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.

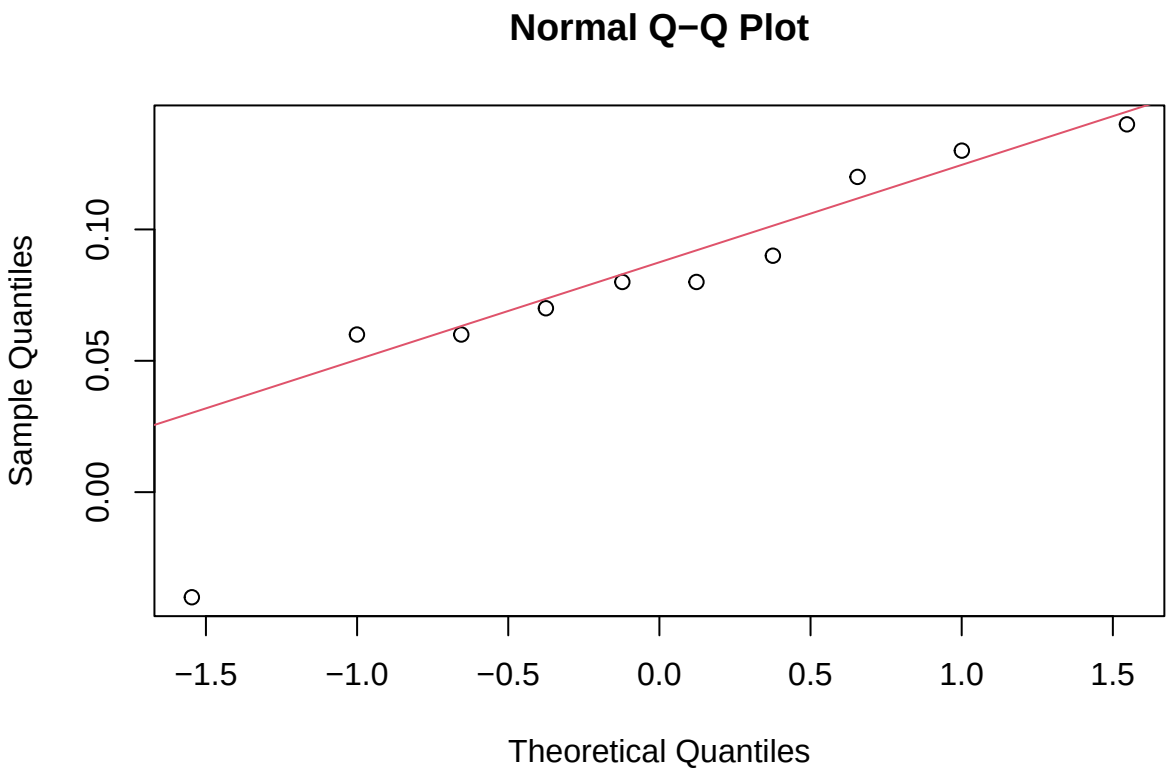
```
before <- c(1.36, 1.37, 1.29, 1.22, 1.38, 1.31, 1.40, 1.39, 1.30, 1.37)
after <- c(1.29, 1.25, 1.20, 1.26, 1.25, 1.23, 1.26, 1.31, 1.24, 1.31)

before_woOut <- c(1.36, 1.37, 1.29, 1.38, 1.31, 1.40, 1.39, 1.30, 1.37)
after_woOut <- c(1.29, 1.25, 1.20, 1.25, 1.23, 1.26, 1.31, 1.24, 1.31)

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

```
##  [1]  0.07  0.12  0.09 -0.04  0.13  0.08  0.14  0.08  0.06  0.06
```

```
shapiro.test(diff)
```

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

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



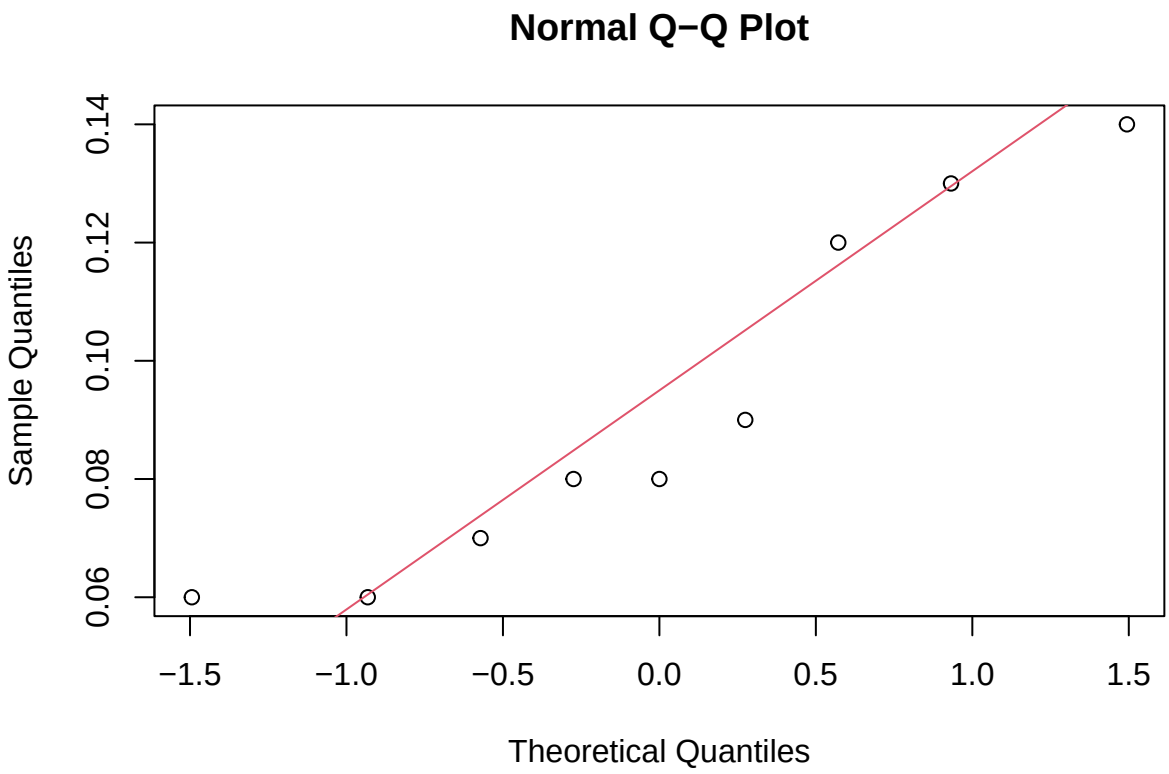

```
# ERWEITERUNG - Entfernen von Outlier -0.04
diff_woOut <- before_woOut - after_woOut
diff_woOut
```

```
## [1] 0.07 0.12 0.09 0.13 0.08 0.14 0.08 0.06 0.06
```

```
shapiro.test(diff_woOut)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diff_woOut
## W = 0.88433, p-value = 0.1743
```

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



Da es sich in diesem Beispiel um die Untersuchung von zwei gepaarten Stichproben handelt, deren Differenz auf Signifikanz überprüft werden soll, sind die Differenzen auf Normalverteilung zu prüfen. Der Q-Q-Plot zeigt hier einen Ausreißer in der linken unteren Ecke, welcher in weiterer Folge entfernt wird. Für beide Datensätze, jenem mit und jenem ohne Auseißer, liefert der Shapiro-Wilk Test keine ausreichenden Hinweise, welche ein Verwerfen der Nullhypothese rechtfertigen würde, allerdings ist das Testergebnis des kleineren Datensatzes deutlich robuster.
Die Daten sind somit annähernd normalverteilt und können mittels t-Test analysiert werden. Es soll im ersten Schritt überprüft werden, ob das Training einen Effekt hatte. Hier kann entweder der gepaarte T-test mit beiden Datensätzen oder der einfache T-Test mit den Differenzen durchgeführt 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!

Um zu überprüfen, ob das Training die mittlere Arbeitszeit verringert hat, wird die Nullhypothese um mu=0 (= tatsächliche Differenz zwischen den Datensätzen gleich 0) und die alternative = "greater" (–> 'before' hat einen höheren Mittelwert als 'after') erweitert und die Wahrscheinlichkeit des Szenarios mittels T-Test errechnet

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
```

```r
# 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
```

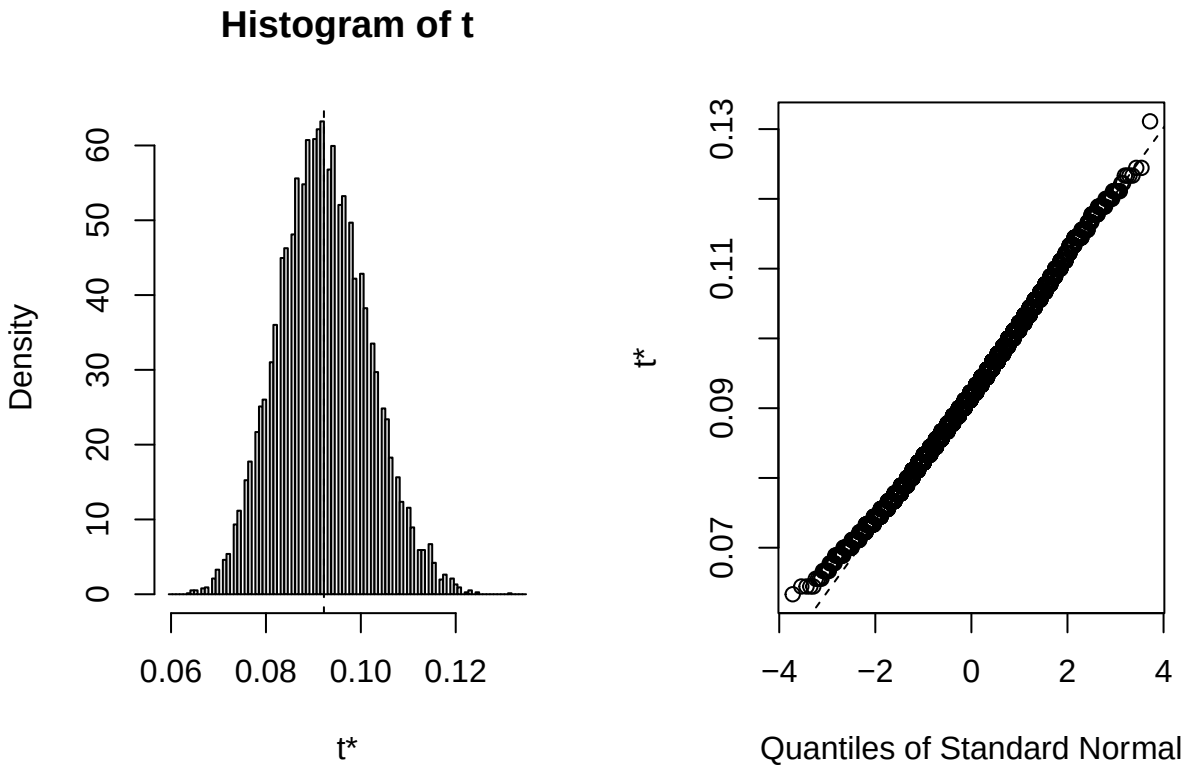
Die Teststatisik beträgt t = 4.9322. Der p-Wert liegt bei 0.0004055, somit kann die Nullhypothese sowohl auf dem 5% als auch auf dem 1% Niveau verworfen werden.

–> Das Training hat die mittlere Arbeitszeit verringert.

ERWEITERUNG - BOOTSTRAPPING

Es werden 10.000 Bootstrap-Stichproben generiert (seed 1234) und die Häufigkeitsverteilung ihrer Mittelwerte sowie das zugehörige Konfidenzintervall für den tatsächlichen Mittelwert auf 95% und 99% Signifikanz bestimmt.

```r
library(boot)

bootstrap_diff_mean <- function(data, indices)
  {
  sample <- data[indices]
  return(mean(sample))
}

set.seed(1234)
num_resample <- 10000

bootstrap_res1 <- boot(diff_woOut, bootstrap_diff_mean, num_resample)
plot(bootstrap_res1)
```




```
(bootstrap_res1_95 <- boot.ci(bootstrap_res1, type = "perc", conf = 0.95))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootstrap_res1, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 0.0744,  0.1122 )
## Calculations and Intervals on Original Scale
```

```
(bootstrap_res1_99 <- boot.ci(bootstrap_res1, type = "perc", conf = 0.99))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootstrap_res1, conf = 0.99, type = "perc")
##
## Intervals :
## Level     Percentile
## 99%   ( 0.0700,  0.1178 )
## Calculations and Intervals on Original Scale
```

Das Histogram der Bootstrap-Strichproben zeigt eine unimodale Verteilung mit einem zentralen Peak und guter Symmetrie und der Q-Q-plot zeigt kaum Abweichungen von der theoretischen Normalverteilung. Wie bei dem t-Test, zeigt auch Bootstrapping einen deutlichen Effekt des Training, sowohl für Konfidenzintervall 95% als auch 99%. Für das Konfidenzintervall von 99% liegt das Intervall der tatsächlichen Differenz (=Trainingseffekt) im Bereich [0.0700, 0.1178].

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.

$$p(Frauen) > p(Männer)$$

```r
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.

$$HA : p(Kinder) > p(Männer)$$

```r
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.

ERWEITERUNG - BAYES STATISTIK

Die Analyse wird mittels bayes.prop.test um eine simulationsbasierte Methode erweitert.

```r
library(BayesianFirstAid)
```

```
## Lade nötiges Paket: rjags
```

```
## Lade nötiges Paket: coda
```

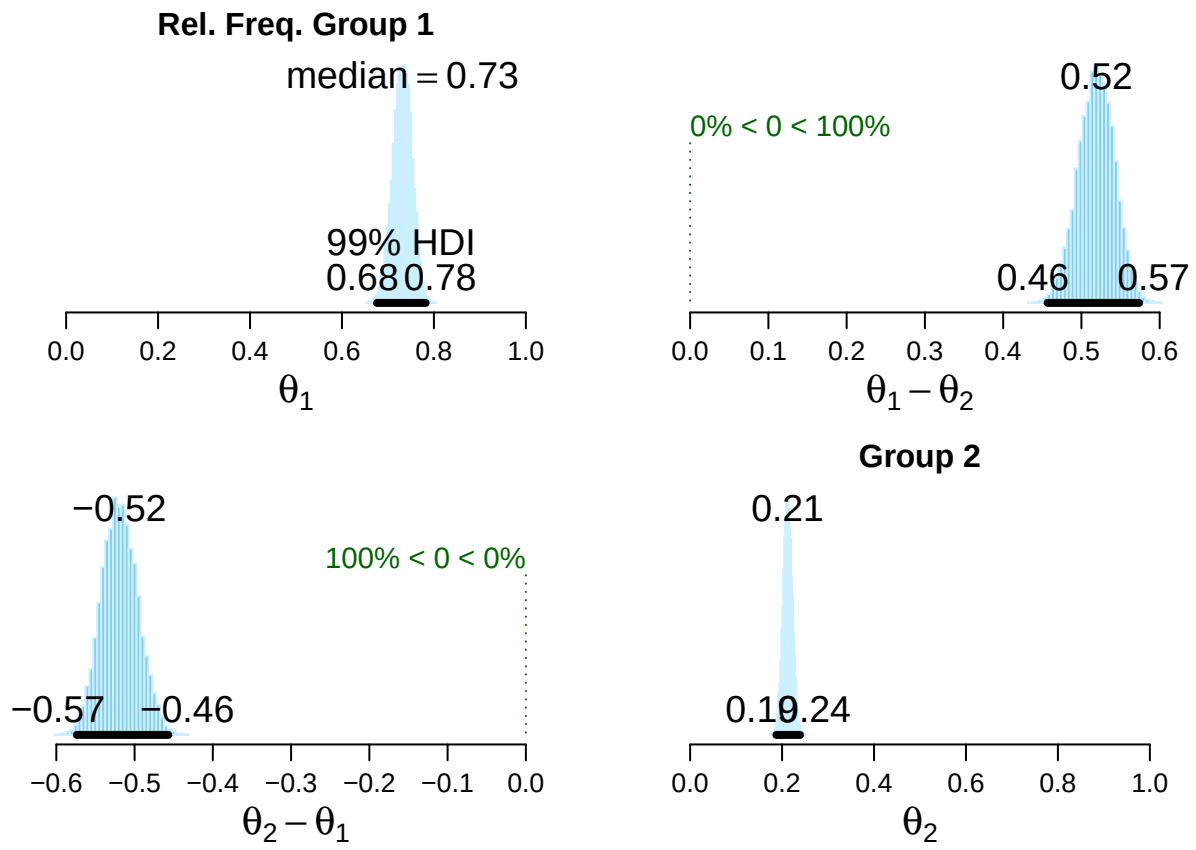
```
## Linked to JAGS 4.3.2
```

```
## Loaded modules: basemod,bugs
```

```r
bayesFit_Titanic_woman <- BayesianFirstAid::bayes.prop.test(
  c(women["Yes"],men["Yes"]),c(sum(women), sum(men)), cred.mass = 0.99)
summary(bayesFit_Titanic_woman)
```

```
##   Data
## number of successes:   344,  367
## number of trials:      470, 1731
##
##   Model parameters and generated quantities
## theta[i]: the relative frequency of success for Group i
## x_pred[i]: predicted number of successes in a replication for Group i
## theta_diff[i,j]: the difference between two groups (theta[i] - theta[j])
##
##   Measures
##                     mean     sd   HDIlo   HDIup %<comp %>comp
## theta[1]           0.731  0.021   0.676   0.782      0      1
## theta[2]           0.212  0.010   0.188   0.239      1      0
## x_pred[1]        343.684 13.761 308.000 377.000      0      1
## x_pred[2]        367.594 24.093 305.000 428.000      0      1
## theta_diff[1,2]    0.519  0.023   0.457   0.574      0      1
##
## 'HDIlo' and 'HDIup' are the limits of a 99% HDI credible interval.
## '%<comp' and '%>comp' are the probabilities of the respective parameter being
## smaller or larger than 0.5 (except for the theta_diff parameters where
## the comparison value comp is 0.0).
##
##   Quantiles
##                    q2.5%    q25%  median    q75%  q97.5%
## theta[1]           0.689   0.717   0.732   0.746   0.770
## theta[2]           0.193   0.206   0.212   0.219   0.232
## x_pred[1]        316.000 335.000 344.000 353.000 370.000
## x_pred[2]        322.000 351.000 368.000 384.000 416.000
## theta_diff[1,2]    0.472   0.503   0.519   0.535   0.563
```

```r
plot(bayesFit_Titanic_woman)
```




```
df_Titanic_woman <- as.data.frame(bayesFit_Titanic_woman)
```

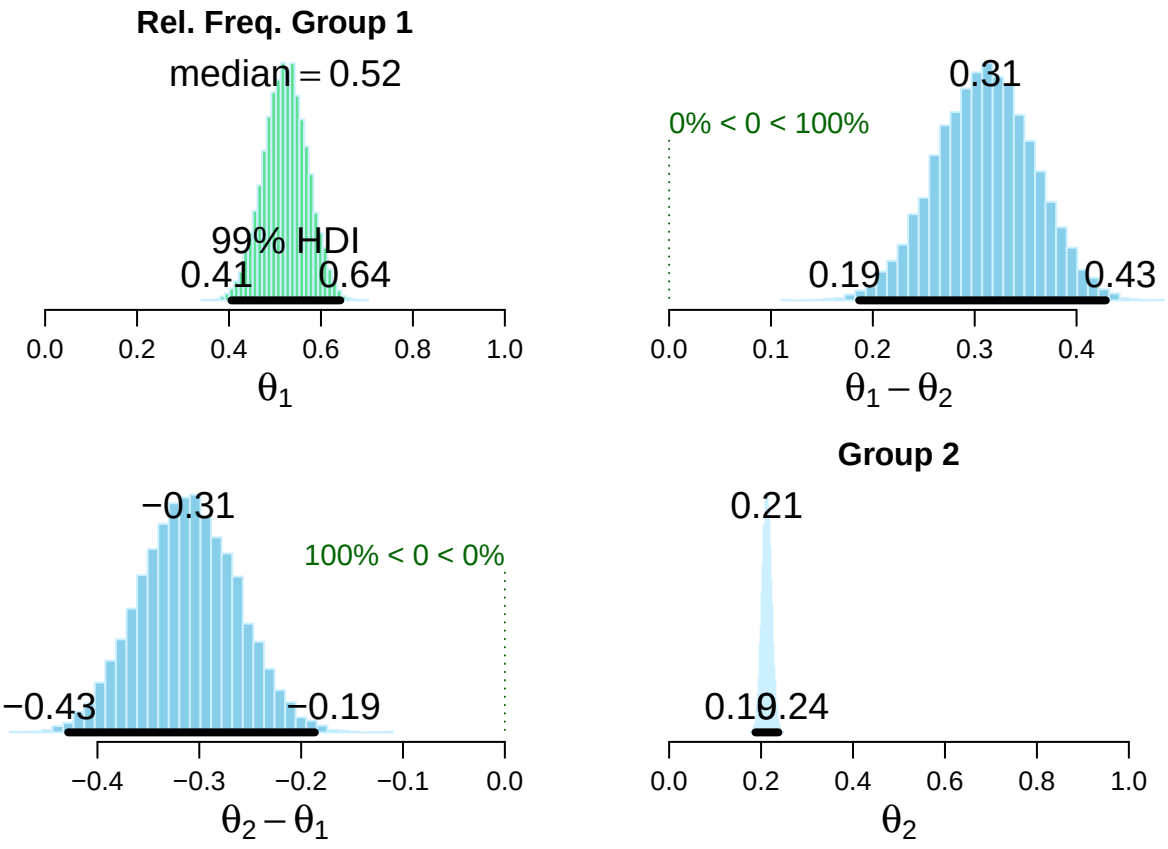
Die Histogramme der A-posterior-Verteilungen zweigen unimodale, symmetrische Form mit geringer Streuung. Das Modell berrechnet eine mittlere Überlebenswahrscheinlichkeit für Frauen von 0.73, wobei das 99%-Konfidenzintervall für die tatsächliche Wahrscheinlichkeit im Bereich [0.676,0.782] liegt. Für Männer wird eine Überlebenswahrscheinlichkeit von 0.212 errechnet, wobei das 99%-Konfidenzintervall im Bereich [0.188,0.236] liegt. Somit ergibt sich eine durchschnittliche Differenz der Überlebensraten von 0.519, mit einem 99%-Konfidenzintervall für die tatsächliche Differenz im Bereich [0.457, 0.574]. Frauen hatten somit deutlich höhere Überlebenschancen.

```
bayesFit_Titanic_children <- BayesianFirstAid::bayes.prop.test(
  c(children["Yes"],men["Yes"]),c(sum(children), sum(men)), cred.mass = 0.99)
summary(bayesFit_Titanic_children)
```

```
##   Data
## number of successes:    57,  367
## number of trials:      109, 1731
##
##   Model parameters and generated quantities
## theta[i]: the relative frequency of success for Group i
## x_pred[i]: predicted number of successes in a replication for Group i
## theta_diff[i,j]: the difference between two groups (theta[i] - theta[j])
##
##   Measures
##                    mean     sd   HDIlo   HDIup %<comp %>comp
```

```
## theta[1]          0.523  0.047   0.405   0.642  0.316  0.684
## theta[2]          0.212  0.010   0.188   0.238  1.000  0.000
## x_pred[1]        56.965  7.331  39.000  75.000  0.000  1.000
## x_pred[2]       367.922 24.049 306.000 428.000  0.000  1.000
## theta_diff[1,2]   0.310  0.048   0.187   0.429  0.000  1.000
##
## 'HDIlo' and 'HDIup' are the limits of a 99% HDI credible interval.
## '%<comp' and '%>comp' are the probabilities of the respective parameter being
## smaller or larger than 0.5 (except for the theta_diff parameters where
## the comparison value comp is 0.0).
##
##   Quantiles
##                    q2.5%    q25%  median    q75%  q97.5%
## theta[1]           0.431   0.491   0.523   0.555   0.613
## theta[2]           0.193   0.206   0.212   0.219   0.232
## x_pred[1]         43.000  52.000  57.000  62.000  71.000
## x_pred[2]        321.000 351.000 368.000 384.000 415.000
## theta_diff[1,2]    0.216   0.277   0.311   0.343   0.403
```

```
plot(bayesFit_Titanic_children)
```




```
df_Titanic_children <- as.data.frame(bayesFit_Titanic_children)
```

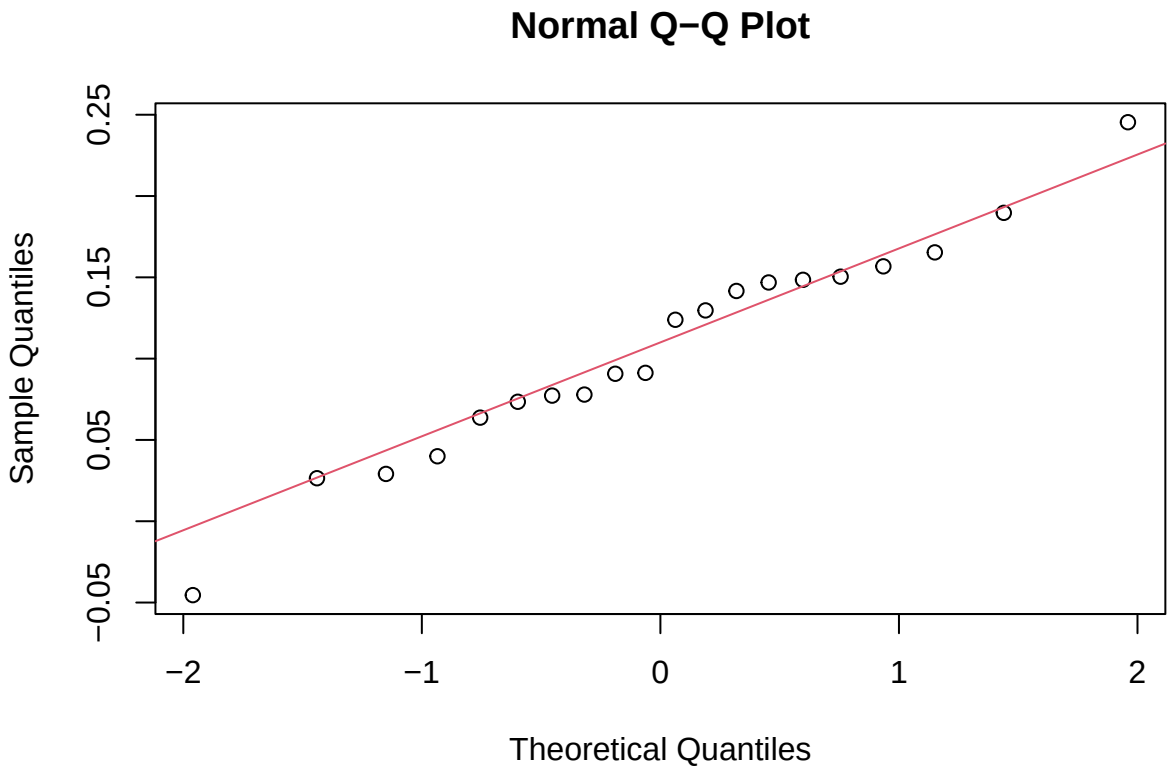
Das Histogramm der A-posterior-Verteilung für Kinder zeigt eine unimodale, symmetrische Form, wobei die Streuung deutlich größer ist als bei den Frauen oder Männern aufgrund der kleineren Stichprobengröße. Daraus ergibt sich auch die breitere Streuung der Differenzen. Das Modell berrechnet eine mittlere

Überlebenswahrscheinlichkeit für Kinder von 0.523, wobei das 99%-Konfidenzintervall für die tatsächliche Wahrscheinlichkeit im Bereich [0.405, 0.642] liegt. Somit ergibt sich eine durchschnittliche Differenz der Überlebensraten von 0.310, mit einem 99%-Konfidenzintervall für die tatsächliche Differenz im Bereich [0.187, 0.429]. Kinder hatten somit ebenfalls deutlich höhere Überlebenschancen.

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.

```
growth_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(growth_data)
qqline(growth_data,col=2)
```




```
shapiro.test(growth_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  growth_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 $\mu = 0.01$ 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(growth_data, mu = 0.01,
alternative = c("greater"))
```

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

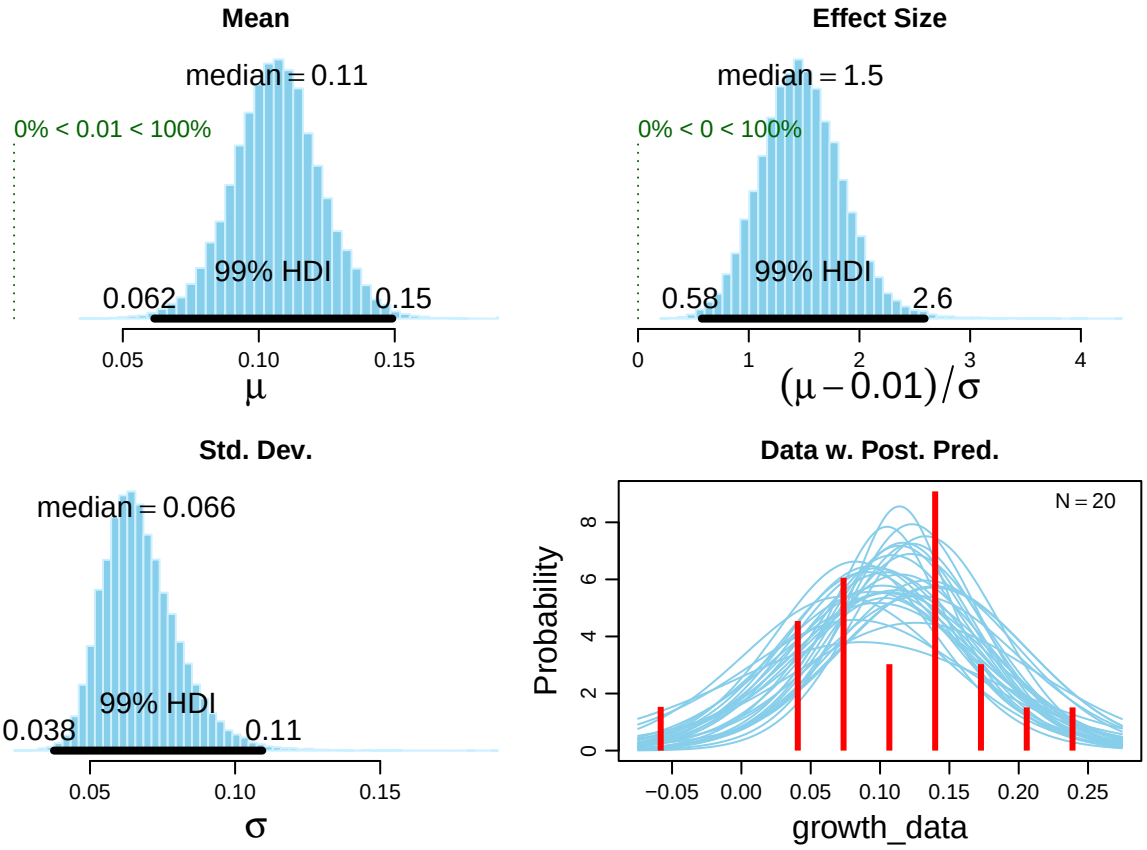
ERWEITERUNG - BAYES STATISTIK

```
bayesFit_growth <- bayes.t.test(growth_data, mu = 0.01, cred.mass = 0.99)
summary(bayesFit_growth)
```

```
##   Data
## growth_data, n = 20
##
##   Model parameters and generated quantities
## mu: the mean of growth_data
## sigma: the scale of growth_data , a consistent
##   estimate of SD when nu is large.
## nu: the degrees-of-freedom for the t distribution fitted to growth_data
## eff_size: the effect size calculated as (mu - 0.01) / sigma
## x_pred: predicted distribution for a new datapoint generated as growth_data
##
##   Measures
##            mean     sd  HDIlo   HDIup %<comp %>comp
## mu        0.107  0.016  0.062   0.149  0.000  1.000
## sigma     0.068  0.013  0.038   0.109  0.000  1.000
## nu       32.907 29.919  1.165 141.259  0.000  1.000
```

```
## eff_size  1.483  0.375  0.576   2.589  0.000  1.000
## x_pred    0.107  0.081 -0.117   0.324  0.092  0.908
##
## 'HDIlo' and 'HDIup' are the limits of a 99% HDI credible interval.
## '%<comp' and '%>comp' are the probabilities of the respective parameter being
## smaller or larger than 0.01.
##
##   Quantiles
##           q2.5%   q25% median   q75%  q97.5%
## mu        0.075  0.096  0.107  0.117   0.139
## sigma     0.046  0.058  0.066  0.075   0.098
## nu        3.328 12.068 23.910 44.089 113.429
## eff_size  0.794  1.229  1.468  1.719   2.253
## x_pred   -0.046  0.060  0.107  0.154   0.259
```

```
plot(bayesFit_growth)
```



Das bayesianische Modell liefert einen a-posterior-Mittelwert von 0.107 mit 99%-HDI im Bereich von [0.062, 0.149]. Somit bietet das Medium einen deutlichen Wachstumsvorteil.

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.

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

```
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
```

```
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
```

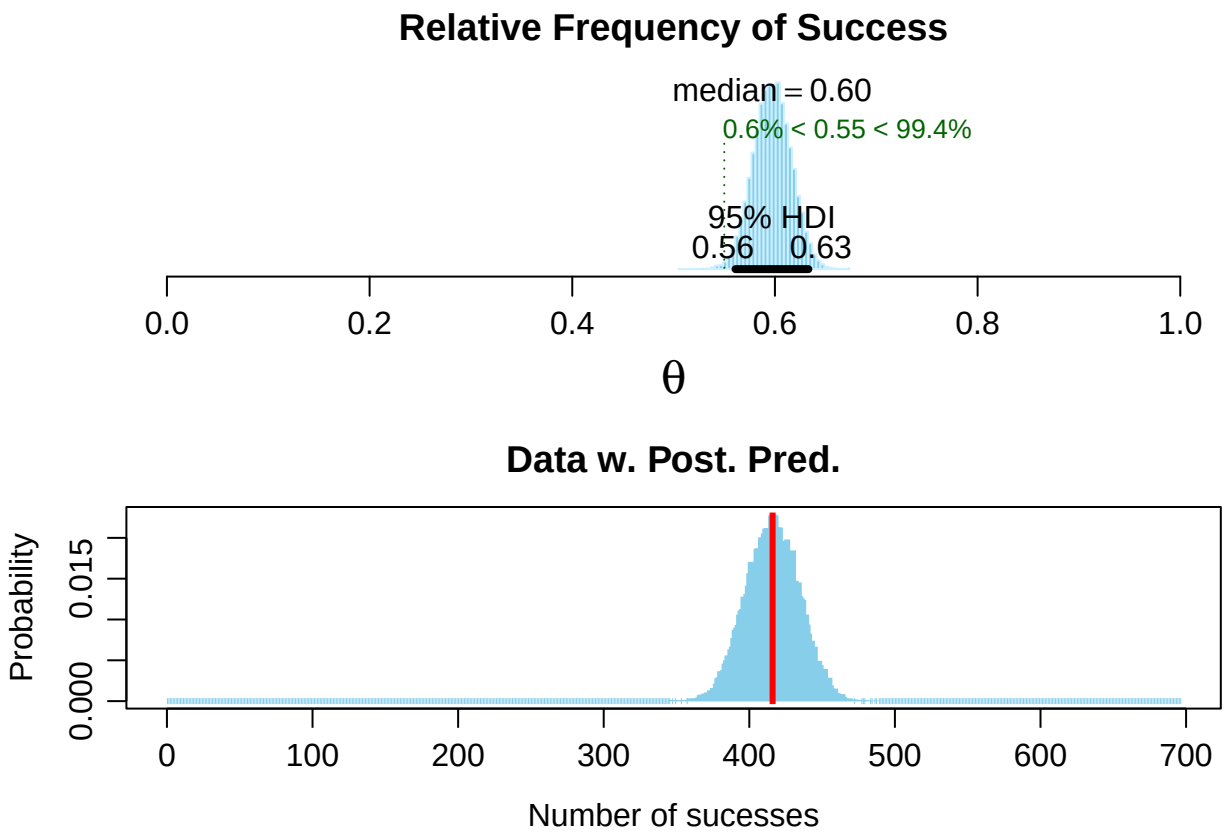
Die Teststatistik beträgt X-squared = 5.1207, der p-Wert liegt bei 0.02364. Es handelt sich um einen 2-seitigen Test, daher kann die Nullhypothese auf dem 5% Niveau verworfen werden, muss auf dem 1% Niveau jedoch beibehalten werden. Dies wird auch anhand der zugehörigen Konfidenzintervalle sichtbar. 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.

ERWEITERUNG - BAYES STATISTIK

```r
bayesFit_pathway95 <- BayesianFirstAid::bayes.prop.test(416, 696, comp.theta = 0.55, cred.mass = 0.95)
bayesFit_pathway99 <- BayesianFirstAid::bayes.prop.test(416, 696, comp.theta = 0.55, cred.mass = 0.99)
summary(bayesFit_pathway95)
```

```
##   Data
## number of successes = 416, number of trials = 696
##
##   Model parameters and generated quantities
## theta: the relative frequency of success
## x_pred: predicted number of successes in a replication
##
##   Measures
##           mean     sd   HDIlo   HDIup %<comp %>comp
## theta    0.597  0.018   0.561   0.633  0.006  0.994
## x_pred 415.882 18.153 380.000 450.000  0.000  1.000
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## '%<comp' and '%>comp' are the probabilities of the respective parameter being
## smaller or larger than 0.55.
##
##   Quantiles
##          q2.5%    q25%  median   q75%  q97.5%
## theta    0.561   0.585   0.597   0.61   0.633
## x_pred 380.000 404.000 416.000 428.00 451.000
```

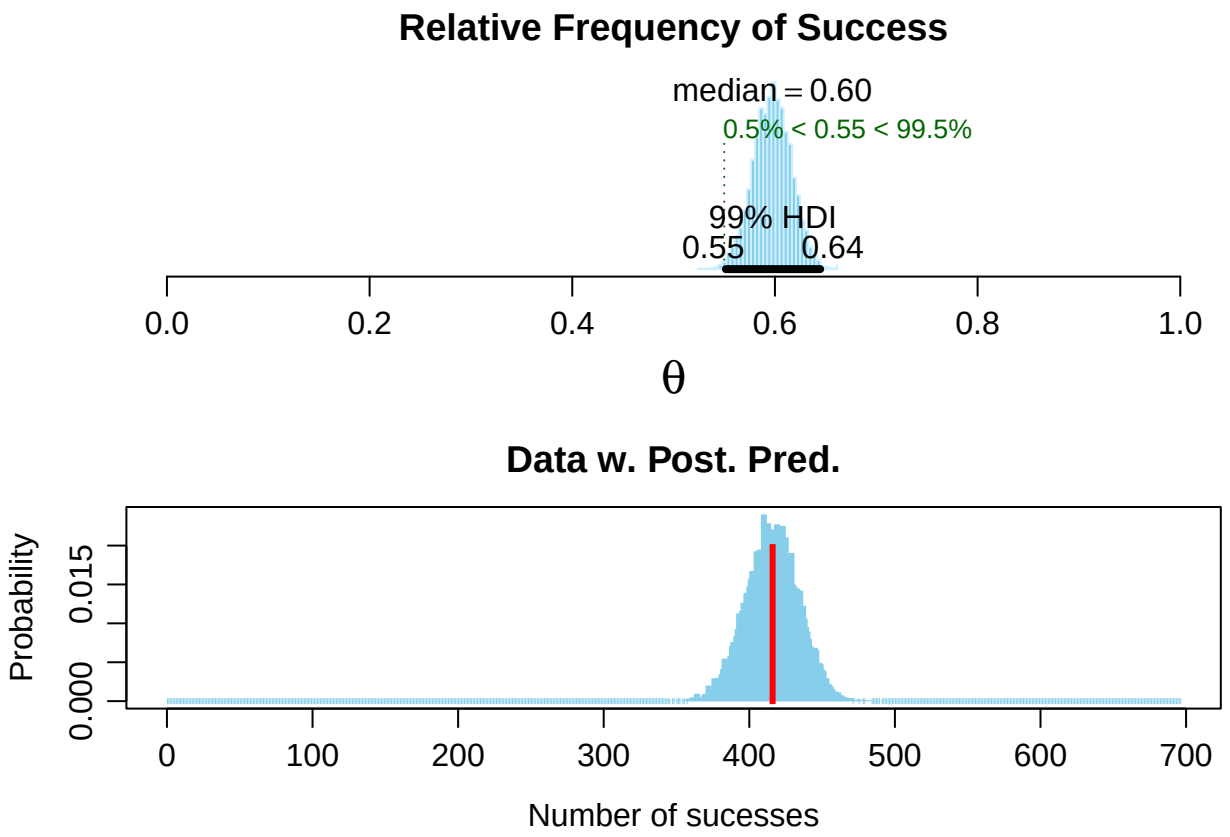
```r
plot(bayesFit_pathway95)
```




```
summary(bayesFit_pathway99)
```

```
##   Data
## number of successes = 416, number of trials = 696
##
##   Model parameters and generated quantities
## theta: the relative frequency of success
## x_pred: predicted number of successes in a replication
##
##   Measures
##           mean     sd   HDIlo   HDIup %<comp %>comp
## theta    0.597  0.019   0.551   0.645  0.005  0.995
## x_pred 415.793 18.244 368.000 461.000  0.000  1.000
##
## 'HDIlo' and 'HDIup' are the limits of a 99% HDI credible interval.
## '%<comp' and '%>comp' are the probabilities of the respective parameter being
## smaller or larger than 0.55.
##
##   Quantiles
##         q2.5%    q25%  median   q75%  q97.5%
## theta    0.56   0.585   0.598   0.61   0.633
## x_pred 379.00 404.000 416.000 428.00 451.000
```

```
plot(bayesFit_pathway99)
```



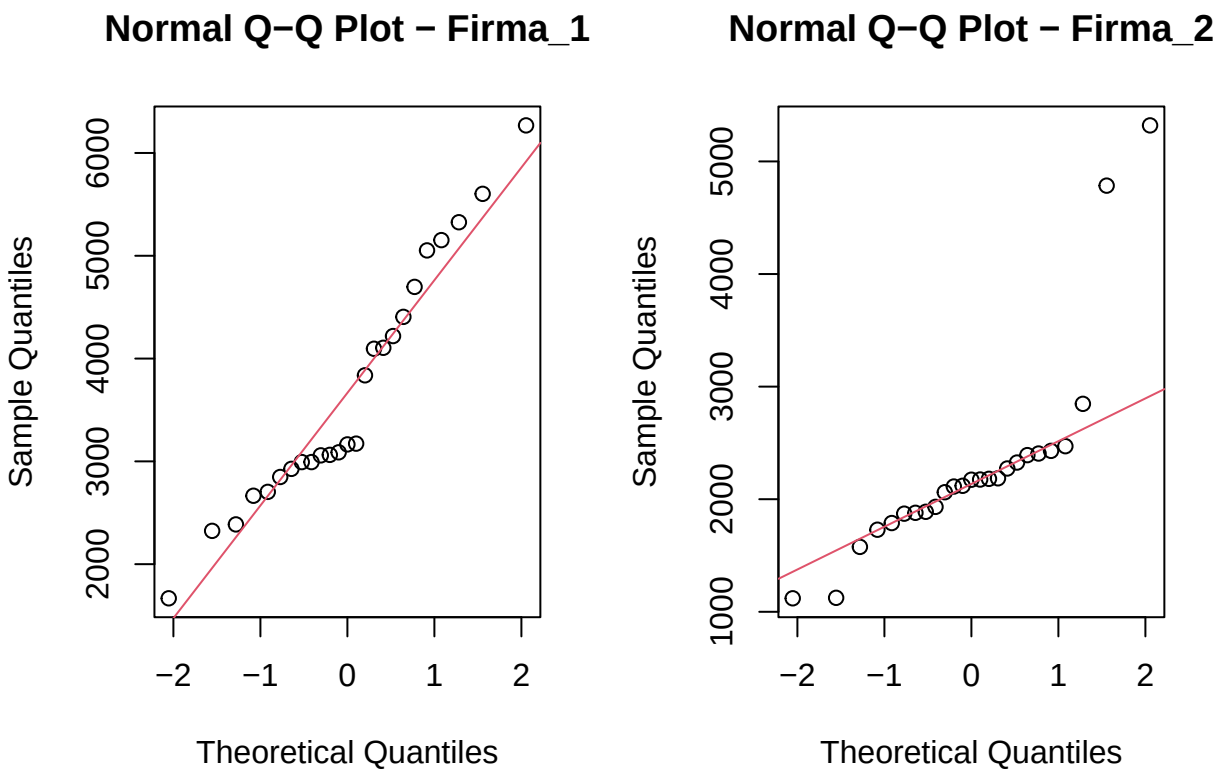
Das bootstrapping Modell liefert einen theta-Wert von 0.597 bei einem 99%-HDI im Bereich [0.551, 0.645]. Die Anzahl der gefundenen Gene ist somit auf dem 1% Niveau signifikant höher als bei der vorhergehenden Studie, im Mittel um 0.048.

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?

```
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)

comp2_woOut <- 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)
```

```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
shapiro.test(comp2_woOut)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  comp2_woOut
## W = 0.9391, p-value = 0.1719
```

```r
summary(comp1)
```

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

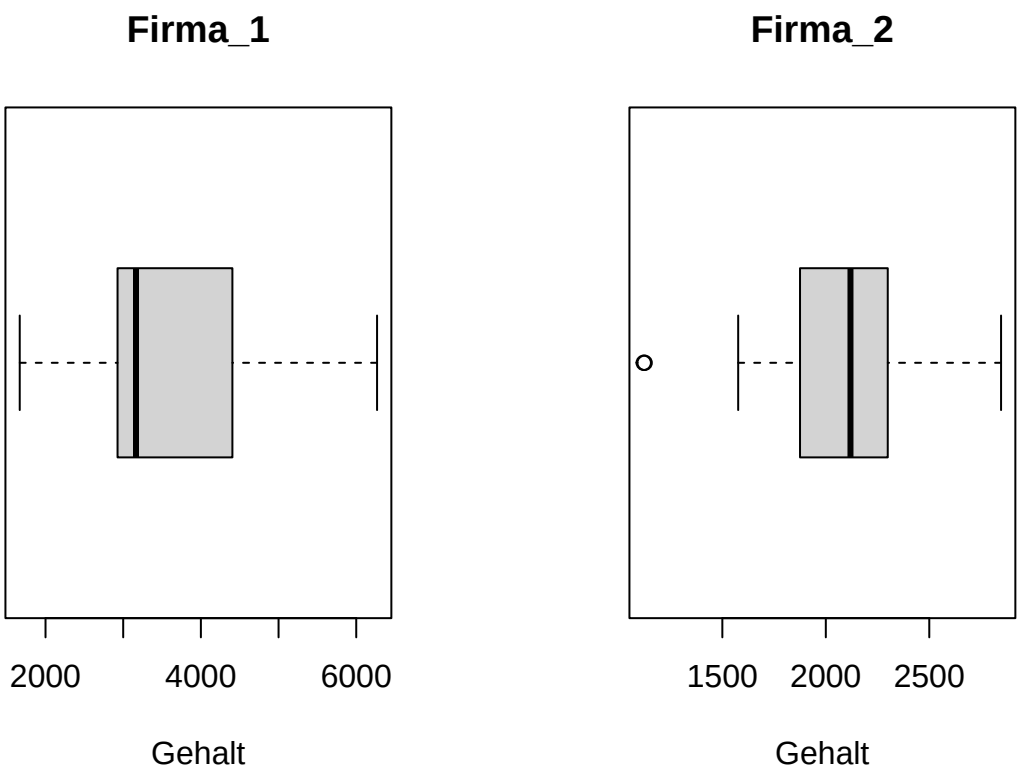
```r
summary(comp2)
```

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

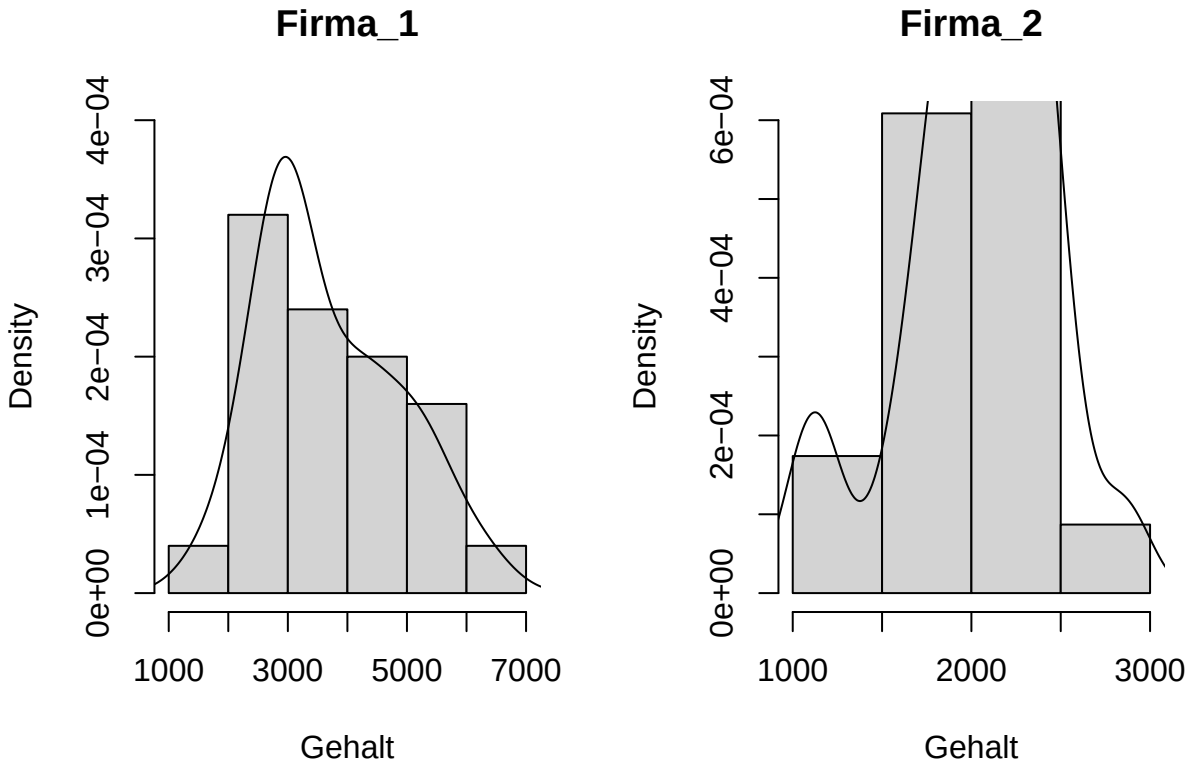
```r
summary(comp2_woOut)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1120    1875    2119    2046    2299    2847
```

```r
boxplot(comp1, horizontal = T, main = "Firma_1", xlab = "Gehalt")
boxplot(comp2_woOut, 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_woOut, freq = F, main = "Firma_2", ylim = c(0,0.0006), xlab = "Gehalt")
lines(density(comp2_woOut))
```



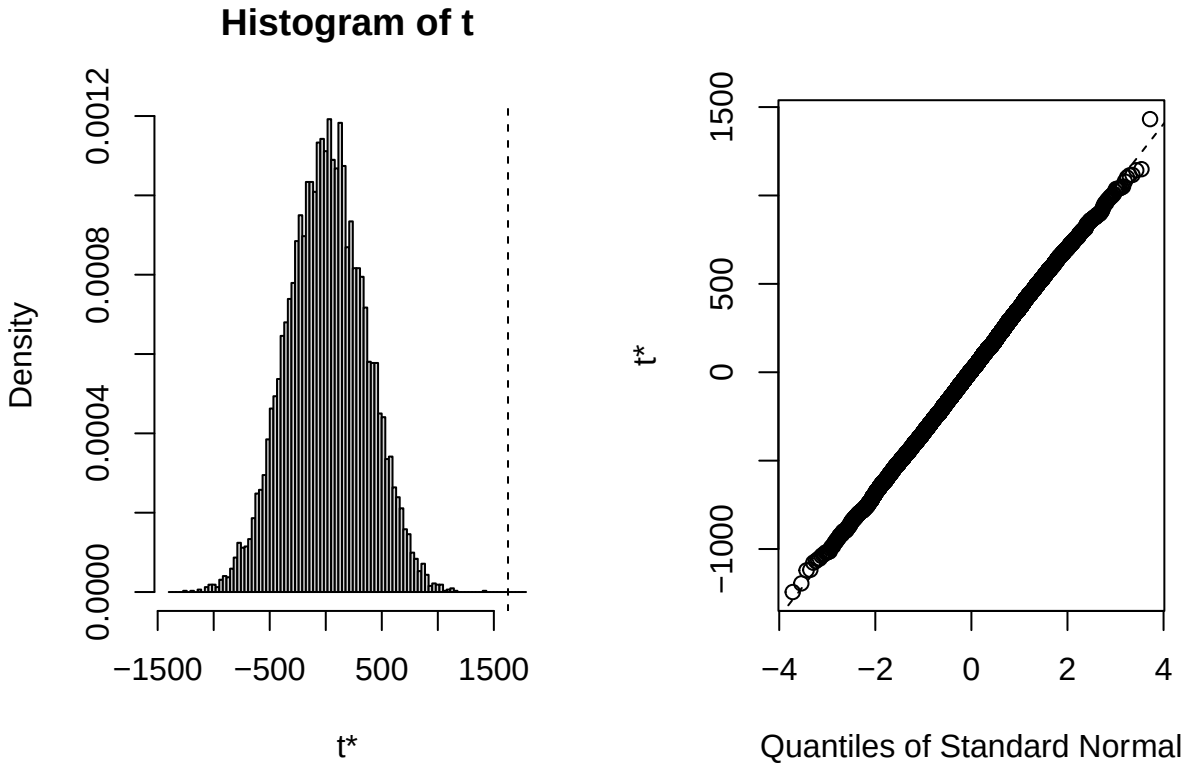
Im Q-Q-Plot von Firma 2 sind zwei deutliche Ausreißer am rechten oberen Rand zu sehen, welche in weiterer Folge entfernt werden. Der Q-Q-Plot und das Histogramm zu den Daten von Firma 1 zeigen jeweils 2 Modi, somit kann der Shapiro-Test nicht angewandt werden und von Normalverteilung ist nicht auszugehen. Aufgrund dieser Datenverteilung ist kein statistischer Test anwendbar und es muss auf simulations-basierte Methoden zurückgegriffen werden.

ERWEITERUNG - BOOTSTRAPPING

Somit wird für die Analyse auf Bootstrapping zurückgegriffen.

```r
salaries <- c(comp1, comp2_woOut)

boot_fn <- function(data, indices) {
  comp1_indices <- indices <= 25
  comp2_indices <- indices > 25

  # Data subsets based on indices
  group1_salaries <- data[comp1_indices]
  group2_salaries <- data[comp2_indices]

  #Calculate
  mean_group1 <- mean(group1_salaries)
```

```
  mean_group2 <- mean(group2_salaries)
  diff_means <- mean_group1 - mean_group2

  return(diff_means)
}

# Set seed for reproducibility
set.seed(123)

# Perform bootstrap resampling
boot_res5 <- boot(salaries, boot_fn, R = 10000)
plot(boot_res5)
```




```
# Calculate bootstrap confidence intervals
(boot_res5_ci95 <- boot.ci(boot_res5, type = "perc", conf = 0.95))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res5, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   (-668,  691 )
```

```
## Calculations and Intervals on Original Scale
```

```
(boot_res5_ci95 <- boot.ci(boot_res5, type = "perc", conf = 0.99))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res5, conf = 0.99, type = "perc")
##
## Intervals :
## Level     Percentile
## 99%   (-878,  878 )
## Calculations and Intervals on Original Scale
```

Das Histogram der Bootstrap-Strichproben zeigt eine unimodale Verteilung mit einem zentralen Peak nahe Null und guter Symmetrie und der Q-Q-plot zeigt kaum Abweichungen von der theoretischen Normalverteilung. Bootstrapping liefert ein 99%-Konfidenzintervall für die tatsächlich Differenz im Bereich von [-878, 878], somit kann kein signifikanter Unterschied zwischen den Gehältern der beiden Firmen festgestellt werden.