# Classical and Bayesian Statistics

## Problems 6

### Problem 6.1

Below you will find several examples of comparisons of two samples. Give *short* answers to the following questions for each example:

- Is the sample paired or unpaired? Justify your answer!
- Is the test to be performed one-sided or two-sided? Justify your answer!
- What is the null hypothesis in words?
- What is the alternative hypothesis in words?

(∗∗) a) One experiment was to investigate the effect of cigarette smoking on platelet aggregations. Blood samples were taken from 11 people before and after smoking a cigarette, and the amount of platelet accumulation was measured. We are interested in whether platelet accumulation is increased by smoking.

(∗∗) b) The next data are from a study by Charles Darwin on cross-pollination and self-insemination. 15 pairs of seedlings of the same age, one produced by self-insemination and one by cross-pollination, were bred. Both parts of each pair had almost identical conditions. The aim was to see whether the cross-pollinated plants had more vitality than the self-pollinated ones, i.e. whether they grew bigger. The height of each plant was measured after a fixed period of time.

(∗∗) c) Does the calcium content in the diet affect the systolic blood pressure? To test this question, a trial group of 10 men were given calcium supplementation for 12 weeks. A control group of 11 men were given a placebo.

(∗∗) d) One experiment investigated whether mice had different levels of iron intake in two forms ($Fe^{2+}$ and $Fe^{3+}$). For this purpose 36 mice were divided into two groups of 18 each and one group was "fed" $Fe^{2+}$ and the other $Fe^{3+}$. As the iron was radioactively marked, both the initial concentration and the concentration could be measured some time later. From this, the proportion of iron absorbed was calculated for each mouse.

### Problem 6.2

Two depth gauges measure the following values for the depth of lakes at 9 different locations:

We assume that the measurements are normally distributed.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Gauge $A$ | 120 | 265 | 157 | 187 | 219 | 288 | 156 | 205 | 163 |
| Gauge $B$ | 127 | 281 | 160 | 185 | 220 | 298 | 167 | 203 | 171 |

Earlier studies show that gauge $B$ systematically measures larger values than gauge $A$. Do the readings confirm this assumption or is a random fluctuation plausible as an explanation?

(∗∗) a) Are the samples paired or independent?

(∗∗) b) Do we perform a one- or two-sided test? Justify your answer.

(∗∗) c) Perform a $t$-test at significance level of $\alpha = 0.05$. Formulate explicitly the model assumptions, the null hypothesis, the alternative hypothesis, and the test result.

## Problem 6.3

The following table shows the jaw lengths of 10 male and 10 female golden jackals:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Male $x_i$ | 120 | 107 | 110 | 116 | 114 | 111 | 113 | 117 | 114 | 112 |
| Female $y_j$ | 110 | 111 | 107 | 108 | 110 | 105 | 107 | 106 | 111 | 111 |

We want to investigate whether the male and female jackals have different jaw lengths.

**R**:

```
# Read in dataset
jackals <- read.table(file="jackals.txt", header=TRUE)
```

(∗∗) a) Are the samples paired or unpaired? Justify your answer.

(∗) b) Formulate null and alternative hypothesis.

(∗∗) c) Perform a $t$-test with the help of **R**. Interpret the $p$-value and the resulting test decision.

d) Perform a Wilcoxon test with the help of **R**. Again, interpret the $p$-value and make the test decision.

e) If the results of the two tests are different, which one would you rather trust? Why?

## Problem 6.4

A U.S. magazine, Consumer Reports, conducted an investigation into the calorie and salt content of various hot dog brands. There were three different types of hot dogs: beef, "meat" (beef, pork, mixed poultry) and poultry.

The results below list the calorie content of different brands of beef and poultry hot dogs.

Beef hot dog:

$$186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132$$

Poultry hot dog:

$$129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142, 86, 143, 152, 146, 144$$

Do the two types of hot dog have a significantly different calorie content? We answer this question with a hypothesis test.

(∗) a) Is it a paired or unpaired test? Justify your answer.

(∗∗) b) Is it a one- or two-sided test? Justify your answer.

(∗) c) Formulate null and alternative hypothesis.

(∗) d) Calculate the means of the two datasets. What is your assumption?

(∗∗) e) Which test would you choose, a $t$-test or Wilcoxon-test? Justify your answer.

(∗∗) f) Perform the corresponding test with **R**. Interpret the $p$-value.

## Problem 6.5

In the year 2013, within the framework of a international cooperation under the leadership of EAWAG in Dübendorf, concentrations of illegal substances in waste water from 42 European cities during one week were investigated (Ort C. et all, *Spatial differences and temporal changes in illicit drug use in Europe quantified by wastewater analysis*, Addiction 2014 Aug).

The median concentrations of ecstasy (MDMA) in waste water were measured on 7 consecutive days (6-12 March) in addition to other substances. On the basis of this study, a widely read Swiss free newspaper stated that a lot more drugs are consumed in Zurich than elsewhere.

The following table shows for the cities of Zurich and Basel the quantities of MDMA that were extracted on the days of the week - the values can be found in the file *mdma.txt*. The values are in mg per 1000 inhabitants per day.

| Weekdays | Wed | Thu | Fri | Sat | Sun | Mon | Tue |
|---|---|---|---|---|---|---|---|
| Zurich | 16.3 | 12.7 | 14.0 | 53.3 | 117 | 62.6 | 27.6 |
| Basel | 10.4 | 8.91 | 11.7 | 29.9 | 46.3 | 25.0 | 29.4 |

Assume that the daily differences $D_i$ between the quantities of MDMA extracted per thousand inhabitants in the wastewater of Zurich and Basel are independently normally distributed with expected value $\mu_D$ and standard deviation $\sigma_D$.

*Hint:*

```
... <- read.table("...", header = TRUE)
```

(∗∗) a) Estimate (calculate) from the data the mean and standard deviation of the differences, i.e. $\widehat{\mu}_D$ and $\widehat{\sigma}_D$.

(∗) b) Are the samples paired or unpaired? Justify your answer.

(∗) c) Formulate the null hypothesis and the alternative hypothesis if you want to check the statement of the said free newspaper.

(∗∗) d) Perform a statistical test with the help of **R** on the significance level 5 %, assuming that the data are normally distributed.

What is your test decision?

(∗∗) e) Specify the 95 % confidence interval for the differences $D_i$ (using **R**).

How do you interpret this confidence interval?

(∗∗) f) Now perform a statistical test with the help of **R** at significance level of 5 %, assuming that the data are not normally distributed. What is your conclusion?

## Problem 6.6

(Continuation of Problem 2.3)

From our own experience, we have the impression that in married couples the husband tends to be older than his wife. Now we want to examine with a hypothesis test whether this is the case.

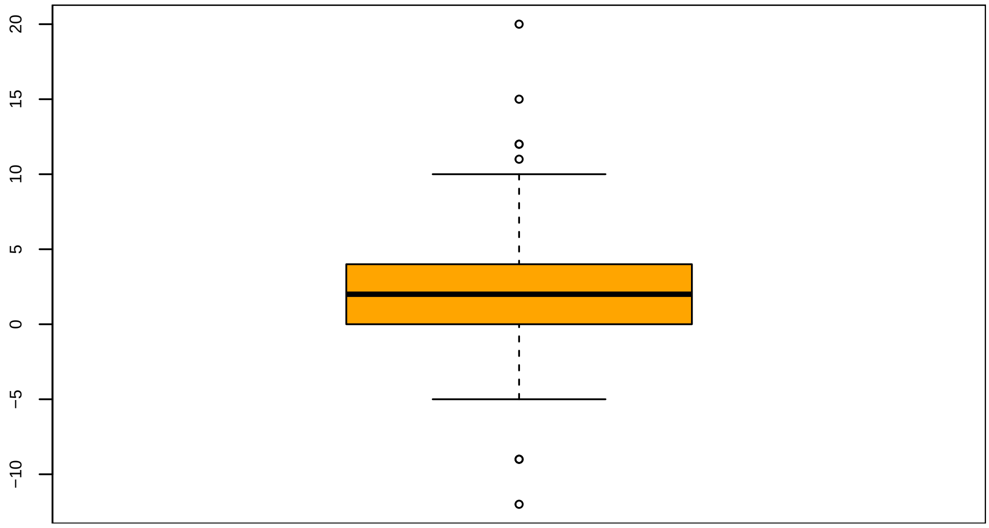
The data set husband_wife.csv contains the values of height and age for men and women of 170 British married couples. The height of husbands and wives is given in cm and the age in years.

*Note:*

```
mf <- read.csv(".../husband_wife.csv")
```

The `...` represent the path where the file was saved.

In Problem 2.3, we have already seen the box plot for the age difference of the married couples.




```r
diff <- mf$age.husband - mf$age.wife

boxplot(diff, col = "orange")
```

Remark **R**: The expression

```r
mf$age.husband

# is equivalent to

mf[, "age.husband"]
```

In about 50 % of the married couples the age difference is between 0 and about 5 years. In about 25 % of the married couples the wife is older than the husband. Our assumption seems to be correct, but the question still remains whether the difference is statistically significant.

a) We want to investigate our assumption that husbands are more likely to be older their wives with a hypothesis test.

(∗) i) Do you choose a paired or unpaired test? Justify your answer.

(∗) ii) Do you choose a one- or two-sided test? Justify your answer.

(∗∗) iii) We assume normal distribution of the data. Carry out a hypothesis test at a significance level of 5 %.

Formulate the null and alternative hypothesis and perform the test and make the test decision.

Make the test decision with the confidence interval.

(∗∗) iv) If you do not assume a normal distribution, which test do you choose? Carry out this test and interpret the result.

b) We are studying the differences in height between women and men. In general, men are larger than women. In England, according to Wikipedia, men are on average 13 cm taller than women.

(∗) i) Which test is appropriate here (one- or two-sided, paired or unpaired)? Justify your answer.

(∗) ii) Formulate the null and alternative hypothesis.

(∗∗) iii) Is the statement that on average the men are 13 cm taller than the women statistically significantly refuted by our data set at a significance level of 5 %? Perform the test and interpret the result. We assume that the body heights are normally distributed.

Make the test decision with the confidence interval.

## Problem 6.7

The body temperature of 10 patients is measured at the time of administration of a drug ($T_1$) and 2 hours later ($T_2$). The aim is to test with a hypothesis test whether this drug has a fever-lowering effect.

| Patient-Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Temp. 1 in °C | 39.1 | 39.3 | 38.9 | 40.6 | 39.5 | 38.4 | 38.6 | 39.0 | 38.6 | 39.2 |
| Temp. 2 in °C | 38.1 | 38.3 | 38.8 | 37.8 | 38.2 | 37.3 | 37.6 | 37.8 | 37.4 | 38.1 |

(∗) a) Is it a paired or unpaired test? Justify your answer.

(∗∗) b) Is it a one- or two-sided test? Justify your answer.

(∗) c) Formulate the null and alternative hypothesis.

(∗∗) d) Assume that the data are normally distributed. Which test do you choose? Carry out the test with R at significance level 5 %. What is your conclusion?

(∗∗) e) If we cannot assume that the data are normally distributed, which test do you choose? Perform this at significance level 5 %.

(∗∗) f) Explain the difference of the $p$-values in subtasks d) and e).

## Problem 6.8

Consider a one-sided $t$-test of $H_0 : \mu = 0$ against $H_A : \mu > 0$ at the significance level of 0.05.

Although the observed $n$ data points have an empirical mean greater than 0, the calculations show that the null hypothesis is not rejected.

Decide whether the following statements are *true or false*.

*Hint:* Make useful sketches including all relevant information.

$(***)$ a) We reject $H_0$ for no level $\alpha < 0.05$.

$(***)$ b) There is a level $\alpha < 1$ where we discard $H_0$.

$(***)$ c) The $p$-value is strictly smaller than 0.5.

$(***)$ d) If we perform a two-sided test at the level 0.05 instead of a one-sided test, we do not discard $H_0$.

$(***)$ e) If we copy the data more and more often (i.e. we look at each data point $k$ times, so that we obtain a total of $k \cdot n$ data points with the same mean as for $n$ data points), we discard $H_0$ for a large $k$ at significance level of 0.05.

# Classical and Bayesian Statistic

## Sample solution for Problems 6

### Solution 6.1

a) *Paired sample*: Each platelet count *before* smoking corresponds to the platelet count of the same person *after* smoking.

*One-sided test*: We do not want to know whether the platelet count has *changed*, but whether it has *increased*.

$H_0$: Smoking has no influence on the accumulation of platelets. ($\mu_{\mathrm{S}} = \mu_{\mathrm{NS}}$)

$H_A$: Smoking increases the accumulation of platelets. ($\mu_S > \mu_{NS}$)

b) *Paired sample*: To each height of a self-pollinated seedling belongs the height of the cross-pollinated "partner".

*One-sided test*: We do not want to know whether the heights *differ*, but whether the cross-pollinated seedlings become *larger* than the self-pollinated ones.

$H_0$: The heights do not differ.($\mu_c = \mu_s$)

$H_A$: Cross-pollinated seedlings become larger than self-pollinated ones. ($\mu_c > \mu_s$)

c) *Unpaired sample*: Unequal numbers in the groups. One blood pressure measurement from the experimental group does not correspond to a specific one from the control group.

*Two-sided test*: We just want to know if the calcium has an effect on the blood pressure, *no matter* if the blood pressure is higher or lower.

$H_0$: Calcium has no effect on blood pressure. ($\mu_{\mathrm{Calcium}} = \mu_{\mathrm{Contrast}}$)

$H_A$: Calcium has an effect on blood pressure. ($\mu_{\mathrm{Calcium}} \neq \mu_{\mathrm{Contrast}}$)

d) *Unpaired sample*: The numbers in the two groups need not be of equal size. The iron measurement of a "$Fe^{2+}$-mouse" does not correspond to a specific measurement of a "$Fe^{3+}$-mouse".

*Two-sided test*: We just want to know if the mice absorb the different forms of iron *differently*.

$H_0$: Iron absorption is independent of iron variety. ($\mu_2 = \mu_3$)

$H_A$: The iron absorption depends on the iron variety. ($\mu_2 \neq \mu_3$)

### Solution 6.2

a) These are *paired* samples. Measurements are taken at the same location with both gauges.

b) It is assumed that gauge $B$ has the higher values than gauge $A$. So we perform a one-sided test.

c) • *Model:*

$$D_1, \ldots, D_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

$\sigma$ is estimated by $\widehat{\sigma}$ and $d_i = x_A - x_B$.

• *Null hypothesis*

$$H_0: \quad \mu_D = \mu_0 = 0$$

*Alternative hypothesis:*

$$H_A: \mu_D < \mu_0$$

• *Significance level:*

$$\alpha = 5\,\%$$

• *p-value*

$$0.01168$$

```r
A <- c(120, 265, 157, 187, 219, 288, 156, 205, 163)
B <- c(127, 281, 160, 185, 220, 298, 167, 203, 171)

t.test(A, B, paired = TRUE, alternative = "less")


    Paired t-test

data:  A and B
t = -2.7955, df = 8, p-value = 0.01168
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
     -Inf -1.93449
sample estimates:
mean difference
      -5.777778
```

• *Test decision*

The $p$-value is less than 0.05 and thus the null hypothesis is rejected. The gauge $B$ produces indeed statistically significantly larger values than gauge $A$.

## Solution 6.3

Load the dataset

```
# Read in dataset
jackals <- read.table(file="jackals.txt",header=TRUE)

# View dataset
head(jackals)

    M   F
1 120 110
2 107 111
3 110 107
4 116 108
5 114 110
6 111 105
```

a) The samples are unpaired, as the individual males do not correspond to a specific female. The numbers in the two samples need not be the same.

b) We introduce the following terms:

- $X_i$: $i$th value of the jaw length of the males, $i = 1, \ldots, n = 10$
- $Y_j$: $j$th value of the jaw length of the female, $j = 1, \ldots, m = 10$

Model:

$$X_i \text{ i.i.d. } \mathcal{N}(\mu_{\text{M}}, \sigma_{\text{M}}^2), \qquad Y_i \text{ i.i.d. } \mathcal{N}(\mu_{\text{F}}, \sigma_{\text{F}}^2)$$

Null hypothesis:

$$H_0: \quad \mu_{\text{M}} = \mu_{\text{F}}$$

Alternative hypothesis:

$$H_A: \quad \mu_{\text{M}} \neq \mu_{\text{F}}$$

c) The **R** output for the $t$-test:

```
t.test(jackals[, "M"], jackals[, "F"])


    Welch Two Sample t-test

data:  jackals[, "M"] and jackals[, "F"]
t = 3.4843, df = 14.894, p-value = 0.00336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.861895 7.738105
sample estimates:
mean of x mean of y
    113.4     108.6
```

The $p$-value is $0.0034 < 0.05$, hence the null hypothesis is rejected. The male and female jackals have statistically significantly different jaw length.

d) The R output for the Wilcoxon-test looks:

```
wilcox.test(jackals[, "M"], jackals[, "F"])


    Wilcoxon rank sum test with continuity correction

data:  jackals[, "M"] and jackals[, "F"]
W = 87.5, p-value = 0.004845
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(jackals[, "M"], jackals[, "F"]) :
  cannot compute exact p-value with ties
```

Note that you can ignore the warning.

The $p$-value is $0.0048 < 0.05$, hence the null hypothesis is rejected in this test as well.

e) The result of the Wilcoxon-test is more trustworthy because, unlike the $t$-test, it does not assume that the data are normally distributed and we cannot verify this condition in any way.

However, the very different standard deviations in the two groups may be problematic for both tests.

## Solution 6.4

a) We cannot unambiguously assign the observations of one data set to the values of the other data set. So it is an unpaired test. Moreover, the data sets have different lengths.

b) No preference a priori between poultry and beef hot dogs is evident, i.e. we perform a two-sided test.

c) As this is an unpaired test, the means $\mu_{\text{Beef}}$ and $\mu_{\text{Poultry}}$ are compared.

Null hypothesis (no difference in calorie content)

$$H_0: \quad \mu_{\text{Beef}} = \mu_{\text{Poultry}}$$

Alternative hypothesis (difference in calorie content)

$$H_A: \quad \mu_{\text{Beef}} \neq \mu_{\text{Poultry}}$$

d) R output:

```
beef <- c(186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111,
          141, 153, 190, 157, 131, 149, 135, 132)

poultry <- c(129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142,
       86, 143, 152, 146, 144)

mean(beef)
[1] 156.85

mean(poultry)
[1] 122.4706
```

The calorie content of beef hot dogs seems to be much higher than that of poultry hot dogs. The null hypothesis should be rejected.

Note that is *not* good practice to decide from the data whether to perform a one- or two-sided test. We have to make this decision *before* collecting data.

We *could* argue that we already "know" ("studies has shown that...") , that beef is more fatty than poultry and *then* a one-sided test would be appropriate. But this information comes from *outside* the data sets.

e) Since there is no indication whether the data are normally distributed, we choose a Wilcoxon test as a precautionary measure.

f) R output:

```
wilcox.test(beef, poultry, paired = FALSE)



    Wilcoxon rank sum test with continuity correction

data:  beef and poultry
W = 285.5, p-value = 0.0004549
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(beef, poultry, paired = FALSE) :
  cannot compute exact p-value with ties
```

Ignore the warning.

The $p$-value is 0.00046 and thus far below the significance level of 0.05. Hence, the null hypothesis that the two types of hot dog have the same statistically significant calorie content is *rejected*.

*Now* we can argue based on d) that beef hot dogs has statistically significantly more calories than poultry hot dogs.

## Solution 6.5

Load the file:

```
mdma <- read.table("mdma.txt", header = TRUE)

head(mdma)

  Zurich Basel
1   16.3 10.40
2   12.7  8.91
3   14.0 11.70
4   53.3 29.90
5  117.0 46.30
6   62.6 25.00
```

a) We estimate with **R** the mean and the standard deviation as follows:

```
d <- mdma$Zurich - mdma$Basel

mean(d)
[1] 20.27

sd(d)
[1] 26.2723
```

We see that the standard deviation is very large compared to the mean. This could be an indication that the data are problematic for a hypothesis test.

b) We can consider the *days* as an experiment units, then these are paired samples because we have two observations per day.

However, it could also be argued that the *cities* are experiment units. In this case the samples are considered unpaired.

c) The null hypothesis is that there is no difference between the two cities in terms of the quantity of MDMA extracted, i.e.

$$H_0 : \quad \mu_D = \mu_0 = 0$$

The alternative hypothesis is

$$H_A : \quad \mu_D \neq \mu_0 = 0$$

Note that we *cannot* decide on a one-sided test because of the claim of the free newspaper, namely that more drugs are consumed in Zurich and therefore more MDMA is extracted, therefore

$$\mu_D > \mu_0 = 0$$

That statement is based on the *same* data which is not good practice (see DoE).

d) Output:

```
t.test(mdma$Zurich,
       mdma$Basel,
       paired = TRUE)


    Paired t-test

data:  mdma$Zurich and mdma$Basel
t = 2.0413, df = 6, p-value = 0.08729
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -4.027829 44.567829
sample estimates:
mean difference
          20.27
```

From the **R** output we can see that the $p$-value is 0.08729 and therefore greater than $\alpha = 0.05$. So at 5 % level of significance we do not reject the null hypothesis. Hence, there is not significantly more ecstasy is consumed in Zurich than in Basel. The claim of the newspaper is not valid.

If we interpret the samples as unpaired, then

```
t.test(mdma$Zurich,
       mdma$Basel,
       paired = FALSE)


    Welch Two Sample t-test

data:  mdma$Zurich and mdma$Basel
t = 1.3273, df = 7.5245, p-value = 0.2233
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.33677  55.87677
sample estimates:
mean of x mean of y
 43.35714  23.08714
```

We have the same test decision, although the $p$-value is much higher

e) From the **R** output for a paired test, we can see that the 95 % confidence interval is given by:

$$[-4.028, 44.57]$$

With 95 % probability, the true difference between the values of MDMA per thousand inhabitants in Zurich and Basel lies in the interval $[-4.028, 44.57]$.

Because 0 is contained in the 95 % confidence interval, we cannot reject the null hypothesis at the 5 % significance level.

f) Output:

```
wilcox.test(mdma$Zurich,
            mdma$Basel,
            paired = TRUE)



    Wilcoxon signed rank exact test

data:  mdma$Zurich and mdma$Basel
V = 27, p-value = 0.03125
alternative hypothesis: true location shift is not equal to 0
```

In this case we reject the null hypothesis, because the $p$-value is 0.0312 and thus lower than the significance level 5 %.

The different results for $t$-test and Wilcoxon-Test are probably because of the huge standard deviation compared to the mean. The normal distribution assumption may not be satisfied, so the Wilcoxon-Test is more trustworthy and the claim of the newspaper is statistically "correct".

## Solution 6.6

a) i) It is a paired test. For each test unit (married couple) there are two associated measurements (age husband, age wife).

ii) We are not sure whether the husbands are really older than their wives. It is simply our *impression* and *not* a fact. So perform do a two-sided test.

iii) Let $D$ denote the age difference between husband and wife.

Null hypothesis

$$H_0: \quad \mu_D = 0$$

Alternative hypothesis:

$$H_0: \quad \mu_D \neq 0$$

```
t.test(diff)



   One Sample t-test

data:  diff
t = 7.1518, df = 169, p-value = 2.474e-11
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
 1.618286 2.852302
sample estimates:
mean of x
 2.235294
```

or

```
t.test(mf$age.husband,
       mf$age.wife,
       paired = TRUE)


    Paired t-test

data:  mf$age.husband and mf$age.wife
t = 7.1518, df = 169, p-value = 2.474e-11
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 1.618286 2.852302
sample estimates:
mean difference
       2.235294
```

The $p$-value is far below the significance level of 5 % and thus the null hypothesis is rejected. The husbands are statistically significantly older than their wives.

The confidence interval is

$$(1.61, \; 2.85)$$

With a probability of 95 % the true mean lies in this interval. The null hypothesis was $\mu_D = 0$, i.e. no age difference. This value does *not* lie in the confidence interval and therefore the null hypothesis is rejected here as well. There is a statistically significant age difference within the married couples.

iv) Wilcoxon-test:

```
wilcox.test(diff)


    Wilcoxon signed rank test with continuity correction

data:  diff
V = 9460, p-value = 3.977e-12
alternative hypothesis: true location is not equal to 0
```

or

```
wilcox.test(mf$age.husband,
            mf$age.wife,
            paired = TRUE)


    Wilcoxon signed rank test with continuity correction

data:  mf$age.husband and mf$age.wife
V = 9460, p-value = 3.977e-12
alternative hypothesis: true location shift is not equal to 0
```

Again, the $p$-value is far below the significance level of 5 % and thus the null hypothesis is rejected. The husbands are statistically significantly older than their wives.

b) i) In this case only the average heights of men and women are compared, so it is an unpaired test. Since we do not know whether the deviation is upwards or downwards from 13 cm, we again perform a two-sided test.

Note that it is *not* the question whether men are taller than women. The question whether men are on average 13 cm taller than women.

ii) Let $\mu_W$ be the average height of the women and $\mu_M$ be the average height of the men. The null hypothesis is

$$H_0 : \quad \mu_W = \mu_M - 13$$

and the alternative hypothesis is

$$H_A : \quad \mu_W \neq \mu_M - 13$$

iii) We assume normal distribution of body heights:

```
t.test(mf$height.husband - 13,
       mf$height.wife,
       paired = FALSE)


    Welch Two Sample t-test

data:  mf$height.husband - 13 and mf$height.wife
t = -0.63293, df = 336.53, p-value = 0.5272
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.812281  0.929928
sample estimates:
mean of x mean of y
 159.8471  160.2882
```

or

```
t.test(mf$height.husband,
       mf$height.wife,
       mu = 13,
       paired = FALSE)


    Welch Two Sample t-test

data:  mf$height.husband and mf$height.wife
t = -0.63293, df = 336.53, p-value = 0.5272
alternative hypothesis: true difference in means is not equal to 13
95 percent confidence interval:
 11.18772 13.92993
sample estimates:
mean of x mean of y
 172.8471  160.2882
```

The $p$-value is far greater than the significance level and thus the null hypothesis is not rejected. The data do *not* refute the difference in height of 13 cm statistically significant.

The confidence interval is

$$(-1.81,\ 0.92)$$

With a probability of 95 % the true mean lies in this interval. The null hypothesis was $\mu_W = \mu_M - 13$, i.e. no statistically significant deviation from the height difference of 13 cm. This value 0 lies in the confidence interval and thus the null hypothesis is *not* rejected. The data do *not* refute the difference in height of 13 cm in a statistically significant way.

If we do not assume normal distribution, we choose an unpaired Wilcoxon-test (Mann-Whitney-U).

```
wilcox.test(mf$height.husband - 13, mf$height.wife, paired = FALSE)


    Wilcoxon rank sum test with continuity correction

data:  mf$height.husband - 13 and mf$height.wife
W = 13760, p-value = 0.4461
alternative hypothesis: true location shift is not equal to 0
```

Again, the null hypothesis is clearly *not* rejected.

## Solution 6.7

a) It is a paired test, as two measurements were taken on one test unit (patient).

b) We want to test the fever-lowering effectiveness. For this purpose we calculate the average of the $\mu_D$ of the differences $D_i$ (Temp. 1 − Temp. 2). In order to be able to prove the effectiveness

$$\mu_D > 0$$

Note that we are only interested in the fever-*lowering* and *not* the fever-rising effect. Therefore a one-sided test.

c) Null hypothesis (drug has no effect)

$$H_0: \quad \mu_D = 0$$

Alternative hypothesis (drug is fever-lowering)

$$H_A: \quad \mu_D > 0$$

d) R output:

```
t_1 <- c(39.1, 39.3, 38.9, 40.6, 39.5, 38.4, 38.6, 39.0, 38.6, 39.2)
t_2 <- c(38.1, 38.3, 38.8, 37.8, 38.2, 37.3, 37.6, 37.8, 37.4, 38.1)

t.test(t_1, t_2, paired=T, alternative="greater")


    Paired t-test

data:  t_1 and t_2
t = 5.6569, df = 9, p-value = 0.0001554
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 0.7976252       Inf
sample estimates:
mean difference
           1.18
```

The $p$-value 0.0001554 is less than 0.05 here. Therefore the difference is statistically significant. We can therefore assume that the drug is antipyretic (lowering the fever).

e) R output:

```
wilcox.test(t_1, t_2, paired=T, alternative="greater")


^^IWilcoxon signed rank test with continuity correction

data:  t_1 and t_2
V = 55, p-value = 0.002865
alternative hypothesis: true location shift is greater than 0
```

```
Warning message:
In wilcox.test.default(t_1, t_2, paired = T, alternative = "greater") :
  cannot compute exact p-value with ties
```

The $p$-value 0.002865 is less than 0.05. Therefore the difference is statistically significant. We can therefore assume that the drug is antipyretic.

f) The $p$-value of the Wilcoxon-test is greater than the $p$-value of the $t$-test. Since the Wilcoxon-test assumes less (no normal distribution) than the $t$-test, there is an additional uncertainty. The null hypothesis is less strongly rejected.

   However, the $t$-test suggests that it is more “precise”. This is true, but only if the data are normally distributed, which is often not known.

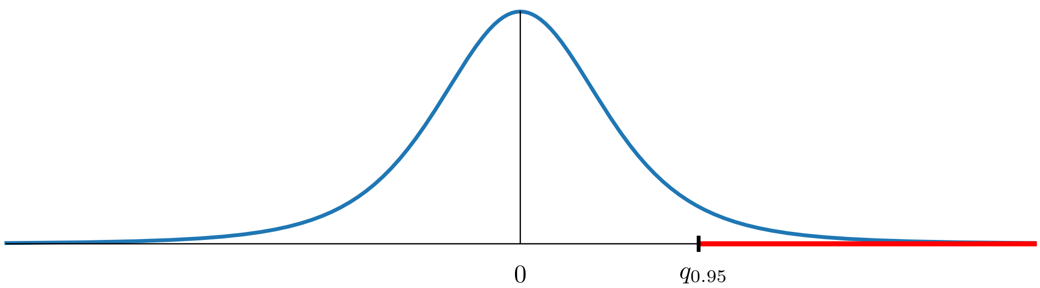
   Therefore the Wilcoxon test is often preferable to the $t$-test.

## Solution 6.8

This task is not easy and we have to read the text *very* carefully:

*Consider a one-sided t-test of $H_0 : \mu = 0$ against $H_A : \mu > 0$ against $H_A$ at the level of 0.05.*

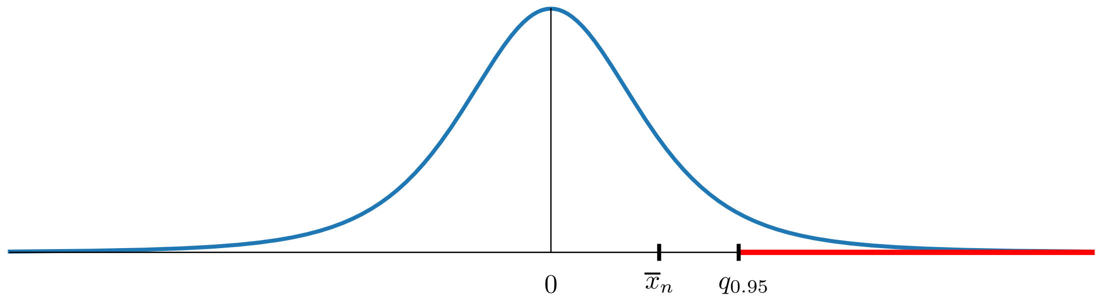
We draw a $t$-distribution with $\mu = 0$ and degree of freedom 4 (these assumptions are not relevant) and with the rejection zone (right-tailed test).




*Although the observed n data points have an empirical mean greater than zero, the calculations show that the null hypothesis is not rejected.*

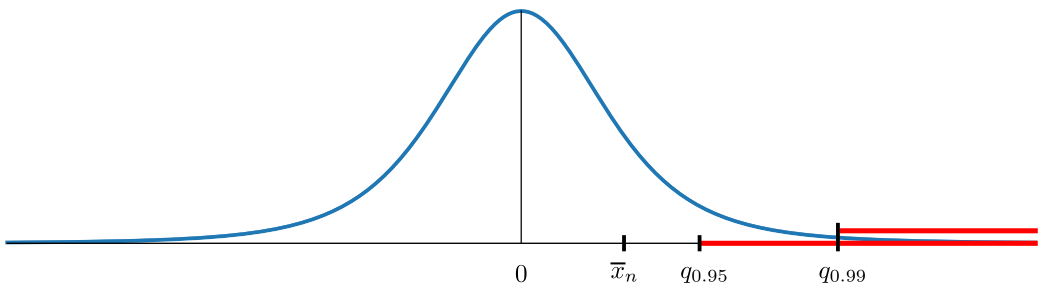
What does this mean?

- First of all, the mean is $\overline{x}_n > 0$ and is to the right of 0 on the $x$-axis.
- $H_0$ is *not* rejected, hence $\overline{x}_n$ is not in the rejection zone.
- $\overline{x}_n$ lies somewhere between 0 and the boundary of the rejection zone.




a) *We discard $H_0$ for no level $\alpha < 0.05$.*

What does that mean? Our significance level is $\alpha = 0.05$. Now we choose a significance level $\alpha^*$ that is *less* than 0.05, for example $\alpha^* = 0.01$. The boundary of the rejection zone moves to the right from $q_{0.95}$ to $q_{0.99}$ and the rejection zone becomes smaller. This means that $\overline{x}_n$ is still *not* in the rejection zone (see Figure below).



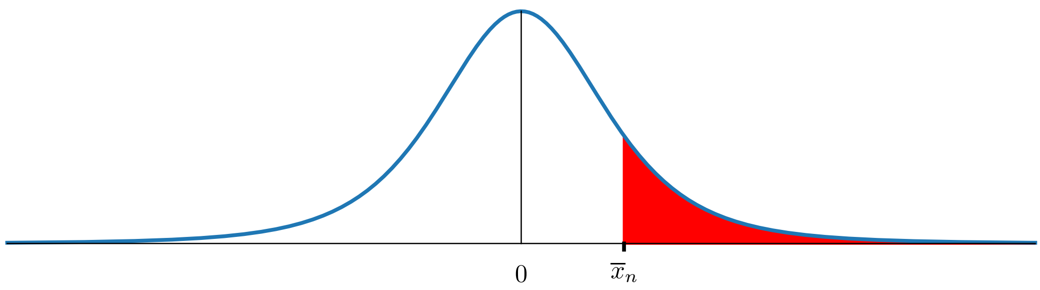

So for all $\alpha < 0.05$ the null hypothesis is not rejected, because $\overline{x}_n$ can never be in the rejection zone. So the statement is correct.

b) *There is a level $\alpha < 1$ where $H_0$ is rejected.*

The reasoning is the same as in a) but in the opposite direction: Let $\alpha = 0.05$ and we *increase* $\alpha$, then the boundary of the rejection zone moves to the left and for some $\alpha$, $\overline{x}_n$ is in the rejection zone and the null hypothesis is rejected. So the statement is correct.

c) *The p-value is strictly smaller than 0.5.*

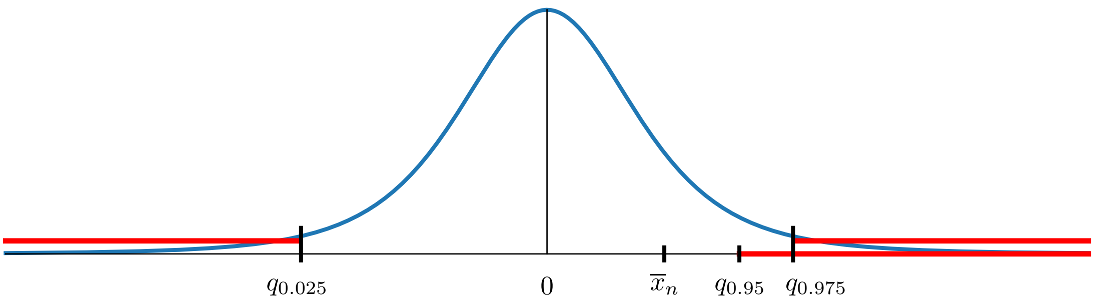
Now what is the $p$-value? This is the probability of observing a certain value (here $\overline{x}_n$) or a more extreme value towards the alternative hypothesis. We can represent probabilities as areas in continuous distributions and this is the area under the curve to the *right* of $\overline{x}_n$ (see Figure below).




Now the total area under the curve is 1, the area to the right of 0 is 0.5 and then the red area ($p$-value) must be less than 0.5. So the statement is correct.

d) *If we perform a two-sided test on significance level of 0.05 instead of a one-sided test, we do not discard* $H_0$.

Similar answer as in a): If we switch from a one-sided to a two-sided test, the rejection zone on the right-hand side becomes smaller. Since $\overline{x}_n$ is not in the rejection zone of the one-sided test, it cannot be in the rejection zone of the two-sided test. So the statement is correct.




e) *If we copy the data more and more often (i.e. we look at each data point k times, so that we obtain a total of $k \cdot n$ data points with the same mean as for n data points), we discard* $H_0$ *for a large k at significance level of 0.05.*

For example, if we quadruple the data points ($k = 4$), then the mean remains the same and the standard deviation *decreases* by a factor $\sqrt{k} = 2$.

Hence $\overline{x}_n = \overline{x}_{nk}$, but the number of observations increases. That means the curve becomes narrower and the boundary of the rejection zone moves to the left. For a sufficiently large $k$, the mean $\overline{x}_n$ lies in the rejection zone and the null hypothesis is rejected. So the statement is correct.