of this textbook; page numbers referring to its 6^{th} edition are also given. At the end of these Lecture Notes you will find the tables of the binomial distributions, instructions for the use of graphing calculators as well as other annexes.

Educational aids

- Book: *An Introduction to Statistical Methods and Data Analysis* by R. Lyman Ott and Michael Longnecker, 7^{th} edition, CENGAGE Learning. The book is available from Acco in printed version. There is also an online version available (not admitted to the exam). The 6^{th} edition, Brooks/Cole, may also be used.
- MAT14303 Basic Statistics Lecture Notes
- PowerPoint presentations of the tutorials (available as handouts via Brightspace, see below)
- Calculator: a simple pocket calculator or a graphing calculator. However, a graphing calculator is not essential, neither at tutorials or practicals nor at the exam.

Brightspace site of MAT14303

On Brightspace, you will find, among other things, the answers to exercises treated during tutorials, handouts of the PowerPoint presentations given at the tutorials, clips, and other material where applicable.

The easiest way to access this site is usually from your personal myWURtoday page or app. You need to have an account and to be registered for the course. The URL for Brightspace is https://brightspace.wur.nl/

Learning outcomes

The learning outcomes describe what a student is expected to know and to be able to do. They specify in detail what is expected of the students, and therefore what subject matter they may get questions about at the exam. The learning outcomes can be used by students as a checklist to check whether all subject matter relevant to Basic Statistics has been understood and processed. The detailed tutorial learning objectives can be found at the beginning of each respective tutorial section.

In general, the learning objectives of Basic Statistics can be described as:

- remember and understand basic ideas of statistical inference and data collection
- determine and explain the appropriate statistical procedure, given the description of the experiment, the research question, and the type of data
- carry out the needed analyses for the discussed standard situations and assess the results in terms of the problem
- perform a hypothesis test for intercept and slope and validate the model assumptions of a simple linear model
- independently analyze data with the computer software R Commander

Procedure of the tutorials

The purpose of the tutorials is:

a. to introduce the topic with a practical application,
b. to explain the relevant subject matter, and
c. to put the subject matter in perspective and connect it to the material in other tutorials.

Self-study, clips and exercises

For each tutorial, the Lecture Notes indicates what students need to do in terms of self-study (reading in the O&L textbook and in these Lecture Notes) and watching knowledge clips on Brightspace before and/or after the tutorial. The Lecture Notes also indicate which exercises the students are expected to do.

Students can raise questions at the question hours that are organized during the teaching period. See Brightspace for the schedule.

Compulsory computer practicals

Attendance of the practicals is compulsory. The reason for this is that this is the only way for us to make sure that you will acquire some basic statistical computer skills. You will not be allowed to use a computer for calculations at the exam – only tables and/or an ordinary or graphing calculator are allowed.

Your attendance of the computer practicals will be registered. Apart from attending them, you will also be required to participate actively in the practicals and to follow directions correctly; your practical teacher will assess your involvement and performance.

You will work on exercises in a fixed 'team' of two persons, as far as possible. As a team, you shall save all edited files, preferably on your own OneDrive; your teacher can ask you to show these files.

Assessment and written examination

a. The assessment is based on a written exam and conditional on a pass for the computer practicals.

b. The examination lasts 3 hours and consists of multiple-choice questions as well as open questions about the subject matter of the tutorials and computer practicals. The only aids allowed at the exam are:

   i) the textbook by Ott and Longnecker, and
   ii) the **Basic Statistics** Lecture Notes, provided that the latter does not include any complete computations and/or answers to exercises.
   iii) Also allowed is a ***self-written*** summary of not more than one double-sided A4 sheet of paper. Writing a summary is recommended, as it forces you to reduce the subject matter to its most important elements. We consider this a great help when preparing for the exam. Therefore, ***only self-written summaries*** may be used. To facilitate checks on this, ***only hand-written summaries*** are allowed. Summaries produced with the help of a text editor and photocopies of summaries are not allowed. ***Your summary shall not include any complete or partial computations.***
   iv) You will need to bring a calculator to use at the examination.

c. After the examination, you will find on Brightspace (https://Brightspace.wur.nl) under MAT14303 information on an opportunity for you to inspect your marked work. This inspection is primarily meant

as an opportunity for students who failed the examination to learn which elements require special attention when preparing for re-examination.

d. You will get a pass for the computer practical only if, according to your teacher, you have attended and actively participated in all 6 practicals. The mark granted for the final examination will only be valid once your computer practical, the attendance of which is compulsory, has resulted in a pass.

e. If you failed the computer practical but passed the written examination, you will be awarded an 'incomplete' as mark; this mark will remain valid for 2 years. If you failed the written examination, you will be awarded an ordinary mark, no matter whether you passed the computer practical or not.

Coordinators

Name	Building/Floor	E-mail
dr. S. (Sabine) K. Schnabel	Radix West (107) / 4^{th} floor	sabine.schnabel@wur.nl
dr. M. (Maikel) P.H. Verouden	Radix West (107) / 4^{th} floor	maikel.verouden@wur.nl

Part I

Tutorials

Tutorial 1

Learning objectives

After this tutorial the student should be able to:

- identify: population, sample, unit, variables (quantitative: discrete, continuous; qualitative: nominal, ordinal);
- recognize and interpret a bar chart and a histogram;
- mention and draw (by hand) an appropriate plot for a given variable;
- interpret and construct a frequency table and a relative frequency table;
- interpret and calculate cumulative frequencies;
- choose the correct measure for central tendency for a given variable;
- determine and interpret the mode, median and mean.

Important concepts

Read for an introduction to the important concepts of population, sample, unit and variable:

☐ **O&L 7th Edition**:

- paragraphs 1.1 and 1.2 pp.2-9, and
- paragraph 4.6 pp.164-166, **or**

☐ **O&L 6th Edition:**

- paragraphs 1.1, and 1.2 pp.2-8,
- paragraph 4.6 pp.155-157.

Descriptive analysis for one variable: visualization

Read:

☐ **O&L 7th Edition**:

- paragraphs 3.1 and 3.2 pp.60-66, and
- paragraph 3.3 pp.66-75 (up to stem-and-leaf plot), **or**

☐ **O&L 6th Edition**:

- paragraphs 3.1, and 3.2 pp.56-62, and
- paragraph 3.3 pp.62-72 (first 9 lines).

Two different graphical representations for a single variable are discussed: the bar chart and the histogram.

> A *bar chart* or *bar plot* is applied for categorical or discrete data.
>
> A bar chart displays the count for each distinct category or value as a separate bar, allowing you to compare categories visually.

There are small gaps between the bars. They indicate that the data is categorical or discrete. There are many variations of the bar chart.

Example 1.1: college majors

University officials periodically review the distribution of undergraduate majors within the colleges of the university to help determine a fair allocation of resources to departments within the colleges. At one review, the following data were obtained (see Table 2), which were presented in a bar chart as shown in Figure 1.

Table 2: Number of majors per college with collegename abbreviation (colAbbrev).

college	noMajors	colAbbrev
Agriculture	1500	Agric.
Arts and Sciences	11000	ArtsandSc.
Business Administration	7000	BusAdm.
Education	2000	Educ.
Engineering	5000	Engin.

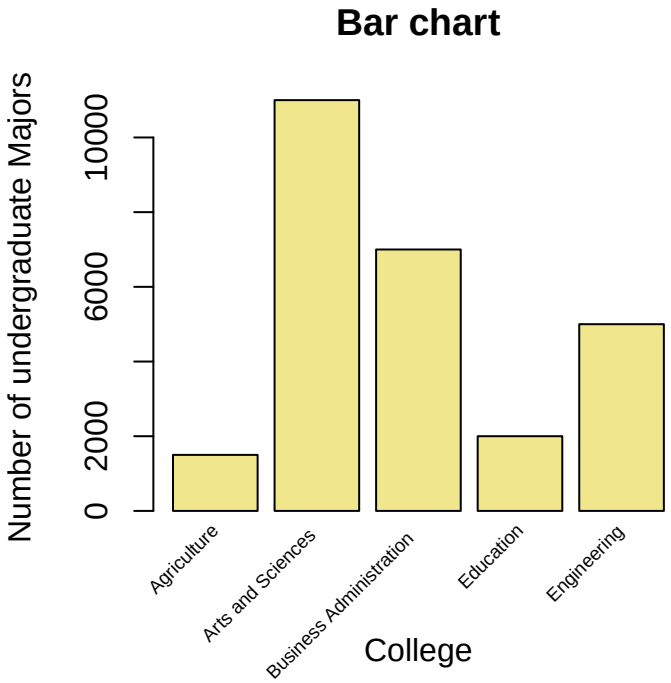


Figure 1: Bar chart of number of undergraduate majors per college.

Note: In newspapers and non-scientific journals, data like these are often presented in a so-called pie chart (see Figure 2). However, in scientific papers bar charts are preferred, because they are often more clear.

A *histogram* can be used for continuous data and for discrete data with many different outcomes.

It displays the (relative) frequency of the data. The range is subdivided into classes, usually of equal width. The height of each rectangle (bar) corresponds to the count of values of the variable falling within the interval.

! Remarks about histograms.

- A histogram also has rectangles but now these cover the full class interval without gaps in between; the rectangles are plotted along an interval scale.
- A histogram shows the shape, center, and spread of the distribution. The choice of class width or the number of classes can heavily influence the shape/impression of the histogram.

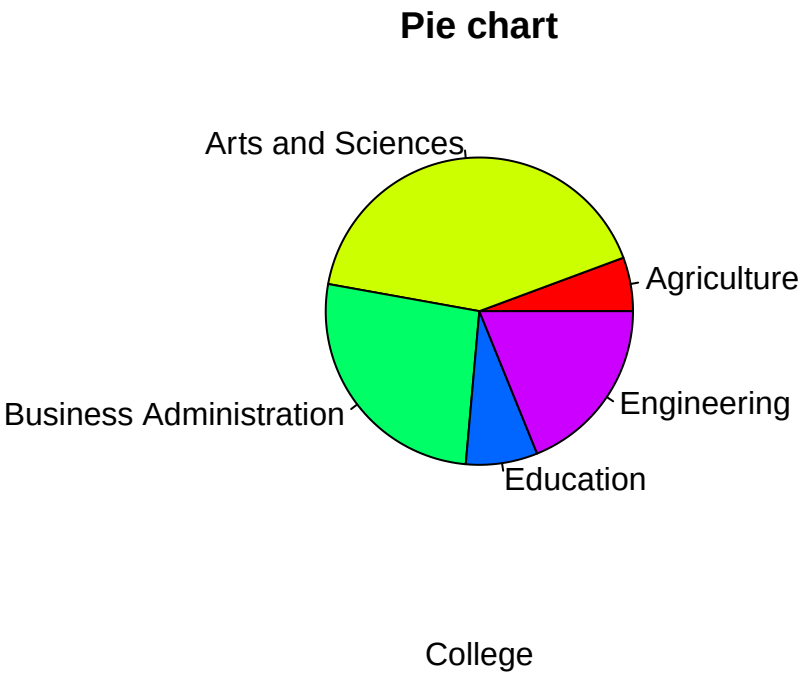


Figure 2: Pie chart of number of undergraduate majors per college.

- Also for a discrete variable with many distinct outcomes, measured in classes (by approximation a continuous variable), a histogram may be suitable.

Table 3: Area of cultivation under glass for 33 growers.

53	39	73	98	49	50	42	63	61	63	19
30	39	100	30	30	20	20	40	59	25	22
44	25	22	24	36	49	39	35	29	43	31

Example 1.2: Cultivation in greenhouses

A researcher did a small study about the cultivation under glass. He asked 33 growers the area of cultivation under glass. The results are shown in Table 3.

Most statistical software programs, like R, will make classes automatically, when creating a histogram (see Figure 3). R has chosen classes with a width of 10 units. In the histogram, you can see that the sample distribution is skewed to the right. The two largest observations could be called outliers (or extreme values).

Descriptive analysis for one variable: measures of central tendency

Read:

- ☐ **O&L 7th Edition**:
  - paragraph 3.4 pp.82-90 (skip grouped data median and Example 3.4), **or**
- ☐ **O&L 6th Edition**:
  - paragraphs 3.4 pp.78-85 (skip grouped data median and Example 3.4).

Measures of central tendency for a sample are discussed: the **mode**, the **median**, and the **(*arithmetic*) mean**.

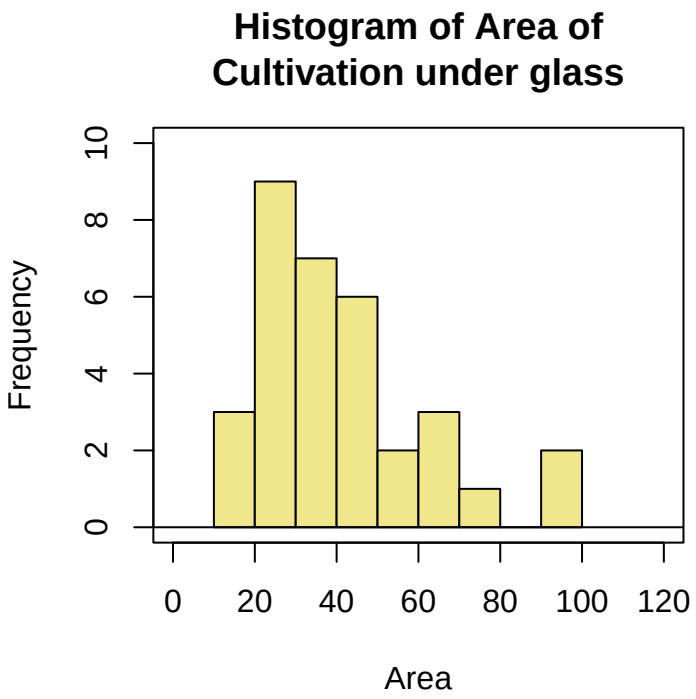


Figure 3: Histogram of area of cultivation under glass.

The *mode* of a set of measurements is the measurement value with the highest frequency.

The *median* of a set of measurements is the middle value when the measurements are arranged from lowest to highest. Additional rules are needed to determine the median in the case of an even number of discrete observations.

The *(arithmetic) mean* $\bar{y}$ of a set of measurements y is the sum of the measurements divided by the total number of measurements.

Example 1.3: Number of plots

A researcher registers the number of plots (variable y) from 55 farmers, having companies of nearly the same size. The results are given in Table 4, and the corresponding bar chart is displayed in Figure 4.

Table 4: Frequency of the number of plots per farmer (total of 55 farmers).

Number of plots	Frequency
1	3
2	5
3	5
4	7
5	9
6	7
7	8
8	4
9	3
10	2
11	0
12	2

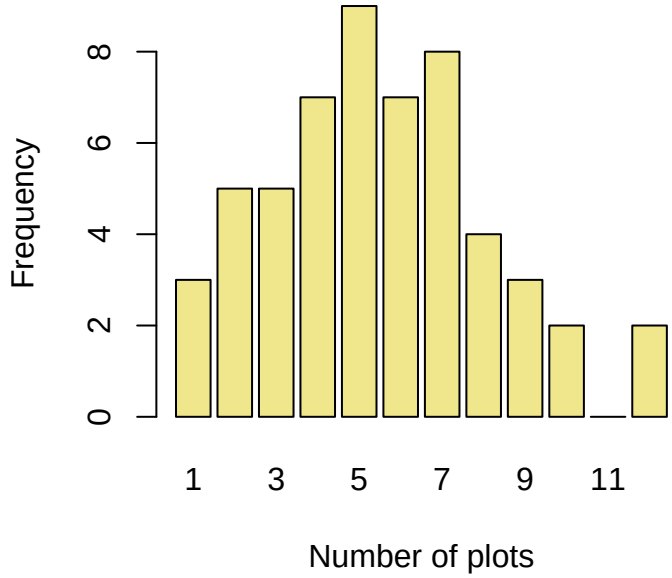


Figure 4: Bar chart of frequencies for number of plots per farmer.

The mode equals 5.
The median is the 28^{th} observation. Therefore, the median is equal to 5.
The mean is $\bar{y} = (3 \times 1 + 5 \times 2 + \ldots + 2 \times 12) \ / \ 55 \approx 5.491$.

R and R Commander can both provide a convenient summary of the data, using the shown commands.

```
# R
summary(object = farmers_count$plots)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   1.000   4.000   5.000   5.491   7.000  12.000
```

```
# R Commander
numSummary(data = farmers_count$plots)
```

```
#>      mean       sd IQR 0% 25% 50% 75% 100%  n
#>  5.490909 2.64486   3  1   4   5   7   12 55
```

Exercises to be done during the tutorial

Exercise 1.1 up to and including **Exercise 1.5** are in the presentation handouts of Tutorial 1. For answers/feedback check Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 1.6

Do either

- ☐ **Exercise 3.15 O&L 7^{th} Edition p.130**, or
- ☐ **Exercise 3.15 O&L 6^{th} Edition p.122**.

Exercise 1.7

Do either

- ☐ **Exercise 3.16 O&L 7^{th} Edition p.130**, or
- ☐ **Exercise 3.16 O&L 6^{th} Edition p.122**.

Tutorial 2

Learning objectives

After this tutorial the student should be able to:

- name the correct measure for variability given the variable;
- determine and interpret the value for the range, interquartile range, percentiles, variance, and standard deviation;
- interpret a box plot, side-by-side box plots, and a scatter plot;
- choose the correct graph to use (box plot, side-by-side box plots, or scatter plot) given the data.

Descriptive analysis for one variable: measures of variability

Read:

☐ **O&L 7th Edition**:

- paragraph 3.5 pp.90-98 (up to Example 3.10, skip on p.91 'grouped data' and pp.92 'The computation of percentiles:. . .'-95 'Example 3.8'), **or**

☐ **O&L 6th Edition**:

- paragraph 3.5 pp.85-93 (up to Example 3.10, skip on p.87 'grouped data' and pp.87 'The computation of percentiles:. . .'-90 'Example 3.8').

The following measures of variability for a sample are discussed: the range, percentiles, the interquartile range, the variance and the standard deviation.

The *range* of a set of measurements is the difference between the largest and the smallest measurement of the data set.

The *p-th percentile* of a set of n measurements arranged in increasing order is that value that has at most p% of the measurements below it and at most $(100 - p)$% above it.

The 25th percentile is called the *first (lower) quartile* (denoted by Q_1), the 50th percentile is the *median* (second or middle quartile, and sometimes denoted by Q_2) and the 75th percentile is called the *third (upper) quartile* (denoted by Q_3).

The *interquartile range* (denoted by IQR) of a set of measurements is the difference between the third (upper) quartile and the first (lower) quartile.

$$\text{IQR} = Q_3 - Q_1$$

The *variance* s^2 of a set of n measurements $y_1, y_2, \ldots, y_n$ with mean μ is the sum of the squared deviations

of the observations from the sample mean, divided by $n-1$.

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

The *standard deviation* s of a set of n measurements $y_1, y_2, \ldots, y_n$ is the square root of the variance s^2.

$$s = \sqrt{s^2}$$

! Remark about variance and standard deviation.

Some books use for the definition of s^2 and s in the denominator n instead of $n-1$. (Graphing) Calculators usually have both versions, e.g., designated σ_n and σ_{n-1} for s.

Example 2.1 (continuation from Example 1.3)

The *range* equals: $12 - 1 = 11$.

The *first (lower) quartile* is the 14^{th} observation, therefore equal to 4; the *third (upper) quartile* is the 42^{th} observation, therefore equals 7. The interquartile range is then IQR $= 7 - 4 = 3$.

The variance is $s^2 \approx 6.9953$ and the standard deviation is $s \approx \sqrt{6.9953} \approx 2.645$. Use your (graphing) calculator to check this.

Asking for a numerical summary in R or R Commander, provides not only the mean and the median (as shown in Example 1.3) but also the quartiles. The 'numerical summary' function in R Commander *numSummary()* additionally provides the standard deviation,the IQR, and the number of observations n. These can be also be obtained using the following commands in R:

```
sd(x = farmers_count$plots)
```

```
#> [1] 2.64486
```

```
IQR(x = farmers_count$plots)
```

```
#> [1] 3
```

```
quantile(x = farmers_count$plots)
```

```
#>   0%  25%  50%  75% 100%
#>    1    4    5    7   12
```

```
length(x = farmers_count$plots)
```

```
#> [1] 55
```

Descriptive analysis for one variable: box plot

Read about the box plot:

- ☐ **O&L 7$^{\text{th}}$ Edition**:
  - paragraph 3.6 pp.104-109, **or**
- ☐ **O&L 6$^{\text{th}}$ Edition**:
  - paragraph 3.6 pp.97-102.

The *box plot* (also called *box-and-whiskers plot*) is a summary plot of a quantitative variable based on the median, quartiles, and extreme values.

The box represents the interquartile range, which contains 50% of the values. The whiskers are lines that

extend from the box to the highest and lowest values, when there are no outliers (extreme values) in the data set. A bold line inside the box represents the median.

Example 2.2 (continuation from Example 1.2)

A researcher did a small study about the cultivation under glass. He asked 33 growers the area of cultivation under glass. The results are shown in Table 3.

The corresponding box plot, constructed in R and displayed in Figure 5, shows that the sample distribution is skewed to the right and that 2 outlying observations (values 98 and 100) are present.

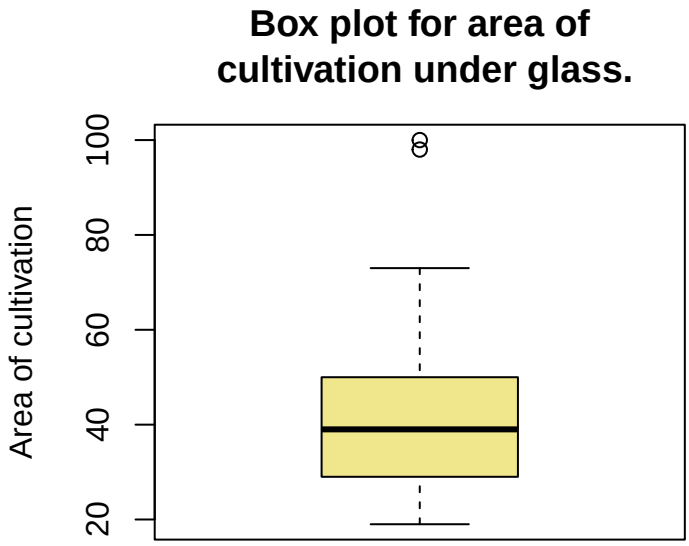


Figure 5: Box plot for area of cultivation under glass.

! Remark about box plots.

Box plots should not be created for small sample sizes.

Descriptive analysis for two quantitative variables: scatter plot

Read about the scatter plot:

- ☐ **O&L 7th Edition**:
  - paragraph 3.7 pp.111 (last two lines above Table 3.15)-113 (first 13 lines), **or**
- ☐ **O&L 6th Edition**:
  - paragraph 3.7 pp.104 (last two lines)-106 (first 18 lines).

Descriptive analysis for one qualitative and one quantitative variable: side-by-side box plots

Read about the side-by-side box plots:

- ☐ **O&L 7th Edition**:
  - paragraph 3.7 pp.115 (last three lines)-119 (skip Figure 3.31 p.116), **or**
- ☐ **O&L 6th Edition**:
  - paragraph 3.7 pp.108 (last ten lines)-112 (skip Figure 3.31 p.109).

Exercises to be done during the tutorial

Exercise 2.1 up to and including **Exercise 2.4** are in the presentation handouts of Tutorial 2. For answers/feedback check Brightspace.

Post-class activity

Watch:

- ☐ '*About the box plot*' (Duration: 2:12 min.)

The clip is linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 2.5

A researcher registers the number of plots (variable y) from 55 farmers having companies of nearly the same size. The results are shown in Table 4.

The R Commander numerical summary output for the number of plots for 55 farmers is given in Table 5.

Table 5: Numerical summary for the number of plots for 55 farmers.

mean	sd	IQR	0%	25%	50%	75%	100%	n
5.490909	2.64486	3	1	4	5	7	12	55

The default R summary output for the number of plots for 55 farmers is:

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   1.000   4.000   5.000   5.491   7.000  12.000
```

a. Use the R Commander numerical summary output in Table 5 to determine the range and the IQR. Check your answer with respect to the IQR with the default R summary output provided above.

b. Find the median in the default R summary output provided above and check your answer with the R Commander numerical summary output in Table 5.

c. Check, using the R Commander numerical summary output in Table 5 and a (graphing) calculator, that the variance $s^2 \approx 6.995$.

Exercise 2.6

Use the data from either

- ☐ **Exercise 3.15 O&L 7th Edition p.130**, or
- ☐ **Exercise 3.15 O&L 6th Edition p.122**.

to calculate the range, interquartile range, variance, and standard deviation.

Exercise 2.7

Do either

- ☐ **Exercise 3.16 O&L 7th Edition p.130**, or
- ☐ **Exercise 3.16 O&L 6th Edition p.122**.

Replace the two data points as mentioned in the exercise. What will be the result from these changes with respect to the range, interquartile range, variance, and standard deviation? Do not recalculate the requested measures, but provide an argumentation.

Exercise 2.8

Do either

- ☐ **Exercise 3.29 O&L 7th Edition p.133, or**
- ☐ **Exercise 3.29 O&L 6th Edition p.125**.

This exercise is slightly modified compared to the version presented in O&L, use the ordered data as provided in Table 6.

The treatment times (in minutes) for patients at a health clinic are as follows:

Table 6: Treatment times at a health clinic for 50 patients.

7	11	13	15	17	20	21	24	28	32
8	12	13	16	17	20	22	24	29	33
10	12	14	16	18	21	22	24	29	35
10	12	15	16	18	21	24	26	29	45
11	12	15	16	19	21	24	27	31	54

Exercise 2.9

Do either

- ☐ **Exercise 3.34b O&L 7th Edition p.135**, or
- ☐ **Exercise 3.34b O&L 6th Edition p.126**.

Exercise 2.10

Do either

- ☐ **Exercise 3.35 O&L 7th Edition p.135**, or
- ☐ **Exercise 3.35 O&L 6th Edition p.127**.

Tutorial 3

Learning objectives

After this tutorial the student should be able to:

- mention, explain and apply the three probability properties, the general addition rule, and the product rule for independent events;
- recognize a random variable;
- distinguish between a discrete probability distribution and a continuous probability distribution;
- recall the properties of a binomial experiment;
- mention and apply the formula of the expected outcome, the variance, and the standard deviation of a binomial distribution;
- determine the probability of an event resulting from a binomial experiment using a binomial distribution table.

Probability laws

Read:

☐ **O&L 7th Edition**:

- paragraph 4.1 pp.149-152 up to the abstract of the research question, and
- paragraph 4.3 pp.155-158, **or**

☐ **O&L 6th Edition**:

- paragraph 4.1 pp.140–143 up to the abstract of the research question, note: $P(\text{even } E)$ should read $P(\text{event } E)$ on p. 142, and
- paragraph 4.3 pp.146-149.

The three *probability properties* are:

- $P(S) = 1$ where S is the sample space (set of all possible outcomes)
- $0 \leq P(A) \leq 1$ for any event A
- The complement $\bar{A}$ of an event consists of all outcomes that are not occurring in A implying $P(A) = 1 - P(\bar{A})$ (complement rule).

Two rules for probabilities are:

- the general addition rule (in O&L referred to as 'probability of union'):
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  with the special case of mutually exclusive events A and B:
  $P(A \cup B) = P(A) + P(B)$.
- the product rule for independent events: $P(A \cap B) = P(A) \times P(B)$

Random variables

(Re-)Read:

- ☐ **O&L 7th Edition**:
  - paragraph 4.6 pp.164-166, **or**
- ☐ **O&L 6th Edition**:
  - paragraph 4.6 pp.155–157,

where the definition of a *random variable* is given. A random variable can be discrete or continuous.

Random variables: discrete random variables

Read:

- ☐ **O&L 7th Edition**:
  - paragraph 4.7 pp.166-167, **or**
- ☐ **O&L 6th Edition**:
  - paragraph 4.7 pp.157–158,

where the term *probability distribution* is discussed. This is also called *probability mass function*.

Discrete probability distributions: the binomial distribution

Read:

- ☐ **O&L 7th Edition**:
  - paragraph 4.8 pp.167-175 up to Poisson distribution, **or**
- ☐ **O&L 6th Edition**:
  - paragraph 4.8 pp.158–166 up to Poisson distribution,

using the *binomial distribution* as an example of a discrete probability distribution.

The *binomial distribution* with parameters n and π:

- Assumptions for a binomial experiment:
  - n identical trials, where n denotes the sample size;
  - the n trials are independent of each other;
  - each trial results in one of two possible outcomes (one labelled as a success, and the other as a failure);
  - the probability of success in a single trial is denoted by π, and the value of π is the same from trial to trial (constant);
  - the random variable y denotes the number of successes in n trials.
- Notation: $y \sim \text{Bin}(n, \pi)$ or $y \sim \text{B}(n, \pi)$
- Expected value: $\text{E}(y) = \mu_y = n \times \pi$
- Variance: $\text{var}(y) = \sigma_y^2 = n \times \pi \times (1 - \pi)$ and standard deviation $\sigma(y) = \sqrt{n \times \pi \times (1 - \pi)}$
- Probability distribution: $P(y = k) = \frac{n!}{k! \times (n-k)!} \times \pi^k \times (1 - \pi)^{n-k} \; \forall \; k \in \{0, 1, 2, \ldots, n\}$

! Remarks about binomial probability calculation and shape of a binomial distribution.

- $n!$ is the notation for the factorial of n, i.e., $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1 = n \times (n-1)!$. For example: $4! = 4 \times 3 \times 2 \times 1 = 24$
- For $\pi = 0.5$ the binomial distribution is symmetric, see Figure 6 for a plot of an example.

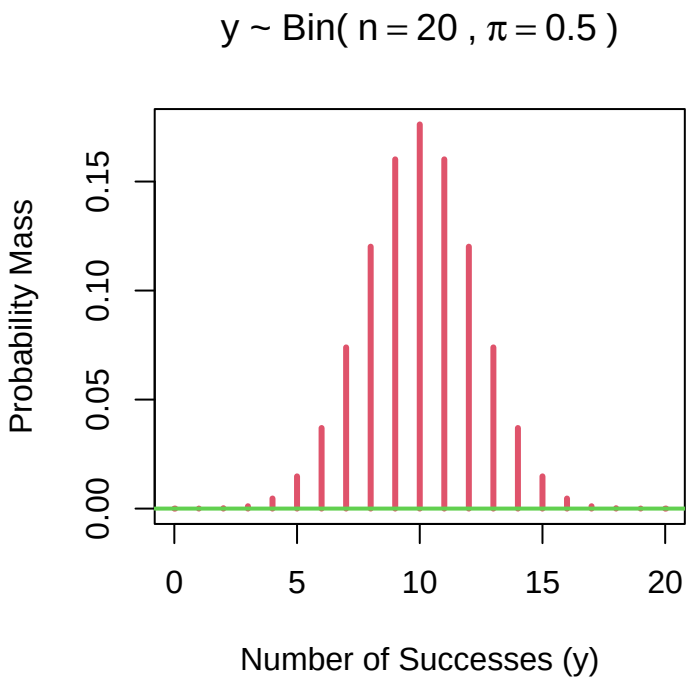


Figure 6: Binomial distribution for $n = 20$ and $\pi = 0.5$

Exercises to be done during the tutorial

Exercise 3.1 up to and including **Exercise 3.4** are in the presentation handouts of Tutorial 3. For answers/feedback check Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 3.5

Do either

- ☐ **Exercise 4.24ab O&L 7th Edition p.217**, or
- ☐ **Exercise 4.24 O&L 6th Edition pp.206-207**.

Exercise 3.6

Do either

- ☐ **Exercise 4.25ab O&L 7th Edition p.217**, or
- ☐ **Exercise 4.25a O&L 6th Edition p.207**.

Exercise 3.7

Do either

- ☐ **Exercise 4.26ac O&L 7th Edition pp.217-218**, or
- ☐ **Exercise 4.26ac O&L 6th Edition p.207**.

Exercise 3.8

Do either

- ☐ **Exercise 4.40 O&L 7th Edition p.220**, or
- ☐ **Exercise 4.40 O&L 6th Edition p.209**.

Exercise 3.9

Do either

- ☐ **Exercise 4.44 O&L 7th Edition p.221, or**
- ☐ **Exercise 4.44 O&L 6th Edition p.210**.

Exercise 3.10

Do either

- ☐ **Exercise 4.45 O&L 7th Edition p.221**, or
- ☐ **Exercise 4.45 O&L 6th Edition pp.210-211**.

Exercise 3.11

Do either

- ☐ **Exercise 4.89a O&L 7th Edition p.226**, or
- ☐ **Exercise 4.89a O&L 6th Edition p.216**.

Tutorial 4

Learning objectives

After this tutorial the student should be able to:

- recognize a continuous probability distribution (i.e., probability density function - PDF);
- mention the three properties and the statistical notation of a Normal distribution;
- mention and explain the statistical notation of the Standard Normal distribution;
- mention and apply the formula for the Z-transformation;
- determine the probability for a given z-value using the table of the Standard Normal Distribution;
- determine the z-value for a given probability using the table of the Standard Normal Distribution;
- give the estimators for the population mean μ, the population variance σ^2, and the population standard deviation σ of a continuous random variable;
- explain the concept of bias with respect to the estimators for μ, σ^2, and σ of a continuous random variable;
- determine the estimates for the population mean μ, the population variance σ^2, and the population standard deviation σ given the data;
- mention and apply the rule of thumb for validity of the normal approximation, i.e., the empirical rule;
- interpret a Q-Q plot.

Pre-class activity

Watch:

☐ *'Random variables and probability distributions.'* (Duration: 4:38 min.)

The clip is linked on Brightspace.

The empirical rule

Read:

☐ **O&L 7th Edition**:

- paragraph 3.5 pp.100-103 starting at line 5 on p.100 up to end of example 3.13, **or**

☐ **O&L 6th Edition**:

- paragraph 3.5 pp.93–96 starting just below example 3.10 to the end of example 3.12,

for a discussion of the empirical rule and its application.

The empirical rule

Given a set of n measurements possessing a bell-shaped histogram (representing the density function), then:

- the interval $\bar{y} \pm 1 \times s$ contains approximately 68% of the measurements,
- the interval $\bar{y} \pm 2 \times s$ contains approximately 95% of the measurements,
- the interval $\bar{y} \pm 3 \times s$ contains approximately 99.7% of the measurements.

Random variables: continuous random variables

Read:

☐ **O&L 7th Edition**:

- paragraph 4.9 pp.177-180, **or**

☐ **O&L 6th Edition**:

- paragraph 4.9 pp.168–171,

where the term *probability density function* $f(y)$ for a continuous random variable is discussed.

Read:

☐ **O&L 7th Edition**:

- paragraph 4.10 pp.180-187, **or**

☐ **O&L 6th Edition**:

- paragraph 4.9 pp.171–178,

discussing the most known example of a continuous distribution: the *normal distribution* (also called Gaussian distribution) and the *standard normal distribution*.

The *normal distribution* with parameters μ and σ is denoted as $y \sim \mathrm{N}(\mu,\ \sigma)$.

- Probability density function: $f(y) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$
- $\mathrm{E}(y) = \mu,\ \mathrm{var}(y) = \sigma^2,\ \sqrt{\mathrm{var}(y)} = \sqrt{\sigma^2} = \sigma$
- Variable y follows a *standard normal distribution* , when $\mu_y = 0$ and $\sigma_y = 1$. This is denoted as: $y \sim \mathrm{N}(\mu = 0,\ \sigma = 1)$
- When $y \sim \mathrm{N}(\mu, \sigma)$, then for the standardized variable $z = \frac{y-\mu}{\sigma}$ holds $z \sim \mathrm{N}(\mu = 0,\ \sigma = 1)$.

! Remark about calculating normal probabilities.

For calculating normal probabilities see O&L Table 1 (at the inside of the cover of both editions and:

- **O&L 7th Edition** pp.1086-1087, **or**
- **O&L 6th Edition** pp.1170–1171,

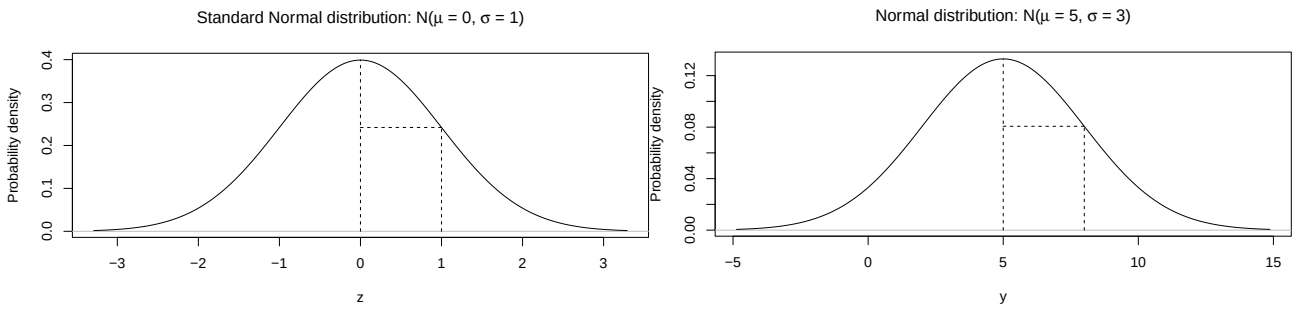
or use a graphing calculator.

The probability density function of the normal distribution is symmetric, see Figure 7a and Figure 7b for examples.

Estimators for the population mean, variance and standard deviation of a continuous variable

This section will introduce a few extra theoretical considerations, which are not included in the textbook.

Population parameters (such as mean and variance) can be measured with certainty only, when all possible outcomes are known, i.e., when the the whole population is known. Usually, when undertaking a study, only



(a) $\mu = 0,\ \sigma = 1$ (b) $\mu = 5,\ \sigma = 3$

Figure 7: Normal distributions, with

a random sample from the population is available. The population parameters are then estimated based on the sample.

As an example, consider the population mean μ of a quantitative variable y. When taking a random sample, the population mean μ is estimated by the sample mean $\bar{y}$. Notation: $\hat{\mu} = \bar{y}$.

A desired property of the estimator of a population parameter is that this estimator is ***unbiased***, which means that the estimate of the population mean (or expected value) is equal to the parameter itself. The estimator $\bar{y}$ is an ***unbiased*** estimator for μ, because $\mathrm{E}(\bar{y}) = \mu$. It is very important to note that $\bar{y}$ is a random variable itself, of which the value will vary from sample to sample even though the samples are taken from the same population.

If you have a random sample of independent observations $y_1, y_2, \ldots, y_n$ with $\mathrm{E}(y) = \mu$ and $\mathrm{var}(y) = \sigma^2$, then:

- $\hat{\mu} = \bar{y}$ is an unbiased estimator of μ.
- $\hat{\sigma}^2 = s^2 = \frac{\sum(y-\bar{y})^2}{n-1}$ is an unbiased estimator of σ^2.
- $\hat{\sigma} = s = \sqrt{s^2}$ is an unbiased estimator of σ.

Please note, that statisticians abuse the usage of the word "mean" for the population mean. The population mean is actually an expected value. From the context can be derived, whether mean refers to a population mean or a sample mean. Anyhow, it is of vital importance for statistical inference to make the distinction between these two means. The sample mean is an unbiased estimator for the population mean (i.e., the expected value).

Checking whether or not a population distribution is normal: Q-Q plot

For checking the normality assumption of observations, a normal probability plot is used, also called Quantile-Quantile plot or Q-Q plot. Observations are plotted vertically against a percentile score on the horizont axis. The horizontal probability scale is non-linear and approximately such that the expected dots for normally distributed data are relatively close to a straight line. The aim of this type of graph is to decide by eye whether (isolated) outliers or extremes occur in the data, and whether the shape (strongly) deviates from the straight line representing normality. R shows a linear quantile scale, hence the name Quantile-Quantile plot or Q-Q plot.

For a more elaborated explanation read:

☐ **O&L 7th Edition**:

- paragraph 4.14 pp.203-206 including lines 6-10 under Figure 4.28, but skipping the part about the R code in lines 1-5, **or**

☐ **O&L 6th Edition**:

- paragraph 4.14 pp.194-197 including lines 1-3 under Figure 4.28,

and watch:

☐ '*Q-Q plots*' (Duration: 5:24 min.) as linked on Brightspace.

Exercises to be done during the tutorial

Exercise 4.1 and **Exercise 4.2** are in the presentation handouts of Tutorial 4. For answers/feedback check Brightspace.

Post-class activity

Watch:

- ☐ '*Random variables and probability distributions.*' (Duration: 4:38 min.), when not done as pre-class activity,
- ☐ '*The Normal Distribution*' (Duration: 6:00 min.),
- ☐ '*Galton Board Demonstration*' (Duration: 2:25 min.),
- ☐ '*The Empirical Rule*' (Duration: 4:26 min.).

All of the clips are linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 4.3

A machine produces packages of coffee. The random variable y is the weight of a randomly selected package of coffee: $\mathrm{E}(y) = \mu$ and $\mathrm{var}(y) = \sigma^2$.

A random sample of 10 packages is taken; the weights (in g) of these packages are: 485, 540, 505, 510, 465, 455, 515, 560, 525, 510

Use the R Commander output in Table 7 to answer the questions below.

Table 7: Numerical summary of the coffee weights.

mean	sd	IQR	0%	25%	50%	75%	100%	n
507	32.0763	32.5	455	490	510	522.5	560	10

a. Give the estimate of μ.

b. Give the estimate of σ.

c. Calculate the estimate of σ^2.

d. Give the median.

e. Although 10 is a very small sample size and we can't be sure, would you expect - based on the weights observed in this sample - that the weight of the packages coffee is normally distributed? Explain your answer.

Exercise 4.4

This exercise uses the slightly modified data of Exercise 3.29 O&L 7^{th} Edition p.133 [O&L 6^{th} Edition p.125]. You can use Table 8 (with the modified ordered data) given below, to answer this question.

The mean and the standard deviation of this sample of treatment times, as given in Table 8, are: $\bar{y} = 20.58$, $s \approx 9.19$.

Table 8: Treatment times at a health clinic for 50 patients.

7	11	13	15	17	20	21	24	28	32
8	12	13	16	17	20	22	24	29	33
10	12	14	16	18	21	22	24	29	35
10	12	15	16	18	21	24	26	29	45
11	12	15	16	19	21	24	27	31	54

a. Check whether the empirical rule applies to this data (i.e., Are 68% of the measurements between $\bar{y} \pm 1 \times s$; Are 95% of the measurements between ...; Are ...?).

b. Is it likely that this sample comes from a normal distribution?

Exercise 4.5

Do either

- ☐ **Exercise 4.65 O&L 7th Edition p.223**, or
- ☐ **Exercise 4.64 O&L 6th Edition p.212**.

Exercise 4.6

Do either

- ☐ **Exercise 4.70 O&L 7th Edition p.223**, or
- ☐ **Exercise 4.70 O&L 6th Edition p.213**.

Tutorial 5

Learning outcomes

After this tutorial the student should be able to:

- explain the concept of the sample distribution for the mean;
- mention and apply the formula (given simple situations) of the expected value, the variance, and the standard deviation of the sample distribution of the mean;
- explain and apply the concept of standard error;
- distinguish the standard error from the standard deviation;
- explain the concept of the sample distribution for the sum;
- mention and apply the formula (given simple situations) of the expected value, the variance, and the standard deviation of the sample distribution of the sum;
- mention and apply the Central Limit Theorem;
- use the Student t-distributions;
- apply the concept of degrees of freedom (denoted by df or ν);
- interpret a 'confidence interval';
- determine the confidence interval for μ of a one sample situation.

Pre-class activity

Watch:

☐ '*Confidence level and Margin of Error*' (Duration: 5:30 min.).

The clip is linked on Brightspace.

Sampling Distribution

Read:

☐ **O&L 7th Edition**:

- paragraph 4.12 pp.190-200, **or**

☐ **O&L 6th Edition**:

- paragraph 4.12 pp.181-190,

where the sampling distribution of the **sample mean** $\bar{y}$ is discussed as well as the role of the Central Limit Theorem. Additionally the book presents the sampling distribution of the **sample sum** $\sum y$.

Rules for the sample mean $\bar{y}$

Let $\bar{y}$ denote the sample mean computed from a random sample of n independent measurements from a population having an expected value (mean) μ and variance σ^2.
It follows that:

- $E(\bar{y}) = \mu_{\bar{y}} = \mu$ and $\text{var}(\bar{y}) = \sigma^2(\bar{y}) = \sigma^2/n$
- The standard deviation $\sigma(\bar{y})$ is also a measure of precision for the estimate of $\bar{y}$ and for that reason is referred to as *standard error of* $\bar{y}$ or $\text{SE}(\bar{y})$.
- $\sigma(\bar{y}) = \text{SE}(\bar{y}) = \sigma/\sqrt{n}$
- When the population distribution is normal, the distribution of $\bar{y}$ is exactly normal.
- When the population distribution is not normal, the Central Limit Theorem says that when n goes to ∞,
  - the distribution of $\bar{y}$ will (in some sense) converge to a normal distribution, and
  - the distribution of the standardized $\bar{y}$ (i.e., $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$) will converge to the standard normal distribution.

! Remark about approximation of a standardized sample mean.

In practice it is often assumed that the distribution of a standardized sample mean is well approximated by a standard normal distribution, when the sample size n is large enough, see:

- **O&L 7th Edition** pp.197-198, or
- **O&L 6th Edition** pp.188–189.

Rules for the sample sum $\sum y$

Let $\sum y$ denote the sample sum computed from a random sample of n independent measurements from a population having an expected value (mean) μ and variance σ^2.
It follows that:

- $E(\sum y) = \mu_{\sum y} = n \times \mu$ and $\text{var}(\sum y) = \sigma^2(\sum y) = n \times \sigma^2 \rightarrow \sigma(\sum y) = \sqrt{n} \times \sigma$.
- When the population distribution is normal, the distribution of $\sum y$ is exactly normal.
- When the population distribution is not normal, the Central Limit Theorem says that when n goes to ∞,
  - the distribution of $\sum y$ will (in some sense) converge to a normal distribution and
  - the distribution of the standardised $\sum y$ (i.e., $\frac{\sum y - n\times\mu}{\sqrt{n}\times\sigma}$) will converge to the standard normal distribution.

Confidence interval for μ (one sample: quantitative continuous random variable y)

Read:

☐ **O&L 7th Edition**:

- paragraphs 5.1 and 5.2 pp.232-240, **or**

☐ **O&L 6th Edition**:

- paragraph 5.1 and 5.2 pp.222-230.

For a random sample with observations $y_1, y_2, \ldots, y_n$ with $E(y) = \mu$ and $\text{var}(y) = \sigma^2$.

Confidence interval for μ

- A confidence interval gives, based on the data of a random sample, all likely values for the unknown value of μ.
- An unbiased estimator for μ is $\hat{\mu} = \bar{y}$ with $\text{SE}(\bar{y}) = \sigma \ / \ \sqrt{n}$.
- In general σ will be unknown and have to be estimated using the sample standard deviation s (see

also Tutorial 4); in that case the standard error for $\bar{y}$ becomes: $\text{SE}(\bar{y}) = s/\sqrt{n}$.
- When sample sizes are large (say $n > 120$) we may assume, based on the Central Limit Theorem, that the sample mean $\bar{y}$ is normally distributed: $\bar{y} \sim \text{N}(\mu, \sigma)$.
  In that case the limits of a $(1-\alpha) \times 100\%$ Confidence Interval for μ are: $\bar{y} \pm z_{\alpha/2} \times s/\sqrt{n}$, with $z_{\alpha/2}$ coming from the standard normal distribution $\text{N}(\mu = 0,\ \sigma = 1)$.
- When sample sizes are small (say $n \leq 120$) we may not assume that the sample mean $\bar{y}$ is normally distributed. In that case we have to use a t-distribution and the limits of a $(1-\alpha) \times 100\%$ Confidence Interval for μ are: $\bar{y} \pm t_{\alpha/2} \times s/\sqrt{n}$, with $t_{\alpha/2}$ coming from the t-distribution with $\nu = n - 1$ degrees of freedom.

The value z_α is the upper α-point of a standard normal distribution, so $z_{0.05} = 1.645$ and $z_{0.05/2} = z_{0.025} = 1.960$. The value t_α is the upper α-point of a t-distribution with the appropriate degrees of freedom ($= \nu$): e.g., when $\nu = 10:\ t_{0.05} = 1.812$ and $t_{0.05/2} = t_{0.025} = 2.228$ (See O&L: Table 1, with the Standard Normal Distribution, and Table 2 with the Inverse Student's t Distributions).

Exercises to be done during the tutorial

Exercise 5.1 and **Exercise 5.2** are in the presentation handouts of Tutorial 5. For answers/feedback check Brightspace.

Post-class activity

Watch:

☐ '*Confidence level and Margin of Error*' (Duration: 5:30 min.), when not done as pre-class activity,

☐ '*Sampling distribution and estimation of population mean*' (Duration: 4:56 min.),

☐ The first 3:30 min. of '*Understanding Confidence Intervals*' (Duration: 4:02 min.).

All of the clips are linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 5.3

A factory delivers packages of sugar. A shop owner suspects that the weight of these packages is systematically less than 1000 g.

To prove this, the shop owner takes a random sample of 30 packages and weighs each package individually. The observed weights are denoted by $y_1, y_2, \ldots, y_{30}$. The 30 observed weights can be considered as independent and normally distributed with $\text{E}(y) = \mu$ and $\sqrt{\text{var}(y)} = \sqrt{\sigma_y^2} = \sigma$.

Computational results: sample mean $\bar{y} = 998.62$ g and sample standard deviation $s = 5$ g.

a. Determine the expected value $\text{E}(\bar{y})$ and the standard deviation $\sigma_{\bar{y}}$ for the sampling distribution of the mean.

b. Determine the 90% Confidence Interval for the population mean μ.

c. Give two case specific interpretations for the Confidence Interval from b.

Exercise 5.4

Do either

☐ **Exercise 4.85 O&L 7th Edition p.225**, or
☐ **Exercise 4.85 O&L 6th Edition p.215**.

Exercise 5.5

Do either

- ☐ **Exercise 5.5 O&L 7th Edition p.286** [Note: Typographical error in the text on line 4, "$\hat{\sigma} = 7.1$ n" should be "$\sigma = 7.1$ mg".], or
- ☐ **Exercise 5.5 O&L 6th Edition p.276** [Note: Before answering a., and b., answer the following additional question: "Identify the population about which inferences can be made from the sample data.". This is question a. in the O&L 7th Edition.].

Exercise 5.6

Do either

- ☐ **Exercise 5.6 O&L 7th Edition p.286**, or
- ☐ **Exercise 5.7 O&L 6th Edition p.276**.

Exercise 5.7

Do either

- ☐ **Exercise 5.7 O&L 7th Edition p.286**, or
- ☐ **Exercise 5.6 O&L 6th Edition p.276** [Note: Typographical error in the text of Exercise 5.6c on line 2, "how many of the 95% confidence intervals…" should be "how many of the 99% confidence intervals…"].

Tutorial 6

Learning outcomes

After this tutorial the student should be able to:

- explain and interpret the general concepts of hypothesis testing: H_0, H_a, test statistic, α;
- explain and interpret the concept of a p-value;
- mention the eight steps of hypothesis testing using a p-value;
- mention the eight steps of hypothesis testing using a Rejection Region;
- apply a test for a population mean μ for the one sample situation using a p-value;
- apply a test for a population mean μ for the one sample situation using a Rejection Region.

Pre-class activity

Watch:

☐ '*Statistics, Student's t-distribution, Gosset, and Guinness*' (Duration: 2:49 min.)

The clip is linked on Brightspace.

Hypothesis testing

In this tutorial hypothesis testing will be introduced.

In the next section the concepts related to hypothesis testing are given (see first box) and the general idea of hypothesis testing is discussed. Afterwards the level of significance (= p-value) is introduced and in the subsequent section the eight steps of hypothesis testing with the p-value are given. Additionally this is applied to a hypothesis about the population mean μ: the *one-sample t-test*. The theoretical part is illustrated with an example including the R/R Commander output of the one-sample t-test.

In the concluding section of this tutorial the eight steps of hypothesis testing using the rejection region (R.R.) are given. This is followed up by an example using the same data as before and applying a one-sample t-test again to illustrate, that using the R.R. is just another way to answer your research question.

Read:

☐ **O&L 7th Edition**:

- paragraph 5.4 pp.242-249 to the end of example 5.7, **or**

☐ **O&L 6th Edition**:

- paragraph 5.4 pp.232-239 to the end of example 5.7.

General concepts of hypothesis testing

concept	description
null hypothesis H_0	the hypothesis to be tested, which is rejected or not rejected
alternative hypothesis (H_a, or H_1)	the research hypothesis, which may or may not be proven based on the random sample
test statistic	a function of the observations, which is used for testing H_a (test statistic is often abbreviated as T.S.)
type I error	H_0 is rejected, while H_0 is true (see Tutorial 7)
type II error	H_0 is not rejected, while H_a is true (see in Tutorial 7)
α (significance level)	maximum probability allowing a type I error
p-value or level of significance	the observed p-value is the probability to find the observed outcome or a more extreme outcome for the test statistic, given that the null hypothesis is true if p-value $> \alpha$ then H_0 is not rejected and H_a is not shown/not accepted if p-value $\leq \alpha$ then H_0 is rejected and H_a is shown/accepted
rejection region	a collection of all possible values of the test statistic for which H_0 has to be rejected (rejection region is often abbreviated as R.R.)

Please note that any hypothesis test can be conducted by using a p-value or by using a rejection region. In scientific papers, almost always a p-value is used. There is an ongoing discussion about the advantages and disadvantages of the p-value.

The level of significance (p-value) of a test

Read:

- ☐ **O&L 7th Edition**:
  - paragraph 5.6 pp.257-260, **or**
- ☐ **O&L 6th Edition**:
  - paragraph 5.6 pp.246-249.

Test procedure in eight steps using the p-value

When testing with a test statistic, and a p-value mention:

Definition of the parameter(s)!

1. the null-hypothesis H_0 versus the alternative hypothesis H_a
2. the test statistic (T.S.)
3. the distribution of the test statistic under H_0
4. the behavior of the test statistic under H_a (i.e., "higher"/"lower"/"higher or lower" values than under H_0)
5. the type (right-, left-, two-tailed) p-value
6. the outcome of the test statistic based on the observations
7. the level of significance (p-value)
8. the conclusion:
   - when p-value $\leq \alpha$ then H_0 is rejected and H_a is shown;
   - when p-value $> \alpha$ then H_0 is **not** rejected and H_a is **not** shown.
   Formulate the conclusion in words with respect to the research question (in terms of H_a).

! Remarks about the hypothesis testing procedure steps.

The book O&L uses 5 steps instead of 8 steps. See **O&L 7th Edition** p.243, or **O&L 6th Edition** p.233. In the tutorials, and Lecture Notes (examples and exercises) the eight steps are used to provide a more clear overview and understanding.

With respect to the 8 steps in the hypothesis testing procedure:
- The first five steps (of the eight) need to be written down before collecting (and analyzing) the data
- The collected data will be used only in step 6 and onwards!
- The null distribution of the test statistic is the distribution of the test statistic, assuming the null hypothesis to be true.

One-sample *t*-test: hypothesis testing for a population mean μ

The one-sample t-test is applied to, as the name suggests, a single simple random sample, in which the variable (e.g., y) is continuous quantitative.

Hypothesis testing for μ: the one-sample t-test

Definition: μ = (population) mean...

1. null hypothesis $H_0: \ \mu \leq \mu_0$, or $\mu \geq \mu_0$, or $\mu = \mu_0$ (μ_0 is a placeholder for the hypothesized value)
2. Test Statistic (T.S.): $t = \frac{\bar{y}-\mu_0}{\text{SE}(\bar{y})} = \frac{\bar{y}-\mu_0}{s \ / \ \sqrt{n}}$
3. Under H_0 T.S. t follows a Student's t-distribution with $\nu = n - 1$ degrees of freedom

   etc.

Example 6.1

A lecturer claims that students can do the exam of his course within two hours. For this reason he wants to shorten the exam time from 3 to 2 hours in the coming academic year. The student board doubts the claim of the lecturer and thinks that it takes more than two hours to complete the exam. To show that the student board is right, a student asked 25 random students who took the exam recently, how much time in minutes y it took them to complete the exam. Test ($\alpha = 0.05$) whether the student board is right. Use the output below for step 6 – 8. You may assume that the 25 observations are independent and normally distributed with $\text{E}(y) = \mu$.

Solution:
Before starting the actual test procedure, formulate the research question (RQ), define the parameter of interest, and decide (based on the available information) which test is the most appropriate.

RQ: Is the expected value in the population for the time to complete the exam more than 120 minutes?
Parameter of interest: μ: population mean time in minutes to complete the exam

Available information: There is one random sample of 25 students ($n = 25$); the variable y is the time in minutes to complete the exam; σ is not mentioned and hence its value is unknown; y is normally distributed; the research question is about the population mean μ (expected value). From the combination of these facts we know that we may apply a one-sample t-test.

1. $H_0: \ \mu \leq 120$ versus $H_a: \ \mu > 120$
2. The test statistic (T.S.): $t = \frac{\bar{y}-\mu_0}{s \ / \ \sqrt{n}} = \frac{\bar{y}-120}{s \ / \ \sqrt{25}}$
3. Under H_0 T.S. t follows a t-distribution with $\nu = n - 1 = 24$ degrees of freedom.
4. Under H_a T.S. t tends to higher values than under H_0.
5. The p-value is right-tailed.

This is what can be stated without using the data set itself. The data (see Table 10) as well as the output from a one-sample t-test is shown below.

Table 10: Time in minutes needed by 25 students to complete the exam.

170.4767	132.8770	184.02669	146.67294	104.78553
178.4164	115.8079	133.20909	164.94575	157.45583
138.8814	125.9051	99.87364	90.46134	144.35779
139.3965	113.2711	123.57045	182.56808	86.37873
158.9425	108.8226	121.47880	104.32277	126.86655

```r
t.test(examtime_min, alternative = "greater", mu = 120, conf.level = 0.95)
```

```
#>
#>  One Sample t-test
#>
#> data:  examtime_min
#> t = 2.485, df = 24, p-value = 0.01016
#> alternative hypothesis: true mean is greater than 120
#> 95 percent confidence interval:
#>  124.4084      Inf
#> sample estimates:
#> mean of x
#>  134.1509
```

6. Outcome T.S.: $t = \frac{\bar{y}-120}{s\ /\ \sqrt{25}} \approx 2.485$ (from output above).
7. p-value: $P(t \geq 2.485) \approx 0.0102$ (In general, p-values are reported with 4 decimals.)
8. p-value $\approx 0.0102 < 0.05$. Therefore, H_0 is rejected, H_a has been shown. It is shown (with $\alpha = 0.05$) that the expected time to complete the exam is more than two hours.

Test procedure in eight steps using the rejection region

When testing with a test statistic, and a rejection region mention:

Definition of the parameter(s)!

1. the null hypothesis H_0 versus the alternative hypothesis H_a
2. the test statistic (T.S.)
3. the distribution of the test statistic under H_0
4. the behavior of the test statistic under H_a (i.e., "higher"/"lower"/"higher or lower" values than under H_0)
5. the type (right-, left-, two-tailed) rejection region (R.R.)
6. the outcome of the test statistic based on the observations
7. the rejection region (R.R.)
8. the conclusion:
   - when the outcome of the T.S. is in the rejection region (R.R.) then H_0 is rejected and H_a is shown;
   - when the outcome of the T.S. is **not** in the rejection region (R.R.) then H_0 is **not** rejected and H_a is **not** shown.
     Formulate the conclusion in words with respect to research question (in terms of H_a).

Example 6.2

For the description of the study as well as the research question please see Example 6.1.

Here additionally assume that there is not a computer at hand nor a graphing calculator. Therefore, finding the exact p-value is impossible and the rejection region needs to be used to perform the hypothesis test.

Solution:

1. $H_0: \ \mu \leq 120$ versus $H_a: \ \mu > 120$

2. The test statistic (T.S.): $t = \frac{\bar{y}-\mu_0}{s\ /\ \sqrt{n}} = \frac{\bar{y}-120}{s\ /\ \sqrt{25}}$
3. Under H_0 T.S. t follows a t-distribution with $\nu = 24$ degrees of freedom.
4. Under H_a T.S. t tends to higher values than under H_0.
5. The rejection region (R.R.) is right-tailed.

This is what can be stated without using the data set itself.

The sample mean $\bar{y}$ and the standard deviation s based on the data (see Table 10) shown above can be calculated:

```
mean(examtime_min)
#> [1] 134.1509
sd(examtime_min)
#> [1] 28.47202
```

6. Outcome T.S.: $t = \frac{\bar{y}-120}{s\ /\ \sqrt{25}} \approx \frac{134.1509-120}{28.472\ /\ 5} \approx 2.485$
7. R.R.: $t \geq 1.711$ (from O&L Table 2 with $\alpha = 0.05$ and $\nu = n - 1 = 25 - 1 = 24$ degrees of freedom)
8. The test statistic ($t \approx 2.485$) is in the rejection region. Therefore, reject H_0, and H_a is shown.
   It is shown (with $\alpha = 0.05$) that the expected time to complete the exam is more than two hours.

Exercises to be done during the tutorial

Exercise 6.1 is in the presentation handouts of Tutorial 6. For answers/feedback check Brightspace.

Post-class activity

Watch:

- ☐ '*Statistics, Student's t-distribution, Gosset, and Guinness*' (Duration: 2:49 min.), when not done as pre-class activity,
- ☐ '*Understanding the p-value*' (Duration: 4:42 min.),
- ☐ '*The Rejection Region*' (Duration: 7:29 min.).

All of the clips are linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 6.2

The information as given with Exercise 5.3 is repeated here:
A factory delivers packages of sugar. A shop owner suspects that the weight of these packages is systematically less than 1000 g.

To prove this, the shop owner takes a random sample of 30 packages and weighs each package individually. The observed weights are denoted by $y_1, y_2, \ldots, y_{30}$. The 30 observed weights can be considered as independent and normally distributed with $E(y) = \mu$ and $\sqrt{\text{var}(y)} = \sqrt{\sigma_y^2} = \sigma$.

Computational results: sample mean $\bar{y} = 998.62$ g and sample standard deviation $s = 5$ g.

a. Formulate the research question.

b. What is the appropriate test given the situation and the research question?

c. Apply the in b. mentioned test, to test the hypothesis of the shop owner ($\alpha = 0.05$). Write down all eight steps. Use Figure 8 to determine the *p*-value.

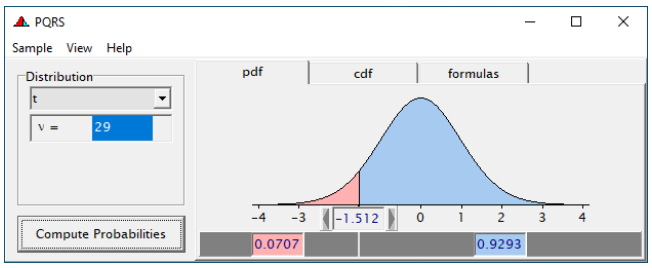


Figure 8: PQRS screenshot to determine the p-value.

d. Suppose you misunderstood the shop owner, who actually asked you to analyse the data with $\alpha = 0.10$. Mention the step(s) of the analysis you should perform again to get the correct result.

e. Perform the step(s) you gave as an answer to d. for $\alpha = 0.10$.

f. Without the PQRS screenshot in Figure 8, or a graphing calculator it would be impossible to determine the p-value. To answer the research question the rejection region must be used. Determine the rejection region ($\alpha = 0.10$) and draw the conclusion based on this rejection region.

Exercise 6.3

Do either

- ☐ **Exercise 5.42ab O&L 7th Edition p.292**, or
- ☐ **Exercise 5.44ab O&L 6th Edition pp.282-283**.

> **i** Notes to Exercise 6.3
>
> Question a. is asking you to apply the appropriate test. Write down all eight steps and use for steps 6.–8. the R output below.
>
> There is output for two different one-sample t-tests. Think carefully, which output is needed for hypothesis testing and which one is needed to read the confidence interval for μ.
>
> With respect to Question b. first calculate the limits of the asked CI yourself, next check your answer using the R output.

R/R Commander output hypothesis tests for Exercise 6.3:

- One-tailed hypothesis test:

```
#>
#>  One Sample t-test
#>
#> data:  volume
#> t = 3.6442, df = 17, p-value = 0.001003
#> alternative hypothesis: true mean is greater than 1600
#> 95 percent confidence interval:
#>  1661.845      Inf
#> sample estimates:
#> mean of x
#>  1718.333
```

- Two-tailed hypothesis test:

```
#>
#>  One Sample t-test
#>
```

```
#> data:  volume
#> t = 3.6442, df = 17, p-value = 0.002007
#> alternative hypothesis: true mean is not equal to 1600
#> 95 percent confidence interval:
#>  1649.824 1786.843
#> sample estimates:
#> mean of x
#>  1718.333
```

Exercise 6.4

Refers to Exercise 6.3. Again perform the one-sample t-test, but now by using the rejection region approach.

Tutorial 7

Learning outcomes

After this tutorial the student should be able to:

- explain Type I and Type II Error;
- explain the effect of changing α on the probability that a Type I Error occurs;
- recognize a situation for which an independent samples t-test for difference between two population means is applicable;
- apply an independent samples t-test for difference between two population means when $\sigma_1^2 = \sigma_2^2$, as well as when $\sigma_1^2 \neq \sigma_2^2$;
- tell how $\sigma_1^2 = \sigma_2^2$ is tested, and can act accordingly the outcome of this test with respect testing $\mu_1 - \mu_2$;
- calculate and interpret a confidence interval for $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$, as well as when $\sigma_1^2 \neq \sigma_2^2$;
- distinguish the one sample t–test situation from the independent samples t–test situation.

Pre-class activity

Watch:

☐ '*False Positives, False Negatives & Type I & II Errors*' (Duration: 2:29 min.)

The clip is linked on Brightspace.

Type I Error and Type II Error

Whenever a hypothesis test is applied, and a decision with respect to the null and alternative hypothesis has been taken based on the outcome of the test, there is a possibility that an error is made. Two types of error can be distinguished: The Type I and the Type II Error. In the mentioned paragraphs both types are explained.

(Re-)Read:

☐ **O&L 7th Edition**:

- paragraph 5.4 p.244 just above definitions 5.1 and 5.2 to p.246 Example 5.5, **or**

☐ **O&L 6th Edition**:

- paragraph 5.4 p.234 under Figure 5.5 to p.236 Example 5.5.

Confidence Interval and Hypothesis testing for $\mu_1 - \mu_2$

Read:

☐ **O&L 7th Edition**:

- paragraph 6.2 pp.303-315, **or**

☐ **O&L 6th Edition**:

– paragraph 6.2 pp.293-305.

This paragraph focuses on the situation of two independent samples. Therefore, one simple random sample with observations $y_{11}, y_{12}, \ldots, y_{1n_1}$ with $E(y_{1i}) = \mu_1$ and $var(y_{1i}) = \sigma_1^2 \; \forall \; i \in \{1, 2, \ldots, n_1\}$ and a second simple random sample with observations $y_{21}, y_{22}, \ldots, y_{2n_2}$ with $E(y_{2i}) = \mu_2$ and $var(y_{2i}) = \sigma_2^2 \; \forall \; i \in \{1, 2, \ldots, n_2\}$.

Assume that both sets of observations within each sample are independent and normally distributed. Both samples are also considered to be independent from each other. The difference $\mu_1 - \mu_2$ is estimated, the construction of a Confidence Interval, and the testing of the hypotheses for this difference are treated.

Confidence interval for $\mu_1 - \mu_2$

1. σ_1^2 and σ_2^2 are not assumed to be equal, i.e., $\sigma_1^2 \neq \sigma_2^2$:

   - an unbiased estimator for $\mu_1 - \mu_2$ is $\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2$ with standard error:

   $$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

   - the boundaries of a $(1 - \alpha) \times 100\%$ Confidence Interval for $\mu_1 - \mu_2$ are given by:

   $$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

   where $t_{\alpha/2}$ comes from a t-distribution with the degrees of freedom provided by the R/R Commander output.

2. σ_1^2 and σ_2^2 are assumed to be equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

   - an unbiased estimator for $\mu_1 - \mu_2$ is $\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2$ with standard error:

   $$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{s_p^2 \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

   where s_p^2 is the pooled variance (an estimator for σ^2) and calculated as:

   $$s_p^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$$

   - the boundaries of a $(1 - \alpha) \times 100\%$ Confidence Interval for $\mu_1 - \mu_2$ are given by:

   $$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \times \sqrt{s_p^2 \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

   where $t_{\alpha/2}$ comes from a Student t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.

Hypothesis testing for $\mu_1 - \mu_2$

1. σ_1^2 and σ_2^2 are not assumed to be equal, i.e., $\sigma_1^2 \neq \sigma_2^2$:

   - Null hypothesis H_0 : $\mu_1 - \mu_2 \leq D_0$, or $\mu_1 - \mu_2 \geq D_0$, or $\mu_1 - \mu_2 = D_0$
   - Test statistic (T.S.): $t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
   - Under H_0 T.S. t follows a t-distribution with ν degrees of freedom, which can be read from the R/R Commander output.

2. σ_1^2 and σ_2^2 are assumed to be equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

   - Null hypothesis $H_0: \mu_1 - \mu_2 \leq D_0$, or $\mu_1 - \mu_2 \geq D_0$, or $\mu_1 - \mu_2 = D_0$
   - Test statistic (T.S.): $t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\text{SE}(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
   - Under H_0 T.S. t follows a Student t-distribution with $\nu = n_1 + n_2 - 2$ degrees for freedom.

! Remarks about equality or unequality of variances.

- It is obvious that when $\sigma_1^2 = \sigma_2^2$, also $\sigma_1 = \sigma_2$ will hold. Generally this is referred to as the assumption of equal variances. To test whether this assumption holds, Levene's test for equality of variance needs to applied before testing with an independent samples t-test. Levene's test for equality of variance tests the hypotheses $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a: \sigma_1^2 \neq \sigma_2^2$. In the R/R Commander output the test statistic is given as a F-statistic and associated degrees of freedom. In this course this test statistic and degrees of freedom can be ignored. The only part of the output from Levene's test, which requires interpretation is the p-value. This is given under `Pr(>F)` in the R/R Commander output for a Levene's test:
  - When this p-value is smaller than or equal to α, reject $H_0: \sigma_1^2 = \sigma_2^2$ and $H_a: \sigma_1^2 \neq \sigma_2^2$ has been shown. Perform the independent samples t-test under the assumption of unequal variances.
  - When the p-value is larger than than α, do **not** reject $H_0: \sigma_1^2 = \sigma_2^2$ and $H_a: \sigma_1^2 \neq \sigma_2^2$ has **not** been shown. Perform the independent samples t-test under the assumption of equal variances.
- When assuming unequal variances, the so-called Welch-Satterthwaite approximation for the degrees of freedom needs to be applied. This is discussed in **O&L 7th Edition** pp.311-312, or **O&L 6th Edition** pp.301-312. The details of the approximation for the degrees of freedom, can be skipped. They can be read from the R/R Commander output.

Exercises to be done during the tutorial

Exercise 7.1 is in the presentation handouts of Tutorial 7. For answers/feedback check Brightspace.

Exercise 7.1

The aim of this research into the effect of weasel scent on hamsters is to show that hamsters, that are exposed to weasel scent, have increased cortisol levels [ng/ml] in their blood.

Use the following notation:

- μ_A = the population mean cortisol level in hamsters not exposed to weasel scent, and
- μ_B = the population mean cortisol level in hamsters exposed to weasel scent.

Assume that the cortisol levels in both groups are normally distributed with means μ_A and μ_B and equal variance σ^2.

Useful R/R Commander plots and output for answering the questions are given below.

- Side-by-side boxplots (see Figure 9)
- Q-Q Plots of cortisol level [ng/ml] in hamsters **not** exposed (see Figure 10a), and exposed (see Figure 10b) to weasel scent.
- Numerical summary for the cortisol level [ng/ml] of hamsters not exposed and exposed to weasel scent (see Table 11)

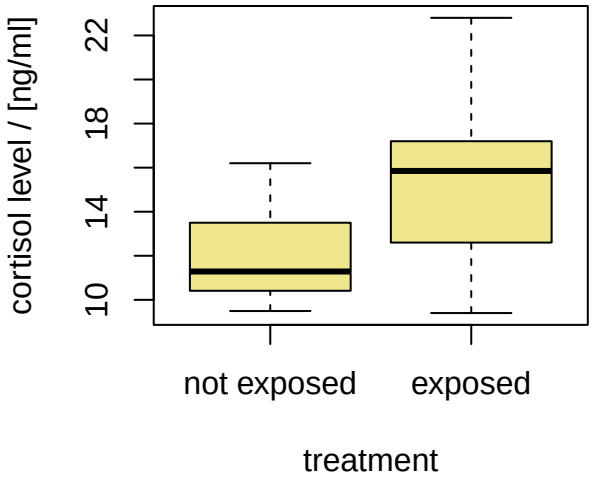


Figure 9: Side-by-side boxplots for the cortisol level [ng/ml] in hamsters not exposed, and exposed to weasel scent.

Table 11: Numerical summary for the cortisol level [ng/ml] of hamsters not exposed and exposed to weasel scent.

	mean	sd	se(mean)	cortisol:n
not exposed	12.10773	2.144313	0.6780915	10
exposed	15.37000	3.715747	1.1750225	10

- Levene's test (see Table 12)

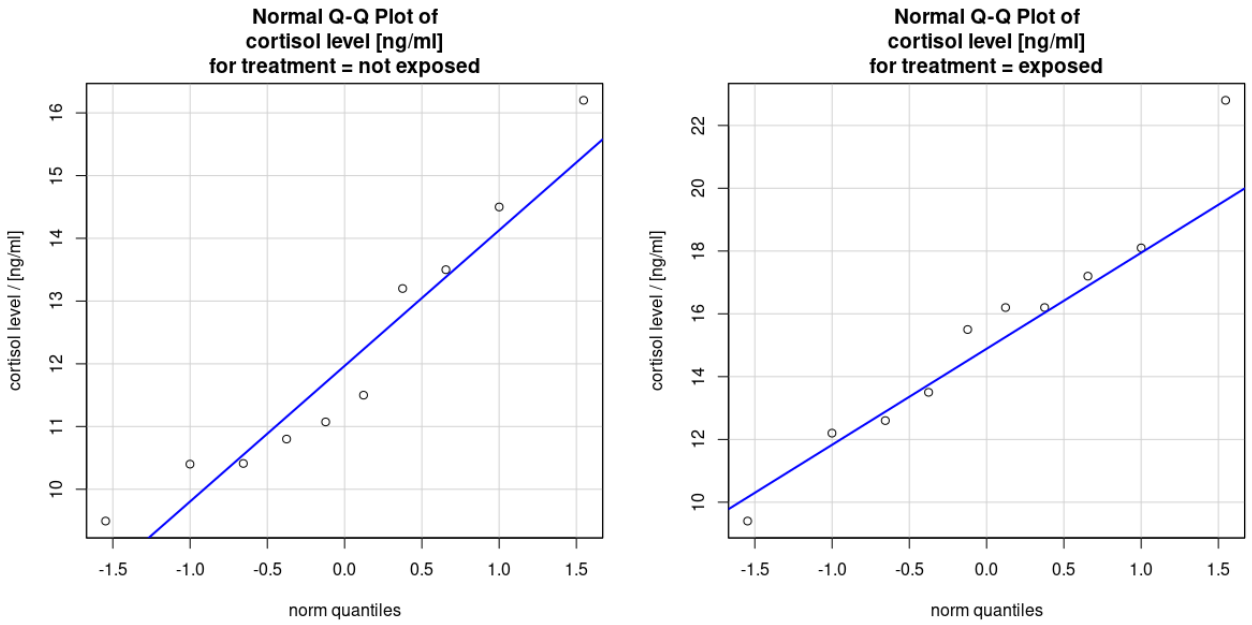
Table 12: Levene's Test for Homogeneity of Variance (center = "mean")

	Df	F value	Pr(>F)
group	1	1.4488	0.2443
	18		

- One-tailed Two Sample t-test:
  - data: cortisol by treatment
  - t = -2.4047, df = 18, p-value = 0.01358
  - alternative hypothesis: true difference in means between group not exposed and group exposed is less than 0
  - 95 percent confidence interval: ($-\infty$, -0.9097619)
  - sample estimates:
    - mean in group not exposed: 12.10773
    - mean in group exposed: 15.37
- Two-tailed Two Sample t-test:
  - data: cortisol by treatment
  - t = -2.4047, df = 18, p-value = 0.02717
  - alternative hypothesis: true difference in means between group not exposed and group exposed is not equal to 0
  - 95 percent confidence interval: (-6.1124762, -0.4120652)
  - sample estimates:
    - mean in group not exposed: 12.10773
    - mean in group exposed: 15.37

a. Check the assumption that the cortisol levels in both groups are normally distributed.

b. Compute the estimated difference between the parameters of interest ($\mu_A - \mu_B$).



(a) Not exposed to weasel scent.

(b) Exposed to weasel scent.

Figure 10: Q-Q Plots of cortisol level [ng/ml] in hamsters.

c. The estimate of σ^2 is 9.20. How is this value calculated based on the information given in the R/R Commander output?

d. Use answer c. to verify that the standard error of the estimator for $\mu_A - \mu_B$ is approximately equal to 1.3566**.**

e. Verify whether it can be shown that the mean cortisol level in group B is higher than that in group A, at a significance level (α) of 0.05**. Give all 8 steps of the test.**

f. Same question in e., but now use the Rejection Region approach instead.

g. Give the 95% **confidence interval for the difference between the population means $\mu_A - \mu_B$.**

Post-class activity

Watch:

- ☐ '*False Positives, False Negatives & Type I & II Errors*' (Duration: 2:29 min.), when not done as pre-class activity,
- ☐ '*Levene's test*' (Duration: 6:31 min.).

All of the clips are linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 7.2

Do either

- ☐ **Exercise 5.16a O&L 7th Edition p.288**, or
- ☐ **Exercise 5.18a O&L 6th Edition p.278**

> **Additional question**
>
> Could you possibly have made a Type II Error in drawing your conclusion in question a. as specified above? Explain your answer.

Exercise 7.3

Read the research study about effects of an oil spill on plant growth:

- ☐ **O&L 7th Edition pp.336-339**, or
- ☐ **O&L 6th Edition pp.325-328**

Useful R/R Commander output for answering the questions of exercise 7.3:

- Levene's test (see Table 13)

Table 13: Levene's Test for Homogeneity of Variance (center = "mean")

	Df	F value	Pr(>F)
group	1	5.209	0.0252
	78		

- One-tailed Two Sample t-test:
  - data: density by tract
  - $t = 3.8209$, df = 78, p-value = 1.3×10^{-4}
  - alternative hypothesis: true difference in means between group control and group oilspill is greater than 0
  - 95 percent confidence interval: (6.5180789, ∞)
  - sample estimates:
    - mean in group control: 38.475
    - mean in group oilspill: 26.925
- Two-tailed Two Sample t-test:
  - data: density by tract
  - $t = 3.8209$, df = 78, p-value = 2.7×10^{-4}
  - alternative hypothesis: true difference in means between group control and group oilspill is not equal to 0
  - 95 percent confidence interval: (5.5319554, 17.5680446)
  - sample estimates:
    - mean in group control: 38.475
    - mean in group oilspill: 26.925
- One-tailed Welch Two Sample t-test:
  - data: density by tract
  - $t = 3.8209$, df = 64.1027923, p-value = 1.5×10^{-4}
  - alternative hypothesis: true difference in means between group control and group oilspill is greater than 0
  - 95 percent confidence interval: (6.5049323, ∞)
  - sample estimates:
    - mean in group control: 38.475
    - mean in group oilspill: 26.925
- Two-tailed Welch Two Sample t-test:
  - data: density by tract
  - $t = 3.8209$, df = 64.1027923, p-value = 3×10^{-4}
  - alternative hypothesis: true difference in means between group control and group oilspill is not equal to 0
  - 95 percent confidence interval: (5.5113368, 17.5886632)

- sample estimates:
  - mean in group control: 38.475
  - mean in group oilspill: 26.925

Perform the hypothesis test as described at O&L 7th Edition p.339 [O&L 6th Edition p.328] yourself following questions a., b., c. and d.

a. Formulate the research question (RQ), define the parameter(s) of interest and give arguments why the independent samples *t*-test is here the appropriate test.

b. Denote/write down the first five steps of the test procedure for the independent samples *t*-test.

c. Next use the appropriate part of the R/R Commander output, to test whether the variances can be assumed unequal or equal. Mention the null and alternative hypothesis for this test, the *p*-value, your conclusion in words, and whether the result implies that you need the Welch-Satterthwaite approximation for the degrees of freedom, when proceeding with the independent sample *t*-test.

d. Finally, use the appropriate output (see your answers to questions b. and c.) to continue with the independent samples *t*-test. Write down all steps (i.e., step 6., 7. and 8.).

e. Give the formula for the 95% Confidence Interval of the difference in (population) mean plant density between control and restored oilspill sites (see O&L p.341 [p.330 in the 6th Edition]) and read the result from the appropriate R/R Commander output.

Exercise 7.4

Do either

- ☐ **Exercise 6.6ab O&L 7th Edition p.345**, or
- ☐ **Exercise 6.6ac O&L 6th Edition p.334**

Notes Exercise 6.6ab O&L 7th Edition

Note to exercise 6.6a.: Start with formulating the research question (RQ), defining the parameter(s) of interest and mentioning the name of the correct hypothesis test. Next apply the hypothesis test and mention all the steps.

For the situation described, the interest is in the difference between two population means. This difference is denoted by D_0 in the equation for the test statistic t. Quite often this difference will be 0, and therefore, denoted as $D_0 = 0$. However, in the given situation the hypothesis test described tests whether the difference in the (population) mean dissolved oxygen level between up- and downstream of a community is more than 0.5. R/R Commander will always test with $D_0 = 0$. Therefore, the output from R/R Commander can not be used. The outcome of the test statistic t would have to be manually calculated and an exact p-value can not be obtained. The solution is to use the rejection region approach.

Though the output for the *t*-test can not be used in this particular case, R/R Commander can of course still be used to calculate sample means, variances, perform Levene's test and make graphical representations.

Note to exercise 6.6b.: Mention the plot(s) on which you based your answer.

R/R Commander output Exercise 6.6ab O&L 7th Edition

- Side-by-side boxplots (see Figure 11)
- Q-Q Plots of dissolved oxygen levels [ppm] for locations downstream (see Figure 12a) and upstream (see Figure 12b) from a riverside community.
- Levene's test (see Table 14)

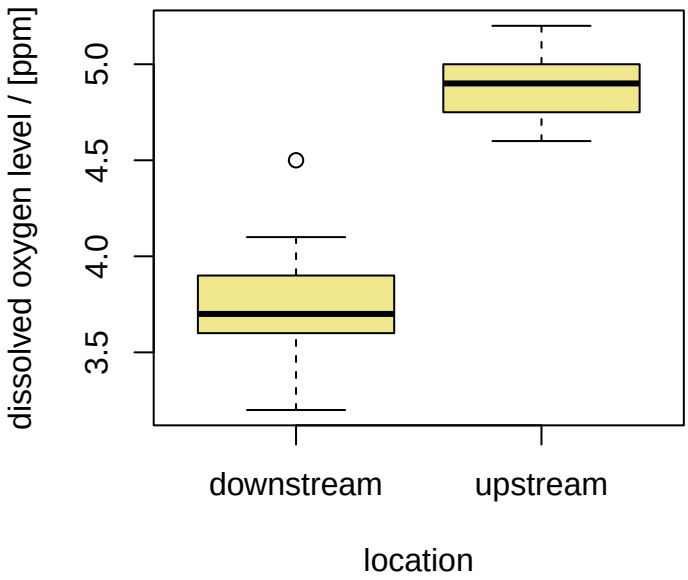
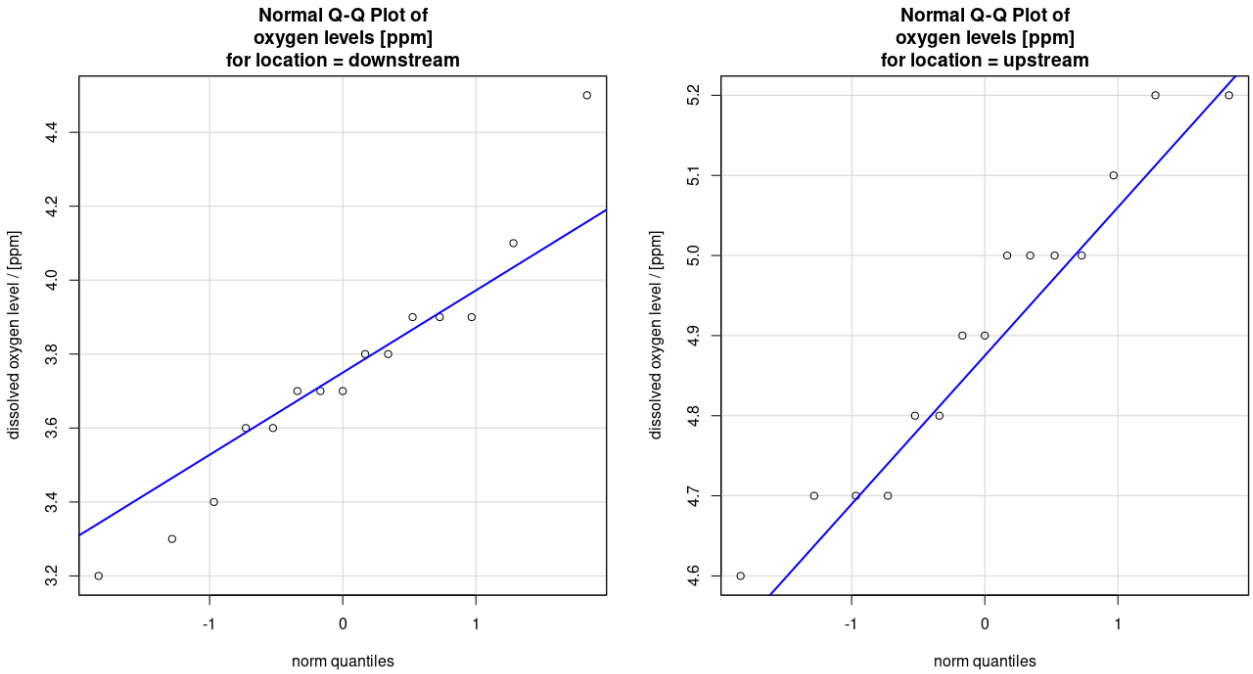


Figure 11: Side-by-side boxplots of dissolved oxygen levels [ppm] for locations up- and downstream from a riverside community.



(a) Downstream from a riverside community.

(b) Upstream from a riverside community.

Figure 12: Q-Q Plots of dissolved oxygen levels [ppm] for locations

Table 14: Levene's Test for Homogeneity of Variance (center = "mean")

	Df	F value	Pr(>F)
group	1	1.5336	0.2259
	28		

- Numerical summary for the dissolved oxygen levels [ppm] in locations down- and upstream from a riverside community (see Table 15)

Table 15: Numerical summary for the dissolved oxygen levels [ppm] in locations down- and upstream from a riverside community.

	mean	sd	oxygen:n
downstream	3.740000	0.3202677	15
upstream	4.906667	0.1869556	15

Additional questions Exercise 6.6ab O&L 7th Edition

I Give the 99% Confidence Interval for the difference between the (population) mean dissolved oxygen level up- and downstream from a community.

II Would a 95% Confidence Interval be wider or narrower than the 99% Confidence Interval? Provide argument to support your answer.

III Suppose the 95% Confidence Interval was based on two samples of size 20. Would this change the confidence coefficient? Why or why not?

Notes Exercise 6.6ac O&L 6th Edition

Note to exercise 6.6a.: Start with formulating the research question (RQ), defining the parameter(s) of interest and mentioning the name of the correct hypothesis test. Next apply the hypothesis test and mention all the steps.

In the independent samples *t*-test R/R Commander uses an alphabetical order for the groups, when calculating the difference between group means. Therefore, in this question the difference is calculated as `aboveTown - belowTown`.

Note to exercise 6.6c.: Mention the plot(s) on which you based your answer.

R/R Commander output Exercise 6.6ac O&L 6th Edition

- Side-by-side boxplots (see Figure 13)
- Q-Q Plots of dissolved oxygen readings [ppm] for locations above (see Figure 14a) and below (see Figure 14b) a community town.
- Levene's test (see Table 16)

Table 16: Levene's Test for Homogeneity of Variance (center = "mean")

	Df	F value	Pr(>F)
group	1	2.8932	0.1
	28		

- Two-tailed Two Sample t-test:
  - data: oxygen by location
  - $t = 1.9551$, df = 28, p-value = 0.06062

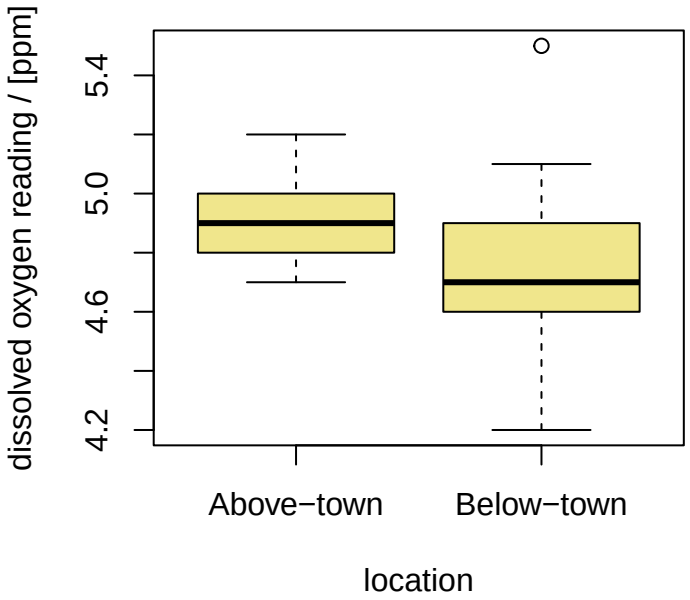
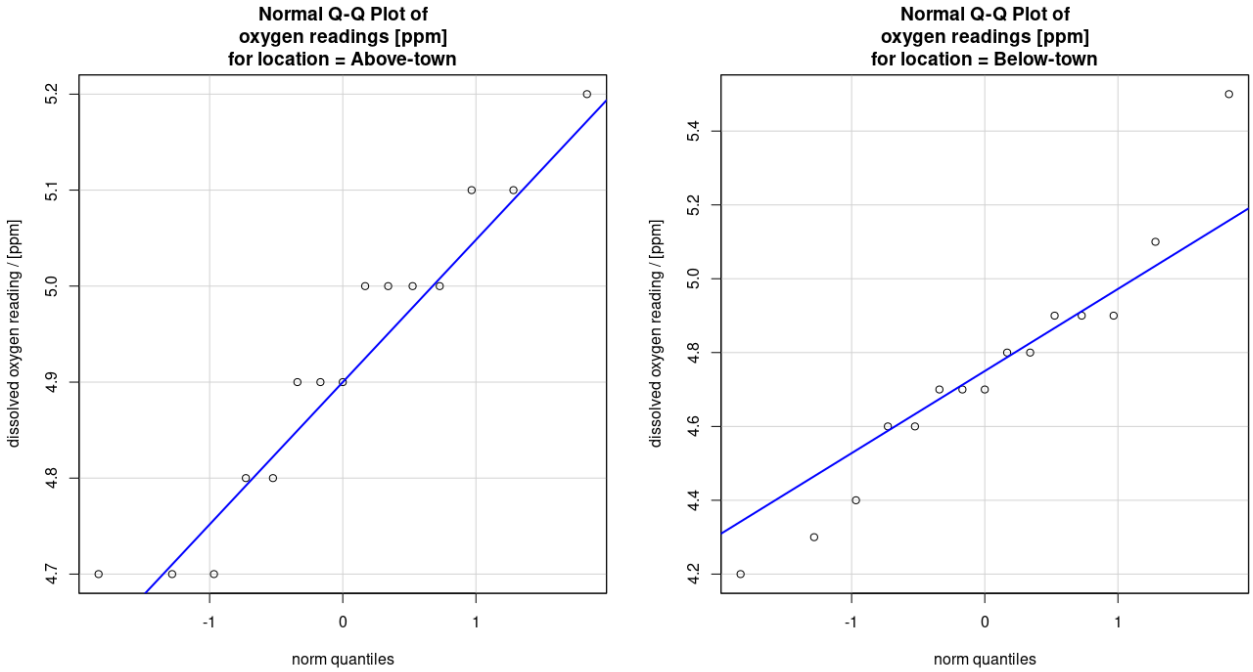


Figure 13: Side-by-side boxplots of dissolved oxygen readings [ppm] for locations above and below from a community town.



(a) Above a community town.

(b) Below a community town.

Figure 14: Q-Q Plots of dissolved oxygen readings [ppm] for locations

- alternative hypothesis: true difference in means between group Above-town and group Below-town is not equal to 0
- 95 percent confidence interval: (-0.0085891, 0.3685891)
- sample estimates:
  - mean in group Above-town: 4.92
  - mean in group Below-town: 4.74

Additional questions Exercise 6.6ac O&L 6th Edition

I Give the 95% Confidence Interval for the difference between the (population) mean dissolved oxygen level above and below a community town.

II Would a 95% Confidence Interval be wider or narrower than the 99% Confidence Interval? Provide argument to support your answer.

III Suppose the 95% Confidence Interval was based on two samples of size 20. Would this change the confidence coefficient? Why or why not?

Tutorial 8

Learning objectives

After this tutorial the student should be able to:

- recognize a situation for which a paired sample t-test for population mean difference is applicable;
- apply a paired sample t-test for population mean difference;
- calculate and interpret a confidence interval for μ_d;
- distinguish the paired sample t-test situation from the one sample and the independent sample t–test situation;
- mention and is aware of some ethical issues concerning research findings;
- determine the probability to get false positive or false negative research findings.

Pre-class activity

Watch:

☐ '*Why Most Published Research Findings Are False*' (Duration: 14:50 min.)

The clip is linked on Brightspace.

Hypothesis testing and confidence interval for μ_d: one sample with paired observations

Read:

☐ **O&L 7th Edition**:

- paragraph 6.4 pp.325-329, **or**

☐ **O&L 6th Edition**:

- paragraph 6.4 pp.314-319.

In this paragraph the situation of one sample with paired observations is treated. Two variables are measured simultaneously on the same experimental units. This yields a single sample with paired observations, for example observations before and after a treatment or observations on married couples. To analyse the data of one sample with paired observations $(y_{11}, y_{21}), (y_{12}, y_{22}), \ldots, (y_{1n}, y_{2n})$, we have to calculate the paired differences $d_i = y_{1i} - y_{2i}$. For the differences $d_i \; \forall \; i \in \{1, 2, \ldots, n\}$ we can apply the theory of the one sample situation.

Hypothesis testing for μ_d

Definition μ_d = (population) mean difference between...

1. null hypothesis $H_0: \; \mu_d = (\mu_d)_0$, where $(\mu_d)_0$ is a placeholder for a hypothesized value.

2. test statistic (T.S.): $t = \frac{\bar{d}-(\mu_d)_0}{\mathrm{SE}(\bar{d})} = \frac{\bar{d}-(\mu_d)_0}{s_d \ / \ \sqrt{n}}$
3. Under H_0 T.S. t follows a Student's t-distribution with $\nu = n - 1$ degrees of freedom.

etc.

Confidence interval for μ_d

- An unbiased estimator for μ_d is $\hat{\mu}_d = \bar{d}$ with standard error $\mathrm{SE}(\bar{d}) = s_d \ / \ \sqrt{n}$.
- The limits of a $(1-\alpha) \times 100\%$ confidence interval for μ_d are $\bar{d} \pm t_{\alpha/2} \times s_d/\sqrt{n}$, where $t_{\alpha/2}$ comes from a Student's t-distribution with $\nu = n - 1$ degrees of freedom.

Some ethics concerning research findings

Some ethical issues concerning research findings are introduced and shortly discussed in class, to make you aware of these issues.

Exercises to be done during the tutorial

Exercise 8.1 up to and including **Exercise 8.3** are in the presentation handouts of Tutorial 8. Check Brightspace for answers/feedback.

Exercise 8.1

a. Is the following experiment an example of 2 independent samples or paired observations?

In the United States of America an experiment was carried out to evaluate the effectiveness of a treatment against tapeworms in the stomach of sheep.

Twenty-four infected sheep (of similar age and health), were randomly assigned, either to a control or a treatment group. After 6 months all sheep were slaughtered, and the number of tapeworms were counted.

b. Is the following experiment an example of 2 independent samples or paired observations?

A river may be contaminated by dispersion of zinc.

Zinc possibly originates from the riverbed. Therefore, near the riverbed higher concentrations are expected.

Zinc concentrations were measured at 6 locations, at the water surface of the river as well as near the riverbed.

Exercise 8.2

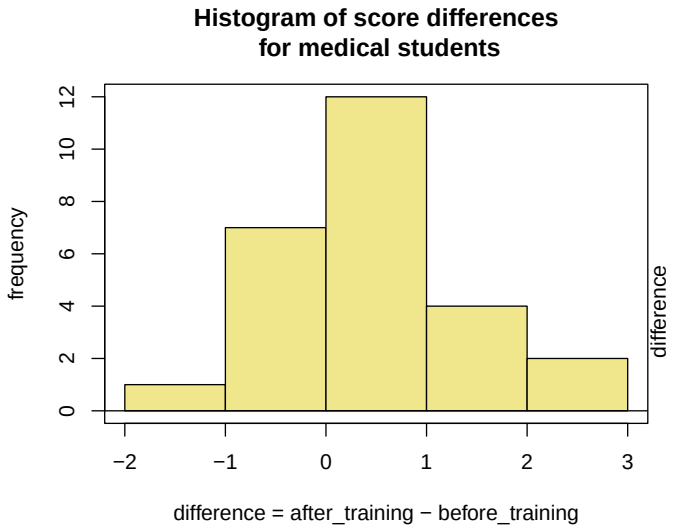
R/R Commander output useful for answering the questions:

- Numerical summaries of the medical students scores (see Table 17)

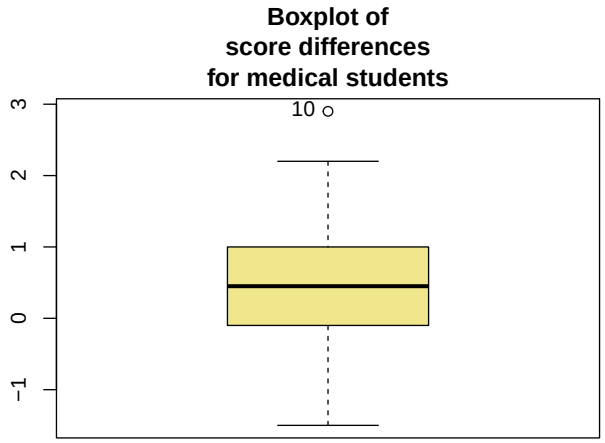
Table 17: Numerical summary of the medical students scores.

	mean	sd	se(mean)	n
after_training	7.14231	1.04045	0.20405	26
before_training	6.60385	1.29876	0.25471	26
difference	0.53846	0.94703	0.18573	26

- Histogram (Figure 15a) and boxplot (Figure 15b) of the differences in scores after and before communication training for medical students.



(a) Histogram



(b) Boxplot

Figure 15: Visualisation of the differences in scores after and before communication training for medical students:

- Q-Q plot of score differences for medical students (Figure 16)

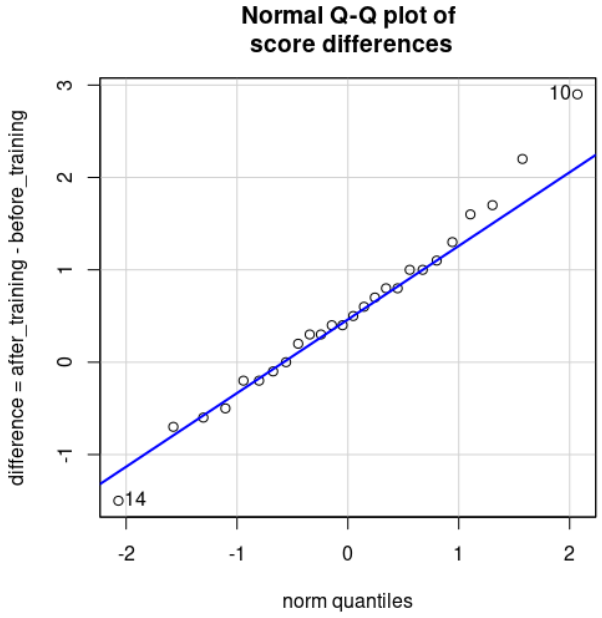


Figure 16: Q-Q plot of differences in scores after and before communication training for medical students.

a. Use hypothesis testing to investigate the research question: Does communication training improve the communication skills of medical students? Check the assumptions, use $\alpha = 0.05$ and mention all steps.

b. Construct a 95% Confidence Interval for the (population) mean difference in scores after and before communication training for the medical students.

Exercise 8.3

Apart from ethical issues in research, there is a probability of wrong conclusions in research.

> Suppose that at Wageningen University 550 hypotheses (550 different studies) are researched every year. From these 250 (alternative) hypotheses are true, 300 (alternative) hypotheses are not true. Probability Type I error = 0.05, Probability Type II error = 0.30. We assume correct research, and no publication bias.

Template for answering the questions:

Decision	H_0 true	H_a true	Total
Reject H_0	$P(\text{Type I Error}) = ?$ $? \times \ldots = \ldots$ (FP)	Correct (TP) $(1-??) \times \ldots = \ldots$	$\ldots$
Do not reject H_0	Correct (TN) $(1-?) \times \ldots = \ldots$	$P(\text{Type II Error}) = ??$ $?? \times \ldots = \ldots$ (FN)	$\ldots$
Total	$\ldots$	$\ldots$	$\ldots$

In this table:

- Reject H_0, and H_a has been shown:
  - When in reality H_0 is true, there will be false positives (FP).
  - When in reality H_a is true, there are true positives (TP).
- Do not reject H_0, and H_a has been not shown:
  - When in reality H_0 is true, there are true negatives (TN).
  - When in reality H_a is true, there will be false negatives (FN).

Calculate:

- the percentage of wrong conclusions of all the conclusions where H_0 is rejected, and H_a has been shown.
- the percentage of correct conclusions (i.e., H_a correctly shown or correctly not shown).

Post-class activity

Watch:

☐ '*Why Most Published Research Findings Are False*' (Duration: 14:50 min.), when not done as pre-class activity,

☐ '*How to distinguish a two independent samples situation from a paired sampled situation.*' (Duration: 7:15 min.)

All of the clips are linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 8.4

Do either

☐ **Exercise 6.28ac O&L 7th Edition p.351**, or
☐ **Exercise 6.28ac O&L 6th Edition p.342**

R/R Commander output useful for answering the questions:

- Numerical summary for the SENS values (Table 19).

Table 19: Numerical summary for the SENS values.

	mean	sd	se(mean)	n
after	5.402	5.155896	1.630437	10
before	7.986	8.123847	2.568986	10
difference	2.584	9.490733	3.001233	10

- Q-Q Plot of the differences in SENS values before and after treatment (Figure 17).

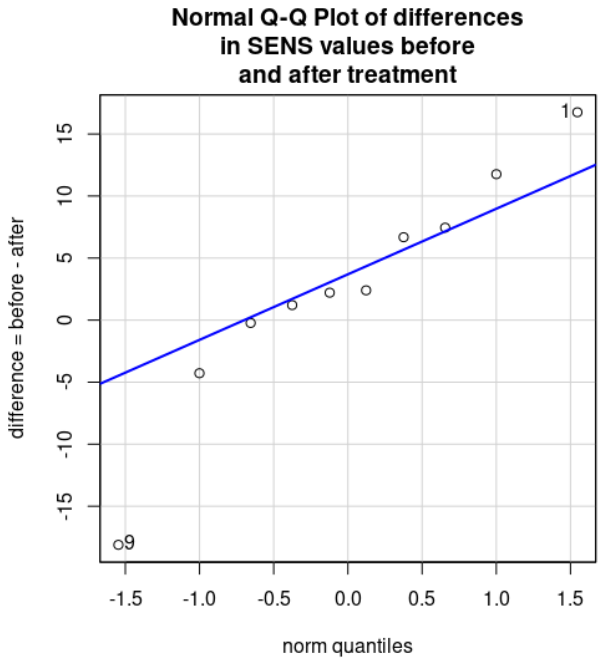


Figure 17: Q-Q Plot of the differences in SENS values before and after treatment.

- One-tailed Paired t-test:
  - data: before and after
  - $t = 0.86098$, df = 9, p-value = 0.2058
  - alternative hypothesis: true mean difference is greater than 0
  - 95 percent confidence interval: (-2.9175993, ∞)
  - sample estimates:
    - mean difference: 2.584

Exercise 8.5

Do either

☐ **Exercise 6.57 O&L 7th Edition p.359**, or
☐ **Exercise 6.57 O&L 6th Edition p.351**

Additional questions

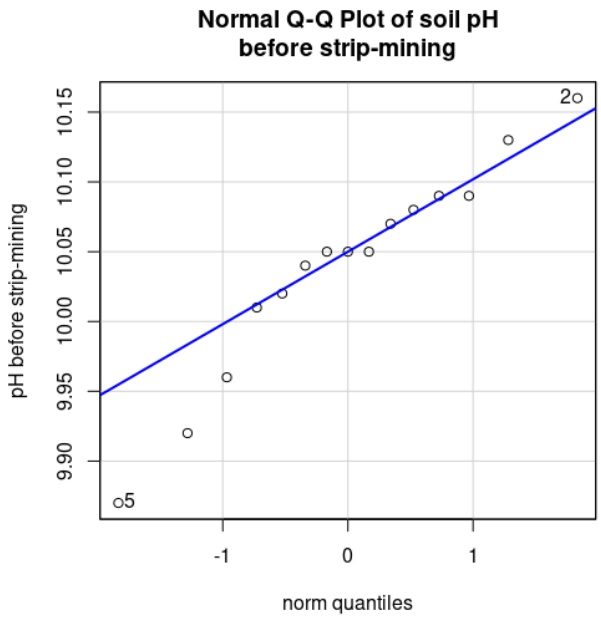
I. Which of the plots (Figure 18) is used to check whether the assumption of normality is met?

II. Provide arguments and draw your conclusion with respect to this assumption.

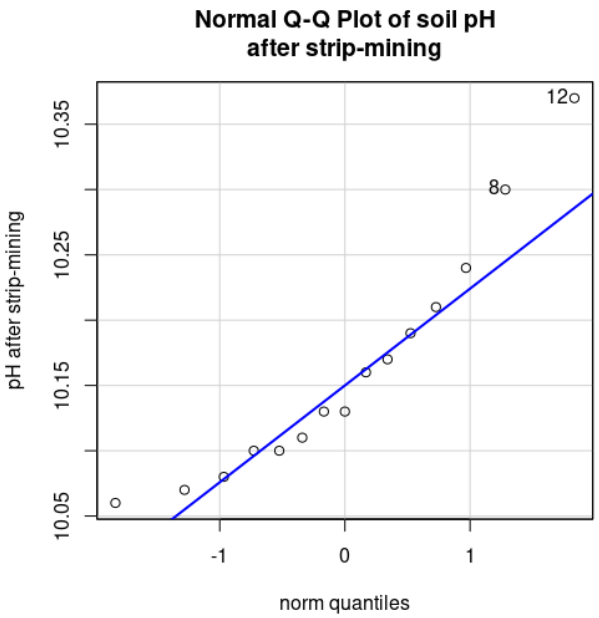
Answer the questions in the following order: **b.**, **a.**, **c.**, **d.**, **Additional question I**, **Additional question II**. Use $\alpha = 0.01$.

R/R Commander output useful for answering the questions:

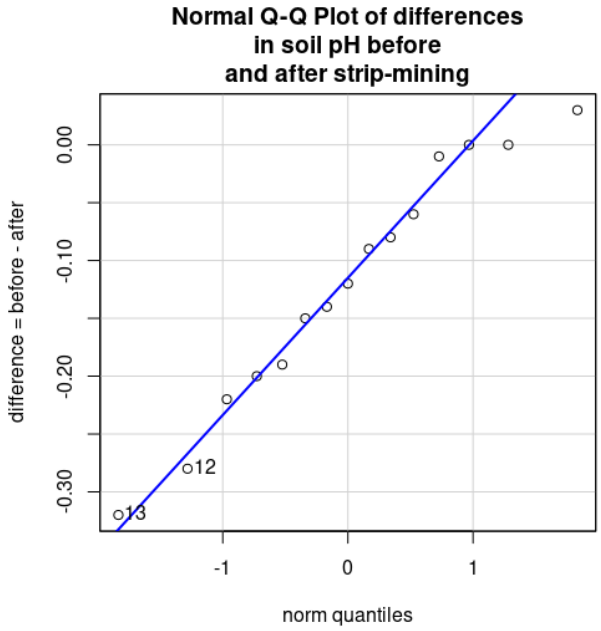
- Left-tailed Paired t-test:
  - data: before and after
  - t = -4.45283, df = 14, p-value = 2.7×10^{-4}
  - alternative hypothesis: true mean difference is less than 0
  - 99 percent confidence interval: ($-\infty$, -0.0500933)
  - sample estimates:
    - mean difference: -0.122
- Right-tailed Paired t-test:
  - data: before and after
  - t = -4.45283, df = 14, p-value = 0.99973



(a) soil pH before strip-mining



(b) soil pH after strip-mining



(c) differences in soil pH before and after strip-mining

Figure 18: Three different Q-Q-plots

  - alternative hypothesis: true mean difference is greater than 0
  - 99 percent confidence interval: (-0.1939067, ∞)
  - sample estimates:
    - mean difference: -0.122

- Two-tailed Paired t-test:
  - data: before and after
  - t = -4.45283, df = 14, p-value = 5.5×10^{-4}
  - alternative hypothesis: true mean difference is not equal to 0
  - 99 percent confidence interval: (-0.2035604, -0.0404396)
  - sample estimates:
    - mean difference: -0.122

Exercise 8.6

Do either

☐ **Exercise 6.58 O&L 7th Edition p.359**, or
☐ **Exercise 6.58 O&L 6th Edition p.351**

Provide a reasoning for your chosen answer.

Exercise 8.7

In a large German university 1500 research hypothesis were tested in the past year. Assume that 400 alternative hypotheses are true. Furthermore assume the probability of a Type I Error to be 0.05 and the probability of a Type II Error to be 0.30.

a. Determine how many of the studies may have a false negative.

b. What is the percentage of false negatives?

c. And what is the percentage of false positives?

d. What will happen with the percentage false positives, when the probability of a Type I Error would be 0.01 instead of 0.05?

Tutorial 9

Learning objectives

After this tutorial the student should be able to:

- recognize a situation for which the binomial test is the most appropriate test;
- give the estimator for π and the associated standard error;
- determine the estimate for π and the associated standard error;
- determine the expected number of successes and the associated standard deviation;
- apply by hand the exact binomial test for π and interpret the result.
- explain correlation;
- mention five properties of Pearson's correlation coefficient;
- mention the estimator for the population correlation ρ ;
- interpret the direction and strength of the correlation based on graph and estimated correlation coefficient.

Pre-class activity

Watch:

☐ '*The danger of mixing up causality and correlation*' (Duration: 5:56 min.)

The clip is linked on Brightspace.

Hypothesis testing for π (one sample; binomial distribution)

(Re-)Read the paragraph about the Binomial distribution from Tutorial 3:

☐ **O&L 7th Edition**:

- paragraph 4.8 pp.167-175 up to Poisson distribution, **or**

☐ **O&L 6th Edition**:

- paragraphs 4.8 pp.158-166 up to Poisson distribution.

Read:

☐ **O&L 7th Edition**:

- paragraph 10.2 pp.483-485 the first six lines, **or**

☐ **O&L 6th Edition**:

- paragraph 10.2 pp.500-502 the first six lines.

In an binomial situation the population proportion "successes" is denoted with π. When π is unknown, use an estimator to get an estimate based on a sample.

Estimator for π

An unbiased estimator for π is $\hat{\pi} = \frac{y}{n}$ with $\text{SE}(\hat{\pi}) = \sqrt{\frac{\hat{\pi} \times (1-\hat{\pi})}{n}}$.

This estimator can be used in a test (the so-called approximate z-test), however in this course we will apply the exact binomial test for answering research questions with respect to π. The exact binomial test simply uses the number of "successes" in the sample (denoted by y) as a test statistic. The null distribution of this test statistic y is a binomial distribution with parameters n and π_0.

Hypothesis testing for π

1. null hypothesis $\text{H}_0: \ \pi = \pi_0$
2. test statistic (T.S.): y = number of *successes* in the sample
3. Under H_0 T.S. y follows a binomial distribution with parameters n and $\pi_0: \ y \sim \text{Bin}(n, \pi_0)$
   - the expected value of this distribution: $\text{E}(y) = n \times \pi_0$
   - the standard deviation of this distribution: $\sigma(y) = \sqrt{n \times \pi_0 \times (1 - \pi_0)}$

! Remark about calculating cumulative binomial probabilities.

For sample sizes $n \leq 20$, cumulative probabilities for the binomial distributions can be found at the end of the Lecture Notes in the appendix using the tables with binomial distributions. Probabilities (for all sample sizes n) can also be calculated by using R/R Commander, or a graphing calculator. During Computer Practical 5 the exact binomial test will be applied using R/R Commander.

Correlation

Read:

☐ **O&L 7th Edition**:

- paragraph 3.7 pp.111-115 from the last two lines above Table 3.16 to side-by-side boxplots,
- paragraph 11.6 pp.591-598 from the sixth line under Example 11.13 **or**

☐ **O&L 6th Edition**:

- paragraph 3.7 pp.104-108 from the last two lines above Table 3.15 to side-to-side boxplots
- paragraph 11.7 pp.613-616 from the fourth line.

We have n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

The sample correlation coefficient is $r_{xy} = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$.

Remark:

r_{xy} is also denoted by $r(x, y)$.

Rules for the correlation coefficient

- $r(x, y) = r(y, x)$
- $-1 \leq r(x, y) \leq +1$
- $r(a \times x + b, c \times y + d) = r(x, y)$ when $a \times c > 0$
- $r(a \times x + b, c \times y + d) = -r(x, y)$ when $a \times c < 0$ where a, b, c, and d are real numbers.
- $r_{xy} = \hat{\beta}_1 \times \sqrt{\frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}}$, or $\hat{\beta}_1 = r_{xy} \times \sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}} = r_{xy} \times \frac{s_y}{s_x}$

! Remark about calculating correlation.

The $r(a \times x + b, c \times y + d) = r(x, y)$ when $a \times c > 0$, or more specifically $r(a \times x, c \times y) = r(x, y)$ when $a \times c > 0$ means that the unit of the measurements has no influence on the value of the correlation coefficient. E.g., when you are interested in the correlation coefficient between the height and the weight of persons, it does not matter whether you measure the weight in gram or in kilogram (and the height in centimeter or meter).

Exercises to be done during the tutorial

Exercise 9.1 and **Exercise 9.2** are in the presentation handouts of Tutorial 9. Check Brightspace for answers/feedback.

Exercise 9.1

Obstructive sleep apnea is a sleep disorder, that causes a person to stop breathing momentarily and then awaken briefly. Sleep researcher theorize that 25% of the general population suffers from this disorder.

Researchers from a university found that 7 out of 20 commercial truck drivers suffered from obstructive sleep apnea.

a. Investigate, using the p-value approach, whether the proportion of truck drivers suffering from sleep apnea is larger than 25%. Mention all the steps.

b. Instead of using the *p*-value approach, the rejection region approach can be used. Using the rejection region approach provide steps 5, 6,7 and 8 to investigate, whether the proportion of truck drivers suffering from obstructive sleep apnea is larger than 25%.

Exercise 9.2

Based on:

- ☐ **Exercise 11.44 O&L 7th Edition p.612**, or
- ☐ **Exercise 11.53 O&L 6th Edition p.342**

Random sample of employees: $n = 52$.
Figure 19 displays a scatter plot of of first-year salary after graduation and years of work experience prior to obtaining their MBA and Table 20 shows the correlation matrix for all variables.

Table 20: Correlation matrix for the variables salary and experience.

	experience	salary
experience	1.0000000	0.6946505
salary	0.6946505	1.0000000

a. Judge the strength of the correlation based on the scatter plot shown in Figure 19.

b. Give the estimated correlation coefficient.

Post-class activity

Watch:

- ☐ *'The danger of mixing up causality and correlation'* (Duration: 5:56 min.), when not done as pre-class activity.

The clip is linked on Brightspace.

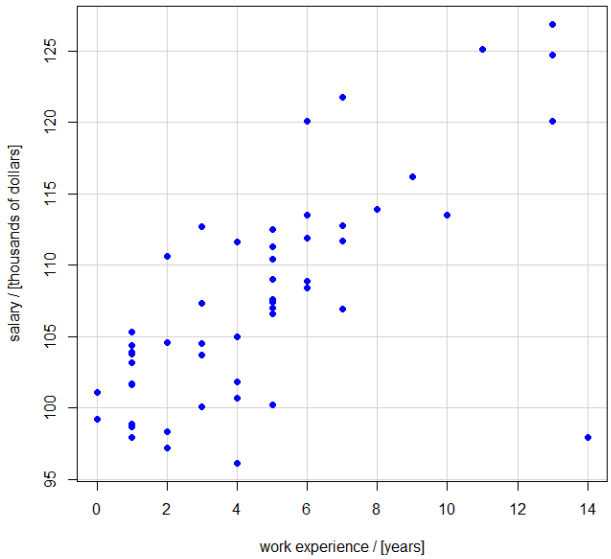


Figure 19: Scatter plot of first-year salary after graduation and years of work experience prior to obtaining their MBA.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 9.3

In 2006 it was shown that 6% of Dutch children between 7 and 9 years old suffer from dyslexia. In 2018 these children were among the students of Dutch universities. Any student suffering from dyslexia is entitled to have more time to finish her/his exam. Based on the number of requests for extra time a teacher thinks that less than 6% of the students at Wageningen University & Research have dyslexia. She takes a random sample of 100 students out of the student population at Wageningen University & Research and finds that 3 of them suffer from dyslexia.

You may use R Commander to find the p-value for this exercise: **Distributions** > **Discrete distributions** > **Binomial distribution** > **Binomial tail probabilities**.... Enter the correct values and choose the correct tail, where lower tail gives the probability: $P(y \leq k)$, and upper tail gives the probability: $P(y > k)$.

a. Formulate the research question.

b. Apply the appropriate test, using $\alpha = 0.10$, to answer the research question.

c. Are you surprised by the outcome of the test?

d. Use R Commander, to find how many students out of 100 should have dyslexia to reject the null hypothesis: Distributions > Discrete distributions > Binomial distribution > Binomial quantiles.... Enter the correct values and choose the correct tail.

Exercise 9.4

This exercise is based on **Exercise 4.45 O&L 6th Edition pp.210-211**, which is not available in **O&L 7th Edition**. Therefore, the exercise is provided below. Use the correct binomial distribution table to answer the questions.

It was claimed that in an inspection of automobiles in Los Angeles, 60% of all automobiles did not meet the EPA regulations. A garage owner thinks that this percentage must be smaller. He takes a sample of 20 automobiles, from which 9 did not meet the EPA regulations.

a. Formulate the research question.

b. Apply the appropriate test , mentioning all steps, to answer the research question. Use $\alpha = 10\%$.

c. How many cars out of a sample of 20 should meet the EPA regulations, when the garage owner is right?

Exercise 9.5

Based on:

- ☐ **Example 11.12 O&L 7th Edition pp.589-590**, or
- ☐ **Example 11.13 O&L 6th Edition pp.610-611**

Read the example in the book and have a look at the table with the data. Answer the following questions with help of the R/R Commander output:

- Scatter plot of the number of eggs produced versus the body weight (Figure 20).

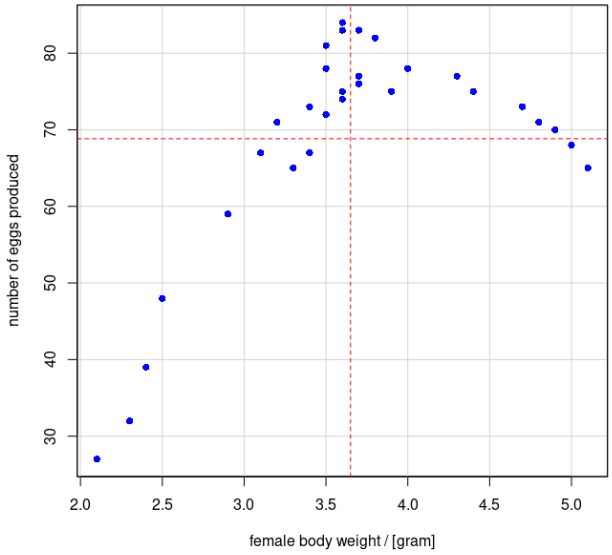


Figure 20: Scatter plot of the number of eggs produced versus the body weight of female grashoppers.

- Correlation matrix for the number of eggs produced and the body weight of female grasshoppers (Table 21)

Table 21: Correlation matrix for the variables number of eggs produced and the body weight of female grasshoppers.

	eggs	weight
eggs	1.0000	0.6059
weight	0.6059	1.0000

a. Based on the scatterplot (Figure 20), would you say that there is a relation between the weight of the female grasshopper and the number of eggs (give arguments)?

b. When a. is answered with 'yes'; would you say this is a straight line relationship? (give arguments)

c. Note: Irrespective the answers given in a. and b., proceed pretending that all assumptions for a straight line relationship are met. Give the estimator and the estimate for the population correlation ρ.

d. Suppose the observations for which weight ≥ 4.0 would be deleted from the data set and the correlation would be calculated again. Will the correlation be unchanged, larger or smaller? Give arguments.

Exercise 9.6

Based on:

☐ **Example 11.14 O&L 7th Edition pp.593-594**, or
☐ **Example 11.15 O&L 6th Edition pp.613-614**

Read the example in the book and have a look at the table with the data. Answer the following questions with help of the R/R Commander output:

- Scatter plot of the productivity index versus the aptitude test score (Figure 21).

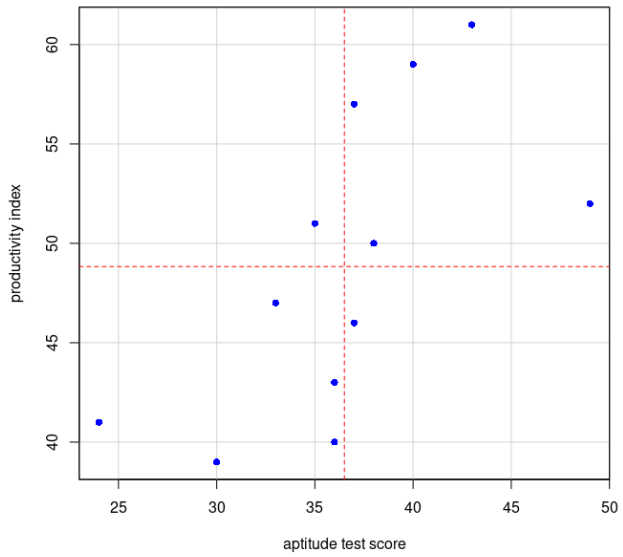


Figure 21: Scatter plot of the productivity index versus the aptitude test score for employees.

- Correlation matrix for the productivity index and the aptitude test score of employees (Table 22).

Table 22: Correlation matrix for the variables productivity index and the aptitude test scores of employees.

	productivity	aptitude
productivity	1.0000	0.6456
aptitude	0.6456	1.0000

a. Does the scatter plot suggest a straight line relation between the productivity index and the aptitude test score of employees?

b. If yes, is it a positive or negative relationship? Give arguments.

Tutorial 10

Learning objectives

After this tutorial the student should be able to:

- recognize a situation and research question for which a simple linear regression is the appropriate analysis;
- mention and explain in own words the five model assumptions of the simple linear regression model;
- give the model equation;
- mention and use the associated terms (i.e., y, x, β_0, β_1, σ and ε) appropriately;
- give the least square estimators for the three model parameters (β_0, β_1, σ);
- interpret a scatter plot and regression line;
- interpret the regression coefficients;
- give (based on R/R Commander output) the estimated regression model;
- give (based on R/R Commander output) the estimated σ:
- apply (based on R/R Commander output) the omnibus F-test for the model.

Pre-class activity

Watch:

☐ '*Statistics Motivation Clip Landscape Architecture and Spatial Planning*' (Duration: 3:36 min.)

The clip is linked on Brightspace.

Simple Linear Regression

Read:

☐ **O&L 7th Edition**:

- paragraph 11.1 pp.555-559 up to smoothers,
- paragraph 11.2 pp.564-568 up to and including Example 11.2, **or**

☐ **O&L 6th Edition**:

- paragraph 11.1 pp.572-576 up to smoothers,
- paragraph 11.2 pp. 581-585 up to and including Example 11.2.

The Simple Linear Regression Model

Model: $\mathrm{E}(y) = \beta_0 + \beta_1 \times x$, or $y = \beta_0 + \beta_1 \times x + \varepsilon$ with $\mathrm{E}(\varepsilon) = 0$.

Assumptions:

- Both y and x are quantitative variables.
- There is a linear relationship between y and x : $\mu_y = \beta_0 + \beta_1 \times x$.
- $\text{var}(\varepsilon) = \sigma_\varepsilon^2$, so $\text{var}(\varepsilon)$ does not depend on the value of x.
- The observations $y_1, y_2, \ldots, y_n$ are independent, or equivalently the residuals $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent.
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are normally distribution with a population mean (or expected value) 0 and **constant variance** σ_ε^2.

Least Squares estimators for the slope and the intercept:

- slope $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ or $\hat{\beta}_1 = r_{xy} \times \frac{s_y}{s_x}$
- intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x}$
- The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and they are the best linear unbiased estimators for β_0 and β_1.

Least Squares estimator for the residual variance

$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \frac{\sum(y_i - \hat{y})^2}{n-2}$, where $\hat{\sigma}_\varepsilon^2$ is an unbiased estimator for σ_ε^2 (the residual variance).

! Remarks about the simple linear regression.

- There is no need to calculate the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_\varepsilon^2$ by hand. For this the R/R Commander output will be used.
- The estimated regression line always passes through the center of all data points $(\bar{x}, \bar{y})$.
- The book (O&L) uses the term *standard error of estimate* for $\hat{\sigma}_\varepsilon$, whereas R/R Commander uses the term *Residual standard error*. These terms are confusing, because "standard error of estimate" is generally reserved for the precision of an unbiased estimator.

Hypothesis test for the model $y = \beta_0 + \beta_1 \times x + \varepsilon$ (omnibus F-test)

Under the assumptions given above:

1. Null hypothesis H_0 : $\beta_1 = 0$ versus H_a : $\beta_1 \neq 0$ (i.e., the alternative hypothesis states that the model has predictive value.)
2. Test statistic (T.S.): $F = \dfrac{\text{MS}_{\text{regression}}}{\text{MS}_{\text{residual}}}$
3. Under H_0 T.S. F follows a F-distribution with for the numerator 1, and for the denominator $n-2$ degrees of freedom.

etc.

! Remarks about the hypothesis test fot the model.

- The F-test above is only used for H_a : $\beta_1 \neq 0$ (and not for H_a : $\beta_1 > 0$ or H_a : $\beta_1 < 0$, or any test value other than 0), and is within the frame work of simple linear regression for a straight line equivalent to the t-test for H_a : $\beta_1 \neq 0$ (as will be explained in Tutorial 11).
- The F-test above is only used for a *two-tailed* alternative hypothesis H_a : $\beta_1 \neq 0$. However, the **rejection region** is (due to the characteristics of the F-distribution) always right-tailed. Therefore, the p-value equals $P(F \geq \text{outcome test statistic})$.

Exercises to be done during the tutorial

Exercise 10.1 and **Exercise 10.2** are in the presentation handouts of Tutorial 10. Check Brightspace for answers/feedback.

Exercise 10.1

Based on the research of Ruben Dijkhof (MSc Thesis Landscape Architecture and Spatial Planning, see clip linked on Brightspace).

Research Question: What is the effect of the distance to the national ecological network (x), measured in kilometers, on the price of agricultural land (y) in the province Limburg?

It is assumed that μ_y and x are linearly related.

Write down the (mathematical) model and describe all used symbols.

Exercise 10.2

Part of the linear model summary for the Rhizotron potato example:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.027743   4.769476   1.054     0.34
thermal_time 0.088391   0.007811  11.316 9.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.364 on 5 degrees of freedom
Multiple R-squared:  0.9624,    Adjusted R-squared:  0.9549
F-statistic: 128.1 on 1 and 5 DF,  p-value: 9.422e-05
```

Using the partial R/R Commander output above:

a. Provide the equation for the estimated model of the root depth explained by thermal time.

b. Provide an estimate for the (population) mean root depth, when the thermal time equals 0.

c. Provide an estimate for the (population) mean root depth, when the thermal time equals 1.

d. What will be the estimated effect on the (population) mean root depth, when the thermal time increases with 4 degree days?

e. Is there any evidence that the model for root depth explained by thermal time has predictive value?

Post-class activity

Watch:

- ☐ *'Statistics Motivation Clip Landscape Architecture and Spatial Planning'* (Duration: 3:36 min.), when not done as pre-class activity.

The clip is linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

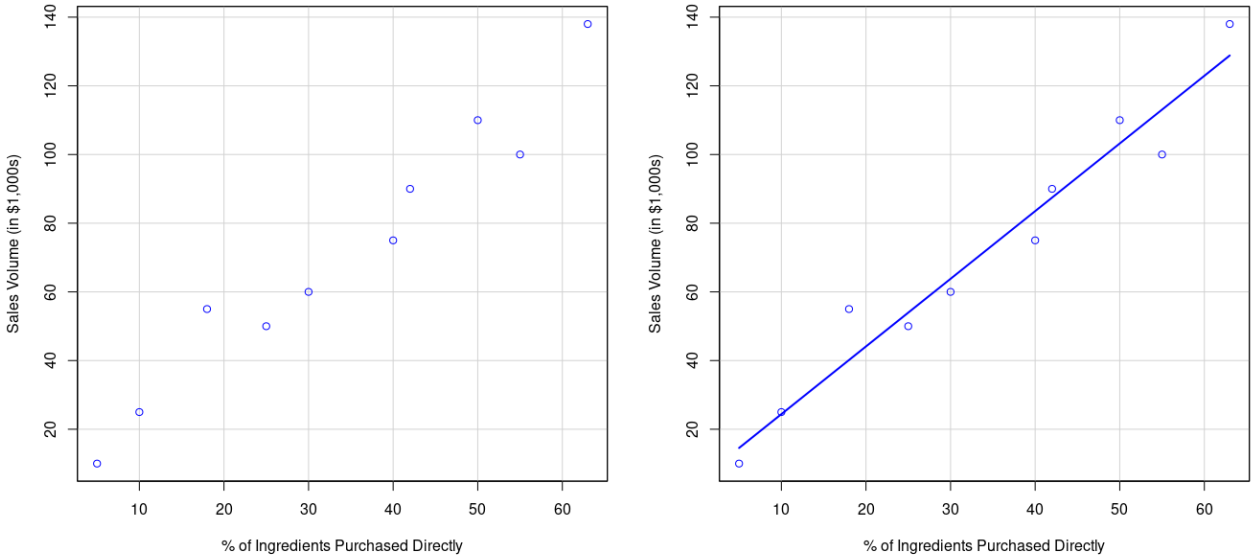
Exercise 10.3

Based on:

- ☐ **Example 11.2 O&L 7th Edition pp.566-568**, or
- ☐ **Example 11.2 O&L 6th Edition pp.583-585**

Read the introduction of this example (not the questions below it). Use the R/R Commander output below to answer the following questions:

- Scatter plots



(a) without Least Squares Line.

(b) with Least Squares Line.

Figure 22: Scatter plots of Sales Volume (in $1,000s) versus % of Ingredients Purchased Directly,

- Summary Simple Linear Regression model (straight line model)

```
Call:
lm(formula = Sales_Volume ~ Purchased_Directly, data = example11_2)

Residuals:
    Min      1Q  Median      3Q     Max
-13.074  -4.403  -1.607   5.719  14.834

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.6979     5.9520   0.789    0.453
Purchased_Directly 1.9705     0.1545  12.750 1.35e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.022 on 8 degrees of freedom
Multiple R-squared:  0.9531,    Adjusted R-squared:  0.9472
F-statistic: 162.6 on 1 and 8 DF,  p-value: 1.349e-06
```

a. What is the research question in example 11.2?

b. Have a look at the provided scatter plots. What can you read from these plots with respect to the strength of the linear relationship?

c. Give, based on the description of the example in the book, the Simple Linear Regression model (straight line model) and describe all symbols used in terms of the actual problem.

d. Test ($\alpha = 0.05$) whether the model has any predictive value. Mention all 8 steps.

e. Is the test performed in d. suitable for the research question (answer a.)? Give arguments.

f. Give the estimated simple linear regression model (straight line model) as an equation.

Exercise 10.4

In a study conducted to examine the quality of fish after 7 days of storage on ice, ten raw fish of the same kind and approximately the same size were caught and prepared for storage on ice. Two of the fish were placed in storage immediately after being caught, two were placed in storage 3 hours after being caught, and two each were placed in storage at 6, 9 and 12 hours after being caught.

Let y denote a measurement of fish quality (on a 10-point scale) after 7 days of storage on ice, and let x denote the time after being caught that the fish were placed in storage on ice. The sample data are (see Table 23):

Table 23: Fish quality data

y	x
8.5	0
8.4	0
7.9	3
8.1	3
7.8	6
7.6	6
7.3	9
7.0	9
6.8	12
6.7	12

The following model is assumed: $y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$

Furthermore assume that the residuals are independent and normally distributed with standard deviation σ_ε. Use, where appropriate, the provided R/R Commander output to answer the questions:

- Summary Simple Linear Regression model (straight line model):

```
Call:
lm(formula = y ~ x, data = fish_quality)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18500 -0.06000  0.01500  0.05875  0.19000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.460000   0.066097  128.00 1.55e-14 ***
x           -0.141667   0.008995  -15.75 2.64e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1207 on 8 degrees of freedom
Multiple R-squared:  0.9688,    Adjusted R-squared:  0.9649
F-statistic: 248.1 on 1 and 8 DF,  p-value: 2.638e-07
```

a. Plot the sample data (by hand). Does there seem to be a linear relation between y and x?

b. Formulate the research question.

c. Give the least squares estimate for β_0 and its estimated standard error.

d. Give the least squares estimate for β_1 and its estimated standard error.

e. Interpret the value of $\hat{\beta}_1$ in the context of this problem.

f. Give an estimate for σ_ε

g. Calculate the estimated population mean fish quality after 7 days storage for a fish placed on ice 7 hours after having been caught.

Tutorial 11

Learning objectives

After this tutorial, within the frame work of a single regression analysis, the student should be able to:

- apply a t-test for the slope;
- construct a confidence interval for the regression coefficients;
- determine for any x-value the predicted expected value $\hat{\mu}$, given the estimated regression equation;
- determine for any x-value the predicted $\hat{y}$ for a random unit, given the estimated regression equation;
- explain the difference in interpretation between $\hat{\mu}$ and $\hat{y}$;
- determine the confidence interval for any predicted mean outcome given the standard error of the predicted value;
- read the fitted value, the standard error of the fitted value and the confidence interval for μ_y from the R/R Commander output;
- read the fitted value and the prediction interval for y for a new value of x from the R/R Commander output;
- distinguish the confidence interval for μ_y from the prediction interval for y;
- explain why the confidence interval for μ_y is smaller than the prediction interval for y .

Hypothesis testing and confidence interval for a regression coefficient

Read:

☐ **O&L 7th Edition**:

- paragraph 11.3 pp.574-577 **or**

☐ **O&L 6th Edition**:

- paragraphs 11.3 pp.590-594.

Hypothesis test for any regression coefficient β_i with $i = 0$ or $i = 1$

Model: $\mathrm{E}(y) = \beta_0 + \beta_1 \times x$, or $y = \beta_0 + \beta_1 \times x + \varepsilon$ with $\mathrm{E}(\varepsilon) = 0$.

Assumptions:

- Both y and x are quantitative variables (by design).
- The observations $y_1, y_2, \ldots, y_n$ are independent, or equivalently the residuals $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent (by design).
- There is a linear relationship between y and x: $\mu_y = \beta_0 + \beta_1 \times x$ (check required).
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are normally distributed with expected value 0 and **constant variance** σ_ε^2, i.e., $\mathrm{var}(\varepsilon) = \sigma_\varepsilon^2$ does not depend on the value of $\hat{y}$ (check required).

Hypothesis testing with β_0 (when $i = 0$), or β_1 (when $i = 1$):

Step 1 Null hypothesis $H_0: \beta_i = d$, where d is a placeholder for a hypothesized value (often $d = 0$). The alternative hypothesis H_a can be left-/right- or two-tailed.
Step 2 Test Statistic (T.S.): $t = \frac{\hat{\beta}_i - d}{SE(\hat{\beta}_i)}$
Step 3 Under H_0 T.S. t follows a Student's t-distribution with $\nu = n - 2$ degrees of freedom.

etc.

Confidence Interval for a regression coefficient

For the model above, with the given assumptions, the limits of a $(1 - \alpha) \times 100\%$ Confidence Interval for a regression coefficient (β_0 when $i = 0$, or β_1 when $i = 1$) are:

$$\hat{\beta}_i \pm t_{\alpha\ /\ 2} \times SE(\hat{\beta}_i)$$

with $P(t \geq t_{\alpha/2}) = \alpha/2$ and where t follows a Student's t distribution with $\nu = n - 2$ degrees of freedom.

Predicting new y values, confidence and prediction intervals

Read:

- ☐ **O&L 7th Edition**:
  - paragraph 11.4 pp.577-581 **or**
- ☐ **O&L 6th Edition**:
  - paragraphs 11.4 pp.594-598.

In this paragraph the prediction of the expected value $E(\mu)$ for a given value of x, the prediction of an individual outcome $E(y)$ as well as the associated confidence and prediction intervals are discussed. Related to prediction, the book also discusses extrapolation.

Exercises to be done during the tutorial

Exercise 11.1 is in the presentation handouts of Tutorial 11. Check Brightspace for answers/feedback.

Exercise 11.1

Based on the data from:

- ☐ **Example 11.14 O&L 7th Edition p.593**, or
- ☐ **Example 11.15 O&L 6th Edition p.613**

A director of a company wants to know, whether there is a positive relationship between the productivity of employees (y) and the score on an aptitude test (x). He has data of 12 employees and assumes the following linear relationship: $y = \beta_0 + \beta_1 \times x + \varepsilon$

R/R Commander output for the simple linear regression model:

```
Call:
lm(formula = y ~ x, data = exercise11_1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4457 -4.9575  0.4418  4.2791  7.7791

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.5394    10.7251   1.915   0.0845 .
x             0.7752     0.2900   2.673   0.0234 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.992 on 10 degrees of freedom
Multiple R-squared:  0.4168,    Adjusted R-squared:  0.3584
F-statistic: 7.146 on 1 and 10 DF,  p-value: 0.02337
```

a. Apply the appropriate test to answer the Research Question. Mention all steps (use $\alpha = 0.05$)

b. Calculate the 95% CI for β_1.

Post-class activity

Watch:

- ☐ *'Prediction Interval in Linear Regression.'* (Duration: 5:40 min.),
- ☐ *'Prediction & Confidence Interval; further details'* (Duration: 12:50 min.).

All of the clips are linked on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 11.2

In preparation of Tutorial 12 please answer the Multiple Choice Questions of the example exam (available on Brightspace > Tutorial 12). The answers will be given and discussed in Tutorial 12.

Exercise 11.3

Based on:

- ☐ **Example 11.2 O&L 7th Edition pp.566-568**, or
- ☐ **Example 11.2 O&L 6th Edition pp.583-585**

This example was used for Exercise 10.3 as well. (Re-)Read the introduction of this example (not the questions below it). Use the R/R Commander output below to answer the following questions:

- Scatter plots → see Tutorial 10, Exercise 10.3, Figure 22
- Summary Simple Linear Regression model (straight line model) → see Tutorial 10, Exercise 10.3
- 95% Confidence Intervals for β_0 and β_1 (see Table 24)

Table 24: 95% Confidence Intervals for β_0 and β_1

	Estimate	2.5 %	97.5 %
(Intercept)	4.6979	-9.0275	18.4232
Purchased_Directly	1.9705	1.6141	2.3269

a. What is the research question in example 11.2? (You have answered this question already in Exercise 10.3a., but good to start with it again.)

b. In Exercise 10.3e. you have applied the omnibus *F*-test to test $H_0: \beta_1 = 0$ against $H_a: \beta_1 \neq 0$. In Tutorial 11 you have seen that you could apply a *t*-test for a regression coefficient as well. Find the outcome of the T.S. *t* for this test in the R output.

c. Check that the squared outcome of the test statistic t is equal to the outcome of the test statistic F of the omnibus F-test.

d. Suppose, in advance it was hypothesized that there is a positive relationship between the percentage of prescription ingredients purchased directly from the supplier and the prescription sales volume. Perform a test ($\alpha = 0.05$) for this hypothesis. Mention all steps. Use the R output for steps 6 and 7 only!

e. Suppose that from previous research it is known that $\beta_1 = 2$. Test ($\alpha = 0.05$) whether this has changed. Mention all steps. Use the appropriate parts of the R output to answer this question.

f. Read the 95% Confidence Interval for β_1 from the R output.

g. Calculate the 90% confidence interval for β_1.

h. Give the interpretation for the confidence interval constructed in question g.

Exercise 11.4

Same case as in Exercise 10.4. However, this time with different questions (except for some useful repetition).

In a study conducted to examine the quality of fish after 7 days of storage on ice, ten raw fish of the same kind and approximately the same size were caught and prepared for storage on ice. Two of the fish were placed in storage immediately after being caught, two were placed in storage 3 hours after being caught, and two each were placed in storage at 6, 9 and 12 hours after being caught.

Let y denote a measurement of fish quality (on a 10-point scale) after 7 days of storage on ice, and let x denote the time after being caught that the fish were placed in storage on ice. The sample data are given in Table 23.

The following model is assumed: $y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$

Furthermore assume that the residuals are independent and normally distributed with standard deviation σ_ε. Use, where appropriate, the provided R/R Commander output to answer the questions:

- Summary Simple Linear Regression model (straight line model) → See Tutorial 10 Exercise 10.4.
- Prediction of new samples:
  - 95% Confidence Interval for $\mu_y = \beta_0 + \beta_1 \times x_{n+1}$ (see Table 25)
  - 95% Prediction Interval for future y at x_{n+1} (see Table 26)

Table 25: 95% Confidence Interval for $\mu_y = \beta_0 + \beta_1 \times x_{n+1}$

x_{n+1}	fit	lwr	upr	se.fit	df
7	7.468333	7.377923	7.558744	0.03920654	8
10	7.043333	6.922390	7.164276	0.05244706	8

Table 26: 95% Prediction Interval for future y at x_{n+1}

x_{n+1}	fit	lwr	upr	se.fit	df
7	7.468333	7.175737	7.760929	0.03920654	8
10	7.043333	6.739910	7.346756	0.05244706	8

a. Formulate the research question.

b. Why is the omnibus F-test not appropriate to answer the research question?

c. Apply the appropriate test ($\alpha = 0.05$) to answer the research question.

d. Give the estimate for σ_ε.

e. Calculate the estimated population mean fish quality score after 7 days of storage on ice, when the fish was stored on ice 10 hours after being caught.

f. Read from the R output the fish quality score after 7 days of storage on ice, when the fish was stored on ice 10 hours after being caught.

g. Explain why the answers to e. and f. are the same.

h. Determine the 95% Confidence Interval for the population mean fish quality score after 7 days of storage on ice, when the fish was stored on ice 10 hours after being caught.

i. Read from the R output the 95% Prediction Interval for the fish quality score after 7 days of storage on ice, when the fish was stored on ice 10 hours after being caught.

j. Explain why the 95% Prediction Interval for y at x_{n+1} is always wider than the Confidence Interval for $\mu_y = \beta_0 + \beta_1 \times x_{n+1}$.

Tutorial 12

Learning objectives

After Tutorial 12, within the framework of a single regression analysis, the student should be able to:

- explain what is meant by extrapolation;
- mention the risks of extrapolation related to prediction;
- tell how correlation and simple linear regression are 'related';
- mention, determine, and interpret the coefficient of determination;
- mention how to check the assumptions with respect to simple linear regression;
- recognize violations of the assumptions.

Relation between correlation and simple linear regression

Read:

☐ **O&L 7th Edition**:

- paragraph 11.6 pp.587-591 (up to "The sample correlation r_{yx} is..."), **or**

☐ **O&L 67th Edition:**:

- paragraph 11.7 pp.608-613 (first 3 lines)

The relation between correlation and simple linear regression for a straight line becomes apparent by:

- the t-test for the slope β_1 is equivalent to the t-test for ρ;
- $r_{yx} = \hat{\beta}_1 \times \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$, or $\hat{\beta}_1 = r_{yx} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = r_{yx} \times \frac{s_y}{s_x}$;
- the coefficient of determination R^2 which is equal to the squared correlation coefficient r_{yx}^2.

Checking the model assumptions of simple linear regression

Read:

☐ **O&L 7th Edition**:

- paragraph 11.2 pp.569 (below Example 11.3)-573, **or**

☐ **O&L 6th Edition:**:

- paragraph 11.2 pp.586 (below Example 11.3)-590

Exercises to be done during the tutorial

The Tutorial 12 presentation contains some of the multiple choice questions from the example exam. In a separate document all answers to the example exam will be made available on Brightspace.

Exercises to be done after the tutorial

For answers/feedback check Brightspace.

Exercise 12.1

Based on:

- ☐ **Exercise 11.57 O&L 7th Edition p.616**, or
- ☐ **Exercise 11.65 O&L 6th Edition pp.649-650**

Read the introduction to this exercise as specified above.

R/R Commander output for this exercise:

- scatter plot with simple linear regression line (Figure 23)

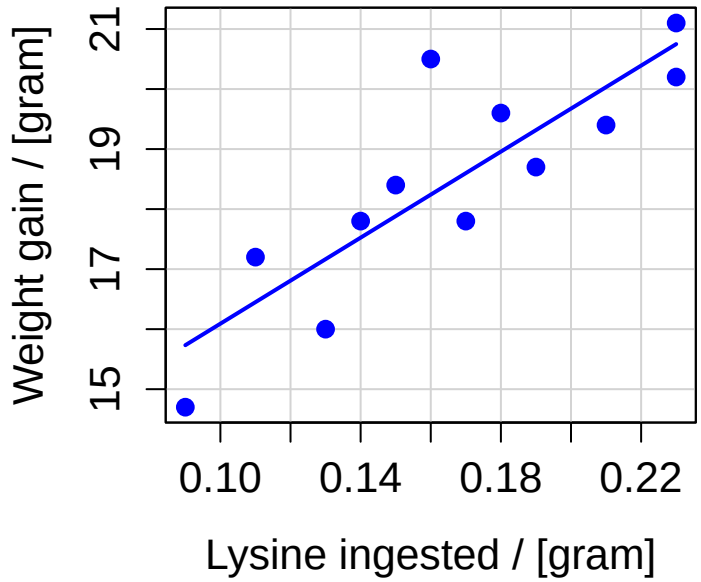


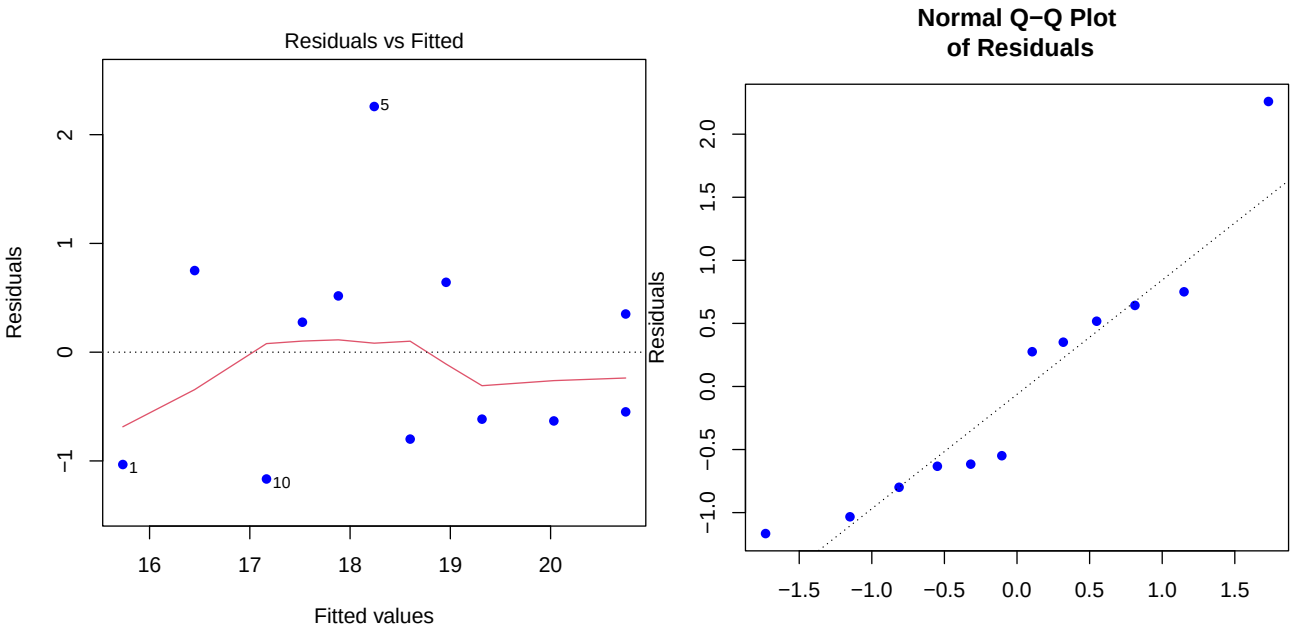
Figure 23: Scatter plot of weight gain (in grams) versus amount of lysine ingested (in grams) with the estimated simple linear regression line.

- Summary Simple Linear Regression model (straight line model)

```
Call:
lm(formula = weight_gain ~ lysine_ingested, data = ex11_57)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1662 -0.6741 -0.1367  0.5486  2.2590

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       12.509      1.192   10.50 1.02e-06 ***
lysine_ingested   35.828      6.957    5.15 0.000431 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.034 on 10 degrees of freedom
Multiple R-squared:  0.7262,    Adjusted R-squared:  0.6988
F-statistic: 26.52 on 1 and 10 DF,  p-value: 0.0004315
```

- Basic diagnostic plots for the simple linear regression straight line model (Figure 24)

Use (where possible) the R/R Commander output to answer the following questions:



(a) Residual plot.

(b) Q-Q plot.

Figure 24: Basic diagnostic plots for `lm(weight_gain ~ lysine_ingested, data = ex11_57);`

a. Mention the symbol of the coefficient of determination and read from the R/R Commander output the estimated value.

b. Give an interpretation for the estimated coefficient of determinination in terms of the actual situation.

c. Explain why the coefficient of determination will increase when the sum of squares of the residuals decreases.

d. Determine the estimated correlation coefficient ρ.

e. When calculated on the same data, will R and r always be exactly the same? Explain your answer.

f. Mention the assumptions for a simple linear regression straight line model.

g. Use (where possible) the R/R Commander output to check the model assumptions one by one. Mention what part of the output you have used for the different assumptions and explain why/why not the assumption holds.

Part II

Computer Practicals

Introduction Computer Practicals

The name of the teacher of your practicum group will be announced at the start of the **Basic Statistics** course. When necessary, contact your teacher by email.

Attendance of the practicals is **compulsory**. The reason for this is, that this is the only way for us to make sure that you will acquire some basic statistical computer skills.

Your attendance of the computer practicals will be registered. Apart from attending them, you will also be required to participate actively in the practicals and to follow directions correctly; your teacher will assess your involvement and performance. For the sake of clarity: going on a winter sports holiday, waiting tables in a restaurant on an afternoon, attending a training for some match, or paying a visit to your parents, who live far away, is no excuse for missing a practicum. The course coordinator will only deviate from the rules in exceptional, distressing situations.

You will work on exercises in a fixed 'team' of two persons, as far as possible. Of course, you are expected to change places at the keyboard regularly. As a team, you shall save all edited files, preferably on your own Wageningen University & Research OneDrive; your teacher can ask you to show these files. Assignments shall be submitted according to your teacher's directions.

You are not allowed to surf the internet, to chat, email, *etc.* during practicals, unless this is specifically required in the context of an assignment.

Contact your practicum teacher in case you cannot attend a computer practicum. In consultation with you, your teacher will look for an opportunity for you to make up for this practicum, possibly under supervision of different teacher.

You will get a pass for the computer practicals only if, according to your teacher, you have attended and actively participated in all practicals. The mark granted for the written examination will only be valid once your computer practicals, the attendance of which is compulsory, have resulted in a pass.

If you failed the computer practicals but passed the written examination, you will be awarded an 'incomplete' as mark; this mark will remain valid for 2 years. Contact the course coordinator, on how to fulfill the computer practicals requirement in order to finish the course and obtain the study credits. If you failed the written examination, you will be awarded that mark, no matter whether you passed the computer practical or not.

Computer Practicum 1

This computer practicum contains the following five parts:

- ☐ Retrieving and saving data files for this and the other computer practicals.
- ☐ Carrying out the *'Smoking and pregnancy'* case study with R Commander.
- ☐ Carrying out the *'Car emissions'* case study with R Commander.
- ☐ Collecting and analyzing data yourself: *'Reaction Time'* with R Commander.
- ☐ Analyzing data with R Commander for the *'UNICEF Education'* case study.

Learning objectives

After this computer practicum the student should be able to do the following in R Commander:

- Retrieve and save data;
- Calculate a new variable;
- Calculate and interpret numerical summaries;
- Make and interpret bar graphs, histograms, and (side-by-side) box plots;
- Enter a new data set.

Part 1 - Downloading the data

The files, needed in the computer practicals of this course, can be downloaded from Brightspace at https://brightspace.wur.nl by choosing the course **Basic Statistics**:

1. Select 'Content' in the top menu.
2. Go to 'Computer Practicals' in the menu on the left of the **Basic Statistics** Brightspace site.
3. In the right side of the screen there will be a blue header named 'BS_Practicals_Data' with specification 'Zip Compressed File'. Click on this blue header named 'BS_Practicals_Data'.
4. The window will change, and display a button labeled **Download** for the '`BS_Practicals_Data.zip`' file. Click on the **Download** button, which will fetch the file and save it into the Downloads folder on your computer.
5. In your personal WUR OneDrive create a folder for the course 'MAT14303 Basic Statistics'. Within this folder create a sub folder for storing the data files for the practicals, e.g., named 'data_practicals'.
6. Go to the Downloads folder on your computer and open the downloaded '`BS_Practicals_Data.zip`' file. Select all files (e.g., using CTRL+A) and copy them (either using mouse right-click > Copy, or CTRL+C).
7. Go to the created folder for storing the data files of the computer practicals in your personal WUR OneDrive and paste the copied files (mouse right-click > Paste, or CTRL+V).

> **!** IMPORTANT
>
> Unzip the files! All course data files needed for the computer practicals are downloaded as **"Zip Compressed file"**.

Part 2 - Analyzing smoking and pregnancy data

Researchers in the United States of America (USA) studied the effect of smoking by pregnant women upon their unborn child between 1960 and 1967. The data set contains data about mothers and their babies. The mothers gave birth in several hospitals in Northern California. The data used here is a subset of the total data set. The file named "`BSP1_Smoking_Pregnancy.RData`" contains the data for this computer practicum. The following variables are given for each mother:

- Length of pregnancy ("`gestation`")
- Birth weight of the baby ("`weight`", in ***ounces*** not ***grams***!)
- Highest level of education of the mother ("`education`")
- Mother smokes (yes/no) during pregnancy ("`smoke`")

The data set does not contain multiple pregnancies for the same woman. Women, who did not smoke during pregnancy, also never smoked before.

Open R (***NOT RStudio***). In the top menu bar go to: **Packages > Load package**... Select the package named "**Rcmdr**" from the list of available packages and click the **OK** button to confirm your selection (a faster way is typing the command: `library(Rcmdr)` at the prompt ">" and executing it by pressing "Enter" on your keyboard). The window as displayed in Figure 25 should appear, which is named R Commander.

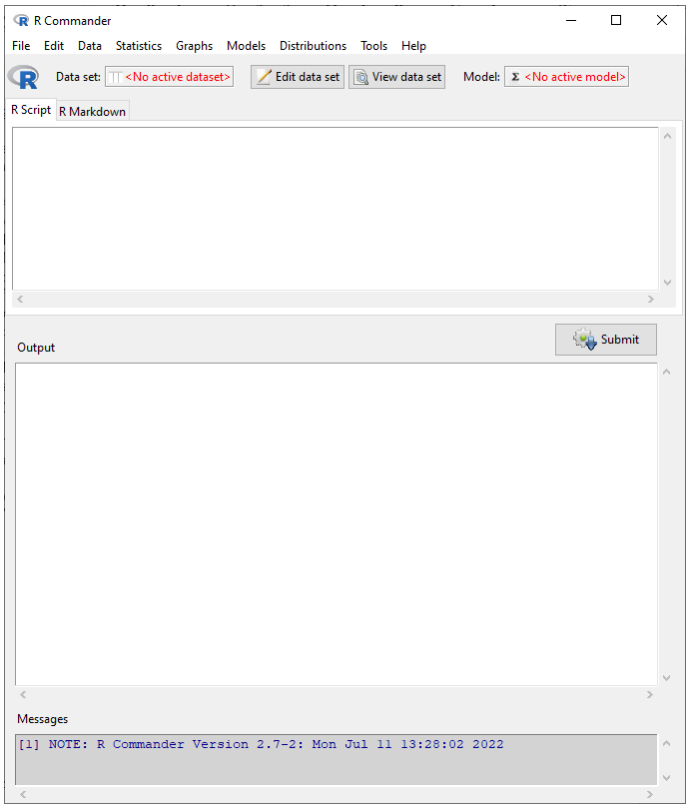


Figure 25: GUI R Commander

Go to: **Data > Load data set**... and choose the file "`BSP1_Smoking_Pregnancy.RData`" with the data set to load. Go to: **Data > Active data set > Select active data set**... and select "`pregnancies_data`" to set as active set. To quickly select or change the active data set click on the field next to **Data set:** just below the top menu bar.

a) Click on **View data set** to get an impression about the data in the set. Indicate, on the answer form, for each of the variables if they are quantitative (discrete or continuous) or qualitative (ordinal or nominal).
b) What are the statistical units?
c) Is this research experimental or observational? Why?
d) For the variable weight convert ounces into grams (1 ounce $\approx$ 28.35 gram and keep in mind that R Commander uses a decimal point.). To achieve this conversion click **Data > Manage variables in active data set > Compute new variable**... Replace the value "`variable`" in the field **New variable name** with a new sensible variable name, e.g., "`weight_grams`". The conversion from ounces to grams should

be written in the field **Expression to compute**, by selecting variables (by double-clicking the required variable(s) from the field **Current variables**) and typing the needed expression for the conversion. Click the **OK** button to confirm and close the window for computing new variables. Use **View data set** to inspect, that a new variable is computed and added to the data set.

e) Give for both groups (non-smoke and smoke) for the variable "`weight_grams`" the mean, the median, the standard deviation, the 1^{st}, and 3^{rd} quartile. The values are found by clicking **Statistics > Summaries > Numerical summaries**... In the window, which opens, select the variable "`weight_grams`" (selecting multiple by pressing CTRL and clicking variable names). Click the button **Summarize by groups**..., select "`smoke`" as groups variable and click **OK** to confirm the selection. Next click the **OK** button to get the numerical summaries by variable "`smoke`". Look at the Output part of the R Commander window for the results and fill in the table on the form.

> Tip: Copying values
>
> By double-clicking on the numbers in the Output part of the R Commander window a number can be selected to copy and paste it into another program, for example a Word or Powerpoint document.

Visualizing data: bar chart, histogram and box plot

When a variable is qualitative or discrete, how many times each possible outcome occurs in the sample can simply be counted directly. A graphical representation can then be drawn, with the outcomes on the horizontal x-axis and the corresponding frequencies or relative frequencies on the vertical y-axis. Such a graph is called a *bar chart*.

For a continuous variable, such as "`weight`", it is more difficult to draw a graph. After all, no two individuals in a sample will have exactly the same weight when weighed very accurately. For a sample of 150 individuals, it makes no sense to create a bar chart with 150 weights on the horizontal axis that all have the same relative frequency of $\frac{1}{150} \approx 0.0067$ on the vertical axis. In this case, the outcomes of the "`weight`" variable are categorized into a limited number of 'bins', for example, < 55, $55 - 65$, $65 - 75$,..., > 115 kg. Relative frequencies for the bins are then calculated, and the visualization proceeds more or less in the same way as for a discrete variable. The graph obtained in this way is called a *histogram*. The bars in the bar chart are now rectangles, with no gap between adjacent bars, where the surface area of the bars (rectangles) is proportionate to the corresponding relative frequencies (and the total area is equal to 1 or 100%). When the bins are not of equal size, a wider bin corresponds to a wider bar (rectangle).

A continuous variable can also be visualized by means of a box plot. The box plot is based on the 5-number summary of John Wilder Tukey (minimum, first quartile, median, third quartile, and maximum).

In the next questions these three different visualizations of the data will be applied to the "`pregnancies_data`" data object in R Commander.

f) Make a separate histogram of the birth weight of babies of mothers, who smoked during their pregnancy, and of mothers, who did not smoke during their pregnancy. Go to: **Graphs > Histogram**..., select the variable "`weight_grams`"'. Click on **Plot by groups**..., and select as **Groups variable:** "`smoke`". Press the **OK** button to confirm the groups variable selection and next click the **OK** button to create the histograms. You can find the graphs in a separate R Graphics window. To copy graphs into another program (e.g., Word/Powerpoint) right-click your mouse cursor on the R Graphics window and select "Copy as bitmap". Use paste (CTRL+V) in the other program (e.g., Word/Powerpoint), where you want the graph to be placed.

> Tip: Record graphics history
>
> If you go in the top menu of the R Graphics window to **History** and choose **Recording**, it is possible to go back- and forward between all graphs made during a R Commander session.

g) Make side-by-side box plots for the birth weight in grams of babies for mothers who smoked during the pregnancy, and for those, who did not smoke during pregnancy. Go to: **Graphs > Boxplot**... Make a sketch of the side-by-side box plots on your answer from.

> **Tip: Cycle graphics history**
>
> If you have set **History > Recording** in the R Graphics window previously, then you can go back to the bar graph you made earlier by **History > Previous** (Page Up button on your keyboard) and go forward to the side-by-side box plots with **History > Next** (Page Down button on your keyboard).

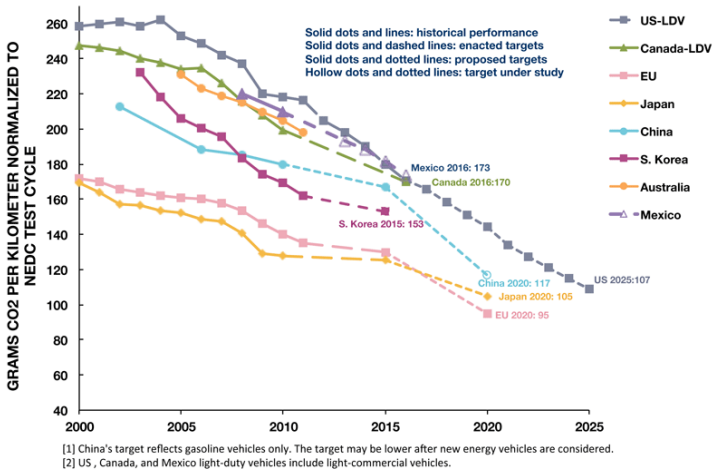
h) If the graph is correct, you will see single points and numbers outside the whiskers of the box plot. What do the points and numbers in the plot represent?
i) Compare the first quartile of both groups calculated in question e) with the side-by-side box plots of question g).
j) Make a bar chart for the variable "`education`" for mothers, who smoked during the pregnancy, and for those who did not smoke during pregnancy. Use **Graphs > Bar graph**. . .. Change on the **Options** tab the **Axis Scaling:** "`Percentages`" by changing the radio button selection in front. Select as **Variable:** "`education`" and **Plot by:** "`smoke`".

Don't close R Commander, because in part 3 R Commander will also be used.

Part 3 - Car emissions

The transport sector is accounting for 14% of the total global greenhouse gas emissions, as reported by the Energy Protection Agency (EPA) in their August 2016 update of the 'Global Greenhouse Gas Emissions Data'. The car industry is under pressure to step-up emissions reduction.

So far, thanks to aggressive emissions targets being set by regulators in the major markets, the industry has managed to achieve considerable cuts in car emissions. (https://www.automotive-iq.com/powertrain/articles/emissions-regulations-force-manufacturers-across-world-clean-their-act)



In Europe, cars are responsible for about 12% of total EU emissions of carbon dioxide (CO_2), the main greenhouse gas. EU legislation sets mandatory emission reduction targets for new cars. The average emissions level of a new car sold in 2016 was 118.1 grams of CO2 per kilometer, considerably below the target of 130 g. If the average CO_2 emissions of a manufacturer's fleet exceed its limit value in any year from 2012, the manufacturer has to pay an excess emissions premium for each car registered.

In the third part of this computer practicum the CO_2 emission from different brands and types of cars will be investigated. These data are found at the website of the Environmental Protection Agency (EPA) of the United States. The data used here are part of a larger data set. In the file "`BSP1_Car_Emissions.RData`" the data for this part of the computer practicum are stored. The following variables are given for each car:

- Brand of car (variable: "`Brand_of_Car`")
- Type of car (variable: "`Type_of_Car`")
- CO_2 emission (grams per mile) (variable: "`CO2gmi`")

Go to: **Data > Load data set**... to load the file named "`BSP1_Car_Emissions.RData`". Change, when necessary, the active data set to "`car_emissions`".

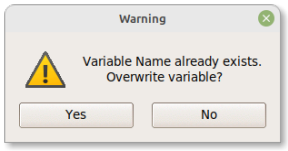
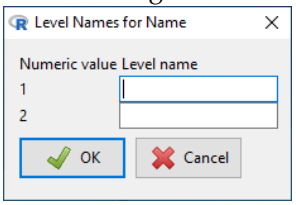
a) Indicate for each of the variables whether they are quantitative (discrete or continuous) or qualitative (ordinal or nominal).
b) The CO_2 emissions in the data set are expressed in grams per mile. Compute a new variable (named: "`CO2gkm`"), that expresses the CO_2 emission in grams per kilometers (1 mile $\approx$ 1.609 kilometer$\rightarrow$ 1.609 kilometer/mile).
c) Make a separate histogram of the CO_2 emissions (g/km) per brand of car. What stands out?
d) Make side-by-side box plots for the CO_2 emissions (g/km) per brand of car. Give a possible reason why Toyota shows a very large spread in CO_2 emissions.
e) Repeat assignment d), but now per type of car. What stands out?

Part 4 - Reaction Time

In some sports reaction time – the ability to respond quickly to a stimulus – is very important. Squash, fencing, table tennis are examples of sports in which quick reaction times are very important. But also in traffic, the driver's reaction time can be extremely important. In this part your reaction time will be measured and recorded. Online there are many tests available for testing your reaction time. The following website was chosen to measure your reaction time: https://www.justpark.com/creative/reaction-time-test/.

Go to: **Data > Load data set**... and load the file named "`BSP1_Reaction_Time.RData`". Change, when necessary, the active data set to "`reaction_time`".

a) Go to: **Data > Manage variables in active data set > Convert numeric variables to factors**... Leave the settings as given and click on the **OK** button.

> ⚠ R Commander Warning
>
> When you pressed the **OK** button, a warning message "Variable Name already exists. Overwrite variable?" is shown. Click the **Yes** button to continue.
>
> 
>
>
> After clicking the **Yes** button, the following window appears:
>
> 
>
>
> Fill in your name and the name of your computer practicum partner. When working alone use, for example, "`Series A`", and "`Series B`".

b) Click the **Edit data set** button, located just below the top menu bar of the R Commander window. Do the online Reaction Time Test and enter the reaction time in the data set "`reaction_time`". Repeat this 10 times. When you work in pairs, each person is doing the online Reaction Time Test 10 times. When you work alone, please determine two series of 10 reaction times (you can, for example, see, whether the reaction times in the second series is faster than the first series).
c) Try to think about an adequate visualization,whether the reaction times of you and your partner (or the two series) are about the same or different.
d) Calculate summary statistics of the reaction time for each person separately.
e) Compare the mean and median of the reaction time for each person separately. Would you use the median or the mean? Why?
f) Compare the Inter Quartile Range (IQR) and standard deviation of the reaction time for each person (or each series) separately. Would you use the IQR or the standard deviation? Why?

g) Who, you or your partner (or which of the two series), has the lowest variability in reaction time?

Part 5 - UNICEF Education case study.

Go to: **Data > Load data set**... and load the file "`BSP1_UNICEF_Education.RData`". Change, when necessary, the active data set to "`UNICEF_education`".

The data comes from the website: https://data.unicef.org, and tells something about the attendance of children at secondary schools. The data is from 60 different countries on four different continents. The variable "`Total`" represents the percentage of children attending secondary school within each country. The variables "`Male`", "`Female`", "`Urban`", and "`Rural`" give percentages for the genders or sub-groups, indicated by the name of the variable, of secondary school attendance within each country.

To get a feeling about the data contained, click on the **View data set** button.

a) Make an appropriate graph (histogram, box plot or bar chart) to visualize the number of countries from each continent.
b) Make a side-by-side box plot for the variable "`Total`" to see the difference in attendance of children at secondary schools between the four continents in the data set.
c) Compute a new variable ("`d_Female_Male`"), which indicates the difference in attendance between females and males at secondary school (use for **Expression to compute:** `Female - Male`).
d) Make a side-by-side box plot of the new variable for the four continents.
e) Suppose that UNICEF would like to invest in a country, in which girls are more likely not to be in secondary school than boys. Which country should they select, if they would like to reduce the largest gap observed? In other countries the attendance at secondary school by boys is less boys than for girls. Give the country with the largest gap in this way.
f) Can you think of reasons, why there are such marked differences in secondary school attendance between boys and girls in different countries?
g) Compute a new variable ("`d_Urban_Rural`"), which indicates the differences in attendance at secondary school between urban and rural areas. Make a side-by-side box plot for the four continents for this variable.
h) When UNICEF has money to reduce the gap between education in urban and rural areas in the continent Americas, which country would be the obvious choice?

Computer Practicum 2

This computer practicum contains the following four parts:

- ☐ Binomial Distribution.
- ☐ Normal Distribution.
- ☐ Estimators and estimates for (population) mean μ, and (population) standard deviation σ.
- ☐ Binomial Distribution: Testing a hypothesis with predatory mites.

Learning objectives

After this computer practicum the student should be able to do the following in R Commander:

- Plot and interpret the Binomial distribution;
- Determine probabilities from the Binomial distribution;
- Perform a simulation using a Binomial situation;
- Produce and interpret a plot of a normal distribution;
- Determine probabilities from a normal distribution;
- Indicate tail probabilities;
- Identify estimates for the (population) mean μ and (population) standard deviation σ;
- Make a Q-Q plot.

Part 1 - Binomial Distribution

The following questions are related to the binomial distribution. Use R Commander to answer the questions. Use the answer sheet, that is handed out by the teacher, to write down your answers.

You can use the following information to answer questions a) to g):
When rolling a fair die, the probability of obtaining the outcome 6 equals $\frac{1}{6} \approx 0.1667$. Assume the fair die is rolled five $(n = 5)$ times, and let y be the number of times obtaining the outcome 6 in 5 rolls of a fair die. Then $y \sim \text{Bin}(n = 5,\ \pi = \frac{1}{6})$.

a) Visualize the binomial distribution mentioned above; $\text{Bin}(n = 5,\ \pi = \frac{1}{6})$. Go to: **Distributions > Discrete distributions > Binomial distribution > Plot binomial distribution**... Use for **Probability of success:** 0.1667 and provide the desired n in the field behind **Binomial trials**. Make a sketch of the graph on the answering form.

> **i** Notes about the resulting Binomial Distribution plot.
>
> The resulting graph will be displayed in a separate R Graphics window:
>
> - If necessary resize the window to see all text of the title above the binomial distribution!
> - On the x-axis the Number of successes 5 is not included, because this probability is extremely

small. However, 5 times obtaining the outcome 6 out of 5 rolls of a fair die is a possibility even when the probability is extremely small.

b) Calculate the probability of obtaining 0 times the outcome six in 5 rolls of the fair die. Go to: **Distributions > Discrete distributions > Binomial distribution > Binomial probabilities**... The probabilities for each number of successes, $P(y = k) \; \forall \; k \in \{0, 1, \ldots, 5\}$, are shown in the part Output of the R Commander window. After you have found the answer in the part Output of the R Commander window and have filled in the answer on the form, try to get the same answer:

- from the graph you have made under a).
- using your graphing calculator (if you have one).

c) Calculate the probability of obtaining 3 times the outcome six in 5 rolls of the fair die. After you have found the answer in the part Output of the R Commander window and have filled in the answer on the form, try to get the same answer:

- from the graph you have made under a).
- using your graphing calculator (if you have one).

d) Calculate the probability of obtaining at most 1 time the outcome six in 5 rolls of the fair die. Go to: **Distributions > Discrete distributions > Binomial distribution > Binomial tail probabilities**... Fill the fields behind **Variable value(s)**, **Binomial trials**, **Probability of success** with the appropriate values and select the appropriate tail for the calculation. After you have found the answer in the part Output of the R Commander window and have filled in the answer on the form, try to get the same answer:

- from the graph you have made under a).
- using your graphing calculator (if you have one).

e) Determine the probability of obtaining at least 2 times the outcome six in 5 rolls of the fair die by using the answer of d) and the complement rule. When you have found the answer, try also to find the answer by using R Commander.

i Note on calculating upper tail probabilities in R Commander and R in general

When selecting the **Upper tail** in probability calculations, R Commander (and R in general) calculates $P(y > k)$ and not $P(y \geq k)$. Think carefully, what you should enter in the field behind **Variable value(s)**, which represents k, to get the correct answer.

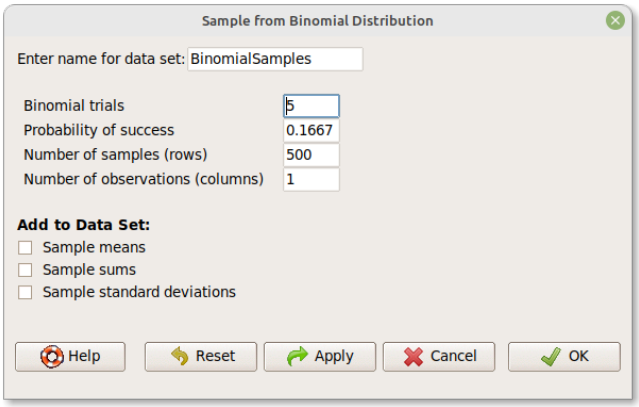


Figure 26: Settings for Part 1 Question f).

f) Simulate 500 times rolling a fair die 5 times and counting the number of times of obtaining the outcome six. Go to: **Distributions > Discrete distributions > Binomial distribution > Sample from binomial distribution**..., use the settings shown in Figure 26 and click the **OK** button to execute. Next:

1. Have a look at the created data object "BinomialSamples", using **View data set**.

2. Calculate the mean of the variable "`obs`" with R Commander (calculating the mean of a variable in R Commander was a topic in Computer Practicum 1).
3. Calculate the (population) mean (or expected value) for the number of times obtaining the outcome six in 5 rolls with a fair die, as discussed in Tutorial 3.
4. Use the table in your answer form to calculate the same answer as above in 3. Hint: fill the table using the output created in question b).

- Are the three answers in 2., 3., and 4. exactly the same? Comparable? Can you explain why / who not?

> Tip: R Commander as advanced calculator
>
> R Commander can be used as a (statistical) calculator. Try it:
>
> - Type on a new line in the R Script part of the R Commander window: `100 * 0.256` (keep the blinking cursor on the same line)
> - Press the **Submit** button situated on the right side between the R Script and Output part in the R Commander window
> - The answer will appear in the Output part of the R Commander window

g) Plot the simulated results. Go to: **Graphs > Plot discrete numeric variable**..., because of the contents of the "`BinomialSamples`" object R Commander does not allow making a bar graph. Compare the resulting plot with the plot made in Part 1 a), and explain similarities / differences.

Part 2 - Normal Distribution

The following questions are related to the Normal Distribution. Use R Commander to answer the questions.

a) Determine $P(y > 12)$, when it is given that $y \sim \mathrm{N}(\mu = 12,\ \sigma = 5)$. Go to: **Distributions > Continuous distributions > Normal distribution > Normal probabilities**... Fill the fields **Variable value(s)**: 12, **Mean**: 12, **Standard deviation**: 5, and choose **Upper tail** by switching the radio button. Click the **OK** button to execute.

> ! Important
>
> Try to understand, what the filled numbers mean with respect to the asked probability, and why to use **Upper tail**.

b) Determine $P(y < 10) \rightarrow y \sim \mathrm{N}(12,\ 5)$.

c) Give the 95^{th} percentile of the distribution $y \sim \mathrm{N}(12,\ 5)$. Go to: **Distributions > Continuous distributions > Normal distribution > Normal quantiles**....

d) Give the 95ˆth percentile of the distribution $x \sim \mathrm{N}(0,\ 1)$ (Standard Normal Distribution).

e) Use the answer of d) to determine the 5^{th} percentile of the distribution $x \sim \mathrm{N}(0,\ 1)$.

Use the following information to answer questions f) to i):
Assume that the height of a random male student (y) is normally distributed with expected value $\mu = 182$ and standard deviation $\sigma = 7$ (cm).

f) Display a graphical representation of this normal distribution. Go to: **Distributions > Continuous distributions > Normal distribution > Plot normal distribution**...

g) Calculate the probability of a student being taller than 190 cm.

h) Make a visualization of the probability of question f. Go to: **Distributions > Continuous distributions > Normal distribution > Plot normal distribution**... and fill in the numbers like as given in the screenshot shown in Figure 27. By clicking behind **color** on "`#BEBEBE`", the color of the area below the density function can be changed.

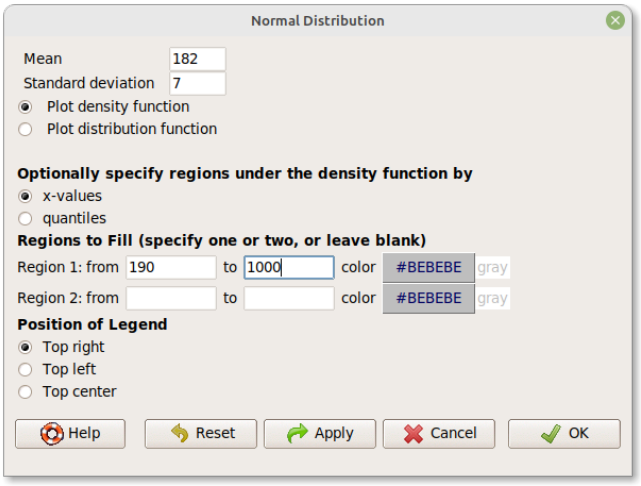


Figure 27: Settings for Part 2 Question h).

i) Calculate the probability, that a student is deviating more than one standard deviation from the (population) mean (or expected) height of a male student.

j) Make a visualization of the probability of Part 2 Question i). Color the probability in the left-tail green ("`#00FF00`") and in the right-tail orange ("`#FFA500`").

> **Note: Coloring regions under a Normal Distribution**
>
> **Region 1: from** ... **to** ..., and **Region 2: from** ... **to** ... should be read from left to right.
>
> Therefore, for the green left-tail fill **Region 1: from** 0 **to** ... (specify your upper boundary).
> For the orange right-tail fill **Region 2: from** ... **to** 1000 (specify your lower boundary).

Part 3 - Estimators and estimates for μ, and σ

A machine fills packages with sugar. The probability distribution of the weight y of a random pack of sugar is a normal distribution with mean $\mathrm{E}(y) = \mu_y$ and variance $\mathrm{var}(y) = \sigma_y^2$. To estimate the parameter μ_y, 10 packs of sugar are selected randomly from the production line and weighed (in g).

The weights are as follows: 510, 525, 560, 515, 455, 465, 510, 505, 540, 485.

a) Create a new data set in R Commander by going to: **Data > New data set**...:

   1. Give it a sensible name, e.g., "`sugar_packages`".
   2. Next click the **Add row** button to get 10 rows in the data set.
   3. Click the field labeled "`V1`" and enter a sensible variable name (e.g., "`weight`").
   4. Enter the weights of the randomly selected packages of sugar, and finally click the **OK** button to create the data set.

b) Use R Commander to calculate the (sample) mean and (sample) standard deviation.

> **Tip: Selecting which numerical summary values to calculate**
>
> **Statistics > Summaries > Numerical Summaries**... allows for selection of the numerical summary values to calculate on the tab **Statistics**. Tick or untick the boxes in front of the numerical summary value you want to select or remove.

c) Fill in the sentence on the answer sheet regarding the (sample) mean $\bar{y}$, by circling the correct answers.

d) Fill in the sentence on the answer sheet regarding the (sample) standard deviation s, by circling the correct answers.

e) Make a Q-Q plot, and judge whether the normality assumption for the weights of randomly selected sugar packages holds here. Go to: **Graphs > Quantile-comparison plot**. . ., and click the **OK** button to create the plot.

Part 4 - Binomial Distribution: Testing a hypothesis with predatory mites

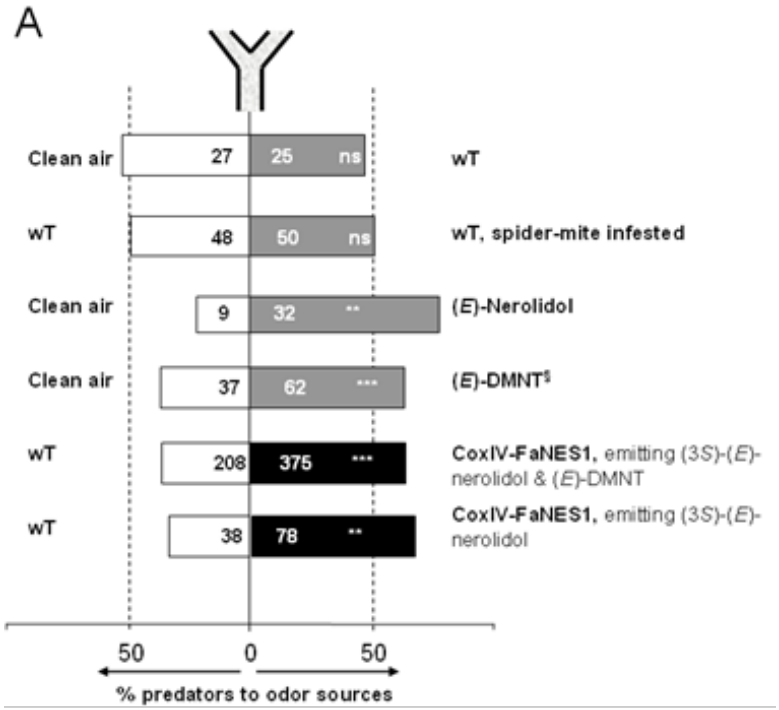


Figure 28: Taken with permission from: Kappers, I.F. *et al.*, 2005, *Science* **309**, pp.2070–2072

Figure 28 is taken from an article in Science authored by the Laboratory of Entomology at Wageningen University & Research on odorous substances to attract predatory mites to *Arabidopsis thaliana* plants (commonly known as thale cress, mouse-ear cress, or arabidopsis).

The figure presents the results of six different experiments with an olfactometer, or Y-shaped tube (as shown at the top in Figure 28). The upper 'bar' (in Figure 28) represents an experiment in which, for each predatory mite, clean air was blown into one end and air with volatiles from a specific plant (in this case wild Type, or **wT** thale cress) into the other end. A total of 52 predatory mites walked upwards to the end of the olfactometer; 27 mites preferred clean air and 25 air with volatiles from the **wT** thale cress plant. This outcome appeared to be non-significant (ns as indicated in the 'bar' of the experiment).

In this part of the computer practicum use R Commander to take a closer look at the outcomes of a part of the thale cress experiments presented in Figure 28. Denote the answers on the anwser form, handed to you by your computer practicals teacher.

Experiment comparing Clean air to (*E*)-Nerolidol

Note that this is the third bar from the top in Figure 28

Let parameter π be the probability, that a predatory mite chooses the odorous substance (*E*)-nerolidol. The researchers wish to show that predatory mites are attracted by the odorous substance (*E*)-nerolidol.

a) Suppose that the mites are **not** attracted by the odorous substance: what proportion of mites are expected to walk to the end of the Y-shaped tube with the odorous substance (*E*)-nerolidol?

b) If the mites are attracted by the odorous substance (*E*)-nerolidol, do you expect the proportion you mentioned under a) to be higher of lower, or can it be either one of them?

c) Given your answers to a) and b) formulate the hypothesis of the researchers (i.e., what the researchers wish to show) in terms of population proportions.

d) What is the probability distribution of the number of mites, that choose for the odorous substance, under the assumption that predatory mites do not have a preference? Denote your answer in the form: $y \sim \ldots$.

e) How many predatory mites (out of the 41 used in this experiment) are expected to choose air with the odorous substance, when predatory mites do not have a preference?

f) Do you expect a higher or a lower number of predatory mites to choose the odorous substance, when predatory mites do have a preference?

g) How many mites eventually did choose the odorous substance in this experiment? In other words: how many 'successes' ($= y$) were actually observed?

h) According to Figure 28, the p-value is significant (**: p-value < 0.01). The p-value will be discussed in an upcoming tutorial, but here it is the probability to observe at least the number of successes as in answer g), when the predatory mites do not have a preference. Use R Commander to calculate this probability (i.e., the p-value). When using **Upper tail**, remember that R Commander calculates $P(y > k)$ (where k represents the number of successes).

i) Beforehand the researchers wanted to have a probability (here the p-value) of 0.05 or smaller, in order to state that it was proven that mites have a preference for the odorous substance (E)-nerolidol. Give your conclusion based on the probability you have found.

The questions, just answered, are actually hypothesis testing. In upcoming tutorials hypothesis testing will discussed in more detail, including all associated terms and concepts. The questions of Part 4 can be considered as a preview or an 'amuse' of what is about to follow.

Computer Practicum 3

This computer practicum contains the following three parts:

- ☐ Simulating a sampling distribution of the mean.
- ☐ Sampling distribution of the sum and the mean.
- ☐ Hypothesis test and $(1-\alpha) \times 100\%$ Confidence Interval (CI) for a (population) mean μ_y.

Learning objectives

After this computer practicum the student should be able to do the following in R Commander:

- Simulate a sampling distribution for the sum and the mean;
- Identify estimates for the (population) mean μ and (population) standard deviation σ;
- Construct a $(1-\alpha) \times 100\%$ Confidence Interval (CI) for population mean μ;
- Apply the Central Limit Theorem;
- Apply a t-test for a population mean μ.

Part 1 - Simulate sampling a normal distribution: Sampling distribution of the mean

The following questions are related to the sampling distribution of the mean. Use R Commander to answer the questions.

Use the following information to answer questions a) up to and including c):
Assume that the height of a random male student (y) is normally distributed with expected value $\mu = 182$ and standard deviation $\sigma = 7$ (cm).

a) Generate 1000 random samples each with same sample size $n = 20$. So, 1000 times a sample is taken of 20 male students with a random height from the distribution $\mathrm{N}(\mu = 182,\ \sigma = 7)$. Go to: **Distributions > Continuous distributions > Normal distribution > Sample from normal distribution**... Enter the correct value for the **Mean**, **Standard deviation**, **Number of samples (rows)**, and **Number of observations (columns)**. Place a check mark in the box in front of **Sample standard deviations**. The box in front of **Sample means** will contain a check mark by default. When filled in correctly, the settings should match Figure 29. Click the **OK** to create the data set "`NormalSamples`". Take a look at the generated data set "`NormalSamples`", and make an appropriate graph of the variable "`mean`". Sketch the graph on the answer form.

b) Explain why the sample means of the 1000 samples of size $n = 20$ observations from the normal distribution are not all exactly equal to 182. Fill in the blank in the statement on the answer form, to indicate what the value of the mean of one row in the data set represents in terms of the (population) mean (or expected) value μ.

c) Fill in the blank in the statement on the answer form, to indicate what the value of the standard deviation of one row in the data set represents in terms of the (population) standard deviation σ. What value do you expect, when you average the 1000 sample variances? Compute the mean sample variance to check

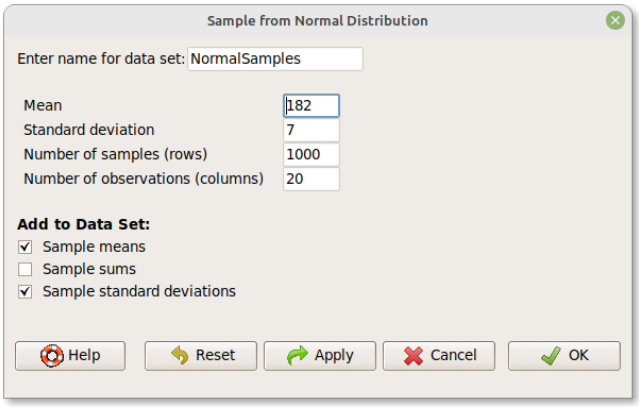


Figure 29: Settings for Part 1 Question a).

this. [Hint: first compute a new variable, e.g., "`variance`", which represents the sample variance of each row. To calculate the squared value of variable "`x`" use "`x^2`".]

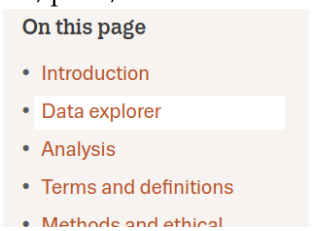
d) Calculate the mean and standard deviation of the distribution of the 1000 samples means [Hint: **Statistics > Summaries > Numerical summaries**... for the variable "`mean`" , while asking for **Mean** and **Standard Deviation** on the **Statistics** tab]. Write your answers on the answer form.

Part 2 - Sampling distribution of the sum and the mean

In the article "Spatial differences and temporal changes in illicit drug use in Europe quantified by waste water analysis" (*Addiction* (2014), **109**, pp.1338-1352) waste water is analysed to monitor the use of drugs in 11 countries and 42 cities in Europe. This research provides complementary evidence on drug consumption to traditional surveillance data. Waste water analysis can measure drug use more quickly, and regularly, than the current national surveys.

This part of the computer practicum will focus on the use of cocaine. This can be measured by determining the concentration Benzoylecgonine (BE load [mg / 1000 people / day]).

a) Open a web browser and surf to the following URL: https://www.emcdda.europa.eu/publications/html/pods/waste-water-analysis_en. When the website appears click in the sidebar : "Data explorer".



Name five cities, which are the five cities with the highest cocaine consumption in 2017, by selecting **Study year**: "`2017`", and **Select a city**: "`NL:Amsterdam`" in the top right corner of the data explorer.
If you look specifically at Amsterdam, can you conclude that the use of cocaine in 2017 is higher in the weekends than on weekdays?

Use the following information for the questions b) to g). Suppose the BE load in waste water in Amsterdam (y) for a random day of the year has a normal distribution with an expected value of 650 and a standard deviation of 250 [mg/ 1000 people / day]. Use R Commander to answer the questions.

b) Calculate the probability that on a random day the BE load is less than 500 [mg / 1000 people / day]. Visualize this probability in a graph.

c) Calculate the probability that the mean BE load of 16 randomly selected days in a year is less than 500 [mg / 1000 people / day]. Note that this question is about the sampling distribution of the mean, in which the expected value is ... and the standard deviation is ... $\rightarrow \bar{y} \sim \mathrm{N}(\ldots, \ldots/\sqrt{\ldots})$. Write your answers on the answer form.

d) Determine the probability, that the mean BE load of 16 randomly selected days in a year is between 600

and 700 [mg / 1000 people / day]. Visualize this probability in a graph.

e) Determine the probability, that the total sum of the BE loads of 16 randomly selected days in a year is above 11000 [mg / 1000 people/ 16 days].). Note that this question is about the sample distribution of the sum, in which the expected value is ... and the standard deviation is ... $\rightarrow \sum y \sim \mathrm{N}(n \times \mu,\ \sqrt{n} \times \sigma)$.

f) A statistician is raising questions about applying the normal distribution to model the BE load. In the weekends there are a lot of parties going on in Amsterdam. Therefore, she is thinking that the distribution of BE load is skewed to the ...

g) When the sample size is raised to 49 randomly selected days, the statistician agrees with the application of the normal distribution for the mean BE load of 49 days. Can you explain why the statistician would accept the application of the normal distribution for the mean BE load of 49 randomly selected days?

Part 3 - Hypothesis test and $(1-\alpha) \times 100\%$ CI for a (population) mean μ_y

In a recent study at Wageningen University & Research a research group focused on the influence of different light conditions on plant growth for the species *Arabidopsis thaliana* (commonly known as thale cress, or mouse-ear cress).

The researchers have compared a fluctuating light setting, and a slow increase and decline of the light intensity (parabolic light) to the more standard constant lighting during daytime. While the parabolic lighting seems to have a positive impact on the weight of the plants, the fluctuating light suggests an adverse effect. From literature it is known, that the average weight of a plant under standard light conditions is about 0.18 g. It can be assumed, that plant weight is normally distributed.

The research group wants to show using 25 *Arabidopsis thaliana* plants, that the fluctuating light treatment will indeed result in a smaller average weight. When testing, use $\alpha = 0.05$.

a) What is the appropriate test to apply in this situation? Give arguments for your answer.

b) Apply the first five steps of the test procedure for the appropriate test.

After the first 5 steps of the procedure, it is time to conduct the experiment/research. For the following steps the data can be used.

c) Load the data in the file "`BSP3_Weight_AT.RData`" into R Commander.

d) Next apply the one sample *t*-test in R Commander. Click in the top menu bar of the R Commander window **Statistics > Means > Single-sample t-test**... Fill **Null hypothesis: mu =** 0.18. Choose, by setting the radio button, the correct Alternative Hypothesis for what you have denoted at step 1 of the procedure. **Confidence level:** 0.95. Click the **OK** button to execute the test.

e) Now proceed with the test procedure and fill in the answer form.

f) One of the assumptions for the one-sample *t*-test is that the variable y is normally distributed. Check this assumption in R Commander.

g) Generate new output to get the appropriate 95% Confidence Interval for μ.

h) Give an interpretation of this Confidence Interval in terms of the actual problem.

Computer Practicum 4

This computer practicum contains the following three parts:

- ☐ Tasting coffee.
- ☐ Weight of bags with coffee beans.
- ☐ Distinguishing between organic and regular coffee.

So far during the tutorials three different *t*-tests have been introduced. The difficult part is not applying the tests, but choosing the appropriate test.

In this computer practicum three different situations are present, where the appropriate test needs, or tests need, to be chosen to answer the research question. To aid you, an answer form will be supplied to you by your computer practicum teacher with a specific structure.

The example, provided in the next section, shows how the answer form should be filled in. Review the example and answers carefully, and try to understand it, before you start with Part 1.

Example

In 1996 the American Department For International Development (DFID) has donated 283.300 cubic ton of soy beans enriched with vitamin C (CSB). The enrichment was such that the expected amount of vitamin C equaled 40 mg / 100 g.

In 8 random samples containing 100 g of CSB each, the measured amounts of vitamin C in mg per 100 g CSB were found to be: 26, 31, 23, 22, 11, 22, 14, 31

Table 27 shows the answer form filled for the example.

Table 27: Answer form filled for the example

	To be filled before collecting data
Research question	*Does the expected amount of vitamin C per* 100 *g CSB deviate from* 40 *mg?*
Name of the test	*one-sample t-test*
variable(s) (symbol(s) & definition(s))	y = *amount of vitamin C (mg /* 100 *g CSB)*
parameter(s) of interest (symbol(s) & definition(s))	μ_y = *(population) mean amount of vitamin C (mg /* 100 *g CSB)*
null hypothesis (H_0)	$H_0: \ \mu = 40$
alternative hypothesis (H_a)	$H_a: \ \mu \neq 40$
Test Statistic (T.S.) (symbol & formula)	$t = \frac{\bar{y}-\mu_0}{SE(\bar{y})} = \frac{\bar{y}-40}{(s_y/\sqrt{8})}$
Distribution T.S. under H_0	*Student's t-distribution with* $\nu = 7$ *degrees of freedom.*
Behavior T.S. under H_a	~~*higher*~~ / ~~*lower*~~ / *higher or lower*
sidedness *p*-value	~~*right tailed*~~ / ~~*left tailed*~~ / *two-tailed*
	Use collected data in R Commander and fill in the results

Levene's test	*Applicable: ~~yes~~ / no* H_0 : H_a : *Outcome T.S.:* $F \approx \ldots$ *p-value:...* *Assume equal variances: yes / no*
(point) estimate parameter	$\bar{y} = 22.5$
SE of the estimate	$\text{SE}(\bar{y}) \approx 2.54$
outcome T.S.	$t \approx -6.88$
p-value	$2 \times P(t \leq -6.88) \approx 0.0002$
Statistical conclusion	$(p\text{-value} \approx 0.0002) < (\alpha = 0.05)$ *Reject* H_0*, and* H_a *has been shown.*
Conclusion	*It is shown (when* $\alpha = 0.05$*) that the expected vitamin content in 100 grams CSB deviates from* 40 *mg, in such a way that it is actually less than* 40 *mg (the estimate is lower than the hypothesized value).*
Confidence Interval	95% *CI for* $\mu_y = (16.49,\ 28.51)$
Possible error Type I or II	*It is possible, that a Type I error has been made.*
Rejection Region (R.R)	$\lvert t \rvert \geq 2.365$ *(from O&L Table* 2 *with* $\alpha/2 = 0.025$ *and* $\nu = 7$ *degrees of freedom)*

Learning objectives

After this computer practicum the student should be able to perform the following tests in R Commander, explain why the chosen test is the appropriate one given the research question and data, and interpret the R Commander output:

- t-test for a (population) mean μ, a.k.a one sample t-test;
- Levene's test;
- t-test for the difference in (population) means $\mu_1 - \mu_2$, a.k.a. independent samples t-test;
- paired samples t-test.

Part 1 - Tasting Coffee

An experiment investigated, whether consumers were more positive about the taste of a certain brand of coffee in a test situation at home compared to tasting the coffee in a laboratory environment. Ten randomly selected consumers were asked to score their judgement about the coffee at home and in a laboratory environment by moving an arrow on a vertical axis with numbers between 0 and 100.

a) Fill in the part "**To be filled before collecting data**" of the answer form for **Part 1**.

The data collected from the experiment is given in Table 28.

Table 28: Judgement scores by consumers for tasting at home and in a laboratory

Consumer	1	2	3	4	5	6	7	8	9	10
Judgement score at home	50	76	81	60	30	70	74	64	76	70
Judgement score in the lab	55	74	79	49	34	65	68	66	73	64

b) Create a new data set in the appropriate way, with a sensible name, e.g., "`tasting_coffee`", in R Commander by going to: **Data > New data set**... using the data from Table 28. Think about the number of rows and columns required to properly reflect the given case.

> **i** Part 1c)
>
> An additional variable or additional variables may be required to check the normality assumption.

c) Check with the appropriate graph(s), whether the normality assumption holds for the variable(s). Mention the graph(s), the variable(s) you used to make the graph(s), and your conclusion(s) with respect to normality. Write your answers on the answer form.

d) Check the mean(s), standard deviation(s), and standard error(s), i.e., make summaries of the variable(s) in R Commander and write down the values on the answer form.

e) Perform Levene's test if appropriate and fill in the row for Levene's test on the answer form. For Levene's test go to: **Statistics > Variances > Levene's test**. . ., select **Center**: "`mean`" by changing the radio button. Click the **OK** button to perform the test.

f) Continue with the test procedure to answer the research question about tasting coffee: perform the appropriate analysis in R Commander. Use $\alpha = 0.05$ and indicate to create a 95% Confidence Interval.

g) On the answer form, fill in the part of the large table in the first column for Part 1 up to 'Confidence Interval'.

h) Decide, whether the 95% Confidence Interval can be read from the output generated at Part 1f), or that new output is required. In case of the latter, please do so. Denote the correct 95% Confidence Interval for the parameter(s) of interest on the answer form.

i) Could an error of Type I or II have been made here? Explain your answer.

j) Instead of using the p-value, also the rejection region (R.R.) could have been used. Use R Commander to find the correct rejection region. Go to: **Distributions**, choose quantiles for the correct distribution.

Part 2 - Weight of bags with coffee beans

A machine at Lavazza, a famous brand of coffee, packages coffee beans in bags weighing half a kilo (500 grams). A random selected half-kilo bag with coffee beans has a net weight y, which is assumed to be normally distributed with expectation μ_y and standard deviation σ_y. Both μ_y and σ_y are unknown, and expressed in the physical unit grams [g].

From the packaging machine at the Lavazza factory 10 half-kilo bags with coffee beans are randomly selected to check whether the weight is in line with the machine settings. These 10 bags are, therefore, a simple random sample.

a) Fill in the part "**To be filled before collecting data**" of the answer form for **Part 2**.

b) Load the data in the file "`BSP4_Weight_Coffee_Bags.RData`" into R Commander.

c) Check with the appropriate graph(s), whether the normality assumption holds for the variable(s). Mention the graph(s), the variable(s) you used to make the graph(s), and your conclusion(s) with respect to normality. Write your answers on the answer form.

d) Check the mean(s), standard deviation(s), and standard error(s), i.e., make summaries of the variable(s) in R Commander and write down the values on the answer form.

e) Perform Levene's test if appropriate and fill in the row for Levene's test on the answer form. For Levene's test go to: **Statistics > Variances > Levene's test**. . ., select **Center**: "`mean`" by changing the radio button. Click the **OK** button to perform the test.

f) Continue with the test procedure to answer the research question about weight of coffee bean bags: perform the appropriate analysis in R Commander. Use $\alpha = 0.10$ and indicate to create a 90% Confidence Interval.

g) On the answer form, fill in the part of the large table in the first column for Part 2 up to 'Confidence Interval'.

h) Decide, whether the 90% Confidence Interval can be read from the output generated at Part 2f), or that new output is required. In case of the latter, please do so. Write down the correct 90% Confidence Interval for the parameter(s) of interest on the answer form.

i) Could an error of Type I or II have been made here? Explain your answer.

j) Instead of using the p-value, also the rejection region (R.R.) could have been used. Use R Commander to find the correct rejection region. Go to: **Distributions**, choose quantiles for the correct distribution.

Part 3 - Distinguishing between organic and regular coffee

The following is based on an article form *Resource* at Wageningen University & Research.

A laboratory test can help distinguish between organic coffee and regular (i.e, non-organic) coffee, and, as a result, can help expose fraud concerning food authenticity. Researchers at Wageningen Food Safety Research have shown this by letting a machine 'smell' coffee aroma. Various types of coffee could be distinguished by looking at the mixture of substances responsible for the aroma. The researchers extracted air from a bottle containing half a gram of coffee. A so-called mass spectrometer determined exactly which aromatic substances were present in the extracted air. Because coffee can release about 900 different substances for the aroma, each type of coffee has its own aromatic profile, which can be considered like a fingerprint of the coffee. These aromatic profiles appeared to be very different for organic coffee and regular coffee.

In this part of the computer practicum, focus will be on one specific aromatic substance: ethyl-dimethyl-pyrazine, also referred to ion 137 (variable: "`ion137`"). This substance has a real coffee aroma. A significant difference in measured intensity of ion 137 between organic coffee and regular coffee, when running a two-tailed hypothesis test, would indicate that this aromatic substance can be of help to distinguish between organic coffee and regular coffee.

The data set consists of 65 observations for regular coffee and 43 observations for organic coffee. The observations can assumed to be independent and normally distributed within both groups, as well as independent between both groups.

a) Fill in the part "**To be filled before collecting data**" of the answer form for **Part 3**.

b) Load the data in the file "`BSP4_Regular_Organic.RData`" into R Commander.

c) Check with the appropriate graph(s), whether the normality assumption holds for the variable(s). Mention the graph(s), the variable(s) you used to make the graph(s), and your conclusion(s) with respect to normality. Write your answers on the answer form.

d) Check the mean(s), standard deviation(s), and standard error(s), i.e., make summaries of the variable(s) in R Commander and write down the values on the answer form.

e) Perform Levene's test if appropriate and fill in the row for Levene's test on the answer form. For Levene's test go to: **Statistics > Variances > Levene's test**. . ., select **Center**: "`mean`" by changing the radio button. Click the **OK** button to perform the test.

f) Continue with the test procedure to answer the research question about tasting coffee: perform the appropriate analysis in R Commander. Use $\alpha = 0.05$ and indicate the create a 95% Confidence Interval.

g) On the answer form, fill in the part of the large table in the first column for Part 2 up to 'Confidence Interval'.

h) Decide, whether the 95% Confidence Interval can be read from the output generated at Part 2f), or that new output is required. In case of the latter, please do so. Write down the correct 95% Confidence Interval for the parameter(s) of interest on the answer form.

i) Could an error of Type I or II have been made here? Explain your answer.

j) Instead of using the p-value, also the rejection region (R.R.) could have been used. Use R Commander to find the correct rejection region. Go to: **Distributions**, choose quantiles for the correct distribution.

Computer Practicum 5

This computer practicum contains the following four parts:

- ☐ Exact Binomial testing.
- ☐ A simulation of correlation.
- ☐ Correlation applied to Tasting Coffee.
- ☐ Simple Linear Regression: Training Forest.

Learning objectives

After this computer practicum the student should be able to perform the following in R Commander, explain why the chosen test is the appropriate one given the research question and data, and interpret the R Commander output:

- Hypothesis test for π: Exact Binomial test,
- Correlation coefficient,
- Simple Linear Regression: provide estimates for β_0, β_1 and σ_ε,
- paired samples t-test.

Part 1 - Exact Binomial Testing: Binge Drinking

This part is based on:

- ☐ **Example 10.5 O&L 7th Edition p.489**, or
- ☐ **Example 10.5 O&L 6th Edition p.506**

Questions:

a) Read the example introduction, but not the solution.

b) Formulate the research question and write it on the answer form.

c) Explain why the exact binomial test is the appropriate test to answer the research question.

d) Define the parameter π, i.e., what is defined as the proportion (probability) of success in the given situation.

e) Denote the first five steps of the exact binomial test to answer the research question.

f) Load the data set into R Commander using the Excel file "`BSP5_Binge_Drinking.xlsx`" via: **Data > Import data > from Exel file**... **File** Provide a name for the data set behind **Enter name of data set:**, e.g., "`binge_drinking`". Remove the check mark in front of **Variable names in first row of spreadsheet**, and click the **OK** button to confirm. Select the Excel file, as mentioned above, and click **Open** to load the data set.

g) Before applying the binomial test the variable, currently a discrete quantitative variable containing only 0 and 1 values, needs to be converted into a nominal qualitative variable with two levels, a so-called factor variable in R/R Commander. Go to: **Data > Manage variables in active data set > Convert numerical variables into factors**.... Provide a new variable name in the field be **New variable name or prefix for**

multiple variables:, e.g., "`student`", and click **OK** to confirm. Supply level names for the numerical values, e.g., for 0 level name: "`non-binge`", and for 1 level name: "`binge`".

h) Click the **View data set** button. The data set will now contain two variables, one numerical column with the obscure name "`...1`" and one named "`student`" displaying the drinking behavior as "`non-binge`" or "`binge`" for each student in the set. Verify that there are 1300 non-binge drinking students and 1200 binge drinking students in the data set by: **Statistics > Summaries > Frequency distributions**..., and click the **OK** button to execute. The results are shown as counts and percentages in the Output part of the R Commander window.

i) Analyses in R/R Commander are performed based on the numerical values underlying qualitative variables. Here 0 is the lowest numerical value representing "`non-binge`" drinking students. Therefore, all analysis are default done on the group "`non-binge`" drinking students in this particular case. Reformulate the first five steps of the exact binomial testing procedure to meet the calculations performed in R/R Commander.

j) Perform the exact binomial test in R Commander using $\alpha = 0.05$. Go to: **Statistics > Proportions > Single-sample proportion test**.... Click on the **Options** tab, select the required **Alternative Hypothesis** as well the correct **Type of Test** using the radio button selection, and fill in the **Null hypothesis: p =** matching the steps written down for question i).

k) Provide the last three steps of the test procedure for this exact binomial test.

Part 2 - A Simulation of Correlation

This part of the computer practicum will give you a sense about the shape shown in a scatter plot and the strength of the straight line relationship, a.k.a. Pearson's correlation.

Start with opening the file "`BSP5_Simulation_Correlation_Scatterplots.xlsx`" in **Microsoft Excel**.

In the blue cell, "`D4`" in the spreadsheet, the value for the population correlation ρ can be changed. Initially it will be 0.50 (do not change it yet!). By pressing the **F9** function key a new sample, out of the population for variables x and y, is drawn. The estimated Pearson's correlation coefficient r, as estimator for ρ based on the sample, will be given in the green cell ("`D5`" in the spreadsheet). The data for x and y are given in the columns "`A`","`B`" and are visualized in the scatter plot.

a) Press **F9** five times sequentially, observe the shape of the scatter plot and the estimate of the correlation coefficient r.

b) What is the shape of the scatter plots? Are the estimates close to the population value of ρ? Explain why or why not.

c) Change the value for the population correlation ρ such that there is a strong negative correlation in the population. Write the value you used here for ρ on the answer form.

d) Now again press **F9** five times sequentially, observe the shape of the scatter plot and the estimate of the correlation coefficient r.

e) What is the shape of the scatter plots for the strong negative correlation you have chosen? Are the estimates close to the chosen population value of ρ? Explain why or why not.

f) Next change the population correlation ρ such that there is a very weak correlation in the population (choose whether the correlation is positive or negative yourself). Write down the chosen value for ρ on the answer form.

g) Describe the expected shape of the scatter plot, when the correlation is weak. Explain the expected shape with arguments.

h) Press **F9** once (i.e., draw one new sample) and check the shape of the scatter plot. Does it match your expectation?

i) Next press **F9** five times and write down the five estimated correlation coefficients r.

j) Are the estimates close to the population value of ρ? Explain why or why not.

Part 3 - Correlation : Tasting Coffee

For this part of the computer practicum the data from Part 1 of Computer Practicum 4 are used about tasting coffee at home in a laboratory environment by consumers. The case considered paired data, where the observations are linked per consumer, i.e., the judgement score given for the taste at home and the judgement score given in the laboratory environment are dependent (coming from the same consumer). When this trend is linear, Pearson's correlation coefficient can be used to determine the strength of this linear relationship.

a) Load the data "`BSP5_Tasting_Coffee.RData`" into R Commander and view the data.

b) Make the appropriate plot to check whether there is a linear relation between the judgement scores for tasting in the laboratory environment and at home.

c) Calculate Pearson's correlation coefficient using **Statistics > Summaries > Correlation matrix**... Choose the two variables for which you want Pearson's correlation coefficient, select both by using **Shift** on your keyboard and click the desired variables (**CTRL** + clicking allows individual selection of multiple). Click the **OK** button to execute the calculation of the correlation matrix.

d) What is the estimated Pearson's correlation coefficient? Write the value on your answer form using 4 decimals.

e) Interpret the strength of the linear relationship between the judgement scores at home and in the laboratory environment (see form).

Part 4 - Simple Linear Regression : Training Forest

Wageningen University has several training forests, where students can carry out measurements during field work. In this computer practicum, data will be used that were collected by students in such a forest in 2015. Whether there is a linear relationship between y, the height of the trees (in meters), and x, the diameter at breast height (in centimeters at 1.30 meters height) will be studied. Assume that the y values are independent and normally distributed with constant variance σ_ε^2.

a) Write down the above-mentioned assumptions in mathematical terms. This will provide you a concise statistical model for the observed y values. Indicate for each element, whether it belongs to the systematic part or the stochastic (random) part of the model.

b) Load the file "`BSP5_Training_Forest.RData`" into R Commander and view the data. How many cases are there?

c) Type in the "R Script" part of the R Commander window: `attr(training_forest, "variable.labels")`, keep the cursor on the same line and click **Submit** to get information on the meaning of the variables in the columns of the data set. Fill in the answer form.

d) Generate the appropriate plot to display the relationship between y and x: **Graphs > Scatter plot**..., choose the correct variable for the y- and x-axis. On the **Options** tab place a check mark in the box in front of **Least-squares line**. Is there a (roughly) linear relationship between the height of trees (in meters) and the diameter of trees at breast height (in cm at 1.30 m) in 2015?

e) Generate simple linear regression output for the relationship in d): **Statistics > Fit models > linear regression**... Choose the correct **Response variable (pick one)**, representing y, and **Explanatory variables (pick one or more)**, representing x. Click the **OK** button to generate the model output.

f) Give the equation of the least squares regression line (often referred to as the estimated regression line).

g) Give the interpretation of the estimated coefficients of the least squares regression line, using the height of trees and diameter of trees at breast height in 2015 in your description.

h) In order to answer the research question, whether the diameter of a tree at breast height has predictive value for the height of a tree a test should be applied. Start with mentioning the first five steps of the appropriate test ($\alpha = 0.05$).

i) Find the needed values in the R/R Commander output produced at question e) to proceed with the test procedure and fill in the answer form.

j) Give the estimate for σ_ε^2.

Computer Practicum 6

This computer practicum contains the following three parts:

- ☐ Simple Linear Regression: Effect of Fertilizer on Lettuce Plants.
- ☐ Simple Linear Regression: Species per Island.
- ☐ Simple Linear Regression: Paying for Bread.

Learning objectives

After this computer practicum the student should be able to do the following in R Commander:

- Apply a simple linear regression model.
- Perform a t-test for the regression coefficents β_0 and β_1.
- Give Confidence Intervals for the regression coefficents β_0 and β_1.
- Give the estimated population mean and predicted value for a single random unit given a value for x.
- Give the standard error for the population mean given a value for x.
- Give a Confidence Interval for the population mean $\mu_x = \beta_0 + \beta_1 \times x_{n+1}$.
- Give a Prediction Interval for a future response y at x_{n+1}.
- Produce plots to check the assumptions for a simple linear regression model.

! R Commander Plugin for Statistical Analysis and Data Display: Heiberger and Holland

Before starting with Part 1 a plug-in for R Commander needs to be loaded. This plug-in is needed to make confidence and prediction intervals for any value of x in Simple Linear Regression.

To load the plug-in use: **Tools > Load Rcmdr plug-in(s)**... Select the "`RcmdrPlugin.HH`" and click the **OK** button to load the plug-in. A message will appear: **The plug-in(s) will not be available until the Commander is restarted. Restart now?**. Click the **Yes** button to proceed.

When the option for "`RcmdrPlugin.HH`" is not available in the list of plug-ins:

- Close R Commander.
- In the R Console go to: **Packages > Install package(s)**...
- Select "0-cloud" as CRAN Mirror.
- In the list of packages select "`RcmdrPlugin.HH`" and click the **OK** button to install.
- Close the R Console.
- Restart the R GUI (**R** 4.3.1 in the **R** folder inside your start menu).
- Restart R Commander.
- Load the plug-in as described above.

Part 1 - Simple Linear Regression: Effect of Fertilizer on Lettuce Plants

In an experiment, 10 lettuce plants are grown in soil to which different amounts of fertilizer are added. The relative amounts of fertilizer used are 0, 1.0, 2.0, and 3.0. The weights (g) of the lettuce plants are measured after 1 week.

The measured weight for the 10 plants is assumed to be an outcome of normally distributed variable, the mean depends linearly on the relative amount of added fertilizer. The following general research questions are addressed in this part:

1. Is there a linear relationship between mean weight and the amount of fertilizer added?
2. Assuming that a possible relationship will be linear, is there a relationship?
3. How strong is this relationship?
4. How uncertain is the estimate of the mean weight of a lettuce plant, given the amount of added fertilizer?
5. How uncertain is the estimate of the weight of an individual lettuce plant, given the amount of added fertilizer?

The data are in the file "`BSP6_Lettuce_Plants.Rdata`".

a) Load the data and inspect the data by viewing.

b) Display the data in a useful graph. Does the relationship appear linear?

c) Should fertilizer be considered a qualitative or a quantitative variable and why?

d) Assume that a straight line linear relationship applies and fit the model using R Commander. Give estimates for the regression parameters and (when applicable) the corresponding estimated standard errors. [Hint: the three regression parameters are... (see also the handouts of Tutorial 10)].

e) Determine a 95% Confidence Interval for β_1: **Models > Confidence intervals**...

f) Test (with $\alpha = 0.05$) whether adding fertilizer has a positive effect on the weight of lettuce plants. Begin with the first five steps on the answer form; next have a look at the output generated for question d) to fill in the final steps of the appropriate test.

g) Get the estimated mean weight, the associated standard error and a 95% Confidence Interval for the mean of a randomly selected lettuce plant grown on soil with a relative amount of added fertilizer equal to 2.5: **Models > Prediction Intervals**...**(HH)**. Fill **Enter X values** with the desired value for fertilizer (note that the **confidence interval for mean** is the default setting). Place a check mark in the box in front of **Standard Error** and click the **OK** button to execute.
Write the answers on the answer form.

h) Give the estimated weight of a randomly selected lettuce plant grown in soil with a relative amount of added fertilizer equal to 2.5. Also give a 95% Prediction Interval (check the appropriate option in R Commander). Compare your answer with the answer given under g).

i) The Confidence Intervals for the mean and the Prediction Intervals for individual data-points can be plotted in one graph: **Models > Confidence interval Plot**... Why is the confidence interval narrower near $\bar{x}$?

j) To check the quality of the regression model create some plots: **Models > Graphs > Basic diagnostic plots**. In this course only the top two plots have been discussed, the other two plots can be ignored.

k) Look at the Normal Q-Q plot. What assumption does it test? Explain, whether the assumption is met or not.

l) Check the Residuals vs Fitted plot. What assumptions does it test? Explain, whether the assumption are met or not.

Part 2 - Simple Linear Regression: Species per Island

In this second part data from the article "Plant species richness – The effect of the Island Size and Habitat Diversity" (Kohn and Walsh, Journal of Ecology, 1994) will be used. The article describes a study about

the plant richness of the Shetland islands. Three variables are measured for each of 47 of the Shetland Islands: the area of the island ("Area", in hectares), the number of dicotyledon plant species per island ("Species_per_island"), and the number of different habitat types per island ("Number_of_habitat_types"). This part will investigate whether "Species_per_island" can be predicted by the area of the island ("Area") using simple linear regression, or by the variable "Number_of_habitat_types" also by using simple linear regression.

The data set is available in the file named "BSP6_Species_Island.RData".

a) Load and view the data.

b) Generate two useful graphical displays to visualize the relations between:

- "Species_per_island" and "Area"
- " 'Species_per_island' " and "Number_of_habitat_types"

c) Are both relations linear?

d) To overcome the problem of the nonlinear relation between "Species_per_island" and "Area', the variable"Area" will be transformed with the natural logarithmic function: **Data > Manage variables in active data set > Compute new variable**... Use as **Expression to compute**: "log(Area)", and use, e.g., for **New variable name**: "log_area".

e) Make a new visualisation to see the relation between "Species_per_island" and "log_area". Can simple linear regression be applied now?

f) Carry out a simple linear regression analysis to predict the number of species per island with log(Area) as explanatory variable. Give the estimated regression line. Do not forget to denote what the variables in your equation mean.

g) Give the coefficient of determination this simple linear regression model.

h) Generate basic diagnostic plots.Are the assumptions for the simple linear regression model of question f) met?

i) Carry out a simple linear regression analysis to predict the number of species per island with "Number_of_habitat_types" as explanatory variable. Give the estimated regression line. Do not forget to denote what the variables in your equation mean.

j) Give the coefficient of determination for this simple linear regression model.

k) Are the assumptions for the simple linear regression model of question i) met?

l) Which model (f or i) would you prefer to predict the number of species per island? Why?

m) Estimate the population mean number of species of a Shetland island with 10 different types of habitats. Give also the standard error of this estimate: **Models > Prediction intervals**...**(HH)**. Set the radio button to **point estimate only** and place a check mark in front of **Standard Error**, next click the **OK** button to execute.

Part 3 - Simple Linear Regression: Paying for Bread

Load and view "BSP6_Pay_for_Bread.RData" in R Commander. This data set contains two variables of a survey. Consumers were asked to give their age ("Age" in years) and the maximum price, they are willing to pay, for a whole loaf of bread, that is more healthy than regular bread ("Willing_to_Pay" in euros). Research question is: "Is there a negative relationship between maximum price, consumers are willing to pay, for a more healthy loaf of bread and age?".

a) Give the estimated equation of the least squares regression line, with description of the variables, as well as the coefficient of determination for this problem. Does the coefficient of determination indicate a strong relationship between "Willing_to_Pay" and "Age"? Why (not)?

b) Give the null- and alternative hypothesis, p-value and conclusion corresponding to the formulated research question (use $\alpha = 0.05$).

c) Does the conclusion of question b) indicate that there is a strong relationship between "`Willing_to_Pay`" and "`Age`"? Why (not)?

d) Generate a plot to visualize this problem.

e) Reflect on the practical relevance of this significant relationship. Use in your answer the words **relevant**, **significant**, and **sample size**.

Appendix A

Binomial Distribution Tables

Table A.1: $P(y \leq k)$ for $y \sim \text{Bin}(n = 1, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
1	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.2: $P(y \leq k)$ for $y \sim \text{Bin}(n = 2, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
2	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
2	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.3: $P(y \leq k)$ for $y \sim \text{Bin}(n = 3, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
3	1	0.9928	0.9720	0.9392	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000
3	2	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
3	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.4: $P(y \leq k)$ for $y \sim \text{Bin}(n = 4, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
4	1	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
4	2	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
4	3	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
4	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.5: $P(y \leq k)$ for $y \sim \text{Bin}(n = 5, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
5	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
5	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
5	3	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
5	4	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
5	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.6: $P(y \leq k)$ for $y \sim \text{Bin}(n = 6, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
6	1	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
6	2	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3437
6	3	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6562
6	4	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
6	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
6	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.7: $P(y \leq k)$ for $y \sim \text{Bin}(n = 7, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
7	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
7	2	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
7	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
7	4	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
7	5	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
7	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
7	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.8: $P(y \leq k)$ for $y \sim \text{Bin}(n = 8, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
8	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
8	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
8	3	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
8	4	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
8	5	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
8	6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
8	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
8	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.9: $P(y \leq k)$ for $y \sim \text{Bin}(n = 9, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
9	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
9	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
9	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
9	4	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
9	5	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
9	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
9	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
9	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
9	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.10: $P(y \leq k)$ for $y \sim \text{Bin}(n = 10, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
10	1	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
10	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
10	3	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
10	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
10	5	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
10	6	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
10	7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
10	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
10	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
10	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.11: $P(y \leq k)$ for $y \sim \text{Bin}(n = 11, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
11	0	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005
11	1	0.8981	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059
11	2	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
11	3	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
11	4	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
11	5	1.0000	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000
11	6	1.0000	1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
11	7	1.0000	1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
11	8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
11	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941
11	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995
11	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.12: $P(y \leq k)$ for $y \sim \text{Bin}(n = 12, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
12	0	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
12	1	0.8816	0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032
12	2	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
12	3	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730
12	4	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
12	5	1.0000	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
12	6	1.0000	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
12	7	1.0000	1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
12	8	1.0000	1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270
12	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807
12	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968
12	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
12	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.13: $P(y \leq k)$ for $y \sim \text{Bin}(n = 13, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
13	0	0.5133	0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
13	1	0.8646	0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0049	0.0017
13	2	0.9755	0.8661	0.6920	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112
13	3	0.9969	0.9658	0.8820	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461
13	4	0.9997	0.9935	0.9658	0.9009	0.7940	0.6543	0.5005	0.3530	0.2279	0.1334
13	5	1.0000	0.9991	0.9925	0.9700	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905
13	6	1.0000	0.9999	0.9987	0.9930	0.9757	0.9376	0.8705	0.7712	0.6437	0.5000
13	7	1.0000	1.0000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095
13	8	1.0000	1.0000	1.0000	0.9998	0.9990	0.9960	0.9874	0.9679	0.9302	0.8666
13	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539
13	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9959	0.9888
13	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983
13	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
13	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.14: $P(y \leq k)$ for $y \sim \text{Bin}(n = 14, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
14	0	0.4877	0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0001
14	1	0.8470	0.5846	0.3567	0.1979	0.1010	0.0475	0.0205	0.0081	0.0029	0.0009
14	2	0.9699	0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.0170	0.0065
14	3	0.9958	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287
14	4	0.9996	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898
14	5	1.0000	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.2120
14	6	1.0000	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953
14	7	1.0000	1.0000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047
14	8	1.0000	1.0000	1.0000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.7880
14	9	1.0000	1.0000	1.0000	1.0000	0.9997	0.9983	0.9940	0.9825	0.9574	0.9102
14	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9961	0.9886	0.9713
14	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9978	0.9935
14	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991
14	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
14	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.15: $P(y \leq k)$ for $y \sim \text{Bin}(n = 15, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	0	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
15	1	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
15	2	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
15	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
15	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
15	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
15	6	1.0000	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
15	7	1.0000	1.0000	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000
15	8	1.0000	1.0000	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
15	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
15	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9937	0.9824
15	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963
15	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
15	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.16: $P(y \leq k)$ for $y \sim \text{Bin}(n = 16, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
16	0	0.4401	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0001	0.0000
16	1	0.8108	0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003
16	2	0.9571	0.7892	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021
16	3	0.9930	0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106
16	4	0.9991	0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384
16	5	0.9999	0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051
16	6	1.0000	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2272
16	7	1.0000	0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018
16	8	1.0000	1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982
16	9	1.0000	1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728
16	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949
16	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616
16	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894
16	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979
16	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997
16	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.17: $P(y \leq k)$ for $y \sim \text{Bin}(n = 17, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
17	0	0.4181	0.1668	0.0631	0.0225	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000
17	1	0.7922	0.4818	0.2525	0.1182	0.0501	0.0193	0.0067	0.0021	0.0006	0.0001
17	2	0.9497	0.7618	0.5198	0.3096	0.1637	0.0774	0.0327	0.0123	0.0041	0.0012
17	3	0.9912	0.9174	0.7556	0.5489	0.3530	0.2019	0.1028	0.0464	0.0184	0.0064
17	4	0.9988	0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.1260	0.0596	0.0245
17	5	0.9999	0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.0717
17	6	1.0000	0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662
17	7	1.0000	0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145
17	8	1.0000	1.0000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5000
17	9	1.0000	1.0000	1.0000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855
17	10	1.0000	1.0000	1.0000	0.9999	0.9994	0.9968	0.9880	0.9652	0.9174	0.8338
17	11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9970	0.9894	0.9699	0.9283
17	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9975	0.9914	0.9755
17	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9936
17	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9988
17	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
17	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.18: $P(y \leq k)$ for $y \sim \text{Bin}(n = 18, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
18	0	0.3972	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
18	1	0.7735	0.4503	0.2241	0.0991	0.0395	0.0142	0.0046	0.0013	0.0003	0.0001
18	2	0.9419	0.7338	0.4797	0.2713	0.1353	0.0600	0.0236	0.0082	0.0025	0.0007
18	3	0.9891	0.9018	0.7202	0.5010	0.3057	0.1646	0.0783	0.0328	0.0120	0.0038
18	4	0.9985	0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.0942	0.0411	0.0154
18	5	0.9998	0.9936	0.9581	0.8671	0.7175	0.5344	0.3550	0.2088	0.1077	0.0481
18	6	1.0000	0.9988	0.9882	0.9487	0.8610	0.7217	0.5491	0.3743	0.2258	0.1189
18	7	1.0000	0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403
18	8	1.0000	1.0000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073
18	9	1.0000	1.0000	0.9999	0.9991	0.9946	0.9790	0.9403	0.8653	0.7473	0.5927
18	10	1.0000	1.0000	1.0000	0.9998	0.9988	0.9939	0.9788	0.9424	0.8720	0.7597
18	11	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811
18	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9986	0.9942	0.9817	0.9519
18	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9846
18	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962
18	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993
18	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
18	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.19: $P(y \leq k)$ for $y \sim \text{Bin}(n = 19, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
19	0	0.3774	0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0001	0.0000	0.0000
19	1	0.7547	0.4203	0.1985	0.0829	0.0310	0.0104	0.0031	0.0008	0.0002	0.0000
19	2	0.9335	0.7054	0.4413	0.2369	0.1113	0.0462	0.0170	0.0055	0.0015	0.0004
19	3	0.9868	0.8850	0.6841	0.4551	0.2631	0.1332	0.0591	0.0230	0.0077	0.0022
19	4	0.9980	0.9648	0.8556	0.6733	0.4654	0.2822	0.1500	0.0696	0.0280	0.0096
19	5	0.9998	0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.0777	0.0318
19	6	1.0000	0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.0835
19	7	1.0000	0.9997	0.9959	0.9767	0.9225	0.8180	0.6656	0.4878	0.3169	0.1796
19	8	1.0000	1.0000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.4940	0.3238
19	9	1.0000	1.0000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.6710	0.5000
19	10	1.0000	1.0000	1.0000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762
19	11	1.0000	1.0000	1.0000	1.0000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204
19	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165
19	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9969	0.9891	0.9682
19	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9904
19	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9978
19	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
19	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
19	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
19	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.20: $P(y \leq k)$ for $y \sim \text{Bin}(n = 20, \pi)$. Please note that this table provides cumulative probabilities.

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
20	0	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
20	1	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000
20	2	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002
20	3	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013

n	k	$\pi = 0.05$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
20	4	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059
20	5	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207
20	6	1.0000	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577
20	7	1.0000	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316
20	8	1.0000	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517
20	9	1.0000	1.0000	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119
20	10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881
20	11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483
20	12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684
20	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9935	0.9786	0.9423
20	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9936	0.9793
20	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941
20	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987
20	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
20	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Appendix B

Standard Normal Distribution Table

Table B.1: $P(Z \leq z)$, where $Z \sim \mathrm{N}(\mu = 0, \sigma = 1)$ for $z \in \{-3.49, \ldots, 0\}$ as shown in Figure B.1a. Please note that this table provides cumulative probabilities.

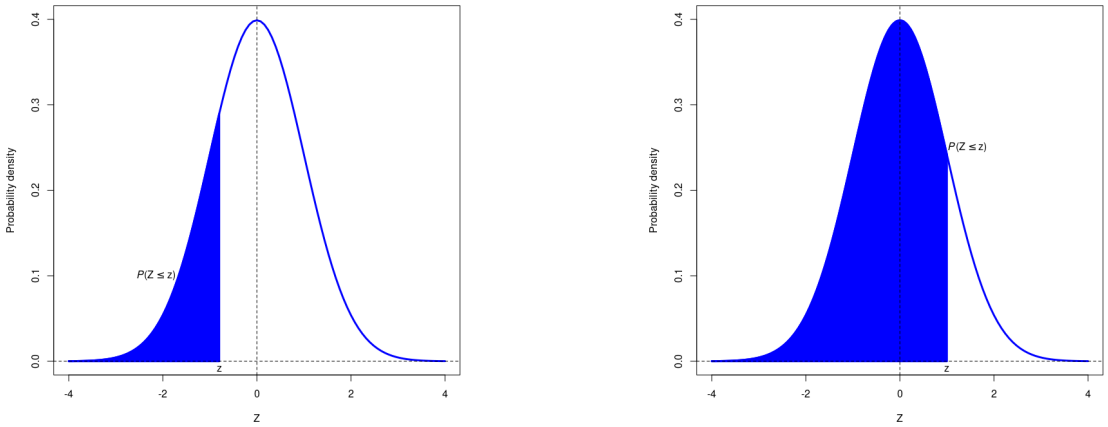
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table B.2: $P(Z \leq z)$, where $Z \sim \mathrm{N}(\mu = 0, \sigma = 1)$ for $z \in \{0, \ldots, 3.49\}$ as shown in Figure B.1b. Please note that this table provides cumulative probabilities.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table B.3: $P(Z \leq z)$, where $Z \sim \mathrm{N}(\mu = 0, \sigma = 1)$ for discrete values of z. Please note that this table provides cumulative probabilities.

z	$P(Z \leq z)$	z	$P(Z \leq z)$
$-\infty$	0.00000000	3.50	0.99976737
-5.00	0.00000029	4.00	0.99996833
-4.50	0.00000340	4.50	0.99999660
-4.00	0.00003167	5.00	0.99999971
-3.50	0.00023263	∞	1.00000000



(a) for negative z-scores

(b) for positive z-scores

Figure B.1: Shaded Area $P(Z \leq z)$ under the standard normal distribution

Appendix C

Inverse Student's t Distributions Table

Table C.1: Critical values for Student's t distribution: $P(t \geq t_{\alpha,\nu}) = \alpha$, as shown in Figure C.1.

ν or df	$\alpha = 0.40$	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.682	1.309	1.696	2.040	2.453	2.744	3.375	3.633
32	0.255	0.682	1.309	1.694	2.037	2.449	2.738	3.365	3.622
33	0.255	0.682	1.308	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.682	1.307	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
36	0.255	0.681	1.306	1.688	2.028	2.434	2.719	3.333	3.582
37	0.255	0.681	1.305	1.687	2.026	2.431	2.715	3.326	3.574
C.L.[1]	20%	50%	80%	90%	95%	98%	99%	99.8%	99.9%

ν or df	$\alpha = 0.40$	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
38	0.255	0.681	1.304	1.686	2.024	2.429	2.712	3.319	3.566
39	0.255	0.681	1.304	1.685	2.023	2.426	2.708	3.313	3.558
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
41	0.255	0.681	1.303	1.683	2.020	2.421	2.701	3.301	3.544
42	0.255	0.680	1.302	1.682	2.018	2.418	2.698	3.296	3.538
43	0.255	0.680	1.302	1.681	2.017	2.416	2.695	3.291	3.532
44	0.255	0.680	1.301	1.680	2.015	2.414	2.692	3.286	3.526
45	0.255	0.680	1.301	1.679	2.014	2.412	2.690	3.281	3.520
46	0.255	0.680	1.300	1.679	2.013	2.410	2.687	3.277	3.515
47	0.255	0.680	1.300	1.678	2.012	2.408	2.685	3.273	3.510
48	0.255	0.680	1.299	1.677	2.011	2.407	2.682	3.269	3.505
49	0.255	0.680	1.299	1.677	2.010	2.405	2.680	3.265	3.500
50	0.255	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.254	0.678	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	0.254	0.677	1.291	1.662	1.987	2.368	2.632	3.183	3.402
100	0.254	0.677	1.290	1.660	1.984	2.364	2.626	3.174	3.390
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
140	0.254	0.676	1.288	1.656	1.977	2.353	2.611	3.149	3.361
160	0.254	0.676	1.287	1.654	1.975	2.350	2.607	3.142	3.352
180	0.254	0.676	1.286	1.653	1.973	2.347	2.603	3.136	3.345
200	0.254	0.676	1.286	1.653	1.972	2.345	2.601	3.131	3.340
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291
C.L.[1]	20%	50%	80%	90%	95%	98%	99%	99.8%	99.9%

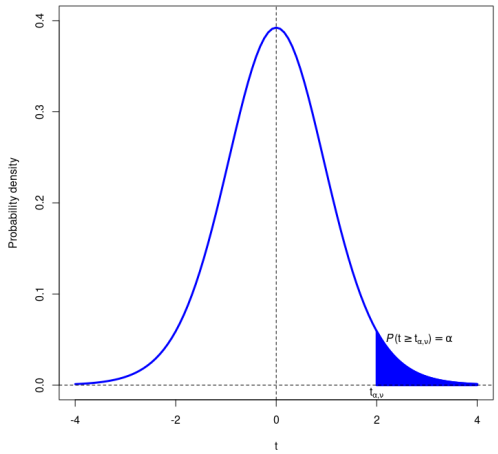


Figure C.1: Shaded right-tail probability in a Student's t distribution: $P(t \geq t_{\alpha,\nu}) = \alpha$

! Two-tailed test and Confidence Intervals

For a two-tailed test with significance level α and $(1-\alpha) \times 100\%$ Confidence Intervals use the values in the column headed by the number obtained by computing $\alpha/2$.

$$t_{\alpha/2,\nu=\infty} = z_{\alpha/2}$$

[1] Confidence Level

Appendix D

Manual Graphing Calculators

D.1 Graphing Calculator Texas Instruments TI-83/TI-83 Plus/TI-84 Plus

D.1.1 Measures of central tendency and variability

Choose: `STAT` > `EDIT`

- Enter the outcomes in list `L1`, or
- Enter the different outcomes in list `L1` and the corresponding frequencies in list `L2`
- Leave `L1` (or `L1`, and `L2`) via `2nd MODE (= QUIT)`

Choose: `STAT` > `CALC` > `1-VAR STATS` > `List: L1`, or `STAT` > `CALC` > `1-VAR STATS` > `List: L1`, and `FreqList: L2`

You will get, among other values:

- $\bar{x}$ = sample mean
- s_x = sample standard deviation (based on division by $n-1$)
- σ_x = population standard deviation (based on division by n)
- `minX` = minimum
- `maxX` = maximum
- $\mathtt{Q}_1$ = first Quartile
- `Med` = median (or second Quartile)
- $\mathtt{Q}_3$ = third Quartile

The variance is equal to the standard deviation squared; Inter Quartile Range (IQR) = $\mathtt{Q}_3 - \mathtt{Q}_1$.

D.1.2 Binomial Distribution

Choose: `2nd VARS (= DISTR)`

Let $y \sim \text{Bin}(n,\ \pi)$ or $y \sim \text{B}(n,\ \pi)$.

- $P(y = k) = \binom{n}{k} \times \pi^k \times (1-\pi)^{n-k} = \frac{n!}{k! \times (n-k)!} \times \pi^k \times (1-\pi)^{n-k}$: `binompdf(`$n$`,`$\pi$`,`$k$`)`
- $P(y \leq k)$: `binomcdf(`n`,`π`,`k`)`

Where n denotes the sample size, π denotes the population proportion of success, and k denotes the number of successes counted in the sample of size n.

D.1.3 Normal Distribution

Choose: `2nd VARS (= DISTR)`

Let $y \sim \text{N}(\mu,\ \sigma)$.

- $P($"lowerbound" $\leq y \leq$ "upperbound"$)$: `normalcdf(lower, upper,`μ`,`σ`)`
- $P(y \geq$ "lowerbound"$)$: `normalcdf(lower, 1E99,`μ`,`σ`)`

- $P(y \leq$ "upperbound"): `normalcdf(-1E99, upper,`μ`,`σ`)`

> **!** Remarks about calculating probabilities using a Normal Distributtion.
>
> - `-1E99`, being -1×10^{99}, denotes negative infinity ($-\infty$), and `1E99` (1×10^{99}) denotes positive infinity (∞)
> - Use $\mu = 0$ and $\sigma = 1$ for the Standard Normal Distribution.

D.1.4 Inverse Normal Distribution

Choose: `2nd VARS (= DISTR)`

Let $y \sim \mathrm{N}(\mu,\ \sigma)$.

- `invNorm(area,`μ`,`σ`)` calculates the value q of y for which: $P(y \leq q) =$ `area`

Therefore, `area` denotes **a left-tailed probability under the normal distribution**.

> **!** Remarks about calculating a quantile using the inverse Normal Distributtion.
>
> - Use $\mu = 0$ and $\sigma = 1$ for the Standard Normal Distribution.

D.1.5 Student *t*-distribution

Choose: `2nd VARS (= DISTR)`

Let $t \sim$ Student t-distribution with $\nu =$ `df` degrees of freedom.

- $P($"lowerbound" $\leq t \leq$ "upperbound"): `tcdf(lower, upper, df)`
- $P(t \geq$ "lowerbound"): `tcdf(lower, 1E99, df)`
- $P(t \leq$ "upperbound"): `tcdf(-1E99, upper, df)`

> **!** Remarks about calculating probabilities using a Student t-distributtion.
>
> - `-1E99`, being -1×10^{99}, denotes negative infinity ($-\infty$), and `1E99` (1×10^{99}) denotes positive infinity (∞)

D.1.6 Inverse Student *t*-distribution

Choose: `2nd VARS (= DISTR)`

Let $t \sim$ Student t-distribution with $\nu =$ `df` degrees of freedom.

- `invT(area, df)` calculates the value q of t for which: $P(t \leq q) =$ `area` under a Student t-distribution with $\nu =$ `df` degrees of freedom.

Therefore, `area` denotes **a left-tailed probability under the Student *t*-distribution**.

The `invT()` function is by default present in the TI-84 Plus graphing calculator. However, not by default in the TI-83/TI-83 Plus. Fortunately, this function can be easily added by programming it into the TI-83/TI-83 Plus as explained in the next section.

D.1.6.1 Programming the INVT() function into the TI-83/TI-83 Plus

Steps to program the `INVT()` function into the TI-83/TI-83 Plus graphing calculator, are shown in https://youtu.be/5Ft5eZVJtPk.

To execute the program select:

1. `PRGM`, `EXEC`, got to `INVT`, and press `ENTER` to select.
2. Press `ENTER` again and fill behind `AREA LEFT:` the value for the left-tailed probability.

3. Press ENTER and fill behind DF: the value for the degrees of freedom $\nu =$ df.
4. After pressing ENTER the value for INVT(area, df) will be calculated as explained above.

D.2 Graphing Calculator Casio CFX-9850/fx-9750GII/fx-9860G Series

D.2.1 Measures of central tendency and variability

Choose: MENU > STAT or press 2 on the keypad, if necessary, empty existing lists via **F6** > **F4** DEL-A.

- Type the different outcomes in List 1 and the corresponding frequencies in List 2.
- Choose **F2** CALC, and then **F6** SET for setting the lists to use when needed:
  - 1 Var Xlist : List1
  - 1 Var Freq: List2 and press EXE
- Select **F1** 1VAR.
- The following values, among others, are displayed:
  - $\bar{x} =$ sample mean
  - $s_x =$ sample standard deviation (based on division by $n - 1$)
  - $\sigma_x =$ population standard deviation (based on division by n)
  - minX = minimum
  - Q1 = first Quartile
  - Med = median
  - Q3 = third quartile
  - maxX = maximum

The variance is equal to the standard deviation squared; Inter Quartile Range (IQR) = Q3 ˇ Q1.

D.2.2 Binomial Distribution

Choose: MENU > STAT or press 2 on the keypad > **F5** DIST

Let $y \sim \text{Bin}(n,\ \pi)$ or $y \sim \text{B}(n,\ \pi)$.

- $P(y = k) = \binom{n}{k} \times \pi^k \times (1 - \pi)^{n-k} = \frac{n!}{k! \times (n-k)!} \times \pi^k \times (1 - \pi)^{n-k}$:
  - Press **F5** BINM and then **F1** Bpd.
  - Press **F2** Var to switch form list (**F1** List) to variable mode, when needed.
  - Behind x, enter the value for the number of successes k and press EXE.
  - Behind Numtrial, enter the value the number of trials n and press EXE.
  - Behind p, enter the value for the probability of success π and press EXE.
  - Navigate to Execute and select **F1** CALC, or press EXE to calculate the answer.
- $P(y \leq k)$:
  - Press **F5** BINM and then **F2** Bcd.
  - Press **F2** Var to switch form list (**F1** List) to variable mode, when needed.
  - Behind x, enter the value for the number of successes k and press EXE.
  - Behind Numtrial, enter the value the number of trials n and press EXE.
  - Behind p, enter the value for the probability of success π and press EXE.
  - Navigate to Execute and select **F1** CALC, or press EXE to calculate the answer.

D.2.3 Normal Distribution

Choose: MENU > STAT or press 2 on the keypad > **F5** DIST

Let $y \sim \text{N}(\mu,\ \sigma)$.

- Calculation of $P(\text{"lowerbound"} \leq y \leq \text{"upperbound"})$, $P(y \geq \text{"lowerbound"})$, or $P(y \leq \text{"upperbound"})$:
  - Press **F1** NORM, and then **F2** Ncd.
  - Press **F2** Var to switch form list (**F1** List) to variable mode, when needed.

- Behind `Lower`, type the "lowerbound" of the interval and press `EXE`. If there is no lowerbound use `-1E99` (-1×10^{99}) for negative infinity ($-\infty$)
- Behind `Upper`, type the "upperbound" of the interval and press `EXE`. If there is no upperbound use `1E99` (1×10^{99}) for positive infinity (∞)
- Provide the values for σ, and μ.
- Navigate to `Execute` and select **F1** `CALC`, or press `EXE` to calculate the answer.

> **!** Remarks about calculating probabilities using a Normal Distributtion.
>
> - Use $\mu = 0$ and $\sigma = 1$ for the Standard Normal Distribution.

D.2.4 Inverse Normal Distribution

Choose: `MENU` > `STAT` or press `2` on the keypad > **F5** `DIST`

Let $y \sim \mathrm{N}(\mu,\ \sigma)$.

- Calculation of the value q of y for which: $P(y \leq q) =$ `Area`:
  - Press **F1** `NORM` and then **F3** `InvN`.
  - Press **F2** `Var` to switch form list (**F1** `List`) to variable mode, when needed.
  - For the fx-9750GII/fx-9680G Series, please pay special attention to the side (`Tail`) used:
    - When **F1** `LEFT` is used, the area from negative infinity ($-\infty$) to the upperbound q will be used, that is the **left-tailed probability under the normal distribution** $P(y \leq q) =$ `Area`.
    - When **F2** `RIGHT` is used, the area from the lowerbound q to positive infinity (∞) will be used, that is the **right-tailed probability under the normal distribution** $P(y \geq q) =$ `Area`.
    - When **F3** `CNTR` is used, the area from the lowerbound $-q$ to the upperbound q will be used, that is the **two-tailed probability under the normal distribution** $P(-q \leq y \leq q) =$ `Area`.
  - Behind `Area`, fill the value for the probability under the normal distribution. In case of using a **Casio CFX-9850 Series** enter the **left-tailed probability under the normal distribution**, being $P(y \leq q) =$ `Area`.
  - Provide the values for σ, and μ.
  - Navigate to `Execute` and select **F1** `CALC`, or press `EXE` to calculate the answer.

> **!** Remarks about calculating probabilities using a Normal Distributtion.
>
> - Use $\mu = 0$ and $\sigma = 1$ for the Standard Normal Distribution.

D.2.5 Student *t*-distribution

Choose: `MENU` > `STAT` or press `2` on the keypad > **F5** `DIST`

Let $t \sim$ Student t-distribution with $\nu =$ `df` degrees of freedom.

- Calculation of $P($"lowerbound" $\leq y \leq$ "upperbound"$)$, $P(y \geq$ "lowerbound"$)$, or $P(y \leq$ "upperbound"$)$:
  - Press **F2** `t`, and then **F2** `tcd`.
  - Press **F2** `Var` to switch form list (**F1** `List`) to variable mode, when needed.
  - Behind `Lower`, type the "lowerbound" of the interval and press `EXE`. If there is no lowerbound use `-1E99` (-1×10^{99}) for negative infinity ($-\infty$)
  - Behind `Upper`, type the "upperbound" of the interval. If there is no upperbound use `1E99` (1×10^{99}) for positive infinity (∞)
  - Provide the values for $\nu =$ `df` degrees of freedom.
  - Navigate to `Execute` and select **F1** `CALC`, or press `EXE` to calculate the answer.

D.2.6 Inverse Student *t*-distribution

Choose: `MENU` > `STAT` or press `2` on the keypad > **F5** `DIST`

Let $t \sim$ Student t-distribution with $\nu =$ `df` degrees of freedom.

- Calculate the value q of t for which: $P(t \leq q) =$ `Area`:
  - Press **F2** `t`, and then **F3** `Invt`.
  - Press **F2** `Var` to switch form list (**F1** `List`) to variable mode, when needed.
  - Behind `Area`, fill the value for the **right-tailed probability under the Student *t*-distribution**, that is $P(y \geq q) =$ `Area`.
  - Provide the value for the degrees of freedom $\nu =$ `df`.
  - Navigate to `Execute` and select **F1** `CALC`, or press `EXE` to calculate the answer.