Prof. Dr. Wolfgang Bischof
TH Augsburg
Wirtschaftspsychologie

2. Übungsblatt zur Vorlesung „Data Analytics II“

Aufgabe 2.1 *(Chiquadrat-Anpassungstest zur Prüfung einer Verteilungsannahme I)*

Sie nehmen an, dass die Anzahl der Menschen, die täglich Ihrer Beratungsstelle aufsuchen, Poisson-verteilt ist mit Erwartungswert 4.2. Prüfen Sie diese Hypothese anhand folgender Häufigkeitstabelle, die sie über die letzten 150 Tage erfasst haben:

Anzahl Besucher	0	1	2	3	4	5	6	7	8	9	10	≥ 11
Anzahl der Tage	0	5	22	23	32	22	19	13	6	4	4	0

Prüfen Sie zusätzlich, ob die täglichen Besucheranzahlen mit einer Poisson-Verteilung mit unbekannter Ankunftsrate verträglich sind. Der Maximum-Likelihood-Schätzer für λ entspricht hier einfach dem arithmetischen Mittel der Ankünfte und beträgt 4,57.

Lösung:

```
# zunächst Wsk.ten ausrechnen
i  <- c(0:10)
p  <- dpois(i,4.2)
p  <- c( p , 1-ppois(10,4.2))
chisq.test(c(0,5,22,23,32,22,19,13,6,4,4,0), p=p)

##
##  Chi-squared test for given probabilities
##
## data:  c(0, 5, 22, 23, 32, 22, 19, 13, 6, 4, 4, 0)
## X-squared = 16.56, df = 11, p-value = 0.1216

# with simulation
chisq.test(c(0,5,22,23,32,22,19,13,6,4,4,0), p=p,
           simulate.p.value = TRUE, B = 20000)

##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 20000 replicates)
##
## data:  c(0, 5, 22, 23, 32, 22, 19, 13, 6, 4, 4, 0)
## X-squared = 16.56, df = NA, p-value = 0.1263

# Der p-Wert liegt jeweils über 10%. Die Nullhypothese kann auf einem
# üblichen Niveau von 5% nicht abgelehnt werden.
#
# Variante mit unbekannter Ankunftsrate
lambda_est <- 4.57
p  <- dpois(i,lambda_est)
p  <- c( p , 1-ppois(10,lambda_est))
result <- chisq.test(c(0,5,22,23,32,22,19,13,6,4,4,0), p=p)
# falscher p-Wert, da lambda aus den Daten geschätzt wurde
# korrekter (approx.) p-Wert
1-pchisq(result$statistic, result$parameter-1)

## X-squared
## 0.4474673

# Der p-Wert liegt über 44%. Die Nullhypothese kann ein einem
# üblichen Niveau von 5% nicht abgelehnt werden.
```

Aufgabe 2.2 *(Chiquadrat-Anpassungstest zur Prüfung einer Verteilungsannahme II)*

150 Studierende schrieben die Mathe-Klausur mit. Die Gesamtpunktzahl lag bei 8 Studierenden unter 20; 18 erreichten zwischen 20 und unter 25 Punkten; 49 lagen zwischen 25 und unter 30; 52 waren zwischen 30 und unter 35; 22 waren zwischen 35 und unter 40; und einer erreichte mehr als 40 Punkte.

(i) Prüfen Sie, ob die Daten mit einer $N_{30;25}$-Verteilung verträglich sind.

(ii) Prüfen Sie nun die Hypothese, dass die Punkte überhaupt mit einer Normalverteilung verträglich sind. Die aus der Häufigkeitstabelle berechneten Maximum-Likelihood-Schätzer betragen 29,67 für μ und 5,36 für σ.

Lösung:

```
nb_obs <- c(8,18,49,52,22,1)
# (i)
chisq.test( nb_obs,
            p=c( pnorm(20,30,5), pnorm(25,30,5)-pnorm(20,30,5),
                 pnorm(30,30,5)-pnorm(25,30,5), pnorm(35,30,5)-pnorm(30,30,5),
                 pnorm(40,30,5)-pnorm(35,30,5), 1-pnorm(40,30,5)))


##
##  Chi-squared test for given probabilities
##
## data:  nb_obs
## X-squared = 8.3867, df = 5, p-value = 0.1362

# p-Wert über 10%. Auf einem Nivau von 5% sind die Daten mit einer N_(30,25)-Vtlg.
# verträglich. Auf einem Niveau von 20% nicht.
# (ii)
result <- chisq.test( nb_obs, p=c(pnorm(20,29.67,5.36),
                                  pnorm(25,29.67,5.36)-pnorm(20,29.67,5.36),
                                  pnorm(30,29.67,5.36)-pnorm(25,29.67,5.36),
                                  pnorm(35,29.67,5.36)-pnorm(30,29.67,5.36),
                                  pnorm(40,29.67,5.36)-pnorm(35,29.67,5.36),
                                  1-pnorm(40,29.67,5.36)))
# Achtung: der p-Wert aus dem chisq.test-Befehl ist falsch,
# weil wir Erwartungswert und Varianz geschätzt haben
# Der (approximativ) richtige p-Wert lautet
1-pchisq(result$statistic, 3)

## X-squared
## 0.1348062

# Wieder liegt der p-Wert über 10%. Es gelten die gleichen Schlussfolgerungen wie
# oben
```

Aufgabe 2.3 *(Chiquadrat-Unabhängigkeitstest)*

In einer Schulklasse wurden die Schüler nach ihrem Lieblingsfach befragt. Dabei ergab sich folgendes Ergebnis:

	Jungen	Mädchen
Deutsch	1	5
Mathematik	4	1
Kunsterziehung	2	5
Sport	18	16

Es soll nun untersucht werden, auf welchem Signifikanzniveau die These, dass die Lieblingsfächer unabhängig vom Geschlecht gewählt werden, abgelehnt werden kann.

(i) Berechnen Sie den approximierten p-Wert des χ^2-Unabhängigkeitstests.

(ii) Prüfen Sie die Approximation durch Simulation.

(iii) Berechnen und simulieren Sie den p-Wert des exakten Tests von Fisher.

Lösung:

```
##
## Chi-Quadrat-Test auf stoch. Unabhängigkeit
##
lieblingsfach_geschlecht <- matrix( c(1,4,2,18,5,1,5,16),4,2 )
# (i)
chisq.test(lieblingsfach_geschlecht)

##
##  Pearson's Chi-squared test
##
## data:  lieblingsfach_geschlecht
## X-squared = 5.8017, df = 3, p-value = 0.1217

##
## (ii) Simulation, da die Zellhäufigkeiten relativ klein sind
chisq.test(lieblingsfach_geschlecht, simulate.p.value = TRUE)

##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  lieblingsfach_geschlecht
## X-squared = 5.8017, df = NA, p-value = 0.1209

# 1 Mio Simulationsdurchlaeufe
chisq.test(lieblingsfach_geschlecht, simulate.p.value = TRUE, B = 1000000)

##
##  Pearson's Chi-squared test with simulated p-value (based on 1e+06
##  replicates)
##
## data:  lieblingsfach_geschlecht
## X-squared = 5.8017, df = NA, p-value = 0.1312

##
# Zusammenfassung:
# X^2-Approx.:    p-Wert: 12,2%
# X^2-Unabh.test: p-Wert: 13,1%
##
#
# (iii) Fisher-Test
#
fisher.test(lieblingsfach_geschlecht)

##
##  Fisher's Exact Test for Count Data
##
## data:  lieblingsfach_geschlecht
## p-value = 0.1419
## alternative hypothesis: two.sided
```

```r
# nur zur Kontrolle: 1 Mio Simulationsdurchlaeufe
fisher.test(lieblingsfach_geschlecht, simulate.p.value = TRUE,
            B=1000000)


##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  lieblingsfach_geschlecht
## p-value = 0.142
## alternative hypothesis: two.sided


#
# Zusammenfassung:
# Fisher-Test: exakt: p-Wert: 14,2%
# simuliert: p-Wert: 14,2%
#
# Bemerkung: Die Differenz des p-Werts zwischen dem exakten Test von Fisher und
# der Chiquadrat-Approximation kommt daher, dass manchen Häufigkeitstabellen
# zwar eine kleinere Wsk. haben als die beobachtete (bei Fisher also zum
# p-Wert beitragen), aber eine kleinere Chiquadrat-Prüfgröße (beim Chi-Quadrat-
# Test also nicht zum p-Wert beitragen)
```

Aufgabe 2.4 *(Chiquadrat-Homogenitätstest)*

Am Abend der Bundestagswahl 2017 präsentierten die Fernsehanstalten ARD und ZDF um 18:00 eine letzte Prognose der Wahlergebnisse[1]. Wenn man annimmt, dass die ARD 1200 Menschen und das ZDF 1000 Menschen befragt hat, ergibt sich folgende 2×7-Feldertafel:

	CDU/CSU	SPD	Linke	Grüne	FDP	AfD	Andere	Sum.
ARD	390	240	108	114	126	162	60	1200
ZDF	335	210	90	90	100	130	45	1000
Sum.	725	450	198	204	226	292	105	2200

Weichen die Ergebnisse signifikant voneinander ab?

Lösung:

```r
##
## Chi-Quadrat-Homogenitätstest zum Vgl. zweier Multinomialvtlg.n
##
## p-Wert per Approximation
umfrage.tab.beob <- matrix( c(390,335,240,210,108,90,114,90,126,100,
                              162,130,60,45),nrow=2 )
chisq.test(umfrage.tab.beob)


##
##  Pearson's Chi-squared test
##
## data:  umfrage.tab.beob
## X-squared = 1.1004, df = 6, p-value = 0.9815


# keine signifikante Abweichung (p-Wert: 98%)
##
## p-Wert per Simulation ( 1 Mio. Simulationsdurchläufe )
chisq.test(umfrage.tab.beob, simulate.p.value = TRUE, B=1000000)
```

[1] Quelle: https://www.wahlrecht.de/news/2017/bundestagswahl-2017.html#prognosen -hochrechnungen

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 1e+06
##  replicates)
##
## data:  umfrage.tab.beob
## X-squared = 1.1004, df = NA, p-value = 0.9816

# analoges Ergebnis
##
## Appendix: p-Wert mit exaktem Test von Fisher (geht nur per
##           Simulation, da die Anzahlen zu groß sind)
fisher.test(umfrage.tab.beob, simulate.p.value = TRUE, B=1000000)

##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  umfrage.tab.beob
## p-value = 0.9822
## alternative hypothesis: two.sided

## gute Übereinstimmung, p-Wert wieder 98%
```

Aufgabe 2.5 *(Korrelationstests)*

In der Datei „WeiterbildungOriginalitaet.xlsx“ finden Sie die Umfrageergebnisse von 102 Personen zu den zwei Items:

- Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.
- Mir fällt es leicht originelle Lösungsansätze zu finden.

Die Antworten reichen von 0 („trifft überhaupt nicht zu“) bis 5 („trifft voll und ganz zu“).

(i) Berechnen Sie die Stichprobenkorrelation nach Pearson und die Rangkorrelation nach Spearman.

(ii) Prüfen Sie, ob die Ergebnisse zu jeder Frage mit einer Normalverteilungsannahme verträglich sind. Prüfen Sie anschließend, ob die Ergebnisse zu beiden Fragen mit einer bivariaten Normalverteilungsannahme verträglich sind.

(iii) Prüfen Sie mit den Korrelationstests nach Pearson, Spearman und Kendall, ob eine Korrelation auf einem Signifikanzniveau von 1% nachgewiesen werden kann.

(iv) Führen Sie den Korrelationstest nach Pearson mit den Rängen der Umfrageergebnisse durch und vergleichen Sie das Ergebnis mit dem in (iii) durchgeführten Korrelationstest nach Spearman.

(v) Wie groß ist die Effektstärke beim Korrelationstest nach Pearson? Wie groß ist die Teststärke bei dieser Effektstärke und einem Signifikanzniveau von 5%?

(vi) Wie viele Personen hätten befragt werden müssen, um bei einem Signifikanzniveau von 5% eine Teststärke von 80% zu erreichen?

Lösung:

```
library(readxl)
umfrageErgebnisse <- read_xlsx("WeiterbildungOriginalitaet.xlsx",1)
# (i)
cor(pull(umfrageErgebnisse,2), pull(umfrageErgebnisse,3), method = "pearson")
```

```
## [1] 0.2911803

cor(pull(umfrageErgebnisse,2), pull(umfrageErgebnisse,3), method = "spearman")

## [1] 0.284269

# (ii)
shapiro.test(pull(umfrageErgebnisse,2))

##
##  Shapiro-Wilk normality test
##
## data:  pull(umfrageErgebnisse, 2)
## W = 0.75745, p-value = 1.097e-11

shapiro.test(pull(umfrageErgebnisse,3))

##
##  Shapiro-Wilk normality test
##
## data:  pull(umfrageErgebnisse, 3)
## W = 0.90254, p-value = 1.524e-06

# Die Daten stimmen sicher nicht aus einer normal-verteilten Grundgesamtheit
library(MVN)
mvn(umfrageErgebnisse[,2:3])

## $multivariateNormality
##            Test       HZ    p value MVN
## 1 Henze-Zirkler 1.180801 0.01325946  NO
##
## $univariateNormality
##               Test
## 1 Anderson-Darling
## 2 Anderson-Darling
##                                                                  Variable
## 1 Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.
## 2          Mir fällt es leicht originelle Lösungsansätze zu finden.
##   Statistic   p value Normality
## 1    8.0042  <0.001      NO
## 2    4.1736  <0.001      NO
##
## $Descriptives
##                                                                    n
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern. 102
## Mir fällt es leicht originelle Lösungsansätze zu finden.                   102
##                                                                     Mean
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern. 4.049020
## Mir fällt es leicht originelle Lösungsansätze zu finden.                   3.431373
##                                                                   Std.Dev
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern. 0.9374227
## Mir fällt es leicht originelle Lösungsansätze zu finden.                   1.0099486
##                                                                   Median
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.      4
## Mir fällt es leicht originelle Lösungsansätze zu finden.                        3
##                                                                      Min
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.   0
## Mir fällt es leicht originelle Lösungsansätze zu finden.                     1
```

```
##                                                                    Max
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.   5
## Mir fällt es leicht originelle Lösungsansätze zu finden.                     5
##                                                                   25th
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.    4
## Mir fällt es leicht originelle Lösungsansätze zu finden.                     3
##                                                                   75th
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.    5
## Mir fällt es leicht originelle Lösungsansätze zu finden.                     4
##                                                                       Skew
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern. -1.7383498
## Mir fällt es leicht originelle Lösungsansätze zu finden.                   -0.1285478
##                                                                     Kurtosis
## Ich bilde mich gerne weiter und versuche mein Horizont stets zu erweitern.  5.0282733
## Mir fällt es leicht originelle Lösungsansätze zu finden.                   -0.7124599

# Der bivariaten Normalverteilungstest besitzt einen p-Wert von 1,3%
# lehnt also auf einem Niveau von 1% nicht ab!
#
# (iii)
cor.test(pull(umfrageErgebnisse,2), pull(umfrageErgebnisse,3),
         method = "pearson")

##
##  Pearson's product-moment correlation
##
## data:  pull(umfrageErgebnisse, 2) and pull(umfrageErgebnisse, 3)
## t = 3.0437, df = 100, p-value = 0.002986
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1025103 0.4596278
## sample estimates:
##       cor
## 0.2911803

cor.test(pull(umfrageErgebnisse,2), pull(umfrageErgebnisse,3),
         method = "spearman")

##
##  Spearman's rank correlation rho
##
## data:  pull(umfrageErgebnisse, 2) and pull(umfrageErgebnisse, 3)
## S = 126578, p-value = 0.003785
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.284269

cor.test(pull(umfrageErgebnisse,2), pull(umfrageErgebnisse,3),
         method = "kendall")

##
##  Kendall's rank correlation tau
##
## data:  pull(umfrageErgebnisse, 2) and pull(umfrageErgebnisse, 3)
## z = 2.8955, p-value = 0.003785
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2492216
```

```r
# alle drei Tests weisen eine Korrelation auf einem Niveau von 1% nach.
# Der (nicht zulässige) Test nach Pearson hat den kleinsten
# p-Wert.
#
# (iv)
cor.test(rank(pull(umfrageErgebnisse,2)), rank(pull(umfrageErgebnisse,3)),
         method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank(pull(umfrageErgebnisse, 2)) and rank(pull(umfrageErgebnisse, 3))
## t = 2.965, df = 100, p-value = 0.003785
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09504868 0.45366390
## sample estimates:
##      cor
## 0.284269
```

```r
# Korrelation stimmt mit der Korrelation aus dem Korrelationstest
# nach Spearman überein (wie zu erwarten war), p-Wert stimmt auch überein
# KI darf nicht verwendet werden (da auch die Ränge sicher nicht
# normalverteilt sind)
#
# (v)
# Die Effektstärke entspricht der Korrelation, also r=0,29
# Dies entspricht nach Cohen einem moderate Effekt.
library(pwr)
pwr.r.test(n=102, r=0.2911803, sig.level = 0.05 )
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 102
##               r = 0.2911803
##       sig.level = 0.05
##           power = 0.8500312
##     alternative = two.sided
```

```r
# Die Teststärke beträgt 85%.
#
# (vi)
pwr.r.test(r=0.2911803, sig.level = 0.05, power = 0.8 )
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 89.43686
##               r = 0.2911803
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```r
# Es hätten 90 Leute ausgereicht.
```