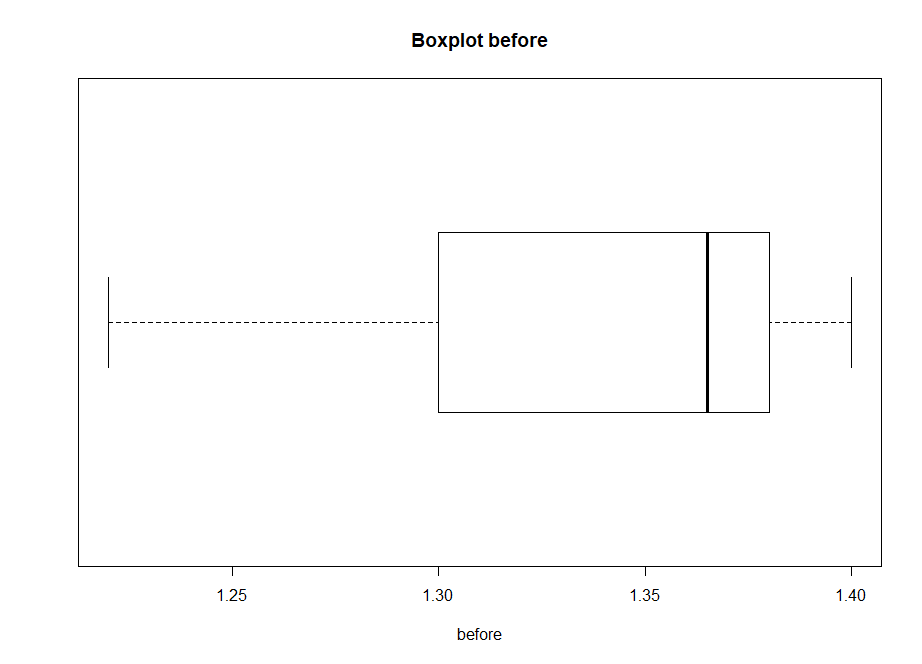
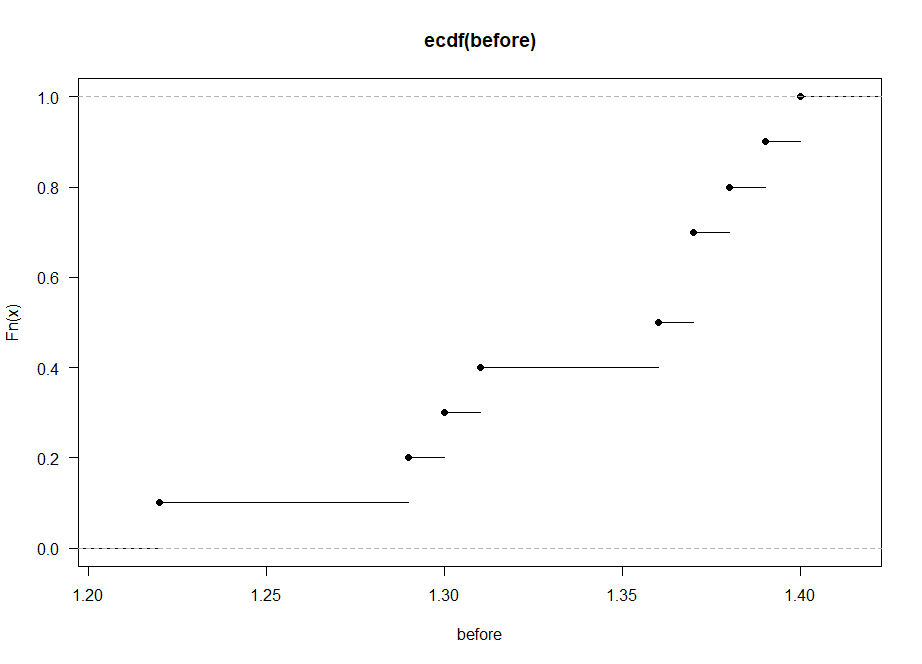
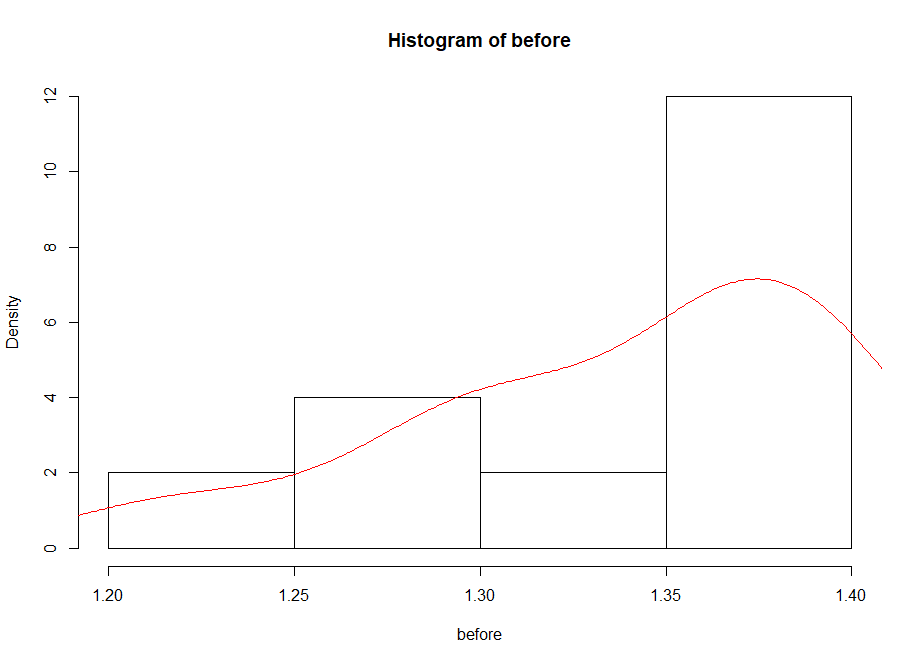
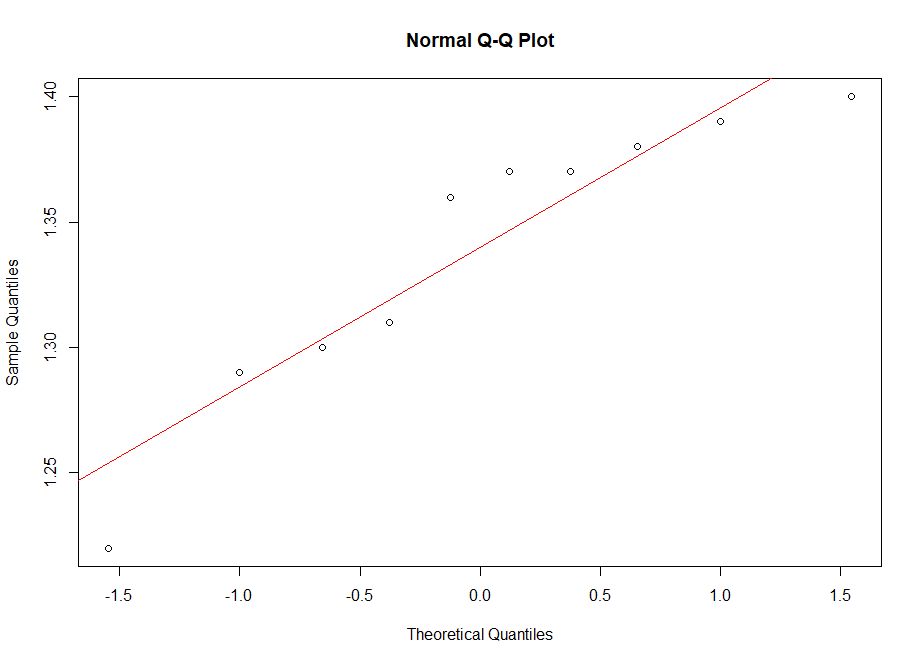
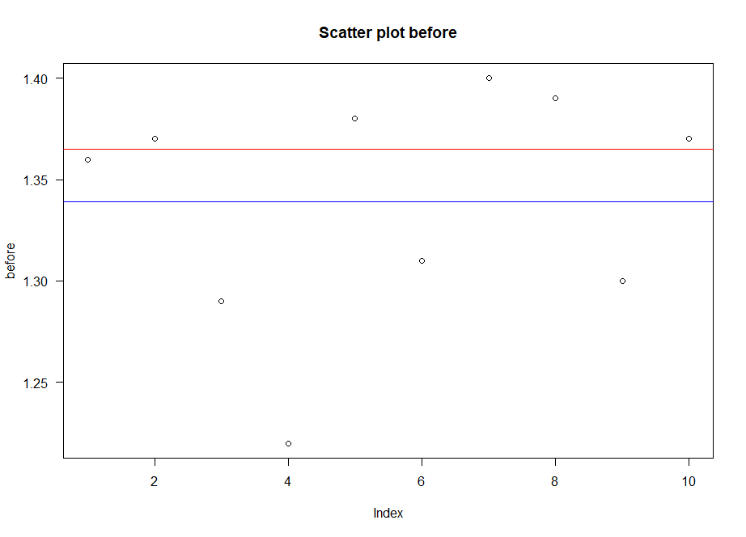
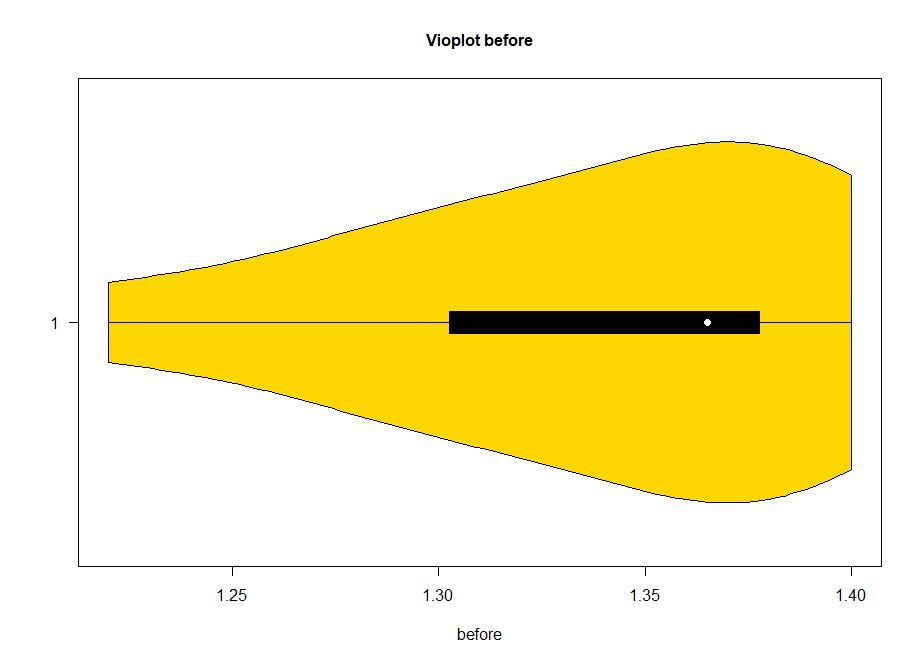
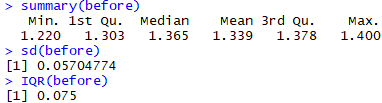
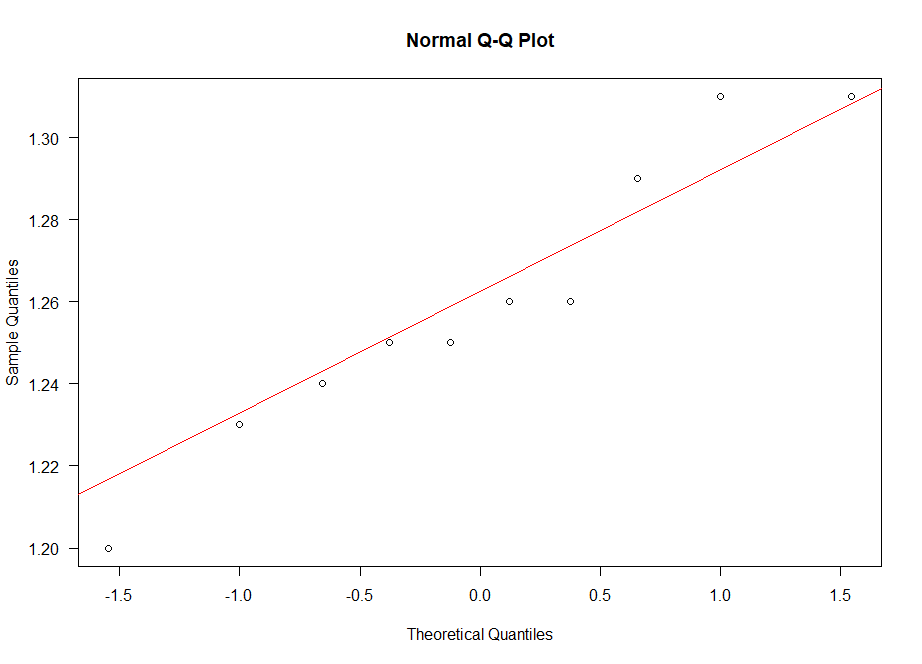
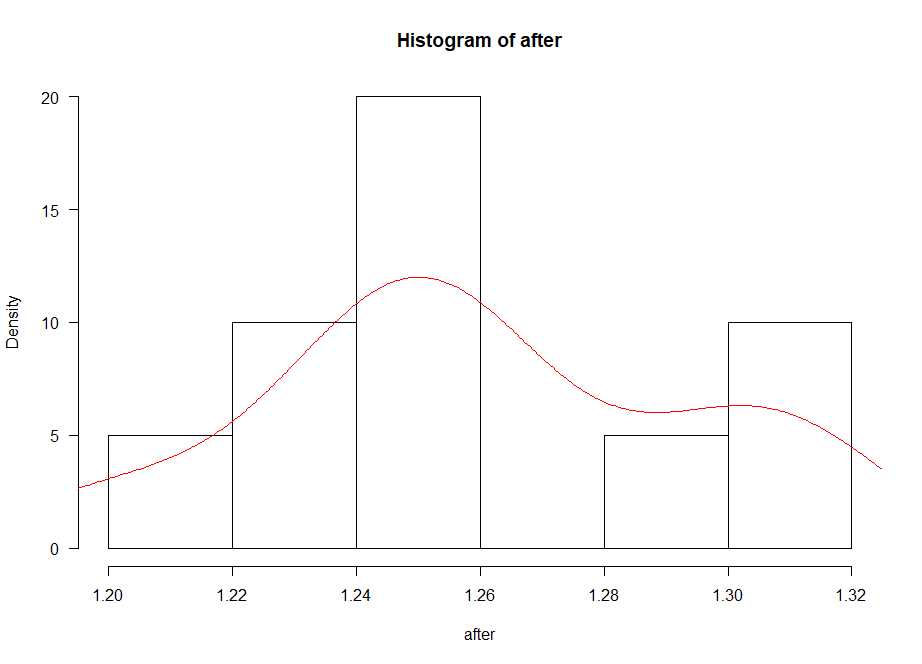
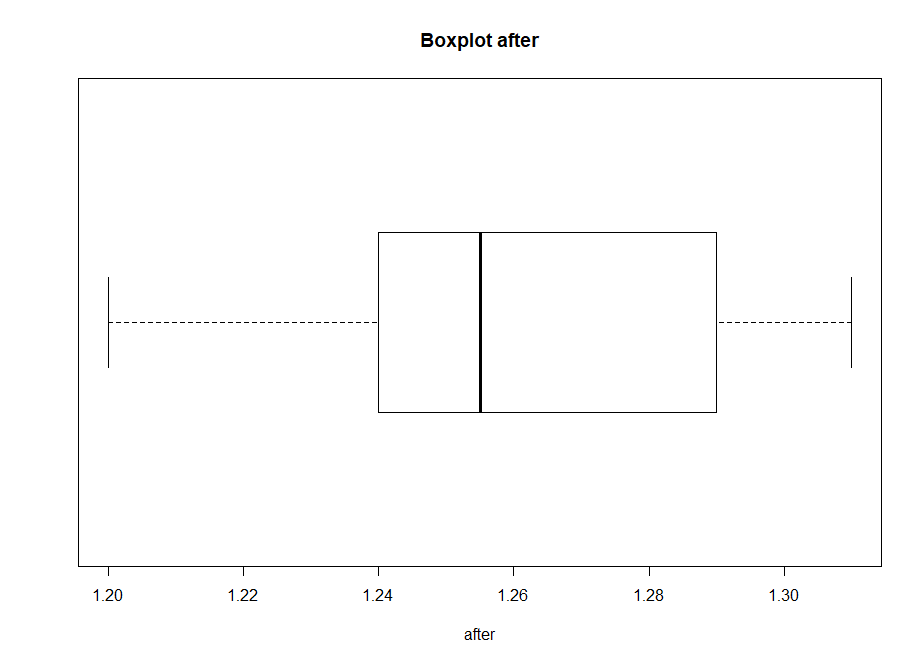
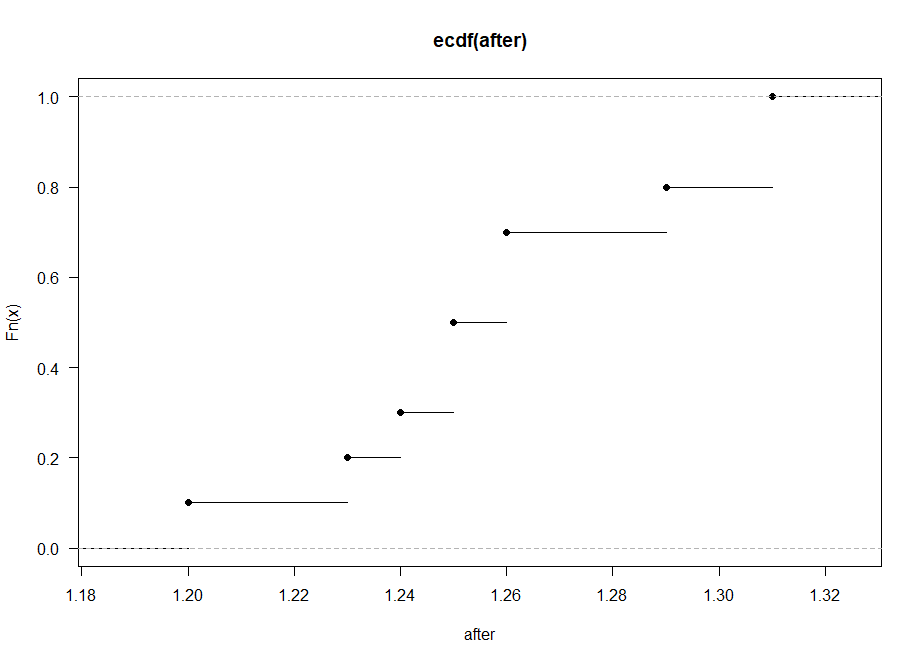
**HÜ3**

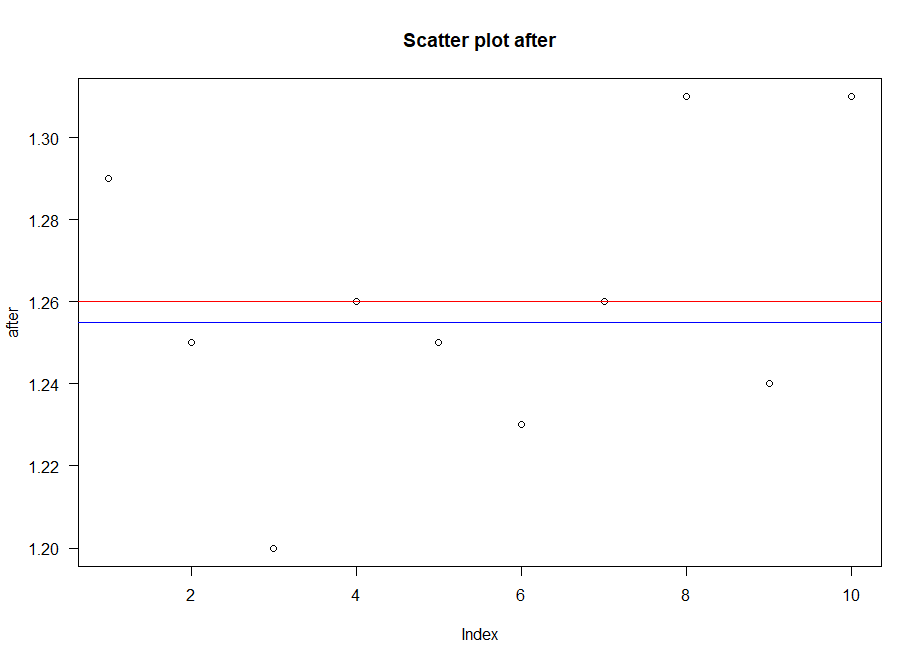
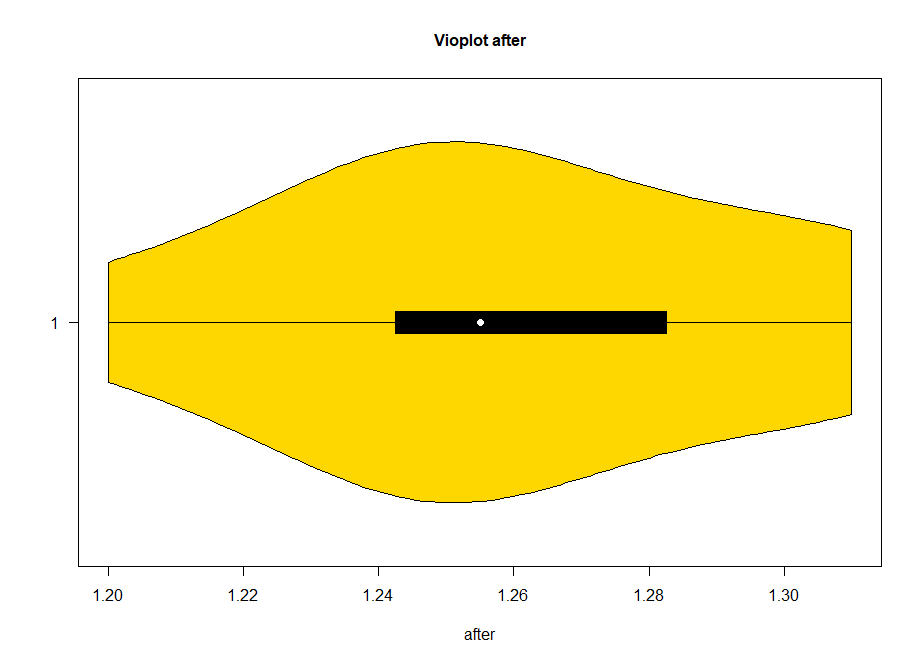
1.

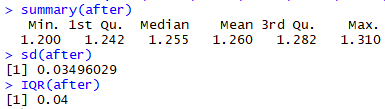




Die Kombination aus dem Histogramm und Dichtefunktion (rote Dichtelinie) veranschaulicht, dass die Daten bimodal sind. QQ-Plot, ECDF, Violin-Plot und Boxplot weisen auf einen links schiefen Rand. Aus dem QQ-Plot lässt sich erkennen, dass links unten auf einen schweren Rand und rechts oben auf einen leichten Rand vorhanden sind. Die Tendenz ist gering, da sich die Werte am Ende wieder zurückpendeln, obwohl ein paar Werte in der Mitte von der Gerade abweichen. Insgesamt liegen zwei Werte nicht optimal auf dem QQ-Plot. Bei Scatterplot sieht man, dass der Median (roter Strich; 1.365) und der Mittelwert (schwarzer Strich;1,339) kaum voneinander abweichen. Diese machen aber keinen Sinn für die Lageabschätzung. Das heißt die Daten sind nicht normalverteilt.

After:





Die Kombination aus dem Histogramm und Dichtefunktion (rote Dichtelinie) veranschaulicht, dass die Daten bimodal sind. Aus dem QQ-Plot lässt sich erkennen, dass links unten ein schwerer Rand und rechts zwei abweichende Werte die einen schweren Rand bilden. Aus dem Violine-Plot lässt sich der Effekt kaum erkennen. Aus dem Boxplot lässt sich die Bimodalitätverhalten besser erkennen als bei Violineplot. Die Tendenz zum schweren Rand ist gering. Die Daten aus ECDF und Scatterplot weisen auf rechts schiefe Verteilung auf. Bei Scatterplot sieht man, dass der Median (roter Strich; 1.255) und der Mittelwert (schwarzer Strich;1,260) kaum voneinander abweichen. Da die Daten schwere Ränder aufweisen, machen Lageschätzer keinen Sinn. Das heißt die Daten sind nicht normalverteilt.

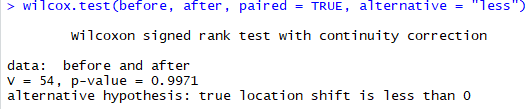
Aus beiden Messungen lässt sich schließen, dass es zu wenig Stichprobenzahl hat. Es wäre aussagekräftiger, wenn wir die Daten für 25 Messungen, oder mehr, pro Mitarbeiter hätten und nicht nur die Mittelwerte pro Mitarbeiter.

Hypothesentestung:

HO: Die mittleren Arbeitszeiten sind vor und nach dem Training ident

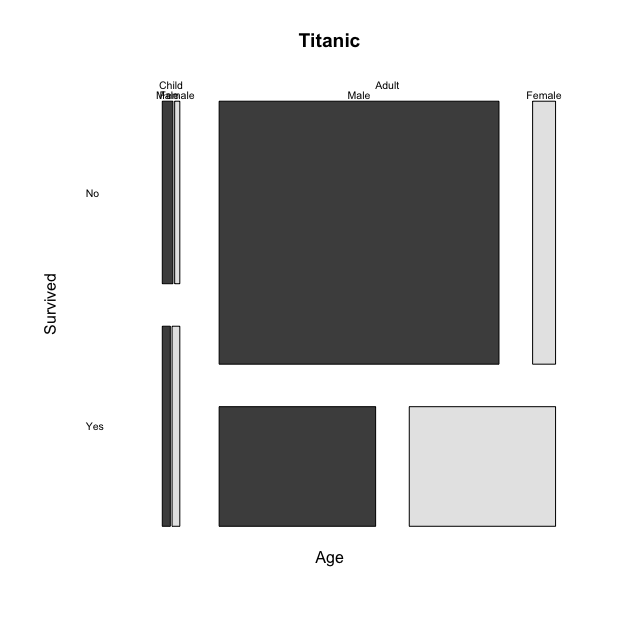
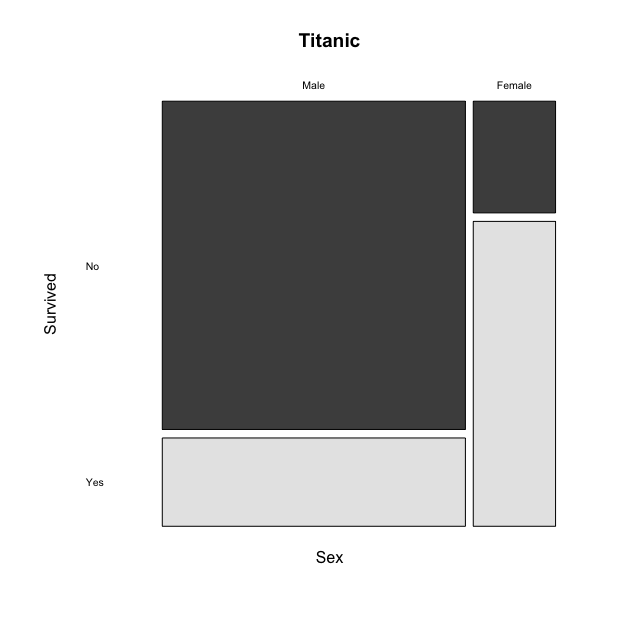
H1: Die mittleren Arbeitszeiten sind nach dem Training signifikant geringer als davor.

Da die Daten nicht parametrisch bzw. nicht normalverteilt sind, werden die Mittelwerte mit Wilcoxon Test verglichen. Außerdem wurde diese Testung durchgeführt, weil man zwei unterschiedliche Stichproben vergleichen kann. Konf.intervalle von 95% und 99% wird bei Wilcoxon nicht berechnet bzw. nicht relevant.



V=54 ist Teststatistik; Ha= Location shift ist die Verschiebung von before auf after. True location steht bei Wilcoxon Rankensummen Test für die Nachfrage über die Lage und nicht nach der Mitte (wird bei t-test gefragt).

Der p-Wert 0.9971 ist größer als 0.05. Das sagt aus, dass die H0 nicht verworfen werden kann. Somit heißt, dass das Training keinen signifikanten Einfluss auf die mittleren Arbeitszeiten der Pipettiervorgängen. Das Training ist somit nicht notwendig.

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.)

Es wird hier graphisch dargestellt, dass die Überlebenschance der Kinder verglichen mit Männern und Frauen verglichen mit Männern. Frauen haben mehr als die Hälfte überlebt und Kinder auch.

|  |  |  |
| --- | --- | --- |
| **Survival** | | |
| **Sex** | **Yes** | **No** |
| Men | 338 | 1329 |
| Women | 316 | 109 |
| Children | 57 | 52 |

Hinweis: Daten aus der Hausübung 2, Beispiel 4.

Test (Propotions test) Männer-Frauen:

Hypothesenaufstellung:

H0: Die Frauen (p1) haben keine höhere Überlebenswahrscheinlichkeit als die Männer (p2). D.h. p1≤ p2.

HA: Die Frauen (p1) haben höhere Überlebenswahrscheinlichkeit als die Männer (p2). D.h. p1 > p2.

> dead<- c(109, 1329) #109 tote Frauen; 1329 tote Männer

> total<- c(425,1667) #Gesamt Stichprobengröße (tote + überlebende) 425 Frauen und gesamte Stichprobengröße (tote + überlebende) 1667 Männer

> prop.test(dead,total, alternative = "less", conf.level = 0.99)

2-sample test for equality of proportions with continuity correction

data: dead out of total

X-squared = 458.35, df = 1, p-value < 2.2e-16

alternative hypothesis: less

99 percent confidence interval:

-1.0000000 -0.4849515

sample estimates:

prop 1 prop 2

0.2564706 0.7972406

X-squared von 458.35 ist die Teststatistik von X2-Verteilung; df ist die Abkürzung für die Freiheitsgrade minus 1. Das heißt die Orginale-Stichprobengröße ist 2. 2-1=1 Freiheitsgrad. Der Konfidenzintervall im Bereich von -1 bis -0.4849 befinden sich 99% der Daten. Geschätzter Wert von Frauen ist 0.2565 und bei Männern 0.7972. Wenn man die zwei Werte miteinander addiert, dann kommt ein negativer Wert raus. Das ist erstes Zeichen, dass H0 verworfen werden kann. P-Wert von 2.2e-16 sagt aus, dass dieH0 verworfen werden kann da der p-Wert deutlich kleiner als der 1% Signifikanzniveau. Frauen hatten somit signifikant höhere Überlebenswahrscheinlichkeit als die Männer.

Test (Prop.test) Männer-Kinder:

H0: Die Kinder (p1) haben keine höhere Überlebenswahrscheinlichkeit als die Männer. D.h. p1≤ p2.

HA: Die Kinder (p1) haben eine höhere Überlebenswahrscheinlichkeit als die Männer (p2). D.h. p1 > p2.

> dead\_child\_men <-c(52, 1329)

> total\_child\_men<-c(109,1667)

> prop.test(dead\_child\_men,total\_child\_men, alternative = "less", conf.level = 0.99)

2-sample test for equality of proportions with continuity correction

data: dead\_child\_men out of total\_child\_men

X-squared = 58.807, df = 1, p-value = 8.694e-15

alternative hypothesis: less

99 percent confidence interval:

-1.0000000 -0.2016615

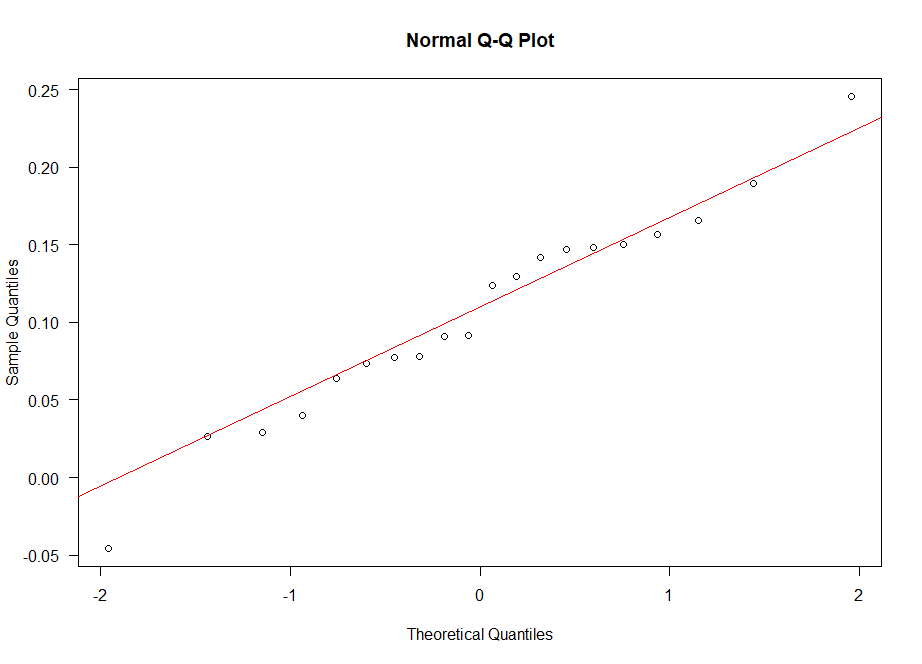
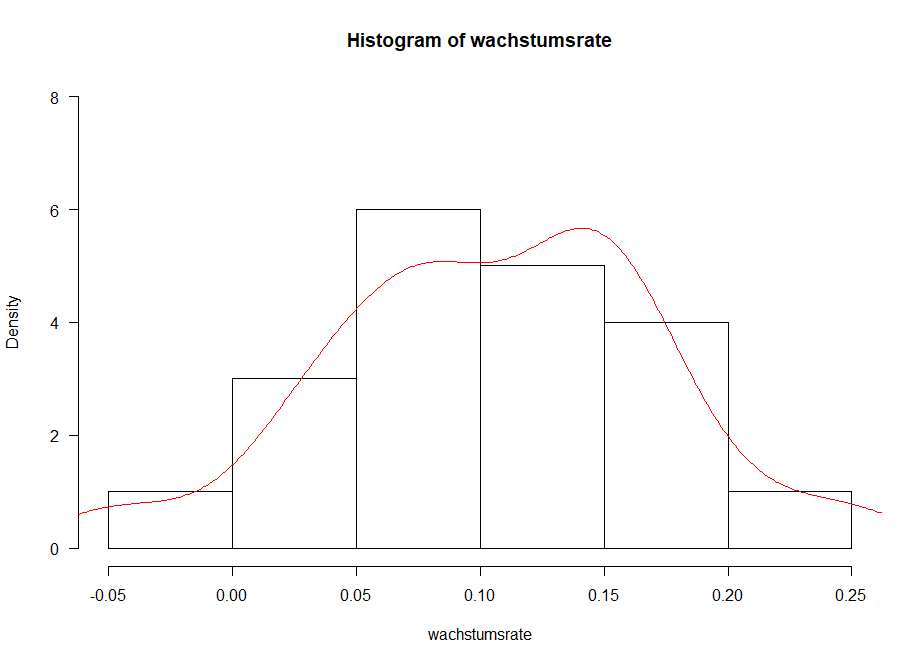
sample estimates:

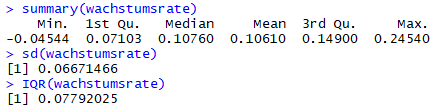
prop 1 prop 2

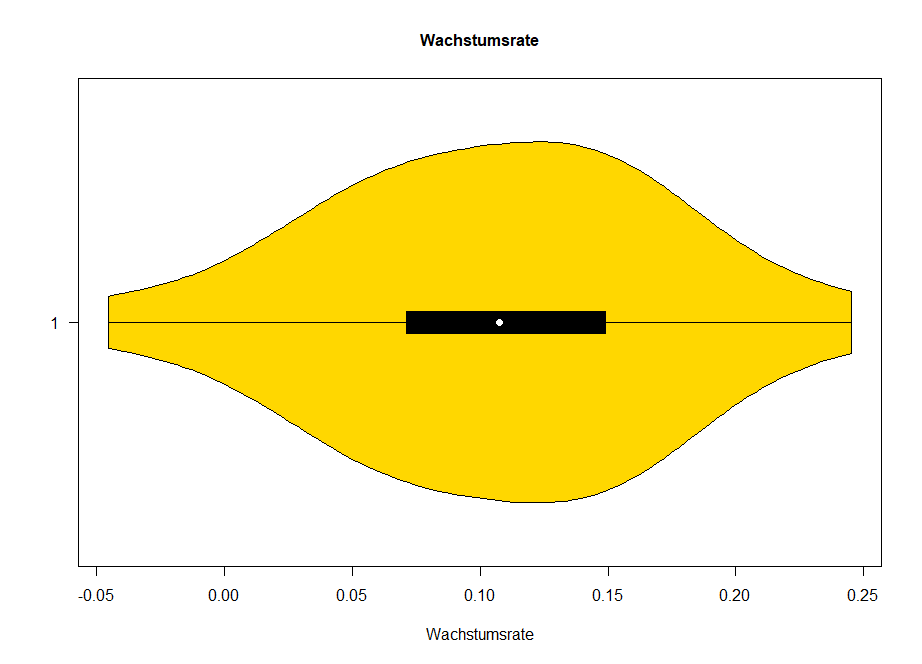
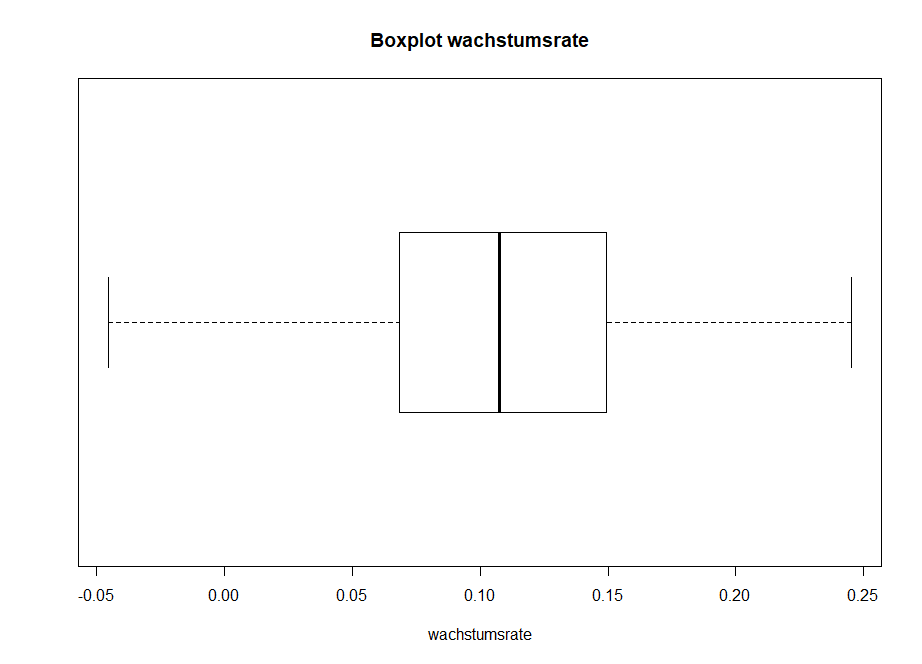
0.4770642 0.7972406

X-squared von 58,81 ist die Teststatistik von X2-Verteilung; df ist die Abkürzung für die Freiheitsgrade minus 1. Das heißt die Orginale-Stichprobengröße ist 2. 2-1=1 Freiheitsgrad. Der Konfidenzintervall im Bereich von -1 bis -0.2017 befinden sich 99% der Daten. Geschätzter Wert von Kindern ist 0.4771 und bei Männern 0.7972. Wenn man die zwei Werte miteinander addiert, dann kommt ein negativer Wert raus. Das ist erstes Zeichen, dass H0 verworfen werden kann. P-Wert von 8.694e-15 sagt aus, dass dieH0 verworfen werden kann da der p-Wert deutlich kleiner als der 1% Signifikanzniveau. Die Kinder hatten somit signifikant höhere Überlebenswahrscheinlichkeit als die Männer.

3.







Die Kombination aus dem Histogramm und Dichtefunktion (rote Dichtelinie) veranschaulicht, dass die Daten leicht bimodal sind. Aus dem Violine-Plot und Boxplot lässt sich die Bimodalität kaum erkennen. Aus dem QQ-Plot lässt sich erkennen, dass links unten und rechts oben ein schwerer Rand vorhanden ist. Dennoch die Tendenz ist gering, da nur zwei Werte von der Geradelinie abweichen. Die Daten zeigen dass der Mittelwert und der Median kaum voneinander abweichen. Sie können hier aber nicht als robuste Lageschätzer angenommen werden, weil die Daten nicht normalverteilt sind. Die Daten sind annährend symmetrisch, daher kann t-test verwendet werden.

Hypothesentestung:

H0: Wachstumsrate ist ≤ 1%

HA: Wachstumsrate ist > 1%

One-sample t-Test wird ausgeführt, da die Daten NV sind und Sigma unbekannt ist.

> t.test(wachstumsrate, alternative="greater", mu=0.01, conf.level = 0.95)

One Sample t-test

data: wachstumsrate

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

t-Wert (6.44) ist test-Statistik für die t-Verteilung; df ist die Abkürzung für die Freiheitsgrade minus 1. Das heißt die Orginale Stichprobengröße ist 20. 20-1=19 Freiheitsgrade.

Der Konfidenzintervall ist im Bereich von 0.08 bis unendlich, liegen 95% der Daten. Das heißt dass 1% Wachstumsrate nicht drinnen ist bzw. dass die Wachstumsrate höher als 1% ist. Geschätzter Wert von Wachstumsrate ist 0.106. H0 kann verworfen werden, weil p-Wert= 1.774 \*10^-6 kleiner als alpha (0.05) des Signifikantsniveau.

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 meiner 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 Gene sich signifikant gegenüber einer früheren Studie verändert hat, die 55% der Gene als beteiligt gefunden hat.

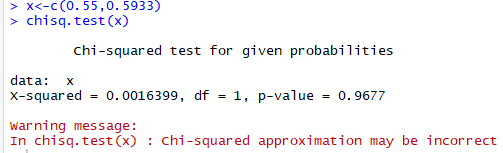
Ergebnis Studie 1: 55% der Gene beteiligt

Ergebnis Studie 2: 413/696 \*100= 59,33%

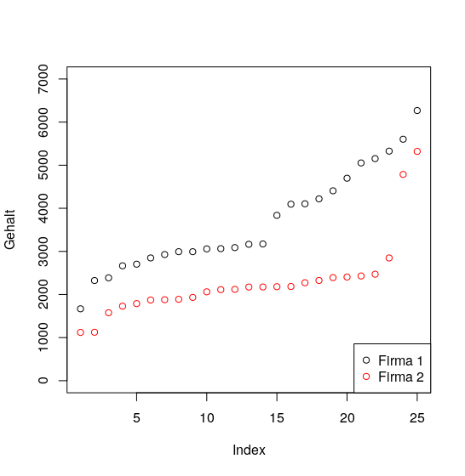
H0: Anteil der beteiligen Genen aus der Studie 2 hat sich gegenüber dem Anteil der beteiligten Genen aus der Studie 1 nicht verändert.

HA: Anteil der beteiligten Gene aus der Studie 2 gegenüber dem Anteil der beteiligten Genen aus der Studie 1 hat sich verändert.

Um die Unabhängigkeit von den zwei Variablen zu testen, wird X2 Test verwendet.

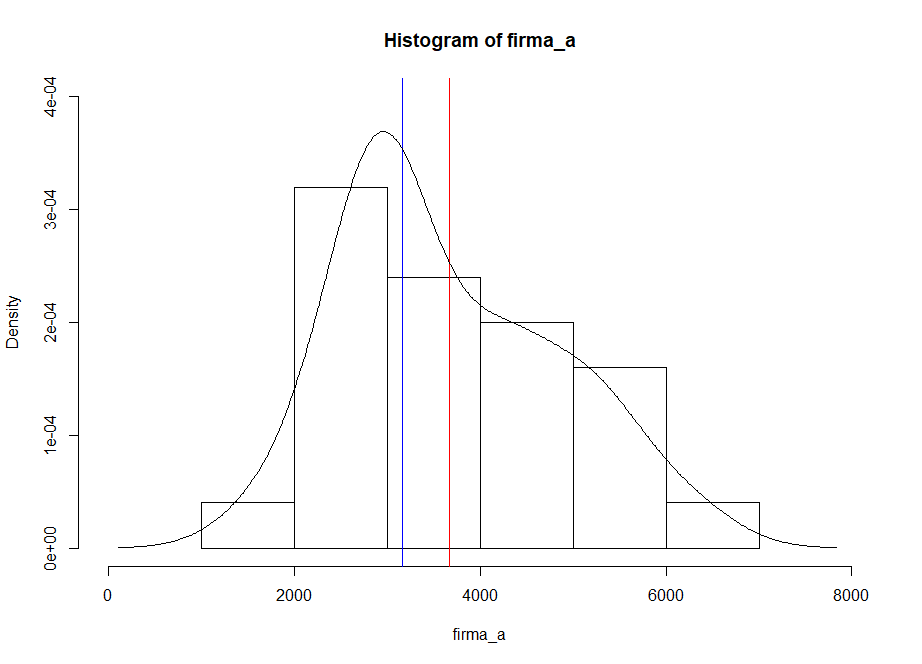
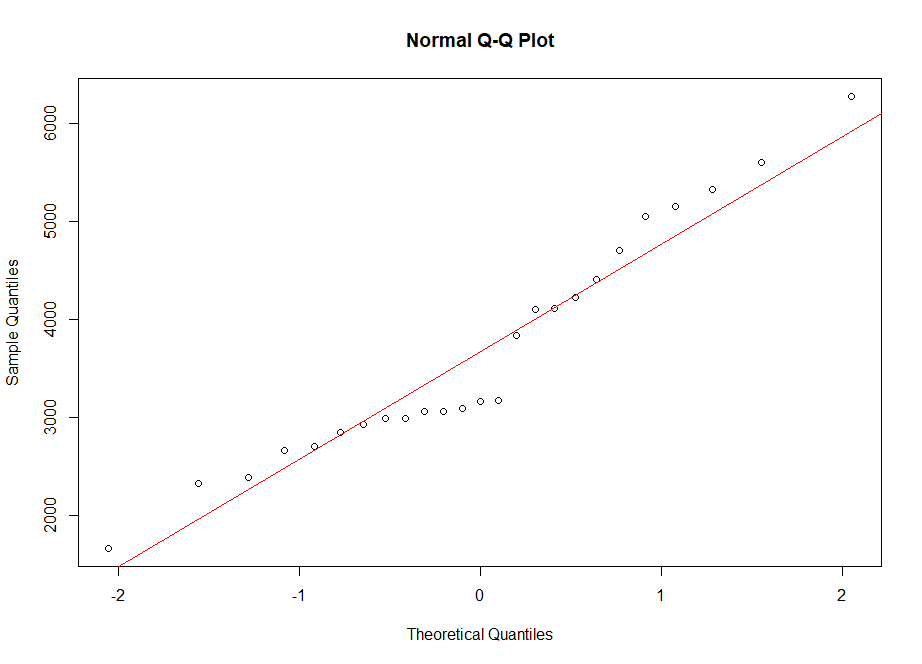


5.



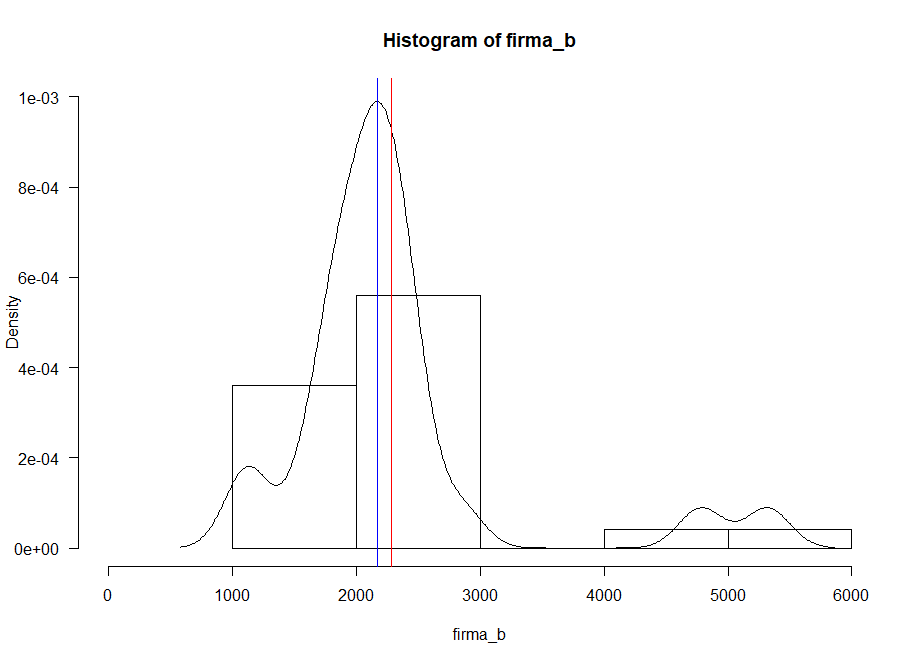
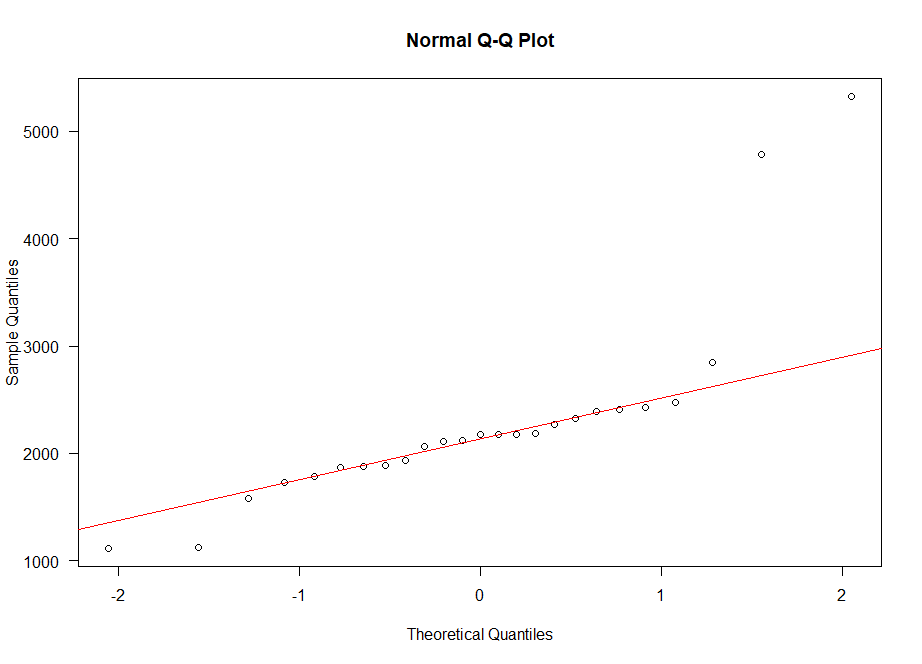
Aus dem Scatterplot lässt sich erkennen, dass die Gehälter in der Firma1 höher sind als die Gehälter in der Firma 2.

Explorative Datendarstellung der Firma 1:



Die Darstellungen zeigen rechts schiefe Verteilung der Daten. Die Kombination aus dem Histogramm und Dichtefunktion (schwarze Dichtelinie) veranschaulicht, dass die Daten bimodal sind. QQ-Plot zeigt links unten einen leichten Rand. Um genauer zu sein, es handelt sich hier um rechts Schiefe-Verteilung. Daher macht die Lageschätzung keinen Sinn. Die Daten sind annährend symmetrisch, daher kann t-test verwendet werden.

Explorative Datendarstellung der Firma 2:

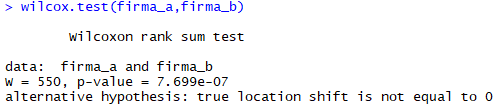


Die Darstellungen zeigen rechts schiefe Verteilung der Daten. Die Kombination aus dem Histogramm und Dichtefunktion (schwarze Dichtelinie) veranschaulicht, dass die Daten bimodal sind. Das QQ-Plot zeigt sowohl links unten als auch rechts oben schwere Ränder. Daher sind die Daten nicht normalverteilt. Es macht Sinn, dass die Daten nicht normalverteilt sind, weil es sich um Gehälter handelt.

Da die Daten der Firma 2 nicht parametrisch sind, wird ein Test für 2 ungepaarte Proben, Wilcoxon Mann Whitney Test, durchgeführt.

H0: es gibt keinen Unterschied zwischen den Gehälter µ1 = µ2

HA: Das Gehalt-Firma 1 ist signifikant größer als das Gehalt-Firma 2 µ1 > µ2



W-Wert von 550 ist die Test Statistik. Der p-Wert ist 7.699e-07, daher kann H0 auf allen Niveaus verworfen werden. Das bedeutet dass das Gehalt der Firma 1 ist signifikant größer als das Gehalt der Firma 2.