Genetische Statistik

Präsenzübung 3 - Deskription

Dr. Janne Pott (janne.pott@uni-leipzig.de)

November 16, 2021

Fragen

Gibt es bereits Fragen zu

- Vorlesung,
- Übung,
- Seminar?

Plan heute

Besprechung der zweiten R-Übungsblatts

- Deskriptive Statistiken in R

Keine weitere Aufgaben

Abschnitt 1

Deskriptive Statistiken in R

Aufgabe 1: Deskriptive Statistik

Datensatz *ergometer.RData*

- Berechnung *BMI*, Deskription *ergometer*, *lactate*, *BMI* und *Alter* für Männer und Frauen getrennt.
- Erstellung QQ-Plots und Histrogramme; Test auf Normalverteilung.
- Vergleich *ergometer* zwischen den Geschlechtern
- Korrelation *ergometer* mit *lactate*, *BMI* und *Alter*.

Aufgabe 1: Lösung a) - BMI

```
# BMI
class(myDat[,weight])
```

```
## [1] "numeric"
```

```
class(myDat[,height])
```

```
## [1] "numeric"
```

```
myDat[,BMI:=round(weight/height^2,2)]
```

Aufgabe 1: Lösung a) - Deskription

```
myCols <- c("ergometer","lactate","alter","BMI")
tab1<-myDat[sex==1, sapply(.SD, summary),.SDcols=myCols]
tab2<-myDat[sex==1, sapply(.SD, sd),.SDcols=myCols]
tab3<-myDat[sex==1, sapply(.SD, var),.SDcols=myCols]
tab_male<-rbind(tab1,tab2,tab3)
rownames(tab_male)[c(7,8)]<-c("SD","Var")

tab1<-myDat[sex==2, sapply(.SD, summary),.SDcols=myCols]
tab2<-myDat[sex==2, sapply(.SD, sd),.SDcols=myCols]
tab3<-myDat[sex==2, sapply(.SD, var),.SDcols=myCols]
tab_female<-rbind(tab1,tab2,tab3)
rownames(tab_female)[c(7,8)]<-c("SD","Var")
```

Aufgabe 1: Lösung a) - Deskription Männer

Tabelle 1: Deskriptive Statistiken - Männer

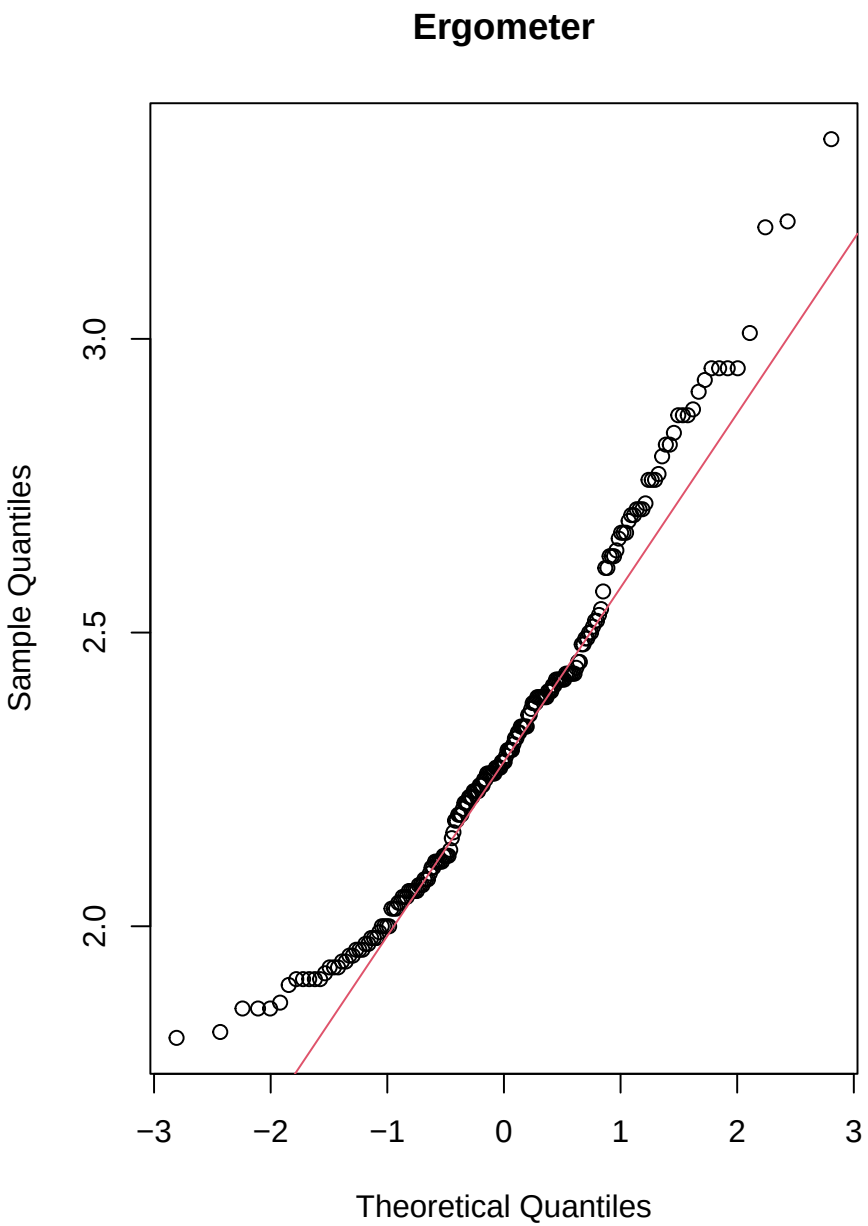
	Ergometer	Laktat	Alter	BMI
Min.	1.81	8.00	51.93	21.05
1st Qu.	2.07	15.00	58.92	23.85
Median	2.25	17.00	67.91	25.29
Mean	2.28	17.09	66.11	25.25
3rd Qu.	2.42	19.00	72.15	26.79
Max.	3.20	26.00	76.90	30.52
SD	0.29	3.17	6.99	2.10
Var	0.08	10.06	48.84	4.42

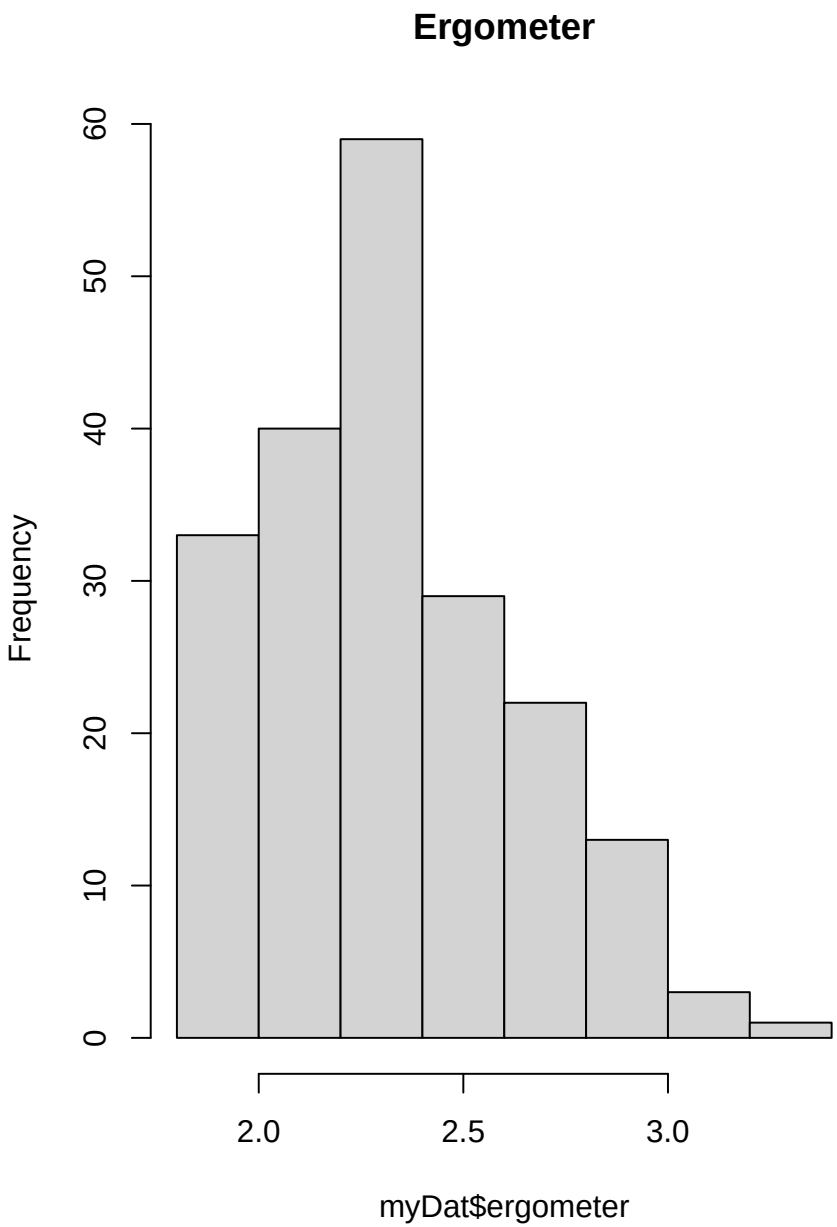
Aufgabe 1: Lösung a) - Deskription Frauen

Tabelle 2: Deskriptive Statistiken - Frauen

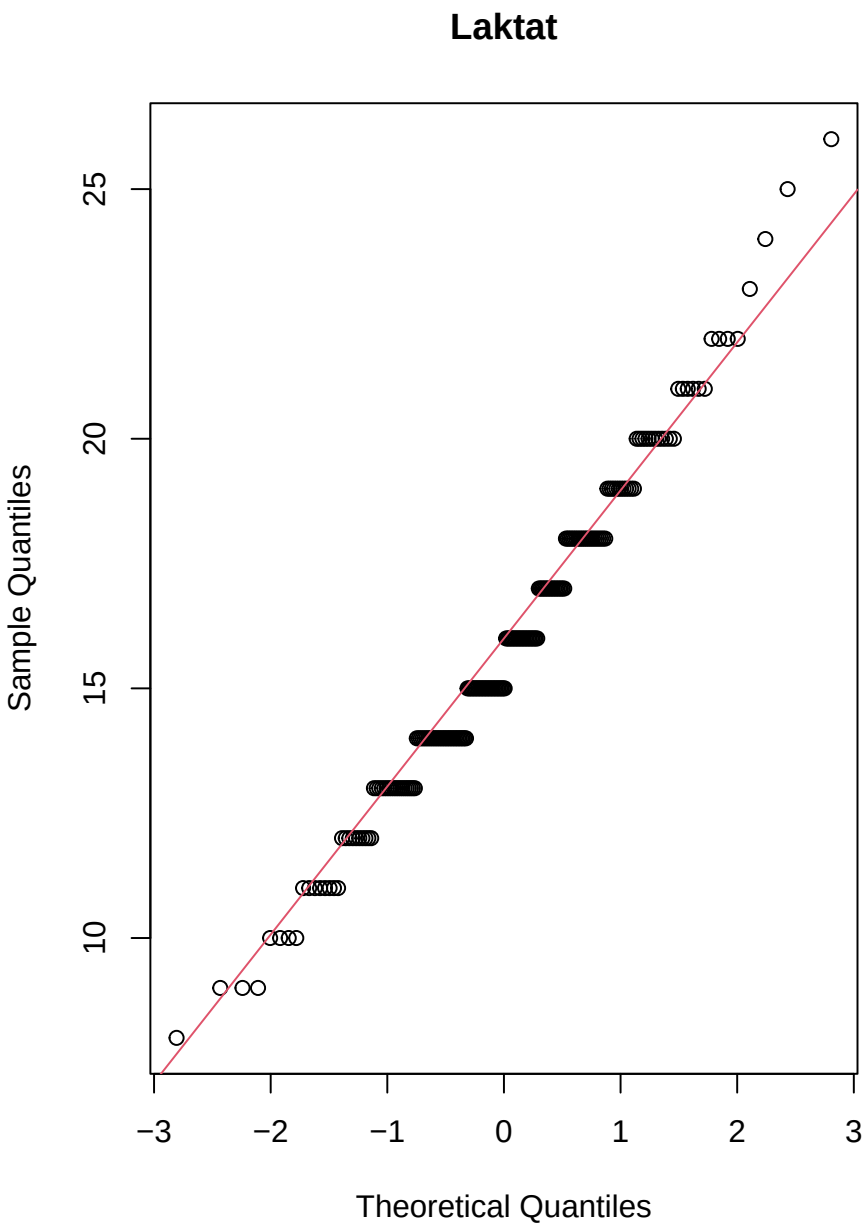
	Ergometer	Laktat	Alter	BMI
Min.	1.86	9.00	46.94	19.46
1st Qu.	2.10	13.00	58.92	22.20
Median	2.34	14.50	63.91	23.24
Mean	2.36	14.68	63.64	23.28
3rd Qu.	2.51	16.00	69.91	24.55
Max.	3.34	24.00	76.90	27.58
SD	0.32	2.75	7.52	1.78
Var	0.10	7.59	56.61	3.18

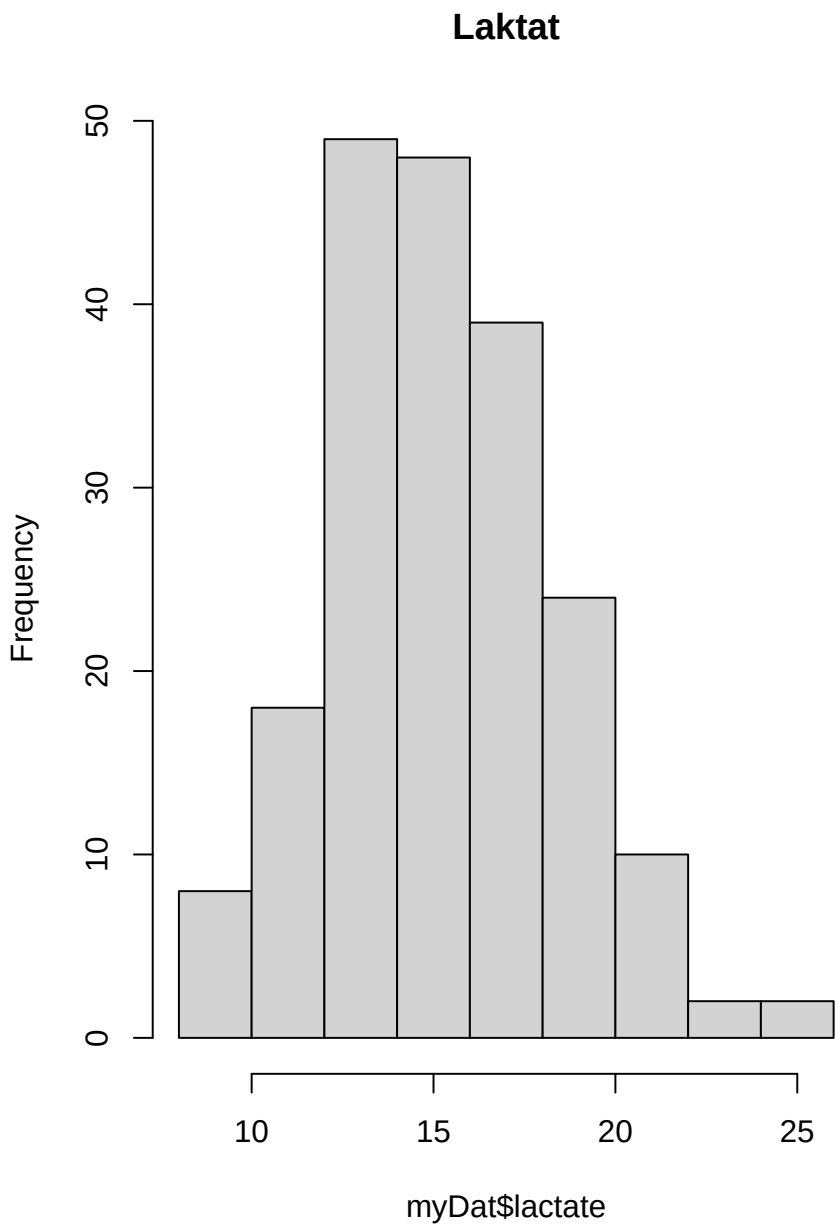
Aufgabe 1: Lösung b) - Plots Ergometer



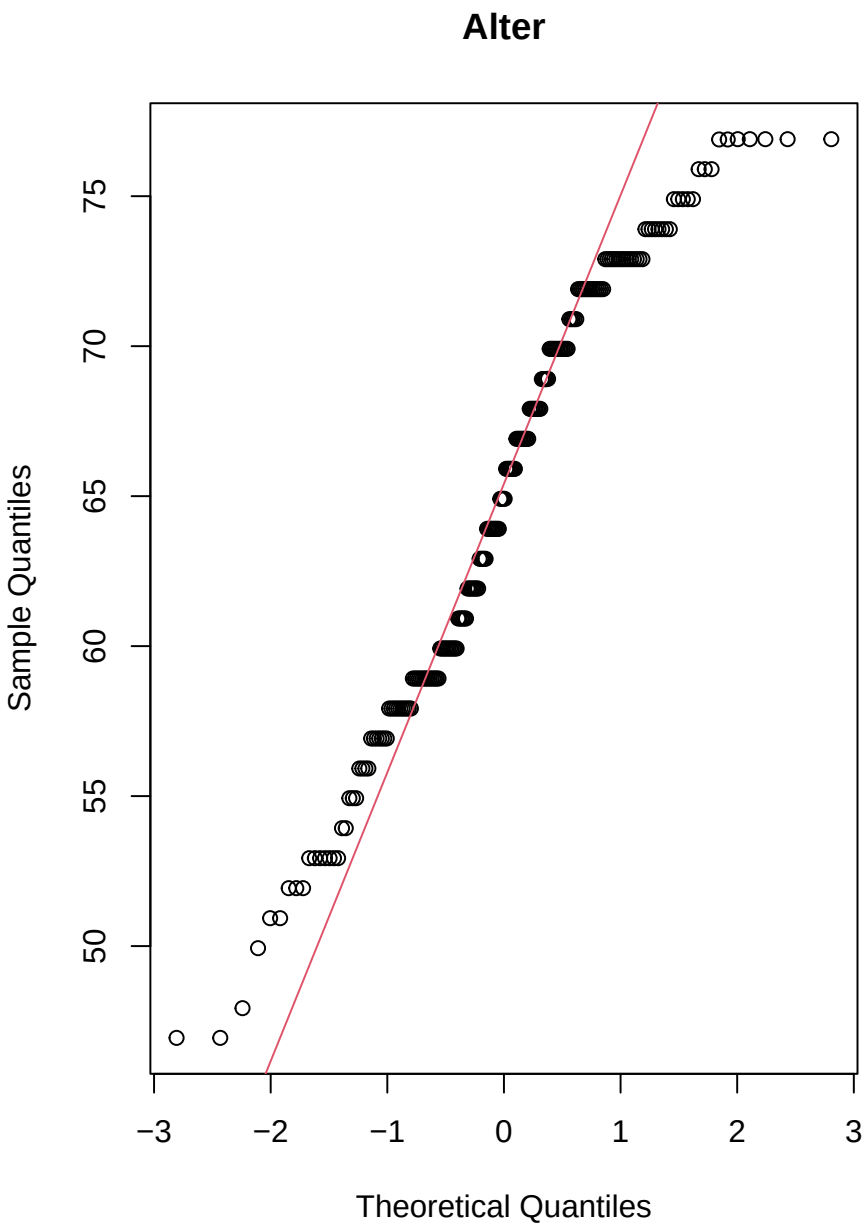


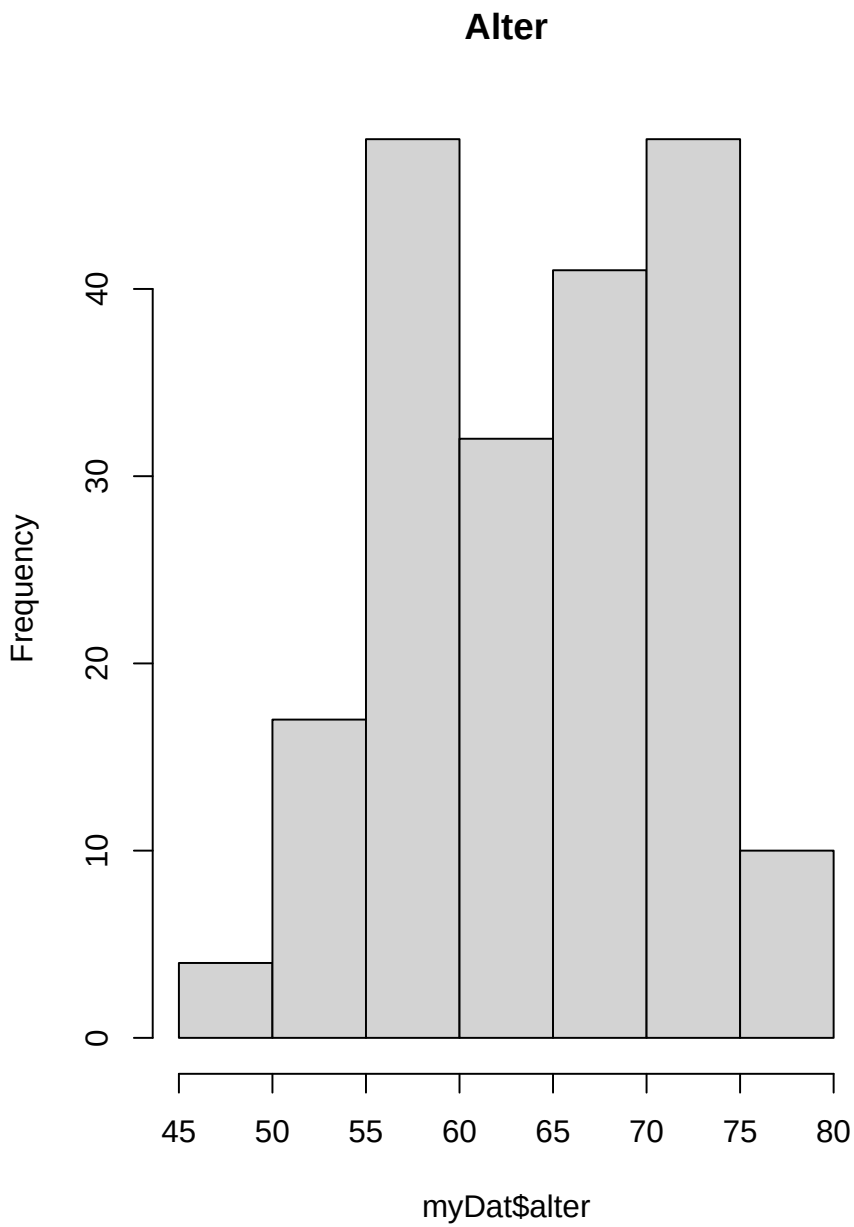
Aufgabe 1: Lösung b) - Plots Laktat



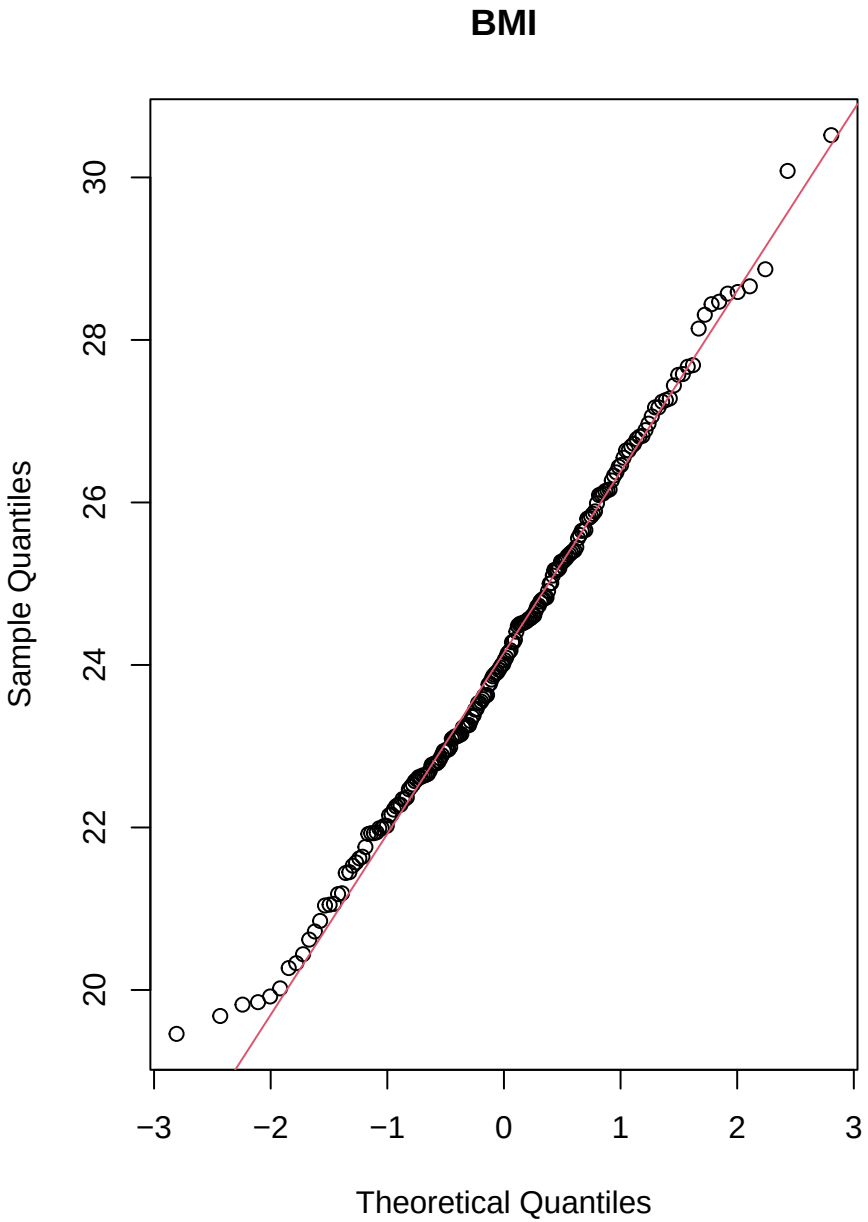


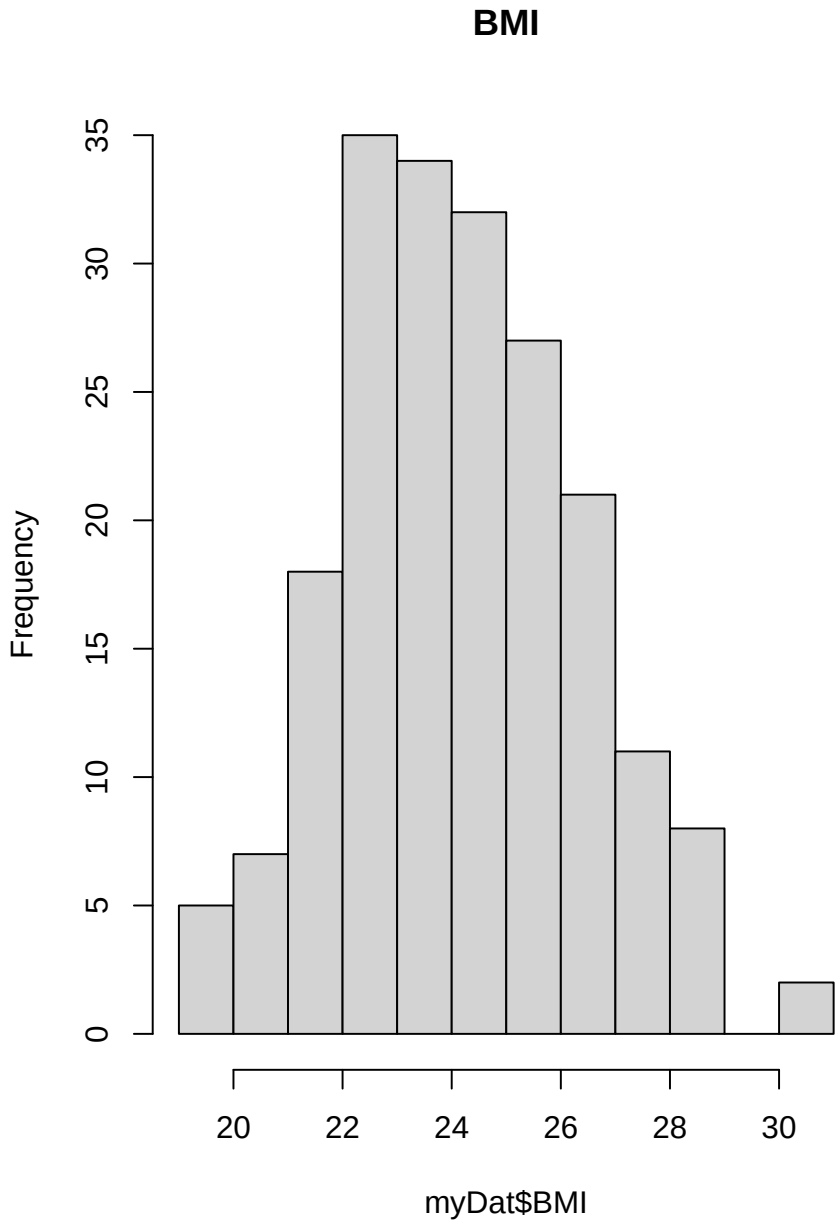
Aufgabe 1: Lösung b) - Plots Alter





Aufgabe 1: Lösung b) - Plots BMI





Aufgabe 1: Lösung b) - Test Normalverteilung

```
p1 = ks.test(myDat$ergometer,
             pnorm,
             mean=mean(myDat$ergometer),
             sd=sd(myDat$ergometer))
p2 = ks.test(myDat$lactate,pnorm,mean=mean(myDat$lactate),sd=s
p3 = ks.test(myDat$alter,pnorm,mean=mean(myDat$alter),sd=sd(my
p4 = ks.test(myDat$BMI,pnorm,mean=mean(myDat$BMI),sd=sd(myDat$

tab4 = data.table(parameter =  c("Ergometer","Laktat",
                                 "Alter","BMI"),
                  KS_Test = c(p1$p.value,p2$p.value,
                              p3$p.value,p4$p.value))
```

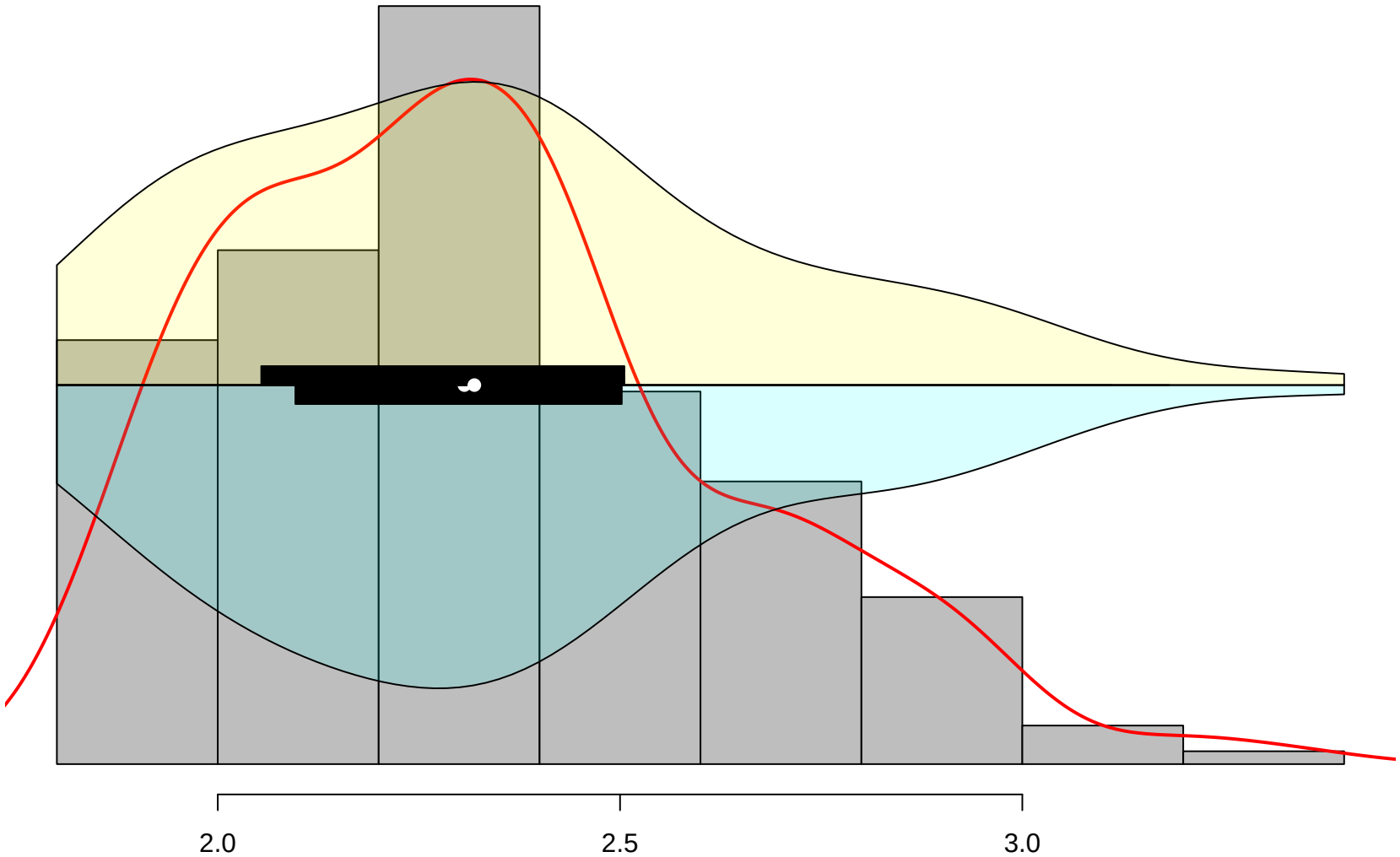
Aufgabe 1: Lösung b) - Test Normalverteilung

```
knitr::kable(t(tab4[,2]), position = "!b",digits =4,
             caption = "Kolmogorov-Smirnov Test auf Normalvert
             col.names = c("Ergometer","Laktat","Alter","BMI")
```

Tabelle 3: Kolmogorov-Smirnov Test auf Normalverteilung

	Ergometer	Laktat	Alter	BMI
KS_Test	0.0666	0.0301	0.0202	0.6896

Aufgabe 1: Lösung c) - Vergleich Ergometer



Aufgabe 1: Lösung c) - Vergleich Ergometer

Welche Annahmen müssen bei einem Vergleich zweier Stichproben gelten?

Aufgabe 1: Lösung c) - Vergleich Ergometer

Welche Annahmen müssen bei einem Vergleich zweier Stichproben gelten?

1. Die zwei Gruppen sind unabhängig voneinander
2. Die zwei Gruppen haben gleiche Varianz oder Streuung
3. Die zwei Gruppen sind normal-verteilt

Aufgabe 1: Lösung c) - Vergleich Ergometer

Welche Annahmen müssen bei einem Vergleich zweier Stichproben gelten?

1. Die zwei Gruppen sind unabhängig voneinander
   - Check, Männer & Frauen sind unabhängig
2. Die zwei Gruppen haben gleiche Varianz oder Streuung
   - Check, visuell via Violinplot; 1a) Varianz sehr ähnlich
3. Die zwei Gruppen sind normal-verteilt

Aufgabe 1: Lösung c) - Vergleich Ergometer

Welche Annahmen müssen bei einem Vergleich zweier Stichproben gelten?

1. Die zwei Gruppen sind unabhängig voneinander
   - Check, Männer & Frauen sind unabhängig
2. Die zwei Gruppen haben gleiche Varianz oder Streuung
   - Check, visuell via Violinplot; 1a) Varianz sehr ähnlich
3. Die zwei Gruppen sind normal-verteilt
   - Check, One-sample Kolmogorov-Smirnov test p-Wert = 0.06 –> Normalverteilung kann nicht abgelehnt werden

–> Student's t-Test

Aufgabe 1: Lösung c) - Vergleich Ergometer

1. Die zwei Gruppen sind unabhängig voneinander
   - Check, Männer & Frauen sind unabhängig
2. Die zwei Gruppen haben gleiche Varianz oder Streuung
   - Check, visuell via Violinplot; 1a) Varianz sehr ähnlich
3. Die zwei Gruppen sind normal-verteilt
   - One-sample Kolmogorov-Smirnov test p-Wert ist grenzwertig

–> Wilcoxon Rank Sum Test

Aufgabe 1: Lösung c) - Vergleich Ergometer

–> Bekannte Verteilung

–> Bekannte Mathematische Formeln

–> Bekannte Parameter, die die Form der Verteilung bestimmen

–> Student's t-Test: parameter-abhängiger Test

`Nullhypothese: Die Mittelwerte der zwei Gruppen sind gleich.`

–> Wilcoxon Rank Sum Test: parameter-freier Test

`Nullhypothese: Die Mediane der zwei Gruppen sind gleich.`

Aufgabe 1: Lösung c) - Student's t-Test

```
t.test(myDat$ergometer ~ myDat$sex)
```

```
##
##  Welch Two Sample t-test
##
## data:  myDat$ergometer by myDat$sex
## t = -1.8014, df = 197.08, p-value = 0.07317
## alternative hypothesis: true difference in means between gr
## 95 percent confidence interval:
##  -0.160775924  0.007273509
## sample estimates:
## mean in group 1 mean in group 2
##        2.281304        2.358056
```

Aufgabe 1: Lösung c) - Wilcoxon Rank Sum Test

```
wilcox.test(myDat$ergometer ~ myDat$sex)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  myDat$ergometer by myDat$sex
## W = 4251, p-value = 0.07899
## alternative hypothesis: true location shift is not equal to
```

Aufgabe 1: Lösung d) - Korrelation

Alter & Laktat sind nicht normalverteilt –> Spearmans rank correlation

```
# Korrelation
# All credits to https://www.r-bloggers.com/more-on-exploring-
cor.prob <- function (X, dfr = nrow(X) - 2) {
  R <- cor(X, use="pairwise.complete.obs",method="spearman")
  above <- row(R) < col(R)
  r2 <- R[above]^2
  Fstat <- r2 * dfr/(1 - r2)
  R[above] <- 1 - pf(Fstat, 1, dfr)
  R[row(R) == col(R)] <- NA
  R
}
```

Aufgabe 1: Lösung d) - Korrelation

```
corTab<-cor.prob(myDat[,7:10])
knitr::kable(corTab, position = "!b",digits =5,
             caption = "Correlation - spearmans rho \ p-value"
             col.names = c("Ergometer","Laktat","Alter","BMI")
```

Tabelle 4: Correlation - spearmans rho p-value

	Ergometer	Laktat	Alter	BMI
ergometer	NA	0.00000	0.0000	0.62813
lactate	-0.47388	NA	0.0000	0.00000
alter	-0.55685	0.76555	NA	0.39244
BMI	-0.03445	0.60313	0.0608	NA

Aufgabe 1: Lösung d) - Korrelation

```
myDat[,watt:=ergometer*weight]
corTab2<-cor.prob(myDat[,7:11])
knitr::kable(corTab2, position = "!b",digits =5,
             caption = "Correlation - spearmans rho \ p-value"
             col.names = c("Ergometer","Laktat","Alter","BMI",
```

Tabelle 5: Correlation - spearmans rho p-value

	Ergometer	Laktat	Alter	BMI	Watt
ergometer	NA	0.00000	0.00000	0.62813	0.00000
lactate	-0.47388	NA	0.00000	0.00000	0.10158
alter	-0.55685	0.76555	NA	0.39244	0.00001
BMI	-0.03445	0.60313	0.06080	NA	0.00000
watt	0.62047	0.11611	-0.31246	0.63012	NA

Aufgabe 1: Lösung c) - Student's t-Test - Wdh

```
t.test(myDat$watt ~ myDat$sex)
```

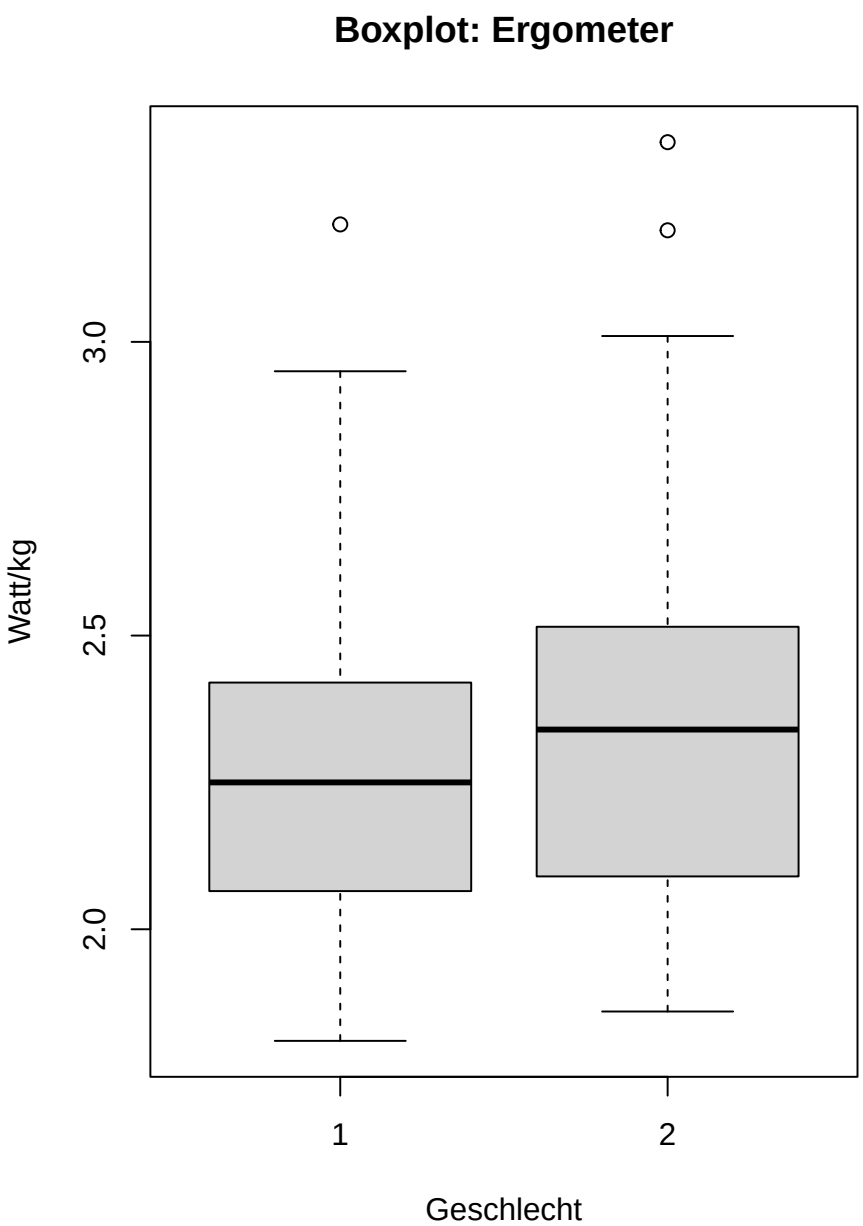
```
##
##  Welch Two Sample t-test
##
## data:  myDat$watt by myDat$sex
## t = 6.5682, df = 188.57, p-value = 4.821e-10
## alternative hypothesis: true difference in means between gr
## 95 percent confidence interval:
##  19.04469 35.39443
## sample estimates:
## mean in group 1 mean in group 2
##        180.0949        152.8754
```

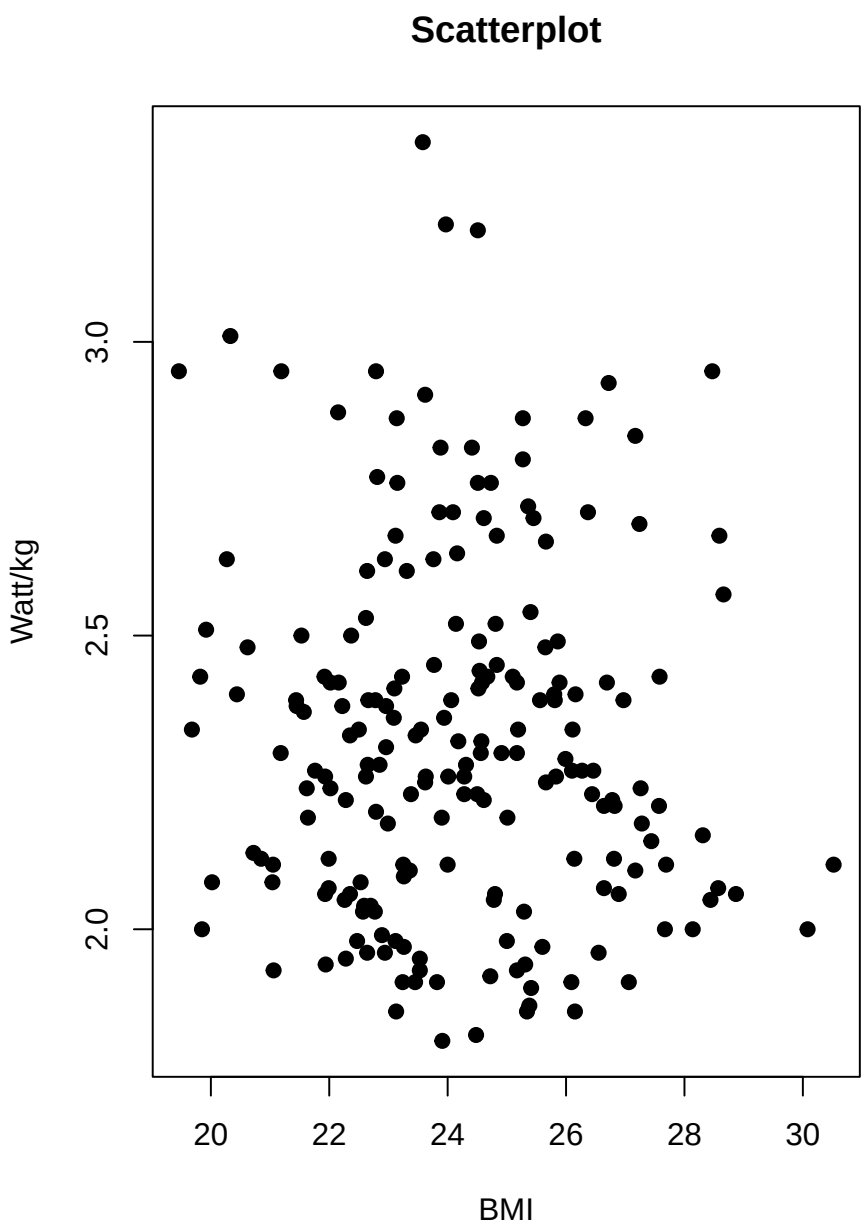
Aufgabe 1: Lösung c) - Wilcoxon Rank Sum Test-Wdh

```
wilcox.test(myDat$watt ~ myDat$sex)
```

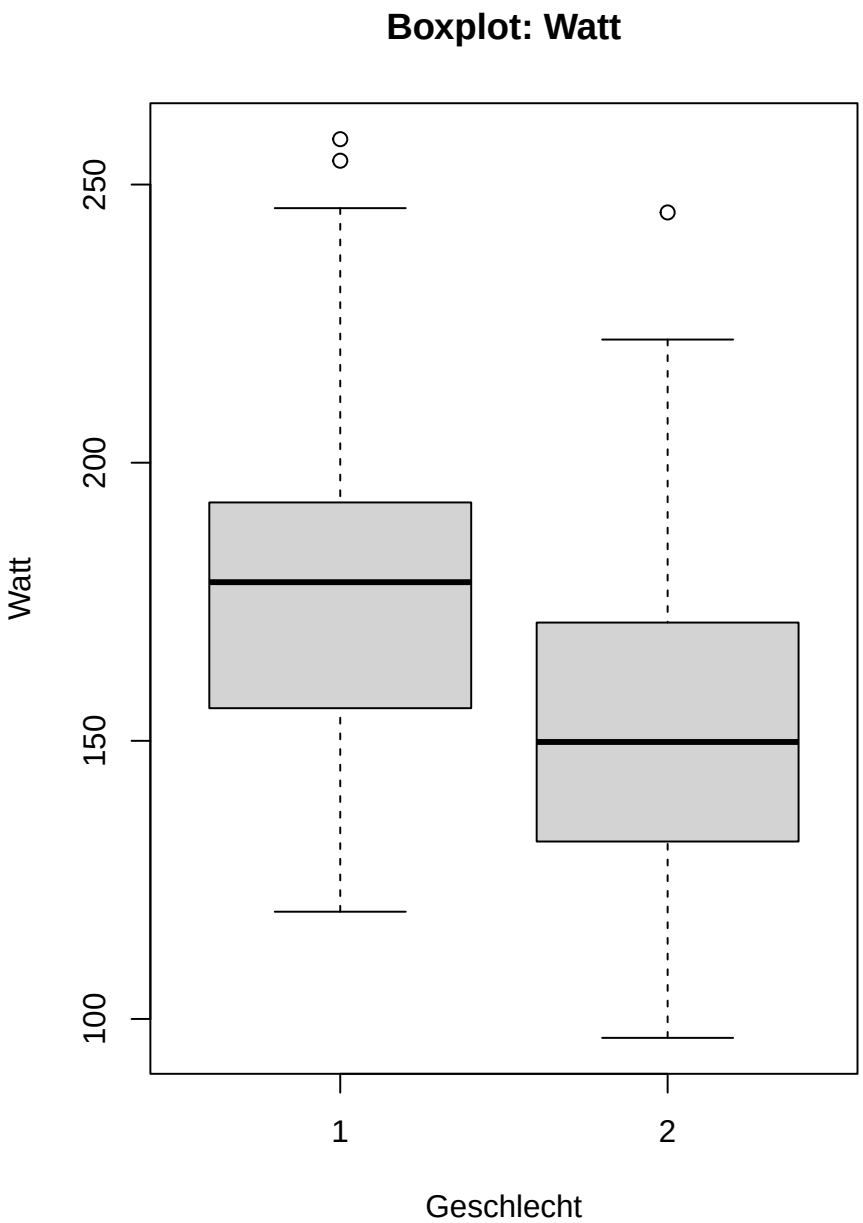
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  myDat$watt by myDat$sex
## W = 7410.5, p-value = 2.151e-09
## alternative hypothesis: true location shift is not equal to
```

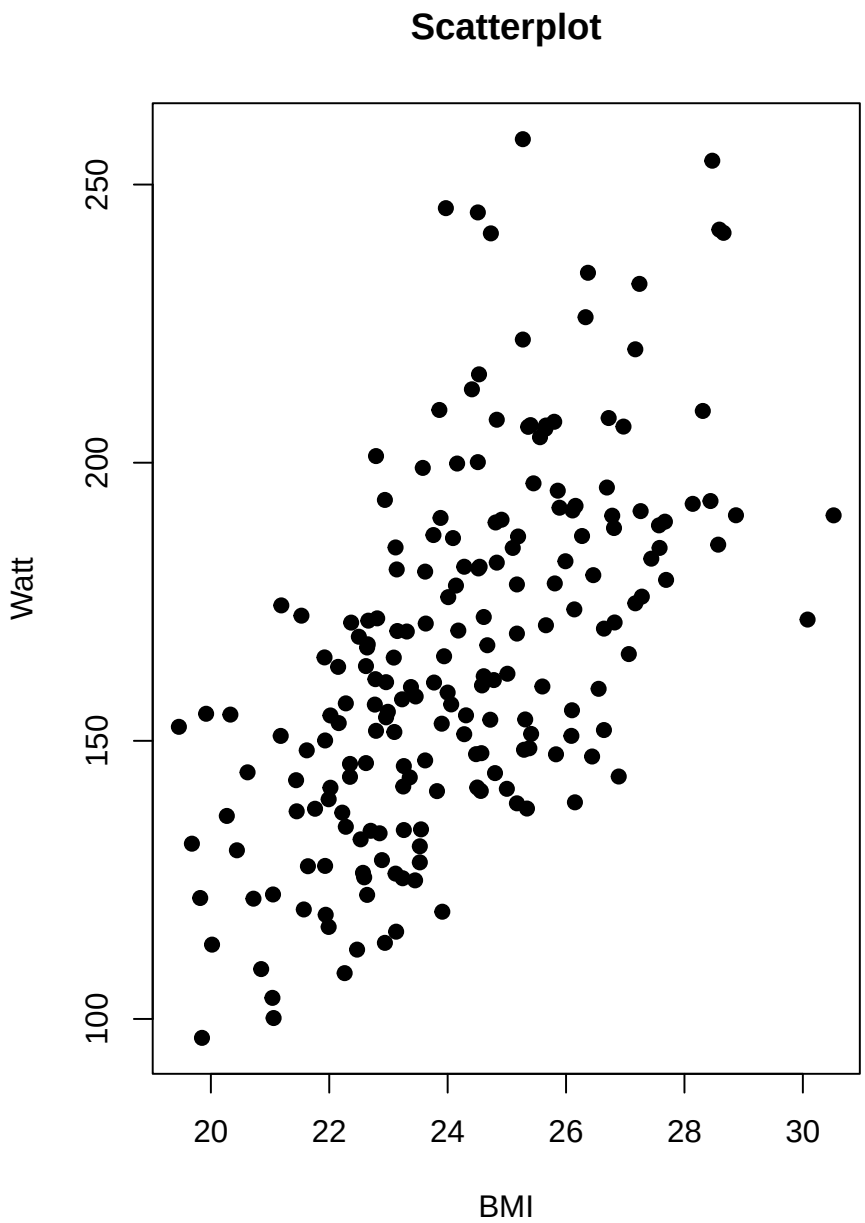
Aufgabe 1: Lösung d) - Korrelation





Aufgabe 1: Lösung d) - Korrelation





Aufgabe 2: Gepaarte Tests

- Deskription
- Test Handlängenunterschied zwischen Männern & Frauen
- Test Längenunterschiede zwischen Schreib- und Nichtschreibhanden für Männer und Frauen getrennt
- Beziehung zwischen Schreibhand, Armverschränkung und Klatschen
- Beziehungen zwischen Größe, Länge der Hand und Unterschied zwischen Schreib-/Nichtschreibhand für Männer und Frauen getrennt

Aufgabe 2: Lösung a) - Deskription binär

Tabelle 6: Händigkeit

	rechts	links
Männer	106	10
Frauen	110	7

Tabelle 7: Klatschen & Armverschränken

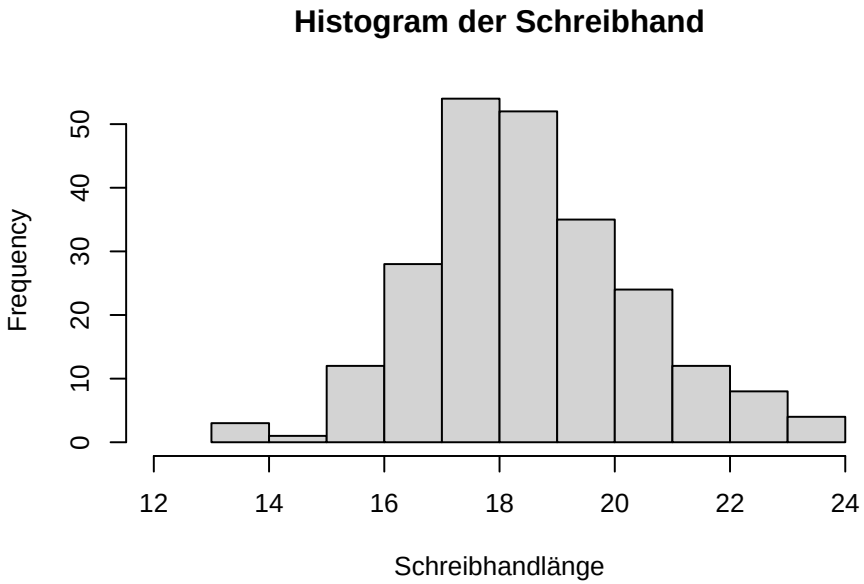
	rechts	egal	links
rechts	71	26	22
egal	16	2	0
links	60	20	16

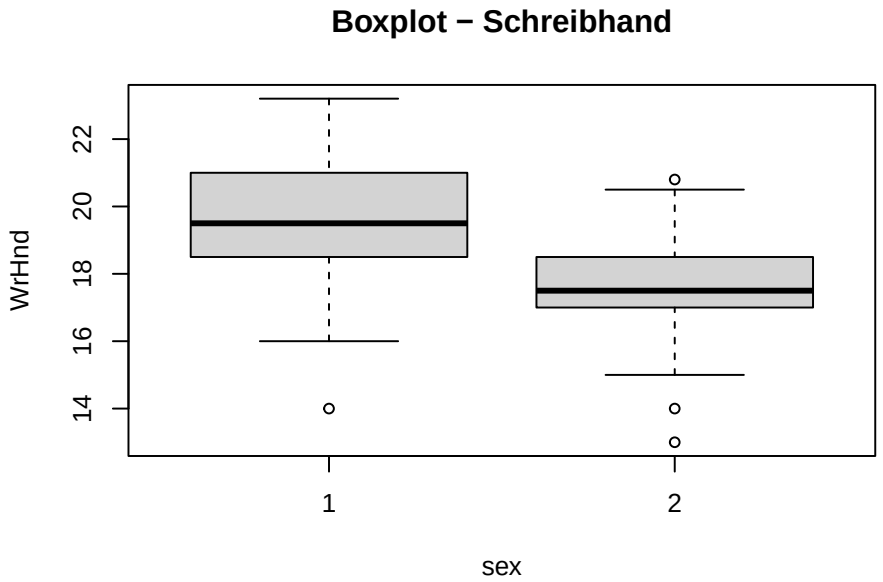
Aufgabe 2: Lösung a) - Deskription kontinuierlich

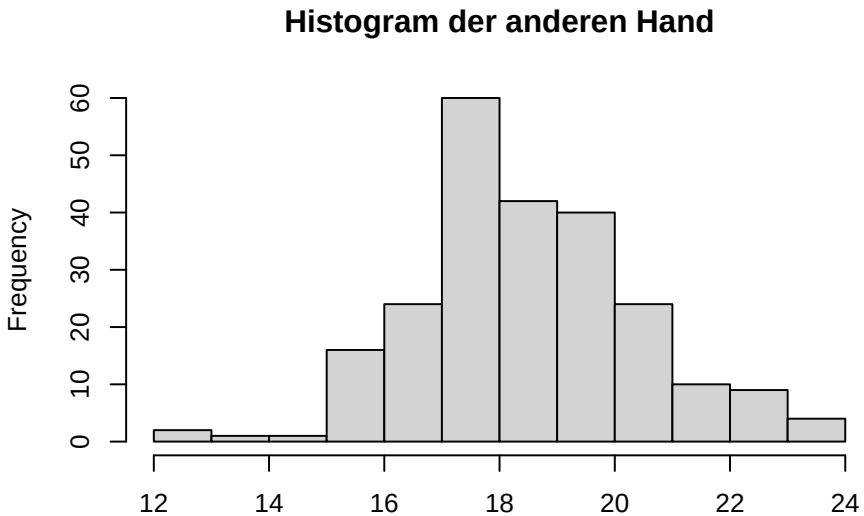
Tabelle 8: Deskriptive Statistiken

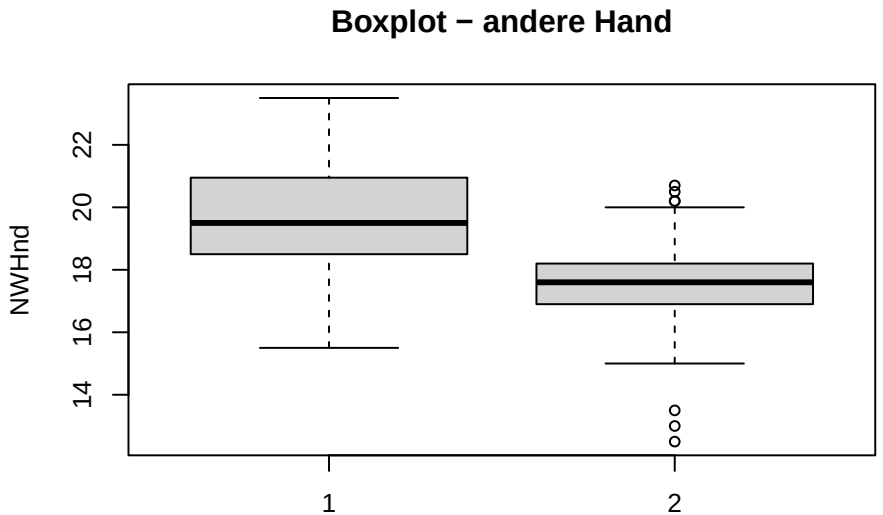
	Länge Schreibhand	Länge der Nichtschreibhand	Größe
Min.	13.000	12.500	1.500
1st Qu.	17.500	17.500	1.650
Median	18.500	18.500	1.710
Mean	18.691	18.627	1.723
3rd Qu.	19.800	19.800	1.800
Max.	23.200	23.500	2.000
SD	1.852	1.914	0.099
Var	3.430	3.662	0.010

Aufgabe 2: Lösung b) - Handlängenunterschiede Männer vs Frauen









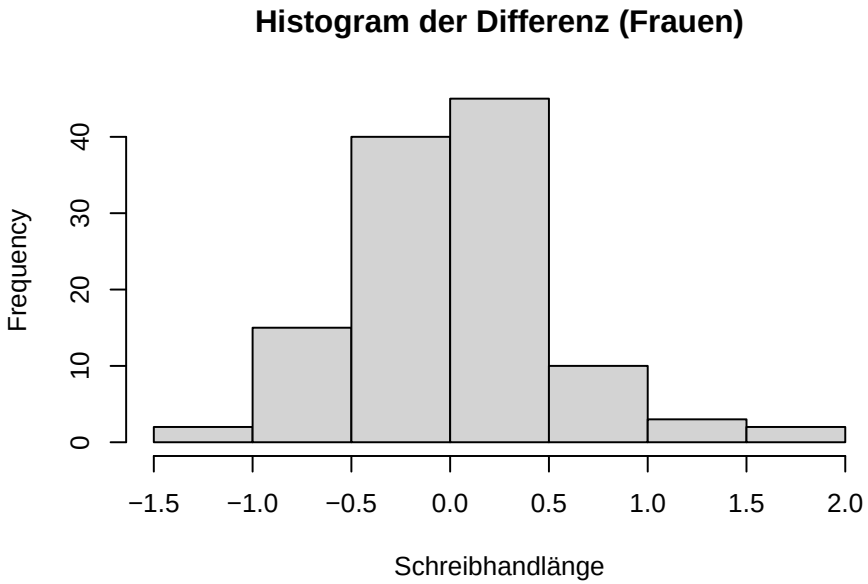
Aufgabe 2: Lösung b) - Handlängenunterschiede Männer vs Frauen

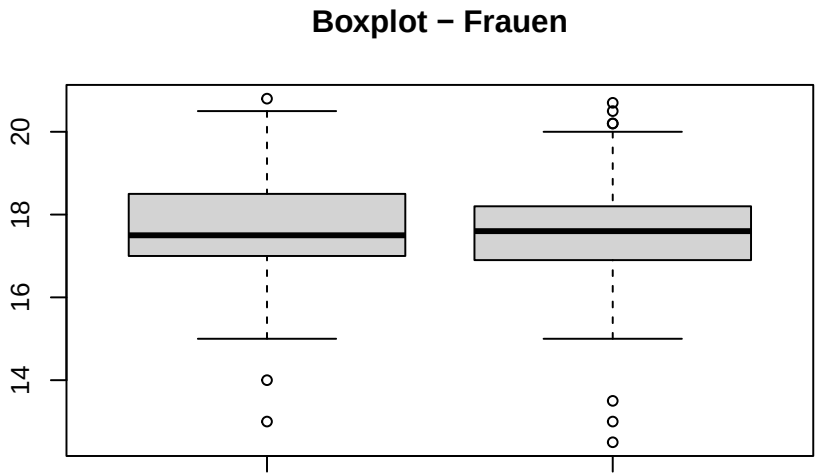
```
p1 = t.test(WrHnd ~ sex)
p2 = t.test(NWHnd ~ sex)
tab5 = data.table(parameter =  c("WrHnd","NWHnd"),
                  t_Test = c(p1$p.value,p2$p.value))
tab5$t_Test<- format(tab5$t_Test, digits = 3)
knitr::kable(t(tab5[,2]), position = "!b",#digits =8,
             caption = "Student's t-Test",
             col.names = c("WrHnd","NWHnd"))
```

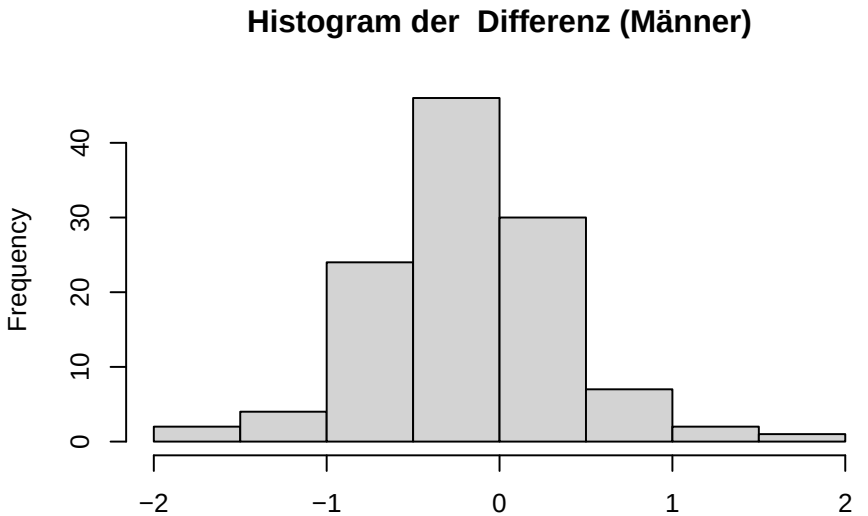
Tabelle 9: Student's t-Test

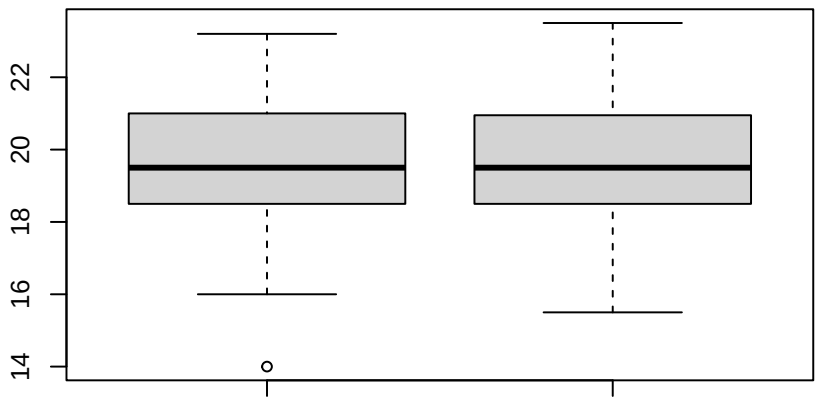
	WrHnd	NWHnd
t_Test	2.16e-21	1.71e-23

Aufgabe 2: Lösung c) - Handlängenunterschiede Schreib -vs Nichtschreibhand









Aufgabe 2: Lösung c) - Handlängenunterschiede Schreib -vs Nichtschreibhand

```
female<-sex==2
p1 = t.test(WrHnd[female],NWHnd[female],paired = T)
p2 = t.test(WrHnd[!female],NWHnd[!female],paired = T)

tab6 = data.table(parameter =  c("WrHnd","NWHnd"),
                  t_Test = c(p1$p.value,p2$p.value))
tab6$t_Test<- format(tab6$t_Test, digits = 3)
knitr::kable(t(tab6[,2]), position = "!b",#digits =8,
             caption = "Student's t-Test der Differenz",
             col.names = c("Frauen","Männer"))
```

Tabelle 10: Student's t-Test der Differenz

	Frauen	Männer
t_Test	0.00587	0.79971

Aufgabe 2: Lösung d) - Beziehung binäre Variablen

Tabelle 11: Fisher's Exact Test

	Fold vs Clap	Fold vs Whnd	Clap vs Whnd
Fisher_Test	0.714998	0.434604	0.000106

Aufgabe 2: Lösung e) - Korrelation Frauen

Tabelle 12: Correlation - spearmans rho p-value

	height	WrHnd	NWHnd	dif
height	NA	0.00059	0.00476	0.42147
WrHnd	0.33591	NA	0.00000	0.87765
NWHnd	0.27874	0.91875	NA	0.00009
dif	0.08086	0.01551	-0.38055	NA

Aufgabe 2: Lösung e) - Korrelation Männer

```
corTab<-cor.prob(myDat2[!female & filt,c(8,3,4,9)])
knitr::kable(corTab, position = "!b",digits =5,
             caption = "Correlation - spearmans rho \ p-value"
             col.names = c("height","WrHnd","NWHnd","dif"))
```

Tabelle 13: Correlation - spearmans rho p-value

	height	WrHnd	NWHnd	dif
height	NA	0.00005	0.00027	0.24226
WrHnd	0.38507	NA	0.00000	0.11309
NWHnd	0.34874	0.95069	NA	0.10632
dif	0.11512	0.15555	-0.15850	NA

Abschnitt 2

Zusammenfassung

Zusammenfassung

- Warum ist Testen auf Normalverteilung wichtig?
- Warum macht es einen Unterschied ob man gepaarte Daten hat oder nicht?