Classical and Bayesian Statistics

Problems 2

Problem 2.1

From our own experience, we have the impression that in married couples the husband is generally older than his wife. we Want to examine statistically whether this is true.

In a study from the UK, the age (in years) and the body height (in cm) of 170 married couples were collected.

(∗) a) Load the file **husband_wife.csv**. Assign it to a variable.

(∗∗) b) Execute the **summary(...)** command. Explain what the command does and interpret the values for the age of husband and wife.

(∗∗) c) Create a box plot of the *difference* of age between husbands and wives.

(∗∗) d) Interpret the median and quartiles in the box plot. What can you say about the outliers?

Problem 2.2

This task is about getting to know more **R** commands and practicing the use of **R**.

We will use the **InsectSprays** data set that is already contained in **R**.

```
head(InsectSprays)


##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

Six different insect sprays were used, which were sprayed on different fields. Then the number of insects was counted that were on the respective field after spraying. (Beall, G., (1942) The Transformation of data from entomological field experiments, Biometrika, 29, 243-262.)

a) First we want to determine the average values of the individual sprays. For this purpose we use the **R** command **tapply(...)**

```
tapply(InsectSprays[, "count"], InsectSprays[, "spray"], FUN = mean)

##         A         B         C         D         E         F
## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

This command applies (apply) the function **FUN**, in this case **mean** to the column **count** (**InsectSprays[, "count"]**). The average are taken ordered by the column **spray** (**InsectSprays[, "spray"]**). This means that the average values for **count** are calculated separately for the sprays $A, B, \ldots, F$.

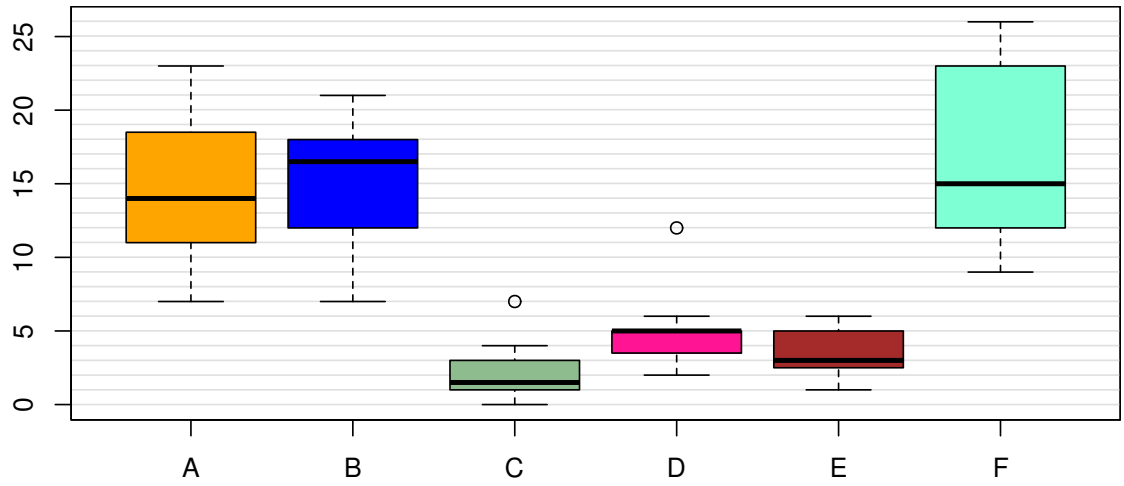
The mean values are very different. The sprays C, D and E seem to be much more efficient than the sprays A, B and F.

The following notation is somewhat easier:

```
tapply(InsectSprays$count, InsectSprays$spray, mean)

##         A         B         C         D         E         F
## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

b) Now we want to make a box plot of the data. Since the data is ordered by the column **spray**, **R** requires the input **boxplot(y ~ x)**, where **y** are the values of which **R** is to take the box plot and **x** are the names by which the values are to be ordered.

```
plot(NULL,xlim=c(0,10),ylim=c(0,27),xaxt="n", yaxt="n")
for (i in 0:26){
lines(c(-1,11),c(1.04*i,1.04*i),col="gray88")
}
par(new=T)
boxplot(count ~ spray,
        data = InsectSprays,
        col=c("orange", "blue", "darkseagreen", "deeppink",
              "brown", "aquamarine")
)
```



Again, it is obvious that the sprays *C*, *D* and *E* appear to be much more efficient than the sprays *A*, *B* and *F*.

Problem 2.3

In the file **Diet.csv** 76 persons are listed, who each followed one of the diets 1,2 or 3 for 6 weeks.

```
diet <- read.csv("~/Dropbox/Statistics_me/Classical_Bayesian_Statistics_W/Testat/Diet.c

head(diet)


##   Person gender Age Height pre.weight Diet weight6weeks
## 1     25     NA  41    171         60    2         60.0
## 2     26     NA  32    174        103    2        103.0
## 3      1      0  22    159         58    1         54.2
## 4      2      0  46    192         60    1         54.0
## 5      3      0  55    170         64    1         63.3
## 6      4      0  33    171         64    1         61.1
```

The file shows the weight **pre.weight** before taking the diet and the weight after 6 weeks **weight6weeks**. We are interested in the weight loss. Therefore we add a column **weight.loss** to the file. This is done as the following way:

```
diet$weight.loss <- diet$weight6weeks - diet$pre.weight

head(diet)


##   Person gender Age Height pre.weight Diet weight6weeks weight.loss
## 1     25     NA  41    171         60    2         60.0         0.0
```

```
## 2     26      NA  32    174         103    2        103.0         0.0
## 3      1       0  22    159          58    1         54.2        -3.8
## 4      2       0  46    192          60    1         54.0        -6.0
## 5      3       0  55    170          64    1         63.3        -0.7
## 6      4       0  33    171          64    1         61.1        -2.9
```

`R` recognises `diet$weight.loss` automatically as a new column and adds it at the end of the dataset.

Now perform the subtasks in the task before for `weight.loss` and `Diet`. Interpret the results in each case.

Problem 2.4

What's wrong with the following statements? Discuss.

(∗) a) The probabilities of a rigged (biased) coin were determined as $P(\text{heads}) = 0.32$ and $P(\text{tails}) = 0.73$.

(∗) b) The probability of winning in a lottery is $-3 \cdot 10^{-6}$.

(∗∗) c) A survey investigated the following events:

S: The interviewed person is pregnant.

M: The interviewed person is male.

It was found that $P(S) = 0.1$, $P(M) = 0.5$ and $P(S \cup M) = 0.7$

Problem 2.5

In a random experiment a red and a blue die are thrown simultaneously. We assume that both die are "fair" (unbiased), i.e. the numbers 1 to 6 of a die occur with the same probability.

(∗) a) Describe the sample space in the form of elementary events.

(∗) b) What is the probability of a single elementary event occuring?

(∗) c) Calculate the probability that the event E_1 "The sum of the eyes is 7" occurs.

(∗) d) What is the probability that the event E_2 "The eye sum is less than 4" occurs?

(∗) e) Calculate $P(E_3)$ for the event E_3 "Both numbers are odd".

(∗∗) f) Calculate $P(E_2 \cup E_3)$.

Problem 2.6

The events A and B are independent with probabilities $P(A) = 3/4$ and $P(B) = 2/3$. Calculate the probabilities of the following events. Make suitable freehand sketches using Venn diagramms.

$(*)$ a) Both events occur.

$(**)$ b) At least one of the two events occurs.

$(**)$ c) At most one of the two events occurs.

$(**)$ d) None of the two events occurs.

$(**)$ e) Exact one of the events occurs.

Problem 2.7

$(**)$ The collapse of a building in Tokyo can be caused by two independent events.

- E_1: Strong earthquake
- E_2: Big typhoon

The annual probabilities of these two events are $P(E_1) = 0.04$ and $P(E_2) = 0.08$.

Calculate the annual probability of the building collapsing.

Classical and Bayesian Statistic

Sample solution for Problems 2

Solution 2.1

a) Load the file

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

Use `head(...)` to check if the file was read correctly:

```
head(hw)
```

```
##   age.husband height.husband age.wife height.wife
## 1          49            180       43         159
## 2          25            184       28         156
## 3          40            165       30         162
## 4          52            177       57         154
## 5          58            161       52         142
## 6          32            169       27         166
```

b)

```
summary(hw)
```

```
##   age.husband    height.husband     age.wife      height.wife
##  Min.   :20.00   Min.   :155.0   Min.   :18.00   Min.   :141.0
##  1st Qu.:33.00   1st Qu.:169.0   1st Qu.:32.00   1st Qu.:156.0
##  Median :43.50   Median :172.0   Median :41.00   Median :160.0
##  Mean   :42.92   Mean   :172.8   Mean   :40.68   Mean   :160.3
##  3rd Qu.:53.00   3rd Qu.:177.0   3rd Qu.:50.00   3rd Qu.:165.0
##  Max.   :64.00   Max.   :190.0   Max.   :64.00   Max.   :176.0
```

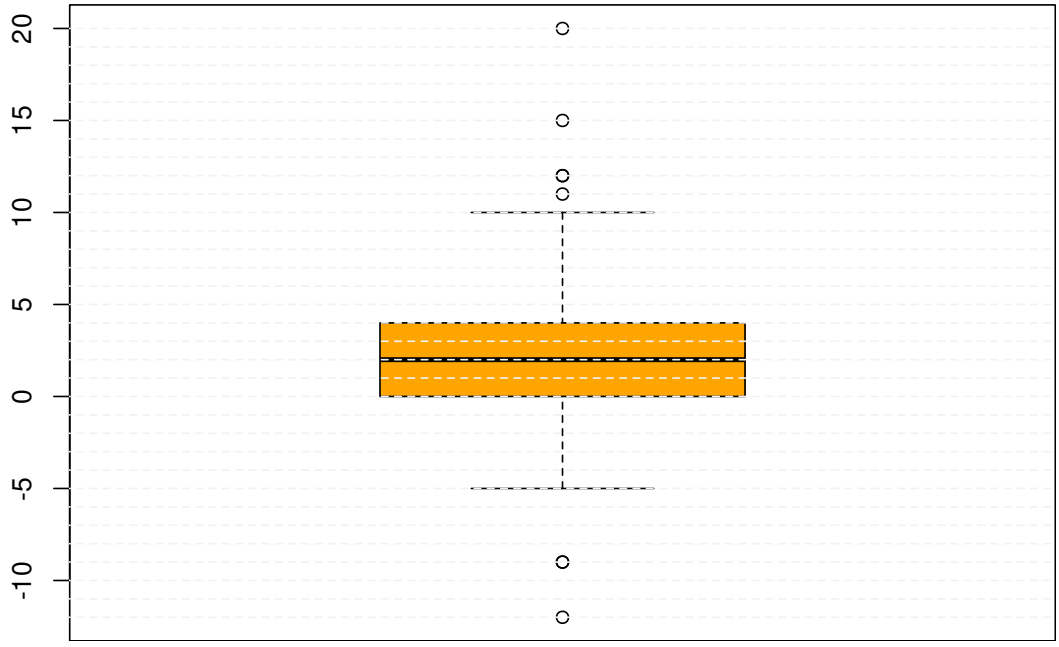
With the command `summary()` we get a short statistical overview of the data. It lists the smallest value (`min.`), the lower quartile (`1st Qu.`), the median (`Median`), the mean value (`Mean`), the upper quartile (`3rd Qu.`) and the maximum value (`Max.`) of the data.

For the 170 husbands, 20 is the lowest age and 64 is the highest. The average age is almost 43 years. The lower quartile is 33 years, so 25 % of the husbands are 33 or younger and 75 % are 33 or older. The median is 43.5 years, so 50 % of the husbands are 43.5 or younger and 50 % are 43.5 or older. The upper quartile is 53 years, so 75 % of the husbands are 53 or younger and 25 % are 53 or older.

The key figures for the wives can be interpreted analogously. However, a quick glance shows that the women tend to be a little younger than the men. This does *not* mean that the husbands are generally older than their wives.

c) Code:

```
age_man <- hw[, 1]
age_woman <- hw[, 3]

boxplot(age_man-age_woman, col="orange")
for (i in  -12:20){
lines(c(-2,2),c(i,i),col="gray95",lty="dashed")
}
```



d) The median is about 2, so the age difference is 2 or less for the half of the married couples and 2 or greater for the other half.

The lower quartile is at about 0, i.e. for 25 % of all couples, the wife is older than her husband.

The upper quartile is at 4, i.e. 25 % of all investigated married couples the husband is 4 or more years older than his wife.

The "middle" half of all married couples has an age difference (husband older than his wife) between 0 and 4 years.

The maximum difference is 20 years and the minimum is about -12. In the latter case the wife is about 12 years older than her husband.

Solution 2.2

Solution 2.3

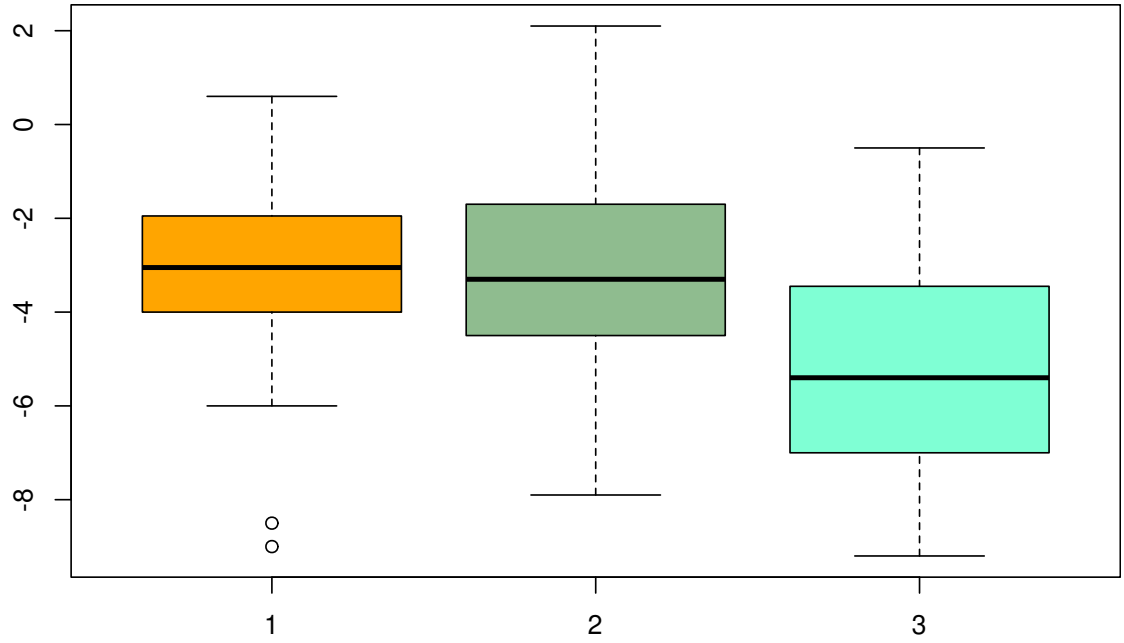
a) Group means:

```r
tapply(diet$weight.loss, diet$Diet, mean)

##         1         2         3
## -3.300000 -3.025926 -5.148148
```

Diets 1 and 2 lead to an average weight loss of about 3 kilograms. For Diet 3 on the other hand, the average weight loss is 5 kilograms and seems to be more efficient than the other two.

b) Graphical representation by a boxplot:

```r
boxplot(weight.loss ~ Diet,
        data = diet,
        col = c("orange", "darkseagreen", "aquamarine")

)
```



Boxplot confirms assumption from subtask a).

Solution 2.4

a) Because "head" and "tail" are the only possible outcomes, the probabilities have to add up to 1. This is not the case:

$$P(\Omega) = P(\text{Head}) + P(\text{Tail}) = 1.05$$

Axiom 2 is violated.

b) The probability is less than zero. Axiom 1 is violated.

c) Because $S \cap M = \{\}$ it follows $P(S) + P(M) = P(S \cup M)$ because of Axiom 3. But this is not the case.

Solution 2.5

a) $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,6)\}$, $|\Omega| = 36$

b) $P(\text{elementary event}) = \frac{1}{|\Omega|} = \frac{1}{36}$

c)
- $E_1 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$
- Number of favorable elementary events: $|E_1| = 6$
- Number of possibe elemetary events: $|\Omega| = 36$
- Probability of E_1 occuring:

$$P(E_1) = \frac{|E_1|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

d) $E_2 = \{(1,1),(2,1),(1,2)\}$:

$$P(E_2) = \frac{|E_2|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$$

e) $E_3 = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$:

$$P(E_3) = \frac{|E_3|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}$$

f) With the addition theorem:

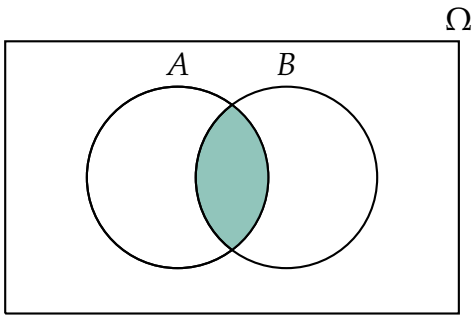
$$\begin{aligned} P(E_2 \cup E_3) &= P(E_2) + P(E_3) - P(E_2 \cap E_3) \\ &= P(E_2) + P(E_3) - P(\{(1,1)\}) \\ &= \frac{3}{36} + \frac{9}{36} - \frac{1}{36} \\ &= \frac{11}{36} \end{aligned}$$

Be careful: The rule $P(E_2 \cap E_3) = P(E_2)P(E_3)$ does not apply in this case as the events E_2 and E_3 are *not* independent.

Solution 2.6

```
A <- 3/4
B <- 2/3
```

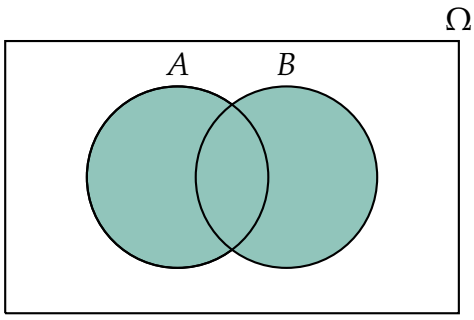
a)



$P(\text{both events}) = P(A \cap B) = P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{2}{3} =$

```
library(MASS)
fractions(A * B)

## [1] 1/2
```

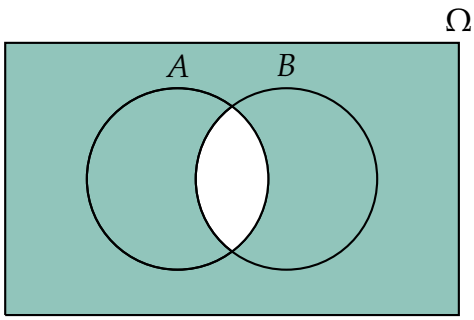
b)



$$\begin{aligned} P(\text{at least one}) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A) \cdot P(B) \\ &= \end{aligned}$$

```
fractions(A + B - A*B)

## [1] 11/12
```

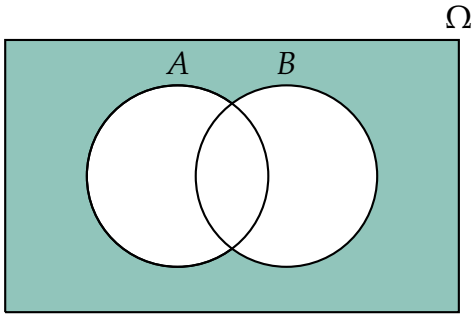
c)



$P(\text{at most one}) = 1 - P(A \cap B) = 1 - P(A) \cdot P(B)$

```
fractions(1 - A*B)

## [1] 1/2
```

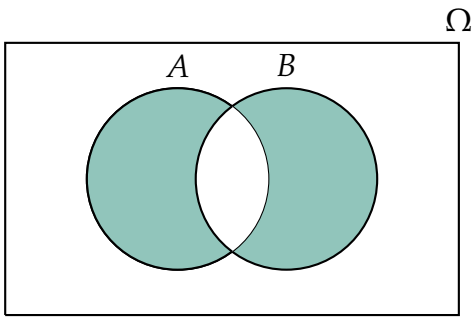
d)



$$\begin{aligned} P(\text{no event}) &= P(\overline{A \cup B}) \\ &= 1 - P(A \cup B) \\ &= 1 - (P(A) + P(B) - P(A) \cdot P(B)) \\ &= \end{aligned}$$

```
fractions(1-(A + B - A*B))

## [1] 1/12
```

e)



$$\begin{aligned} P(\text{exactly one event}) &= P(A \cup B) - P(A \cap B) \\ &= P(A) + P(B) - 2P(A) \cdot P(B) \\ &= \end{aligned}$$

```
fractions(A + B - 2*A*B)

## [1] 5/12
```

Solution 2.7

The annual probability of collapse $E_1 \cup E_2$ can be calculated as follows

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.04 + 0.08 - 0.04 \cdot 0.08 = 0.1168$$

where $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$, since E_1 and E_2 are independent.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.