Assignment 1 - Group 27

Joost Driessen, Emma van Lipzig, Rohan Zonneveld

2023-02-26

Opmerkingen

1a - Geef wat meer context bij wat je doet. VB: Nu begin je met r code om de 96% CI te vinden zonder dat ergens staat dat je dat gaat doen en hoe je dat gaat doen. r code moet denk ik meer als toevoeging zijn en niet de hele opdracht.

1b - zie 1a, leg uit waarom je een t-test gaat doen

2c - Zouden jullie nog willen kijken naar de vraag: 'Can you improve this CI?'?

3 - Vraag weghalen bovenaan

3a - Shappiro Wilk does not .... etc. alinea moet er uit vm

3b - label op x-as boxplot

3c - labels op interaction plot zien er raar uit. Liever geen data$diet als label maar liever diet, etc.

4c - Moet je niet eerst onderzoeken of er een interaction effect is en als dat er niet is pas kijken of er een main effect is? Nu staat het andersom.

4 - in Rmd format zetten

Excercise 1. Birthweight

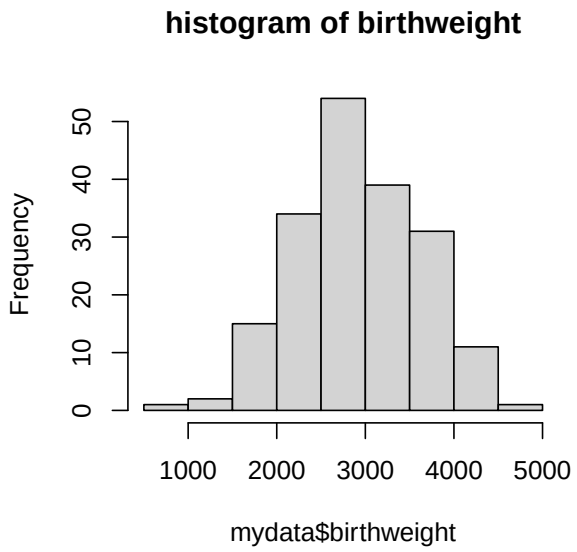
a)

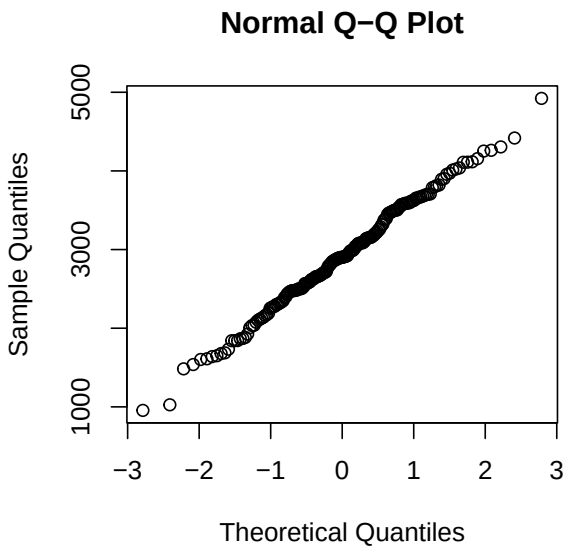
To test for normality we perform a *shapiro-wilk test* and look at the histogram and QQ-plot.

```
p=shapiro.test(mydata$birthweight)[[2]]; p
```

```
## [1] 0.9
```

$p > 0.05$, which means the Shapiro-Wilk test supports normality





Both the Shapiro-Wilk test and the plots support normality

```
n <- length(mydata$birthweight)
alpha <- 0.04
t_value <- qt(1-alpha/2, df=n-1)
t_value
```

```
## [1] 2.07
```

```
lower_bound <- mu - t_value*sigma/sqrt(n)
upper_bound <- mu + t_value*sigma/sqrt(n)
```

The 96% confidence interval lies between 2808.084 g and 3018.501 g.

TODO: Evaluate the sample size needed to provide that the length of the 96%-CI is at most 100. (Formule uit college 0/1 slide 49)

```
boot_fn <- function(data, index) mean(data[index])
boot_data <- boot(mydata$birthweight, boot_fn, R=1000)
quantile(boot_data$t, c(0.02, 0.98))
```

```
##   2%  98%
## 2806 3016
```

The 96% confidence interval using bootstrap lies between 2807.927 and 3018.938 These two confidence intervals are very similar. Normally a bootstrapped CI is constructed because of non-normality or outliers, but this data is distributed normally, so the bootstrap yields the same results as the regular method.

b)

```r
t.test(mydata$birthweight, mu = 2800, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  mydata$birthweight
## t = 2, df = 187, p-value = 0.01
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829  Inf
## sample estimates:
## mean of x
##      2913
```

This t-test supports the expert's claim that the mean birthweight is higher than 2800 g, since the CI is entirely above this value.

```r
num_above <- sum(mydata$birthweight > 2800)
num_below <- sum(mydata$birthweight < 2800)
```

Perform the sign test

```r
binom.test(num_above, num_above + num_below, p = 0.5, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  num_above and num_above + num_below
## number of successes = 107, number of trials = 188, p-value = 0.03
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.507 1.000
## sample estimates:
## probability of success
##                  0.569
```

The sign test too indicates that the expert's claim was right, since $p < 0.05$.

c)

```r
sig.level <- 0.05
effect.size <- (2913.293 - 2800) / sigma
power_t <- pwr.t.test(n = n, d = effect.size, sig.level = sig.level, type = "one.sample", alternative =
birthweight <- c(2786, 2977, 3076, 3277, 2826, 3106, 3016, 3131, 2846, 3011)
```

TODO: Write down the power

d)

The $1-\alpha$-confidence interval for p is found by $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (based on CLT). Since we do not know the true value of $\hat{p}$, we can resolve the equation $\hat{p}_l = \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and find a value for $\hat{p}$.

```
f <- function(x) x - qnorm(0.975)*sqrt((x*(1-x))/n) - 0.25
p.hat <- uniroot(f, c(0.25, 1))$root

me=qnorm(0.9975)*sqrt((p.hat*(1-p.hat))/n)
p.hat_right=p.hat+me
p.hat_right
```

```
## [1] 0.412
```

So, the confidence interval is [0.25,0.412] and the confidence level is 95%, since the $1-\alpha$-confidence interval was constructed and α is 5%.

e)

We want to test about the difference in population proportion of male and female babies weighing less than 2600 grams. We divide the sample into two groups, babies weighing less than and more than 2600 grams. The proportion of male babies weighing less than 2600 grams equals $34/62 = 0.548$ and babies weighing more equals $61/126 = 0.484$. To test if this difference is significant we perform a proportion test. The number of 'successes' equals the number of male babies in each group.

```
p=prop.test(c(34,61),c(62,126))[[3]]; p
```

```
## [1] 0.501
```

Since $p > 0.05$, H0 is not rejected which indicates that there is no significant difference in the population proportion of male and female babies. So, the claim of the expert is false.

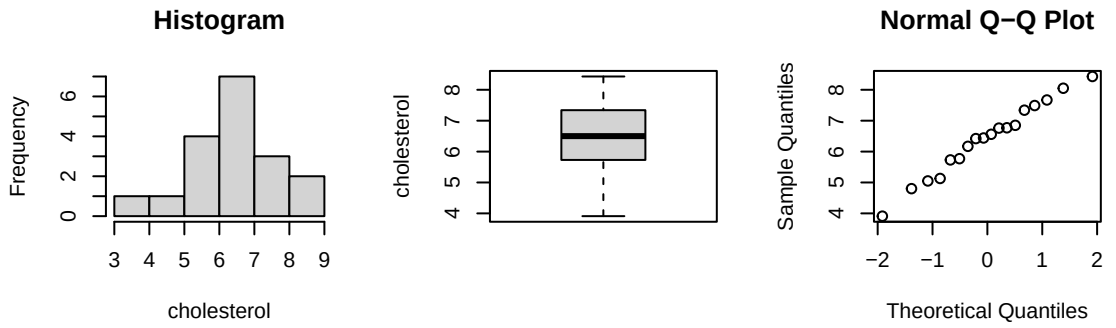
Excercise 2. Cholesterol

a)

We summarise the data by giving a summary and by making a histogram, a boxplot and a QQ-plot of both columns.

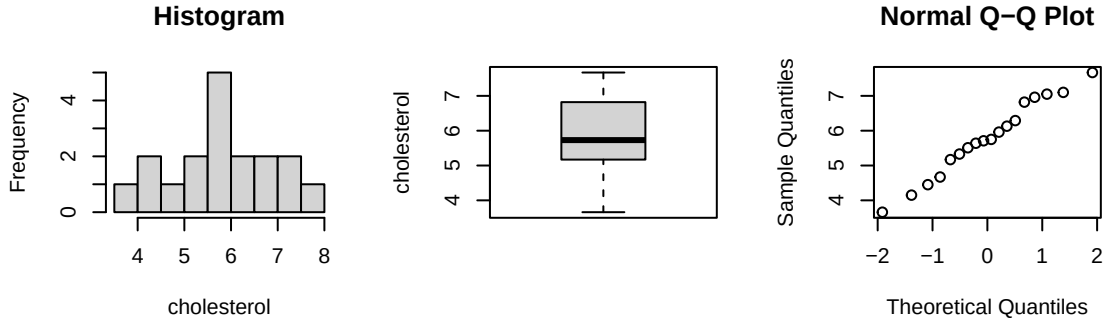
Before Diet

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.91    5.74    6.50    6.41    7.22    8.43
```



After 8 weeks of diet

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.66    5.21    5.73    5.78    6.69    7.67
```

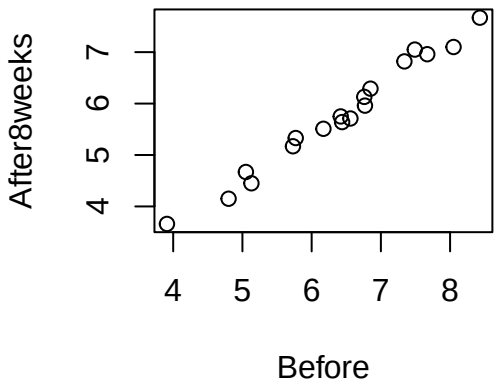


Both histograms look normally distributed, the boxplots are symmetrical with the mean in the center of the box and the QQ-plots are linear. In conclusion, the plots give no reason to suspect that the data is drawn from a non-normal distribution.

To investigate if the columns are correlated we calculate the correlation according to the literature. To visualize the correlation a plot is created with before on the x-axis and after8weeks on the y-axis.

```
cor(Before,After8weeks)
```

```
## [1] 0.991
```



Both the calculated and the plotted correlation indicate that the columns are strongly correlated.

b)

Since the study measured two sets of observations obtained from the same individuals the data is paired. In this case, each individual's before and after cholesterol levels are paired observations. Two relevant statistical tests for paired samples are: paired t-test and Wilcoxon signed-rank test.

A *paired t-test* is performed when the data is normally distributed. Looking back to *a)* we can see that the data is indeed normally distributed.

```
p=t.test(After8weeks, Before, paired = TRUE)[[3]]; p
```

```
## [1] 3.28e-11
```

The p-value is lower than 0.05, which means H0 can be rejected meaning that there is indeed a true difference in means between the two variables.

The *Wilcoxon signed-rank test* does not assume that the data is normally distributed and is therefore a useful alternative to the paired t-test when this assumption is violated.

```
p=wilcox.test(After8weeks, Before, paired = TRUE)[[3]]; p
```

```
## [1] 7.63e-06
```

The p-value is lower than 0.05, so H0 is rejected meaning that the true location of the population is different between columns.

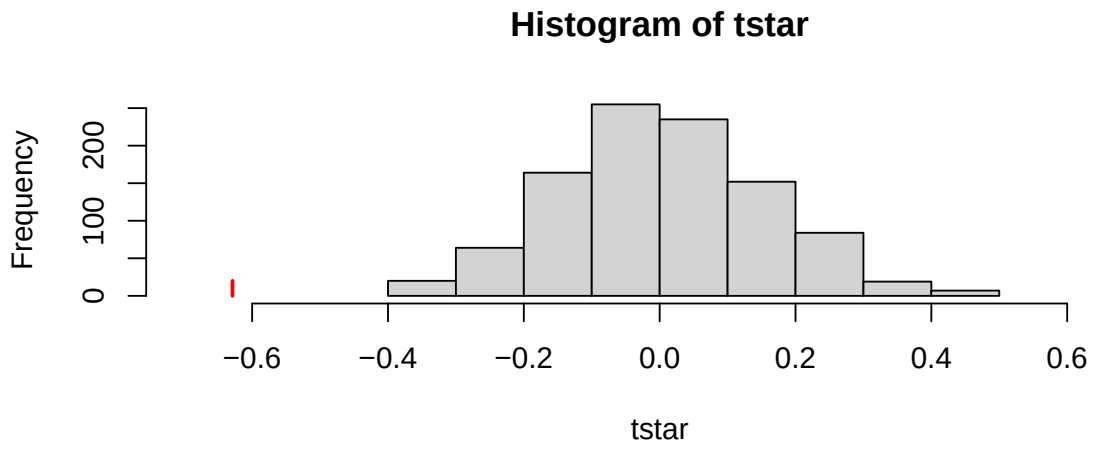
Permutation test

A permutation test is performed when the data is paired and we want to test for a difference between groups. Also, a permutation test does not assume any particular distribution of the data, therefore we can apply a permutation test to determine whether the diet has an effect. The R code to perform this test is as follows:

```
mystat=function(x,y) {mean(x-y)}
B=1000; tstar=numeric(B)
for (i in 1:B) {
  dietstar=t(apply(cbind(After8weeks,Before),1,sample))
  tstar[i]=mystat(dietstar[,1],dietstar[,2]) }
myt=mystat(After8weeks,Before)

myt
```

```
## [1] -0.629
```

In the histogram below the distribution of t* is depicted. The distribution is normal, due to CLT. The actual value of the T statistic is depicted in red.




```
pl=sum(tstar<myt)/B
pr=sum(tstar>myt)/B
p=2*min(pl,pr); p
```

```
## [1] 0
```

The value of p is equal to zero, which means H0 can be rejected. This means that there is a significant difference between before and after the diet.

c)

The mean of a uniform sample is given by $E[\bar{X}] = (1/n) * \sum Xi = 1/18 * \sum Xi$. The variation of a uniform sample is also given, $Var(Xi) = ((\theta - 3)^2)/12$. So the variance of the sample mean is given by $Var(\bar{X}) = Var((1/18) * \sum Xi) = (1/18^2) * \sum Var(Xi) = ((\theta - 3)^2)/(12 * 18 * 18)$. According to the central limit theorem, as n goes to infinity, the distribution of $\bar{X}$ approaches a normal distribution with mean $E[\bar{X}]$ and variance $Var(\bar{X})$. The mean of a uniform distribution is in the center, so to estimate the maximum of a uniform distribution $(\bar{\theta})$ one can simply multiply the mean by 2 and subtract the start of the interval.

```
x.bar = mean(After8weeks)
theta.hat = 2 * x.bar - 3
```

A 95% confidence interval for θ is found using the central limit theorem by calculating the margin of error and adding and subtracting it to $\bar{\theta}$.

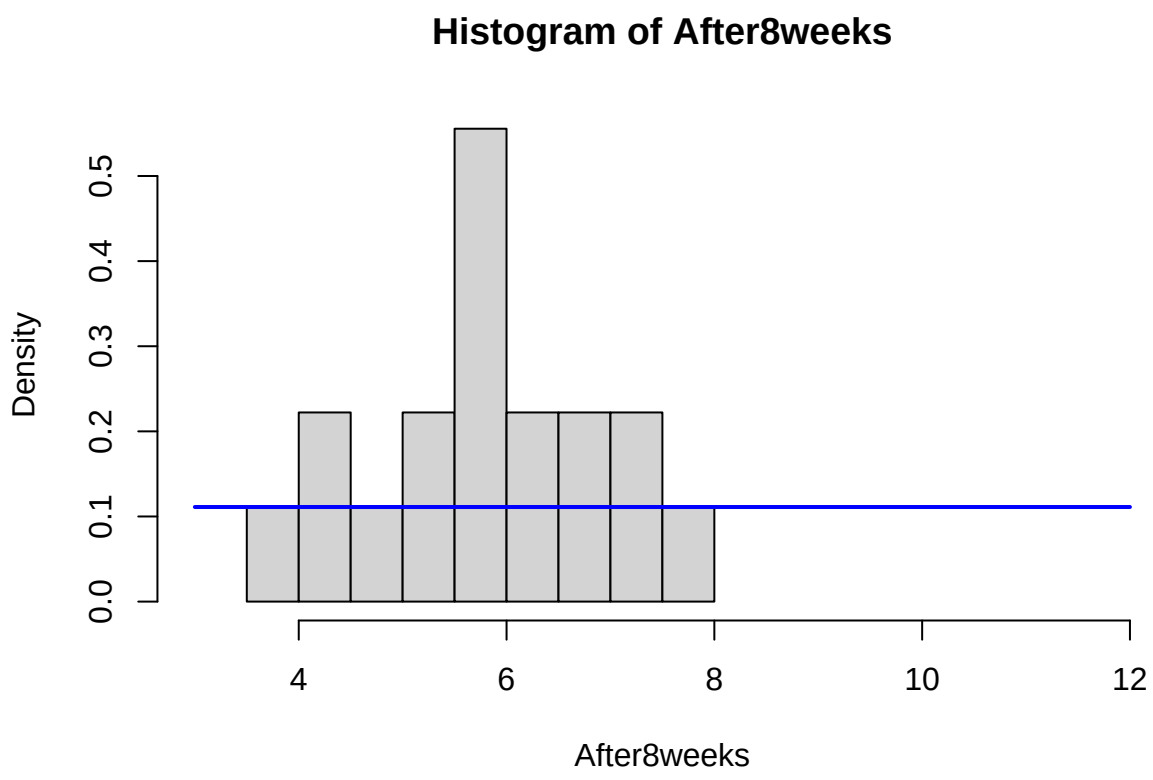
```
z = qnorm(0.975)
sigma = sqrt(((theta.hat - 3)^2)/(12*18*18))
me = z * sigma

ci = c(theta.hat - me, theta.hat + me)
ci
```

```
## [1] 8.38 8.73
```

This confidence interval means that for values of $8.38 < \theta < 8.73$ the data is not significantly different than if the data for the cholesterol after eight weeks of diet was drawn from $Unif[3, \theta]$.

Can you improve this CI?

d)



The After8weeks column does not appear to be uniformly distributed. Since there are no values bigger than 8 in the data it is expected that values of $\theta > 8$ will reject the null hypothesis. To investigate if After8weeks is

drawn from an uniform distribution in [3,θ] a bootstrap test is performed. Surrogate T-values are generated that are representative of values of T under H0. The test statistic (in this case the maximum) is computed. This process is repeated 1000 times. Finally, the T-value is compared to the surrogate T*-values to determine a p-value.

```
H0=rep(1, 9)
for(theta in 4:12){
  n=length(After8weeks); t=max(After8weeks)
  B=1000; Tstar=numeric(B)
  for(i in 1:B) {
   Xstar=runif(n,3,theta)
   Tstar[i]=max(Xstar)}
  pl=sum(Tstar<t)/B;pr=sum(Tstar>t)/B
  p=2*min(pl,pr)
  if(p<0.05){H0[theta-3]=0}}
H0
```

```
## [1] 0 0 0 0 1 0 0 0 0
```

θ	4	5	6	7	8	9	10	11	12
H0	0	0	0	0	1	0	0	0	0

The H0 is rejected for all values of θ except 8. This result was to be expected as the data ranges from 3 to 8 as can be seen in the histogram.

The *Kolmogorov-Smirnov* is used to test if two samples are from the same distribution. The actual data can be compared to data generated from a uniform distribution. So it is perfectly suited to use in this situation.

```
H0=rep(1,9)
for(theta in 4:12){
  n=length(After8weeks)
  sample=runif(n,3,theta)
  p=ks.test(After8weeks,sample)$p.value
  if(p<0.05){H0[theta-3]=0}
}
H0
```

```
## [1] 0 0 0 1 1 1 1 1 1
```

θ	4	5	6	7	8	9	10	11	12
H0	0	0	0	1	1	1	1	0	0

e)

To test whether the median cholesterol level after 8 weeks of low fat diet is less than 6 a *Wilcoxon signed-rank test* is performed.

```
p=wilcox.test(After8weeks,mu=6,alt="l")[[3]]; p
```

```
## [1] 0.223
```

The p-value is bigger than 0.05 so the null hypothesis can not be rejected. This means that it is not statistically significant that the population median is lower than 6.

```
n=sum(After8weeks<4.5)
prop=n/length(After8weeks);prop
```

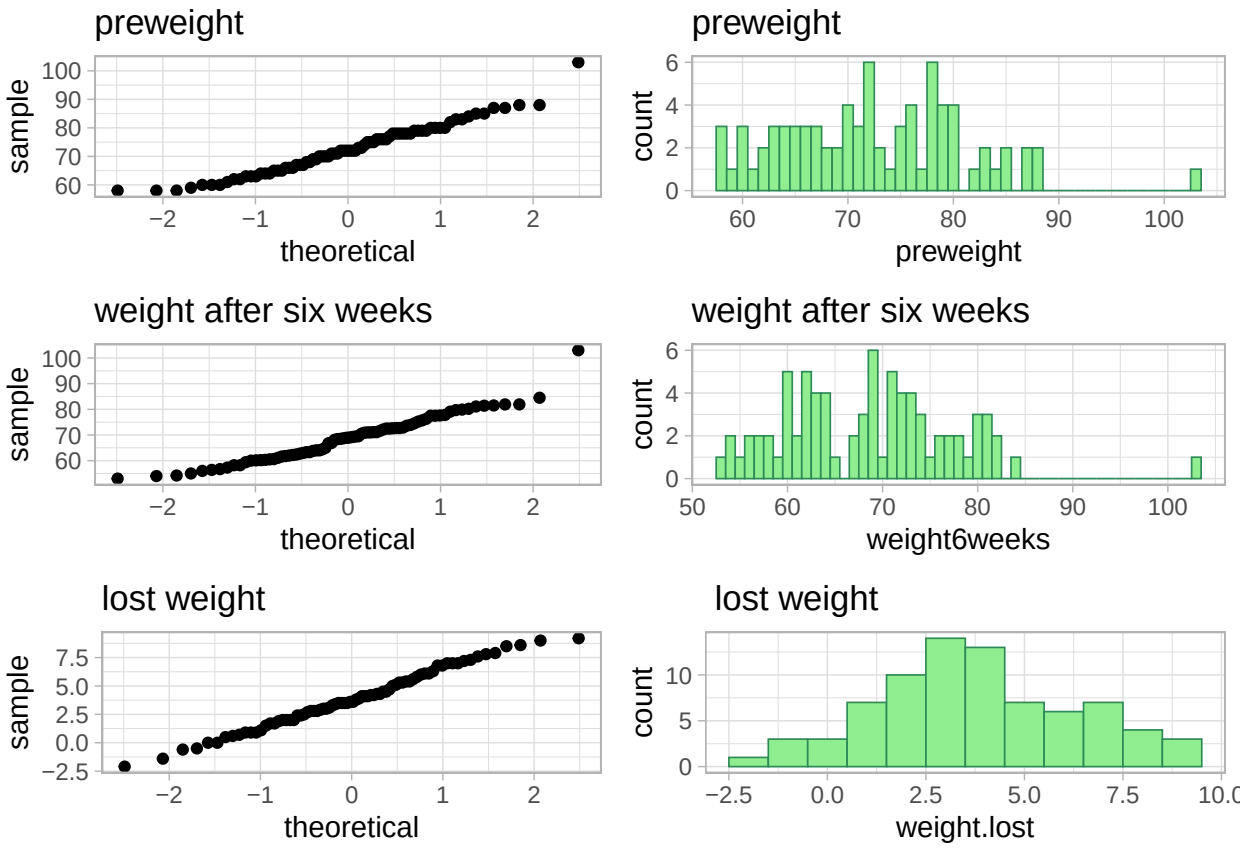
```
## [1] 0.167
```

The proportion of the values lower than 4.5 is less than 25%

Exercise 3 - Diet

```
##   person gender age height preweight diet weight6weeks weight.lost
## 1      1      0  22    159        58    1         54.2         3.8
## 2      2      0  46    192        60    1         54.0         6.0
## 3      3      0  55    170        64    1         63.3         0.7
```

a)



The qqplot and histogram of the weigth before diet show that this data does not appear to be normally

distributed, as there is one large outlier of 103. The same goes for the weight after six weeks, so after the diet. This again does not seem normally distributed because of the outlier 103. However, the differences between the two, represented as weight.lost, do appear to be distributed normally. This can be seen in both the QQ plot and histogram.

The Shapiro-Wilk test does not reject normality (if we take alpha = 0.05), for the preweight data. As this test does not always reject normality for non-normal distributions, this distribution may still be non-normal (as indicated by the histogram and qqplot).

The Shapiro-Wilk test rejects normality (if we take alpha = 0.05), for the weight at six weeks data. When this test rejects normality, the distribution should indeed be non-normal, which corresponds with the qqplot and histogram of the data.

Data	p-value for Shapiro-Wilk test
preweight	0.055
weight6weeks	0.011

The boxplots below shows that there may be a difference in the weight before and after dieting. The first boxplot shows the distribution of weight data before and after a diet, the second boxplot shows the distribution of the difference between these weights, which is the distribution of the weight loss.

Because the differences between the pre and post diet weight data is distributed normally, a t-test can be used to determine whether these two groups differ significantly from each other. Because the data is the pre and post weight from the same subjects, in this situation a paired t-test is applicable. H0 is that the groups are not different.

```
t_test1 <- t.test(data$preweight, data$weight6weeks, data = data, paired = TRUE)
t_test1$p.value
```

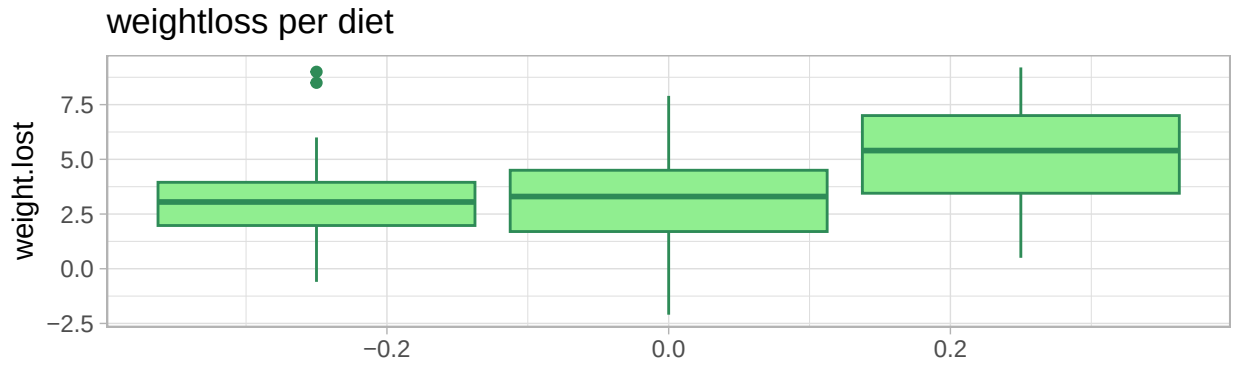
```
## [1] 1.17e-21
```

The p value of the t-test is < 0.05, therefore H0 is rejected. This means that the means of preweight and weight6weeks statistically differ. As the weightloss over the six weeks of dieting was significant, it could be that diet has an influence on weightloss. However, this does not take into account other factors included in the experiment like gender and age.

b)

Apply one-way ANOVA to test whether type of diet has an effect on the lost weight. Do all three types diets lead to weight loss? Which diet was the best for losing weight? Can the Kruskal-Wallis test be applied for this situation?

The boxplots below show that that different diets may have different effects on weightloss.



The summary below shows that the study is not a balanced design. Diet 1 has less participants than diet 2 and 3.

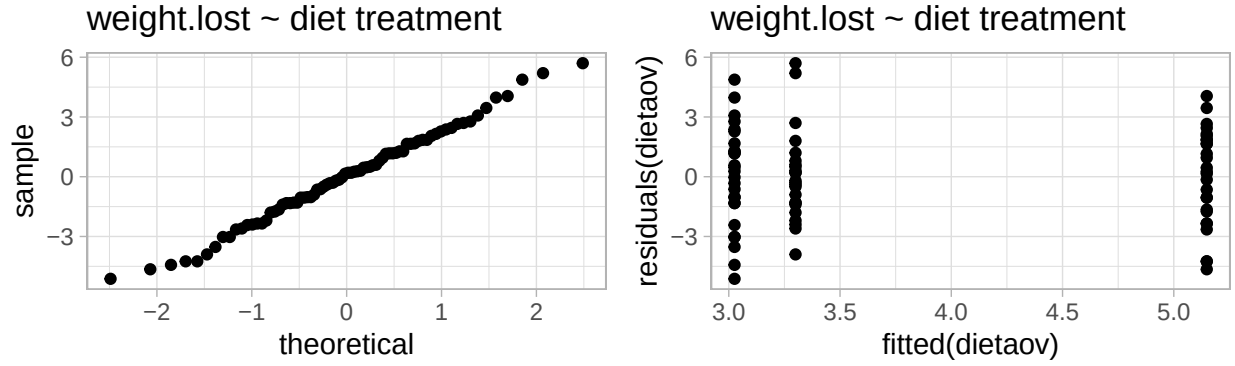
The ANOVA with treatment parametrization and diet 1 as base level shows that the other diets differ significantly from diet 1. This means that some diets are more effective for weightloss than others. If the study had a 'no diet' treatment, this would have been a good base level to use, however, this is not the case. As it is not clear which of the diets should be the base level, it is better to use sum parametrization.

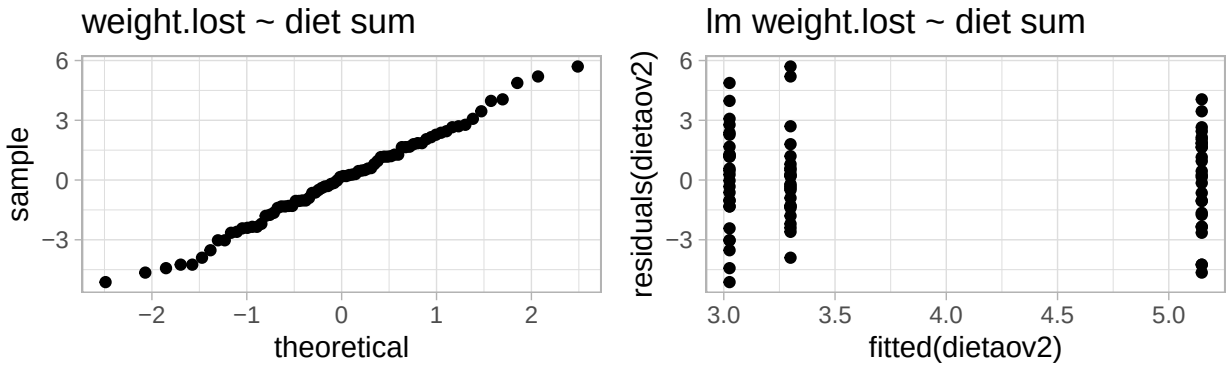
Relevant columns have been set to factor using `data$diet <- as.factor(data$diet)`. The following code has been used for the anova: `dietaov <- lm(weight.lost ~ diet, data=data) ; anova(dietaov).`

A summary of the treatment parametrization model shows that diet 3 is better for weightloss than diet 1. A summary of the model with sum parametrization shows that diet 2 differs significantly from the global diet mean (p = 0.039). Although all diets can be used for weightloss, diet 2 and 3 are better than diet 1. A t-test shows that diet 2 and 3 are significantly different (p APPROX 0.003), which implies that diet 3 is best for losing weight.

parametrization	p-value
treatment	0.003
sum	0.003

Below the model assumptions for the two ANOVA's are tested. The qqplots show that the residuals for both models seem normally distributed, and neither plot of the fitted against actual residuals shows particular patterns. Therefore the model assumptions are met for both ANOVA's. Because these assumptions are met, a non-parametric test like Kruskall-Wallis would be weaker to use. This is based on ranks, which means a lot of information is thrown out when using Kruskal-Wallis. It is still possible to use this test, but in this case it would not provide better insights than ANOVA.





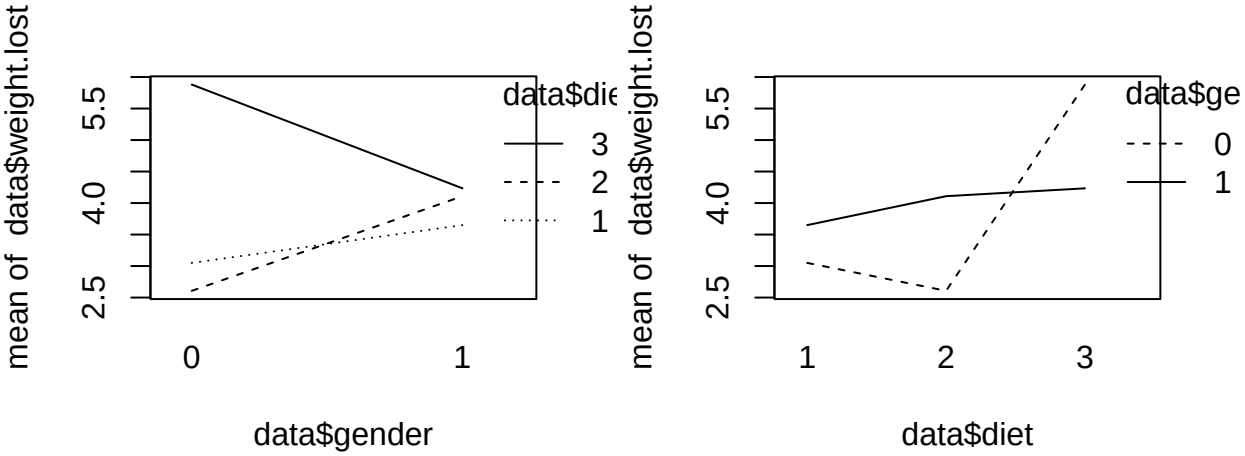
c)

Use two-way ANOVA to investigate effect of the diet and gender (and possible interaction) on the lost weight, using the following code: `genderlmm <- lm(weight.lost ~ diet + gender, data = data) ; anova(genderlmm)`, `genderlmi <- lm(weight.lost ~ diet * gender, data = data) ; anova(genderlmi)`.

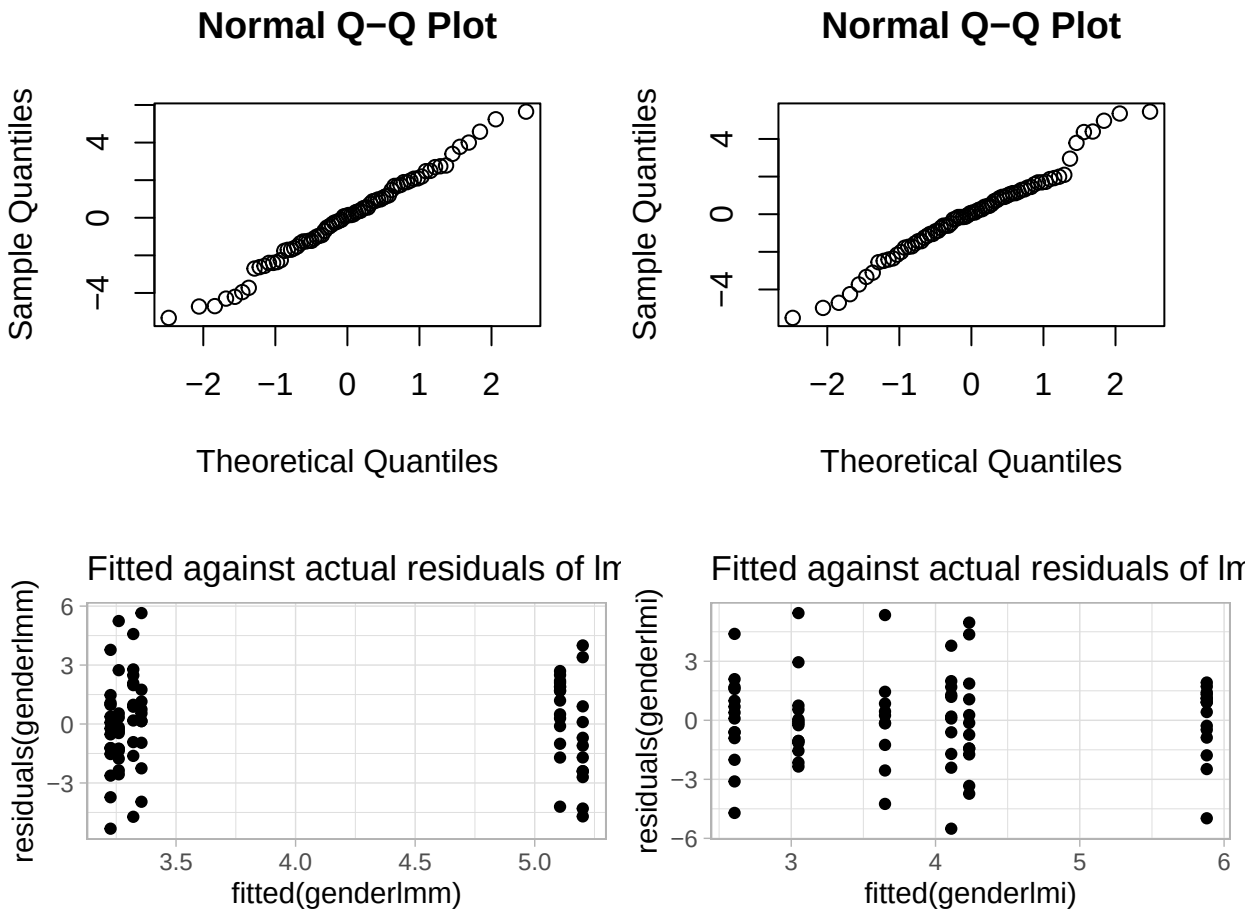
The two-way ANOVAs below show a main effect of diet on weight loss, no main effect of gender on weight loss and an interaction effect between gender and diet on weight loss. The factor diet is sum parametrized because there is no clear reason to pick one of the diets as a base level.

model	effect	p-value
diet + gender	diet	0.007
diet + gender	gender	0.877
diet + gender	diet:gender	0.093

The interaction plots show that indeed diet matters a lot. The specifically for diet 2 and 3 the difference between the genders is very pronounced. It looks like gender matters less as the lines for diet 1 and 2 look somewhat similar, but the line for diet 3 is very different.



Both the fitted against actual residual plot and qq residual plot indicate that the additive model does not violate the assumption of normality. However, the spread of the residuals looks different in the last column from the other columns for the interaction model. The QQ residual plot also does not look completely normal. Therefore it is doubtful that the assumption of normality is met for the interaction model.



e)

I prefer the approach from c), because this takes into account more of the data that is gathered during the experiment. Therefore this could lead to more accurate expectations of how different diets work for different people. Because an interaction effect between gender and diet was found, the interaction model is used to predict the weight loss per diet for both genders with respective average preweights, using the code `predict(genderlmi, newdata)`. As expected, for both genders, the predicted weightloss is highest for diet 3.

gender	preweight	diet	predicted weightloss
0	67.12	1	3.05
0	67.12	2	2.607
0	67.12	3	5.88
1	79.03	1	3.65
1	79.03	2	4.109
1	79.03	3	4.233

Excercise 4. Yield of peas