Übungen zu Data Analytics II

Prof. Dr. Wolfgang Bischof

TH Augsburg

7. Oktober 2024

Prof. Dr. Wolfgang Bischof
TH Augsburg
Wirtschaftspsychologie

1. Übungsblatt zur Vorlesung „Data Analytics II“

Aufgabe 1.1 *(Wilcoxon-Vorzeichen-Rang-Test)*

Betrachten Sie die Situation aus Aufgabe 8.1 der Übungsblätter zu Data Analytics 1.

(i) Testen Sie mit einem Test Ihrer Wahl, ob die Differenzen mit der Normalverteilungsannahme verträglich sind.

(ii) Nehmen Sie nun an, dass die Differenzen nicht als Realisierungen einer Normalverteilung angesehen werden können und prüfen Sie grafisch, ob die Annahme einer symmetrischen Verteilung der Differenzen zum Median gerechtfertigt ist.

(iii) Prüfen Sie auf einem Signifkanznivau von 5%, ob ein positiver Effekt der Therapie nachgewiesen werden kann.

Lösung:

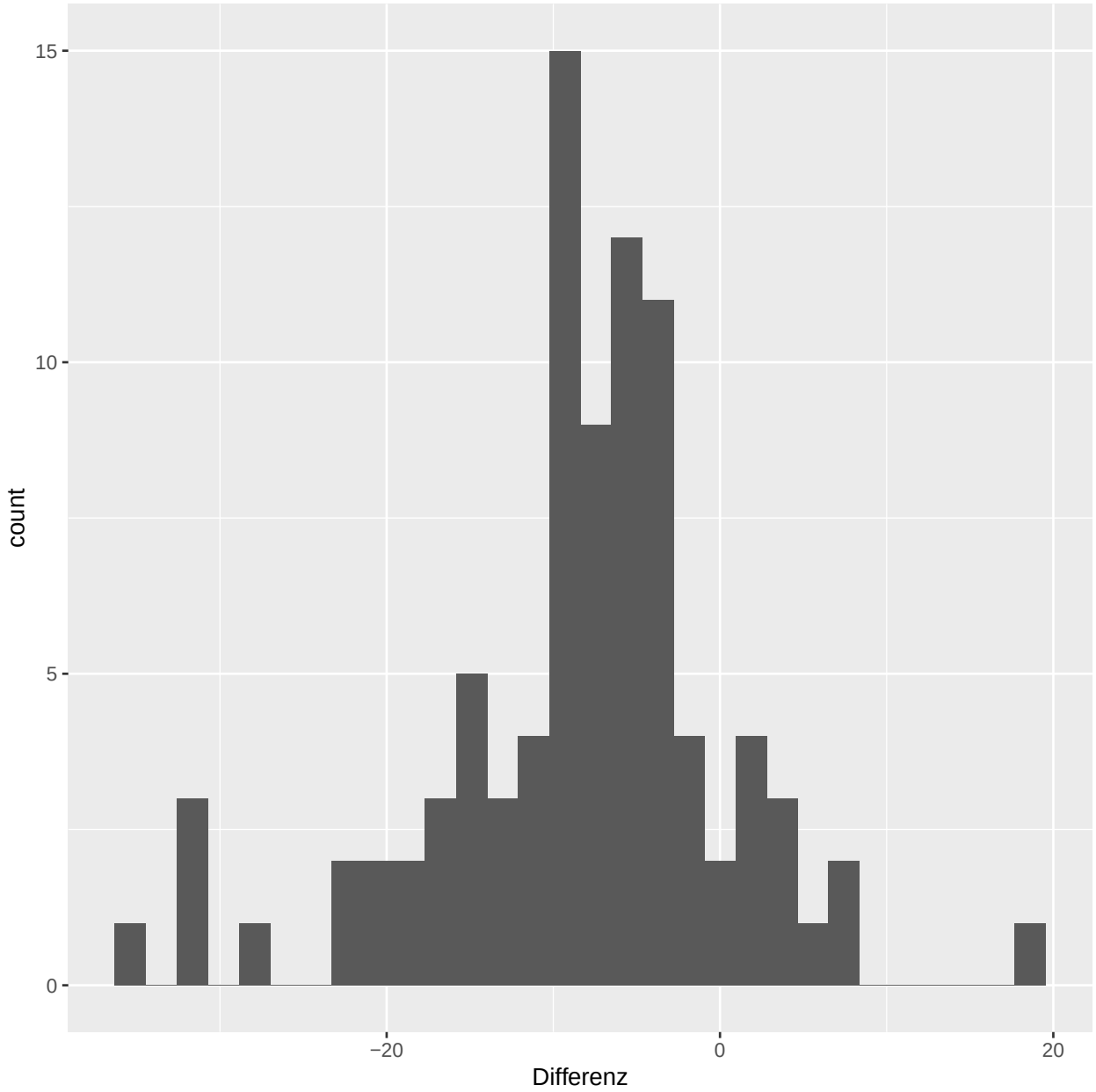
```
#
#
## Aufgabe 1.1
## ===========
library(haven)
BDI <- read_sav("BDI.sav")
#
# (i)
Differenz <- BDI$BDI_nach_Therapie - BDI$BDI_vor_Therapie
shapiro.test( Differenz )

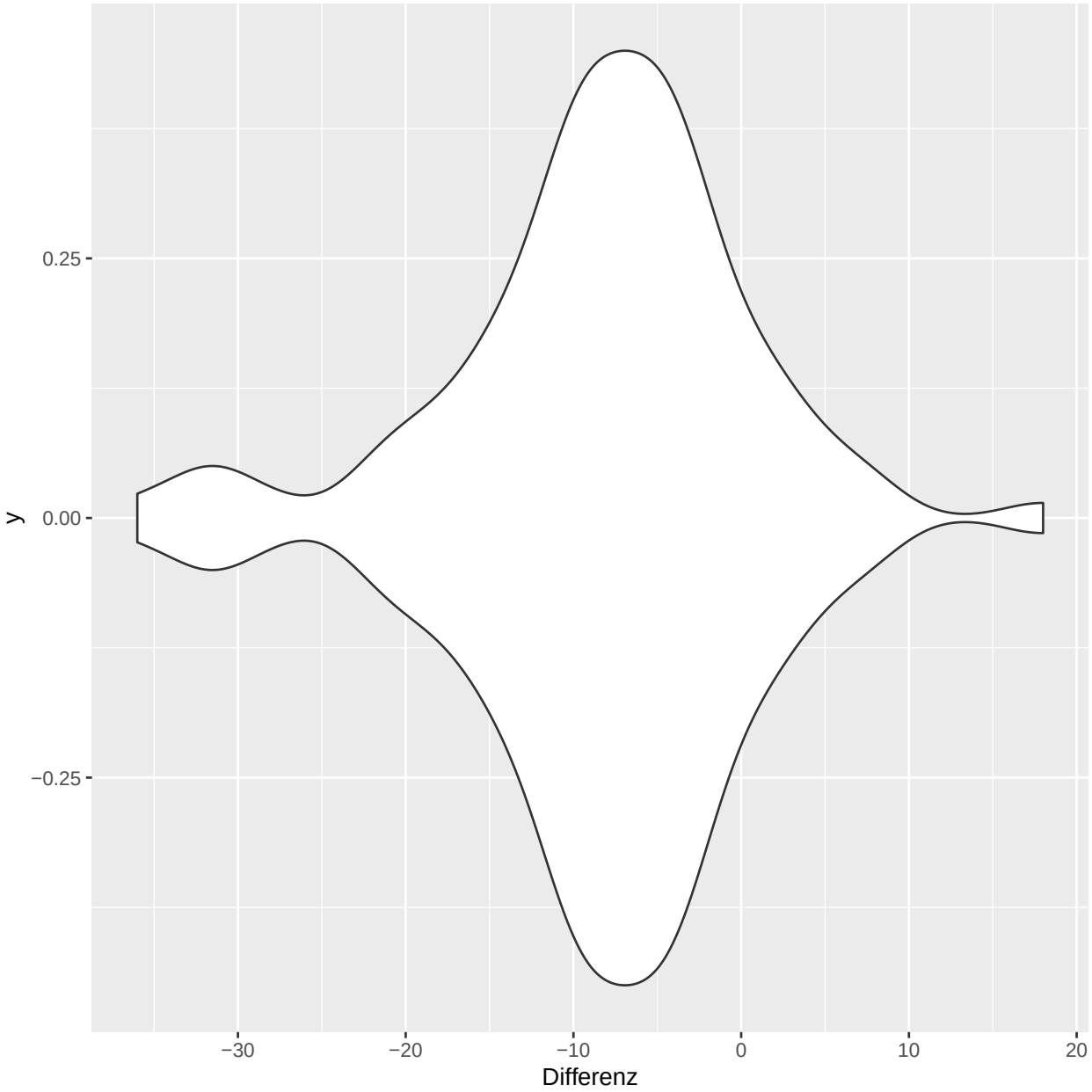
##
##  Shapiro-Wilk normality test
##
## data:  Differenz
## W = 0.95, p-value = 0.002

# Die Werte sind auffallend unverträglich mit der
# Normalverteilungsannahme
# (ii)
# graphische Überprüfung, ob die Differenzen symmetrisch zum Median
# liegen
library(tidyverse)
ggplot(BDI) + geom_histogram( aes( x=Differenz ))
ggplot(BDI) + geom_violin( aes( x=Differenz, y=0 ))
# sieht beides ok aus
# (iii)
wilcox.test( BDI$BDI_nach_Therapie, BDI$BDI_vor_Therapie,
             alternative = "less", paired = TRUE )

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  BDI$BDI_nach_Therapie and BDI$BDI_vor_Therapie
## V = 257, p-value = 7e-13
## alternative hypothesis: true location shift is less than 0

# auch ohne Normalverteilungsannahme kann die positive Wirkung
# hochsignifikant nachgewiesen werden
#
```


15
10
5
0
count
−20
0
20
Differenz



Aufgabe 1.2 *(Wilcoxon-Mann-Whitney-Rangsummentest)*
Sven Red-Selig aus der fünften Klasse behauptet, dass die Jungs seiner Klasse viel schneller laufen als die Mädels. Da ihm die Mädchen widersprechen, bittet Sven seinen Mathelehrer Gerd Genau auf einem Signifikanzniveau von 5% nachzuweisen, dass seine Behauptung stimmt. Versuchen Sie, Svens Behauptung aufgrund folgender Daten zu belegen, in dem Sie annehmen, dass die Zeiten normalverteilt sind.

Wie gehen sie vor, wenn Sie lediglich wissen, dass die Zeiten kontinuierlich verteilt sind?

Jungen	8,4	8,1	10,0	9,0	8,9	8,5	8,6	9,4	9,3	8,7
Mädchen	10,2	9,7	8,2	9,5	10,1	10,8	8,0	8,8	9,6	9,8

Prüfen Sie, ob die Normalverteilungsannahme auf einem Signifikanzniveau von 10% abgelehnt werden kann. Verwenden Sie dazu drei verschiedene Tests auf Normalverteilung.

Lösung:

```
## Aufgabe 3.3
##
## Approx. t-Test
##
jz <- c(8.4, 8.1, 10.0,  9.0, 8.9, 8.5, 8.6, 9.4,  9.3, 8.7)
mz <- c(10.2, 9.7, 8.2, 9.5, 10.1, 10.8, 8.0, 8.8, 9.6, 9.8)
t.test( jz,mz, alternativ="less" )
```

```
##
##  Welch Two Sample t-test
##
## data:  jz and mz
## t = -1.7479, df = 15.157, p-value = 0.05035
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.001319293
## sample estimates:
## mean of x mean of y
##      8.89      9.47

##
## Wilcoxon-Rangsummentest
##
wilcox.test( jz,mz, alternativ="less" )

##
##  Wilcoxon rank sum test
##
## data:  jz and mz
## W = 28, p-value = 0.05256
## alternative hypothesis: true location shift is less than 0

##
## Tests auf Normalverteilung
##
#
# 1. Variante: Normalscorestest
#
library("ppcc")
jz <- c(8.4, 8.1, 10.0,  9.0, 8.9, 8.5, 8.6, 9.4,  9.3, 8.7)
mz <- c(10.2, 9.7, 8.2, 9.5, 10.1, 10.8, 8.0, 8.8, 9.6, 9.8)
set.seed(1907)
ppccTest(jz)

##
##  Probability Plot Correlation Coefficient Test
##
## data:  jz
## ppcc = 0.98104, n = 10, p-value = 0.7602
## alternative hypothesis: jz differs from a Normal distribution

ppccTest(mz)

##
##  Probability Plot Correlation Coefficient Test
##
## data:  mz
## ppcc = 0.97227, n = 10, p-value = 0.5452
## alternative hypothesis: mz differs from a Normal distribution

# Achtung: Die p-Werte entstehen durch Simulation.
# Um reproduzierbare Ergebnisse zu erhalten, muss man den "seed" des
# Zufallsgenerators setzen
#
#
# 2. Variante: Lilliefors-Test
library("nortest")
lillie.test(jz)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  jz
## D = 0.13311, p-value = 0.884

lillie.test(mz)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mz
## D = 0.21347, p-value = 0.2203

##
## 3. Variante: Shapiro-Wilk
##
shapiro.test(jz)

##
##  Shapiro-Wilk normality test
##
## data:  jz
## W = 0.96644, p-value = 0.856

shapiro.test(mz)

##
##  Shapiro-Wilk normality test
##
## data:  mz
## W = 0.94018, p-value = 0.555

#
# Vergleich der p-Werte
#
# Die p-Werte für die Jungenzeiten schwanken zwischen 76% und 88%
# für die Mädchenzeiten zwischen 22% und 56%.
# Die Normalverteilungsannahme kann also mit keiner Variante
# abgelehnt werden.
```

Aufgabe 1.3 *(Kruskall-Wallis-Test)*

(i) Laden Sie das Paket `DA.students` und betrachten Sie die Datentabelle `dat.studenten`. Weisen Sie signifikant nach, dass die Mathenote abhängig von der Zufriedenheit mit der Mathenote ist. Bestimmen Sie den genauen p-Wert des Testproblems.

Lösung:

```
library(tidyverse)
library(DA.students)

summary(dat.studenten)

##      Alter          Groesse       Geschlecht      AlterV          AlterM
```

```
##  Min.   :16.0   Min.   :150   weiblich:743   Min.   :38.0   Min.   :37.0
##  1st Qu.:20.0   1st Qu.:166   männlich:511   1st Qu.:50.0   1st Qu.:48.0
##  Median :21.0   Median :173                  Median :54.0   Median :51.0
##  Mean   :21.8   Mean   :173                  Mean   :54.4   Mean   :51.6
##  3rd Qu.:23.0   3rd Qu.:180                  3rd Qu.:58.0   3rd Qu.:55.0
##  Max.   :37.0   Max.   :199                  Max.   :87.0   Max.   :70.0
##  NA's   :5      NA's   :1                    NA's   :4      NA's   :2
##    AnzSchuhe       NoteMathe             MatheZufr     Wohnform
##  Min.   :  1.0   Min.   :1.00   sehr zufrieden:171   Length:1254
##  1st Qu.:  8.0   1st Qu.:2.70   zufrieden     :211   Class :character
##  Median : 15.0   Median :3.30   geht so       :249   Mode  :character
##  Mean   : 20.4   Mean   :3.32   unzufrieden   :345
##  3rd Qu.: 27.0   3rd Qu.:4.00   NA's          :278
##  Max.   :275.0   Max.   :5.00
##  NA's   :1       NA's   :296
```

```
# es gibt einige NAs; verwende nur die Zeilen, bei denen sowohl die
# Mathe-Note als auch die Zufriedenheit mit der Mathe-Note vorhanden sind.
dat.studenten <- dat.studenten %>%
  filter( !is.na(NoteMathe) & !is.na(MatheZufr))

kruskal.test(dat.studenten$NoteMathe, dat.studenten$MatheZufr)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dat.studenten$NoteMathe and dat.studenten$MatheZufr
## Kruskal-Wallis chi-squared = 495, df = 3, p-value <2e-16
```

```
# Manuelle Bestimmung des p-Werts
pchisq(495.28, 3, lower.tail = FALSE)
```

```
## [1] 5.03e-107
```

(ii) Betrachten Sie die Datentabelle `mpg`. Können Sie auf einem Signifikanzniveau vo 5% nachweisen, dass der Hubraum (`displ`) abhängig vom Hersteller ist. Ist das Baujahr (`year`) signifikant abhängig vom Hersteller?

Lösung:

```
library(tidyverse)

kruskal.test(mpg$displ, mpg$manufacturer)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg$displ and mpg$manufacturer
## Kruskal-Wallis chi-squared = 156, df = 14, p-value <2e-16
```

```
kruskal.test(mpg$year, mpg$manufacturer)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg$year and mpg$manufacturer
## Kruskal-Wallis chi-squared = 8.2, df = 14, p-value = 0.9
```

Aufgabe 1.4 *(Chiquadrat-Anpassungstest zur Prüfung von Klassenwsk.en, Ross2006: Bsp. 11.2a, S.439f)*

Es wird vermutet, dass sich berühmte Leute auf ihren Geburtstag freuen (beispielsweise, weil sie da besondere Aufmerksamkeit erhalten). Folglich sind sie vor ihrem Geburtstag in guter mentaler Verfassung, was die Wahrscheinlichkeit dafür senkt, dass sie in den drei Monaten vor Ihrem Geburtstag sterben. Um diese Vermutung zu testen, wurde ein Stichprobe von 1251 Amerikaner:innen zufällig aus dem *Who Was Who in America* ausgewählt und die Monate vor oder nach ihrem Geburtstag gezählt:

Monate, die der Todestag vor (-) oder nach (+) dem Geburtstag liegt	Anzahl der Beobachtungen
-6,-5,-4	277
-3,-2,-1	283
0, 1, 2	358
3, 4, 5	333

(i) Prüfen Sie die Vermutung mit dem Chiquadrat-Anpassungstest auf einem Signifikanzniveau von 1% und berechnen Sie den p-Wert durch die Approximation durch die Chiquadrat-Verteilung.

(ii) Berechnen Sie den p-Wert nun durch 10000-fache Simulation von 1251 Todestagen.

(iii) Wie groß ist die Effektstärke?

(iv) Wie groß ist Teststärke bei der eben berechneten Effektstärke und einem Signifikanzniveau von 5%?

(v) Welche Stichprobengröße hätte ausgereicht, wenn die Effektstärke wie in (iii) berechnet ist, das Signifikanzniveau 5% und die Teststärke 75% beträgt?

Lösung:

```
  options("width"=60)
nb_obs <- c(277,283,358,333)
# (i)
chisq.test( nb_obs )

##
##  Chi-squared test for given probabilities
##
## data:  nb_obs
## X-squared = 14.775, df = 3, p-value = 0.00202

# Da der p-Wert unter 1% liegt, kann die Nulhypothese "Geburtstag hat keinen
# Einfluss" signifikant abgelehnt werden.
#
# (ii) alternativ mit Simulation
chisq.test( nb_obs, simulate.p.value = TRUE, B = 10000 )

##
##  Chi-squared test for given probabilities with
##  simulated p-value (based on 10000 replicates)
##
## data:  nb_obs
## X-squared = 14.775, df = NA, p-value = 0.0022

# nahezu gleiches Ergebnis
#
# (iii)
library(pwr)
(w = ES.w1( rep(1/4,4), nb_obs/1251 ) )
```

```
## [1] 0.1086748

# nur ein schwacher Effekt
# (iv)
pwr.chisq.test( w, N=1251, df=3, sig.level = 0.05)

##
##      Chi squared power calculation
##
##               w = 0.1086748
##               N = 1251
##              df = 3
##       sig.level = 0.05
##           power = 0.912661
##
## NOTE: N is the number of observations

# aber dank 1251 Beobachtungen eine gute Teststärke
# (v)
pwr.chisq.test( w, df=3, sig.level = 0.05, power=0.75)

##
##      Chi squared power calculation
##
##               w = 0.1086748
##               N = 826.8214
##              df = 3
##       sig.level = 0.05
##           power = 0.75
##
## NOTE: N is the number of observations

# Es hätten 827 Personen gereicht.
```

Bemerkung: Bedenken Sie, dass die Nullhypothese unbedingt *vor* der Analyse aufgestellt werden muss. Wenn Sie die Daten zur Formulieren der Nullhypothese verwenden, müssen sie eine zweite (Test-)Stichprobe verwenden, um H_0 zu prüfen.