Hausübung 03 - Statistische Tests

Baier Sebastian, Magdalena Figlmüller, Alice Schwarzböck

2023-11-29

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.

-) 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?

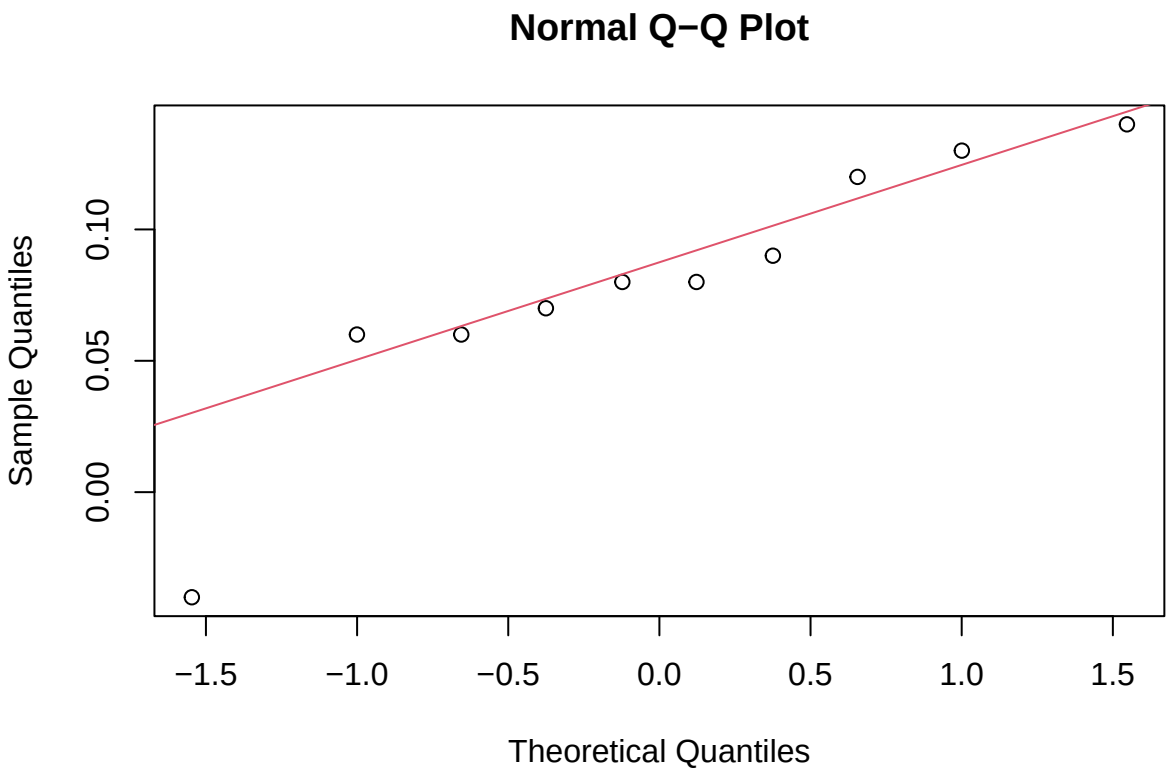
-) Beantworten Sie diese Fragen auf dem 5% und 1% Niveau.

```
before <- c(1.36, 1.37, 1.29, 1.22, 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)

# Test auf Normalverteilung der Daten (Differenzen)
diff <- before - after
shapiro.test(diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diff
## W = 0.86853, p-value = 0.09609
```

```
qqnorm(diff)
qqline(diff,col=2)
```



Die Überprüfung der beiden Datensätze mittels Q-Q-Plot und Shapiro-Wilk Test liefert keine ausreichenden Hinweise, welche ein Verwerfen der Nullhypothese rechtfertigen würde. Ein Ausreißer ist im Q-Q-Plot zu erkennen, sein Einfluss auf die Verteilung liegt laut Shapiro-Test allerdings im tolerierbaren Bereich. Die Daten sind somit annähernd normalverteilt und können mittels t-Test analysiert werden.

Nullhypothese: Das Training hatte keinen Effekt.

$$H0 : \mu_{before} = \mu_{after}$$

Alternativhypothese: Das Training hatte einen Effekt.

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

```
# T-test Variante1 (mit 2 gepaarten Stichproben)
t.test(before, after, paired = T, conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 4.9322, df = 9, p-value = 0.0008109
## alternative hypothesis: true mean difference is not equal to 0
## 99 percent confidence interval:
##  0.02694624 0.13105376
## sample estimates:
```

```
## mean difference
##           0.079
```

```
# T-test Variante2 (mit Differenzen)
t.test(diff, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  diff
## t = 4.9322, df = 9, p-value = 0.0008109
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  0.02694624 0.13105376
## sample estimates:
## mean of x
##     0.079
```

Die Teststatisik beträgt t = 4.9322. Der p-Wert liegt bei 0.0008109, somit kann die Nullhypothese sowohl auf dem 5% als auch auf dem 1% Niveau verworfen werden.

–> Das Training hatte einen Effekt!

Nullhypothese: Die Arbeitszeit war vor dem Training kürzer/gleich.

$$H0 : \mu_{before} \leq \mu_{after}$$

Alternativhypothese: Das Training hat die Arbeitszeit verringert.

$$HA : \mu_{before} > \mu_{after}$$

```
# T-test Variante1 (mit 2 gepaarten Stichproben)
t.test(x = before,y=after,mu = 0,conf.level = 0.99,paired=TRUE,
alternative = c("greater"))
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 4.9322, df = 9, p-value = 0.0004055
## alternative hypothesis: true mean difference is greater than 0
## 99 percent confidence interval:
##  0.03380804        Inf
## sample estimates:
## mean difference
##           0.079
```

```
# T-test Variante1 (mit Differenzen)
t.test(diff,mu = 0,conf.level = 0.99,
alternative = c("greater"))
```

```
##
##  One Sample t-test
##
## data:  diff
## t = 4.9322, df = 9, p-value = 0.0004055
## alternative hypothesis: true mean is greater than 0
## 99 percent confidence interval:
##  0.03380804        Inf
## sample estimates:
## mean of x
##     0.079
```

Die Teststatisik beträgt t = 4.9322. Der p-Wert liegt bei 0.0004055, somit kann die Nullhypothese sowohl auf dem 5% als auch auf dem 1% Niveau verworfen werden.

–> Das Training hat die mittlere Arbeitszeit verringert.

Aufgabe 2

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

```
str(Titanic)
```

```
##  'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

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

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

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

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

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

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

Es sollen jeweils 2 Proportionen miteinander verglichen werden (Männer - Frauen / Männer - Kinder), daher wird der Proportionentest angewandt. Auf Basis des Grenzverteilungssatzes wird eine Normalverteilung der Daten angenommen.

1) Vergleich FRAUEN : MÄNNER

Nullhypothese: Frauen hatten keine besseren Überlebenschancen als Männer.

$$p(Frauen) \leq p(Männer)$$

Alternativhypothese: Frauen hatten bessere Überlebenschancen als Männer.

$$p(Frauen) > p(Männer)$$

```
prop.test(c(women["Yes"],men["Yes"]),c(sum(women), sum(men)), alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(women["Yes"], men["Yes"]) out of c(sum(women), sum(men))
## X-squared = 454.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.4812549 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7319149 0.2120162
```

Die Teststatisik beträgt X-squared = 454.5. Aufgrund des p-Wertes mit < 2.2e-16, kann die Nullhypothese auf dem 1% Niveau verworfen werden.

–> Frauen hatten bessere Überlebenschancen als Männer.

2) Vergleich KINDER : MÄNNER

Nullhypothese: Kinder hatten keine besseren Überlebenschancen als Männer.

$$H0 : p(Kinder) \leq p(Männer)$$

Alternativhypothese: Kinder hatten bessere Überlebenschancen als Männer.

$$HA : p(Kinder) > p(Männer)$$

```
prop.test(c(children["Yes"],men["Yes"]),c(sum(children), sum(men)), alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(children["Yes"], men["Yes"]) out of c(sum(children), sum(men))
## X-squared = 54.16, df = 1, p-value = 9.24e-14
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2257103 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.5229358 0.2120162
```

Die Teststatisik beträgt X-squared = 54.16. Aufgrund des p-Wertes mit < 9.24e-14, kann die Nullhypothese auf dem 1% Niveau verworfen werden.

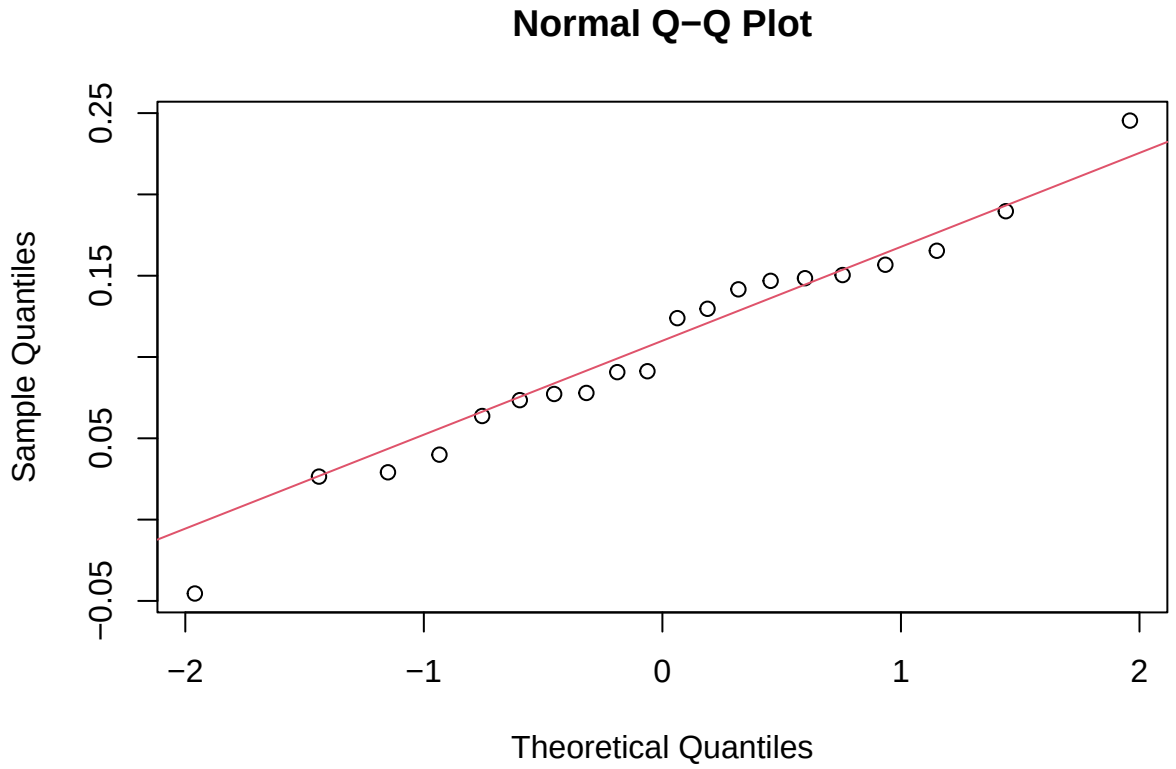
–> Kinder hatten bessere Überlebenschancen als Männer.

Aufgabe 3

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

```
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)

qqnorm(data)
qqline(data,col=2)
```




```
shapiro.test(data)
```

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

Weder der QQ-Plot noch der Shapiro-Test sprechen gegen die Nullhypothese, daher wird eine Normalverteilung der Daten angenommen. Es soll die mittlere Wachstumsrate gegen den Referenzwert von 0.01 verglichen werden, daher wird ein 1-Sample t-test mit μ = 0.01 angewendet.

Nullhypothese: Das Nährmedium bietet keinen zusätzlichen Wachstumsvorteil.

$$H0 : \mu_{normal} \geq \mu_{test}$$

Alternativhypothese: Das Nährmedium bietet einen zusätzlichen Wachstumsvorteil.

$$HA : \mu_{normal} < \mu_{test}$$

```
t.test(data, mu = 0.01,
alternative = c("greater"))
```

```
##
##  One Sample t-test
##
## data:  data
## 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
```

Die Teststatisik beträgt t = 6.4434. Der p-Wert liegt bei 1.774e-06, somit kann die Nullhypothese verworfen werden.

–> Das Nährmedium bietet einen Wachstumsfaktor.

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 einer Studie gefunden wurden und daher glaubwürdig sind. Von diesen haben 413 mit der Krebsentwicklung zu tun.

-) Testen Sie, ob der Anteil der beteiligten Genen sich signifikant gegenüber einer früheren Studie verändert hat, die 55% der Gene als beteiligt gefunden hat.Berechnen Sie das zugehörige 95% bzw. 99% Konidenzintervall für dieses Szenario.

Der Anteil an beteiligten Genen beträgt laut der aktuellen Studie rund 59% (=413/696). Es gilt über den 2-seitigen Proportionentest zu eruieren, ob die Abweichung zu einer früheren Studie (55%) als signifikant anzusehen ist.

Nullhypothese: Die aktuelle Studie unterschiedet sich nicht von der früheren Studie.

$$H0 : p(Gene) = 0.55$$

Alternativhypothese: Die aktuelle Studie unterschiedet sich von der früheren Studie.

$$HA : p(Gene) \neq 0.55$$

```r
prop.test(413, 696, p = 0.55, conf.level = 0.95)
```

```
##
##  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
```

```r
prop.test(413, 696, p = 0.55, conf.level = 0.99)
```

```
##
##  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 beträgt X-squared = 5.1207, der p-Wert liegt bei 0.02364. Es handelt sich um einen 2-seitigen Test, daher kann die Nullhypothese auf dem 5% Niveau verworfen werden, muss auf dem 1% Niveau jedoch beibehalten werden. Dies wird auch anhand der zugehörigen Konfidenzintervalle sichtbar. Das 95%-ige Konfidenzintervall liegt zwischen [0.5557580, 0.6299782], inklusidert das Ergebnis der vorherigen Studie mit 55% somit nicht, während das 99%-ige Konfidenzintervall zwischen [0.5440439 0.6409477] liegt und das Ergebnis der früheren Studie mit 55% einschließt.

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. Diverse Gehälter konnten Sie aufgrund von online Transparenzvorgaben in Erfahrung bringen.Welche der Firmen bietet Ihnen das attraktivere Gehalt?

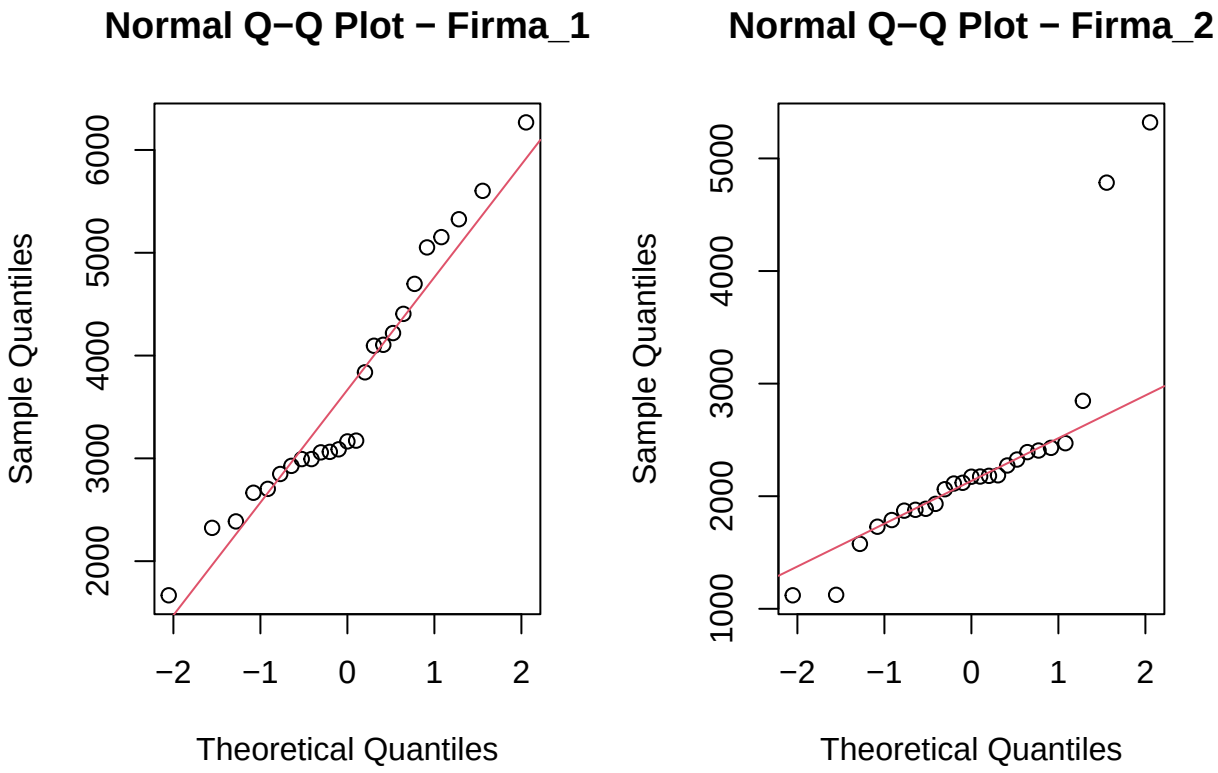
```r
comp1 <- c(4218.874, 2323.970, 4104.761, 3172.519, 3058.287, 2386.729, 4405.709,
           2665.709, 5326.124, 2993.015,5152.121, 3164.876, 2703.269, 3837.005,
           2927.137, 2847.995, 3087.938, 3063.339, 4697.341, 5602.379, 2992.996,
           5052.060, 4095.423, 1668.059, 6268.097)

comp2 <- 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)
```

```r
par(mfrow=c(1,2))

qqnorm(comp1, main = "Normal Q-Q Plot - Firma_1")
qqline(comp1,col=2)

qqnorm(comp2, main = "Normal Q-Q Plot - Firma_2")
qqline(comp2,col=2)
```




```r
shapiro.test(comp1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  comp1
## W = 0.94161, p-value = 0.1612
```

```r
shapiro.test(comp2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  comp2
## W = 0.7182, p-value = 1.293e-05
```

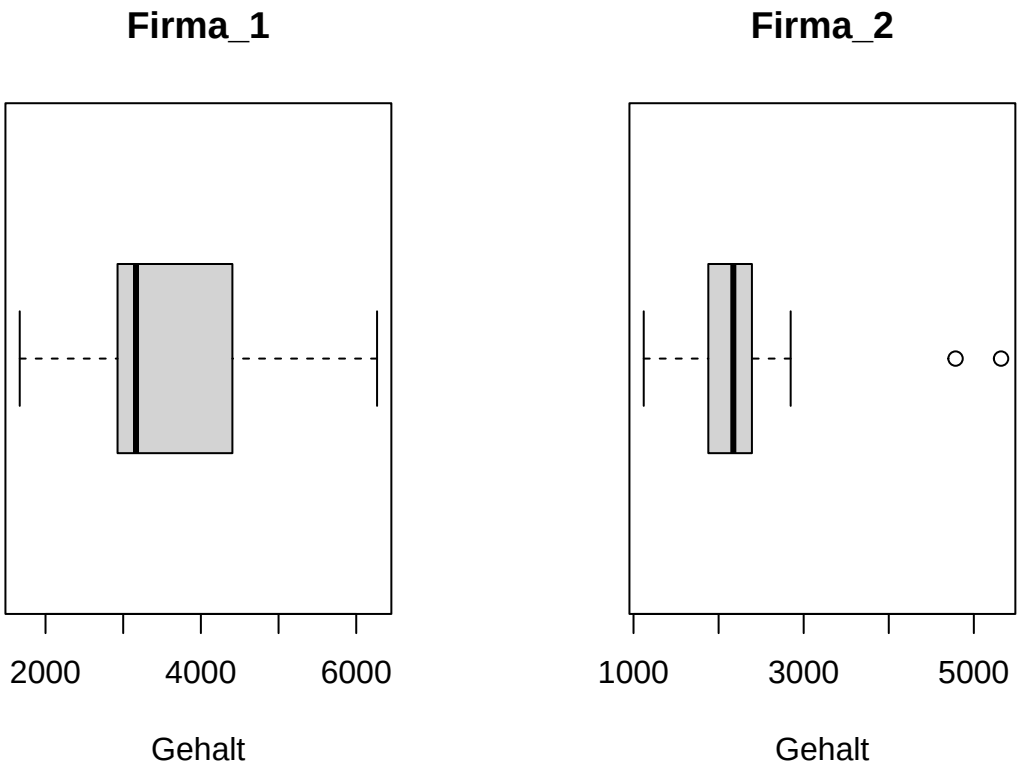
```r
summary(comp1)
```

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

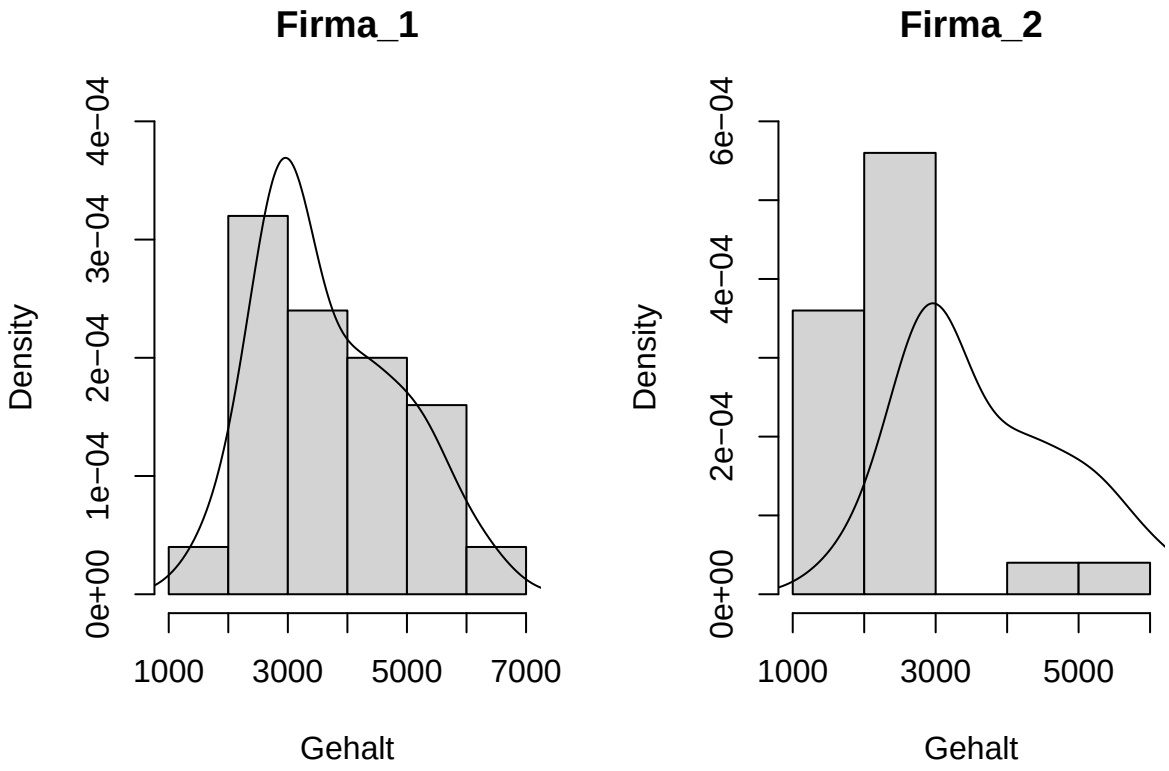
```r
summary(comp2)
```

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

```r
boxplot(comp1, horizontal = T, main = "Firma_1", xlab = "Gehalt")
boxplot(comp2, horizontal = T, main = "Firma_2", xlab = "Gehalt")
```




```r
hist(comp1, freq = F, main = "Firma_1", ylim = c(0,0.0004), xlab = "Gehalt")
lines(density(comp1))
hist(comp2, freq = F, main = "Firma_2", ylim = c(0,0.0006), xlab = "Gehalt")
lines(density(comp1))
```



Sowohl der Q-Q-Plot als auch der Shapiro-Test zeigen, dass die Gehälter von Firma2 im Gegensatz zu jenen der Firma 1 nicht normalverteilt sind. Im Boxplot ist erkennbar, dass die Gehälter der Firma 1 deutlich breiter gestreut sind als jene von Firma 2, welche in einem engen Intervall um 2000 zu finden sind (mit der Ausnahme von 2 deutlichen Ausreißern). Aufgrund dieser Datenverteilung muss für die Analyse auf einen nicht-parametrischen Test zurückgegeriffen werden. Die Wahl fällt hier auf den Wilcoxon Test, welcher lediglich Unimodalität vorsieht. Wie im Histogramm ersichtlich ist diese Voraussetzung erfüllt.

Nullhypothese: Firma_1 zahlt weniger/gleiches Gehalt wie Firma_2

$$H0 : Gehalt(Firma1) \leq Gehalt(Firma2)$$

Alternativhypothese: Firma_1 zahlt mehr Gehalt

$$HA : Gehalt(Firma1) > Gehalt(Firma2)$$

```
wilcox.test(x = comp1, y = comp2, alternative = "greater",
paired = FALSE)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  comp1 and comp2
## W = 550, p-value = 3.849e-07
## alternative hypothesis: true location shift is greater than 0
```

Die Teststatistik ist W=550 und der p-Wert liegt bei 3.849e-07. Somit kann die Nullhypotese sowohl auf dem 5% als auch auf dem 1% Signifikanznivear verworfen werden.

–> Firma 1 zahlt höhere Gehälter als Firma 2.