Hausübung 3

Reutterer Maximilian, Sattler Lukas, Weinzierl Jakob

2023-11-26

Für alle Übungen führen Sie als ersten Schritt eine Datenexploration durch, um die erforderlichen Annahmen des Tests zu Überprüfen. Dann entscheiden Sie, welche Test geeignet ist, um die Fragestellung zu beantworten und mit den gebenen Daten kompatibel ist. Schreiben Sie den Ansatz für das Test Problem (parametrisch, nichtparametrisch, resampling; welche Verteilungsannahmen an die Daten gelten), die Nullhypothese und Alternativhypothese explizit an. Aus dem Testoutput führen Sie den Wert der Teststatistik explizit an und welche Verteilung diese haben soll. Lesen Sie den p-Wert ab und argumentieren anhand dieses Wertes, welche Entscheidung Sie treffen.

Aufgabe 1

Ein Labor schickt seine Mitarbeiten zu einem Pipettiertraining und möchte anschließend testen, ob sich dieses ausgezahlt hat, indem die mittleren Zeiten zur Durchführen von 25 Pipettiervorgängen vor und nach dem Training gemessen werden.

Before training: 1.36, 1.37, 1.29, 1.22, 1.38, 1.31, 1.40, 1.39, 1.30, 1.37 After training: 1.29, 1.25, 1.20, 1.26, 1.25, 1.23, 1.26, 1.31, 1.24, 1.31

Hatte das Training irgendeinen Effekt? Sollte die Firma, die das Labor betreibt, die Mitarbeiter anderer Labors zu einem solchen Training schicken, um ihre mittlere Arbeitszeit zu verringern? Beantworten Sie diese Fragen auf dem 5% und 1% Niveau.

```
before <- c(1.36, 1.37, 1.29, 1.38, 1.31, 1.40, 1.39, 1.30, 1.37)
after <- c(1.29, 1.25, 1.20, 1.26, 1.25, 1.23, 1.26, 1.31, 1.24, 1.31)
summary(before)
```

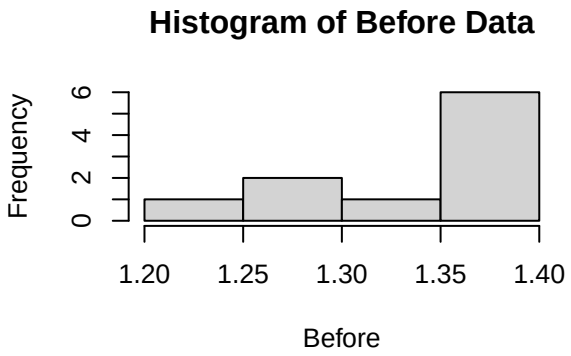
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.290   1.310   1.370   1.352   1.380   1.400
```

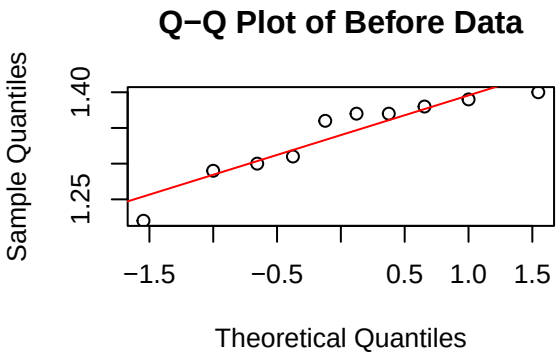
```
summary(after)
```

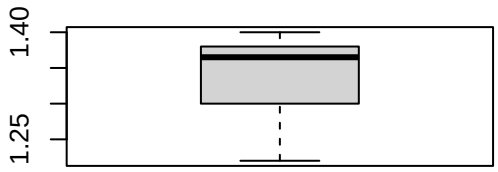
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.242   1.255   1.260   1.282   1.310
```

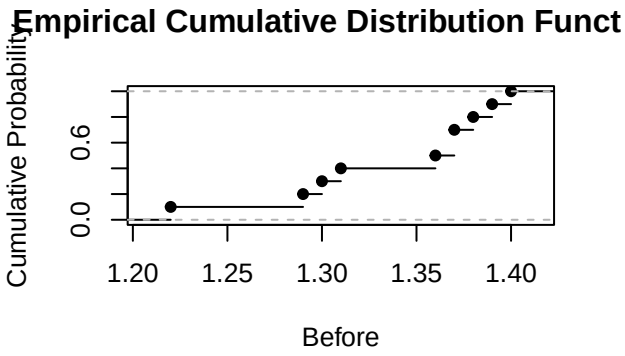
```
library(stats)

before <- c(1.36, 1.37, 1.29, 1.22, 1.38, 1.31, 1.40, 1.39, 1.30, 1.37)
after <- c(1.29, 1.25, 1.20, 1.26, 1.25, 1.23, 1.26, 1.31, 1.24, 1.31)

# Histogram and Q-Q plot for 'before' vector
par(mfrow=c(2, 2)) # Set up the plotting area

# Histogram for 'before'
```

```r
hist(before, main="Histogram of Before Data", xlab="Before")

# Q-Q plot for 'before'

qqnorm(before, main="Q-Q Plot of Before Data")
qqline(before, col="red")
boxplot(before)

ecdf_func <- ecdf(before)

plot(ecdf_func, xlab="Before", ylab="Cumulative Probability",
     main="Empirical Cumulative Distribution Function")
```

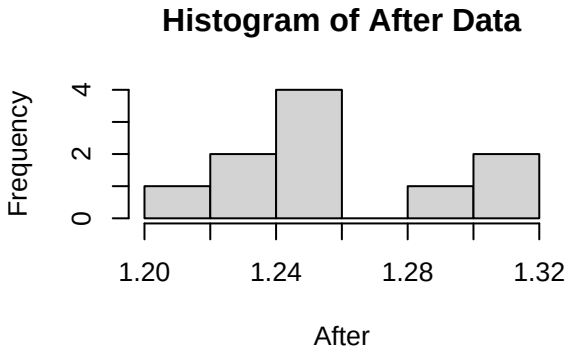


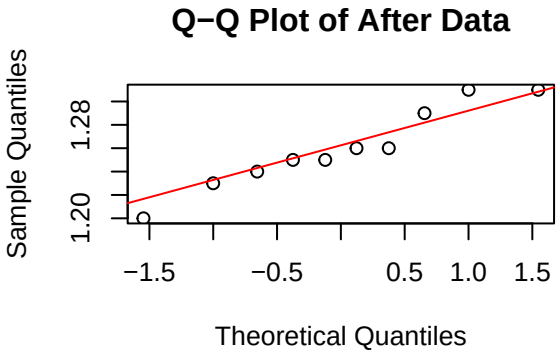


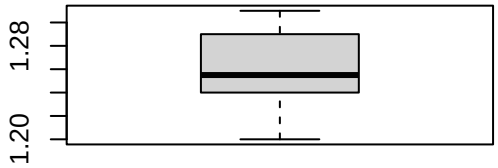


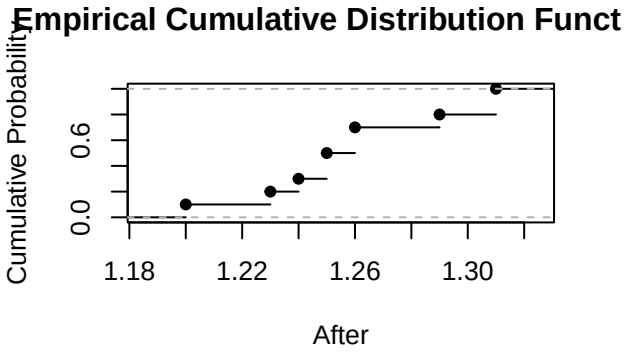



```r
# Histogram for 'after'
hist(after, main="Histogram of After Data", xlab="After")

# Q-Q plot for 'after'
qqnorm(after, main="Q-Q Plot of After Data")
qqline(after, col="red")
boxplot(after)

ecdf_func2 <- ecdf(after)

plot(ecdf_func2, xlab="After", ylab="Cumulative Probability",
     main="Empirical Cumulative Distribution Function")
```










```
par(mfrow=c(1, 1)) # Reset plotting area
```

** Für diese Fragestellung kann ein parametrischer Test durchgeführt werden, da wir annehmen, dass die Messzeiten eines Mitarbeiters einer Normalverteilung folgen.
Sind die Daten normalverteilt können parametrischen Tests wie der t-Test, die ANOVA oder die Pearson-Korrelation durchgeführt werden.**

Im ersten Schritt muss prizipiell geprüft werden, ob ein paramterischer (Anova, t-test) oder ein nicht-parametrischer test (willocoxon oder cruskal wallis test) verwendet werden darf. Dazu ahben wir die Daten als Histogram , Boxplot , QQPLOT und Ecdf dargestellt. Das Histogram lässt aufgrund der wenigen Datenpubnkte keine eindeutige Identifizieurng als Normalverteilung zu. Wir haben uns daher in der Verbesserung dazu entschieden,einen Wilcoxon-test durchzuführen, und anschleißend das Ergebnis mit einer Resampling-Simulation abzusichern. -> Ziehen mit Zurücklegen

Für die erste Hypothese fragen wir uns, ob das Training für den Mitarbeiter einen Effekt hatte. Dabei stellen wir folgende Nullhypothese auf, die keinen Unterschied zwischen den Mittelwerten der zwei Trainingszeiten feststellen soll:

$$H_0 : \mu_{before} = \mu_{after}$$

Die alternative Hypothese, dass das Training einen Unterschied ausmachte, heißt:

$$H_1 : \mu_{before} \neq \mu_{after}$$

Wir verwenden einen paired Wilcoxon-Rangusmmen-Test, da die Messungen von einem Mitarbeiter durchge-

führt wurden und wir den Vorher-Nachher-Effekt messen möchten. Da wir irgendeinen zeitlichen Effekt (schneller oder langsamer) testen möchten, wenden wir ihn in der two-tailed (zweiseitige) Variante an.

```
test_result <- wilcox.test(before, after, paired = TRUE)

# Print the results
print(test_result)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  before and after
## V = 54, p-value = 0.008004
## alternative hypothesis: true location shift is not equal to 0
```

Der Resultat zeigt, dass die Test-Statistik einen Wert von V = 54 hat und aufgrund des p-value = 0.008004 die Nullhypothese (es gibt keinen Effekt) sowohl auf 5% als auch auf 1% Signifikanzniveau verworfen werden kann. Die zentrale Tendenz der Daten (Median) ist daher unterschiedlich, und wir können H1 behalten, dass das Traning einen Effekt hatte.

** Ergänzung SS 2024 **

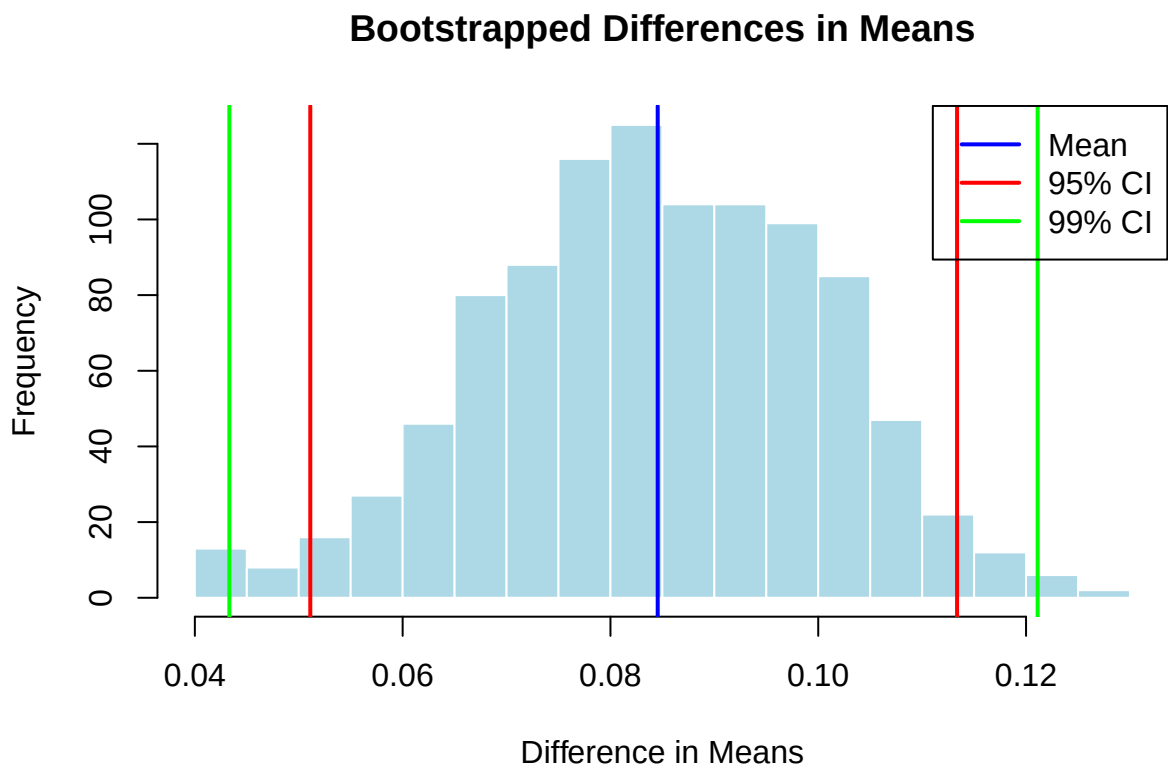
```
before_training <- c(1.36, 1.37, 1.29,  1.38, 1.31, 1.40, 1.39, 1.30, 1.37)
after_training <- c(1.29, 1.25,  1.26, 1.25, 1.23, 1.26, 1.31, 1.24, 1.31)

#Funktion zur Berechnung der Differenz der Mittelwerte
mean_diff <- function(data1, data2) {
  return(mean(data1) - mean(data2))
}

#Bootstrapping durchführen
set.seed(123)
n_boot <- 1000
boot_diffs <- numeric(n_boot)

for (i in 1:n_boot) {
  boot_before <- sample(before_training, length(before_training), replace = TRUE)
  boot_after <- sample(after_training, length(after_training), replace = TRUE)
  boot_diffs[i] <- mean_diff(boot_before, boot_after)
}




#Bootstrapped Konfidenzintervalle berechnen

boot_ci <- quantile(boot_diffs, c(0.025, 0.975))
boot_ci2 <- quantile(boot_diffs, c(0.005, 0.995))

boot_ci
```

```
##       2.5%      97.5%
## 0.05111111 0.11336111
```

```
boot_ci2
```

```
##       0.5%      99.5%
## 0.04332222 0.12111667
```

```
hist(boot_diffs, breaks = 30, main = "Bootstrapped Differences in Means",
     xlab = "Difference in Means", col = "lightblue", border = "white")
abline(v = boot_ci[1], col = "red", lwd = 2)
abline(v = boot_ci[2], col = "red", lwd = 2)

abline(v = boot_ci2[1], col = "green", lwd = 2)
abline(v = boot_ci2[2], col = "green", lwd = 2)


abline(v = mean(boot_diffs), col = "blue", lwd = 2)
legend("topright", legend = c("Mean", "95% CI", "99% CI"),
       col = c("blue", "red", "green"), lwd = 2)
```



Da wir hier im Vorfeld einen nicht-parametirschen test durchgeführt haben, wollten wir die zentrale Mittelwert-Tendenz im Anschluss mit Bootstrapping simulieren, um daraus einen direkten Verrgleich zu bekommen. Wir haben dafür ein Bootstrapping mit Resampling durchgeführt. Dabei haben wir 1000 als Anzahl der Resamplings gewählt. Hier sind die Intervallgrenzen für die erwartet Differenzt des Mittelwerts.

Da in beiden Intervallen 0 nicht enthalten ist, kann man hier erwarten, dass ein Unterschied zwischen den beiden Mittelwerten besteht (sowohl bei 95 als auch 99 CI)

Der Resampling Ansatz deckt sich also ergänzend mit der Annahme (H1) , die beim durchgeführten Wilxocon -Test behalten wurde.

```
2.5%      97.5%
```

0.05111111 0.11336111

```
  0.5%      99.5%
```

0.04332222 0.12111667

-> Bestätigung des hypothesentests durch simulationsverfahren

Aufgabe 2

Hatten auf der Titanic Frauen und Kinder eine signifikant (auf dem 1% Niveau) bessere Überlebenschance als Männer? (Tipp: Vergleichen Sie jeweils Frauen und Kinder separat.)

Da es sich hier um das Vergleichen zweier binomial verteilter Proportionen handelt, werden wir einen Proportionen-Test durchführen. Dabei nehmen wir ein Konfidenzniveau von 1% an.

Ob Frauen bzw. Kinder eine höhere Überlebenschance hatten als Männer, definieren wir die Nullhypothese folgendermaßen:

$$H_0 : P_{Frauen/Kinder} \leq P_{Männer}$$

Die alternative Hypothese dazu lautet:

$$H_1 : P_{Frauen/Kinder} > P_{Männer}$$

```
alle=apply(Titanic, c(3,4), sum); alle
```

```
##        Survived
## Age       No Yes
##   Child   52  57
##   Adult 1438 654
```

```
Kinder=apply(Titanic, c(3,4), sum)[1,]; Kinder
```

```
##  No Yes
##  52  57
```

```
FM=apply(Titanic, c(2,4), sum); FM
```

```
##         Survived
## Sex        No Yes
##   Male   1364 367
##   Female  126 344
```

```
Frauen=apply(Titanic, c(2,4), sum)[2,]; Frauen
```

```
##  No Yes
## 126 344
```

```
Männer=apply(Titanic, c(2,4), sum)[1,]; Männer
```

```
##   No  Yes
## 1364  367
```

```
prop.test(c(Frauen["Yes"],Männer["Yes"]),c(sum(Frauen),sum(Männer)),alternative = "greater", conf.level
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(Frauen["Yes"], Männer["Yes"]) out of c(sum(Frauen), sum(Männer))
## X-squared = 454.5, df = 1, p-value < 2.2e-16
```

```
## alternative hypothesis: greater
## 99 percent confidence interval:
##  0.4658044 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7319149 0.2120162
```

```
prop.test(c(Kinder["Yes"],Männer["Yes"]),c(sum(Kinder),sum(Männer)),alternative = "greater", conf.level
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(Kinder["Yes"], Männer["Yes"]) out of c(sum(Kinder), sum(Männer))
## X-squared = 54.16, df = 1, p-value = 9.24e-14
## alternative hypothesis: greater
## 99 percent confidence interval:
##  0.1924267 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.5229358 0.2120162
```

Im Fall der Frauen beträgt die Test-Statistik (mit Chi-Quadrat Verteilung) 454.5. Mit einem p-value von $<$ 2.2e-16 bei einem 1% Signifikanzniveau kann die Nullhypothese verworfen werden und daher hatten Frauen eine höhere Überlebenschance als Männer.

Bei den Kindern beträgt die Test-Statistik 54.16. Mit einem p-value von 9.24e-14 kann auch hier die Nullhypothese bei 1% Signifikanzniveau verworfen werden und daher hatten auch Kinder eine höhere Überlebenschance als Männer.

Aufgabe 3

Ein Biologe vergleicht die mittleren Wachstumsraten einer Bakterienkultur auf einer Petrischale über einen Zeitraum von 20 Minten minütlich. Es soll dabei untersucht werden, ob der Nährboden die Wachstumsrate gegenüber der durchschnittlich zu erwartenden Wachstumsrate von 1% fördert.

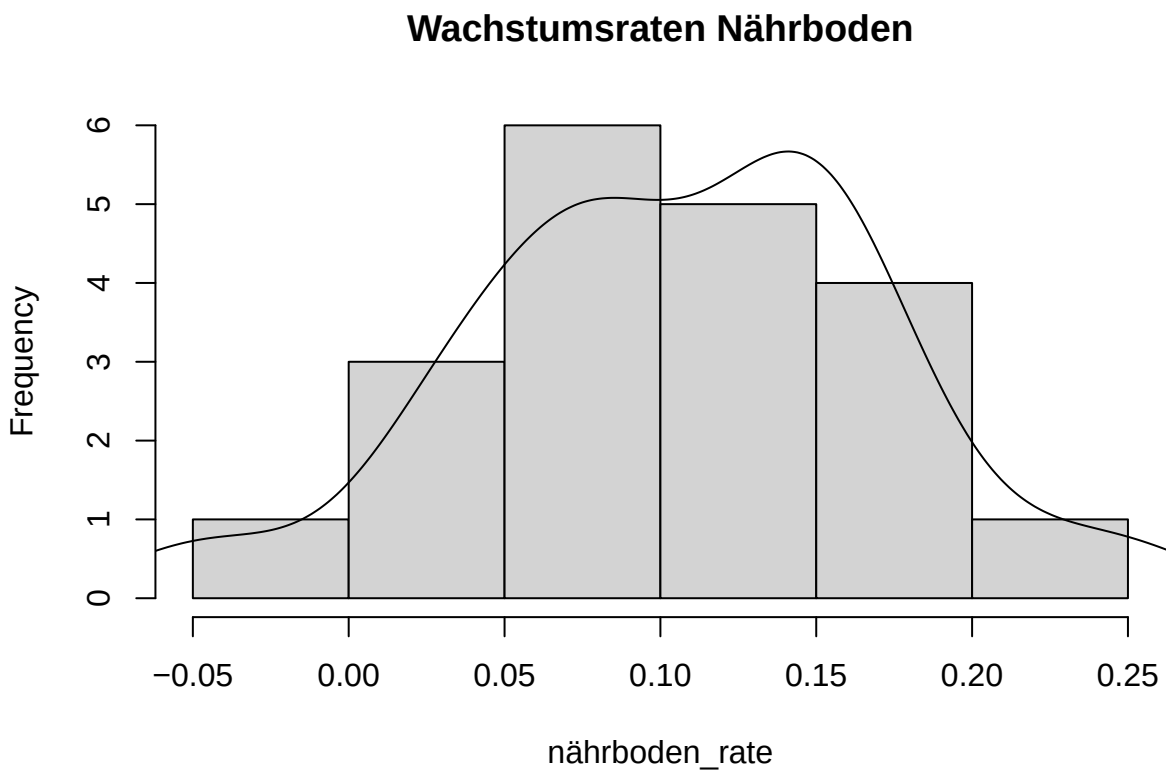
```
t = 20 #minuten
nährboden_rate <- c(0.146842, 0.156757, 0.091255, 0.063720, 0.148471,
            -0.045436, 0.150407, 0.077905, 0.077267, 0.026454,
            0.090700, 0.245384, 0.129650, 0.141617, 0.039957,
            0.165351, 0.029091, 0.073473, 0.189657, 0.123897)
normal_rate <- 0.01
```

Zuerst werden die Daten visualisiert, um auf Normalverteilung zu prüfen und ein Shapiro-Wilk Test durchgeführt

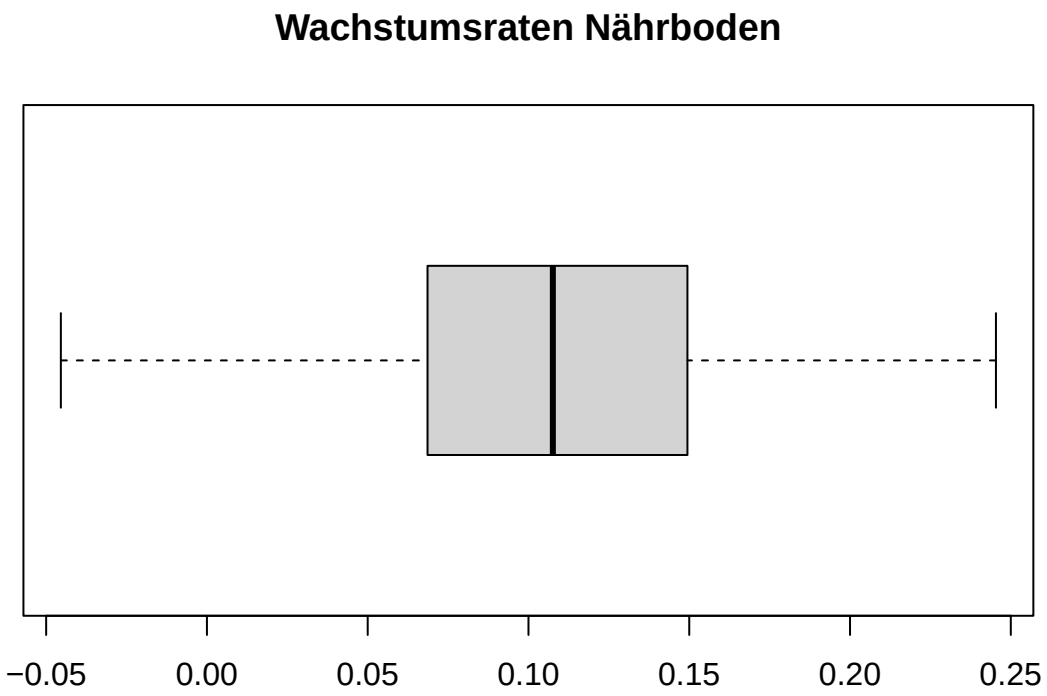
```
summary(nährboden_rate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.04544  0.07103  0.10758  0.10612  0.14896  0.24538
```

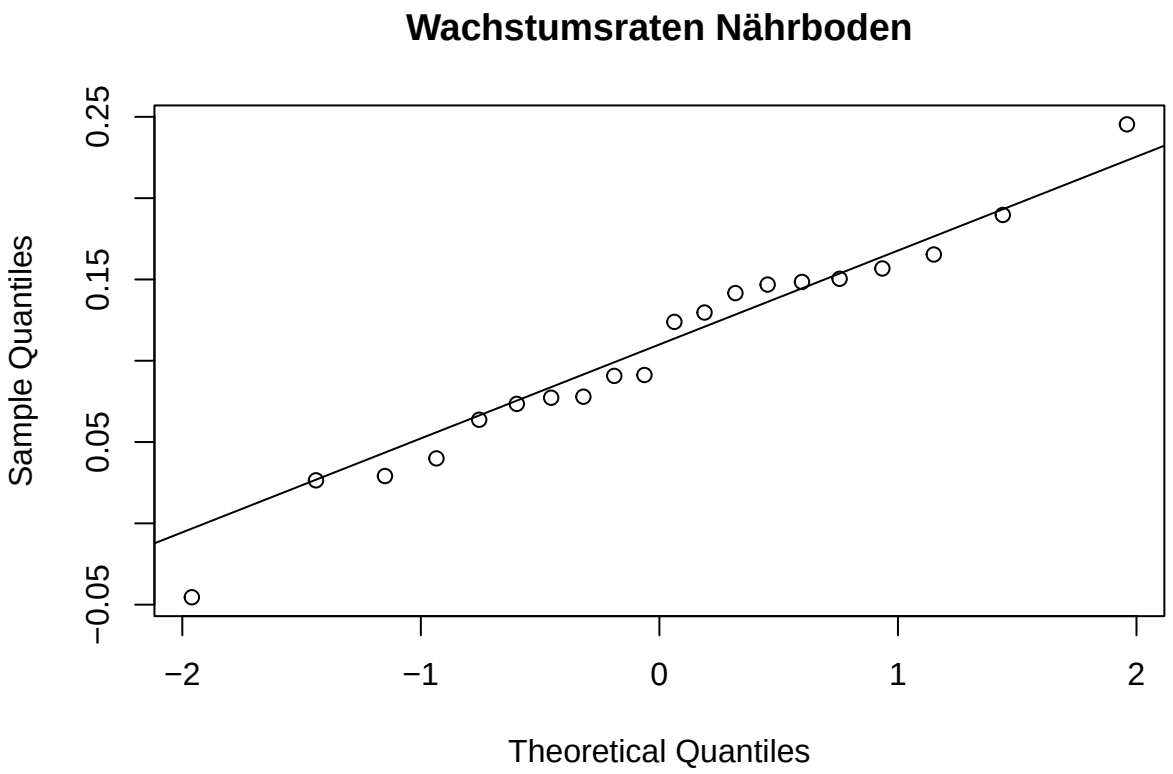
```
hist(nährboden_rate, main="Wachstumsraten Nährboden")
lines(density(nährboden_rate))
```



```r
boxplot(nährboden_rate, main="Wachstumsraten Nährboden", horizontal = T)
```




```r
qqnorm(nährboden_rate, main="Wachstumsraten Nährboden")
qqline(nährboden_rate)
```




```
shapiro.test(nährboden_rate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  nährboden_rate
## W = 0.97869, p-value = 0.916
```

Die Daten scheinen Normalverteilt zu sein, der Minimalwert könnte laut dem QQ-Plot ein Ausreißer sein. Der Shapiro-Wilk normality test zeigt einen p-Wert 0.916, die H0 (die Daten sind Normalverteilt) kann daher nicht verworfen werden. Daher können parametrische Tests angewendet werden. Es wird ein einseitiger, rechtsseitiger Test verwendet.

Die Nullhypothese H0 lautet, die Wachstumsrate ist mit Nährboden gleich/geringer als die Normale mit 1%:

$$H_0 : \mu_{Nährboden} \leq \mu_{Normal}$$

Die Alternativhypothese H1 lautet, die Wachstumsrate ist höher als die Normale von 1%.

$$H_1 : \mu_{Nährboden} > \mu_{Normal}$$

Es wird mit einem Signifikanz-Niveau von 0.05 gearbeitet.

```
t.test(nährboden_rate, mu=normal_rate, conf.level = 0.95, alternative = "greater") #greater=mein mu ist
```

```
##
##  One Sample t-test
##
## data:  nährboden_rate
## t = 6.4434, df = 19, p-value = 1.774e-06
## alternative hypothesis: true mean is greater than 0.01
## 95 percent confidence interval:
##  0.080326      Inf
## sample estimates:
## mean of x
## 0.1061209
```

Der Nährboden weist eine mittlere Wachstumsrate von 10,6% auf, deutlich über der Vergleichsrate. Die Datenexploration lässt den Schluss zu, dass die Daten Normalverteilt (und damit auch unimodal) sind, bestätigt durch einen Shapiro-Wilk normality test. Mittelwert und Median liegen eng beieinander, der QQ-Plot lässt keine schweren/leichten Ränder erkennen. Es wird daher der Mittelwert der normalverteilten Daten mit dem Vergleichswert der Wachstumsrate von 1% verglichen. Dazu kann man den parametrischen Einstichproben-t-Test anwenden, der eine Normalverteilung voraussetzt.

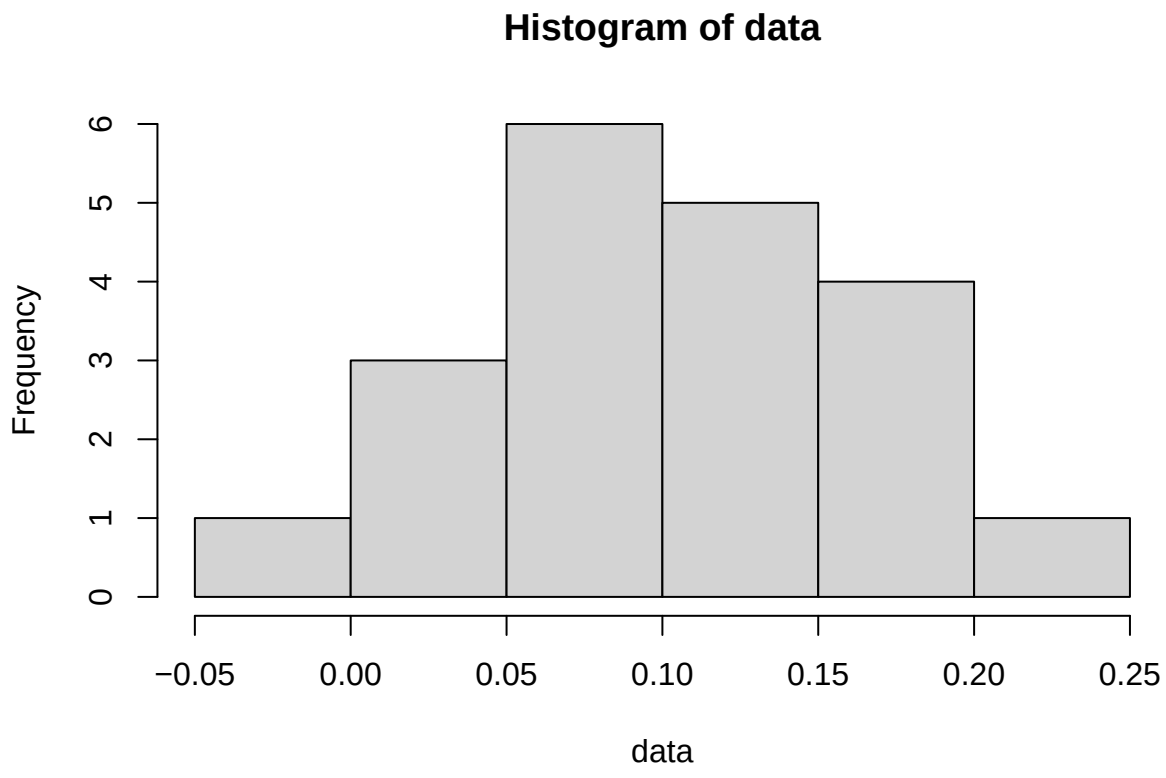
Mit einem p-Wert von 1,744e-6 ist die Wahrscheinlichkeit, die H0 fälschlich zu verwerfen äußerst gering. Sie kann auf den Signifikanz-Niveaus von 5%, 1% und 0.1% verworfen werden. Das Ergebnis ist somit als stark signifikant zu werten. Der verwendete Test weist eine t-Verteilung auf. Die Teststatistik ist 6,443 mit 19 Freiheitsgraden.

Ergänzung SS24

Bayesfactor Im Zuge der Überarbeitung haben wir uns dazu entschieden, einen Hypothesentest mit dem Bayesfaktor durchzuführen. Dazu haben wir das Paket BayesFactor benutzt. Dabei wird ausgehend von der Verteilung der Baysesfaktor aus der psoterioriverteilung gebildet. Wir testen mit h0 = 0.01.

1] Alt., r=0.707 : 5641.645 ±0% spricht dafür, dass der wahre Mittelwert mit sehr hoher WSL nicht bei 0.01 liegt. Dadruch decken sich hier unsere Ergebnise mit der Ausarebitung aus dem WS, bei der wir aufgrund des student t-Tests davon ausgegangen sind, dass der Wahre mittelwert ungleich bzw größer als 0.01 ist.

```
library(BayesFactor)
library(ggplot2)
data <- c(0.146842, 0.156757, 0.091255, 0.063720, 0.148471,
          -0.045436, 0.150407, 0.077905, 0.077267, 0.026454,
          0.090700, 0.245384, 0.129650, 0.141617, 0.039957,
          0.165351, 0.029091, 0.073473, 0.189657, 0.123897)
hist(data)
```




```
priorBF <- ttestBF(x = data, mu = 0.01)
summary(priorBF)
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 5641.645 ±0%
##
## Against denominator:
```

```
##   Null, mu = 0.01
## ---
## Bayes factor type: BFoneSample, JZS
```

```
## Sample from the corresponding posterior distribution
posterior_data <- ttestBF(x = data, posterior = TRUE, iterations = 1000 )
summary(posterior_data)
```

```
##
## Iterations = 1:1000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##            Mean        SD  Naive SE Time-series SE
## mu     0.101326 1.600e-02 5.060e-04      5.060e-04
## sig2   0.005199 1.815e-03 5.740e-05      6.217e-05
## delta  1.467825 3.409e-01 1.078e-02      1.163e-02
## g     22.553780 3.487e+02 1.103e+01      1.103e+01
##
## 2. Quantiles for each variable:
##
##           2.5%      25%      50%     75%     97.5%
## mu    0.071589 0.090868 0.100926 0.11224  0.132369
## sig2  0.002719 0.003918 0.004849 0.00609  0.009753
## delta 0.850682 1.236406 1.461990 1.69586  2.142738
## g     0.344009 0.940406 1.929901 4.43383 45.148428
```

```
#View(posterior_data)
posterior_data = as.data.frame(posterior_data)

bf_pos = ttestBF(x = posterior_data$mu, mu= 0.01)
bf_pos
```

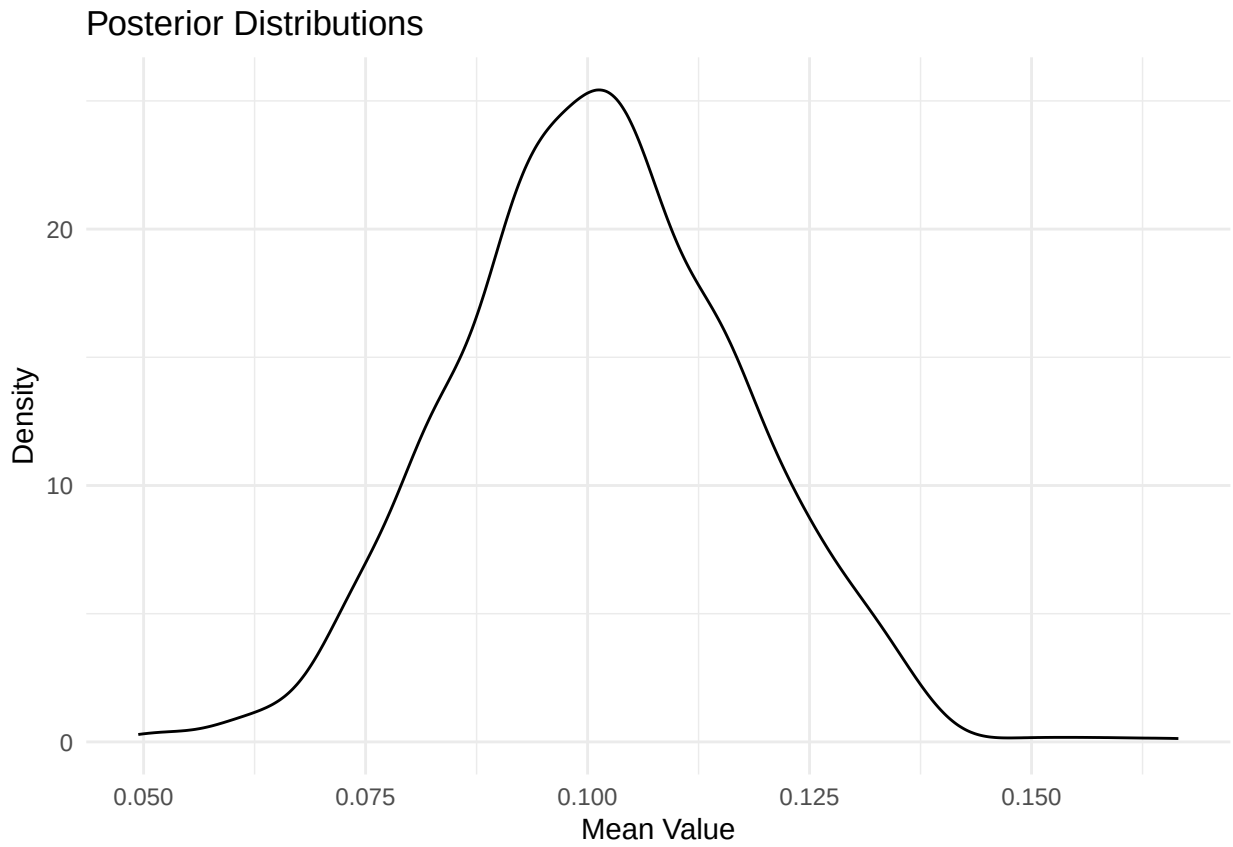
```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 8.343131e+759 ±0%
##
## Against denominator:
##   Null, mu = 0.01
## ---
## Bayes factor type: BFoneSample, JZS
```

```
summary(bf_pos)
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 8.343131e+759 ±0%
##
```

```
## Against denominator:
##   Null, mu = 0.01
## ---
## Bayes factor type: BFoneSample, JZS
```

```
ggplot(posterior_data, aes(x = mu)) +
  geom_density(alpha = 0.5) +
  labs(title = 'Posterior Distributions',
       x = 'Mean Value',
       y = 'Density') +
  theme_minimal()
```



Aufgabe 4

Eine Genontologieanalyse wird durchgeführt, um den Anteil von Genen aus bestimmten Pfades (pathway) zu bestimmen, die an der Entwicklung von Krebs beteiligt sind. Um die Frage zu beantworten, werden 720 mögliche Gene in Betracht gezogen, von denen 696 in mehr als einer Studie gefunden wurden und daher glaubwürdig sind. Von diesen haben 413 mit der Krebsentwicklung zu tun. Berechnen Sie eine Schätzung und das zugehörige 95% bzw. 99% Konfidenzintervall für dieses Szenario. Testen Sie, ob der Anteil der beteiligten Genen sich signifikant gegenüber einer früheren Studie verändert hat, die 55% der Gene als beteiligt gefunden hat.

Dies ist ein Proportionentest, da gefragt wird ob ein Gen beteiligt/nicht beteiligt ist. Die Fragestellung lautet, ob sich der Anteil gegenüber einer vorigen Studie mit 55% VERÄNDERT hat. Wir führen daher einen zweiseitigen Proportionentest durch, der eine Binomialverteilung als Grundlage hat.

Die Nullhypothese H0 lautet:

$$H_0 : p = p0$$

Die Alternativhypothese H1 lautet:

$$H_1 : p \neq p0$$

Die zu testenden Signifikanz-Niveaus sind 0.05 und 0.01.

```
prop.test(413, 696, p = 0.55, alternative = c("two.sided"), conf.level = 0.95, correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  413 out of 696, null probability 0.55
## X-squared = 5.1207, df = 1, p-value = 0.02364
## alternative hypothesis: true p is not equal to 0.55
## 95 percent confidence interval:
##  0.5557580 0.6299782
## sample estimates:
##         p
## 0.5933908
```

```
prop.test(413, 696, p = 0.55, alternative = c("two.sided"), conf.level = 0.99, correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  413 out of 696, null probability 0.55
## X-squared = 5.1207, df = 1, p-value = 0.02364
## alternative hypothesis: true p is not equal to 0.55
## 99 percent confidence interval:
##  0.5440439 0.6409477
## sample estimates:
##         p
## 0.5933908
```

Die Teststatistik X-squared ist 5.1207 mit einem Freiheitsgrad (df=1). Der p-Wert mit 0.02364 reicht nur aus, das erste Signifikanz-Niveau von 5% zu verwerfen, jedoch nicht jenes mit 1%, die Daten sind daher nur schwach signifikant. Es sind 59.33% der Gene an der Krebsentwicklung beteiligt. Dies ist innerhalb des 95% Konfidenzintervalls welches von 55.57% bis 63.00% reicht, aber außerhalb jenes des 99% KI (54.40% bis 64.10%).

Aufgabe 5

Bevor Sie einen Job annehmen, möchten Sie als Kandidat oder Kandidatin die Gehälter in den Firmen vergleichen, die beide bereit wären, Sie anzustellen. Folgende Gehälter können Sie aufgrund von online Transparenzvorgaben in Erfahrung bringen.

Erste Firma: 4218.874 2323.970 4104.761 3172.519 3058.287 2386.729 4405.709 2665.709 5326.124 2993.015 5152.121 3164.876 2703.269 3837.005 2927.137 2847.995 3087.938 3063.339 4697.341 5602.379 2992.996 5052.060 4095.423 1668.059 6268.097

Zweite Firma: 1888.252 2429.395 2062.037 1932.138 1788.335 2119.263 2185.819 2173.098 2391.626 1576.546 1871.540 2405.640 2470.771 1879.237 2181.048 2272.962 2174.767 1729.053 1119.993 2325.788 2112.610 2847.006 1124.272 5320.000 4785.000

Welche der Firmen bietet Ihnen das attraktivere Gehalt?

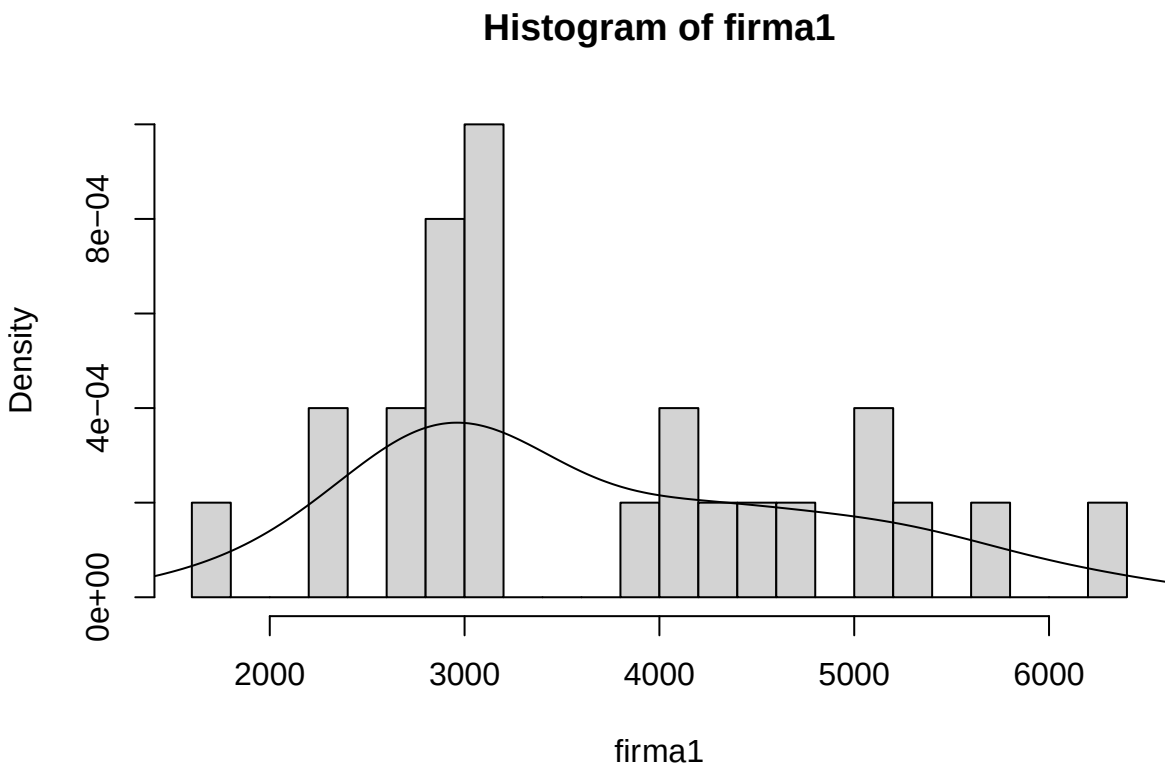
```
library("moments")
firma1 <-c(4218.874, 2323.970,
4104.761, 3172.519, 3058.287, 2386.729 ,4405.709, 2665.709 ,5326.124, 2993.015,
5152.121, 3164.876 ,2703.269 ,3837.005 ,2927.137 ,2847.995 ,3087.938, 3063.339,
4697.341 ,5602.379 ,2992.996 ,5052.060, 4095.423, 1668.059, 6268.097)

firma2 <- c(1888.252, 2429.395, 2062.037, 1932.138, 1788.335 ,2119.263,
2185.819, 2173.098 ,2391.626, 1576.546, 1871.540, 2405.640, 2470.771, 1879.237,
2181.048 ,2272.962 ,2174.767, 1729.053 ,1119.993 ,2325.788 ,2112.610, 2847.006,
1124.272 ,5320.000, 4785.000)

print(summary(firma1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1668    2927    3165    3673    4406    6268
```

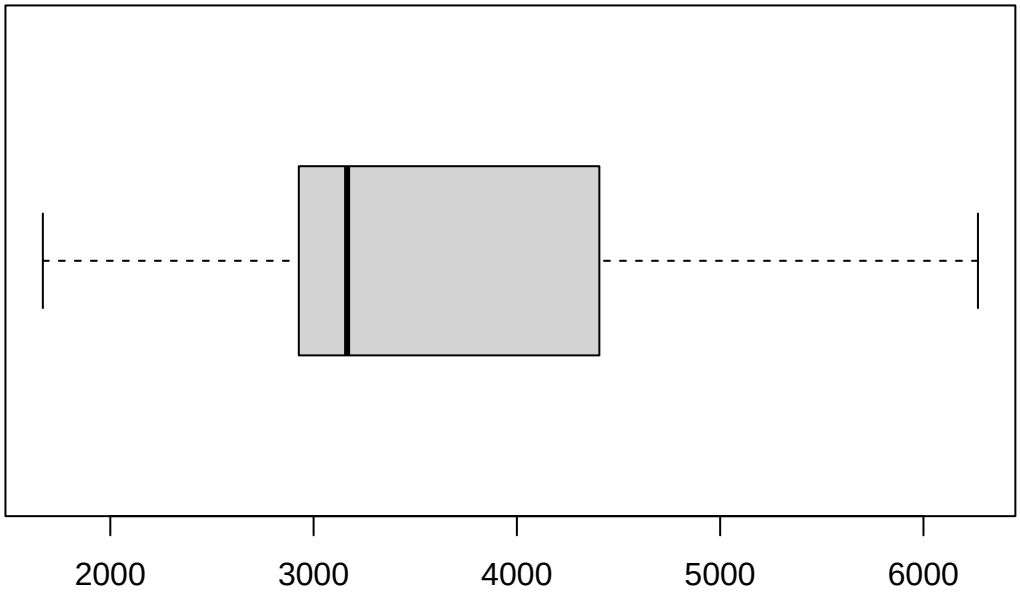
```
print(summary(firma2))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1120    1879    2173    2287    2392    5320
```

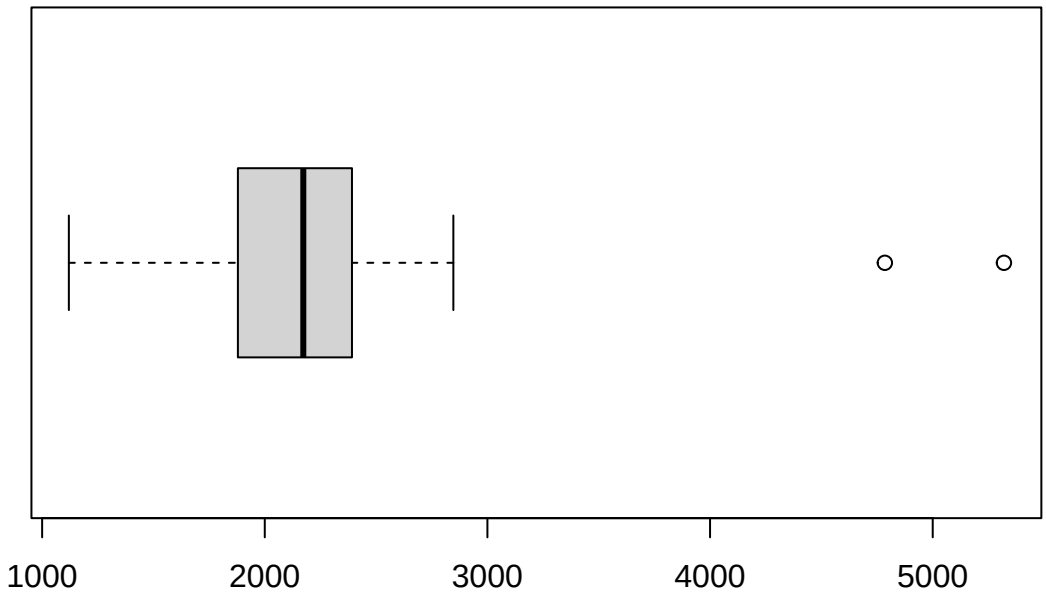
```
hist(firma1, breaks = 20, prob = TRUE)
lines(density(firma1))
```



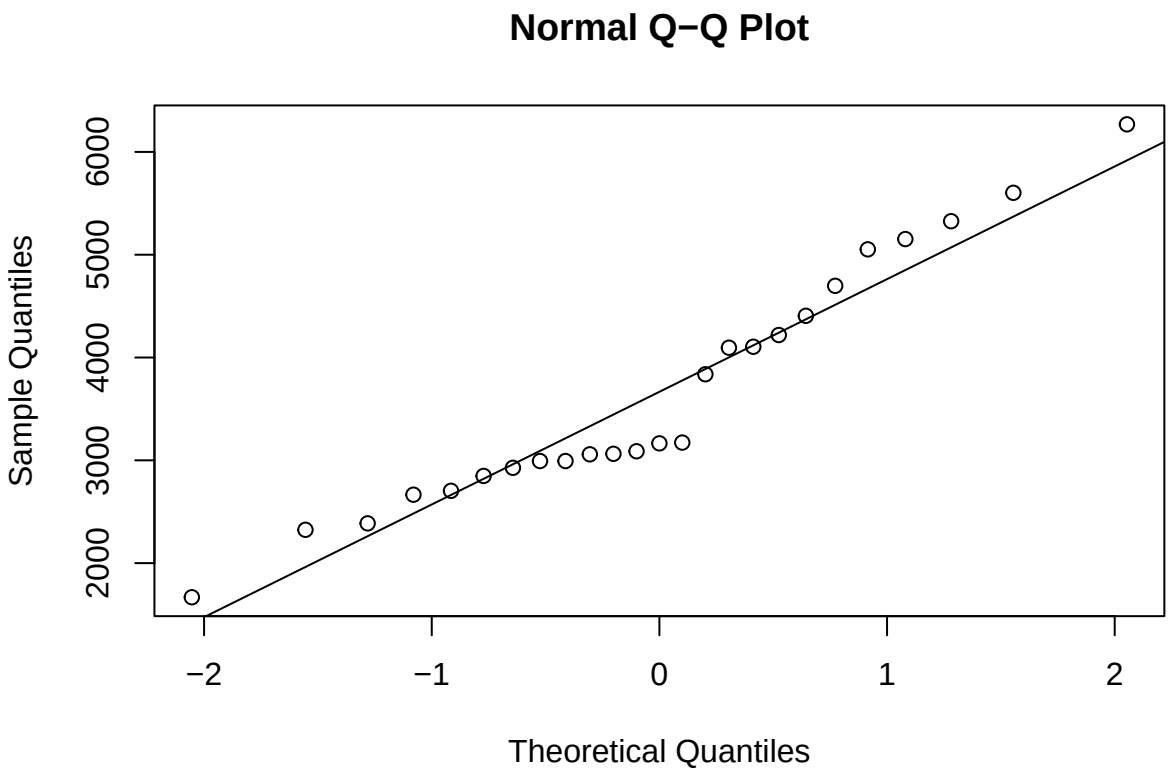

```
boxplot(firma1, horizontal = TRUE, title = "Firma1")
```



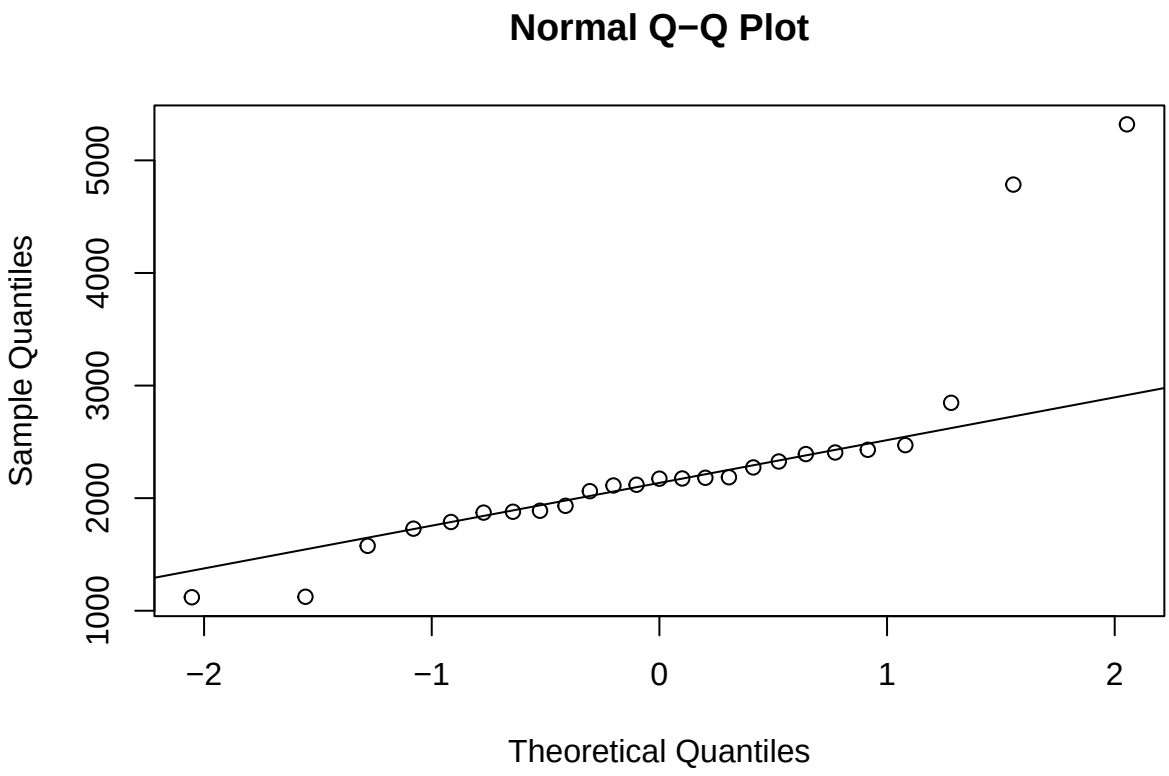

```
boxplot(firma2, horizontal = TRUE)
```



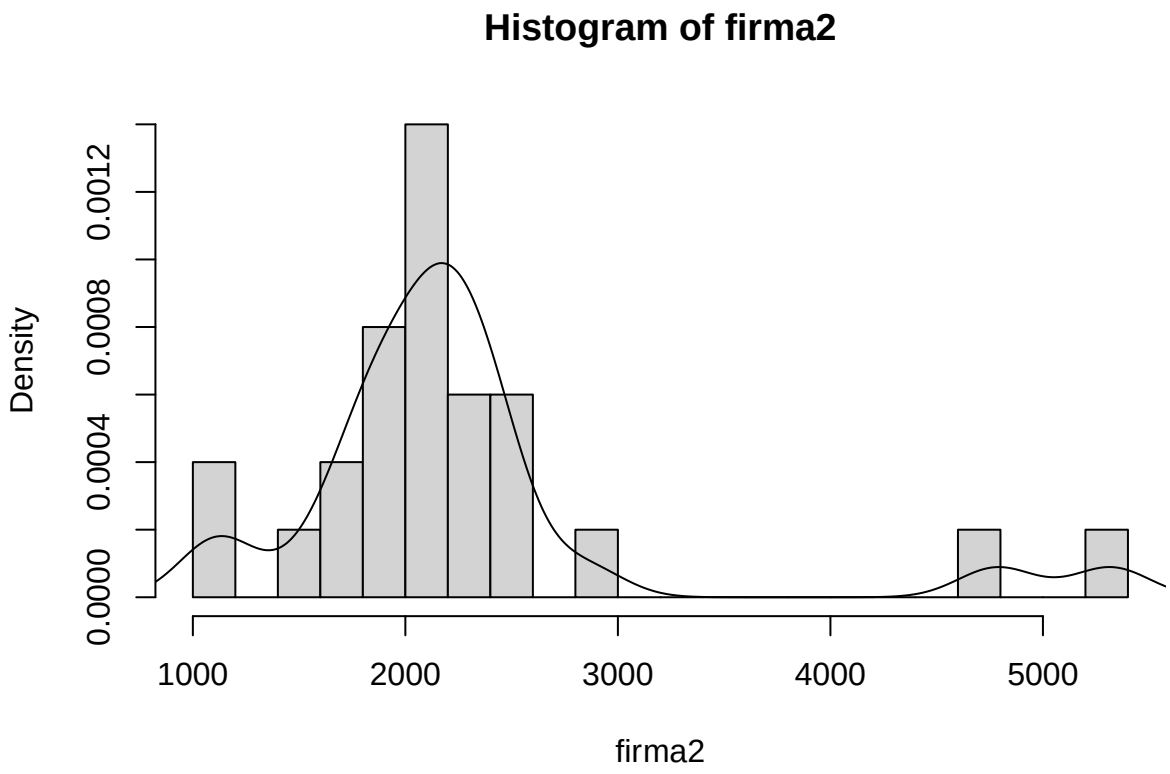

```
qqnorm(firma1)
qqline(firma1)
```




```
qqnorm(firma2)
qqline(firma2)
```




```
hist(firma2, breaks = 20, prob = TRUE)
lines(density(firma2))
```




```
#Normalverteilung? Für eine Normalverteilung müsste der Kurtosis-Werte nahe 3 leigen
print(kurtosis(firma1))
```

```
## [1] 2.386853
```

```
print(kurtosis(firma2))
```

```
## [1] 7.690339
```

```
#Standardabweichung
sd(firma1)
```

```
## [1] 1168.453
```

```
sd(firma2)
```

```
## [1] 921.7741
```

Ausgehend von der Summary, wo der Mittelwert von Firma 1 bei 3673 und von Firma 2 bei 2287 liegt, würde man man davon ausgehen, dass bei Firma 1 das Gehalt lukrativer ist.

Die Nullhypothese H0 lautet (Gehälter haben gleiche Tendenz):

$$H_0 : p = p0$$

Die Alternativhypothese H1 lautet (Gehälter sind unterschiedlich):

$$H_1 : p \neq p0$$

Wir überprüfen die Testbediongungen. Es handelt sich auf keinen Fall um eine NV, daher prüfen wir, ob die Daten Unimodal sind Ausreißer beinhalten annähernd die gleiche Verteilung haben

Firma 2 hat am oberen Ende 2 Werte, die aufgrund des Boxplots eher als ausreißer erkennabr sind. Am unteren Ende scheint es einen 2ten Modus zu geben, weshalb die Testbedinung für einen Mann_whitney U test (unimodal) nciht mehr gegeben ist. Wir verwerfen also unsere Annnahme aus dem WS, hier wilcox.test durchzuführen.

SS24 Stattdessen haben wir uns dazu entschieden, eine Bootsreapping Simulation mit Resampling durchzuführen. Dazu haben wir wieder n=1000 genommen und die Erwartete Differenz des Mittelwerts berechnet.

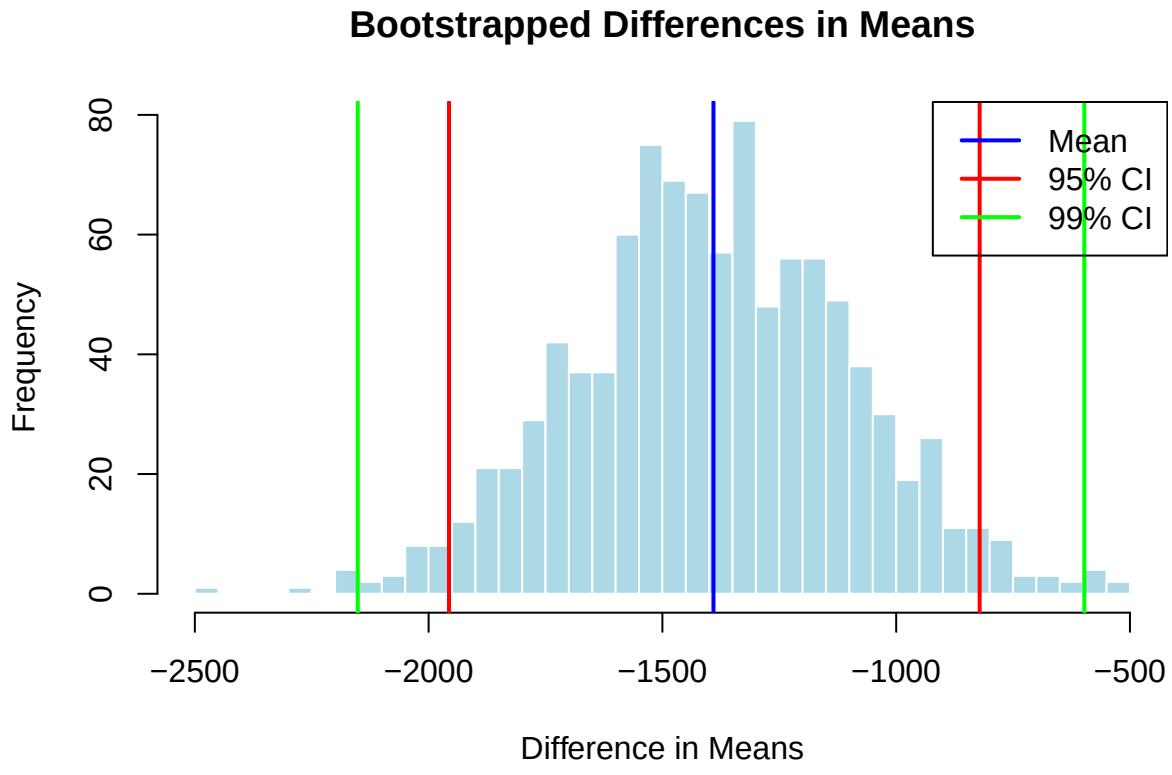
```
firma1 <-c(4218.874, 2323.970,
4104.761, 3172.519, 3058.287, 2386.729 ,4405.709, 2665.709 ,5326.124, 2993.015,
5152.121, 3164.876 ,2703.269 ,3837.005 ,2927.137 ,2847.995 ,3087.938, 3063.339,
4697.341 ,5602.379 ,2992.996 ,5052.060, 4095.423, 1668.059, 6268.097)

firma2 <- c(1888.252, 2429.395, 2062.037, 1932.138, 1788.335 ,2119.263,
2185.819, 2173.098 ,2391.626, 1576.546, 1871.540, 2405.640, 2470.771, 1879.237,
2181.048 ,2272.962 ,2174.767, 1729.053 ,1119.993 ,2325.788 ,2112.610, 2847.006,
1124.272 ,5320.000, 4785.000)

#Funktion zur Berechnung der Differenz der Mittelwerte
mean_diff <- function(data1, data2) {
  return(mean(data1) - mean(data2))
}

#Bootstrapping durchführen
set.seed(123)
n_boot <- 1000
boot_diffs <- numeric(n_boot)

for (i in 1:n_boot) {
  boot_firma1 <- sample(firma1, length(firma1), replace = TRUE)
  boot_firma2 <- sample(firma2, length(firma2), replace = TRUE)
  boot_diffs[i] <- mean_diff(boot_firma2, boot_firma1)
}




#Bootstrapped Konfidenzintervalle berechnen

boot_ci <- quantile(boot_diffs, c(0.025, 0.975))
boot_ci2 <- quantile(boot_diffs, c(0.005, 0.995))

boot_ci
```

```
##       2.5%      97.5%
## -1956.3743  -821.4704
```

```
boot_ci2
```

```
##       0.5%      99.5%
## -2151.5870  -597.5857
```

```
hist(boot_diffs, breaks = 30, main = "Bootstrapped Differences in Means",
     xlab = "Difference in Means", col = "lightblue", border = "white")
abline(v = boot_ci[1], col = "red", lwd = 2)
abline(v = boot_ci[2], col = "red", lwd = 2)

abline(v = boot_ci2[1], col = "green", lwd = 2)
abline(v = boot_ci2[2], col = "green", lwd = 2)


abline(v = mean(boot_diffs), col = "blue", lwd = 2)
legend("topright", legend = c("Mean", "95% CI", "99% CI"),
       col = c("blue", "red", "green"), lwd = 2)
```



Conclusion und Interpretation

Ausgehend von dem 5% und 1% CI sieht man, dass 0 nicht in der Simulation enthalten ist, die Werte aus der DifferenzenBildung Firma2-Firma1 negativ sind. Daher kann man davon ausgehen, dass Firma 1 das lukrativere Gehalt bietet.