

# Homework 2

Pascal Pilz, k12111234

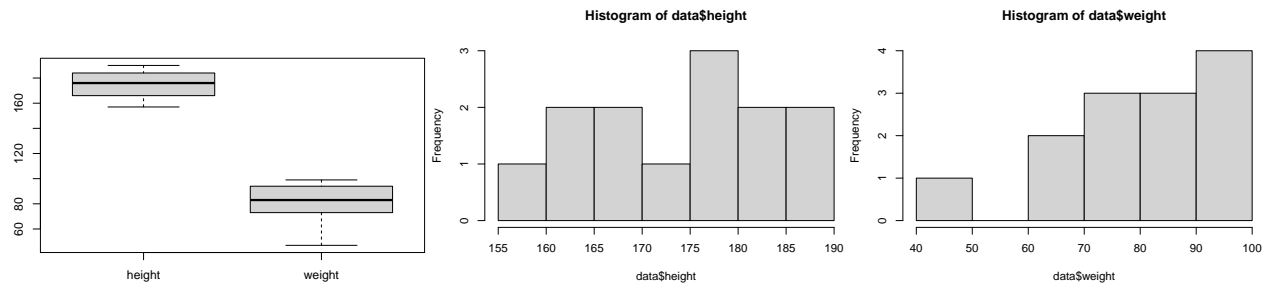
2023-10-20

## Example 5

We create boxplots of the data from example 2:

```
x1 <- c(157,184,162,166,168,163,178,185,190,187,176,174,179)
x2 <- c(47,85,66,83,62,73,84,94,99,96,76,94,74)
data <- data.frame("height"=x1, "weight"=x2)
boxplot(data)

hist(data$height)
hist(data$weight)
```



## Example 6

We have haemoglobin data of 16 patients:

```
x <-c (7.2,7.7,8.0,8.1,8.3,8.4,8.4,8.5,8.6,8.7,9.1,9.1,9.1,9.8,10.1,10.3)
```

a)

As a point estimator for the true mean and true standard deviation we can calculate the sample mean and sample standard deviation:

```
n <- length(x)

sample_mean <- (1/n) * sum(x)
sample_var <- (1/(n-1)) * sum((x-sample_mean)^2)
sample_std <- sqrt(sample_var)

print(sprintf("sample mean: %.3f, sample standard deviation: %.3f",
              sample_mean, sample_std))
```

```
## [1] "sample mean: 8.713, sample standard deviation: 0.844"
```

b)

Calculating a 95% confidence interval for the true mean value of haemoglobin based on the sample above.

For this we have the following formula:

$$\mu \in \left[ \bar{x} \pm t_{n-1, 1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right]$$

where  $\alpha$  is the chosen risk level,  $t_{n-1, 1-\alpha/2}$  is corresponding quantile of the  $t$ -distribution, and  $s$  is the sample variance ( $s/\sqrt{n}$  is the standard error).

We have 16 samples and a confidence level of 95% ( $\alpha = 0.05$ ), we find in the table  $t_{15, 0.975} = 2.13$ . To make it more precise and as an exercise in R we also get the value from the `qt` function.

```
alpha <- 0.05
t <- qt(1-(alpha/2), n-1)

val <- t * sample_std / sqrt(n)

conf_lower <- sample_mean - val
conf_upper <- sample_mean + val

print(sprintf("P(true mean in [%.3f,%.3f]) = %s",
              conf_lower, conf_upper, 1-alpha))
```

```
## [1] "P(true mean in [8.263,9.162]) = 0.95"
```

c)

What would happen with the confidence interval, if the sample size was not 16 patients but 160 patients?

The confidence interval would get smaller, since both the value of the  $t$ -distribution would get smaller ( $t_{159, 0.975} = 1.974$ ) and  $n$ , which is in the denominator of the formula, would increase. If the sample mean and sample variance would remain the same we would get the following:

```

n_new <- 159
alpha <- 0.05
t <- qt(1-(alpha/2), n_new-1)

val <- t * sample_std / sqrt(n_new)

conf_lower <- sample_mean - val
conf_upper <- sample_mean + val

print(sprintf("P(true mean in [%.3f,%.3f]) = %s",
              conf_lower, conf_upper, 1-alpha))

## [1] "P(true mean in [8.580,8.845]) = 0.95"

```