Chapter 1

Introduction, descriptive statistics, R and data visualization (solutions to exercises)

Contents

1.1 Infant birth weight

|||| Exercise 1.1 Infant birth weight

In a study of different occupational groups the infant birth weight was recorded for randomly selected babies born by hairdressers, who had their first child. The following table shows the weight in grams (observations specified in sorted order) for 10 female births and 10 male births:

Females (x)	2474	2547	2830	3219	3429	3448	3677	3872	4001	4116
Males (y)	2844	2863	2963	3239	3379	3449	3582	3926	4151	4356

Solve at least the following questions a)-c) first "manually" and then by the inbuilt functions in R. It is OK to use R as alternative to your pocket calculator for the "manual" part, but avoid the inbuilt functions that will produce the results without forcing you to think about how to compute it during the manual part.

a) What is the sample mean, variance and standard deviation of the female births? Express in your own words the story told by these numbers. The idea is to force you to interpret what can be learned from these numbers.

|||| Solution

We have $n = 10$, hence the sample mean is

$$\begin{aligned}\bar{x} &= \frac{1}{10}(2474 + 2547 + 2830 + 3219 + 3429 + 3448 + 3677 + 3872 + 4001 + 4116)\\ &= 3361.3,\end{aligned}$$

and the sample variance

$$\begin{aligned}s^2 = \frac{1}{9}\big(&(2474 - 3361.3)^2 + (2547 - 3361.3)^2 + (2830 - 3361.3)^2 + (3219 - 3361.3)^2\\ &+ (3429 - 3361.3)^2 + (3448 - 3361.3)^2 + (3677 - 3361.3)^2 + (3872 - 3361.3)^2\\ &+ (4001 - 3361.3)^2 + (4116 - 3361.3)^2)\big)\\ = 344920.5,&\end{aligned}$$

and finally the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{344920.5} = 587.30.$$

```
## In R we compute it by:
x <- c(2474, 2547, 2830, 3219, 3429, 3448, 3677, 3872, 4001, 4116)
mean(x)

[1] 3361


var(x)

[1] 344920


sd(x)

[1] 587.3
```

Interpretation: if we consider the 10 female births as a representative sample from the population of all female births, we estimate the population mean weight μ to be $\hat{\mu} = 3361$ g. Individual female births will not be exactly 3361 g each of them, they will typically differ from that value. They are estimated to differ from the mean by $s = 587$ g on average. Since they are expected to differ both above and below the mean, one would expect most female births to be within plus/minus $2 \cdot 587 = 1174$ g of the mean.

An average absolute difference to the mean (i.e. estimated by the sample standard deviation s) somehow matches (on a linear scale) that individual observations distribute from the mean minus $2s$ to the mean plus $2s$! (at least if they are evenly distributed).

b) Compute the same summary statistics of the male births. Compare and explain differences with the results for the female births.

|||| Solution

For the manual computation, the same three formulas as above should be used. Here we show the R-computations and results:

```
## In R we compute it by:
y <- c(2844, 2863, 2963, 3239, 3379, 3449, 3582, 3926, 4151, 4356)
mean(y)

[1] 3475


var(y)

[1] 283158


sd(y)

[1] 532.1
```

Thus

$$
\begin{aligned}
\bar{x} &= 3475.2,\\
s^2 &= 283158,\\
s &= 532.13.
\end{aligned}
$$

Comparison: the male birth weights are on average a little higher, but the standard deviation is a little smaller.

An important part of the course is to give you methods that would make it possible for you to do a comparison of these numbers in a more elaborate and clever way than above. A concern for the thoughtful reader would be: what might happen if we repeated this study by recording birth weights for another sample of 2×10 births? Would the comparison come out the same way or differently? Actually, it IS possible to answer this question based just on a SINGLE sample, if we include some probability calculations in the statement.

c) Find the five quartiles for each sample — and draw the two box plots with pen and paper (i.e. not using R.)

|||| Solution

Note that the 10 weights are already ordered in the data table, so the first step of finding the quartiles have been carried out for us. With $n = 10$ we get the following values for np:

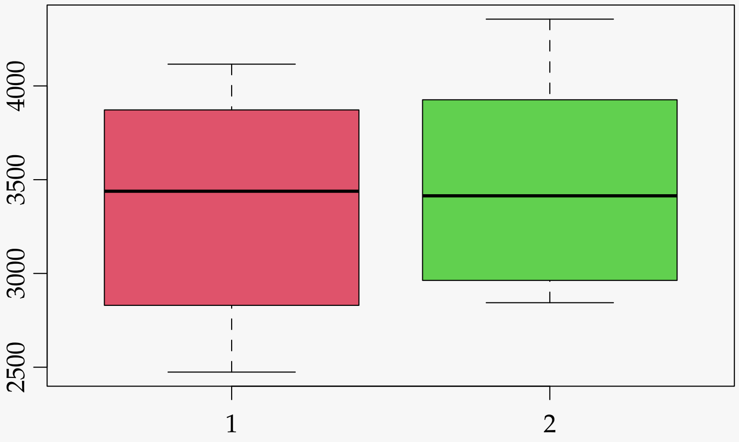
	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$
np	0	2.5	5	7.5	10

This means that we according to the definition of quantiles (or percentiles) can read off the Q_1 and Q_3 as the 3rd and the 8th observation and the median as the average of the 5th and 6th observation:

	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$
np	0	2.5	5	7.5	10
Quartile	Min	Q_1	Median	Q_3	Max
Females	2474	2830	(3429+3448)/2	3872	4116
Males	2844	2963	(3379+3449)/2	3926	4356

Now the two basic box plots could be made from these 2 × 5 numbers:

```
## In R we make the modifed box plots by
boxplot(list(x,y), col=2:3)
```



d) Are there any “extreme” observations in the two samples (use the *modified* *box plot* definition of extremness)?

|||| Solution

As the modified box plot is the default choice in R, and no individual observations are seen beyond the whiskers, there are no extreme observations (which by the way is defined as an observation further than $1.5 \cdot \text{IQR}$) away from the box.

e) What are the coefficient of variations in the two groups?

|||| Solution

The coefficient of variation (CV) is the standard deviation seen relative to the mean, thus for the females it is

$$CV_{\text{female}} = \frac{s_x}{\bar{x}} \cdot 100\% = \frac{587.2993}{3361.3} \cdot 100\% = 17.5\%,$$

and for males it is

$$CV_{\text{male}} = \frac{s_y}{\bar{y}} \cdot 100\% = \frac{532.1261}{3475.2} \cdot 100\% = 15.3\%.$$

1.2 Course Grades

Exercise 1.2 Course grades

To compare the difficulty of 2 different courses at a university the following grades distributions (given as number of pupils who achieved the grades) were registered:

	Course 1	Course 2	Total
Grade 12	20	14	34
Grade 10	14	14	28
Grade 7	16	27	43
Grade 4	20	22	42
Grade 2	12	27	39
Grade 0	16	17	33
Grade -3	10	22	32
Total	108	143	251

a) What is the median of the 251 achieved grades?

Solution

We look at the 251 grades seen from the Total column of the table. Seen from below, these 251 grades are already ordered, so to find the median we should find the 126th ordered observation from below. Since there are 104 grades in the -3, 0, and 2 Grade categories and 42 in the Grade 4 category, the 126th ordered observation from below is a 4, so the answer is: the median is 4.

Just a note about that it is actually the *sample median* which is asked for, however as noted in Remark 1.3, the *sample* is left out. Further, it is noticed that the *sample median* can be used as an estimate of the *population median* in the same way as illustrated for the mean in Figure 1.1, same goes for quantiles, quartiles, IQR, and all other statistics.

b) What are the quartiles and the IQR (Inter Quartile Range)?

Solution

Since $n \cdot 0.25 = 251 \cdot 0.25 = 62.75$ and $n \cdot 0.75 = 251 \cdot 0.75 = 188.25$ we must find the lower and upper quartiles Q_1 and Q_3 as the 63rd and 189th observation from below. Let's look at the accumulated (from below) numbers:

	Total	Acccum. (from below)
Grade 12	34	251
Grade 10	28	217
Grade 7	43	189
Grade 4	42	146
Grade 2	39	104
Grade 0	33	65
Grade -3	32	32

So it becomes clear that

$$\begin{aligned} Q_1 &= 0, \\ Q_3 &= 7, \\ \mathrm{IQR} &= 7 - 0 = 7. \end{aligned}$$

Finally, a notice about that here the quartiles are the actually the *sample quartiles* and they can be thought of as estimates for the *population quartiles*, as illustrated for the *mean* in Figure 1.1. Actually to be consistent in notation we should use a 'hat' to indicate this, e.g. the *first sample quartile* $\hat{Q}_1$ is an estimate of the *first population quartile*, however to simplify and due to tradition this is not done.

1.3 Cholesterol

|||| Exercise 1.3 Cholesterol

In a clinical trial of a cholesterol-lowering agent, 15 patients' cholesterol (in mmol L^{-1}) was measured before treatment and 3 weeks after starting treatment. Data is listed in the following table:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4

a) What is the median of the cholesterol measurements for the patients before treatment, and similarly after treatment?

|||| Solution

To find the medians we need to order both data sets, and then, since $n = 15$, an odd number, the median is the 8th observation $x_{(8)}$ in the ordered set. This is done "maunally" (or call it step by step) by R in the following way:

```
## Reading the data into R
before <- c(9.1, 8.0, 7.7, 10.0, 9.6, 7.9, 9.0, 7.1, 8.3, 9.6,
            8.2, 9.2, 7.3, 8.5, 9.5)
after  <- c(8.2, 6.4, 6.6, 8.5, 8.0, 5.8, 7.8, 7.2, 6.7, 9.8,
            7.1, 7.7, 6.0, 6.6, 8.4)
## Making ordered vectors using the 'sort' function
sortedBefore <- sort(before)
sortedAfter <- sort(after)
## Printing the ordered vectors
sortedBefore

 [1]  7.1  7.3  7.7  7.9  8.0  8.2  8.3  8.5  9.0  9.1  9.2
[12]  9.5  9.6  9.6 10.0


sortedAfter

 [1] 5.8 6.0 6.4 6.6 6.6 6.7 7.1 7.2 7.7 7.8 8.0 8.2 8.4 8.5
[15] 9.8


## Printing the 8th observation in these vectors
sortedBefore[8]

[1] 8.5


sortedAfter[8]

[1] 7.2
```

Giving the results

$$\begin{aligned} \text{'median before'} &= 8.5, \\ \text{'median after'} &= 7.2. \end{aligned}$$

Using the R-function `summary` one would get them directly, together with more info:

```
## Direct summaries
summary(before)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.10    7.95    8.50    8.60    9.35   10.00


summary(after)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.80    6.60    7.20    7.39    8.10    9.80
```

We have also learned that we can use the R-function `quantile` to get the quartiles, and to use the percentile definition given in Definition 1.7, we should use the `type=2` argument:

```
## Now using the quantile function
quantile(before, type=2)

  0%  25%  50%  75% 100%
 7.1  7.9  8.5  9.5 10.0


quantile(after, type=2)

  0%  25%  50%  75% 100%
 5.8  6.6  7.2  8.2  9.8
```

It can be noted that some of the quartiles given here are not exactly the same as those given by the `summary` function. This is due to the fact that `summary` function uses the default setting of the `quantile` function, so NOT the `type=2` option. We will live with this little difference, which will not cause any problems. We consider both results just as valid, just only one of them are defined in the material.

b) Find the standard deviations of the cholesterol measurements of the patients before and after treatment.

|||| Solution

We should use the defining formulae for the sample mean (Def. 1.4) and sample standard deviation (Def. 1.11) for each sample

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

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

In R we find these as:

```
mean(before)

[1] 8.6

mean(after)

[1] 7.387

sd(before)

[1] 0.9024

sd(after)

[1] 1.09
```

c) Find the sample covariance between cholesterol measurements of the patients before and after treatment.

|||| Solution

Define the "before treatment" sample as $x_1, x_2, \ldots, x_{15}$ and the "after treatment" sample $y_1, y_2, \ldots, y_{15}$, then the sample covariance is found using Definition 1.18 as

$$s_{xy} = \frac{1}{14}\sum_{i=1}^{15}(x_i - 8.6)(y_i - 7.3867) = 11.15/14 = 0.79643.$$

In R we find this as:

```
## Calculate the sample covariance 'manually'
sum((before - mean(before)) * (after - mean(after)))/14

[1] 0.7964


sum((before - 8.6) * (after - 7.3867))/14

[1] 0.7964


## or use the inbuilt function
cov(before,after)

[1] 0.7964
```

d) Find the sample correlation between cholesterol measurements of the patients before and after treatment.

|||| Solution

This is Definition 1.19 and simply

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{0.79643}{0.90238 \cdot 1.0901} = 0.8096.$$

In R we find this by:

```
## 'Manually'
0.79643/(0.90238*1.0901)

[1] 0.8096


## or
cov(before, after)/(sd(before) * sd(after))

[1] 0.8096


## or
cor(before, after)

[1] 0.8096
```

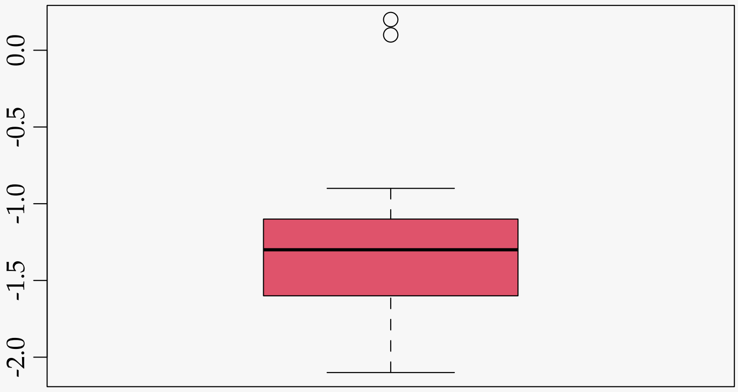
e) Compute the 15 differences (Dif = Before − After) and do various summary statistics and plotting of these: sample mean, sample variance, sample standard deviation, boxplot etc.

|||| Solution

The differences are:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4
Dif	0.9	1.6	1.1	1.5	1.6	2.1	1.2	-0.1	1.6	-0.2	1.1	1.5	1.3	1.9	1.1

```
## Analysis of differences
dif <- after-before
## The summary
summary(dif)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -2.10   -1.60   -1.30   -1.21   -1.10    0.20


## Sample variance
var(dif)

[1] 0.4098


## Sample standard deviation
sd(dif)

[1] 0.6402


## Boxplot
boxplot(dif,col=2)
```



The mean effect (decrease of cholesterol due to treatment) would be estimated at 1.2 nMol/l. But clearly there is also a high degree of differences in what the effect is: the standard deviation of (all) the differences is 0.64. Looking at the boxplot, we find two patients with values identified as extreme, which from the data table is seen to be patient no 8 and 10. The better way, maybe, here to tell the story would be the following: for 2 out of 15 patients (13% of patients) the treatment clearly had no effect. For the remaining 13 out of 15 (87% of patients) the treatment had the following average effect and standard deviation (recomputing the mean and standard deviation for the 13 patients):

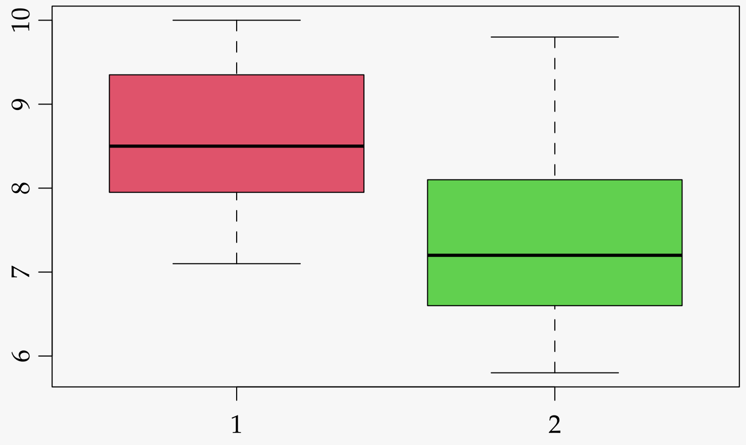
```
## Analysis of 13 non-extreme differences
## Take out observation 8 and 10
dif13 <- dif[-c(8,10)]
## Mean of the 13 differences
mean(dif13)

[1] -1.423


## Standard deviation of the 13 differences
sd(dif13)

[1] 0.3468
```

f) Observing such data the big question is whether an average decrease in cholesterol level can be "shown statistically". How to formally answer this question is presented in Chapter 3, but consider now which summary statistics and/or plots would you look at to have some idea of what the answer will be?

|||| Solution

In the previous question we were studying the differences in the attempt to answer this question. One could also, as we did initially look at the data separately, and e.g. supplement by the grouped boxplot:

```
## The grouped boxplot
boxplot(list(before,after),col=2:3)
```



And we would conclude: the average effect is 1.2 (we see no extreme patients in this plot!), and the standard deviation within each group of data is around 1 (see above: $s_{\text{before}} = 0.9$ and $s_{\text{after}} = 1.1$).

Which of the two approaches do you prefer - the "difference"-approach or the "separate"-approach?

We would definitely recommend the "difference"-approach, or as we will call it later, the "paired" approach, since this match the setup of the study, and in the most correct way uses the relevant information. Note how the difference-approach identifies the outliers/extremes and also ends up with much smaller standard deviations, also seen by the range and/or box-widths(IQR) in the box-plots. The point is that in the differences we have removed the variability stemming from the characteristics of each patient (e.g. body mass, genes, etc.). One phrase used is that in such an experiment like this, a patient acts as his own control, and hence the fact the patients are different does not blur the important effect signal.

1.4 Project start

|||| Exercise 1.4 Project start

a) Go to CampusNet and take a look at the first project and read the project page on the website for more information (02323.compute.dtu.dk/projects or 02402.compute.dtu.dk/projects). Follow the steps to import the data into R and get started with the explorative data analysis.

|||| Solution

There is no results for this exercise - you have to do it as a project.

Chapter 2

Probability and simulation (solutions to exercises)

Contents

2.1 Discrete random variable

Exercise 2.1 Discrete random variable

a) Let X be a stochastic variable. When running the R-command `dbinom(4,10,0.6)` R returns 0.1115, written as:

```
dbinom(4,10,0.6)
[1] 0.1115
```

What distribution is applied and what does 0.1115 represent?

Solution

The distribution applied is the binomial distribution with $n = 10$ observations and $p = 0.6$ the probability for success. The value, 0.1115 output by R is the value of the probability density function (pdf) for $x = 4$, hence the probability of getting exactly 4 successes in 10 draws with replacement with a success probability of 60%.

b) Let X be the same stochastic variable as above. The following are results from R:

```
pbinom(4,10,0.6)
[1] 0.1662

pbinom(5,10,0.6)
[1] 0.3669
```

Calculate the following probabilities: $P(X \leq 5)$, $P(X < 5)$, $P(X > 4)$ and $P(X = 5)$.

|||| Solution

```
## P(X <= 5)
pbinom(5,10,0.6)

[1] 0.3669


## P(X < 5)
pbinom(4,10,0.6)

[1] 0.1662


## P(X > 4)
1 - pbinom(4,10,0.6)

[1] 0.8338


## P(X = 5)
pbinom(5,10,0.6) - pbinom(4,10,0.6)

[1] 0.2007
```

c) Let X be a stochastic variable. From R we get:

```
dpois(4,3)
[1] 0.168
```

What distribution is applied and what does 0.168 represent?

|||| Solution

The Poisson distribution and the value is the probability of getting $x = 4$ events per interval when the average events per interval $\lambda = 3$ (i.e. the mean).

d) Let X be the same stochastic variable as above. The following are results from R:

```
ppois(4,3)
[1] 0.8153

ppois(5,3)
[1] 0.9161
```

Calculate the following probabilities: $P(X \leq 5)$, $P(X < 5)$, $P(X > 4)$ and $P(X = 5)$.

Solution

```
## P(X <= 5))
ppois(5,3)

[1] 0.9161


## P(X < 5)
ppois(4,3)

[1] 0.8153


## P(X > 4)
1 - ppois(4,3)

[1] 0.1847


## P(X = 5)
ppois(5,3) - ppois(4,3)

[1] 0.1008
```

2.2 Course passing proportions

|||| Exercise 2.2 Course passing proportions

a) If a passing proportion for a course given repeatedly is assumed to be 0.80 on average, and there are 250 students who are taking the exam each time, what is the expected value, μ and standard deviation, σ, for the number of students who do not pass the exam for a randomly selected course?

|||| Solution

If X is the number of students not passing a randomly selected course, this random variable follows the binomial distribution with $n = 250$ and $p = 0.20$, so we use the formula for the mean and variance of the binomial

$$\mu = np = 0.2 \cdot 250 = 50, \quad \sigma^2 = np(1-p) = 250 \cdot 0.2 \cdot 0.8 = 40 = 6.32^2.$$

So the answer is: $\mu = 50$ and $\sigma = 6.32$.

2.3 Notes in a box

|||| Exercise 2.3 Notes in a box

A box contains 6 notes:

On 1 of the notes there is the number 1
On 2 of the notes there is the number 2
On 2 of the notes there is the number 3
On 1 of the notes there is the number 4

Two notes are drawn at random from the box, and the following random variable is introduced: X, which describes the number of notes with the number 4 among the 2 drawn. The two notes are drawn without replacement.

a) The mean and variance for X, and $P(X = 0)$ are?

|||| Solution

X follows the hypergeometric distribution (2.24) with $N = 6$, $a = 1$, and $n = 2$ so the mean and variance formula (2.25) for this distribution is used to find

$$\mu_x = n\frac{a}{N} = 2/6,$$

and

$$\sigma_x^2 = n\frac{a(N-a)}{N^2}\frac{N-n}{N-1} = 2\frac{1\cdot(6-1)\cdot(6-2)}{6^2\cdot(6-1)} = \frac{2\cdot 5\cdot 4}{36\cdot 5} = 8/36 = 2/9.$$

And the hypergeometric probability formula (2-25) gives

$$P(X=0) = \frac{\binom{1}{0}\binom{5}{2}}{\binom{6}{2}} = \frac{5\cdot 4\cdot 2}{2\cdot 6\cdot 5} = 2/3.$$

So the correct answer is: $\mu_x = 1/3$, $\sigma_x^2 = 2/9$ and $P(X = 0) = 2/3$.

b) The 2 notes are now drawn with replacement. What is the probability that none of the 2 notes has the number 1 on it?

Solution

The binomial pdf (2-20) is used in R:

```
dbinom(0, size=2, prob=1/6)

[1] 0.6944
```

Another way is, since it is zero successes, then there is 5/6 probability of not getting a success and since we want that to happen in both two draws, then we can simply multiply the probabilities

$$P(X=0)=\frac{5}{6}\cdot\frac{5}{6}=\frac{25}{36}$$

So the correct answer is: 0.694 or $\frac{25}{36}$.

2.4 Consumer survey

|||| Exercise 2.4 Consumer survey

In a consumer survey performed by a newspaper, 20 different groceries (products) were purchased in a grocery store. Discrepancies between the price appearing on the sales slip and the shelf price were found in 6 of these purchased products.

a) At the same time a customer buys 3 random (different) products within the group consisting of the 20 goods in the store. The probability that no discrepancies occurs for this customer is?

|||| Solution

Let X denote the number of discrepancies when purchasing 3 random (different) products within the group of the 20 goods in the store. X then follows the hypergeometric distribution (NOT the binomial!!) (why not binomial: because you don't potentially by two goods of the same kind - you DO buy 3 DIFFERENT ones and hence having bought one - you do NOT "put it back" again and the randomly select - it is WITHOUT replacement). Therefore

$$P(X=0) = \frac{\binom{6}{0}\binom{14}{3}}{\binom{20}{3}} = \frac{14 \cdot 13 \cdot 12 \cdot 3 \cdot 2}{20 \cdot 19 \cdot 18 \cdot 3 \cdot 2} = \frac{91}{15 \cdot 19} = 0.3192982.$$

Hence the answer is: 0.319.

2.5 Hay delivery quality

|||| Exercise 2.5 Hay delivery quality

A horse owner receives 20 bales of hay in a sealed plastic packaging. To control the hay, 3 bales of hay are randomly selected, and each checked whether it contains harmful fungal spores.

It is believed that among the 20 bales of hay 2 bales are infected with fungal spores. A random variable X describes the number of infected bales of hay among the three selected.

a) The mean of X, (μ_X), the variance of X, (σ_X^2) and $P(X \geq 1)$ are?

|||| Solution

The hypergeometric distribution with $N = 20$, $a = 2$ and $n = 3$ is used ("sampling without replacement"). First the mean and variance formulas for the hypergeometric distribution gives

$$\mu_x = 3\frac{2}{20} = 0.3,$$

and

$$\sigma_x^2 = 3\frac{2}{20}(1 - \frac{2}{20})(\frac{20-3}{20-1}) = 0.2415789.$$

Then we find

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{18}{3}\binom{2}{0}}{\binom{20}{3}} = 0.2842,$$

So the answer is: $\mu_x = 0.3$, $\sigma_x^2 = 0.242$ and $P(X \geq 1) = 0.2842$.

b) Another supplier advertises that no more than 1% of his bales of hay are infected. The horse owner buys 10 bales of hay from this supplier, and decides to buy hay for the rest of the season from this supplier if the 10 bales are error-free.
What is the probability that the 10 purchased bales of hay are error-free, if 1% of the bales from a supplier are infected (p_1) and the probability that the 10 purchased bales of hay are error-free, if 10% of the bales from a supplier are infected (p_{10})?

Solution

We use the binomial distribution(s) ("sampling with replacement"="sampling from an infinte population").

We can use the pdf or cdf in R. For p_1:

```
dbinom(x=0, size=10, prob=0.01)

[1] 0.9044


pbinom(q=0, size=10, prob=0.01)

[1] 0.9044
```

and for p_2:

```
dbinom(x=0, size=10, prob=0.1)

[1] 0.3487


pbinom(q=0, size=10, prob=0.1)

[1] 0.3487
```

This 10 independent events in series where each must be non-infected, i.e. a success is then $1-0.01$ and $1-0.1$, and can thus also be calculated simply by multiplying the probability of success in each event

$$p_1 = P(X_1 = 0) = 0.99^{10} = 0.9044,$$

and

$$p_{10} = P(X_{10} = 0) = 0.90^{10} = 0.3487.$$

So the answer becomes: $P_1 = 0.9044$ and $P_{10} = 0.3487$.

2.6 Newspaper consumer survey

|||| Exercise 2.6 Newspaper consumer survey

In a consumer survey performed by a newspaper, 20 different groceries (products) were purchased in a grocery store. Discrepancies between the price appearing on the sales slip and the shelf price were found in 6 of these purchased products.

a) Let X denote the number of discrepancies when purchasing 3 random (different) products within the group of the 20 products in the store. What is the mean and variance of X?

|||| Solution

We must use the hypergeometric distribution, since we draw $n = 3$ of the $N = 20$ products, where $a = 6$ have discrepancies, with no replacement (we cannot draw the same product twice).

The mean and variance of the hypergeometric distribution is found in Theorem 2.25, thus

$$\mu_X = n \cdot \frac{a}{N} = 3\frac{6}{20} = 0.90,$$

and

$$\sigma_X^2 = n \cdot \frac{a}{N}\left(1 - \frac{a}{N}\right)\left(\frac{N-n}{N-1}\right) = 3\frac{6}{20} \cdot \left(1 - \frac{6}{20}\right)\left(\frac{20-3}{20-1}\right) = 0.5636842.$$

Hence the answer is: $\mu_X = 0.90$ and $\sigma_X^2 = 0.56$.

2.7 A fully automated production

Exercise 2.7 A fully automated production

On a large fully automated production plant items are pushed to a side band at random time points, from which they are automatically fed to a control unit. The production plant is set up in such a way that the number of items sent to the control unit on average is 1.6 item pr. minute. Let the random variable X denote the number of items pushed to the side band in 1 minute. It is assumed that X follows a Poisson distribution.

a) What is the probability that there will arrive more than 5 items at the control unit in a given minute is?

Solution

With $\lambda = 1.6$, we find that

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.994 = 0.006$$

where the 0.994 can be found with R by:

```
1-ppois(q=5, lambda=1.6)

[1] 0.00604
```

So the answer is: approximately 0.6%.

b) What is the probability that no more than 8 items arrive to the control unit within a 5-minute period?

Solution

With $\lambda_{5minutes} = 8$, we find that

$$P(X \leq 8) = 0.593$$

where the 0.593 can be found by R:

```
ppois(q=8, lambda=8)

[1] 0.5925
```

So the answer is: approximately 59.3%.

2.8 Call center staff

Exercise 2.8 Call center staff

The staffing for answering calls in a company is based on that there will be 180 phone calls per hour randomly distributed. If there are 20 calls or more in a period of 5 minutes the capacity is exceeded, and there will be an unwanted waiting time, hence there is a capacity of 19 calls per 5 minutes.

a) What is the probability that the capacity is exceeded in a random period of 5 minutes?

Solution

The 60 minutes mean of 180 calls corresponds to a 5 minutes mean of $\mu_{5min} = 180/12 = 15$ and the event of exceeding capacity is the event of observing at least 20 calls within 5 minutes. Let X represent the number of calls within a randomly chosen 5 minutes interval, then we need to find $P(X \geq 20)$, which in R:

```r
1 - ppois(q=19, lambda=15)

[1] 0.1248
```

So the correct answer is: $P(X \geq 20) = 0.125$, where $X \sim Po(15)$.

b) If the probability should be at least 99% that all calls will be handled without waiting time for a randomly selected period of 5 minutes, how large should the capacity per 5 minutes then at least be?

Solution

Let X (as above) represent the number of calls in a randomly chosen 5 minutes interval, i.e. $X \sim Po(15)$. It is required that

$$P(\text{“All calls will be handled”}) = P(X \leq x_{\text{capacity}}) \geq 0.99$$

where x_{capacity} must be the smallest capacity which keeps the probability above 0.99. Using R to find $P(X \leq x_{\text{capacity}})$ for 22,23,...,26:

```
ppois(q=22:26, lambda=15)

[1] 0.9673 0.9805 0.9888 0.9938 0.9967
```

shows that the first (smallest) capacity level achieving this is 25.

So the correct answer is: the capacity must be at least 25 per 5 minutes

2.9 Continuous random variable

|||| Exercise 2.9 Continuous random variable

a) The following R commands and results are given:

```
pnorm(2)
[1] 0.9772

pnorm(2,1,1)
[1] 0.8413

pnorm(2,1,2)
[1] 0.6915
```

Specify which distributions are used and explain the resulting probabilities (preferably by a sketch).

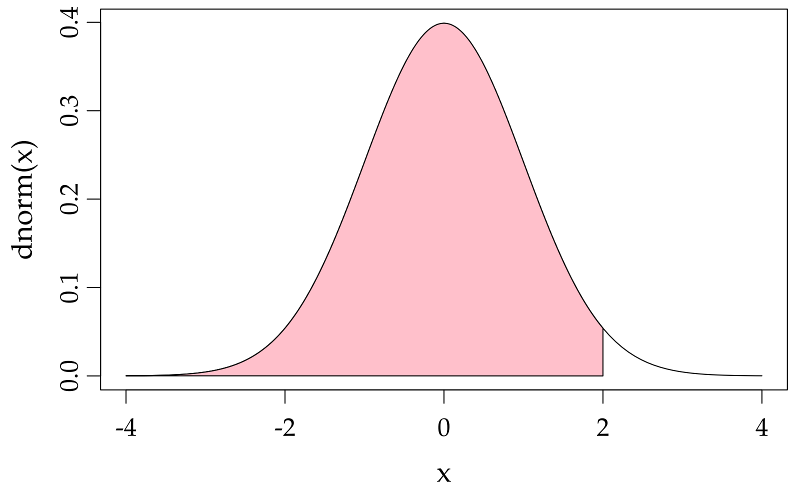
|||| Solution

The normal distribution function (or normal cumulated density function cdf). The found probabilities are

- $P(X \leq 2)$ (or $P(X < 2)$) for $X \sim N(\mu = 0, \sigma^2 = 1)$
- $P(X \leq 2)$ (or $P(X < 2)$) for $X \sim N(\mu = 1, \sigma^2 = 1)$
- $P(X \leq 2)$ (or $P(X < 2)$) for $X \sim N(\mu = 1, \sigma^2 = 4)$

A sketch for the first, $P(Z \leq 2)$ (using Z indicates that is follows the standard normal distribution $N(0,1)$):

```
curve(dnorm, xlim=c(-4,4))
xseq <- seq(-4, 2, len=1000)
polygon(x=c(xseq,2,xseq[1]),
        y=c(dnorm(xseq),0,dnorm(xseq[1])),
        col="pink")
```



b) What is the result of the following command: `qnorm(pnorm(2))`?

Solution

`qnorm` and `pnorm` are each others inverse so the result is the same as the argument: 2

c) The following R commands and results are given:

```
qnorm(0.975)
[1] 1.96

qnorm(0.975,1,1)
[1] 2.96

qnorm(0.975,1,2)
[1] 4.92
```

State what the numbers represent in the three cases (preferably by a sketch).

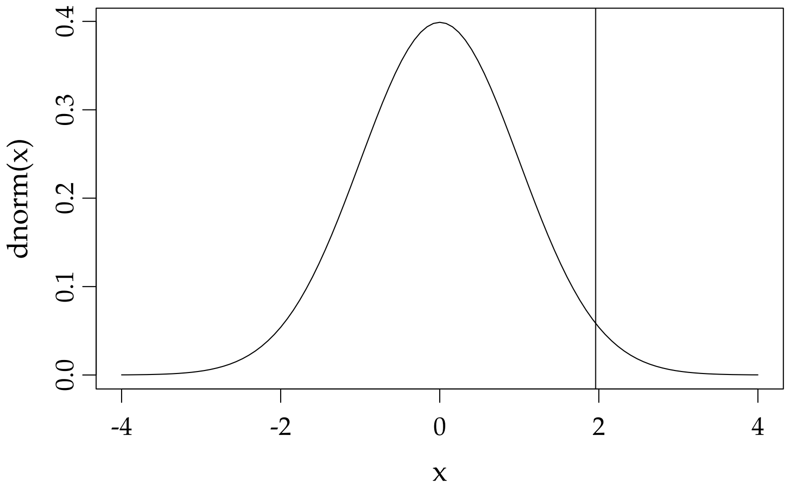
|||| Solution

The 97.5% percentiles for

- $N(\mu = 0, \sigma^2 = 1)$
- $N(\mu = 1, \sigma^2 = 1)$
- $N(\mu = 1, \sigma^2 = 4)$

A sketch for the first:

```
## Plot the standard normal distribution
curve(dnorm, xlim=c(-4,4))
## Add a vertical line at the 0.975 quantile
abline(v=qnorm(0.975))
```



2.10 The normal pdf

|||| Exercise 2.10 The normal pdf

a) Which of the following statements regarding the probability density function of the normal distribution $N(1, 2^2)$ is false?

1. The total area under the curve is equal to 1.0
2. The mean is equal to 1^2
3. The variance is equal to 2
4. The curve is symmetric about the mean
5. The two tails of the curve extend indefinitely
6. Don't know

|||| Solution

We need to find the one false statement, and go through the claims one at a time:

1. True, the total area under the curve is one, since this is true for all probability distributions, see Definition 2.32
2. True. The mean value is one, and we have that $1^2 = 1$
3. False, the standard deviation is two and the variance is four
4. True, the distribution is symmetric around the mean value
5. True, the normal density is defined between $-\infty$ and ∞

Correct answer is 3.

Let X be normally distributed with mean 24 and variance 16

b) Calculate the following probabilities:
   - $P(X \leq 20)$
   - $P(X > 29.5)$
   - $P(X = 23.8)$

|||| Solution

- $P(X \leq 20)$

```
pnorm(20, mean=24, sd=4)
[1] 0.1587
```

- $P(X > 29.5)$

```
1 - pnorm(29.5, mean=24, sd=4)
[1] 0.08457
```

- The probability of a continuous random variable to have the outcome equal to a single value is zero, i.e. $P(X = 23.8) = 0$.

2.11 Computer chip control

Exercise 2.11 Computer chip control

A machine for checking computer chips uses on average 65 milliseconds per check with a standard deviation of 4 milliseconds. A newer machine, potentially to be bought, uses on average 54 milliseconds per check with a standard deviation of 3 milliseconds. It can be used that check times can be assumed normally distributed and independent.

a) What is the probability that the time savings per check using the new machine is less than 10 milliseconds is?

Solution

Let $X_{\text{old}} \sim N(65, 4^2)$ and $X_{\text{new}} \sim N(54, 3^2)$. If we let U denote the time saving per check, we have that $U = X_{\text{old}} - X_{\text{new}}$. We now from Theorem 2.40 that a linear combination of normal random variables is also normal and from Theorem 2.56

$$\mathrm{E}(U) = \mathrm{E}(X_{\text{old}} - X_{\text{new}}) = \mathrm{E}(X_{\text{old}}) - \mathrm{E}(X_{\text{new}}) = 65 - 54 = 11,$$

and

$$\mathrm{V}(U) = \mathrm{V}(X_{\text{old}} - X_{\text{new}}) = \mathrm{V}(X_{\text{old}}) + \mathrm{V}(X_{\text{new}}) = 16 + 9 = 25.$$

Hence $U \sim N(11, 5^2)$.

We are asked to find $P(U < 10)$, so

```r
pnorm(q=10, mean=11, sd=5)
```

```
[1] 0.4207
```

Another way to solve this is via a transformation to the standard normal distribution using Theorem 2.43. Then

$$
\begin{aligned}
P(U < 10) &= P\left(Z < \frac{10 - \mathrm{E}(U)}{\sqrt{\mathrm{V}(U)}}\right) \\
&= P\left(Z < \frac{10 - (65 - 54)}{\sqrt{3^2 + 4^2}}\right) \\
&= P\left(Z < \frac{-1}{5}\right) \\
&= P\left(Z < -0.2\right) \\
&= 0.4207,
\end{aligned}
$$

where the latter can be found in R:

```
z <- (10-11)/5
z

[1] -0.2


pnorm(z)

[1] 0.4207
```

So the answer is: approximately 42%.

b) What is the mean (μ) and standard deviation (σ) for the total time use for checking 100 chips on the new machine is?

|||| Solution

Let U be the total time use for checking 100 chips on the new machine, that is

$$U = \sum_{i=1}^{100} X_i$$

where $X_i \sim N(54, 3^2)$. So we find, using mean and variance identities in Theorem

2.56, that

$$\begin{aligned}\mu = \mathrm{E}(U) = \mathrm{E}\left(\sum_{i=1}^{100} X_i\right) &= \mathrm{E}(X_1 + X_2 + \cdots + X_{100})\\ &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_{100})\\ &= \sum_{i=1}^{100} E(X_i) = \sum_{i=1}^{100} 54 = 100 \cdot 54 = 5400,\end{aligned}$$

and

$$\begin{aligned}\sigma^2 = \mathrm{V}(U) = \mathrm{V}\left(\sum_{i=1}^{100} X_i\right) &= \mathrm{V}(X_1 + X_2 + \cdots + X_{100})\\ &= \mathrm{V}(X_1) + \mathrm{V}(X_2) + \cdots + \mathrm{V}(X_{100})\\ &= \sum_{i=1}^{100} \mathrm{V}(X_i) = \sum_{i=1}^{100} 9 = 100 \cdot 9.\end{aligned}$$

So the answer is: $\mu = 100 \cdot 54 = 5400$ ms and $\sigma = 3 \cdot \sqrt{100} = 30$ ms.

2.12 Concrete items

Exercise 2.12 Concrete items

A manufacturer of concrete items knows that the length (L) of his items are reasonably normally distributed with $\mu_L = 3000$ mm and $\sigma_L = 3$ mm. The requirement for these elements is that the length should be not more than 3007 mm and the length must be at least 2993 mm.

a) The expected error rate in the manufacturing will be?

Solution

The expected error rate is:

$$\begin{aligned} P(L \leq 2993) + P(L \geq 3007) &= P(L \leq 2993) + 1 - P(L \leq 3007) \\ &= 2 \cdot P(L \leq 2993) = 0.01963066. \end{aligned}$$

In R:

```
pnorm(2993, mean=3000, sd=3) + 1 - pnorm(3007, mean=3000, sd=3)

[1] 0.01963


2*pnorm(2993, mean=3000, sd=3)

[1] 0.01963
```

So the answer becomes: approximately 2%.

b) The concrete items are supported by beams, where the distance between the beams is called L_{beam} and can be assumed normal distributed. The concrete items length is still called L. For the items to be supported correctly, the following requirements for these lengths must be fulfilled: $90 \text{ mm} < L - L_{\text{beam}} < 110 \text{ mm}$. It is assumed that the mean of the distance between the beams is $\mu_{\text{beam}} = 2900$ mm. How large may the standard deviation σ_{beam} of the distance between the beams be if you want the requirement fulfilled in 99% of the cases?

|||| Solution

The following should be fulfilled

$$P(90 < L - L_{\text{beam}} < 110) = 0.99.$$

We know that $\text{E}(L - L_{\text{beam}}) = 3000 - 2900 = 100$ and that

$$\text{V}(L - L_{\text{beam}}) = 9 + \sigma_{\text{beam}}^2.$$

So transforming to the standard normal gives

$$0.99 = P(90 < L - L_{\text{beam}} < 110) = P\left(\frac{-10}{\sqrt{9 + \sigma_{\text{beam}}^2}} < Z < \frac{10}{\sqrt{9 + \sigma_{\text{beam}}^2}}\right).$$

So since for the standard normal, we can find that

$$0.99 = P(-z_{0.005} < Z < z_{0.005}),$$

where $z_{0.005} = 2.576$ (in R: `qnorm(0.995)`), we can solve

$$2.576 = \frac{10}{\sqrt{9 + \sigma_{\text{beam}}^2}},$$

for σ_{beam}

$$\sigma_{\text{beam}} = \sqrt{\left(\frac{10}{2.576}\right)^2 - 9} = 2.464.$$

So the answer becomes: $\sigma_{\text{beam}} = 2.46$ mm.

2.13 Online statistic video views

Exercise 2.13 Online statistic video views

In 2013, there were 110,000 views of the DTU statistics videos that are available online. Assume first that the occurrence of views through 2014 follows a Poisson process with a 2013 average: $\lambda_{365days} = 110000$.

a) What is the probability that in a randomly chosen half an hour there is no occurrence of views?

Solution

The half hour intensity is

$$\lambda_{30min} = \lambda_{365days} / (365 \cdot 48) = \frac{110000}{17520} = 6.28.$$

So if X is the number of views in half an hour then, $X \sim Po(6.28)$ and the wanted probability is

$$P(X = 0) = \exp(-6.28) = 0.00187.$$

Or in R:

```
lambda30min <- 110000/(365*24*2)
dpois(x=0, lambda=lambda30min)

[1] 0.001876
```

So the correct answer is: 0.002.

b) There has just been a view, what is the probability that you have to wait more than fifteen minutes for the next view?

|||| Solution

This can be solved either using the Poisson distribution: the 15 minutes rate is

$$\lambda_{15min} = \lambda_{365days}/(365 \cdot 96) = \frac{110000}{2 \cdot 17520} = 3.14.$$

So if X is the number of views in 15 minutes then, $X \sim Po(3.14)$ and the wanted probability is found in R:

```r
lambda30min <- 110000/(365*24*2)
lambda15min <- lambda30min/2
dpois(0,lambda15min)

[1] 0.04331
```

$$P(X = 0) = \exp(-3.14) = 0.043.$$

Or using the exponential distribution: the mean waiting time for a view is (in minutes)

$$\beta = 365 \cdot 24 \cdot 60/110000 = 4.778.$$

which in R:

```r
beta <- 365*24*60/110000
1-pexp(15, rate=1/beta)

[1] 0.04331
```

So, the correct answer is: 0.043.

2.14 Body mass index distribution

|||| Exercise 2.14 Body mass index distribution

The so-called BMI (Body Mass Index) is a measure of the weight-height-relation, and is defined as the weight (W) in kg divided by the squared height (H) in meters:

$$BMI = \frac{W}{H^2}.$$

Assume that the population distribution of BMI is a log-normal distribution with $\alpha = 3.1$ and $\beta = 0.15$ (hence that $\log(BMI)$ is normal distributed with mean 3.1 and standard deviation 0.15).

a) A definition of "being obese" is a BMI-value of at least 30. How large a proportion of the population would then be obese?

|||| Solution

$$P(BMI > 30) = P(\log(BMI) > \log(30)) = P(Z > \frac{\log(30) - 3.1}{0.15}) = P(Z > 2.008) = 0.0223$$

where Z is a standard normal variable $Z \sim N(0,1)$. Or in R:

```
al <- 3.1
be <- 0.15
1-pnorm((log(30)-al)/be)

[1] 0.02232


1-plnorm(30,al,be)

[1] 0.02232
```

So the correct answer is: 2.23%.

2.15 Bivariate normal

|||| Exercise 2.15 Bivariate normal

a) In the bivariate normal distribution (see Example 2.73), show that if Σ is a diagonal matrix then (X_1, X_2) are also independent and follow univariate normal distributions.

|||| Solution

Σ is diagonal so

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix},$$

hence $|\Sigma| = \sigma_1^2 \sigma_2^2$ and

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix},$$

the joint density is therefore

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} e^{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}}, \end{aligned}$$

which can be recognized as the product between two univariate normal pdf's (with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$). As the density is the product of two univariate densities they are independent, see Theorem 2.75.

b) Assume that Z_1 and Z_2 are independent standard normal random variables. Now let X and Y be defined by

$$\begin{aligned} X &= a_{11}Z_1 + c_1, \\ Y &= a_{12}Z_1 + a_{22}Z_2 + c_2. \end{aligned}$$

Show that an appropriate choice of $a_{11}, a_{12}, a_{22}, c_1, c_2$ can give any bivariate normal distribution for the random vector (X, Y), i.e. find $a_{11}, a_{12}, a_{22}, c_1, c_2$ as a function of μ_X, μ_Y and the elements of Σ.

Note that $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ (i.e. here $\Sigma_{12} = \Sigma_{21} = \mathrm{Cov}(X,Y)$), and that any linear combination of random normal variables will result in a random normal variable.

|||| Solution

$$
\begin{aligned}
\mathrm{E}(X) &= c_1 \Rightarrow c_1 = \mu_X, \\
\mathrm{E}(Y) &= c_2 \Rightarrow c_2 = \mu_Y, \\
\mathrm{V}(X) &= a_{11}^2 \Rightarrow a_{11} = \sqrt{\Sigma_{11}}, \\
\mathrm{Cov}(X,Y) &= a_{11}a_{12} \\
&= \sqrt{\Sigma_{11}}a_{12} \Rightarrow a_{12} = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}}}, \\
\mathrm{V}(Y) &= a_{12}^2 + a_{22}^2 \\
&= \frac{\Sigma_{12}^2}{\Sigma_{11}} + a_{22}^2 \Rightarrow a_{22} = \sqrt{\Sigma_{22} - \frac{\Sigma_{12}^2}{\Sigma_{11}}}.
\end{aligned}
$$

c) Use the result to simulate 1000 realization of a bivariate normal random variable with $\boldsymbol{\mu} = (1,2)$ and

$$
\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}
$$

and make a scatter plot of the bivariate random variable.

|||| Solution

In R:

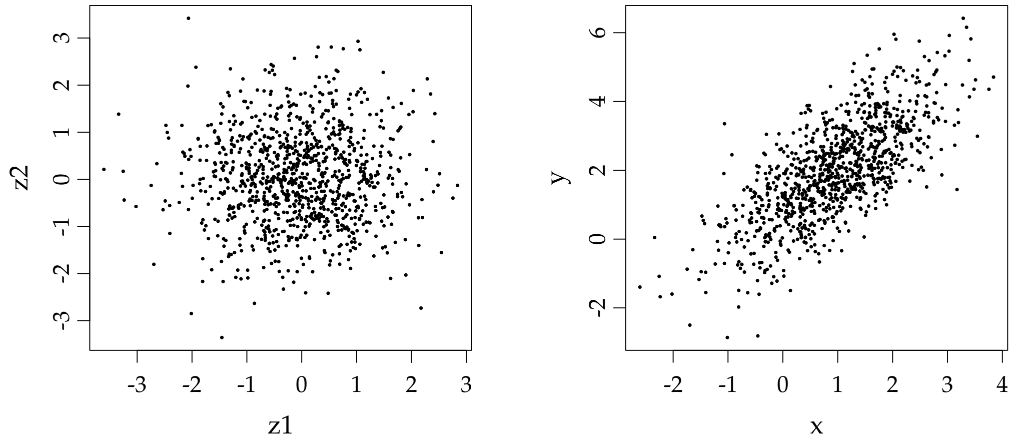
```
## Set the seed
set.seed(125)
## The parameters
Sigma <- matrix(c(1, 1, 1, 2), ncol=2, byrow=TRUE)
Sigma
```

```
     [,1] [,2]
[1,]    1    1
[2,]    1    2
```

```
mu <- c(1, 2)
c1 <- mu[1]
c2 <- mu[2]
a11 <- sqrt(Sigma[1, 1])
a12 <- Sigma[1, 2] / sqrt(Sigma[1, 1])
a22 <- sqrt(Sigma[2, 2] - Sigma[1, 2]^2/Sigma[1, 1])
## Simulate
k <- 1000
z1 <- rnorm(k)
z2 <- rnorm(k)
## The simulation of X an Y
x <- a11 * z1 + c1
y <- a12 * z1 + a22 * z2 + c2
## The sample covariance (\hat{Sigma})
var(cbind(z1, z2))
```

```
        z1      z2
z1 1.01484 0.01453
z2 0.01453 0.97433
```

```
## Make the scatter plots
par(mfrow=c(1,2))
plot(z1, z2, pch=19, cex=0.2)
plot(x, y, pch=19, cex=0.2)
```


z2
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
z1
y
6
4
2
0
-2
-2
-1
0
1
2
3
4
x

2.16 Sample distributions

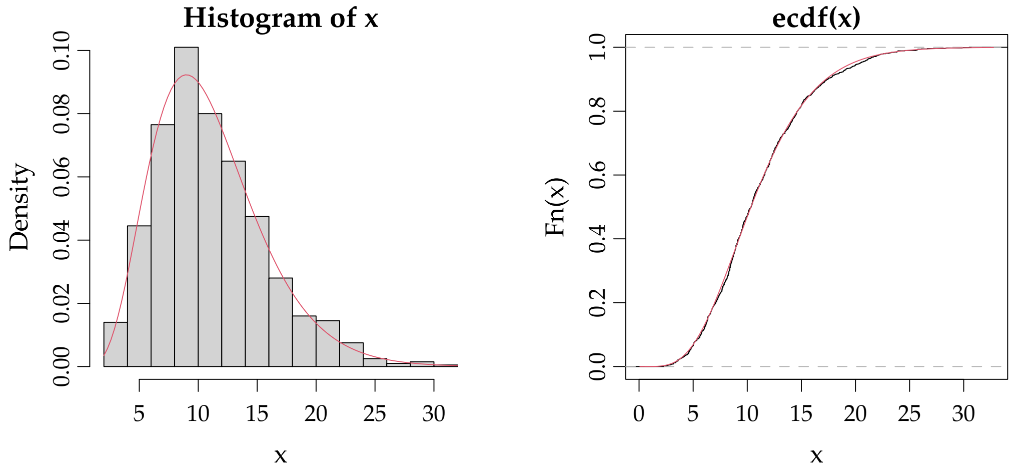
Exercise 2.16 Sample distributions

a) Verify by simulation that $\frac{n_1+n_2-2}{\sigma^2}S_p^2 \sim \chi^2(n_1+n_2-2)$ (See Example 2.85). You may use $n_1 = 5$, $n_2 = 8$, $\mu_1 = 2$, $\mu_2 = 4$, and $\sigma^2 = 2$.

Solution

In R:

```
## Set the seed to get same simulation each time
set.seed(125)
## Set parameters
k <- 1000
n1 <- 5; n2 <- 8
mu1 <- 2; mu2 <- 4
sigma <- sqrt(2)
s1 <- replicate(k, var(rnorm(n1, mean = mu1, sd = sigma)))
s2 <- replicate(k, var(rnorm(n2, mean = mu2, sd = sigma)))
sp <- ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
x <- sp * (n1+n2-2) / sigma^2
## Plot
par(mfrow=c(1,2))
hist(x, freq = FALSE)
curve(dchisq(xseq, df = n1+n2-2), xname="xseq", add =TRUE, col = 2)
plot(ecdf(x))
curve(pchisq(xseq, df = n1+n2-2), xname="xseq", add =TRUE, col = 2)
```



b) Show that if $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$, then

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

Verify the result by simulation. You may use $n_1 = 5$, $n_2 = 8$, $\mu_1 = 2$, $\mu_2 = 4$, and $\sigma^2 = 2$.

|||| Solution

First we consider Theorem 2.87, which makes us think how to find an expression which is standard normal distributed and include the right variables. Since,

$$\begin{aligned} \mathrm{E}(\bar{X} - \bar{Y}) &= \mu_1 - \mu_2, \\ \mathrm{V}(\bar{X} - \bar{Y}) &= \mathrm{V}(\bar{X}) + \mathrm{V}(\bar{Y}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right), \end{aligned}$$

then we can standardize $\bar{X} - \bar{Y}$ by

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1).$$

Then we look for something which is χ^2-distributed and find

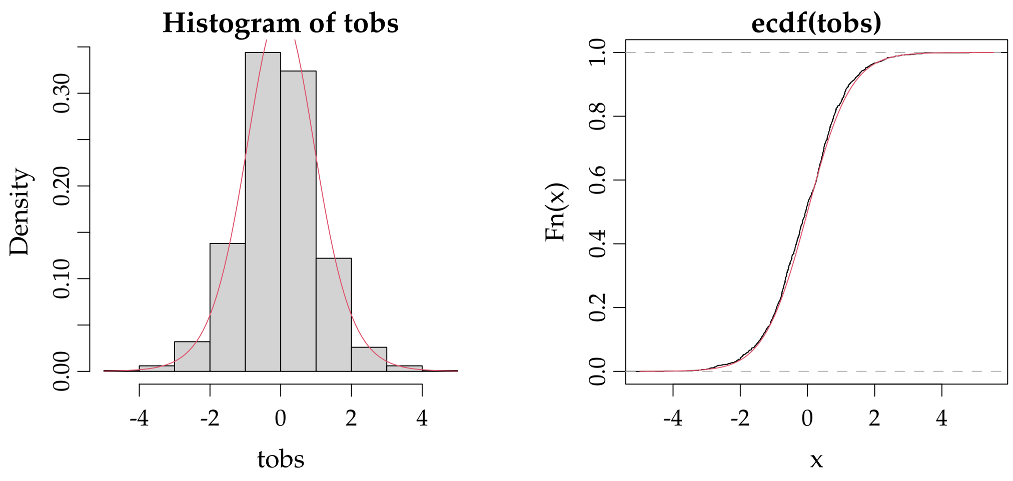
$$\frac{n_1 + n_2 - 2}{\sigma^2} S_p^2 \sim \chi^2(n_1 + n_2 - 2).$$

These two expressions with a little more enables us using Theorem 2.87 to setup

$$\frac{\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}}{\sqrt{\frac{n_1+n_2-2}{\sigma^2}S_p^2\cdot\frac{1}{n_1+n_2-2}}}=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim t(n_1+n_2-2)$$

Verify by simulating the left side and comparing this with the right side:

```
set.seed(125)
## Set parameters
n1 <- 5; n2 <- 8
mu1 <- 2; mu2 <- 4
sigma <- sqrt(2)
## Simulate
k <- 1000
X <- replicate(k, rnorm(n1, mean=mu1, sd=sigma))
Y <- replicate(k, rnorm(n2, mean=mu2, sd=sigma))
s1 <- apply(X, 2, var)
s2 <- apply(Y, 2, var)
m1 <- apply(X, 2, mean)
m2 <- apply(Y, 2, mean)
sp <- ((n1 - 1) * s1 + (n2 - 1) * s2)/(n1 + n2 - 2)
tobs <- (m1 - m2 - (mu1 - mu2))/sqrt(sp * (1/n1 + 1/n2))
## Plot
par(mfrow=c(1,2))
hist(tobs, freq = FALSE)
curve(dt(x, df=n1+n2-2), add=TRUE, col=2)
plot(ecdf(tobs))
curve(pt(x, df=n1+n2-2), add=TRUE, col=2)
```



2.17 Sample distributions 2

|||| Exercise 2.17 Sample distributions 2

Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$, with $X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$ be independent random variables. Hence, two samples before they are taken. S_1^2 and S_2^2 are the sample variances based on the X's and the Y's respectively. Now define a new random variable

$$Q = \frac{S_1^2}{S_2^2} \tag{2-1}$$

a) For n equal $2, 4, 8, 16$ and 32 find:

1. $P(Q < 1)$
2. $P(Q > 2)$
3. $P\left(Q < \frac{1}{2}\right)$
4. $P\left(\frac{1}{2} < Q < 2\right)$

Solution

From Theorem 2.96 we know that Q follows an F-distribution with degrees of freedom $\nu_1 = \nu_2 = n - 1$, and we find the required probabilities with R:

```
## Set n as a vector to get the results for all the n
n <- c(2, 4, 8, 16, 32)
## P(Q<1)
pf(1, df1=n-1, df2=n-1)

[1] 0.5 0.5 0.5 0.5 0.5


## P(Q>2)
1 - pf(2, df1=n-1, df2=n-1)

[1] 0.39183 0.29179 0.19036 0.09553 0.02900


## P(Q<0.5)
pf(0.5, df1=n-1, df2=n-1)

[1] 0.39183 0.29179 0.19036 0.09553 0.02900


## P(0.5<Q<2)
pf(2, df1=n-1, df2=n-1) - pf(0.5, df1=n-1, df2=n-1)

[1] 0.2163 0.4164 0.6193 0.8089 0.9420
```

b) For at least one value of n illustrate the results above by direct simulation from independent normal distributions. You may use any values of μ_1, μ_2 and σ^2.

|||| Solution

In R:

```
set.seed(124)
## Set parameters
mu1 <- 2; mu2 <- 1
sigma <- 2
## Simulate
n <-  8; k <- 100000
S1sq <- replicate(k, var(rnorm(n, mean=mu1, sd=sigma)))
S2sq <- replicate(k, var(rnorm(n, mean=mu2, sd=sigma)))
Q <- S1sq/S2sq
# P(Q<1)
sum(Q < 1) / k
```

```
[1] 0.4987
```

```
# P(Q>2)
1 - pf(2,df1=n-1,df2=n-1)
```

```
[1] 0.1904
```

```
sum(Q > 2) / k
```

```
[1] 0.1894
```

```
# P(Q<0.5)
pf(0.5,df1=n-1,df2=n-1)
```

```
[1] 0.1904
```

```
sum(Q < 0.5) / k
```

```
[1] 0.1919
```

```
# P(0.5<Q<2)
pf(2,df1=n-1,df2=n-1) - pf(0.5,df1=n-1,df2=n-1)
```

```
[1] 0.6193
```

```
sum(Q > 0.5 & Q < 2) / k
```

```
[1] 0.6188
```

Chapter 3

Probability and simulation (solutions to exercises)

Contents

3.1 Concrete items

Exercise 3.1 Concrete items

A construction company receives concrete items for a construction. The length of the items are assumed reasonably normally distributed. The following requirements for the length of the elements are made

$$\mu = 3000 \text{ mm}.$$

The company samples 9 items from a delevery which are then measured for control. The following measurements (in mm) are found:

3003	3005	2997	3006	2999	2998	3007	3005	3001

a) Compute the following three statistics: the sample mean, the sample standard deviation and the standard error of the mean, and what are the interpretations of these statistics?

Solution

From the data we get the sample mean and sample standard deviation

$$\bar{x} = 3002.33\text{mm and } s = 3.708\text{mm}.$$

```
x <- c(3003, 3005, 2997, 3006, 2999, 2998, 3007, 3005, 3001)
mean(x)

[1] 3002


sd(x)

[1] 3.708
```

From Definition 3.7, we get the standard error of the mean as

$$SE_{\bar{x}} = \frac{3.708}{\sqrt{9}} = 1.236.$$

The interpretations of these are:

$\bar{x} = 3002.33$ The best estimate we can make of the true mean length of such concrete items

$s = 3.708$ The population of concrete item lenghts has a standard deviation estimated at 3.7. We estimate the average deviation from the mean for concrete items to be 3.7 mm. So most concrete items will be within the range of approximately ±7.4 mm of the mean

$SE_{\bar{x}} = 1.236$ All of the following are versions of the same story given by this number:

- The standard deviation of the sampling distribution of the sample mean (seen as a random variable) is (estimated at) 1.24
- And also: The standard deviation of the sampling distribution of the difference between sample mean and the population mean (seen as a random variable) is(estimated at) 1.24
- So from sample to sample (of size $n = 9$) the sample mean will be different. And the size of these differences, that is, the difference between the the sample mean and the true population mean is on average 1.24
- The sample mean is on average 1.24 away from the target: the population mean
- The error we will make on average in using the sample mean for estimating the population mean is 1.24

|||| Remark 3.1

Please, think about the difference between the story told by s and the story told by $SE_{\bar{x}}$ (both are estimated standard deviations, but for two VERY different concepts).

b) In a construction process, 5 concrete items are joined together to a single construction with a length which is then the complete length of the 5 concrete items. It is very important that the length of this new construction is within 15 m plus/minus 1 cm. How often will it happen that such a construction will be more than 1 cm away from the 15 m target (assume that the population mean concrete item length is $\mu = 3000$ mm and that the population standard deviation is $\sigma = 3$)?

|||| Solution

Let Y denote the length of the joined construction. So

$$Y = \sum_{i=1}^{5} X_i,$$

where X_i is the length of a randomly selected concrete item. So using the rules for mean and variance calculations from Section 2.7, we can find that

$$\mathrm{E}(Y) = \sum_{i=1}^{5} \mathrm{E}(X_i) = \sum_{i=1}^{5} 3000 = 5 \cdot 3000 = 15000,$$

and

$$\mathrm{V}(Y) = \sum_{i=1}^{5} \mathrm{V}(X_i) = \sum_{i=1}^{5} 3^2 = 5 \cdot 9 = 45.$$

We can also state that, since the concrete item distribution is a normal

$$X_i \sim N(\mu, \sigma^2),\ i = 1, \ldots, 5,$$

then the sum of five (independent) of such will be normal (Theorem 2.40), so

$$Y \sim N(15000, 45).$$

(Actually, the normality result is expressed in Theorem 3.3 for the sample mean, however, the sum is just a simple scaling of the mean, so then the normality also holds for the sum) So we can now find the answer to the question

$$P(|Y - 15000| > 10) = 2 \cdot P(\frac{(Y - 15000)}{\sqrt{45}} > 10/\sqrt{45}) = 2 \cdot P(Z > 1.4907) = 0.136.$$

```
2*(1-pnorm(15010, mean=15000, sd=sqrt(45)))

[1] 0.136


2*pnorm(-1.4907)

[1] 0.136
```

In between 13 and 14 cases out of 100 the joined construction is beyond 1 cm away from the target – maybe a new supplier should be considered!

c) Find the 95% confidence interval for the mean μ.

|||| Solution

Since the 97.5%-quantile, $t_{0.975}$ of the t-distribution with 8 degrees of freedom equals $t_{0.975} = 2.306$ (found in R as: `qt(0.975, 8)`), we get

$$3002.33 \pm 2.306 \cdot \frac{3.708}{\sqrt{9}} \Leftrightarrow [2999.5; 3005.2] \, .$$

Or everything in R:

```
x <- c(3003, 3005, 2997, 3006, 2999, 2998, 3007, 3005, 3001)
mean(x)

[1] 3002


sd(x)

[1] 3.708


qt(0.975, 8)

[1] 2.306


t.test(x)


One Sample t-test

data:  x
t = 2429, df = 8, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2999 3005
sample estimates:
mean of x
     3002
```

d) Find the 99% confidence interval for μ. Compare with the 95% one from above and explain why it is smaller/larger!

|||| Solution

Since the 99.5%-quantile, $t_{0.995}$ of the t-distribution with 8 degrees of freedom equals $t_{0.995} = 3.355$ (found in R as: `qt(0.995, 8)`), we get

$$3002.33 \pm 3.335 \cdot \frac{3.708}{\sqrt{9}} \Leftrightarrow [2998.2; 3006.5] \, .$$

Or everything in R:

```
qt(0.995, 8)

[1] 3.355


t.test(x, conf.level=0.99)


One Sample t-test

data:  x
t = 2429, df = 8, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 2998 3006
sample estimates:
mean of x
     3002
```

It makes good sense that the 99% confidence interval becomes larger than the 95% one, as the consequence of wanting to be more confident on capturing the true mean μ will make us having to state a larger interval.

e) Find the 95% confidence intervals for the variance σ^2 and the standard deviation σ.

Solution

We use the formula for the variance confidence interval

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}};\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right].$$

where the quantiles come from a χ^2-distribution with $\nu = n-1 = 8$ degrees of freedom

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}};\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right] = \left[\frac{8\cdot 3.708^2}{\chi^2_{0.975}};\ \frac{8\cdot 3.708^2}{\chi^2_{0.025}}\right] = \left[\frac{8\cdot 13.75}{17.535};\ \frac{8\cdot 13.75}{2.180}\right]$$

$$= [6.273;\ 50.465]\,.$$

And for the standard deviation:

$$\left[\sqrt{6.273};\ \sqrt{50.465}\right] = [2.50;\ 7.10]\,.$$

In R:

```
qchisq(c(0.975,0.025), 8)

[1] 17.53  2.18


c(8*13.75/qchisq(0.975, 8), 8*13.75/qchisq(0.025, 8))

[1]  6.273 50.465


sqrt(c(8*13.75/qchisq(0.975, 8), 8*13.75/qchisq(0.025, 8)))

[1] 2.505 7.104
```

f) Find the 99% confidence intervals for the variance σ^2 and the standard deviation σ.

Solution

$$\left[\frac{8\cdot 13.75}{21.955};\ \frac{8\cdot 13.75}{1.344}\right] = [5.010;\ 81.820]\,.$$

And for the standard deviation:

$$\left[\sqrt{5.010};\ \sqrt{81.820}\right] = [2.24;\ 9.05]\,.$$

In R:

```
qchisq(c(0.995,0.005), 8)

[1] 21.955  1.344

c(8*13.75/qchisq(0.995, 8), 8*13.75/qchisq(0.005, 8))

[1]  5.01 81.82

sqrt(c(8*13.75/qchisq(0.995, 8), 8*13.75/qchisq(0.005, 8)))

[1] 2.238 9.045
```

3.2 Aluminum profile

|||| Exercise 3.2 Aluminum profile

The length of an aluminum profile is checked by taking a sample of 16 items whose length is measured. The measurement results from this sample are listed below, all measurements are in mm:

180.02	180.00	180.01	179.97	179.92	180.05	179.94	180.10
180.24	180.12	180.13	180.22	179.96	180.10	179.96	180.06

From data is obtained: $\bar{x} = 180.05$ and $s = 0.0959$.

It can be assumed that the sample comes from a population which is normal distributed.

a) A 90%-confidence interval for μ becomes?

|||| Solution

Since the 95%-quantile, $t_{0.95}$ of the t-distribution with 15 degrees of freedom equals $t_{0.95} = 1.753$ (found in R as: `qt(0.95, 15)`), we get

$$180.05 \pm 1.753 \cdot \frac{0.0959}{\sqrt{16}} = [180.00, 180.10]\,.$$

Or everything in R:

```
x <- c(180.02, 180.00, 180.01, 179.97, 179.92, 180.05, 179.94, 180.10,
180.24, 180.12, 180.13, 180.22, 179.96, 180.10, 179.96, 180.06)
mean(x)

[1] 180.1


sd(x)

[1] 0.09592


qt(0.95, 15)

[1] 1.753


t.test(x, conf.level=0.9)


One Sample t-test

data:  x
t = 7509, df = 15, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 180.0 180.1
sample estimates:
mean of x
    180.1
```

b) A 99%-confidence interval for σ becomes?

|||| Solution

We use the formula for the variance confidence interval

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}};\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right],$$

where the quantiles come from a χ^2-distribution with $\nu = n-1 = 15$ degrees of freedom

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right] = \left[\frac{15\cdot 0.0092}{\chi^2_{0.995}}; \frac{15\cdot 0.0092}{\chi^2_{0.005}}\right] = \left[\frac{15\cdot 0.0092}{32.801}; \frac{15\cdot 0.0092}{4.601}\right]$$
$$= [0.00421;\ 0.02999]\,.$$

And for the standard deviation:

$$\left[\sqrt{0.00421};\ \sqrt{0.02999}\right] = [0.065;\ 0.173]\,.$$

In R:

```
qchisq(p=c(0.995,0.005), df=15)

[1] 32.801  4.601


c(15*0.0092/qchisq(p=0.995,df=15), 15*0.0092/qchisq(p=0.005,df=15))

[1] 0.004207 0.029994


sqrt(c(15*0.0092/qchisq(p=0.995,df=15), 15*0.0092/qchisq(p=0.005,df=15)))

[1] 0.06486 0.17319
```

3.3 Concrete items (hypothesis testing)

|||| Exercise 3.3 Concrete items (hypothesis testing)

This is a continuation of Exercise 1, so the same setting and data is used (read the initial text of it).

a) To investigate whether the requirement to the mean is fulfilled (with $\alpha = 5\%$), the following hypothesis should be tested

$$H_0 : \mu = 3000$$
$$H_1 : \mu \neq 3000.$$

Or similarly asked: what is the evidence against the null hypothesis?

|||| Solution

This is a one-sample situation. In R it could be handled by:

```
x <- c(3003,3005,2997,3006,2999,2998,3007,3005,3001)
t.test(x, mu=3000)


One Sample t-test

data:  x
t = 1.9, df = 8, p-value = 0.1
alternative hypothesis: true mean is not equal to 3000
95 percent confidence interval:
 2999 3005
sample estimates:
mean of x
     3002
```

from which the answer can be seen. One could also explixitly do it as

$$t_{\text{obs}} = \frac{3002.333 - 3000}{3.708/\sqrt{9}} = 1.885.$$

And then find the p-value as (using a t-distribution with $\nu = 8$ degrees of freedom)

$$2 * P(T > 1.885) = 0.096.$$

(in R as: `2*(1-pt(1.885,8))`). So although there is a weak evidence aginst the null, cf. the p-value interpretation table in Section 3.1, when using an α of 0.05 the null hypothesis is not rejected, but must be accepted.

b) What would the level $\alpha = 0.01$ critical values be for this test, and what are the interpretation of these?

|||| Solution

The critical values would be $\pm t_{0.995} = \pm 3.355$:

```
qt(p=0.995, df=8)

[1] 3.355
```

This means that, in a new experiment, the standardized difference between the data and the null hypothesis, also called t_{obs}, must be either larger than 3.355 or smaller than -3.355 to lead to a significant result of the experiment.

c) What would the level $\alpha = 0.05$ critical values be for this test (compare also with the values found in the previous question)?

|||| Solution

The critical values would be $\pm t_{0.975} = \pm 2.306$:

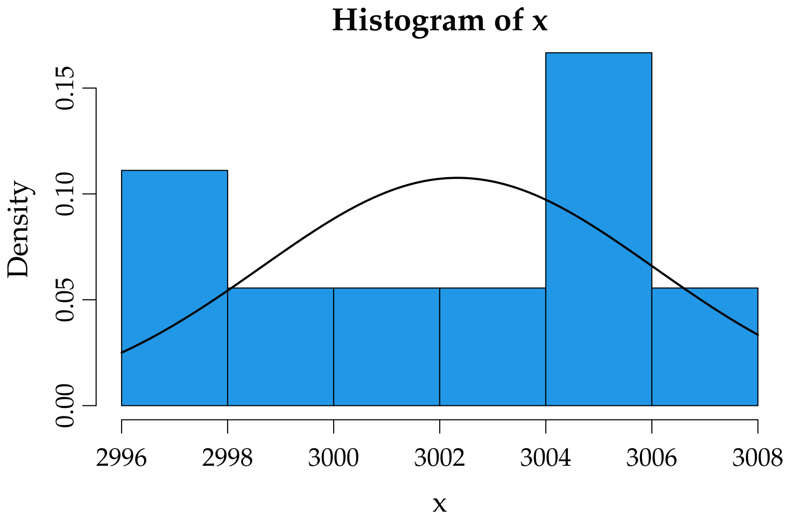
```
qt(p=0.975, df=8)

[1] 2.306
```

This means that, again in a new experiment, it is easier to detect an effect with significance level $\alpha = 0.05$ than on level $\alpha = 0.01$.

d) Investigate, by som plots, whether the data here appears to be coming from a normal distribution (as assumed until now)?

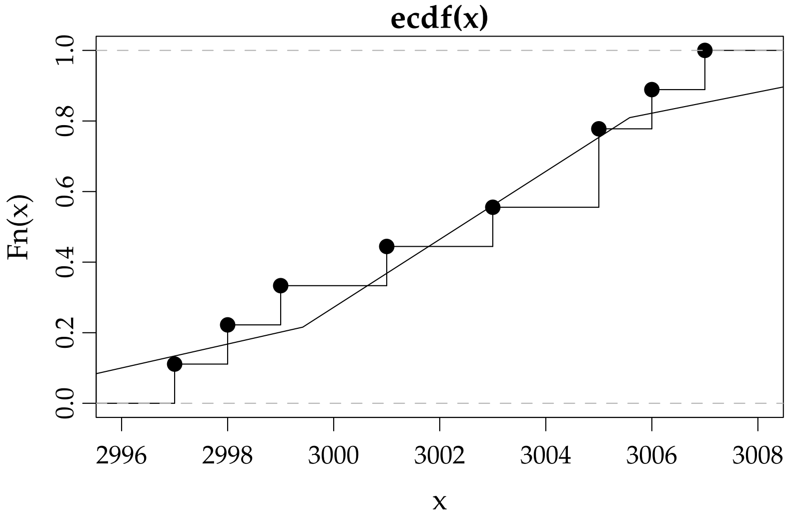
|||| Solution

```
x=c(3003,3005,2997,3006,2999,2998,3007,3005,3001)
hist(x, freq=F, col = 4)
xp <- seq(2996, 3008, 0.1)
lines(xp, dnorm(xp, mean(x), sd(x)), lwd = 2)
```



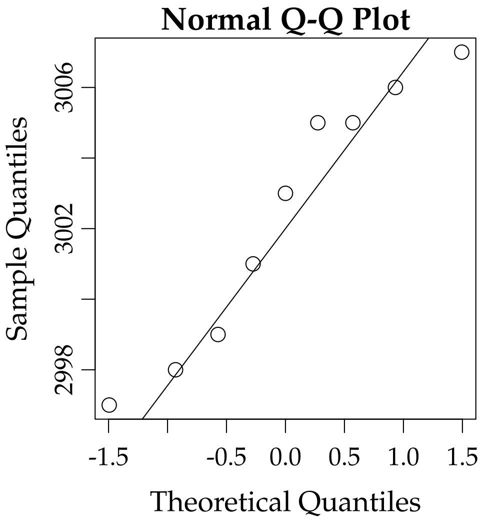
|||| Solution

```
plot(ecdf(x), verticals = TRUE)
xp <- seq(0.9*min(x), 1.1*max(x), length.out = 100)
lines(xp, pnorm(xp, mean(x), sd(x)))
```



|||| Solution

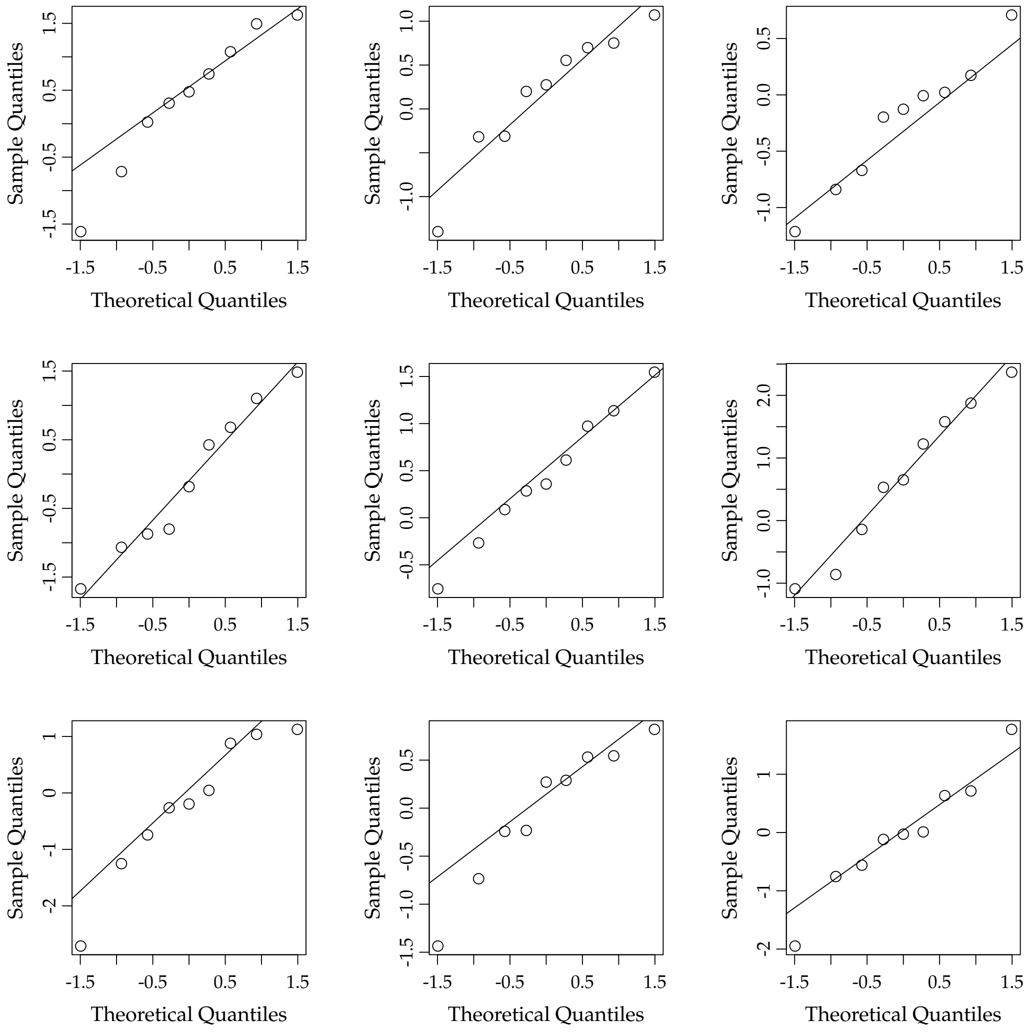
```
qqnorm(x)
qqline(x)
```



|||| Solution

Compare with 9 simulated ones

```r
par(mfrow = c(3, 3))
for (i in 1:9){
    xr <- rnorm(9)
    qqnorm(xr, main="")
    qqline(xr)
}
```



|||| Solution

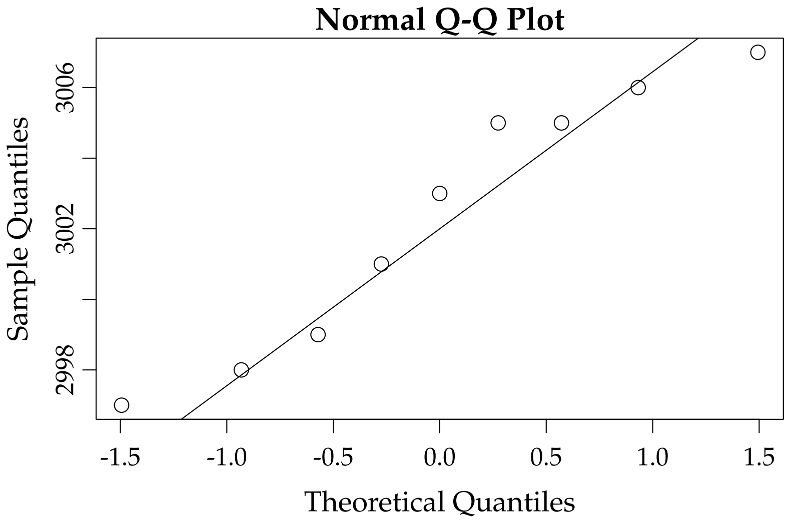
The nine data points do not differ more from the line than what truly normally distributed samples of size $n = 9$ do, so we cannot falsify the normality assumption (Did we prove normality?...no, we accept that they are normal distributed (like when we accept the null hypothesis)).

e) Assuming that you, maybe among different plots, also did the normal q-q plot above, the question is now: What exactly is plotted in that plot? Or more specifically: what are the x- and y-coordinates of e.g. the two points to the lower left in this plot?

|||| Solution

Let us look at the Normal q-q plot again:

```
qqnorm(x)
qqline(x)
```



The y-coordinates of the nine points from left to right in the plot are the ordered observations $x_{(1)}, \ldots, x_{(9)}$:

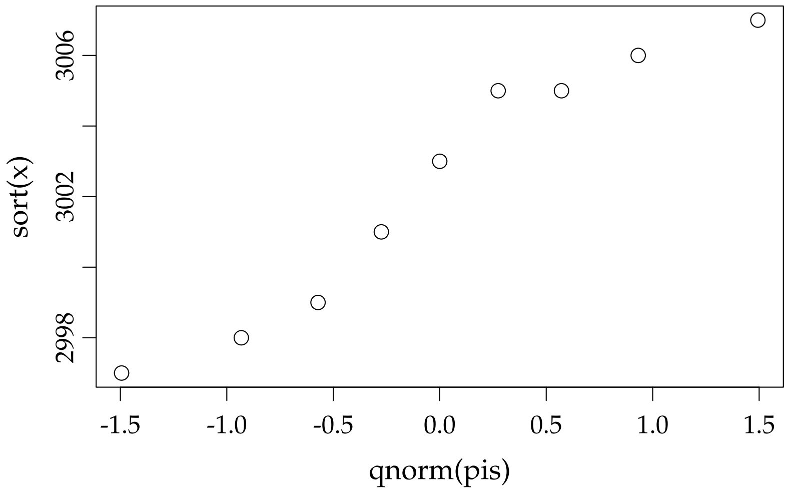
```
sort(x)

[1] 2997 2998 2999 3001 3003 3005 3005 3006 3007
```

|||| Solution

The x-coordinates are quantiles from the standard normal distribution. Method 3.42 tell us exactly which quantiles, that are used by R (in the current case there is less than $n = 10$ observations):

$$p_i = \frac{i - 3/8}{9 + 1/4},\ i = 1, \ldots, 9$$

```
is <- 1:9
pis <- (is-3/8)/(9+1/4)
qnorm(pis)

[1] -1.4942 -0.9320 -0.5716 -0.2744  0.0000  0.2744  0.5716  0.9320
[9]  1.4942


plot(qnorm(pis), sort(x))
```



3.4 Aluminum profile (hypothesis testing)

Exercise 3.4 Aluminium profile (hypothesis testing)

We use the same setting and data as in Exercise 2, so read the initial text of it.

a) Find the evidence against the following hypothesis:

$$H_0 : \mu = 180.$$

Solution

First we compute the observed t-statistic

$$t_{\text{obs}} = \frac{180.05 - 180}{0.0959/\sqrt{16}} = 2.085,$$

and the p-value is

$$p\text{-value} = 2 \cdot P(T > 2.085) = 2 \cdot 0.0273 = 0.055.$$

```
2*(1-pt(2.085, 15))

[1] 0.05457
```

Or in completely solve by R:

```
x <- c(180.02, 180.00, 180.01, 179.97, 179.92, 180.05, 179.94, 180.10,
180.24, 180.12, 180.13, 180.22, 179.96, 180.10, 179.96, 180.06)
t.test(x, mu=180)


One Sample t-test

data:  x
t = 2.1, df = 15, p-value = 0.05
alternative hypothesis: true mean is not equal to 180
95 percent confidence interval:
 180.0 180.1
sample estimates:
mean of x
    180.1
```

Hence there is *weak evidence against* H_0, cf. Table 3.1.

b) If the following hypothesis test is carried out

$$H_0 : \mu = 180,$$
$$H_1 : \mu \neq 180.$$

What are the level $\alpha = 1\%$ critical values for this test?

|||| Solution

The critical values are the 0.005 and 0.995 quantiles of the t-distribution with $\nu = n - 1 = 15$ degrees of freedom, $\pm t_{0.995} = \pm 2.947$:

```
qt(p=0.995, df=15)

[1] 2.947
```

c) What is the 99%-confidence interval for μ?

|||| Solution

The formula gives

$$180.05 \pm t_{0.995} \frac{0.0959}{\sqrt{16}} = [179.98;\ 180.12].$$

```
t.test(x, mu=180, conf.level=0.99)


One Sample t-test

data:  x
t = 2.1, df = 15, p-value = 0.05
alternative hypothesis: true mean is not equal to 180
99 percent confidence interval:
 180.0 180.1
sample estimates:
mean of x
    180.1
```

d) Carry out the following hypothesis test

$$H_0 : \mu = 180,$$
$$H_1 : \mu \neq 180,$$

using $\alpha = 5\%$.

Solution

We allready found the p-value $= 0.055$ above, and as this is larger than α we cannot reject the null hypothesis of $\mu = 180$.

3.5 Transport times

|||| Exercise 3.5 Transport times

A company, MM, selling items online wants to compare the transport times for two transport firms for delivery of the goods. To compare the two companies recordings of delivery times on a specific route were made, with a sample size of n = 9 for each firm. The following data were found:

$$\text{Firm A: } \bar{y}_A = 1.93 \text{ d and } s_A = 0.45 \text{ d,}$$
$$\text{Firm B: } \bar{y}_B = 1.49 \text{ d and } s_B = 0.58 \text{ d.}$$

note that d is the SI unit for days. It is assumed that data can be regarded as stemming from normal distributions.

a) We want to test the following hypothesis

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B$$

What is the p-value, interpretation and conclusion for this test (at $\alpha = 5\%$ level)?

|||| Solution

This is the independent samples (non-directional) t-test (Welch) in Method 3.49 (or Method 3.51). We first find the t-test statistic

$$t_{\text{obs}} = \frac{1.93 - 1.49}{\sqrt{0.45^2/9 + 0.58^2/9}} = 1.8,$$

and then the degrees of freedom

$$\nu = \frac{\left(\frac{0.45^2}{9} + \frac{0.58^2}{9}\right)^2}{\frac{(0.45^2/9)^2}{8} + \frac{(0.58^2/9)^2}{8}} = 15.0,$$

and the p-value is then found to

$$2 \cdot P(T > 1.8) = 0.0922,$$

using a t-distribution with 15.0 degrees of freedom. So even though, according to Table 3.1, there is weak evidence against the null hypothesis, when we use $\alpha = 0.05$ we cannot reject the null hypothesis of the two firms being equally fast. In R it can be found as

```
ms <- c(1.93, 1.49)
vs <- c(0.45^2, 0.58^2)
ns <- c(9,9)
t_obs <- (ms[1]-ms[2])/sqrt(vs[1]/ns[1]+vs[2]/ns[2])
nu <- ((vs[1]/ns[1]+vs[2]/ns[2])^2)/
((vs[1]/ns[1])^2/(ns[1]-1)+(vs[2]/ns[2])^2/(ns[2]-1))
t_obs

[1] 1.798


nu

[1] 15.07


2*(1-pt(t_obs,nu))

[1] 0.09222
```

b) Find the 95% confidence interval for the mean difference $\mu_A - \mu_B$.

|||| Solution

We need the degrees of freedom - we found that number above: $\nu = 15.0$. Since the relevant t-quantile then is, using $\nu = 15.0$,

```
qt(p=0.975, df=15.0)

[1] 2.131
```

$$t_{0.975} = 2.131,$$

the confidence interval becomes

$$1.93 - 1.49 \pm 2.131\sqrt{0.45^2/9 + 0.58^2/9}.$$

Which becomes

```
1.93-1.49 +c(-1,1)*qt(0.975,15.0)*sqrt(0.45^2/9+0.58^2/9)

[1] -0.08156  0.96156
```

Hence the answer is: $[-0.082;\ 0.962]$.

c) What is the power of a study with $n = 9$ observations in each of the two samples of detecting a potential mean difference of 0.4 between the firms (assume that $\sigma = 0.5$ and that we use $\alpha = 0.05$)?

Solution

```
power.t.test(n = 9, delta = 0.4, sd = 0.5, sig.level = 0.05)


     Two-sample t test power calculation

              n = 9
          delta = 0.4
             sd = 0.5
      sig.level = 0.05
          power = 0.3578
    alternative = two.sided

NOTE: n is number in *each* group
```

So the power is only 0.36 - not nearly good enough for a reasonable study.

d) What effect size (mean difference) could be detected with $n = 9$ observations in each of the two samples with a power of 0.8 (assume that $\sigma = 0.5$ and that we use $\alpha = 0.05$)?

Solution

```
power.t.test(n = 9, power = 0.8, sd = 0.5, sig.level = 0.05)


     Two-sample t test power calculation

              n = 9
          delta = 0.7035
             sd = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

So a potential mean difference of 0.70 is detectable with probability 0.8 by such a study.

e) How large a sample size (from each firm) would be needed in a new investigation, if we want to detect a potential mean difference of 0.4 between the firms with probability 0.90, that is with `power=0.90` (assume that $\sigma = 0.5$ and that we use $\alpha = 0.05$)?

|||| Solution

```
power.t.test(power = 0.90, delta = 0.4, sd = 0.5, sig.level = 0.05)


     Two-sample t test power calculation

              n = 33.83
          delta = 0.4
             sd = 0.5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

So $n = 34$ in each sample would do the job!

3.6 Cholesterol

|||| Exercise 3.6 Cholesterol

In a clinical trial of a cholesterol-lowering agent, 15 patients' cholesterol (in mmol/L) has been measured before treatment and 3 weeks after starting treatment. Data are listed in the following table:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4

The following is run in R:

```
x1 <- c(9.1, 8.0, 7.7, 10.0, 9.6, 7.9, 9.0, 7.1,
        8.3, 9.6, 8.2, 9.2, 7.3, 8.5, 9.5)
x2 <- c(8.2, 6.4, 6.6, 8.5, 8.0, 5.8, 7.8, 7.2,
        6.7, 9.8, 7.1, 7.7, 6.0, 6.6, 8.4)
t.test(x1, x2)


Welch Two Sample t-test

data:  x1 and x2
t = 3.3, df = 27, p-value = 0.003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4637 1.9630
sample estimates:
mean of x mean of y
    8.600     7.387


t.test(x1, x2, pair=TRUE)


Paired t-test

data:  x1 and x2
t = 7.3, df = 14, p-value = 0.000004
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8588 1.5678
```

```
sample estimates:
mean of the differences
                  1.213
```

a) Can there, based on these data be demonstrated a significant decrease in cholesterol levels with $\alpha = 0.001$?

|||| Solution

This is clearly a *paired* setting so only the results from the last of the R-calls are relevant, where we can read off the results:

The (non-directional) p-value is 0.00000367, so there is very strong evidence against the null hypothesis, and we can beyond any reasonable doubts conclude that the mean cholesterol level has decreased after the 3 weeks.

3.7 Pulse

Exercise 3.7 Pulse

13 runners had their pulse measured at the end of a workout and 1 minute after again and we got the following pulse measurements:

Runner	1	2	3	4	5	6	7	8	9	10	11	12	13
Pulse end	173	175	174	183	181	180	170	182	188	178	181	183	185
Pulse 1min	120	115	122	123	125	140	108	133	134	121	130	126	128

The following was run in R:

```
Pulse_end <- c(173,175,174,183,181,180,170,182,188,
               178,181,183,185)
Pulse_1min <- c(120,115,122,123,125,140,108,133,134,
                121,130,126,128)
mean(Pulse_end)

[1] 179.5


mean(Pulse_1min)

[1] 125


sd(Pulse_end)

[1] 5.19


sd(Pulse_1min)

[1] 8.406


sd(Pulse_end-Pulse_1min)

[1] 5.768
```

a) What is the 99% confidence interval for the mean pulse drop (meaning the drop during 1 minute from end of workout)?

|||| Solution

We use the paired sample t-test version of the confidence interval (see Section 3.2.3) with $s_{\text{dif}} = 5.768$, $n = 13$ and 12 degrees of freedom for the t-quantile $t_{0.005}$:

```
qt(p=0.995, df=12)

[1] 3.055
```

Or in completely in R:

```
Pulse_end <- c(173,175,174,183,181,180,170,182,188,178,181,183,185)
Pulse_1min <- c(120,115,122,123,125,140,108,133,134,121,130,126,128)

t.test(Pulse_end, Pulse_1min, paired=TRUE, conf.level = 0.99)


Paired t-test

data:  Pulse_end and Pulse_1min
t = 34, df = 12, p-value = 3e-13
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 49.58 59.35
sample estimates:
mean of the differences
                  54.46
```

So, the answer is:

$$54.46 \pm 3.054 \cdot \frac{5.768}{\sqrt{13}} = [49.58; 59.35].$$

b) Consider now the 13 pulse end measurements (first row in the table). What is the 95% confidence interval for the standard deviation of these?

Solution

Using 3.19 we find the 95% confidence interval for the variance to

$$\frac{12 \cdot 5.18998^2}{\chi^2_{0.975}} < \sigma^2 < \frac{12 \cdot 5.18998^2}{\chi^2_{0.025}},$$

which then for the standard deviation becomes

$$\sqrt{\frac{12 \cdot 5.18998^2}{23.34}} < \sigma < \sqrt{\frac{12 \cdot 5.18998^2}{4.40}}.$$

Or in R:

```
t.test(Pulse_end, Pulse_1min, paired=TRUE, conf.level = 0.95)


Paired t-test

data:  Pulse_end and Pulse_1min
t = 34, df = 12, p-value = 3e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 50.98 57.95
sample estimates:
mean of the differences
                  54.46


sqrt(12*var(Pulse_end)/qchisq(0.975,12))

[1] 3.722


sqrt(12*var(Pulse_end)/qchisq(0.025,12))

[1] 8.567
```

So, the answer is that we accept that $\sigma \in [3.72; 8.57]$ or we could write $3.72 < \sigma < 8.57$.

3.8 Foil production

|||| Exercise 3.8 Foil production

In the production of a certain foil (film), the foil is controlled by measuring the thickness of the foil in a number of points distributed over the width of the foil. The production is considered stable if the mean of the difference between the maximum and minimum measurements does not exceed 0.35 mm. At a given day, the following random samples are observed for 10 foils:

Foil	1	2	3	4	5	6	7	8	9	10
Max. in mm (y_{max})	2.62	2.71	2.18	2.25	2.72	2.34	2.63	1.86	2.84	2.93
Min. in mm (y_{min})	2.14	2.39	1.86	1.92	2.33	2.00	2.25	1.50	2.27	2.37
Max-Min (D)	0.48	0.32	0.32	0.33	0.39	0.34	0.38	0.36	0.57	0.56

The following statistics may potentially be used

$$\bar{y}_{\text{max}} = 2.508,\ \bar{y}_{\text{min}} = 2.103,\ s_{y_{\text{max}}} = 0.3373,\ s_{y_{\text{min}}} = 0.2834,\ s_D = 0.09664.$$

a) What is a 95% confidence interval for the mean difference?

|||| Solution

The 95% confidence interval is given by:

```
(2.508 - 2.103) + c(-1, 1) * qt(0.975, df=10-1) * 0.09664 / sqrt(10)

[1] 0.3359 0.4741
```

The confidence interval contains those values of the mean difference that we believe in based on the data. Notice that the CI contains $\mu_D = 0.35$, hence we know that we will not reject the null hypothesis $H_0 = 0.35$, which is the next question.

b) How much evidence is there that the mean difference is different from 0.35? State the null hypothesis, t-statistic and p-value for this question.

Solution

The t-statistic is found using Method 9:

$$t_{\text{obs}} = \frac{(2.508 - 2.103) - 0.35}{0.09664/\sqrt{10}} = 1.80.$$

```
(tobs <- ((2.508 - 2.103) - 0.35) / (0.09664 / sqrt(10)))

[1] 1.8
```

The p-value for this assessment is

$$p\text{-value} = 2 \cdot P(T > t_{\text{obs}}) = 0.1054.$$

where T has a t-distribution with 9 degrees of freedom:

```
2 * pt(abs(tobs), df=10-1, lower.tail=FALSE)

[1] 0.1054
```

According to Table 3.1 there is little or no evidence that against the null hypothesis that $\mu_D = 0.35$.

3.9 Course project

Exercise 3.9 Course project

At a specific education it was decided to introduce a project, running through the course period, as a part of the grade point evaluation. In order to assess whether it has changed the percentage of students passing the course, the following data was collected:

	Before introduction of project	After introduction of project
Number of students evaluated	50	24
Number of students failed	13	3
Average grade point $\bar{x}$	6.420	7.375
Sample standard deviation s	2.205	1.813

a) As it is assumed that the grades are approximately normally distributed in each group, the following hypothesis is tested:

$$H_0 : \mu_{\text{Before}} = \mu_{\text{After}},$$
$$H_1 : \mu_{\text{Before}} \neq \mu_{\text{After}}.$$

The test statistic, the p-value and the conclusion for this test become?

Solution

The (Welch) t-test statistic for this setup is found using Method :

$$t = \frac{6.42 - 7.375}{\sqrt{2.205^2/50 + 1.813^2/24}} = -1.97.$$

From the t-distribution with $\nu = 54.4$ we can find that the (non-directional) p-value is

$$2 \cdot P(T > 1.97) = 0.054.$$

```
2*(1-pt(1.97, 54.4))

[1] 0.05394
```

In R we could do it by the following:

```
ms <- c(6.42, 7.375)
vs <- c(2.205^2, 1.813^2)
ns <- c(50, 24)
t_obs <- (ms[1]-ms[2])/sqrt(vs[1]/ns[1]+vs[2]/ns[2])
nu <- ((vs[1]/ns[1]+vs[2]/ns[2])^2)/
((vs[1]/ns[1])^2/(ns[1]-1)+(vs[2]/ns[2])^2/(ns[2]-1))
t_obs

[1] -1.973


nu

[1] 54.39
```

On a 5% level we cannot conclude a significant difference in the grade point means before and after.

b) A 99% confidence interval for the mean grade point difference is?

|||| Solution

We need the degrees of freedom - we found that number above: $\nu = 54.4$. Since the relevant t-quantile then is, using $\nu = 54.4$,

```
qt(p=0.995, df=54.4)

[1] 2.669
```

Hence $t_{0.995} = 2.669$, and the confidence interval becomes

$$6.42 - 7.375 \pm 2.669\sqrt{2.205^2/50 + 1.813^2/24},$$

Which is:

```
6.42-7.375+c(-1,1)*qt(0.995,54.4)*sqrt(2.205^2/50 + 1.813^2/24)

[1] -2.2468  0.3368
```

and the answer is: we accept that the mean difference in the interval [-2.247; 0.337].

c) A 95% confidence interval for the grade point standard deviation after the introduction of the project becomes?

Solution

The confidence interval formula for a sample variance is used WITH the square-root applied to everything:

$$\frac{\sqrt{(n-1)\cdot s^2}}{\sqrt{\chi^2_{0.975}}} < \sigma < \frac{\sqrt{(n-1)\cdot s^2}}{\sqrt{\chi^2_{0.025}}},$$

so

$$\frac{\sqrt{23\cdot 1.813^2}}{\sqrt{38.076}} < \sigma_{\text{After}} < \frac{\sqrt{23\cdot 1.813^2}}{\sqrt{11.689}}.$$

The values found in R:

```
sqrt(23*1.813^2)/sqrt(qchisq(0.975,23))

[1] 1.409


sqrt(23*1.813^2)/sqrt(qchisq(0.025,23))

[1] 2.543
```

So the answer is, that we accept that $\sigma_{\text{After}} \in [1.41;\ 2.54]$.

3.10 Concrete items (sample size)

|||| Exercise 3.10 Concrete items (sample size)

This is a continuation of Exercise 1, so the same setting and data is used (read the initial text of it).

a) A study is planned of a new supplier. It is expected that the standard deviation will be approximately 3, that is, $\sigma = 3$ mm. We want a 90% confidence interval for the mean value in this new study to have a width of 2 mm. How many items should be sampled to achieve this?

|||| Solution

We use the sample size formula (Method 3.63) with wanted margin of error $ME = 1$ (as the width of the confidence interval is twice the margin of error)

$$n = \left(\frac{z_{0.95}\sigma}{ME}\right)^2 = \left(\frac{1.645 \cdot 3}{1}\right)^2 = 24.35.$$

(in R: `qnorm(0.95)` to get $z_{0.95} = 1.645$). Hence the answer becomes: at least 25.

b) Answer the sample size question above but requiring the 99% confidence interval to have the (same) width of 2 mm.

|||| Solution

The same formula as above:

$$n = \left(\frac{z_{0.995}\sigma}{ME}\right)^2 = \left(\frac{2.576 \cdot 3}{1}\right)^2 = 59.72.$$

(using R: `qnorm(0.995)` to get $z_{0.995} = 2.576$). Hence the answer becomes: at least 60.

c) (Warning: This is a difficult question about a challenging abstraction - do not worry, if you do not make this one) For the two sample sizes found in the two previous questions find the probability that the corresponding confidence interval in the future study will actually be more than 10% wider than planned for (still assuming and using that the population variance is $\sigma^2 = 9$).

|||| Solution

The random width of the confidence interval is due to the randomly changing sample variance in the formula for the (random) half width of the interval

$$\text{'The half width of CI'} = t_{1-\alpha/2}\frac{S}{\sqrt{n}}.$$

The (sampling) distribution of the variance estimator, S, is a χ^2-distribution, as stated in Section 3.1.6 (around Equation (3-17)). Thus: let S^2 be the variance of a sample of size n from a normal distribution with variance $\sigma^2 = 3^2 = 9$. Then

$$\chi^2 = \frac{(n-1)\cdot S^2}{9},$$

is a stochastic variable following the χ^2-distribution with $v = n - 1$ degrees of freedom.

|||| Solution

So, as the wanted margin error was $ME = 1$, we are asked to first find with $\alpha = 0.10$ and $n = 25$, and hence $t_{0.95} = 1.711$ (in R `qt(0.95,24)`)

$$\begin{aligned}
P\left(\text{'Half width of CI'} > 1.1\right) &= P\left(t_{0.95}\cdot\frac{S}{\sqrt{n}} > 1.1\right)\\
&= P\left(t_{0.95}^2\cdot\frac{S^2}{n} > 1.1^2\right)\\
&= P\left(t_{0.95}^2\cdot\frac{9\cdot\chi^2}{24\cdot 25} > 1.1^2\right)\\
&= P\left(\chi^2 > \frac{1.1^2\cdot 24\cdot 25}{t_{0.95}^2\cdot 9}\right)\\
&= P\left(\chi^2 > 27.56\right)\\
&= 0.28.
\end{aligned}$$

```
qt(0.95, 24)

[1] 1.711


1.1^2*25*24/(qt(0.95, 24)^2*9)

[1] 27.56


1-pchisq(1.1^2*25*24/(qt(0.95, 24)^2*9), 24)

[1] 0.2791
```

Therefore in almost 30% of cases an experiment planned for a 90%$ME = 1$ would actually end up wih a confidence interval of half width more than 1.1.

For the 99% case, and $n = 60$ the same computation gives:

```
qt(p=0.995, df=59)

[1] 2.662


1.1^2*60*59/(qt(p=0.995, df=59)^2*9)

[1] 67.18


1 - pchisq(q=1.1^2*60*59/(qt(0.995, 59)^2*9), df=59)

[1] 0.2174
```

So in this case it only happens in 22% of the cases. In the next part of this topic we will learn how we can plan experiments such that we are more in control of the risk of the experiments not really meeting our needs.

d) Now a new experiment is to be planned. In the first part above, given some wanted margin of error (ME) a sample size of $n = 25$ was found. What are each of the probabilities that an experiment with $n = 25$ will detect effects corresponding to ("end up significant for") $\mu_1 = 3001, 3002, 3003$ respectively? Assume that we use the typical $\alpha = 0.05$ level and that $\sigma = 3$?

Solution

We can only solve this by the in-built `power.t.test` function in R. If we specify everything but the power - it will compute the power for us. With e.g. $\mu_1 = 3001$, we have that $\mu_0 - \mu_1 = -1$, so in the R-function the `delta` should be set to either 1 or -1. And in fact one can insert a list of the three relevant deltas to get the three answers by a single call to R:

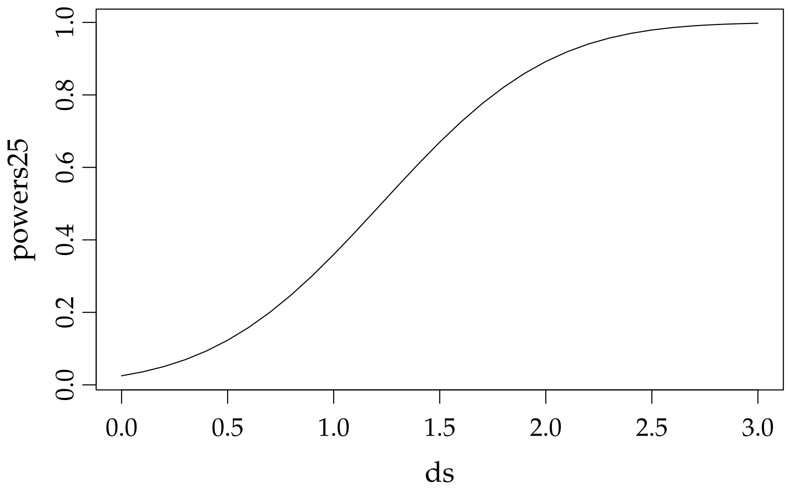
```
power.t.test(n = 25, delta = 1:3, sd = 3, sig.level = 0.05,
type = c("one.sample"))


     One-sample t test power calculation

              n = 25
          delta = 1, 2, 3
             sd = 3
      sig.level = 0.05
          power = 0.3595, 0.8920, 0.9977
    alternative = two.sided
```

So the three probabilities are 0.36, 0.89 and 0.998. A difference of 1 would not be reasonably detectable by this experiment but a difference of 2 has a high probility of being detected and even more so for 3.

Solution

A plot of all possible powers for all possible effect sizes could now easily be made:

```
ds = seq(0, 3, 0.1)
powers25 <- power.t.test(n = 25, delta = ds, sd = 3, sig.level = 0.05,
type = c("one.sample"))$power
plot(ds, powers25, type = "l")
```



e) One of the sample size computation above led to $n = 60$ (it is not so important how/why). Answer the same question as above using $n = 60$.

|||| Solution

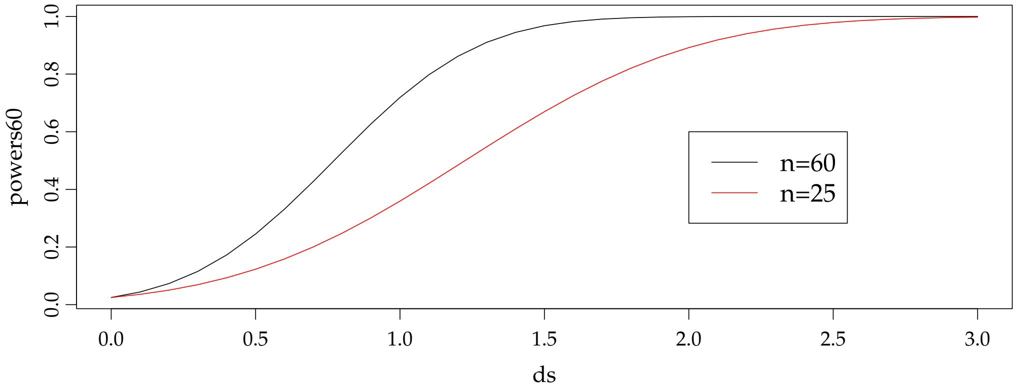
```
power.t.test(n = 60, delta = 1:3, sd = 3, sig.level = 0.05,
type = c("one.sample"))


     One-sample t test power calculation

              n = 60
          delta = 1, 2, 3
             sd = 3
      sig.level = 0.05
          power = 0.7190, 0.9991, 1.0000
    alternative = two.sided
```

So the three probabilities are 0.72, 0.999 and 1.00000. A difference of 1 would still not be reasonably detectable by this experiment but a difference of 2 and 3 has extremely high power.

|||| Solution

A plot of all possible powers for all possible effect sizes could now easily be made:

```
ds = seq(0, 3, 0.1)
powers60 <- power.t.test(n = 60, delta = ds, sd = 3, sig.level = 0.05,
type = c("one.sample"))$power
plot(ds, powers60, type = "l")
lines(ds, powers25, col="red")
legend(2,0.6,c("n=60", "n=25"), col=c("black", "red") , lty=c(1,1))
```



f) What sample size would be needed to achieve a power of 0.80 for an effect of size 0.5?

|||| Solution

The approximate sample size formula, Method 3.47, is only formally given for one-sided tests, so at first sight again we have no formulas for this (as we test two-sided here). However, in fact the formula also works nicely for the two-sided case, substituting $z_{1-\alpha}$ by $z_{1-\alpha/2}$, as the only error that is made would be the left hand tail of the rejection area, cf. the power plots on page 46, which essentially is zero for relevant effect sizes. So IF we use this adapted (and no-where stated) formula we would get:

$$n = \left(3 \cdot \frac{z_{0.8} + z_{0.975}}{0.5}\right)^2 = 282.6.$$

```
(3*(qnorm(p=0.8)+qnorm(p=0.975))/0.5)^2

[1] 282.6
```

However, it would actually be (slightly) better to simply use the R-function again as the R-function dos not rely on any normal approximation of the more correct *t*-distributions:

```
power.t.test(power = 0.8, delta = 0.5, sd = 3, sig.level = 0.05,
type = c("one.sample"))


     One-sample t test power calculation

              n = 284.5
          delta = 0.5
             sd = 3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

This would lead to $n = 285$. So 0.5 is an effect size that it would be pretty expensive to be able to detect.

g) Assume that you only have the finances to do an experiment with $n = 50$. How large a difference would you be able to detect with probability 0.8 (i.e. Power$=$ 0.80)?

Solution

```
power.t.test(n = 50, power = 0.8, sd = 3, sig.level = 0.05,
type = c("one.sample"))


     One-sample t test power calculation

              n = 50
          delta = 1.213
             sd = 3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

The answer is 1.21, so a true alternative mean of 2998.8 (or smaller) or 3001.2 (or larger) would be detected by this experiment with probability 0.80

Chapter 4

Statistics by Simulation (solutions to exercises)

Contents

4.1 Reliability: System lifetime (simulation as a computation tool)

Exercise 4.1 Reliability: System lifetime (simulation as a computation tool)

A system consists of three components A, B and C serially connected, such that A is positioned before B, which is again positioned before C. The system will be functioning only so long as A, B and C are all functioning. The lifetime in months of the three components are assumed to follow exponential distributions with means: 2 months, 3 months and 5 months, respectively (hence there are three random variables, X_A, X_B and X_C with exponential distributions with $\lambda_A = 1/2$, $\lambda_B = 1/3$ and $\lambda_C = 1/5$ resp.). A little R-help: You will probably need (or at least it would help) to put three variables together to make e.g. a $k \times 3$-matrix – this can be done by the `cbind` function:

```
x <- cbind(xA,xB,xC)
```

And just as an example, remember from the examples in the chapter that the way to easily compute e.g. the mean of the three values for each of all the k rows of this matrix is:

```
simmeans <- apply(x, 1, mean)
```

a) Generate, by simulation, a large number (at least 1000 – go for 10000 or 100000 if your computer is up for it) of system lifetimes (hint: consider how the random variable Y = System lifetime is a function of the three X-variables: is it the sum, the mean, the median, the minimum, the maximum, the range or something even different?).

Solution

Note that the lifetime can be seen as the minimal value of the three random component lifetimes:

$$\text{"Lifetime"} = \min(X_A, X_B, X_C).$$

First, note that the generated solution below has been generated with this seed in order to get the same result each time. Note, that when a simulation analysis is carried out, this number should only be set once and set randomly (potentially it is possible to find a seed (see Remark 2.12) that gives a rare simulation result and thus showing a "wrong" result, however if k is high enough this is very hard). The solution below has been generated with the following seed

```
## You might want to set the seed to achieve a particular result
set.seed(82719)
```

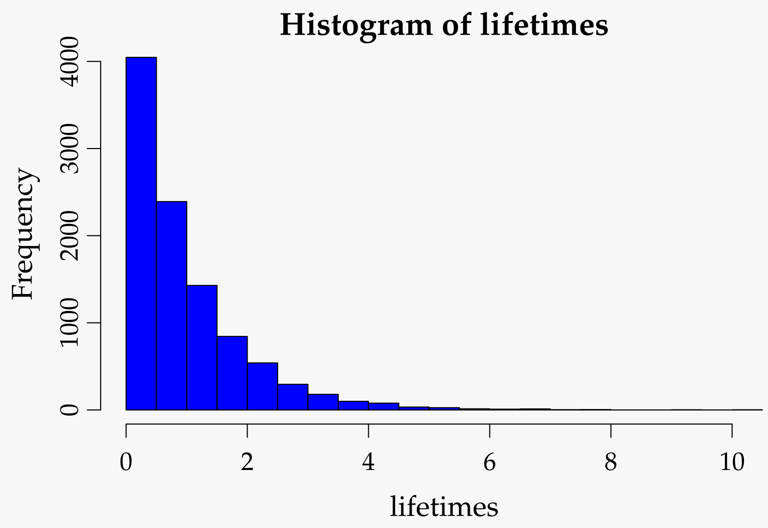
The following R-code generates 10.000 simulated system lifetimes:

```
## Number of simulations
k <- 10000
## Generating k component A lifetimes
xA <- rexp(k,1/2)
## Checking the mean of these
mean(xA)

[1] 2.018


## Generating k component B lifetimes
xB <- rexp(k,1/3)
## Checking the mean of these
mean(xB)

[1] 2.998


## generating k component C lifetimes
xC <- rexp(k,1/5)
## Checking the mean of these
mean(xC)

[1] 5.046


# Putting these three sets of k lifetimes together into a
# single k-by-3 matrix:
x <- cbind(xA,xB,xC)

# Finding the minimum value of the three components
# in each of the k situations:
lifetimes <- apply(x,1,min)
```

|||| Solution

Let us have a look at these simulated lifetimes:

```
## Histogram of the simulated lifetimes
hist(lifetimes, col = "blue", nclass = 30)
```



b) Estimate the mean system lifetime.

|||| Solution

```
## The estimated mean lifetime
mean(lifetimes)

[1] 0.974
```

c) Estimate the standard deviation of system lifetimes.

Solution

```
## The estimated std. dev. of the lifetime
sd(lifetimes)

[1] 0.9842
```

d) Estimate the probability that the system fails within 1 month.

Solution

We need to count how often the lifetimes are smaller than or equal to 1 month – this can in R be achieved by use of a logical operator:

```
## The fraction of times the simulated lifetime was below or equal 1
mean(lifetimes <= 1)

[1] 0.6437
```

In R `FALSE` is a 0 and a `TRUE` is a 1 - this is why we can simply apply the `mean` function directly on the vector of TRUES and FALSES like this.

e) Estimate the median system lifetime

Solution

```
## The estimated median lifetime
median(lifetimes)

[1] 0.6731
```

f) Estimate the 10th percentile of system lifetimes

Solution

```
## The estimated 10% quantile
quantile(lifetimes, 0.10)

   10%
0.1007
```

g) What seems to be the distribution of system lifetimes? (histogram etc)

Solution

We already made the histogram above. It appears that the minimum of the three exponential variables also has a distribution that looks like an exponential. In fact, there is a theoretical result (beoynd the syllabus of this course) that states that the distribution of the minimum of these three exponential distributions is again an exponential distribution but now with

$$\lambda_{min} = \lambda_A + \lambda_B + \lambda_C = 1/2 + 1/3 + 1/5 = 31/30.$$

Note how this matches nicely with the found mean above!

4.2 Basic bootstrap CI

|||| Exercise 4.2 Basic bootstrap CI

(Can be handled without using R) The following measurements were given for the cylindrical compressive strength (in MPa) for 11 prestressed concrete beams:

$$38.43, 38.43, 38.39, 38.83, 38.45, 38.35, 38.43, 38.31, 38.32, 38.48, 38.50.$$

1000 bootstrap samples (each sample hence consisting of 11 measurements) were generated from these data, and the 1000 bootstrap means were arranged on order. Refer to the smallest as $\bar{x}^*_{(1)}$, the second smallest as $\bar{x}^*_{(2)}$ and so on, with the largest being $\bar{x}^*_{(1000)}$. Assume that

$$\begin{aligned}
\bar{x}^*_{(25)} &= 38.3818,\\
\bar{x}^*_{(26)} &= 38.3818,\\
\bar{x}^*_{(50)} &= 38.3909,\\
\bar{x}^*_{(51)} &= 38.3918,\\
\bar{x}^*_{(950)} &= 38.5218,\\
\bar{x}^*_{(951)} &= 38.5236,\\
\bar{x}^*_{(975)} &= 38.5382,\\
\bar{x}^*_{(976)} &= 38.5391.
\end{aligned}$$

a) Compute a 95% bootstrap confidence interval for the mean compressive strength.

|||| Solution

Looking at Method box 4.10, we see that we need to find the 2.5%, and 97.5% quantiles of the 1000 bootstrap samples. According to the rule for finding the 2.5% quantile this should be the average of the 25th andn the 26th observation:

$$q_{0.025} = \frac{\bar{x}^*_{(25)} + \bar{x}^*_{(26)}}{2} = 38.3818,$$

and similarly

$$q_{0.975} = \frac{\bar{x}^*_{(975)} + \bar{x}^*_{(976)}}{2} = \frac{38.5382 + 38.5391}{2} = 38.5387,$$

and hence the 95% bootstrap confidence band is:

$$[38.3818; 38.5387].$$

b) Compute a 90% bootstrap confidence interval for the mean compressive strength.

|||| Solution

As above we get:

$$q_{0.05} = \frac{\bar{x}^*_{(50)} + \bar{x}^*_{(51)}}{2} = \frac{38.3909 + 38.3919}{2} = 38.3914,$$

and similarly:

$$q_{0.95} = \frac{\bar{x}^*_{(950)} + \bar{x}^*_{(951)}}{2} = \frac{38.5218 + 38.5236}{2} = 38.5227,$$

and hence the 90% bootstrap confidence band is:

$$[38.3914; 38.5227].$$

4.3 Various bootstrap CIs

|||| Exercise 4.3 Various bootstrap CIs

Consider the data from the exercise above. These data are entered into R as:

```
x <- c(38.43, 38.43, 38.39, 38.83, 38.45, 38.35,
       38.43, 38.31, 38.32, 38.48, 38.50)
```

Now generate $k = 1000$ bootstrap samples and compute the 1000 means (go higher if your computer is fine with it)

a) What are the 2.5%, and 97.5% quantiles (so what is the 95% confidence interval for μ without assuming any distribution)?

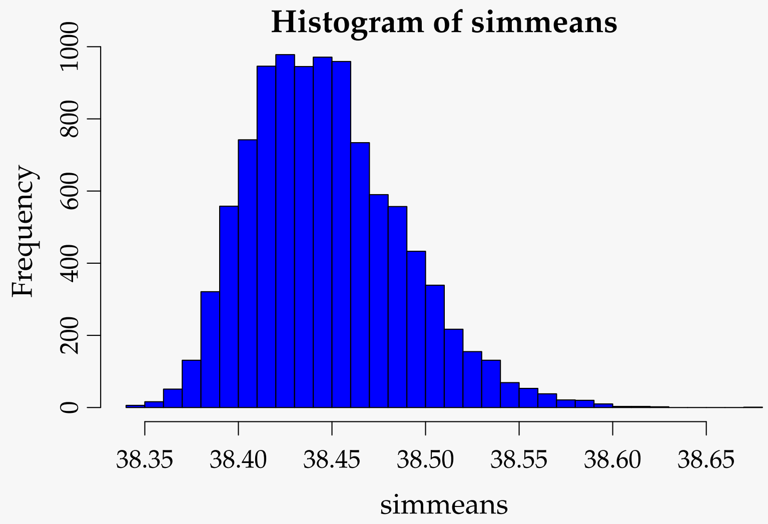
|||| Solution

The solution below has been generated with the following seed (see Remark 2.12)

```
## You might want to set the seed to achieve a particular result
set.seed(6287)
```

```
x <- c(38.43, 38.43, 38.39, 38.83, 38.45, 38.35,
       38.43, 38.31, 38.32, 38.48, 38.50)
k <- 10000
simsamples <- replicate(k, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))

 2.5% 97.5%
38.38 38.54
```

```r
hist(simmeans, col="blue", nclass=30)
```

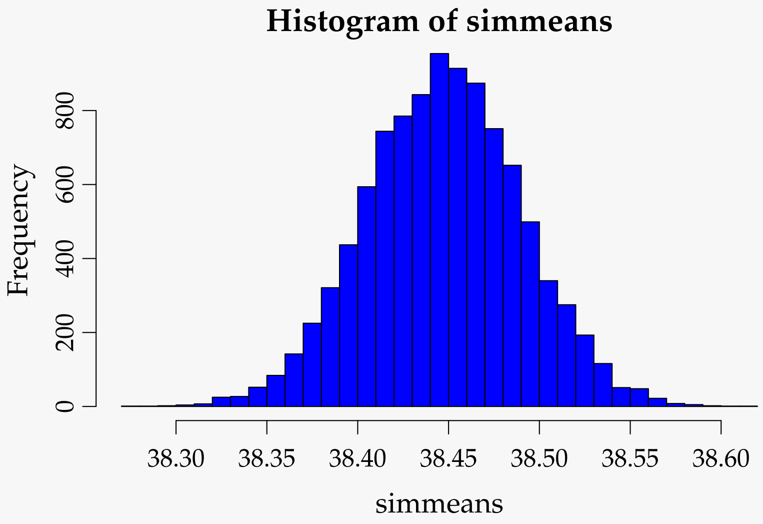


b) Find the 95% confidence interval for μ by the parametric bootstrap assuming the normal distribution for the observations. Compare with the classical analytic approach based on the t-distribution from Chapter 2.

|||| Solution

First we do the parametric bootstrap:

```
k <- 10000
n <- length(x)
simsamples <- replicate(k, rnorm(n, mean(x), sd(x)))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))

 2.5% 97.5%
38.36 38.53


hist(simmeans, col="blue", nclass=30)
```



And the classic *t*-based approach (without simulation):

```
t.test(x)


One Sample t-test

data:  x
t = 904, df = 10, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 38.35 38.54
sample estimates:
mean of x
    38.45
```

c) Find the 95% confidence interval for μ by the parametric bootstrap assuming the log-normal distribution for the observations. (Help: To use the `rlnorm` function to simulate the log-normal distribution, we face the challenge that we need to specify the mean and standard deviation on the log-scale and not on the raw scale, so compute mean and standard deviation for log-transformed data for this R-function)

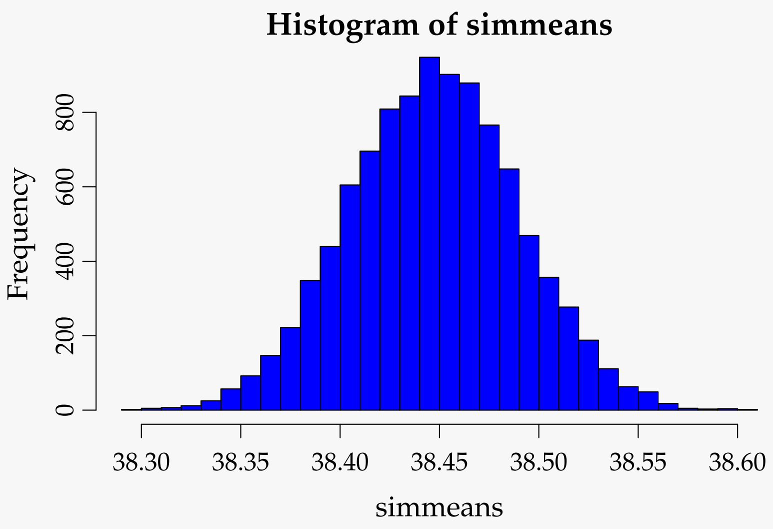
|||| Solution

We do the parametric bootstrap using the log-normal distribution.

```
k <- 10000
simsamples <- replicate(k, rlnorm(n, mean(log(x)), sd(log(x))))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))
```

```
 2.5% 97.5%
38.36 38.53
```

```
hist(simmeans, col="blue", nclass=30)
```



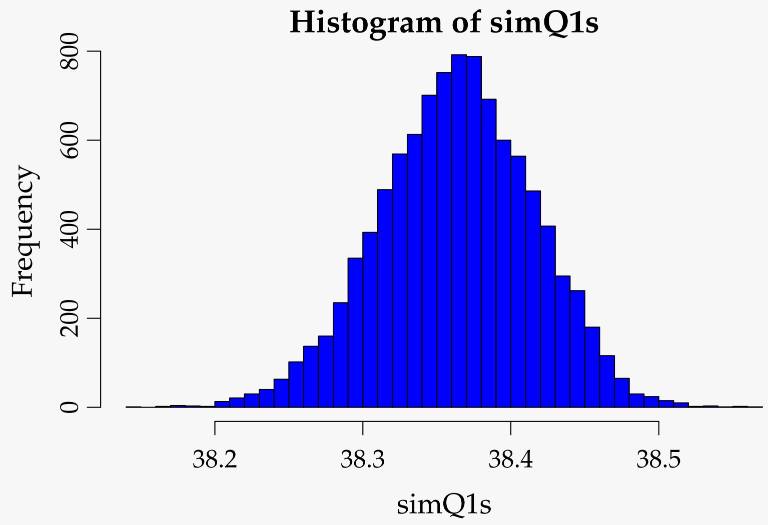
d) Find the 95% confidence interval for the lower quartile Q_1 by the parametric bootstrap assuming the normal distribution for the observations.

|||| Solution

We do the parametric bootstrap of lower quartile Q_1 using the normal distribution by first making a $Q1$-function in R, and then the usual stuff:

```
Q1 <- function(x){ quantile(x, 0.25) }
k <- 10000
simsamples <- replicate(k, rnorm(n, mean(x), sd(x)))
simQ1s <- apply(simsamples, 2, Q1)
quantile(simQ1s, c(0.025, 0.975))

 2.5% 97.5%
38.26 38.46
```

```
hist(simQ1s, col="blue", nclass=30)
```



e) Find the 95% confidence interval for the lower quartile Q_1 by the non-parametric bootstrap (so without any distributional assumptions)

Solution

We simply substitute the sampling line with the non-parametric version:

```
k <- 10000
simsamples <- replicate(k, sample(x, replace = TRUE))
simQ1s <- apply(simsamples, 2, Q1)
quantile(simQ1s, c(0.025, 0.975))

 2.5% 97.5%
38.31 38.43
```

4.4 Two-sample TV data

|||| Exercise 4.4 Two-sample TV data

A TV producer had 20 consumers evaluate the quality of two different TV flat screens - 10 consumers for each screen. A scale from 1 (worst) up to 5 (best) were used and the following results were obtained:

TV screen 1	TV screen 2
1	3
2	4
1	2
3	4
2	2
1	3
2	2
3	4
1	3
1	2

a) Compare the two means without assuming any distribution for the two samples (non-parametric bootstrap confidence interval and relevant hypothesis test interpretation).

Solution

The solution below has been generated with the following seed (see Remark 2.12)

```
## You might want to set the seed to achieve a particular result
set.seed(98273)
```

```
x1 <- c(1, 2, 1, 3, 2, 1, 2, 3, 1, 1)
x2 <- c(3, 4, 2, 4, 2, 3, 2, 4, 3, 2)
## Number of simulated (bootstrapped) samples
k = 10000
## Simulated samples of TV1 group
simx1samples = replicate(k, sample(x1, replace = TRUE))
## Simulate samples of TV2 group
simx2samples = replicate(k, sample(x2, replace = TRUE))
simmeandifs = apply(simx1samples, 2, mean) - apply(simx2samples, 2, mean)
## The quantiles giving the 95% CI
quantile(simmeandifs, c(0.025,0.975))

 2.5% 97.5%
 -1.9  -0.5
```

We reject the null hypothesis of $\mu_1 = \mu_2$, since zero is not included in the CI of the differences.

b) Compare the two means assuming normal distributions for the two samples - without using simulations (or rather: assuming/hoping that the sample sizes are large enough to make the results approximately valid).

Solution

```
t.test(x1, x2)


Welch Two Sample t-test

data:  x1 and x2
t = -3.2, df = 18, p-value = 0.005
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.9987 -0.4013
sample estimates:
mean of x mean of y
      1.7       2.9
```

We reject the null hypothesis of $\mu_1 = \mu_2$.

c) Compare the two means assuming normal distributions for the two samples - simulation based (parametric bootstrap confidence interval and relevant hypothesis test interpretation – in spite of the obviously wrong assumption).

Solution

```
simx1samples <- replicate(k, rnorm(n, mean(x1), sd(x1)))
simx2samples <- replicate(k, rnorm(n, mean(x2), sd(x2)))
simmeandifs = apply(simx1samples, 2, mean) - apply(simx2samples, 2, mean)
quantile(simmeandifs, c(0.025,0.975)) # percentiles

   2.5%   97.5%
-1.9223 -0.4886
```

We reject the null hypothesis of $\mu_1 = \mu_2$.

4.5 Non-linear error propagation

|||| Exercise 4.5 Non-linear error propagation

The pressure P, and the volume V of one mole of an ideal gas are related by the equation $PV = 8.31T$, when P is measured in kilopascals, T is measured in kelvins, and V is measured in liters.

a) Assume that P is measured to be 240.48 kPa and V to be 9.987 L with known measurement errors (given as standard deviations): 0.03 kPa and 0.002 L. Estimate T and find the uncertainty in the estimate.

|||| Solution

This is a almost direct copy of the rectangle example ($A = XY$) (Example 4.5), since $T = PV/8.31$, so since: To use the approximate error propagation rule, we must differentiate the function $f(x,y) = xy/8.31$ with respect to both x and y:

$$\frac{\partial f}{\partial x} = y/8.31 \quad \frac{\partial f}{\partial y} = x/8.31.$$

We get the result: $\hat{T} = 240.48 \cdot 9.987/8.31 = 289.0101$, and the uncertainty is:

$$\sigma_{\hat{T}} = \sqrt{9.987^2 \times 0.03^2 + 240.48^2 \times 0.002^2}/8.31 = 0.0682.$$

b) Assume that P is measured to be 240.48kPa and T to be 289.12K with known measurement errors (given as standard deviations): 0.03kPa and 0.02K. Estimate V and find the uncertainty in the estimate.

|||| Solution

$$V = f(P, T) = 8.31T/P.$$

So:

$$\frac{\partial f}{\partial T} = 8.31/P \ \frac{\partial f}{\partial P} = -8.31\frac{T}{P^2},$$

and hence:

$$\hat{V} = 8.31 \cdot 289.12/240.48 = 9.9908.$$

and

$$\sigma_{\hat{V}} = 8.31\sqrt{1/240.48^2 \times 0.02^2 + 289.12^2/240.48^4 \times 0.03^2} = 0.00143.$$

c) Assume that V is measured to be 9.987 L and T to be 289.12 K with known measurement errors (given as standard deviations): 0.002 L and 0.02 K. Estimate P and find the uncertainty in the estimate.

|||| Solution

Since

$$P = f(V, T) = 8.31T/V,$$

we can simply change the roles of P and V in the above and find similarly

$$\frac{\partial f}{\partial T} = 8.31/V \ \frac{\partial f}{\partial V} = -8.31\frac{T}{V^2},$$

and hence

$$\hat{P} = 8.31 \cdot 289.12/9.987 = 240.5715,$$

and

$$\sigma_{\hat{P}} = 8.31\sqrt{1/9.987^2 \times 0.02^2 + 289.12^2/9.987^4 \times 0.002^2} = 0.0510.$$

d) Try to answer one or more of these questions by simulation (assume that the errors are normally distributed).

Solution

Let's look at 3. The following R-code will do the job:

The solution below has been generated with the following seed (see Remark 2.12)

```
## You might want to set the seed to achieve a particular result
set.seed(28973)
```

```
k <- 10000
Vs <- rnorm(k, 9.987, sd = 0.002)
Ts <- rnorm(k, 289.12, sd = 0.02)
Ps <- 8.31*Ts/Vs
sd(Ps)

[1] 0.05124
```

Rerunning this a few times will show that 0.051 is the proper result. This additional re-running gives a feeling of the error in the simulation - rather small here. Alternatively increase k.

Similarly 2. can be handled as:

```
k <- 10000
Ps <- rnorm(k, 240.28, sd = 0.03)
Ts <- rnorm(k, 289.12, sd = 0.02)
Vs <- 8.31*Ts/Ps
sd(Vs)

[1] 0.001432
```

Providing again basically the same answer as above: 0.0014.

Chapter 5

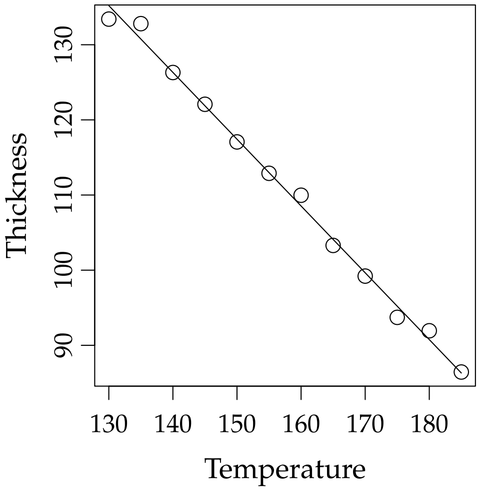
Simple Linear regression (solutions to exercises)

Contents

5.1 Plastic film folding machine

|||| Exercise 5.1 Plastic film folding machine

On a machine that folds plastic film the temperature may be varied in the range of 130-185 °C. For obtaining, if possible, a model for the influence of temperature on the folding thickness, $n = 12$ related set of values of temperature and the fold thickness were measured that is illustrated in the following figure:



a) Determine by looking at the figure, which of the following sets of estimates for the parameters in the usual regression model is correct:

1) $\hat{\beta}_0 = 0, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$

2) $\hat{\beta}_0 = 0, \hat{\beta}_1 = 0.9, \hat{\sigma} = 3.6$

3) $\hat{\beta}_0 = 252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 3.6$

4) $\hat{\beta}_0 = -252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$

5) $\hat{\beta}_0 = 252, \hat{\beta}_1 = -0.9, \hat{\sigma} = 36$

|||| Solution

First of all, the only possible intercept ($\hat{\beta}_0$) among the ones given in the answers is 252. And then the slope estimate of -0.9 in these two options looks reasonable. We just need to decide on whether the estimated standard deviation of the error $s_e = \hat{\sigma}$ is 3.6 or 36. From the figure it is clear that the points are NOT having an average vertical distance to the line in the size of 36, so 3.6 must be the correct number and hence the correct answer is:

3) $\hat{\beta}_0 = 252$, $\hat{\beta}_1 = -0.9$, $\hat{\sigma} = 3.6$

b) What is the only possible correct answer:

1) The proportion of explained variation is 50% and the correlation is 0.98

2) The proportion of explained variation is 0% and the correlation is -0.98

3) The proportion of explained variation is 96% and the correlation is -1

4) The proportion of explained variation is 96% and the correlation is 0.98

5) The proportion of explained variation is 96% and the correlation is -0.98

|||| Solution

The proportion of variation explained must be pretty high, so 0 can be ruled out. Answer 1 and 4 is also ruled out since the correlation clearly is negative. This also narrows the possibilities down to answer 3 and 5. And since the correlation is NOT exactly -1 (in which case the observations would be exactly on the line), the correct answer is:

5) The proportion of explained variation is 96% and the correlation is -0.98

5.2 Linear regression life time model

|||| Exercise 5.2 Linear regression life time model

A company manufactures an electronic device to be used in a very wide temperature range. The company knows that increased temperature shortens the life time of the device, and a study is therefore performed in which the life time is determined as a function of temperature. The following data is found:

Temperature in Celcius (t)	10	20	30	40	50	60	70	80	90
Life time in hours (y)	420	365	285	220	176	117	69	34	5

a) Calculate the 95% confidence interval for the slope in the usual linear regression model, which expresses the life time as a linear function of the temperature.

|||| Solution

Either one could do all the regression computations to find the $\hat{\beta}_1 = -5.3133$ and then subsequently use the formula for the confidence interval for β_1 in Method 5.15

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_1} = \hat{\beta}_1 \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

or just run `lm` in `R` to find:

```
D <- data.frame(t=c(10,20,30,40,50,60,70,80,90),
                y=c(420,365,285,220,176,117,69,34,5))
fit <- lm(y ~ t, data=D)
summary(fit)
```

```

Call:
lm(formula = y ~ t, data = D)

Residuals:
   Min     1Q Median     3Q    Max 
-21.02 -12.62  -9.16  17.71  29.64 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  453.556     14.394    31.5  8.4e-09 ***
t             -5.313      0.256   -20.8  1.5e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 19.8 on 7 degrees of freedom
Multiple R-squared:  0.984, Adjusted R-squared:  0.982 
F-statistic:  432 on 1 and 7 DF,  p-value: 0.000000151
```

and use the knowledge of the information in the R-output that wht is know as the "standard error for the slope" can be directly read off as

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.2558,$$

and $t_{0.025}(7) = 2.364$ - in R:

```
qt(.975,7)
```

```
[1] 2.365
```

to get $-5.31 \pm 2.365 \cdot 0.2558$, or in R:

```
-5.31+c(-1,1)*qt(.975,7)*0.2558
```

```
[1] -5.915 -4.705
```

b) Can a relation between temperature and life time be documented on level

5%?

|||| Solution

Since the confidence interval does not include 0, it can be documented that there is a relationship between life time and temperature, also the p-value is $1.5 \cdot 10^{-7} < 0.05 = \alpha$, which also give strong evidence against the null-hypothesis.

5.3 Yield of chemical process

|||| Exercise 5.3 Yield of chemical process

The yield y of a chemical process is a random variable whose value is considered to be a linear function of the temperature x. The following data of corresponding values of x and y is found:

Temperature in °C (x)	0	25	50	75	100
Yield in grams (y)	14	38	54	76	95

The average and standard deviation of temperature and yield are

$$\bar{x} = 50, \; s_x = 39.52847, \; \bar{y} = 55.4, \; s_y = 31.66702,$$

In the exercise the usual linear regression model is used

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, 5$$

a) Can a significant relationship between yield and temperature be documented on the usual significance level $\alpha = 0.05$?

|||| Solution

It could most easily be solved by running the regression in R as:

```
D <- data.frame(x=c(0,25,50,75,100),
                y=c(14,38,54,76,95))
fit <- lm(y ~ x, data=D)
summary(fit)


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

Residuals:
   1    2    3    4    5 
-1.4  2.6 -1.4  0.6 -0.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15.4000     1.4967    10.3    0.002 ** 
x             0.8000     0.0244    32.7 0.000063 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.93 on 3 degrees of freedom
Multiple R-squared:  0.997, Adjusted R-squared:  0.996 
F-statistic: 1.07e+03 on 1 and 3 DF,  p-value: 0.0000627
```

Alternatively one could use hand calculations and use the formula in Theorem 5.12 for the t-test of the null hypothesis: $H_0 : \beta_1 = 0$.

The relevant test statistic and p-value can be read off in the R output as 32.7 and 0.000063. So the answer is:

Yes, as the relevant test statistic and p-value are resp. 32.7 and $0.00006 < 0.05 = \alpha$.

b) Give the 95% confidence interval of the expected yield at a temperature of $x_{\text{new}} = 80$ °C.

Solution

We use the formula in Equation (5-59) for the confidence limit of the line (the expected value of Y_i for a value x_{new}):

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}},$$

and we have to compute $\hat{\beta}_0$, $\hat{\beta}_1$ and s_e either by hand OR in R as above:

$$\hat{\beta}_0 = 15.4, \ \hat{\beta}_1 = 0.8, \ \hat{\sigma} = 1.932.$$

So the confidence interval becomes

$$(15.4 + 0.8 \cdot 80) \pm 3.182 \cdot 1.932\sqrt{\frac{1}{5} + \frac{(80-50)^2}{6250}},$$

since

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}S_{xx} \Leftrightarrow$$
$$S_{xx} = (n-1)s_x^2 = 4 \cdot 39.528^2 = 6250.$$

Thus the answer is

$$79.40 \pm 3.61 = [75.79, \ 83.01].$$

In R this could be by:

```
predict(fit, newdata=data.frame(x=80), interval="confidence",
        level=0.95)

   fit   lwr   upr
1 79.4 75.79 83.01
```

c) What is the upper quartile of the residuals?

Solution

The five residuals become: -1.4, 2.6, -1.4, 0.6 og -0.4.

We use the basic definition of finding a quantile (from Definition 1.7) and the upper quartile is $q_{0.75}$ (see Definition 1.8). We set $n = 5$, $p = 0.75$, so

$$np = 3.75$$

So the upper quartile is the 4th observation in the ordered sequence:

$$-1.4, -1.4, -0.4, 0.6, 2.6.$$

This is also found in the `summary()` output above under

```
Residuals:
   1    2    3    4    5
-1.4  2.6 -1.4  0.6 -0.4
```

So the answer is: 0.6.

5.4 Plastic material

|||| Exercise 5.4 Plastic material

In the manufacturing of a plastic material, it is believed that the cooling time has an influence on the impact strength. Therefore a study is carried out in which plastic material impact strength is determined for 4 different cooling times. The results of this experiment are shown in the following table:

Cooling times in seconds (x)	15	25	35	40
Impact strength in kJ/m^2 (y)	42.1	36.0	31.8	28.7

The following statistics may be used:

$$\bar{x} = 28.75, \ \bar{y} = 34.65, \ S_{xx} = 368.75.$$

a) What is the 95% confidence interval for the slope of the regression model, expressing the impact strength as a linear function of the cooling time?

|||| Solution

The easiest way to get to the confidence interval is to use the standard error for the slope ($\hat{\sigma}_{\beta_1}$ or denoted with SE_{β_1}) given in the R output:

```
x <- c(15,25,35,40)
y <- c(42.1,36.0,31.8,28.7)
summary(lm(y ~ x))


Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4 
 0.2814 -0.6051  0.4085 -0.0847 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   49.639      0.878    56.5  0.00031 ***
x             -0.521      0.029   -18.0  0.00308 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.556 on 2 degrees of freedom
Multiple R-squared:  0.994, Adjusted R-squared:  0.991 
F-statistic:  324 on 1 and 2 DF,  p-value: 0.00308
```

the standard error for the slope is $\hat{\sigma}_{\beta_1} = 0.029$ (also known as the sampling distribution standard deviation for $\hat{\beta}_1$). Finding the relevant t-quantile (with $\nu = 2$ degrees of freedom (either of):

```
c(qt(0.025, df=2), qt(0.975, df=2))

[1] -4.303  4.303
```

$|t_{0.025}| = 4.303$, which using Theorem 5.15 gives

$$-0.521 \pm 4.303 \cdot 0.029,$$

giving

$$-0.521 \pm 0.125,$$

or, that we say with high confidence that the true parameter value is in the interval, i.e.

$$-0.646 \leq \beta_1 \leq -0.396.$$

b) Can you conclude that there is a relation between the impact strength and the cooling time at significance level $\alpha = 5\%$?

|||| Solution

The relevant p-value can be read off directly from the `summary` output: 0.00308, and we can conclude: *Yes, as the relevant p-value is 0.00308, which is smaller than 0.05.*

c) For a similar plastic material the tabulated value for the linear relation between temperature and impact strength (i.e the slope) is -0.30. If the following hypothesis is tested (at level $\alpha = 0.05$)

$$H_0 : \beta_1 = -0.30$$
$$H_1 : \beta_1 \neq -0.30$$

with the usual t-test statistic for such a test, what is the range (for t) within which the hypothesis is accepted?

|||| Solution

The so-called critical values for the t-statistic with $\nu = 2$ degrees of freedom is found as (or at least the negative one of the two): $t_{0.025} = -4.303$ - in R: `qt(0.975,2)`). So the answer becomes:

$$[-4.303, 4.303].$$

5.5 Water polution

|||| Exercise 5.5 Water polution

In a study of pollution in a water stream, the concentration of pollution is measured at 5 different locations. The locations are at different distances to the pollution source. In the table below, these distances and the average pollution are given:

Distance to the pollution source (in km)	2	4	6	8	10
Average concentration	11.5	10.2	10.3	9.68	9.32

a) What are the parameter estimates for the three unknown parameters in the usual linear regression model: 1) The intercept (β_0), 2) the slope (β_1) and 3) error standard deviation (σ)?

Solution

The question is solved by considering the following R-output:

```
D <- data.frame(concentration=c(11.5, 10.2, 10.3, 9.68, 9.32),
                distance=c(2, 4, 6, 8, 10))
fit <- lm(concentration ~ distance, data=D)
summary(fit)


Call:
lm(formula = concentration ~ distance, data = D)

Residuals:
     1      2      3      4      5
 0.324 -0.488  0.100 -0.032  0.096

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.664      0.365   31.96 0.000067 ***
distance      -0.244      0.055   -4.43    0.021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.348 on 3 degrees of freedom
Multiple R-squared:  0.868, Adjusted R-squared:  0.823
F-statistic: 19.7 on 1 and 3 DF,  p-value: 0.0213
```

Given the knowledge of the R-output structure, the three values can be read off directly from the output.
So the correct answer is: $\hat{\beta}_0 = 11.7$, $\hat{\beta}_1 = -0.244$ and $SE_{\hat{\sigma}} = \hat{\sigma} = 0.348$.

b) How large a part of the variation in concentration can be explained by the distance?

Solution

The amount of variation in the model output (Y) explained by the variable input (x) can be found from the squared correlation, that can be read off directly from the output as `"Multiple R-squared"`. So the correct answer is: $R^2 = 86.8\%$ (it is actually an estimate of the variation in concentration which can be explained by distance, since it is what we found with the particular data at hand. If the sample was taken again, then this value would vary. We should actually calculate a confidence interval for R^2 to understand how accurate this estimate is!).

c) What is a 95%-confidence interval for the expected pollution concentration 7 km from the pollution source?

|||| Solution

The wanted number is estimated by the point on the line (using $x_{\text{new}} = 7$)

$$-0.244 \cdot 7 + 11.664 = 9.96,$$

and the confidence interval is given by

$$9.96 \pm t_{0.025}(3) \cdot \hat{\sigma} \sqrt{\frac{1}{5} + \frac{(7-6)^2}{S_{xx}}},$$

where $S_{xx} = 4^2 + 2^2 + 0^2 + 2^2 + 4^2 = 40$ and $t_{0.025}(3) = 3.182$ (in R: `qt(0.975,3)`) we have that

$$3.182 \cdot 0.348 \sqrt{\frac{1}{5} + \frac{1}{40}} = 0.525,$$

where s_x is:

```
sd(D$distance)

[1] 3.162
```

and thus

$$S_{xx} = (n-1) \cdot s_x^2 = 4 \cdot 3.162^2 = 40.$$

This could also have been found by

```
predict(fit, newdata=data.frame(distance=7), interval="confidence",
        level=0.95)

    fit   lwr   upr
1 9.956 9.431 10.48
```

So the correct answer is:

$$9.96 \pm 0.525 = [9.43, 10.5].$$

5.6 Membrane pressure drop

Exercise 5.6 Membrane pressure drop

When purifying drinking water you can use a so-called membrane filtration. In an experiment one wishes to examine the relationship between the pressure drop across a membrane and the flux (flow per area) through the membrane. We observe the following 10 related values of pressure (x) and flux (y):

	1	2	3	4	5	6	7	8	9	10
Pressure (x)	1.02	2.08	2.89	4.01	5.32	5.83	7.26	7.96	9.11	9.99
Flux (y)	1.15	0.85	1.56	1.72	4.32	5.07	5.00	5.31	6.17	7.04

Copy this into R to avoid typing in the data:

```
D <- data.frame(
    pressure=c(1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99),
    flux=c(1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04)
)
```

a) What is the empirical correlation between pressure and flux estimated to? Give also an interpretation of the correlation.

Solution

The questions are most easily solved by using `lm` in R:

```
D <- data.frame(
  pressure=c(1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99),
  flux=c(1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04)
)
fit <- lm(flux ~ pressure, data=D)
summary(fit)


Call:
lm(formula = flux ~ pressure, data = D)

Residuals:
   Min     1Q Median     3Q    Max 
-0.989 -0.318 -0.140  0.454  1.046 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  -0.1886     0.4417   -0.43      0.68    
pressure      0.7225     0.0706   10.23 0.0000072 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.645 on 8 degrees of freedom
Multiple R-squared:  0.929, Adjusted R-squared:  0.92 
F-statistic:  105 on 1 and 8 DF,  p-value: 0.00000718
```

The found coefficient of determination (see Theorem 5.25) can be read off the R output to be 0.929. The sign of the correlation is the same as the sign of the slope, which can be read off to be positive ($\hat{\beta}_1 = 0.7225$), so the correlation is

$$\hat{\rho} = r = \sqrt{0.929} = 0.964.$$

So the empirical correlation is 0.964, and thus flux is found to increase with increasing pressure.

b) What is a 90% confidence interval for the slope β_1 in the usual regression model?

|||| Solution

We use the formula for the slope (β_1, see Method 5.15) confidence interval, and can actually just realize that the correct t-quantile to use is the $t_{1-0.05}(8) = 1.860$ (in R: `qt(0.95,8)`), and the other values we read of the `summary` output.
So the confidence interval is: $0.7225 \pm 1.860 \cdot 0.0706$.

c) How large a part of the flux-variation ($\sum_{i=1}^{10}(y_i - \bar{y})^2$) is not explained by pressure differences?

|||| Solution

The squared correlation, $r^2 = 0.929$ express the explained variation, this means that $1 - 0.929 = 0.071$ express the unexplained variation by the model.

d) Can you at significance level $\alpha = 0.05$ reject the hypothesis that the line passes through $(0,0)$?

|||| Solution

The hypothesis is the same as:

$$H_0 : \beta_0 = 0$$

which is the hypothesis results provided in the output in the `"intercept"` row of `summary`, so: *No, since the relevant p-value is 0.68, which is larger than α.*

e) A confidence interval for the line at three different pressure levels: $x_{\text{new}}^{\text{A}} = 3.5$, $x_{\text{new}}^{\text{B}} = 5.0$ and $x_{\text{new}}^{\text{C}} = 9.5$ will look as follows:

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{\text{new}}^{U} \pm C_U$$

where U then is either A, B or C. Write the constants C_U in increasing order.

|||| Solution

The formula for the Confidence limits of $\alpha + \beta x_{\text{new}}$ includes the following term:

$$\frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}$$

and this is the ONLY term in C_U that makes C_U different between the three Us. And since $\bar{x} = 5.547$ it is clear that

$$(5.0 - 5.547)^2 < (3.5 - 5.547)^2 < (9.5 - 5.547)^2$$

and hence

$$(x^{\text{B}}_{\text{new}} - 5.547)^2 < (x^{\text{A}}_{\text{new}} - 5.547)^2 < (x^{\text{C}}_{\text{new}} - 5.547)^2$$

So $C_{\text{B}} < C_{\text{A}} < C_{\text{C}}$

5.7 Membrane pressure drop (matrix form)

Exercise 5.7 Membrane pressure drop (matrix form)

This exercise uses the data presented in Exercise 6 above.

a) Find parameters values, standard errors, *t*-test statistics, and *p*-values for the standard hypotheses tests.

   Copy this into R to avoid typing in the data:

```
D <- data.frame(
  pressure=c(1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99),
  flux=c(1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04)
)
```

Solution

```
D <- data.frame(
  pressure=c(1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99),
  flux=c(1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04)
)
fit <- lm(flux ~ pressure, data=D)
summary(fit)
```

```

Call:
lm(formula = flux ~ pressure, data = D)

Residuals:
   Min     1Q Median     3Q    Max 
-0.989 -0.318 -0.140  0.454  1.046 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  -0.1886     0.4417   -0.43      0.68    
pressure      0.7225     0.0706   10.23 0.0000072 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.645 on 8 degrees of freedom
Multiple R-squared:  0.929, Adjusted R-squared:  0.92 
F-statistic:  105 on 1 and 8 DF,  p-value: 0.00000718
```

The parameter estimates are given in the first column, the standard errors in the second column, the t-test statistics are given in the third column and the p-values of the standard hypothesis are given in the last column.

b) Reproduce the above numbers by matrix vector calculations. You will need some matrix notation in R:

- Matrix multiplication (XY): `X%*%Y`
- Matrix transpose (X^T): `t(X)`
- Matrix inverse (X^{-1}): `solve(X)`
- Make a matrix from vectors ($X = [x_1^T; x_2^T]$): `cbind(x1,x2)`

See also Example 5.24.

Solution

```
X <- cbind(1, D$pressure)
y <- D$flux
n <- length(y)
beta <- solve(t(X) %*%X ) %*% t(X) %*% y
beta

        [,1]
[1,] -0.1886
[2,]  0.7225


e <- y - X %*% beta
s <- sqrt(sum(e^2)/(n-2))
Vbeta <- s^2 * solve(t(X) %*%X )
se.beta <- sqrt(diag(Vbeta))
t.obs <- beta / se.beta
p.value <- 2 * (1 - pt(abs(t.obs), df = n-2))

## Collection in a table
analasis.table <- cbind(beta, se.beta, t.obs, p.value)
analasis.table

             se.beta
[1,] -0.1886 0.44171 -0.4269 0.680696710
[2,]  0.7225 0.07064 10.2269 0.000007177


## Put some names on our table
colnames(analasis.table) <- c("Estimates","Std.Error","t.obs","p.value")
rownames(analasis.table) <- c("beta1","beta2")
analasis.table

      Estimates Std.Error   t.obs     p.value
beta1   -0.1886   0.44171 -0.4269 0.680696710
beta2    0.7225   0.07064 10.2269 0.000007177


## Done!!
```

5.8 Independence and correlation

|||| Exercise 5.8 Independence and correlation

Consider the layout of independent variable in Example 5.11,

a) Show that $S_{xx} = \frac{n\cdot(n+1)}{12\cdot(n-1)}$.

Hint: you can use the following relations

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

|||| Solution

$\bar{x}$ becomes

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n}\frac{i-1}{n-1} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(i-1)\\ &= \frac{1}{n(n-1)}\left(\frac{n(n+1)}{2}-n\right) = \frac{1}{2},\end{aligned}$$

and S_{xx} becomes

$$\begin{aligned}S_{xx} &= \sum_{i=1}^{n}\left(\frac{i-1}{n-1}-\frac{1}{2}\right)^2\\ &= -\frac{n}{4}+\frac{1}{(n-1)^2}\sum_{i=1}^{n}(i^2+1-2i)\\ &= -\frac{n}{4}+\frac{1}{(n-1)^2}\left(\frac{n(n+1)(2n+1)-6n^2}{6}\right)\\ &= \frac{n}{(n-1)^2}\left(\frac{4n^2+6n+2-12n-3(n-1)^2}{12}\right)\\ &= \frac{n}{(n-1)^2}\left(\frac{n^2-1}{12}\right) = \frac{n(n+1)}{12(n-1)}.\end{aligned}$$

b) Show that the asymptotic correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\lim_{n\to\infty} \rho_n(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sqrt{3}}{2}.$$

|||| Solution

The correlation between $\hat{\beta}_0$ and $\hat{\beta}_0$ is

$$\begin{aligned}
\rho_n(\hat{\beta}_0, \hat{\beta}_1) &= \frac{\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\mathrm{V}(\hat{\beta}_0)\,\mathrm{V}(\hat{\beta}_1)}} \\
&= -\frac{\sigma^2 \bar{x}/S_{xx}}{\sqrt{\sigma^4 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\frac{1}{S_{xx}}}} \\
&= -\frac{\bar{x}/S_{xx}}{\frac{1}{S_{xx}}\sqrt{\left(\frac{S_{xx}}{n} + \bar{x}^2\right)}} \\
&= -\frac{\bar{x}}{\sqrt{\frac{S_{xx}}{n} + \bar{x}^2}}.
\end{aligned}$$

Notice that the correlation is not a function of the variance (σ^2), but only a function of the independent variables. Now insert the values of $\bar{x}$ and S_{xx}

$$\begin{aligned}
\rho_n(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{1}{2\sqrt{\frac{n+1}{12(n-1)} + \frac{1}{4}}} = -\frac{1}{2\sqrt{\frac{n+1+3(n-1)}{12(n-1)}}} \\
&= -\frac{1}{2\sqrt{\frac{2n-1}{6(n-1)}}} = -\frac{\sqrt{6(n-1)}}{2\sqrt{2n-1}} \\
&= -\frac{1}{2}\sqrt{\frac{6(n-1)}{2(n-1/2)}} = -\frac{\sqrt{3}}{2}\sqrt{\frac{n-1}{n-1/2}}
\end{aligned}$$

.

which converges to $-\frac{\sqrt{3}}{2}$ for $n \to \infty$.

Consider a layout of the independent variable where $n = 2k$ and $x_i = 0$ for $i \leq k$ and $x_i = 1$ for $k < i \leq n$.

c) Find S_{xx} for the new layout of x.

|||| Solution

$$\bar{x} = \frac{1}{2},$$

and

$$\begin{aligned} S_{xx}^{\text{new}} &= \sum_{i=1}^{k}\left(0-\frac{1}{2}\right)^2 + \sum_{i=k+1}^{2k}\left(1-\frac{1}{2}\right)^2 \\ &= \frac{k}{4}+\frac{k}{4}=\frac{k}{2}=\frac{n}{4}. \end{aligned}$$

d) Compare S_{xx} for the two layouts of x.

|||| Solution

$$\frac{S_{xx}}{S_{xx}^{\text{new}}} = \frac{n(n+1)}{12(n-1)}\frac{4}{n} = \frac{(n+1)}{3(n-1)} < 1; \quad for \quad n > 2$$

which imply that $S_{xx}^{\text{new}} > S_{xx}$ for all $n > 2$.

e) What is the consequence for the parameter variance in the two layouts?

|||| Solution

The larger S_{xx} for the new layout imply that the parameter variance is smaller for the new layout (given that data comes from the same model).

f) Discuss pro's and cons for the two layouts.

|||| Solution

The smaller parameter variance for the new layout would suggest that we should use this layout. However, we would not be able to check that data is in fact generated by a linear model. Consider e.g. data generated by the model

$$y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

if we only look at $x_i = 0$ or $x_i = 1$ we will not be able to detect that the relationship is in fact non-linear.

Chapter 6

Multiple Linear Regression (solutions to exercises)

Contents

6.1 Nitrate concentration

|||| Exercise 6.1 Nitrate concentration

In order to analyze the effect of reducing nitrate loading in a Danish fjord, it was decided to formulate a linear model that describes the nitrate concentration in the fjord as a function of nitrate loading, it was further decided to correct for fresh water runoff. The resulting model was

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{6-1}$$

where Y_i is the natural logarithm of nitrate concentration, $x_{1,i}$ is the natural logarithm of nitrate loading, and $x_{2,i}$ is the natural logarithm of fresh water run off.

a) Which of the following statements are assumed fulfilled in the usual multiple linear regression model?

1) $\varepsilon_i = 0$ for all $i = 1, ..., n$, and β_j follows a normal distribution

2) $E[x_1] = E[x_2] = 0$ and $V[\varepsilon_i] = \beta_1^2$

3) $E[\varepsilon_i] = 0$ and $V[\varepsilon_i] = \beta_1^2$

4) ε_i is normally distributed with constant variance, and ε_i and ε_j are independent for $i \neq j$

5) $\varepsilon_i = 0$ for all $i = 1, ..., n$, and x_j follows a normal distribution for $j = \{1, 2\}$

|||| Solution

1) ε_i follows a normal distribution with expectation equal zero, but the realizations are not zero, and further β_j is deterministic and hence it does not follow a distribution ($\hat{\beta}_j$ does), hence 1) is not correct

2)- 3) There are no assumptions on the expectation of x_j and the variance of ε equal σ^2, not β_1^2 hence 2) and 3) are not correct

4) Is correct, this is the usual assumption about the errors

5) Is incorrect since ε_j follow a normal distribution, further the are no distributional assumptions on x_j. In fact we assume that x_j is known

The parameters in the model were estimated in R and the following results are available (slightly modified output from summary):

```
> summary(lm(y ~ x1 + x2))

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.36500    0.22184 -10.661  < 2e-16
x1           0.47621    0.06169   7.720 3.25e-13
x2           0.08269    0.06977   1.185    0.237
---
Residual standard error: 0.3064 on 237 degrees of freedom
Multiple R-squared: 0.3438,Adjusted R-squared: 0.3382
F-statistic: 62.07 on 2 and 237 DF,  p-value: < 2.2e-16
```

b) What are the parameter estimates for the model parameters ($\hat{\beta}_i$ and $\hat{\sigma}^2$) and how many observations are included in the estimation?

Solution

The number of degrees of freedom is equal $n - (p+1)$, and since the number of degrees of freedom is 237 and $p = 2$, we get $n = 237 + 2 + 1 = 240$. The parameters are given in the first column of the coefficient matrix, i.e.

$$\begin{aligned} \hat{\beta}_0 &= -2.365 \tag{6-2} \\ \hat{\beta}_1 &= 0.476 \tag{6-3} \\ \hat{\beta}_2 &= 0.083 \tag{6-4} \end{aligned}$$

and finally the estimated error variance is $\hat{\sigma}^2 = 0.3064^2$.

c) Calculate the usual 95% confidence intervals for the parameters (β_0, β_1, and β_2).

|||| Solution

From Theorem 6.5 we know that the confidence intervals can be calculated by

$$\hat{\beta}_i \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_i},$$

where $t_{1-\alpha/2}$ is based on 237 degrees of freedom, and with $\alpha = 0.05$, we get $t_{0.975} = 1.97$. The standard errors for the estimates is the second column of the coefficient matrix, and the confidence intervals become

$$\hat{\beta}_0 = -2.365 \pm 1.97 \cdot 0.222 \tag{6-5}$$
$$\hat{\beta}_1 = 0.467 \pm 1.97 \cdot 0.062 \tag{6-6}$$
$$\hat{\beta}_2 = 0.083 \pm 1.97 \cdot 0.070 \tag{6-7}$$

d) On level $\alpha = 0.05$ which of the parameters are significantly different from 0, also find the p-values for the tests used for each of the parameters?

|||| Solution

We can see directly from the confidence intervals above that β_0 and β_1 are significantly different from zero (the confidence intervals does not cover zero), while we cannot reject that $\beta_2 = 0$ (the confidence interval cover zero). The p-values we can see directly in the R output: for β_0 is less than 10^{-16} and the p-value for β_1 is $3.25 \cdot 10^{-13}$, i.e. very strong evidence against the null hypothesis in both cases.

6.2 Multiple linear regression model

|||| Exercise 6.2 Multiple linear regression model

The following measurements have been obtained in a study:

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
y	1.45	1.93	0.81	0.61	1.55	0.95	0.45	1.14	0.74	0.98	1.41	0.81	0.89	0.68	1.39	1.53	0.91	1.49	1.38	1.73	1.11	1.68	0.66	0.69	1.98
x_1	0.58	0.86	0.29	0.20	0.56	0.28	0.08	0.41	0.22	0.35	0.59	0.22	0.26	0.12	0.65	0.70	0.30	0.70	0.39	0.72	0.45	0.81	0.04	0.20	0.95
x_2	0.71	0.13	0.79	0.20	0.56	0.92	0.01	0.60	0.70	0.73	0.13	0.96	0.27	0.21	0.88	0.30	0.15	0.09	0.17	0.25	0.30	0.32	0.82	0.98	0.00

It is expected that the response variable y can be described by the independent variables x_1 and x_2. This imply that the parameters of the following model should be estimated and tested

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

a) Calculate the parameter estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$), in addition find the usual 95% confidence intervals for β_0, β_1, and β_2.
You can copy the following lines to R to load the data:

```
D <- data.frame(
  x1=c(0.58, 0.86, 0.29, 0.20, 0.56, 0.28, 0.08, 0.41, 0.22,
       0.35, 0.59, 0.22, 0.26, 0.12, 0.65, 0.70, 0.30, 0.70,
       0.39, 0.72, 0.45, 0.81, 0.04, 0.20, 0.95),
  x2=c(0.71, 0.13, 0.79, 0.20, 0.56, 0.92, 0.01, 0.60, 0.70,
       0.73, 0.13, 0.96, 0.27, 0.21, 0.88, 0.30, 0.15, 0.09,
       0.17, 0.25, 0.30, 0.32, 0.82, 0.98, 0.00),
  y=c(1.45, 1.93, 0.81, 0.61, 1.55, 0.95, 0.45, 1.14, 0.74,
      0.98, 1.41, 0.81, 0.89, 0.68, 1.39, 1.53, 0.91, 1.49,
      1.38, 1.73, 1.11, 1.68, 0.66, 0.69, 1.98)
)
```

Solution

The question is answered by R. Start by loading data into R and estimate the parameters in R

```
fit <- lm(y ~ x1 + x2, data=D)
summary(fit)


Call:
lm(formula = y ~ x1 + x2, data = D)

Residuals:
   Min     1Q Median     3Q    Max
-0.155 -0.078 -0.020  0.050  0.301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43355    0.06598    6.57  1.3e-06 ***
x1           1.65299    0.09525   17.36  2.5e-14 ***
x2           0.00394    0.07485    0.05     0.96
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.113 on 22 degrees of freedom
Multiple R-squared:  0.94, Adjusted R-squared:  0.934
F-statistic:  172 on 2 and 22 DF,  p-value: 3.7e-14
```

Solution

The parameter estimates are given in the first column of the coefficient matrix, i.e.

$$
\begin{aligned}
\hat{\beta}_0 &= 0.434,\\
\hat{\beta}_1 &= 1.653,\\
\hat{\beta}_2 &= 0.0039,
\end{aligned}
$$

and the error variance estimate is $\hat{\sigma}^2 = 0.11^2$. The confidence intervals can either be calculated using the second column of the coefficient matrix, and the value of $t_{0.975}$ (with degrees of freedom equal 22), or directly in R:

```
confint(fit)

              2.5 % 97.5 %
(Intercept)  0.2967 0.5704
x1           1.4555 1.8505
x2          -0.1513 0.1592
```

b) Still using confidence level $\alpha = 0.05$ reduce the model if appropriate.

|||| Solution

Since the confidence interval for β_2 cover zero (and the p-value is much larger than 0.05), the parameter should be removed from the model to get the simpler model

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

the parameter estimates in the simpler model are

```
fit <- lm(y ~ x1, data=D)
summary(fit)


Call:
lm(formula = y ~ x1, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1563 -0.0763 -0.0215  0.0516  0.2999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4361     0.0440    9.91  9.0e-10 ***
x1            1.6512     0.0871   18.96  1.5e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.11 on 23 degrees of freedom
Multiple R-squared:  0.94, Adjusted R-squared:  0.937
F-statistic:  360 on 1 and 23 DF,  p-value: 1.54e-15
```

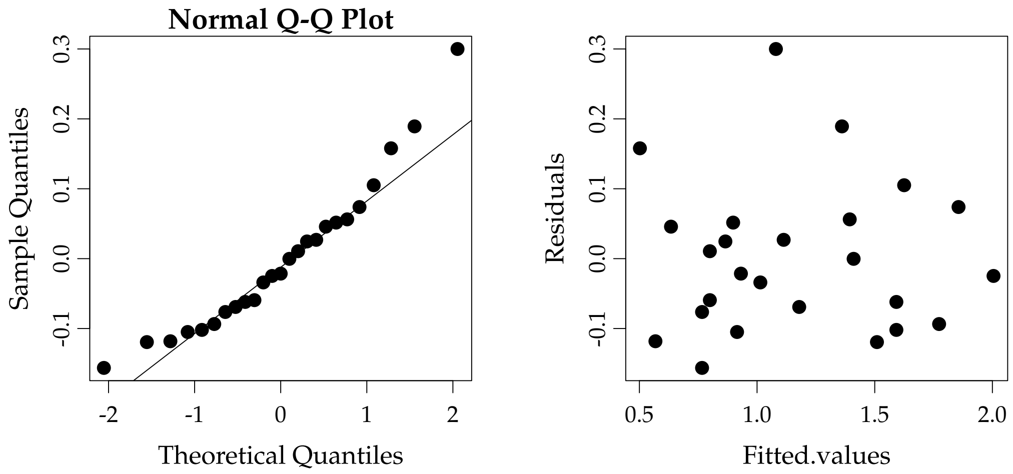
and both parameters are now significant.

c) Carry out a residual analysis to check that the model assumptions are fulfilled.

|||| Solution

We are interested in inspecting a q-q plot of the residuals and a plot of the residuals as a function of the fitted values

```
par(mfrow=c(1,2))
qqnorm(fit$residuals, pch=19)
qqline(fit$residuals)
plot(fit$fitted.values, fit$residuals, pch=19,
     xlab="Fitted.values", ylab="Residuals")
```

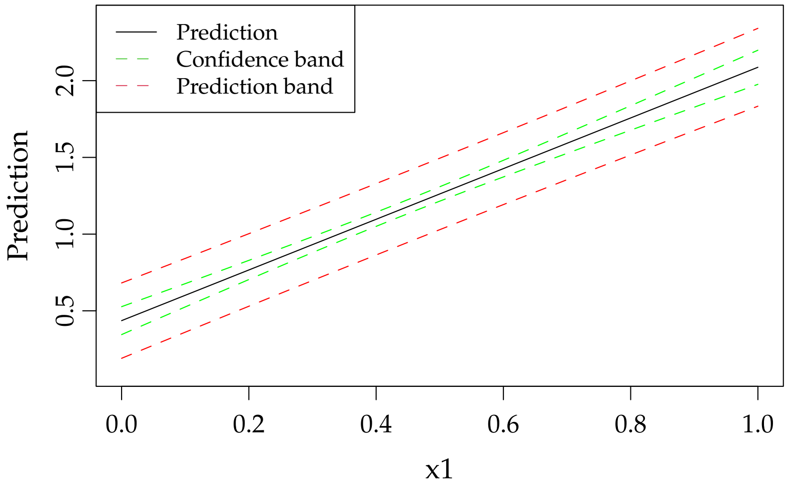


there are no strong evidence against the assumptions, the qq-plot is are a straight line and the are no obvious dependence between the residuals and the fitted values, and we conclude that the assumptions are fulfilled.

d) Make a plot of the fitted line and 95% confidence and prediction intervals of the line for $x_1 \in [0,1]$ (it is assumed that the model was reduced above).

|||| Solution

```r
x1new <- seq(0,1,by=0.01)
pred <- predict(fit, newdata=data.frame(x1=x1new),
                interval="prediction")
conf <- predict(fit, newdata=data.frame(x1=x1new),
                interval="confidence")
plot(x1new, pred[ ,"fit"], type="l", ylim=c(0.1,2.4),
     xlab="x1", ylab="Prediction")
lines(x1new, conf[ ,"lwr"], col="green", lty=2)
lines(x1new, conf[ ,"upr"], col="green", lty=2)
lines(x1new, pred[ ,"lwr"], col="red", lty=2)
lines(x1new, pred[ ,"upr"], col="red", lty=2)
legend("topleft", c("Prediction","Confidence band","Prediction band"),
       lty=c(1,2,2), col=c(1,3,2), cex=0.7)
```



6.3 MLR simulation exercise

|||| Exercise 6.3 MLR simulation exercise

The following measurements have been obtained in a study:

Nr.	1	2	3	4	5	6	7	8
y	9.29	12.67	12.42	0.38	20.77	9.52	2.38	7.46
x_1	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00
x_2	4.00	12.00	16.00	8.00	32.00	24.00	20.00	28.00

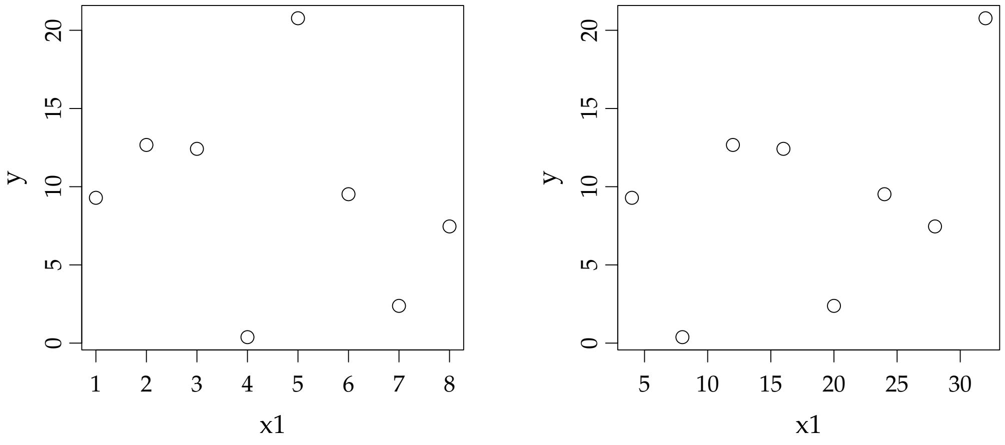
a) Plot the observed values of y as a function of x_1 and x_2. Does it seem reasonable that either x_1 or x_2 can describe the variation in y?
You may copy the following lines into R to load the data

```
D <- data.frame(
  y=c(9.29,12.67,12.42,0.38,20.77,9.52,2.38,7.46),
  x1=c(1.00,2.00,3.00,4.00,5.00,6.00,7.00,8.00),
  x2=c(4.00,12.00,16.00,8.00,32.00,24.00,20.00,28.00)
)
```

|||| Solution

The data is plotted with

```
par(mfrow=c(1,2))
plot(D$x1, D$y, xlab="x1", ylab="y")
plot(D$x2, D$y, xlab="x1", ylab="y")
```



There does not seem to be a strong relation between y and x_1 or x_2.

b) Estimate the parameters for the two models

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

and

$$Y_i = \beta_0 + \beta_1 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

and report the 95% confidence intervals for the parameters. Are any of the parameters significantly different from zero on a 5% confidence level?

Solution

The models are fitted with

```
fit1 <- lm(y ~ x1, data=D)
fit2 <- lm(y ~ x2, data=D)
confint(fit1)

                2.5 % 97.5 %
(Intercept) -0.5426 24.898
x1          -3.1448  1.893


confint(fit2)

                2.5 %  97.5 %
(Intercept) -7.5581 15.9659
x2          -0.2958  0.8688
```

since all confidence intervals cover zero we cannot reject that the parameters are in fact zero, and we would conclude neither x_1 nor x_2 explain the variations in y.

c) Estimate the parameters for the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim (N(0,\sigma^2), \tag{6-8}$$

and go through the steps of Method 6.16 (use confidence level 0.05 in all tests).

|||| Solution

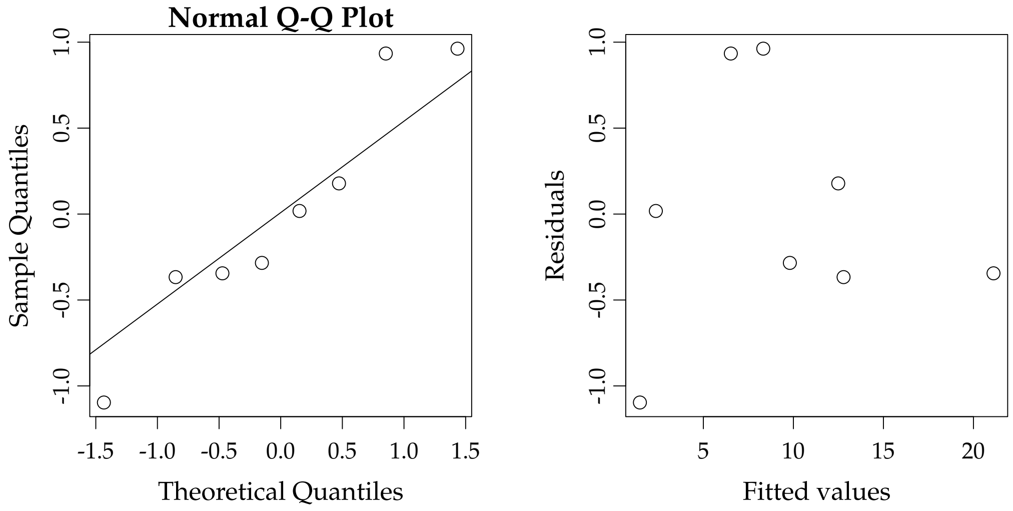
The model is fitted with

```
fit <- lm(y ~ x1 + x2, data=D)
summary(fit)


Call:
lm(formula = y ~ x1 + x2, data = D)

Residuals:
      1       2       3       4       5       6       7       8 
 0.9622  0.1783 -0.3670 -1.0963 -0.3448 -0.2842  0.0178  0.9339 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)   8.0325     0.6728    11.9 0.0000727 ***
x1           -3.5734     0.1955   -18.3 0.0000090 ***
x2            0.9672     0.0489    19.8 0.0000061 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.821 on 5 degrees of freedom
Multiple R-squared:  0.988, Adjusted R-squared:  0.983 
F-statistic:  208 on 2 and 5 DF,  p-value: 0.0000154
```

|||| Solution

Before discussing the parameter let’s have a look at the residuals:

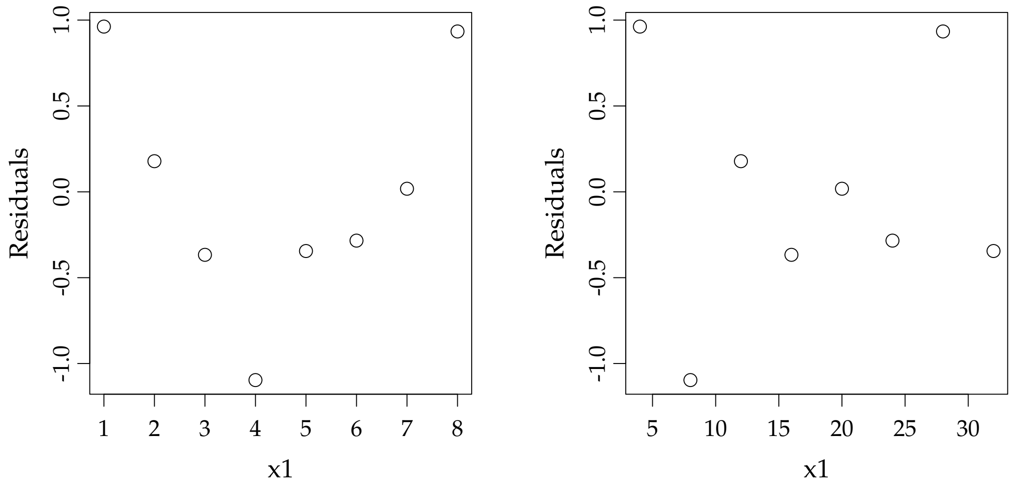
```
par(mfrow=c(1,2))
qqnorm(fit$residuals)
qqline(fit$residuals)
plot(fit$fitted.values, fit$residuals,
     xlab="Fitted values", ylab="Residuals")
```



The are no obvious structures in the residuals as a function of the fitted values and also there does not seem be be serious departure from normality, but lets try to look at the residuals as a function of the independent variables anyway

|||| Solution

```
par(mfrow=c(1,2))
plot(D$x1, fit$residuals, xlab="x1", ylab="Residuals")
plot(D$x2, fit$residuals, xlab="x1", ylab="Residuals")
```



the plot of the residuals as a function of x_1 suggest that there could be a quadratic dependence.

|||| Solution

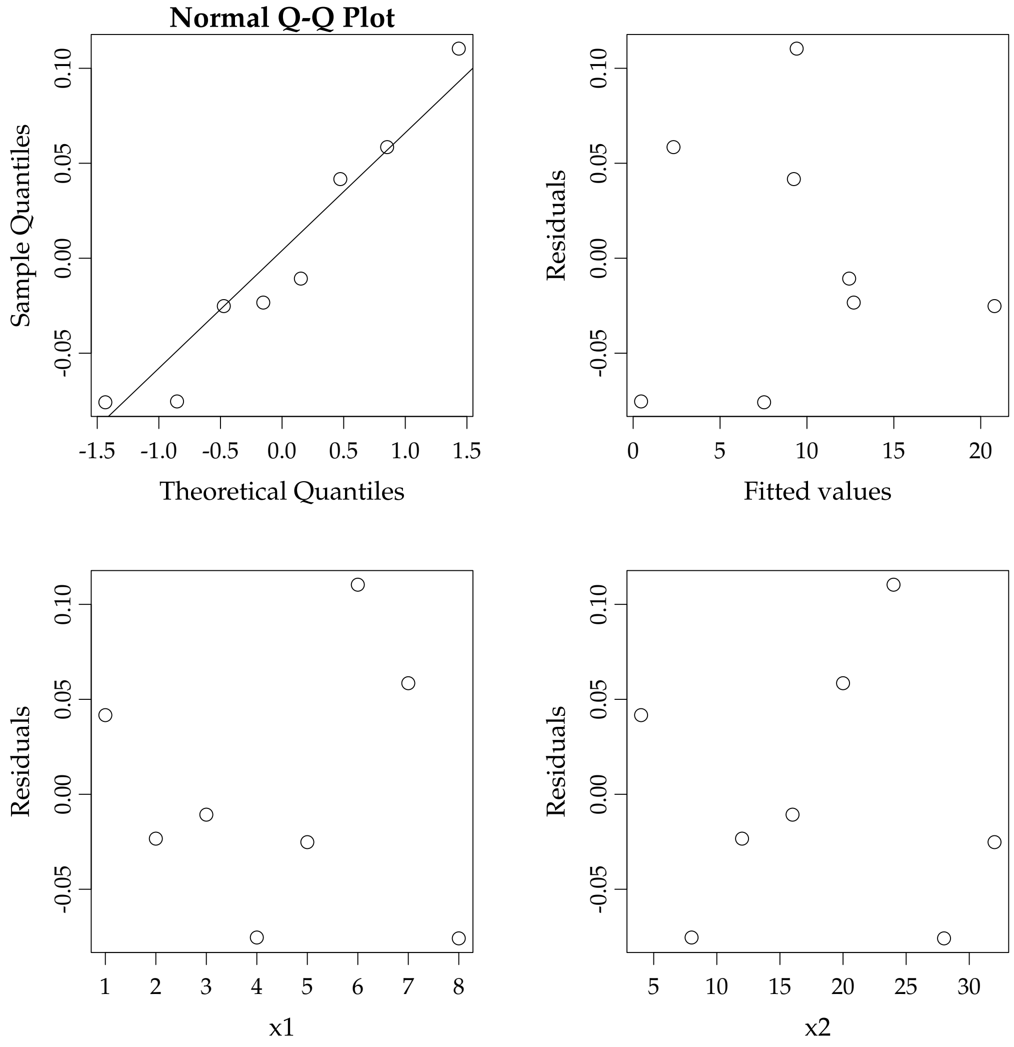
Now include the quadratic dependence of x_1

```
D$x3 <- D$x1^2
fit3 <- lm(y ~ x1 + x2 + x3, data=D)
summary(fit3)


Call:
lm(formula = y ~ x1 + x2 + x3, data = D)

Residuals:
      1       2       3       4       5       6       7       8 
 0.0417 -0.0233 -0.0107 -0.0754 -0.0252  0.1104  0.0585 -0.0758 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.1007     0.1212    83.3  1.2e-07 ***
x1           -5.0024     0.0709   -70.5  2.4e-07 ***
x2            1.0006     0.0054   185.2  5.1e-09 ***
x3            0.1474     0.0070    21.1  3.0e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0867 on 4 degrees of freedom
Multiple R-squared:     1, Adjusted R-squared:     1 
F-statistic: 1.26e+04 on 3 and 4 DF,  p-value: 2.11e-08
```

we can see that all parameters are still significant, and we can do the residual analysis of the resulting model.

|||| Solution

```
   par(mfrow=c(2,2))
   qqnorm(fit3$residuals)
   qqline(fit3$residuals)
   plot(fitted.values(fit3), fit3$residuals,
        xlab="Fitted values", ylab="Residuals")
   plot(D$x1, fit3$residuals, xlab="x1", ylab="Residuals")
   plot(D$x2, fit3$residuals, xlab="x2", ylab="Residuals")
```



There are no obvious structures left and there is no departure from normality, and we can report the finally selected model as

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i}^2 + \varepsilon_i, \quad \varepsilon_i \sim (N(0, \sigma^2),$$

with the parameters estimates given above.

d) Find the standard error for the line, and the confidence and prediction intervals for the line for the points $(\min(x_1), \min(x_2))$, $(\bar{x}_1, \bar{x}_2)$, $(\max(x_1), \max(x_2))$.

|||| Solution

The question is solved by

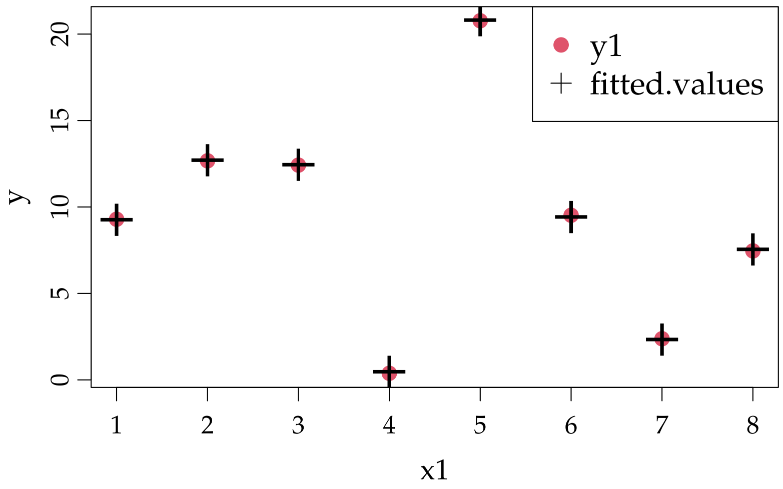
```
## New data
Dnew <- data.frame(x1=c(min(D$x1),mean(D$x1),max(D$x1)),
                   x2=c(min(D$x2),mean(D$x2),max(D$x2)),
                   x3=c(min(D$x1),mean(D$x1),max(D$x1))^2)

## standard error for the line
predict(fit3, newdata=Dnew, se=TRUE)$se

      1       2       3
0.07306 0.04785 0.07985


## Confidence interval
predict(fit3, newdata=Dnew, interval="confidence")

     fit    lwr    upr
1  9.248  9.045  9.451
2  8.587  8.454  8.720
3 11.538 11.317 11.760


## Prediction interval
predict(fit3, newdata=Dnew, interval="prediction")

     fit    lwr    upr
1  9.248  8.934  9.563
2  8.587  8.312  8.862
3 11.538 11.211 11.866
```

e) Plot the observed values together with the fitted values (e.g. as a function of x_1).

|||| Solution

The question is solved by

```
plot(D$x1, D$y, pch=19, col=2, xlab="x1", ylab="y")
points(D$x1, fit3$fitted.values, pch="+", cex=2)
legend("topright", c("y1","fitted.values"), pch=c(19,3), col=c(2,1))
```



Notice that we have an almost perfect fit when including x_1, x_2 and x_1^2 in the model, while neither x_1 nor x_2 alone could predict the outcomes.

Chapter 7

Inference for Proportions

Contents

7.1 Passing proportions

|||| Exercise 7.1 Passing proportions

To compare the level of 2 different courses at a university the following grades distributions (given as number of pupils who achieved the grades) were registered:

	Course 1	Course 2	*Row total*
Grade 12	20	14	34
Grade 10	14	14	28
Grade 7	16	27	43
Grade 4	20	22	42
Grade 2	12	27	39
Grade 0	16	17	33
Grade -3	10	22	32
Column total	108	143	251

The passing proportions for the two courses, p_1 and p_2 should be compared. As the grades -3 and 0 means not passed, we get the following table of the number of students:

	Course 1	Course 2	*Row total*
Passed	82	104	186
Not passed	26	39	65
Column total	108	143	251

a) Compute a 95% confidence interval for the difference between the two passing proportions.

|||| Solution

We use the formula for a (large sample) confidence band for the difference of two proportions (Method 7.15): with $x_1 = 82$, $n_1 = 108$, $x_2 = 104$, $n_2 = 143$ and $\alpha = 0.05$, so

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} = 0.032,$$

and $z_{0.025} = 1.96$. By the way

$$\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1}+\frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}} = \sqrt{\frac{x_1\left(n_1-x_1\right)}{n_1^3}+\frac{x_2\left(n_2-x_2\right)}{n_2^3}}.$$

Therefore the answer is:

$$0.032 \pm 1.96\sqrt{\frac{82\cdot 26}{108^3}+\frac{104\cdot 39}{143^3}} \Leftrightarrow [-0.0768; 0.141].$$

b) What is the critical values for the χ^2-test of the hypothesis $H_0 : p_1 = p_2$ with significance level $\alpha = 0.01$?

|||| Solution

This test has degrees of freedom $(2-1)(2-1) = 1$, with the critical value $\chi^2_{0.99}$, so the correct answer is:

```
qchisq(0.99, 1)

[1] 6.635
```

c) If the passing proportion for a course given repeatedly is assumed to be 0.80 on average, and there are 250 students who are taking the exam each time, what is the expected value, μ and standard deviation, σ, for the number of students who do not pass the exam for a randomly selected course?

|||| Solution

If X is the number of students not passing a randomly selected course, this random variable follows the binomial distribution with $n = 250$ and $p = 0.20$, so we use the formulas from Theorem 2.21 for the mean and variance of the binomial distribution

$$\mu = np = 0.2 \cdot 250 = 50, \ \ \sigma^2 = np(1-p) = 250 \cdot 0.2 \cdot 0.8 = 40 = 6.32^2.$$

Thus the answer is $\mu = 50$ and $\sigma = 6.32$.

7.2 Outdoor lightning

|||| Exercise 7.2 Outdoor lighting

A company that sells outdoor lighting, gets a lamp produced in 3 material variations: in copper, with painted surface and with stainless steel. The lamps are sold partly in Denmark and partly for export. For 250 lamps the distribution of sales between the three variants and Denmark/export are depicted. The data is shown in the following table:

	Country	
	Danmark	Export
Copper variant	7.2%	6.4%
Painted variant	28.0%	34.8%
Stainless steel variant	8.8%	14.8%

a) Is there a significant difference between the proportion exported and the proportion sold in Denmark (with $\alpha = 0.05$)?

|||| Solution

The situation asked about here is a "one sample proportion" case, where 110 (44% out of 250) are sold in Denmark (and hence 250-110=140 for export). Using Method 7.11 the standard statistic for the hypothesis test $H_0 : p = 0.5$, is

$$z_{\text{obs}} = \frac{140 - 250 \cdot 0.5}{\sqrt{250 \cdot 0.5 \cdot 0.5}},$$

the critical values are $\pm z_{0.975} = \pm 1.96$.

So the correct answer is: no, since $15/\sqrt{250/4} = 1.90$ is within ± 1.96.

b) The relevant critical value to use for testing whether there is a significant difference in how the sold variants are distributed in Denmark and for export is (with $\alpha = 0.05$)?

|||| Solution

This is a so-called null hypothesis of homogeneity in a 3×2 frequency table ($r \times c$ table, see Method 7.22). The critical value for the χ^2-test is based on the χ^2 distribution with $(r-1)(c-1) = 2$ degrees of freedom. Hence the correct answer is: $\chi^2_{0.95}(2) = 5.991$.

7.3 Local elections

|||| Exercise 7.3 Local elections

At the local elections in Denmark in November 2013 the Social Democrats (A) had $p = 29.5\%$ of the votes at the country level. From an early so-called exit poll it was estimated that they would only get 22.7% of the votes. Suppose the exit poll was based on 740 people out of which then 168 people reported having voted for A.

a) At the time of the exit poll the p was of course not known. If the following hypothesis was tested based on the exit poll

$$H_0 : p = 0.295$$
$$H_1 : p \neq 0.295,$$

what test statistic and conclusion would then be obtained with $\alpha = 0.001$?

|||| Solution

The one-proportions test statistic found in Method 7.11 is

$$z_{\text{obs}} = \frac{168 - 740 \cdot 0.295}{\sqrt{740 \cdot 0.295 \cdot (1 - 0.295)}} = -4.05,$$

and the critical values are $\pm z_{0.9995} = \pm 3.291$ (in R: `qnorm(0.9995)`). So the correct answer is:

Test statistic: -4.05. Conclusion: we reject the null hypothesis, since $-4.05 < -z_{0.9995} = -3.291$.

b) Calculate a 95%-confidence interval for p based on the exit poll.

|||| Solution

We use Method 7.3

$$\hat{p} = \frac{x}{n} = \frac{168}{740} = 0.227, \tag{7-1}$$

and

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.227 \pm 1.96 \cdot \sqrt{\frac{0.227 \cdot 0.773}{740}}.$$

Which we calculate in R by

```
0.227 + c(-1, 1)*1.96*sqrt(0.227*0.773/740)

[1] 0.1968 0.2572
```

c) Based on a scenario that the proportion voting for particular party is around 30%, how large an exit poll should be taken to achieve a 99% confidence interval having a width of 0.01 in average for this proportion?

|||| Solution

The proportion sample size formula, Method 7.13, using a guess of $p = 0.3$ is

$$0.3 \cdot 0.7 \cdot (z_{0.995}/ME)^2$$

where ME is the marginal error and since the confidence interval is plus/minus the marginal error, we should take $ME = 0.01/2$ and $z_{0.995} = 2.576$ (in R: `qnorm(0.995)`).

So, rounding up to nearest integer, the correct answer is:
$0.3 \cdot 0.7 \cdot (2.576/(0.01/2))^2 \approx 55741$ persons.

7.4 Sugar quality

|||| Exercise 7.4 Sugar quality

A wholesaler needs to find a supplier that delivers sugar in 1 kg bags. From two potential suppliers 50 bags of sugar are received from each. A bag is described as 'defective' if the weight of the filled bag is less than 990 grams. The received bags were all control weighed and 6 defective from supplier A and 12 defective from supplier B were found.

a) If the following hypothesis

$$H_0 : p_{\text{A}} = p_{\text{B}},$$
$$H_1 : p_{\text{A}} \neq p_{\text{B}}.$$

is tested on a significance level of 5%, what is the p-value and conclusion?

|||| Solution

With

$$\hat{p}_{\text{A}} = \frac{6}{50} = 0.12, \quad \hat{p}_{\text{B}} = \frac{12}{50} = 0.24,$$

and the common

$$\hat{p} = \frac{6+12}{50+50} = 0.18.$$

Using the z-test for comparing two proportions Method 7.18 the hypothesis can be tested by

$$z_{\text{obs}} = \frac{\hat{p}_{\text{B}} - \hat{p}_{\text{A}}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{\text{B}}} + \frac{1}{n_{\text{B}}}\right)}} = \frac{0.24 - 0.12}{\sqrt{0.18 \cdot 0.82 \cdot (2/50)}} = 1.5617.$$

Finding the probability of observing this or more extreme that for a standard normal (in R: `1 - pnorm(1.5617)`) is

$$P(Z > 1.5617) = 0.0592,$$

thus the p-value becomes 0.118. This leads to the conclusion that the null hypothesis cannot be rejected, since the p-value is above the significance level $p\text{-value} = 0.118 > 0.05 = \alpha$.

Similarly one could have performed a 2-by-2 table χ^2-test that would give $\chi^2_{\text{obs}} = Z^2 = 2.439$ and the p-value found using the $\chi^2(1)$-distribution, or simply in R run any of the two calls:

```
prop.test(x=c(6,12), n=c(50,50), correct = FALSE)


2-sample test for equality of proportions without continuity
correction

data:  c(6, 12) out of c(50, 50)
X-squared = 2.4, df = 1, p-value = 0.1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.26875  0.02875
sample estimates:
prop 1 prop 2
  0.12   0.24


chisq.test(matrix(c(6,12,44,38), ncol = 2), correct = FALSE)


Pearson's Chi-squared test

data:  matrix(c(6, 12, 44, 38), ncol = 2)
X-squared = 2.4, df = 1, p-value = 0.1
```

b) A supplier has delivered 200 bags, of which 36 were defective. A 99% confidence interval for p the proportion of defective bags for this supplier is:

|||| Solution

The large sample proportion confidence interval formula in Method 7.3 is used

$$0.18 \pm z_{0.995} \cdot \sqrt{\frac{0.18 \cdot 0.82}{200}},$$

thus (in R: `qnorm(0.995)`)

$$0.18 \pm 2.576 \cdot \sqrt{\frac{0.18 \cdot 0.82}{200}}$$

So the answer becomes:

```
0.18 + c(-1, 1)*2.576*sqrt(0.18*0.82/200)

[1] 0.11 0.25
```

```
prop.test(x=36, n=200, correct=FALSE, conf.level=0.99)


1-sample proportions test without continuity correction

data:  36 out of 200, null probability 0.5
X-squared = 82, df = 1, p-value <2e-16
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.1207 0.2599
sample estimates:
   p
0.18
```

c) Based on the scenario, that the proportion of defective bags for a new supplier is about 20%, a new study was planned with the aim of obtaining an average width, B, of a 95% confidence interval. The Analysis Department achieved the result that one should examine 1537 bags, but had forgotten to specify which value for the width B, they had used. What was the value used for B?

|||| Solution

Method 7.13 holds the relevant sample size formular

$$n = 0.2 \cdot 0.8 \cdot \left[\frac{z_{0.975}}{ME}\right]^2$$

which when solved for ME becomes (and plugging in $n = 1537$)

$$ME = \sqrt{0.2 \cdot 0.8 \cdot \frac{z_{0.975}^2}{1537}} = 0.020$$

And since the width of the confidence interval is twice the margin of error, the answer is: $B = 0.040$.

7.5 Physical training

|||| Exercise 7.5 Physical training

A company wants to investigate whether the employees' physical training condition will affect their success in the job. 200 employees were tested and the following count data were found:

	Physical training condition		
	Below average	Average	Above average
Bad job succes	11	27	15
Average job succes	14	40	30
Good job succes	5	23	35

The hypothesis of independence between job success and physical training condition is to be tested by the use of the for this setup usual χ^2-test.

a) What is the expected number of individuals with above average training condition and good job success under H_0 (i.e. if H_0 is assumed to be true)?

|||| Solution

The expected number under the null hypothesis for each cell is found as

$$\text{"column total"} \cdot \frac{\text{"row total"}}{\text{"total"}},$$

for table cell (3,3), which is asked about, so the answer is

$$e_{33} = 80 \cdot \frac{63}{200} = 25.2.$$

b) For the calculation of the relevant χ^2-test statistic, identify the following two numbers:

– A: the number of contributions to the test statistic

– B: the contribution to the statistic from table cell (1,1)

|||| Solution

Since the contingency table is a 3-by-3 table the number of contributions to the test statistic is 9 (one for each cell), and the (1,1) contribution is

$$\frac{(o_{11} - e_{11})^2}{e_{11}},$$

where $o_{11} = 11$ and

$$e_{11} = \frac{30 \cdot 53}{200} = 7.95.$$

Hence

$$\frac{(o_{11} - e_{11})^2}{e_{11}} = \frac{(11 - 7.95)^2}{7.95} = 1.170.$$

So the answer becomes: A is 9 and B is 1.170.

c) The total χ^2-test statistic is 10.985, so the p-value and the conclusion will be (both must be valid):

|||| Solution

The p-value is found by the $\chi^2(4)$-distribution (degrees of freedom is $(r - 1) \cdot (c - 1) = (3 - 1) \cdot (3 - 1)$)

$$P(\chi^2 > 10.985) = 0.027,$$

which can be found in R by: `1-pchisq(10.9849,4)`.

So the answer becomes:

The p-value $= 0.027 < 0.05 = \alpha$ and therefore H_0 is rejected on a 5% level, and thus on this significance level there is a significant dependence between job success and physical training condition.

Chapter 8

Comparing means of multiple groups - ANOVA (solutions to exercise)

Contents

8.1 Environment action plans

|||| Exercise 8.1 Environment action plans

To investigate the effect of two recent national Danish aquatic environment action plans the concentration of nitrogen (measured in g/m^3) have been measured in a particular river just before the national action plans were enforced (1998 and 2003) and in 2011. Each measurement is repeated 6 times during a short stretch of river. The result is shown in the following table:

	N_{1998}	N_{2003}	N_{2011}
	5.01	5.59	3.02
	6.23	5.13	4.76
	5.98	5.33	3.46
	5.31	4.65	4.12
	5.13	5.52	4.51
	5.65	4.92	4.42
Row mean	5.5517	5.1900	4.0483

Further, the total variation in the data is $SST = 11.4944$. You got the following output from `R` corresponding to a one-way analysis of variance (where most of the information, however, is replaced by the letters A-E as well as U and V):

```
> anova(lm(N ~ Year))
Analysis of Variance Table

Response: N
          Df SumSq MeanSq Fvalue    Pr(>F)
Year       A   B      C      U         V
Residuals  D 4.1060   E
```

a) What numbers did the letters A-D substitute?

Solution

One should check the structure of the oneway ANOVA table, so A and D are the degrees of freedom, A $= k - 1 = 3 - 1 = 2$ and D $= n - k = 18 - 3 = 15$. And B is the treatment sum-of-squares

$$SS(Tr) = SST - SSE = 11.4944 - 4.1060 = 7.3884,$$

And finally, C is the $MS(Tr)$-value

$$MS(Tr) = SS(Tr)/2 = 7.3884/2 = 3.6942.$$

b) If you use the significance level $\alpha = 0.05$, what critical value should be used for the hypothesis test carried out in the analysis (and in the table illustrated with the figures U and V)?

Solution

The relevant distribution for testing effects in ANOVA is the F-distribution, here with degrees of freedom $k - 1 = 2$ and $n - k = 15$. So,

$$F_{0.05}(2, 15) = 3.682,$$

found in R as:

```
qf(p=0.95, df1=2, df2=15)

[1] 3.682
```

c) Can you with these data demonstrate statistically significant (at significance level $\alpha = 0.05$) differences in N-mean values from year to year (both conclusion and argument must be valid)?

|||| Solution

U is the F-statistic

$$F = \frac{C}{E} = \frac{(11.4944 - 4.1060)/2}{4.1060/15} = 13.496,$$

and V is the p-value (using the F(2,15)-distribution)

$$P(F > 13.496) = 0.00044,$$

Or in R

```
1-pf(q=13.496, df1=2, df2=15)

[1] 0.0004435
```

So, the answer is, yes, as the number V is less than 0.05.

d) Compute the 90% confidence interval for the single mean difference between year 2011 and year 1998.

|||| Solution

We use the formula for a single pre-planned pairwise post hoc confidence intervals

$$4.0483 - 5.5517 \pm t_{0.05}(15)\sqrt{MSE \cdot (1/6 + 1/6)},$$

$$-1.50 \pm 1.753\sqrt{4.1060/15 \cdot (1/3)}.$$

In R:

```
-1.50+c(-1,1)* 1.753*sqrt(4.1060/15*(1/3))

[1] -2.0295 -0.9705
```

8.2 Environment action plans (part 2)

|||| Exercise 8.2 Environment action plans (part 2)

This exercise is using the same data as the previous exercise, but let us repeat the description here. To investigate the effect of two recent national Danish aquatic environment action plans the concentration of nitrogen (measured in g/m^3) have been measured in a particular river just before the national action plans were enforced (1998 and 2003) and in 2011. Each measurement is repeated 6 times during a short stretch of river. The result is shown in the following table, where we have now added also the variance computed within each group.

	N_{1998}	N_{2003}	N_{2011}
	5.01	5.59	3.02
	6.23	5.13	4.76
	5.98	5.33	3.46
	5.31	4.65	4.12
	5.13	5.52	4.51
	5.65	4.92	4.42
Row means	5.5517	5.1900	4.0483
Row variances	0.2365767	0.1313200	0.4532967

The data can be read into R and the means and variances computed by the following in R:

```
nitrogen <- c(5.01, 5.59, 3.02,
              6.23, 5.13, 4.76,
              5.98, 5.33, 3.46,
              5.31, 4.65, 4.12,
              5.13, 5.52, 4.51,
              5.65, 4.92, 4.42)
year <- factor(rep(c("1998", "2003", "2011"), 6))
tapply(nitrogen, year, mean)

 1998  2003  2011 
5.552 5.190 4.048 


tapply(nitrogen, year, var)

  1998   2003   2011 
0.2366 0.1313 0.4533 
```

```
mean(nitrogen)

[1] 4.93
```

a) Compute the three sums of squares (SST, $SS(Tr)$ and SSE) using the three means and three variances, and the overall mean (show the formulas explicitly).

|||| Solution

The treatment sum-of-squares $SS(Tr)$ can (Theorem 8.2 Equation 8-6) be computed from the three means as

$$\begin{aligned} SS(Tr) &= \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 \\ &= 6 \cdot (5.551667 - 4.93)^2 + 6 \cdot (5.190000 - 4.93)^2 + 6 \cdot (4.048333 - 4.93)^2 \\ &= 7.388439. \end{aligned}$$

The residual error sum-of-squares SSE (see Theorem 8.2) is defined by by

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

So we see that the inner part is the total variance of each group (using Equation 8-15 in Theorem 8.4)

$$SSE = \sum_{i=1}^{k}(n_i - 1)s_i$$

and now we can insert the values

$$\begin{aligned} SSE &= 5s_1^2 + 5s_2^2 + 5s_3^2 = 5 \cdot 0.2365767 + 5 \cdot 0.1313200 + 5 \cdot 0.4532967 \\ &= 4.105967. \end{aligned}$$

Finally, then (not that we would need this in real data analysis when we have the other two)

$$SST = SS(Tr) + SSE = 7.388439 + 4.105967 = 11.49441.$$

b) Find the SST-value in R using the sample variance function `var`.

|||| Solution

The *SST*-value is ”almost” just the variance of the observations ignoring the group information, or rather, it is the numerator of this variance calculation, so: $n - 1 = 17$ times the variance will be *SST* (cf. Theorem 8.4):

```
D <- data.frame(
  nitrogen=c(5.01, 5.59, 3.02,
             6.23, 5.13, 4.76,
             5.98, 5.33, 3.46,
             5.31, 4.65, 4.12,
             5.13, 5.52, 4.51,
             5.65, 4.92, 4.42),
  year=factor(rep(c("1998", "2003", "2011"), 6))
)
tapply(D$nitrogen, D$year, mean)

 1998  2003  2011
5.552 5.190 4.048


tapply(D$nitrogen, D$year, var)

  1998   2003   2011
0.2366 0.1313 0.4533


mean(D$nitrogen)

[1] 4.93


17 * var(D$nitrogen)

[1] 11.49
```

c) Run the ANOVA in R and produce the ANOVA table in R.

|||| Solution

It may be done as follows:

```
fit <- lm(nitrogen ~ year, data=D)
anova(fit)

Analysis of Variance Table

Response: nitrogen
          Df Sum Sq Mean Sq F value  Pr(>F)
year       2   7.39    3.69    13.5 0.00044 ***
Residuals 15   4.11    0.27
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

d) Do a complete post hoc analysis, where all the 3 years are compared pairwise.

|||| Solution

We want to construct the $M = 3 \cdot 2/2 = 3$ different confidence intervals using Method 8.9. As all n_is equal 6 in this case, all 3 confidence intervals will have the same width, and we can use Remark 8.13 and compute the (half) width of the confidence intervals, the LSD-value. And since there are 3 multiple comparisons we will use $\alpha_{\text{Bonferroni}} = 0.05/3 = 0.01667$

$$LSD_{0.01667} = t_{1-(0.05/3)/2} \cdot \sqrt{2 \cdot 0.2737/6} = 0.8136.$$

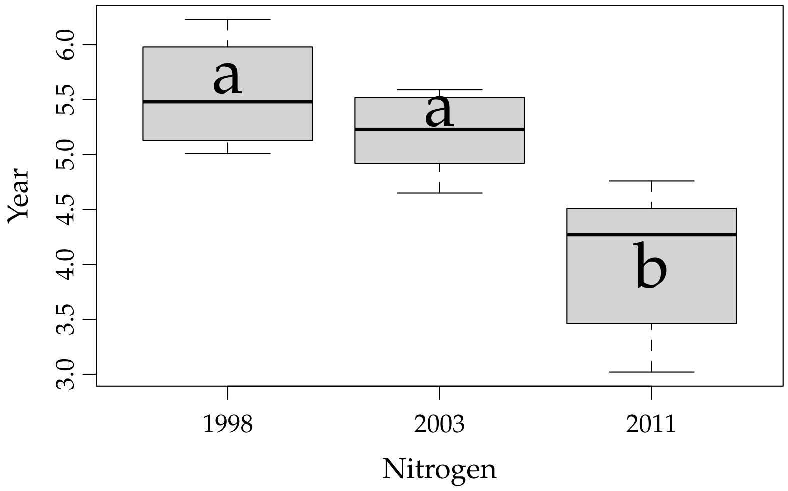
```
LSD_0.01667 <- qt(1-(0.05/3)/2, 15) * sqrt(2*0.2737/6)
LSD_0.01667

[1] 0.8136
```

So, if we again study the three group means, we can see that the nitrogen level in 2011 is significantly smaller than in 2003 and 1998, whereas the level in 1998 and 2003 are not significantly different.

|||| Solution

The differences could also be shown in the following plot, where the compact letter (see page 368 of Chapter 8) display way of telling the story has been added to the box plot:

```
plot(D$year, D$nitrogen, xlab="Nitrogen", ylab="Year")
text(1:3, c(5.7, 5.4, 4), c("a", "a", "b"), cex=2)
```



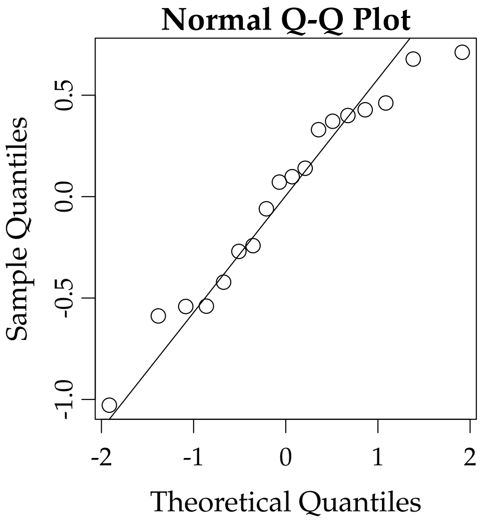
Hence, the groups are sorted from largest sample mean to lowest, and then the groups (here years) which are not significantly different share letters.

e) Use R to do model validation by residual analysis.

|||| Solution

The box plot does not indicate clear variance differences (although it can be a bit difficult to know exactly how different such patterns should be for it to be a problem. Let us check for the normality by doing a normal q-q plot on the residuals:

```
qqnorm(fit$residuals)
qqline(fit$residuals)
```



There appears to be no important deviation from normality. For more detailed investigations, see 8.17.

8.3 Plastic film

|||| Exercise 8.3 Plastic film

A company is starting a production of a new type of patch. For the product a thin plastic film is to be used. Samples of products were received from 5 possible suppliers. Each sample consisted of 20 measurements of the film thickness and the following data were found:

	Average film thickness $\bar{x}$ in μm	Sample standard deviation s in μm
Supplier 1	31.4	1.9
Supplier 2	30.6	1.6
Supplier 3	30.5	2.2
Supplier 4	31.3	1.8
Supplier 5	29.2	2.2

From the usual calculations for a one-way analysis of variance the following is obtained:

Source	Degrees of freedom	Sums of Squares
Supplier	4	$SS(Tr) = 62$
Error	95	$SSE = 362.71$

a) Is there a significant ($\alpha = 5\%$) difference between the mean film thicknesses for the suppliers (both conclusion and argument must be correct)?

|||| Solution

The F-test statistics for one-way ANOVA is

$$F_{\text{obs}} = \frac{MS(Tr)}{MSE} = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} == \frac{62/4}{362.71/95} = 4.06,$$

and the relevant critical value is $F_{0.05}(4,95)$ to be found in R by: `qf(p=0.95, df1=4, df2=95)`. So the answer is:

Yes, the null hypothesis is rejected, since $Fobs = 4.06$ is larger than the critical value 2.47.

Or we could find the p-value:

```
1-pf(4.06, 4, 95)
```

```
[1] 0.004406
```

and conclude that it is this is so small, we have strong evidence against the null hypothesis.

b) Compute a 95% confidence interval for the difference in mean film thicknesses of Supplier 1 and Supplier 4 (considered as a “single pre-planned” comparison).

Solution

The "ANOVA post hoc" confidence interval is to be used

$$31.4 - 31.3 \pm t_{0.975}(95)\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where $MSE = \frac{SSE}{n-k} = \frac{362.71}{95}$, so since $t_{0.975}(95)$ - to be found in R as: `qt(0.975,95)`, it becomes

$$0.1 \pm 1.985\sqrt{\frac{362.71}{95}\left(\frac{1}{20} + \frac{1}{20}\right)}.$$

```
0.1+ c(-1,1)* qt(0.975, 95)*sqrt(362.71/(95*10))

[1] -1.127  1.327
```

So the answer is

$$0.1 \pm 1.985\sqrt{\frac{362.71}{95}\left(\frac{1}{10}\right)} = [-1.2, 1.3].$$

8.4 Brass alloys

|||| Exercise 8.4 Brass alloys

When brass is used in a production, the modulus of elasticity, E, of the material is often important for the functionality. The modulus of elasticity for 6 different brass alloys are measured. 5 samples from each alloy are tested. The results are shown in the table below where the measured modulus of elasticity is given in GPa:

Brass alloys					
M1	M2	M3	M4	M5	M6
82.5	82.7	92.2	96.5	88.9	75.6
83.7	81.9	106.8	93.8	89.2	78.1
80.9	78.9	104.6	92.1	94.2	92.2
95.2	83.6	94.5	87.4	91.4	87.3
80.8	78.6	100.7	89.6	90.1	83.8

In an R-run for oneway analysis of variance:

```
anova( lm(elasmodul ~ alloy) )
```

the following output is obtained: (however some of the values have been substituted by the symbols A, B, and C)

```
> anova( lm(elasmodul ~ alloy) )
Analysis of Variance Table

Response: elasmodul
Df  Sum Sq Mean Sq F value    Pr(>F)
alloy      A 1192.51 238.501  9.9446 3.007e-05
Residuals  B    C     23.983
```

a) What are the values of A, B, and C?

|||| Solution

The A and B are the degrees of freedom, which in the oneway ANOVA is $k-1$ and $n-k$, where $k=6$ is the number of groups and $n=30$ is the number of observations. C can be found by

$$C = SSE = MSE \cdot (n-k) = 23.983 \cdot 24 = 575.59$$

So the answer is:

$A=5$, $B=24$ and $C=575.59$.

b) The assumptions for using the one-way analysis of variance is (choose the answer that lists all the assumptions and that NOT lists any unnecessary assumptions):

1) The data must be normally and independently distributed within each group and the variances within each group should not differ significantly from each other

2) The data must be normally and independently distributed within each group

3) The data must be normally and independently distributed and have approximately the same mean and variance within each group

4) The data should not bee too large or too small

5) The data must be normally and independently distributed within each group and have approximately the same IQR-value in each group

|||| Solution

It is difficult to make a lot of arguments here, but simply emphasize that only in Answer 1 all assumptions needed, and no unnecessary assumptions, are listed.

c) Compute a 95% confidence interval for the single pre-planned difference between brass alloy 1 and 2.

Solution

A pre-planned post hoc 95% confidence interval between two groups in a one-way ANOVA is

$$\bar{y}_1 - \bar{y}_2 \pm t_{0.975}\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

So we have to compute the means of the M1 and M2 groups

$$\bar{y}_1 = 84.62, \quad \bar{y}_2 = 81.14,$$

(=3.48) and then plug in $MSE = 23.983$ and $n_1 = n_2 = 5$.

```
3.48+ c(-1,1)* qt(0.975, 24)*sqrt(23.983*2/5)

[1] -2.912  9.872
```

So the answer is:

$$3.48 \pm t_{0.025}\sqrt{23.983\left(\frac{2}{5}\right)} = [2.91, 9.87].$$

8.5 Plastic tubes

|||| Exercise 8.5 Plastic tubes

Some plastic tubes for which the tensile strength is essential are to be produced. Hence, sample tube items are produced and tested, where the tensile strength is determined. Two different granules and four possible suppliers are used in the trial. The measurement results (in MPa) from the trial are listed in the table below:

	Granule	
	g1	g2
Supplier a	34.2	33.1
Supplier b	34.8	31.2
Supplier c	31.3	30.2
Supplier d	31.9	31.6

The following is run in R:

```
D <- data.frame(
  strength=c(34.2,34.8,31.3,31.9,33.1,31.2,30.2,31.6),
  supplier=factor(c("a","b","c","d","a","b","c","d")),
  granule=factor(c(1,1,1,1,2,2,2,2))
)
anova(lm( strength ~ supplier + granule, data=D))
```

with the following result:

```
Analysis of Variance Table

Response: strength
           Df  Sum Sq Mean Sq F value Pr(>F)
supplier   3  10.0338 3.3446  3.2537  0.1792
granule    1   4.6512 4.6512  4.5249  0.1233
Residuals  3   3.0837 1.0279
```

a) Which distribution has been used to find the p-value 0.1792?

Solution

The p-value is from the F-test from a two-way ANOVA using the $F(3,3)$-distribution.

Hence the correct answer is:

The F-distribution with the degrees of freedom $\nu_1 = 3$ and $\nu_2 = 3$.

b) What is the most correct conclusion based on the analysis among the following options (use $\alpha = 0.05$)?

1) A significant difference has been found between the variances from the analysis of variance

2) A significant difference has been found between the means for the 2 granules but not for the 4 suppliers

3) No significant difference has been found between the means for neither the 4 suppliers nor the 2 granules

4) A significant difference has been found between the means for as well the 4 suppliers as the 2 granules

5) A significant difference has been found between the means for the 4 suppliers but not for the 2 granules

|||| Solution

Since both of the p-values are larger than 0.05 none of the two usual hypothesis tests (of no group difference) are significant. So the correct answer is:

3) No significant difference has been found between the means for neither the 4 suppliers nor the 2 granules

8.6 Joining methods

Exercise 8.6 Joining methods

To compare alternative joining methods and materials a series of experiments are now performed where three different joining methods and four different choices of materials are compared.

Data from the experiment are shown in the table below:

	Material				Row
Joining methods	1	2	3	4	average
A	242	214	254	248	239.50
B	248	214	248	247	239.25
C	236	211	245	243	233.75
Column average	242	213	249	246	

In an R-run for two-way analysis of variance:

```
Strength <- c(242,214,254,248,248,214,248,247,236,211,245,243)
Joiningmethod <- factor(c("A","A","A","A","B","B","B","B","C","C","C","C"))
Material <- factor(c(1,2,3,4,1,2,3,4,1,2,3,4))
anova(lm(Strength ~ Joiningmethod + Material))
```

the following output is generated (where some of the values are replaced by the symbols A, B, C, D, E and F):

```
Analysis of Variance Table

Response: Strength
                      Df  Sum Sq Mean Sq  F value Pr(>F)
Joiningmethod          A    84.5       B  C       0.05041 .
Material               D       E  825.00  F       1.637e-05 ***
Residuals              6    49.5    8.25
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

a) What are the values for A, B and C?

|||| Solution

A is the degrees of freedom for `Joiningmethod`, wich is the number of groups/levels minus 1: A=2 (and then actually the question can allready be answered). BUT also:

$$\mathrm{B} = MS(\text{Joiningmethod}) = 84.5/2 = 84.5/2 = 42.25,$$

and

$$\mathrm{C} = \frac{MS(\text{Joiningmethod})}{MSE} = \frac{42.25}{8.25} = 5.12.$$

So the answer becomes:

A=2, B=42.25 and C=5.12.

b) What are the conclusions concerning the importance of the two factors in the experiment (using the usual level $\alpha = 5\%$)?

|||| Solution

We can read off the answer from the two p-values given in the output - one of them is below α (Material p-value) and one is NOT (Joiningmethod p-value).

So the answer is:

Significant differences between materials can be concluded, but not between joining methods.

c) Do post hoc analysis for as well the Materials as Joining methods (Confidence intervals for pairwise differences and/or hypothesis tests for those differences).

Solution

First we find the treatment and block means (and we print the ANOVA table):

```
options(digits=2)
Strength <- c(242,214,254,248,248,214,248,247,236,211,245,243)
Joiningmethod <- factor(c("A","A","A","A","B","B","B","B","C","C","C","C"))
Material <- factor(c(1,2,3,4,1,2,3,4,1,2,3,4))

fit <- lm(Strength ~ Joiningmethod + Material)
anova(fit)

Analysis of Variance Table

Response: Strength
              Df Sum Sq Mean Sq F value   Pr(>F)
Joiningmethod  2     85      42    5.12     0.05 .
Material       3   2475     825  100.00 0.000016 ***
Residuals      6     49       8
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1


tapply(Strength, Joiningmethod, mean)

  A   B   C
240 239 234


tapply(Strength, Material , mean)

  1   2   3   4
242 213 249 246
```

We can find the 0.05/3 (Bonferroni-corrected) *LSD*-value from the two-way version of Remark Remark 8.13 (see Section 8.3.3) for the comparison of the 3 `Joiningmethods`:

```
LSD_bonf <- qt(1-0.05/6, 6) * sqrt(2*8.25/4)
LSD_bonf

[1] 6.7
```

We see that none of the three Joining methods are different from each other (although close), which matches fine with the p-value just above 0.05.

And we then do the same for the 4 Materials (that is, 6 pairwise comparisons): We can find the 0.05/6 (Bonferroni-corrected) LSD-value from the two-way version of Remark 8.13 for the comparison of the 4 Materials:

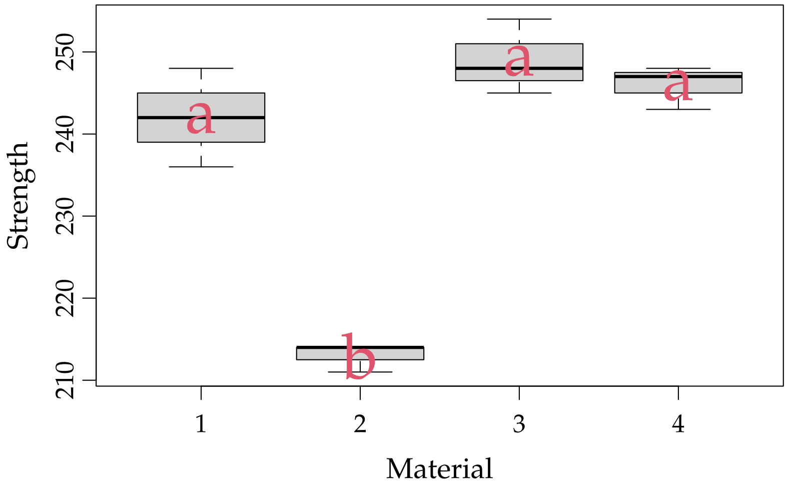
```
LSD_bonf <- qt(1-0.05/12, 6) * sqrt(2*8.25/3)
LSD_bonf
```

```
[1] 9.1
```

|||| Solution

So we see that Material 2 is significantly smaller than each of the other three but none of these 3 are significantly different from each other:

```
plot(Strength ~ Material)
text(1:4, c(242, 213, 249, 246), c("a", "b", "a", "a"), cex=2, col=2)
```



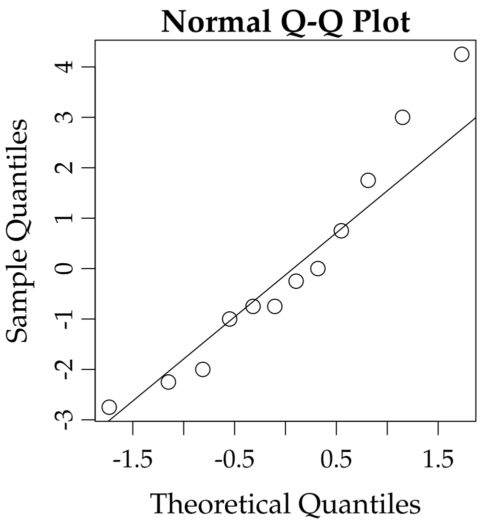
d) Do residual analysis to check for the assumptions of the model:

1. Normality
2. Variance homogeneity

|||| Solution

First the residual normality plot:

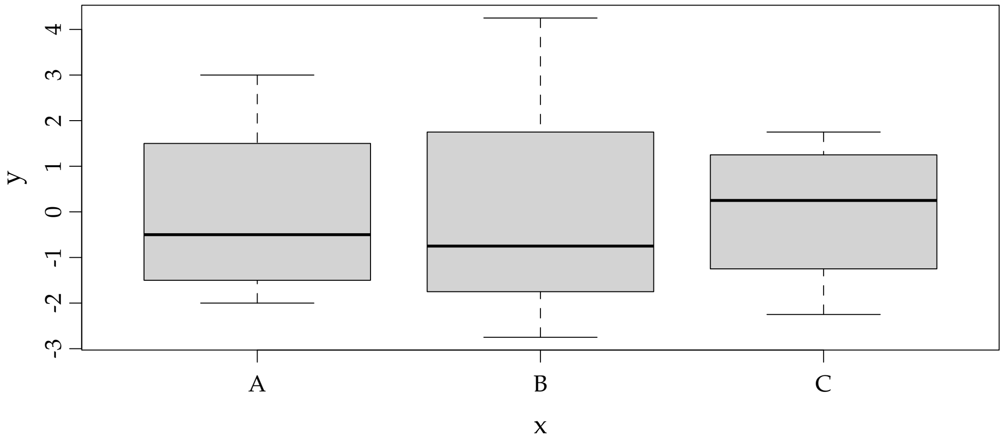
```
qqnorm(fit$residuals)
qqline(fit$residuals)
```



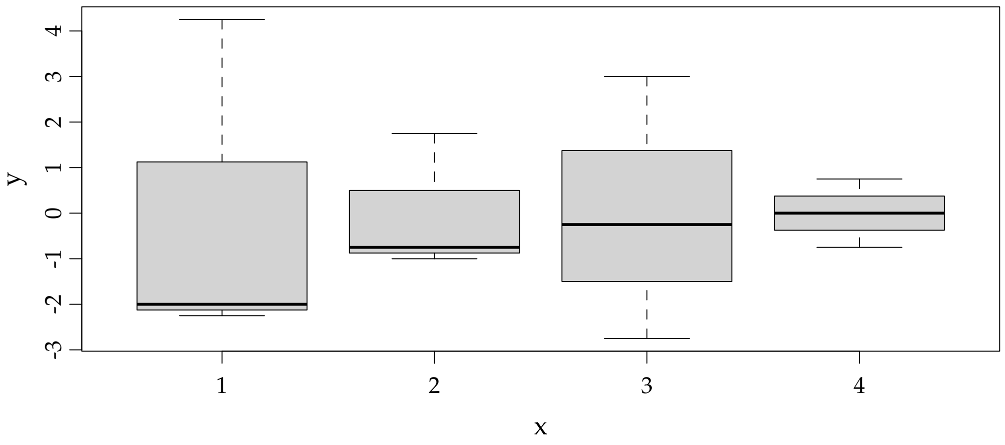
|||| Solution

Then the investigation of variance homogeneity:

```
plot(Joiningmethod, fit$residuals)
```



```
plot(Material, fit$residuals)
```



There may some indications of lower variability within Materials 2 and 4 compared to 1 and 3 (We do not, however, have the methodology (i.e. a test of difference in variance) in the course to deal with this).

8.7 Remoulade

|||| Exercise 8.7 Remoulade

A supermarket has just opened a delicacy department wanting to make its own homemade “remoulade” (a Danish delicacy consisting of a certain mixture of pickles and dressing). In order to find the best recipe a taste experiment was conducted. 4 different kinds of dressing and 3 different types of pickles were used in the test. Taste evaluation of the individual “remoulade” versions were carried out on a continuous scale from 0 to 5.

The following measurement data were found:

	Dressing type				Row
Pickles type	A	B	C	D	average
I	4.0	3.0	3.8	2.4	3.30
II	4.3	3.1	3.3	1.9	3.15
III	3.9	2.3	3.0	2.4	2.90
Column average	4.06	2.80	3.36	2.23	

In an R-run for twoway ANOVA:

```
anova(lm(Taste ~ Pickles + Dressing))
```

the following output is obtained, where some of the values have been substituted by the symbols A, B, C, D, E and F):

```
anova(lm(Taste ~ Pickles + Dressing))
Analysis of Variance Table

Response: Taste
          Df  Sum Sq  Mean Sq  F value  Pr(F)
Pickles    A  0.3267  0.16333     E     0.287133
Dressing   B  5.5367  1.84556     F     0.002273
Residuals  C    D     0.10556
```

a) What are the values of A, B, and C?

|||| Solution

From the general definition of the two-way ANOVA table (see page 356 of Chapter 8) the degrees of freedom are $k-1$, $l-1$ and $(k-1)(l-1)$, where $k=3$ is the number of rows, $l=4$ is the number of columns.

So the answer is:

A=2, B=3 and C=6.

b) What are the values of D, E, and F?

|||| Solution

E and F are the observed F-statistics, which are:

$$F_{\text{obs,Pickles}} = \frac{MS_{\text{Pickles}}}{MSE} = \frac{0.16333}{0.10556} = 1.547$$
$$F_{\text{obs,Dressing}} = \frac{MS_{\text{Dressing}}}{MSE} = \frac{1.84556}{0.10556} = 17.48$$

Actually, only one answer option has these two values. The D= SSE could be found from the total sum of squares

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{4}(y_{ij}-2.23)^2,$$

and then

$$D = SSE = SST - 0.3267 - 5.5367.$$

Or more easily using that the degrees of freedom ($=(r-1)(b-1)=6$) and then

$$D = SSE = 6 \cdot MSE = 6 \cdot 0.10556 = 0.633.$$

In any case, the answer is:

$D = 0.633$, $E = 1.55$ and $F = 17.48$

c) With a test level of $\alpha = 5\%$ the conclusion of the analysis, what is the conclusion of the tests?

|||| Solution

We look at the P-values in the ANOVA table, and observe that the Dressing p-value is BELOW 0.05 and the Pickles p-value is ABOVE 0.05, and hence the answer is:

Only the choice of the dressing type has a significant influence on the taste.

8.8 Transport times

|||| Exercise 8.8 Transport times

In a study the transport delivery times for three transport firms are compared, and the study also involves the size of the transported item. For delivery times in days, the following data found:

	The size of the item			Row
	Small	Intermediate	Large	average
Firm A	1.4	2.5	2.1	2.00
Firm B	0.8	1.8	1.9	1.50
Firm C	1.6	2.0	2.4	2.00
Coulumn average	1.27	2.10	2.13	

In R was run:

```
anova(lm(Time ~ Firm + Itemsize))
```

and the following output was obtained: (wherein some of the values, however, has been replaced by the symbols A, B, C and D)

```
Analysis of Variance Table

Response: Time
          Df  Sum Sq   Mean Sq  F value   Pr(>F)
Firm       2     A        B      4.2857  0.10124
Itemsize   2  1.44667     C        D     0.01929
Residuals  4  0.23333  0.05833
```

a) What is A, B, C and D?

|||| Solution

We have a two-way ANOVA situation. The definition of the terms in the given ANOVA table can all be found in Section 8.3, as:

$$C = \frac{SS(Bl)}{DF(Bl)} = \frac{1.44667}{2} = 0.723,$$

$$D = \frac{MS(Bl)}{MSE} = \frac{C}{MSE} = \frac{0.723}{0.05833} = 12.4.$$

And since B is given easily by A:

$$B = \frac{SS(Tr)}{DF(Tr)} = \frac{A}{2},$$

where $DF(Tr)$ denotes the degrees of freedom for the treatment Firm. We just to find $A = SS(Tr)$. We could use the defining formula, and that the overall average is $\bar{y}_{\cdot\cdot} = 5.5/3 = 1.83$

$$A = SS(Tr) = 3 \cdot (2 - 1.83)^2 + 3 \cdot (1.5 - 1.83)^2 + 3 \cdot (2 - 1.83)^2 = 0.5,$$

but more easily we could find B from the F-value as

$$B = 4.2857 \cdot 0.05833 = 0.25,$$

and then

$$A = 2, B = 0.5.$$

Hence the correct answer is:

A = 0.5, B = 0.25, C =0.723 and D = 12.4.

b) What is the conclusion of the analysis (with a significance level of 5%)?

|||| Solution

We look at the two p-values, and see that the Itemsize p-value is less than 5% ("groups significantly different") and that the Firm p-value is NOT ("groups NOT significantly different") and hence the correct answer is:

Only the size of the item has a significant influence on the delivery time.