Data Analytics 2 für Wirtschaftspsychologie

Nichtparametrische Hypothesentests

Prof. Dr. Wolfgang Bischof

Augsburg

27. Januar 2026

Inhaltsverzeichnis

VIII Nichtparametrische Tests 2
1 Tests auf Normalverteilung . . . 2
2 Q-Q-Plots . . . 2
3 Ränge . . . 4
4 Wilcoxon-Vorzeichen-Rang-Test für gepaarte Beobachtungen . . . 4
5 Wilcoxon-Mann-Whitney-Zweistichproben-Rangsummentest für unabhängige Stichproben . . . 6
6 Kruskall-Wallis-Test . . . 7
7 Chi-Quadrat-Tests . . . 9
8 Effekt- und Teststärke . . . 9
9 Multinomialverteilung . . . 10
10 χ^2-Anpassungstest zum Prüfen der Klassenwahrscheinlichkeiten . . . 11
11 χ^2-Anpassungstest zum Prüfen einer Verteilungsannahme mit bekannten Parametern 13
12 χ^2-Anpassungstest zum Prüfen einer Verteilungsannahme mit unbekannten Parametern . . . 16
13 χ^2-Test auf stochastische Unabhängigkeit . . . 17
14 χ^2-Homogenitätstest . . . 20
15 Korrelationstests . . . 21

VIII Nichtparametrische Tests

1 Tests auf Normalverteilung

Es gibt unzählige Tests, um zu prüfen, ob vorhandenen metrische Daten mit einer Normalverteilungsannahme verträglich sind. Ohne genauer auf die jeweiligen Prüfgrößen einzugehen, sollen hier drei gängige Tests genannt werden.

- Shapiro-Wilk-Test (`shapiro.test`)
- Normalscores-Test (`ppccTest` aus dem Paket `ppcc`)
- Lilliefors-Test (`lillie.test` aus dem Paket `nortest`)

Prinzipiell gilt, dass die Normalverteilungsannahme um so signifikanter abgelehnt werden kann um so kleiner der p-Wert ist. Ein signifikanter Nachweis einer Normalverteilungsannahme ist prinzipiell nicht möglich. Ein hoher p-Wert bedeutet lediglich, dass die Daten gut mit einer Normalverteilungsannahme verträglich sind, sie könnten aber natürlich z.B. auch von einer t_{100}-Verteilung stammen. Beispiele hierzu machen wir in der Übung 1.1; Details zu den Tests finden Sie z.B. über die R-Hilfefunktion.

ABC
1

2 Q-Q-Plots

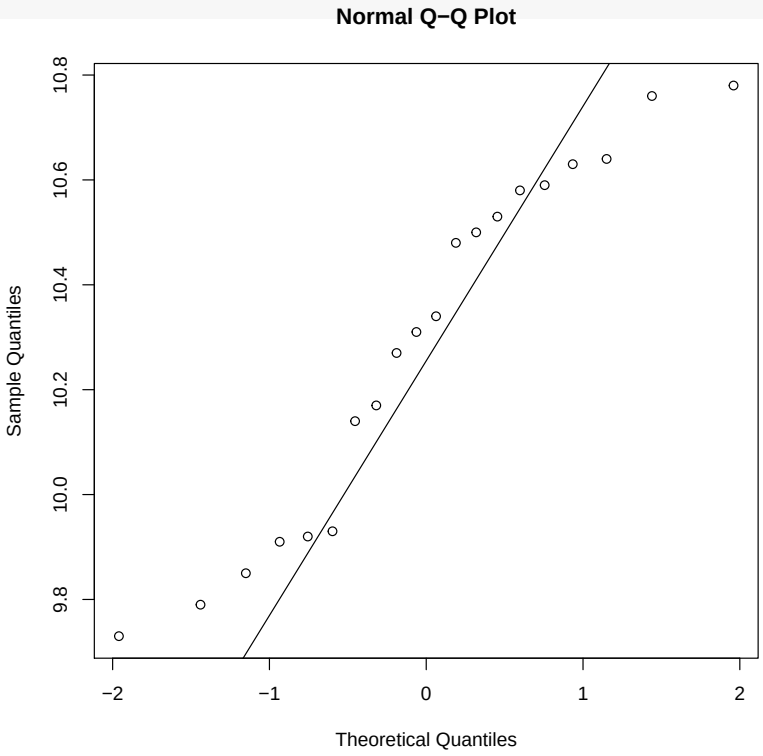
Ein Quantile-Quantile-Plot, kurz Q-Q-Plot, ist ein exploratives, grafisches Werkzeug, um zu prüfen, ob die Verteilung von Beobachtungswerten mit einer gegebenen Verteilung verträglich ist.
Dazu ordnet man die Beobachtungswerte $\{b_i\}_{i=1,\ldots,n}$ der Größe nach und vergleicht sie mit dem Erwartungswert oder Median a_i der $i.$-größten von n Beobachtungen der theoretischen Verteilung, in dem man die Paare (a_i, b_i) in ein Koordinatensystem aufträgt und prüft, ob die Punkte ungefähr auf der Winkelhalbierenden liegen.
Wenn man prüfen möchte, ob die Daten mit einer Normalverteilung verträglich sind, ohne die Parameter der Normalverteilung zu kennen, führt man einen Q-Q-Plot gegen die Standardnormalverteilung durch, in R mit dem Befehl `qqnorm`. Wegen der Eigenschaften der Normalverteilung müssen die Punkte auf einer Geraden (und nicht zwingend auf der Winkelhalbierenden) liegen, wenn die Daten normalverteilt sind.
Führt man nach dem `qqnorm`-Befehl den Befehl `qqline` aus, so wird eine Hilfslinie gezeichnet, die es erleichtert zu prüfen, ob alle Punkte auf einer Geraden liegen (Details dazu findet man in der Hilfefunktion von `qqline`.).
Zum Testen auf andere Verteilungen muss man sich die Erwartungswerte oder Mediane der Ordnungsstatistiken zunächst selbst berechnen bzw. simulieren und kann dann mit `qqplot` einen QQ-Plot zeichnen lassen.
Bequemer ist das Kommando `qqPlot` aus dem Paket `EnvStats`. Hier werden die Schätzwerte für die mittleren Ordnungsstatistiken für eine große Menge von Verteilungen vom Kommando berechnet und man muss nur noch angeben, gegen welche Verteilung man testen möchte.

Beispiel:
Beim Abpacken von Kartoffeln in Zehnkilosäcke kann das Normgewicht nicht genau eingehalten werden. Beim Wiegen von 20 Kartoffelsäcken ergaben sich folgende Werte (in kg):

9.92	10.64	10.59	9.79	10.53	10.14	10.78	10.63	9.73	10.48
10.76	10.17	9.91	10.58	10.31	9.85	10.27	9.93	10.5	10.34

Prüfen Sie, ob die Annahme, dass das Füllgewicht durch eine Normalverteilung adäquat beschrieben werden kann, mit den obigen Beobachtungswerten verträglich ist.
Der QQ-Plot gegen die Normalverteilung sieht so aus:

```
gew <- c(9.92, 10.64, 10.59,  9.79, 10.53, 10.14, 10.78, 10.63,
         9.73, 10.48, 10.76, 10.17,  9.91, 10.58, 10.31,  9.85,
         10.27, 9.93, 10.5, 10.34)
qqnorm(gew)
qqline(gew)
```

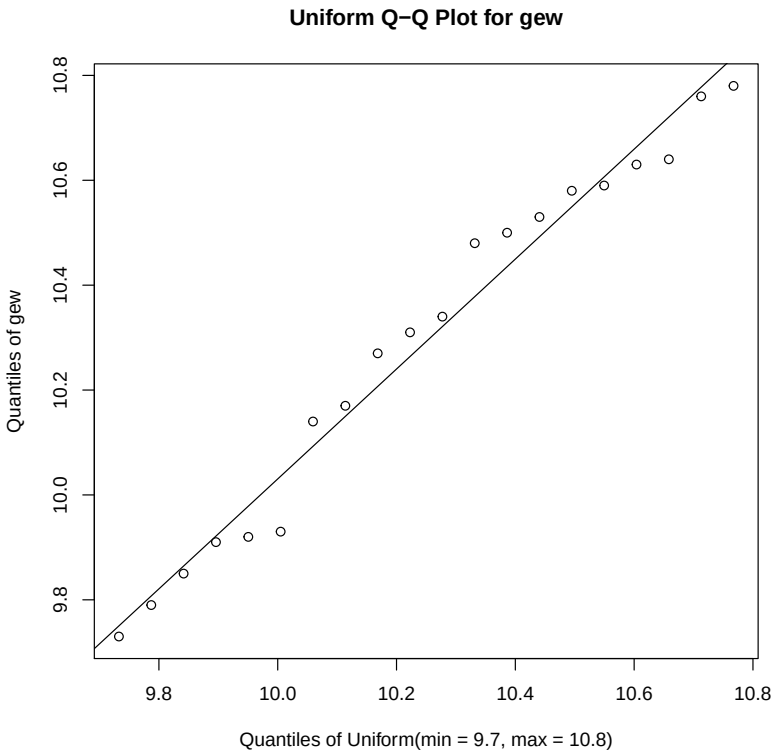


Man sieht, dass die Beobachtungswerte am linken und rechten Rand weniger stark streuen als man dies bei einer Normalverteilung erwarten würde. Der QQ-Plot widerspricht deutlich einer Normalverteilungsannahme.

Sehen wir uns zum Vergleich den QQ-Plot gegen die $U_{(9{,}7;10{,}8)}$-Verteilung an:

```
library(EnvStats)

gew <- c(9.92, 10.64, 10.59,  9.79, 10.53, 10.14, 10.78, 10.63,
         9.73, 10.48, 10.76, 10.17,  9.91, 10.58, 10.31,  9.85,
         10.27, 9.93, 10.5, 10.34)
qqPlot (gew, distribution="unif",
        param.list = list(min=9.7,max=10.8), add.line = TRUE )
```



Die Punkte weichen nicht stark von einer Geraden ab, der Gleichverteilungshypothese kann also nicht signifikant widersprochen werden. ABC 2

3 Ränge

Wir betrachten erneut die Zweistichprobenmodelle aus dem letzten Abschnitt, wobei wir nun die Normalverteilungsannahme fallen lassen. Für den Fall von gepaarten Beobachtungen setzen wir nun nur noch ein metrisches Skalenniveau und eine zum Ursprung symmetrische Verteilung der Differenzen voraus. Für den Fall von unabhängigen Stichproben reicht sogar ein ordinales Skalenniveau, d.h. die Werte müssen nur noch angeordnet werden können.

Da wir die Verteilung der Differenzen und Werte nun nicht mehr kennen, verwenden wir für die Prüfgrößen statt den Werten selbst deren Ränge. Wie in Data Analytics 1 werden bei Bindungen/Ties die arithmetischen Mittel der entsprechenden Ränge verwendet.

4 Wilcoxon-Vorzeichen-Rang-Test für gepaarte Beobachtungen

Voraussetzungen:
$X_{11}, ..., X_{1n}, X_{21}, ..., X_{2n}$ reellwertige Zufallsvariablen mit $D_1 := X_{11} - X_{21}, \ldots, D_n := X_{1n} - X_{2n}$ u.i. symmetrisch zum Median der Stichprobendifferenzen verteilt.

Testproblem:
$\text{med}(D) = 0$ vs. $\text{med}(D) \neq 0$ (bzw. $\text{med}(D) > 0$ oder $\text{med}(D) < 0$
Zur Entscheidung des Testproblem verwenden wir den „Wilcoxon-Vorzeichen-Rang-Test“. Zur Berechnung der Prüfgröße ordnen wir die Absolut-Beträge $|D_1|, \ldots, |D_n|$ der Differenzen der Größe nach und bestimmen die Ränge, in dem wir sie nach der absoluten Differenz sortieren beginnend mit der kleinsten. Jeder positiven Differenz geben wir ihren positiven Rang und jeder negativen Differenz das Negative des Rangs. Die zugeordnete Zahl nennen wir den „signierten Rang R_i“ der Differenz D_i. Ist eine Differenz 0, so erhält diese Differenz keinen signierten Rang.

Als Prüfgröße verwenden wir die Summe V aller positiven signierten Ränge. Den p-Wert des Testproblems erhalten wir in R mit dem Befehl `wilcox.test`.
Bemerkung: Wenn die Annahme der Symmetrie der Differenzen nicht gerechtfertigt ist, kann der sogenannte „Vorzeichentest“ durchgeführt werden. Er berücksichtigt nicht die Ränge, sondern zählt lediglich, wie oft eine positive Differenz auftrat. Man berücksichtigt dabei nur von 0 verschiedene Differenzen. Wenn es m solche Differenzen gibt, ist die Prüfgröße, also die Anzahl der positiven Differenzen $B_{m,0.5}$-verteilt, da jede Differenz zu 50% positiv ist.
Zur Illustration des Wilcoxon-Vorzeichen-Rang-Tests betrachten wir folgendes Beispiel.
Beispiel:

	1	2	3	4	5	6	7	8	9	10
Punkte nach Rep.	41	48	37	46	44	55	50	42	38	35
Punkte vor Rep.	34	42	36	36	39	48	38	44	32	35
Differenz (nach-vor)	7	6	1	10	5	7	12	-2	6	0
signierter Rang	+6.5	+4.5	+1	+8	+3	+6.5	+9	-2	+4.5	

Damit ergibt sich als Prüfgröße $v = 1 + 3 + 2 \cdot 4.5 + 2 \cdot 6.5 + 8 + 9 = 43$.

(a) Sind die Daten mit einer Normalverteilungsannahme für die Differenzen verträglich?

(b) Nehmen Sie nun unabhängig von (i) an, dass die Differenzen nicht normalverteilt sind. Kann dennoch ein signifikanter ($\alpha = 5\%$) positiver Effekt des Repetitoriums nachgewiesen werden?

(c) Nehmen Sie nun an, dass keine symmetrische Verteilung der Differenzen um den Median herum vorliegt. Kann dann immer noch ein signifikanter ($\alpha = 5\%$) positiver Effekt des Repetitoriums nachgewiesen werden?

Lösung:

```
# (i)
shapiro.test(Differenz)

##
##  Shapiro-Wilk normality test
##
## data:  Differenz
## W = 0.95, p-value = 0.7

# (ii)
wilcox.test( Punkte_nach_Rep, Punkte_vor_Rep, alternative = "greater",
             paired = TRUE )

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Punkte_nach_Rep and Punkte_vor_Rep
## V = 43, p-value = 0.009
## alternative hypothesis: true location shift is greater than 0

# Der Effekt ist immer noch sehr signifikant. Wie zu erwarten war ist der
# p-Wert größer als beim entsprechenden t-Test
# (iii)
# Bei neun Differenzen wurde acht mal eine positive Wsk. beobachtet
pbinom(7,9,0.5,lower.tail = FALSE)
```

```
## [1] 0.01953

# Der Effekt ist immer noch signifikant. Wie zu erwarten war, ist der
# p-Wert nochmal deutlich größer
#
```

5 Wilcoxon-Mann-Whitney-Zweistichproben-Rangsummentest für unabhängige Stichproben

Voraussetzungen:
Gegeben seien n stochastisch unabhängige Zufallsvariablen: $X_1, ..., X_n$, wovon die ersten k: $X_1, ..., X_k$ id. vtlt. gem. P sind, und die letzten l: $X_{k+1}, ..., X_{k+l=n}$ id. vtlt. gem. Q mit zugehörigen Rängen $R_1, ..., R_n$.
Es soll nun getestet werden, ob die beiden Verteilungen P und Q verschieden sind bzw., ob P typischerweise kleinere oder größere Ausprägungen liefert als Q. Es sind also folgenden Testprobleme interessant:

- $H_0: \ P = Q \ \text{ gg. } \ H_1: \ P \neq Q$ (falls Nachweis $P \neq Q$ gewünscht)
- $H_0: \ P = Q \ \text{ gg. } \ H_1: \ P \prec Q$ (falls Nachweis $P \prec Q$ gewünscht)
- $H_0: \ P = Q \ \text{ gg. } \ H_1: \ P \succ Q$ (falls Nachweis $P \succ Q$ gewünscht)

Im Folgenden suchen wir eine geeignete Prüfgröße für den Fall, dass über P, Q keine weiteren Kenntnisse vorliegen.

Definition:
Die ZV $W_P := R_1 + ... + R_k$ und $W_Q = R_{k+1} + ... + R_{k+l}$ werden als „Gruppenrangsummen“ bezeichnet.
W_P ist also die Summe der Plätze der Beobachtungen mit Verteilung P;
W_Q ist also die Summe der Plätze der Beobachtungen mit Verteilung Q.

Bemerkungen:

(a) Es gilt:
$$W_P + W_Q = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$
$\Rightarrow$ Es genügt $W := W_P$ zu betrachten.

(b) Im bindungsfreien Fall kann anstatt von W alternativ auch die sog. U-Statistik [U für „Uniform“] verwendet werden, die für jede Beobachtung aus der ersten Gruppe zählt, wie viele Beobachtungen aus der zweiten Gruppe einen Platz mit kleinerem Rang haben und dann über all diese Anzahlen summiert.

Es gilt: $W = U + \frac{k(k+1)}{2}$ mit $U_P := \sum_{i=1}^{k} \sum_{j=k+1}^{k+l} \mathbb{1}_{\{X_i > X_j\}}$.

(c) Im allgemeinen Fall verwenden wir den Zusammenhang aus dem bindungsfreien Fall, um U zu definieren:
$$U := W - \frac{k(k+1)}{2}.$$

Definition:
Ein Test $H_0 : P = Q$ gg. $H_1 : P \prec Q$ mit Teststatistik U bzw. W und Ablehnungsbereich

$$\{U < c\} = \left\{W < c + \frac{k(k+1)}{2}\right\}$$

heißt Mann-Whitney-U-Test oder Wilcoxon-Zweist.-Rangsummentest.
Analog für die anderen oben genannten Testprobleme.

Der Test wird in R mit dem Befehl `wilcox.test` durchgeführt.

Bemerkung: Als Effektstärke wird $U/(k \cdot l)$ verwendet, das ist beim Testproblem $P = Q$ vs. $P \succ Q$ der Anteil aller „in Richtung Alternative günstigen“ paarweisen Vergleiche, das sind beim Test aus der obigen Definition die Vergleiche, bei denen die Beobachtung aus der zweiten Gruppe kleiner ist. Diese Effektstärke wird auch als *common language*-Effektstärke (CL) bezeichnet, weil sie so einfach interpretierbar ist. Cohens Klassifikation der Effektstärken lässt sich auch auf die - von ihm selbst nicht behandelte - *common language*-Effektstärke übertragen. Man erhält folgende Interpretation:

CL=0.55	geringer/kleiner Effekt
CL=0.64	mittlerer Effekt
CL=0.71	starker/großer Effekt

Beispiel: siehe Übungsaufgabe 1.2. ABC 3

6 Kruskall-Wallis-Test

(vgl. Lehn/Wegmann, 5. Aufl., S. 178ff)

Beim Zweistichproben-Test von Wilcoxon-Mann-Whitney werden Messreihen von zwei Gruppen untersucht und geprüft, ob ihre Entstehung durch das gleiche Zufallsgesetz beschrieben werden kann. In den Anwendungen muss man jedoch häufig mehr als zwei Gruppen überprüfen. Wie in den Abschnitten zuvor nehmen wir an, dass für die Messreihen mindestens ein ordinales Skalenniveau vorliegt.
Es soll entschieden werden, ob aus den Daten auf Unterschiede zwischen den Gruppen geschlossen werden kann. Wir stellen uns vor, dass die Messreihe

$$x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2}; \ldots; x_{k1}, \ldots, x_{kn_k}$$

durch Realisierungen von unabhängigen Zufallsvariablen

$$X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2}; \ldots; X_{k1}, \ldots, X_{kn_k}$$

entstanden sind, wobei die Zufallsvariablen $X_{i1}, \ldots, X_{in_i}$ jeweils dieselbe Verteilungsfunktion F_i besitzen $(i = l, ..., k)$.

Als zu testende Nullhypothese wählen wir

$$H_0 : F_1 = F_2 = \cdots = F_k.$$

Trifft H_0 zu, so sind die $n = n_1 + \cdots + n_k$ Zufallsvariablen

$$X_{11}, \ldots, X_{1n_1}; X_{21}, \ldots, X_{2n_2}; \ldots; X_{k1}, \ldots, X_{kn_k}$$

unabhängig und identisch verteilt. Wir ordnen die Gesamtheit der Beobachtungsergebnisse der Größe nach und nummerieren sie beim kleinsten beginnend durch. Jedem Ergebnis ordnen wir seine Nummer als Rang zu. Bei gleichen Ergebnissen wählen wir den mittleren Rang. Dadurch werden den n_i Ergebnissen der i-ten Messreihe n_i Ränge $r_{i1}, \ldots, r_{in_i}$ zugeordnet. Falls die Nullhypothese zutrifft, kann man erwarten, dass die Mittelwerte

$$\overline{r_i} = \frac{1}{n_i}\left(r_{i1} + \cdots r_{i_{n_i}}\right)$$

der jeweiligen Gruppenränge sich nicht sehr voneinander unterscheiden und deshalb alle in der Nähe des Gesamtmittels

$$\begin{aligned}\overline{r} &= \frac{1}{n}\left(r_{11} + \cdots + r_{1n_1} + \cdots + r_{k1} + \cdots + r_{kn_i}\right) \\ &= \frac{1}{n}\left(1 + 2 + \cdots + n\right) = \frac{n+1}{2}\end{aligned}$$

aller Ränge liegen werden. Darum kann man davon ausgehen, dass die Quadrate

$$\left(\overline{r_1} - \frac{n+1}{2}\right)^2, \ldots, \left(\overline{r_k} - \frac{n+1}{2}\right)^2$$

klein sein werden. 1st die Nullhypothese verletzt, so ist damit zu rechnen, dass für einige der Gruppen der Mittelwert der Ränge deutlich unterhalb, für andere deutlich oberhalb des Gesamtmittels liegen wird und deswegen einige der Quadrate größer ausfallen werden. Wir werden deshalb die Nullhypothese verwerfen, wenn die Prüfgröße

$$T(x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2}; \ldots; x_{k1}, \ldots, x_{kn_k}) = c\sum_{i=1}^{k} n_i \left(\overline{r_i} - \frac{n+1}{2}\right)^2$$

eine kritische Schranke s übersteigt. Dabei hängt der Skalierungsfaktor c von der Anzahl der Bindungen ab. Ist c geeignet gewählt (für Details siehe Lehn/Wegmann S. 179f), ist die Prüfgröße unter H_0 näherungsweise χ^2-verteilt mit $k-1$ Freiheitsgraden.
Als kritische Schranke wählt man die kleinste reelle Zahl s mit

$$\mathrm{P}_{H_0}(T > s) \leq \alpha.$$

Das Testverfahren mit der Entscheidungsregel

$$H_0 \text{ ist abzulehnen, falls } T(x_{11}, \ldots, x_{1n_1}; x_{21}, \ldots, x_{2n_2}; \ldots; x_{k1}, \ldots, x_{kn_k}) > s$$

besitzt dann das Niveau α. Es heißt der „Kruskal-Wallis-Test“.

Bemerkung: In Anlehnung an verschiedene χ^2-Tests (siehe nächste Abschnitte) verwenden Bühner/Ziegler $\sqrt{T/n}$ als Effektstärke, wobei $n := \sum_{i=1}^{k} n_i$ die Gesamtzahl der Versuchspersonen sind, und bezeichnen diese als Cohens w.
Beispiel: Um den Erfolg verschiedener Lehrmethoden für die Vorlesung „Wirtschaftsmathematik“ zu überprüfen, werden drei Gruppen mit ca. 30 Studierenden gebildet, die jeweils mit einer anderen Lehrmethode unterrichtet werden. Für jede:n Studierende:n wird festgehalten, wie viele Punkte in der Mathe-Klausur erzielt wurden. Danach wird der Rang unter allen Studierenden aller Gruppen ermittelt. Die Ergebnisse sind auszugsweise in folgender Tabelle zusammengestellt:

klassisch		*Peer instruction*		*JITT*	
Punkte	Rang	Punkte	Rang	Punkte	Rang
2	2	85	75	49	43
52	45	75	50	85	75
32	20	13	15	30	19
⋮	⋮	⋮	⋮	⋮	⋮
$n_1 = 28$		$n_2 = 33$		$n_3 = 31$	
$\overline{r_1} = 29.6$		$\overline{r_2} = 61.3$		$\overline{r_3} = 46.0$	

Damit ergibt sich ($n = n_1 + n_2 + n_3 = 92$):

$$T = c \cdot \left(28 \cdot \left(29.6 - \frac{93}{2} \right)^2 + 33 \cdot \left(61.3 - \frac{93}{2} \right)^2 + 31 \cdot \left(46.0 - \frac{93}{2} \right)^2 \right)$$

Da die Prüfgröße unter der Nullhypothese χ^2-verteilt ist mit $3 - 1 = 2$ Freiheitsgraden, ergibt sich bei einem Signifikanzniveau α von 5% eine kritische Schranke von $s =$ `qchisq(0.95,2)` $= 5.99$. Auf die genaue Berechnung der Prüfgröße soll hier verzichtet werden, da wir zur Berechnung des Skalierungsfaktors c die Anzahl der Bindungen wissen müssten und die Daten nur auszugsweise angegeben sind.
In R wird der Kruskall-Wallis-Test mit den Befehl `kruskal.test` durchgeführt.
Zwei Anwendungen finden Sie in Übungsaufgabe 1.3.

7 Chi-Quadrat-Tests

In diesem Abschnitt wird eine ganze Reihe von Chi-Quadrat-Tests vorgestellt: Der Chi-Quadrat-Test zum Prüfen der Klassen-Wahrscheinlichkeiten in einer Multinomialverteilung, die Chi-Quadrat-Anpassungstests zum Prüfen einer Verteilungsannahme mit bekannten und unbekannten Parametern sowie der Chi-Quadrat-Unabhängigkeits- und Homogenitätstest.
All diesen Tests ist gemeinsam, dass eine $k \times l$-Feldertafel vorliegt, die in jeder Zelle eine beobachtete Anzahl b_{ij} $i = 1 \ldots, k;, j = 1, \ldots, l$ enthält.
Außerdem wird je Zelle eine unter der Nullhypothese erwartete Anzahl e_{ij} ermittelt. Als Prüfgröße T ermittelt man je Zelle die standardisierte quadratische Abweichung $\frac{(b_{ij} - e_{ij})^2}{e_{ij}}$ und summiert über alle diese Werte auf.
In allen Fällen ist die Prüfgröße unter der Nullhypothese asymptotisch Chi-Quadrat-verteilt, wobei sich die Anzahl der Freiheitsgrade aus dem speziellen Testproblem ergeben.
Eine Daumenregel besagt, dass man die Näherung nur dann verwenden soll, wenn jede Zelle mindestens eine erwartet Häufigkeit von fünf (siehe Lehn/Wegmann) aufweist. Im Zweifel sollte man mit der exakten (und nur mit viel Aufwand zu ermittelnden) Verteilung oder mit einem simulierten p-Wert arbeiten. Letzteres ist über eine Option mit dem `chisq.test`-Kommando, das für nahezu alle oben genannten Tests (Ausnahme: Anpassungstest an eine Verteilung mit unbekannten Parametern) verwendet werden kann.

8 Effekt- und Teststärke

Cohen schlägt in seinem Buch „Statistical Power Analysis for the Behavioral Sciences“ in Kapitel 7 auf den Seiten 215 bis 227 folgende Effektstärke w vor:

$$w = \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(p_{0;ij} - p_{1;ij})^2}{p_{0;ij}}}.$$

Dabei ist $p_{0;ij}$ die unter der Nullhypothese erwartete relative Häufigkeit für die Zelle (i, j) und $p_{1;ij}$ die unter einer Alternative erwartete relative Häufigkeit für die Zelle (i, j). Mit absoluten Häufigkeiten ausgedrückt lautet die Formel:

$$w = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(e_{ij}/N - b_{ij}/N)^2}{e_{ij}/N}} = \sqrt{\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(e_{ij} - b_{ij})^2}{e_{ij}}} = \sqrt{\frac{T}{N}},$$

wobei N die Anzahl der Beobachtungen ist.
Cohen schreibt, dass ein Wert von w von 0,1 einem schwachen Effekt, ein Wert von 0,3 einem mittleren Effekt und ein Wert von 0,5 einem starken Effekt entspricht. Er empfiehlt allerdings, dass man diese Werte nur als grobe Richtschnur verwenden sollte und im Einzelfall sich besser selbst die Effektstärke w zu einer Abweichung, die man einschätzen kann, berechnen sollte.
Zur Berechnung der Effektstärke w stehen im bereits erwähnten Paket `pwr` für die Chi-Quadrat-Anpassungstests die Funktion `ES.w1` und für den Chi-Quadrat-Unabhängigkeits- und Homogenitätstest die Funktion `ES.w2` zur Verfügung.
Die Teststärke kann man mit der Funktion `pwr.chisq.test` aus dem gleichen Paket berechnet werden. Mit dieser Funktion kann auch die benötigten Stichprobengröße bei gewünschter Teststärke berechnet werden (und vieles mehr).

9 Multinomialverteilung

Sei X $\mathbb{R}^k$-we ZV mit

$$\mathrm{P}\left(X = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right) = p_1, \ldots, \mathrm{P}\left(X = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}\right) = p_k$$

mit $p_i \geq 0$ und $\sum p_i = 1$.

Definition:

Eine solche Wahrscheinlichkeitsverteilung heißt „Multinomial-Vtlg.“ mit Param.-vektor $p = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}$,

i.Z. $M_{1,p}$.

Beispiele:

- Glücksrad mit k Segmenten
- Würfeln ($k = 6$)
- Verteilung der Stimmanteile aller Parteien
- für k=2 erhält man die Binomialverteilung
- Roulette

10 χ^2-Anpassungstest zum Prüfen der Klassenwahrscheinlichkeiten

s.z.B.: Georgii, S. 298 ff

Wie testet man die Korrektheit eines Würfels?
Wie testet man die Zufälligkeit von Zufallszahlen?
Wie testet man, ob Umfrageergebnisse signifikant vom Ergebnis der letzten Bundestagswahl abweichen?

Seien $X_1, ..., X_n$ u.i.v. gem. $M_{1,p}$ verteilt mit p wie im letzten Abschnitt, d.h. k möglichen Ergebnissen.

Testproblem:
$p = p_0$ gg. $p \neq p_0$

Test:
(χ^2-Anpassungstest zum Prüfen der Klassenwahrscheinlichkeiten einer Multinomialvtlg.)

$$\varphi(\underbrace{\begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1k} \end{pmatrix}}_{=\begin{pmatrix} x_1 \\ \end{pmatrix}}, ..., x_n) = \begin{cases} 1 & \text{für } \underbrace{\sum_{i \leq k} \frac{\left(\overbrace{\sum_{j=1}^{n} x_{ji}}^{\text{beob. Häufigk. für } i} - \overbrace{np_i}^{\text{erw. Häufigk. für i}} \right)^2}{\underbrace{np_i}_{\text{erw. Häufig. für i}}}}_{=:T(x)} \begin{cases} > \\ \leq \end{cases} c \\ 0 & \end{cases}$$

Wie ist die Prüfgröße verteilt?
exakt: schwierig
asymptotisch χ^2_{k-1}-vtlt. (siehe Georgii); man kann sich den einen verlorenen Freiheitsgrad dadurch intuitiv erklären, als sich die beobachtete Anzahl in der letzten Klasse k aus den anderen Anzahlen deterministisch ergibt.

$\Rightarrow c = (1-\alpha)$-Quantil von χ^2_{k-1}.

Bemerkungen:

(i) üblicher Tipp: χ^2-Approx. zufriedenstellend, falls $np_i \geq 5 \; \forall i = 1, \ldots, k$

(ii) mein Tipp: bei Zweifeln: Simulation zur Ermittlung des p-Werts durchführen

Beispiel:
Nach Georgii, Bsp. 11.14, S.303:
Umfrage unter 783 Münchnern:

	CSU	SPD	Grüne	Sonstige
Umfrage	38.0	38.7	11.7	11.6
Wahl	37.9	37.4	9.6	15.1

Weichen die Umfrageergebnisse signifikant von dem Wahlergebnis ab?

$T(x) = \frac{(0.38 \cdot 783 - 0.379 \cdot 783)^2}{0.379 \cdot 783} + ... + \frac{(0.116 \cdot 783 - 0.151 \cdot 783)^2}{0.151 \cdot 783}$

Mit R erhält man:

```
#
chisq.test( c(0.38, 0.387, 0.117, 0.116)*783,
            p=c(0.379, 0.374, 0.096, 0.151) )

##
##  Chi-squared test for given probabilities
##
## data:  c(0.38, 0.387, 0.117, 0.116) * 783
## X-squared = 10.305, df = 3, p-value = 0.016
```

d.h. $T(x) \approx 10.3$.
Das das 98%-Quantil der χ^2_3-Verteilung `qchisq(0.98, 3)` ≈ 9.84
und das 99%-Quantil der χ^2_3-Verteilung `qchisq(0.99, 3)` ≈ 11.34 lautet, kann die Nullhypothese auf einem Signifikanzniveau von 2% abgelehnt werden, nicht aber auf einem Signifikanzniveau von 1%. Zum selben Ergebnis kommt man, wenn man den p-Wert von 1.6% beachtet.
Eine signifikante Abweichung vom Wahlergebnis kann also nur auf einem Niveau von 2% nachgewiesen werden.

Wenn man der Asymptotik misstraut (obwohl alle erwarteten Zellhäufigkeiten mindestens fünf sind), kann man den p-Wert auch simulieren:

```
chisq.test( c(0.38, 0.387, 0.117, 0.116)*783,
            p=c(0.379, 0.374, 0.096, 0.151), simulate.p.value = TRUE, B=100000 )

##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 100000 replicates)
##
## data:  c(0.38, 0.387, 0.117, 0.116) * 783
## X-squared = 10.305, df = NA, p-value = 0.016

#
```

Dies führt zum gleichen p-Wert, was bedeutet, dass die Chi-Quadrat-Approximation sehr gut war. ABC 4
Die Effektstärke dieses Testproblems beträgt:

```
#
library(pwr)
(w= ES.w1( P0=c(0.379, 0.374, 0.096, 0.151), P1=c(0.38, 0.387, 0.117, 0.116) ))

## [1] 0.1147
```

Bei dieser Effektstärke und einem Signifikanzniveau von 5%, beträgt die Teststärke:

```
pwr.chisq.test( w, N=783, df=3, sig.level = 0.05 )

##
##      Chi squared power calculation
##
##               w = 0.1147
##               N = 783
```

```
##              df = 3
##       sig.level = 0.05
##           power = 0.7748
##
## NOTE: N is the number of observations

#
```

Es liegt also ein nur schwacher Effekt und eine zufriedenstellende Teststärke vor. Ü 1

11 χ^2-Anpassungstest zum Prüfen einer Verteilungsannahme mit bekannten Parametern

Der Anpassungstest aus dem letzten Abschnitt kann auch zum Prüfen einer Verteilungsannahme für beliebige Verteilungen auf den reellen Zahlen verwendet werden. Die Verteilung muss dabei mit allen Parametern bekannt sein. Zum Testen auf eine Verteilungsfamilie, wie z.B. die Poisson- oder die Normalverteilung mit unbekannten Parametern, siehe nächster Abschnitt.
Dazu teilt man die reellen Zahlen so in Intervalle ein, dass jedes Intervall unter der Verteilung etwa die gleiche Wahrscheinlichkeit hat.
Dann berechnet man je Intervall die Wahrscheinlichkeit unter der Verteilung gegen die getestet werden soll. Nun zählt man für jedes Intervall, wie viele Beobachtungen in ihm liegen und führt den χ^2-Anpassungstest zum Prüfen der Klassenwahrscheinlichkeiten aus dem letzten Abschnitt durch.
Der Test funktioniert für jede beliebige Verteilung, hat aber eine deutlich geringere Teststärke als spezielle Verteilungstests wie z.B. der Lilliefors-Test oder der Shapiro-Wilk-Test, weil die Verteilung nur für einige Intervalle mit den Beobachtungen verglichen wird.

Beispiel:
Bei der Notengebung wird Lehrern manchmal vorgeworfen, sie neigten dazu, Extremurteile zu vermeiden. In einem Kurs erzielten 17 Schüler folgende Durchschnittsnoten:

1.58	2.84	3.52	4.16	5.36	2.01	3.03	3.56	
4.19	2.35	3.16	3.75	4.60	2.64	3.40	3.99	4.75

Nehmen Sie der Einfachheit halber an, dass sich diese Durchschnittsnoten aus so vielen Einzelnoten ergeben haben, dass sie als kontinuierlich verteilt angesehen werden können. Prüfen Sie zum Niveau $\alpha = 0.1$ mit dem χ^2- Anpassungstest die Nullhypothese, dass obige Daten $N_{3.5,1}$-verteilt sind. Unterteilen Sie hierzu die relativen Häufigkeiten in die sechs Gruppen

$$(-\infty, 1.5] \quad (1.5,2.5] \quad (2.5,3.5] \quad (3.5,4.5] \quad (4.5,5.5] \quad (5.5, \infty)$$

Berechnen Sie abschließend noch die Effektstärke w.

Lösung:

```
## Chi-Quadrat-Anpassungstest
##
## Absolute Notenhäufigkeiten entweder selbst zählen (uncool):
haeuf.noten.kat <- c(0,3,5,6,3,0)
## oder R zaehlen lassen (cool)
noten.dat <- c(1.58,2.84,3.52,4.16,5.36,2.01,3.03,3.56,4.19,
```

```
               2.35,3.26,3.75,4.6,2.64,3.4,3.99,4.75)
noten.kat <- cut(noten.dat,0:6+0.5)
haeuf.noten.kat.R <- table(noten.kat)
# Kommt dasselbe raus ?
all(haeuf.noten.kat == haeuf.noten.kat.R)

## [1] TRUE

# yep!
#
# Gem. N-Vtlg erwartete rel. Notenhäufigkeiten
haeuf.rel.noten.kat.nvtlg <-
  c( pnorm(1.5,3.5,1), pnorm(2.5,3.5,1)-pnorm(1.5,3.5,1),
     pnorm(3.5,3.5,1)-pnorm(2.5,3.5,1),
     pnorm(4.5,3.5,1)-pnorm(3.5,3.5,1),
     pnorm(5.5,3.5,1)-pnorm(4.5,3.5,1), 1-pnorm(5.5,3.5,1) )
##
## Chi-Qu.Anp.test
##
chisq.test( haeuf.noten.kat, p=haeuf.rel.noten.kat.nvtlg )

## Warning in chisq.test(haeuf.noten.kat, p = haeuf.rel.noten.kat.nvtlg): Chi-squared
approximation may be incorrect

##
##  Chi-squared test for given probabilities
##
## data:  haeuf.noten.kat
## X-squared = 1.303, df = 5, p-value = 0.9346

# Bemerkungen:
#  (1) in haeuf.noten.kat muessen absolute Zahlen eingetragen
#      werden
#  (2) in haeuf.rel.noten.kat.nvtlg dürften auch absolute Zahlen
#      stehen; man müsste dann allerdings die Option
#      rescale.p auf TRUE setzen
#  (3) chisq.test( haeuf.noten.kat, haeuf.rel.noten.kat.nvtlg )
#      führt einen Chi-Quadrat-Unabh.test durch. Dabei werden
#      die beiden Argumente als Ausprägungen von zwei
#      Merkmalen angesehen (jeder Vektoreintrag enstricht dabei
#      einem Datensatz). Diese einzelnen Datensätze werden von
#      der Funktion zunächst zu einer Kontingenztabelle
#      aggregiert.
#      Für diese Tabelle wird dann der Chi-Quadrat-Unabh.test
#      durchgeführt (siehe ?chisq.test). Zu unseren Daten (die
#      ja schon aggregiert sind) passt das nat. gar nicht.
##
## p-Wert mit Simulation
##
chisq.test( haeuf.noten.kat, p=haeuf.rel.noten.kat.nvtlg,
            simulate.p.value=TRUE, B=10000 )
```

```
##
##  Chi-squared test for given probabilities with simulated p-value
##  (based on 10000 replicates)
##
## data:  haeuf.noten.kat
## X-squared = 1.303, df = NA, p-value = 0.9728

##
## Achtung: In the goodness-of-fit case simulation is done by
## random sampling from the discrete distribution specified
## by p, each sample being of size n = sum(x).
## This simulation is done in R and may be slow.
## Also B nicht zu groß waehlen, sonst dauert die
## Befehlsausfuehrung sehr lang.
#
# Für gute Laptops:
chisq.test( haeuf.noten.kat, p=haeuf.rel.noten.kat.nvtlg,
            simulate.p.value=TRUE, B=100000 )

##
##  Chi-squared test for given probabilities with simulated p-value
##  (based on 1e+05 replicates)
##
## data:  haeuf.noten.kat
## X-squared = 1.303, df = NA, p-value = 0.9728
```

Auf einem Niveau von 10% kann der Test auf Normalverteilung nicht abgelehnt werden, da der p-Wert ca. 97% beträgt. Die Notenverteilung ist also verträglich mit der Normalverteilungsannahme. Für die Effektstärke ergibt sich $w = \sqrt{\texttt{X-squared}/17} = 0.28$, was nach Cohen einem mittleren Effekt entspricht. Man bräuchte aber die Daten von 168 Schüler, um einen solchen Effekt auf einem Signifikanzniveau von 5% bei einer Teststärke von 80% nachweisen zu können, wie folgende Analyse zeigt:

```
#
library(pwr)
(w= ES.w1( P0=haeuf.rel.noten.kat.nvtlg, P1=haeuf.noten.kat/17 ))

## [1] 0.2768

pwr.chisq.test( w, df=5, sig.level = 0.05, power=0.8 )

##
##      Chi squared power calculation
##
##               w = 0.2768
##               N = 167.4
##              df = 5
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations

#
```

12 χ^2-Anpassungstest zum Prüfen einer Verteilungsannahme mit unbekannten Parametern

Wir können auch Anpassungstests für eine Nullhypothese durchführen, bei der die Wahrscheinlichkeiten $\{p_i, i = 1, \ldots, k\}$ nicht vollständig festgelegt sind. Wenn wir beispielsweise wissen wollen, ob die Anzahl der Menschen, die an einem Tag zu einer Beratungsstelle kommen, Poisson verteilt sind, ohne die Ankunftsrate λ zu kennen, oder wenn wir prüfen wollen, ob unserer Daten normalverteilt sind, ohne μ and σ^2 zu kennen.

Satz

Nehmen wir an, dass es m solcher unspezifizierten Parameter gibt und dass sie aus der Häufigkeitstabelle (nicht aus den Rohdaten!) mit der Maximum-Likelihood-Methode zu schätzen sind. Dann gilt für große n näherungsweise:

$$T \sim \chi^2_{k-1-m}$$

Mit anderen Worten, für jeden Parameter, der geschätzt werden muss, muss ein Freiheitsgrad abgezogen werden.

Der Test lautet also

Lehne H_0 ab, wenn $T \geq F^{-1}_{\chi^2_{k-1-m}}(1-\alpha)$,

und der p-Wert beträgt bei einer Prüfgröße von $T = t$:

$$p\text{-Wert} \approx 1 - F_{\chi^2_{k-1-m}}(t)$$

Bemerkung: Die Schätzung der Parameter aus der Häufigkeitstabelle mit der Maximum-Likelihood-Methode ist meist nur iterativ möglich und liegt außerhalb unserer Vorlesung. In Aufgaben werde ich diese Schätzwerte stets angeben.

Beispiel:

Bei einem IQ-Test unter einer Stichprobe von 120 Studierenden traten folgende Häufigkeiten auf:

Intervall	Anzahl
(-∞,70]	3
(70,85]	18
(85,100]	30
(100,115]	35
(115,130]	32
(130, ∞)	2

Prüfen Sie die Nullhypothese, dass die IQ-Verteilung unter den Studierenden normalverteilt ist, wenn die Maximum-Likelihood-Schätzer 102.625 für μ und 17.12 für σ betragen.

Lösung: Zunächst müssen wir die Wahrscheinlichkeiten für die sechs Intervalle unter der Annahme der Nullhypothese, also einer Normalverteilung der Daten mit $\mu = 102.625$ und $\sigma = 17.12$ berechnen:

Intervall	Wahrscheinlichkeit	
(-∞,70]	0.028	`=pnorm(70,102.625,17.12)`
(70,85]	0.123	
(85,100]	0.288	
(100,115]	0.326	
(115,130]	0.180	
(130, ∞]	0.055	

Die Prüfgröße kann mit dem `chisq.test`-Befehl berechnet werden:

```
#
chisq.test( c(3,18,30,35,32,2), p=c(0.028, 0.123, 0.288, 0.326, 0.180, 0.055) )

##
##  Chi-squared test for given probabilities
##
## data:  c(3, 18, 30, 35, 32, 2)
## X-squared = 10.0, df = 5, p-value = 0.075

#
```

Damit ergibt sich für den p-Wert $1-F_{\chi^2_{6-1-2}}(10) = 1-F_{\chi^2_3}(10)$ =`1-pchisq(10,3)`=1.9%. Die Nullhypothese kann auf einem Niveau von 2%, nicht aber auf einem Nivau von 1% abgelehnt werden.

Achtung: Man darf weder den p-Wert aus dem Ergebnis des `chisq.test`-Kommandos, noch die `simulate.p.value`-Option verwenden, da R nicht weiß, dass wir zuvor zwei Parameter geschätzt haben.

ABC
5

13 χ^2-Test auf stochastische Unabhängigkeit

Es soll nun der exakte Test von Fisher zum Prüfen der Unabhängigkeit in einer Vierfeldertafel (siehe Abschnitt **??**) erweitert werden auf den Fall von Merkmalen mit mehr als zwei Ausprägungen. Da die Berechnung des p-Werts des exakten Tests von Fisher für größere Tafeln sehr aufwendig und irgendwann kaum mehr möglich ist, leiten wir zudem einen approximativen Test her, den χ^2-Test auf stochastische Unabhängigkeit.

Gegeben sei eine $a \times b$-Feldertafel, also $k = ab$ viele Klassen, mit den entsprechenden Wahrscheinlichkeiten π_{ij} für $i = 1, \ldots, a$ und $j = 1, \ldots, b$. Die linke Tabelle zeigt die Beobachtungshäufigkeiten, die rechte die Modellwahrscheinlichkeiten.

	B_1	$\ldots$	B_b	Σ
A_1	n_{11}	$\ldots$	n_{1b}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	n_{a1}	$\ldots$	n_{ab}	$n_{a\bullet}$
Σ	$n_{\bullet 1}$	$\ldots$	$n_{\bullet b}$	n

	B_1	$\ldots$	B_b	Σ
A_1	π_{11}	$\ldots$	π_{1b}	$\pi_{1\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_a	π_{a1}	$\ldots$	π_{ab}	$\pi_{a\bullet}$
A_a	$\pi_{\bullet 1}$	$\ldots$	$\pi_{\bullet b}$	1

Dabei ist $n_{i\bullet} := \sum_{j=1}^{b} n_{ij}$ die i-te Zeilensumme der Beobachtungshäufigkeiten, also die Summe der Beobachtungshäufigkeiten von Ausprägung A_i des ersten Merkmals, und $n_{\bullet j} := \sum_{i=1}^{a} n_{ij}$ die j-te Spaltensumme der Beobachtungshäufigkeiten, also die Summe der Beobachtungshäufigkeiten von Ausprägung B_j des zweiten Merkmals. Entsprechend ist $\pi_{i\bullet} := \sum_{j=1}^{b} \pi_{ij}$ die i-te Zeilensumme der Klassenwahrscheinlichkeiten, also die Summe der Klassenwahrscheinlichkeiten von Ausprägung A_i des ersten Merkmals, und $\pi_{\bullet j} := \sum_{i=1}^{a} \pi_{ij}$ die j-te Spaltensumme der Klassenwahrscheinlichkeiten, also die Summe der Klassenwahrscheinlichkeiten von Ausprägung B_j des zweiten Merkmals. Folglich lautet das Testproblem:

Nullhypothese: Die Merkmale A und B sind stochastisch unabhängig, d.h. $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ für alle $i = 1, \ldots, a$ und $j = 1, \ldots, b$.

Alternative: Die Merkmale A und B sind stochastisch abhängig, d.h. $\pi_{ij} \neq \pi_{i\bullet} \cdot \pi_{\bullet j}$ für mindestens ein Paar $(i, j) \in \{1, \ldots, a\} \times \{1, \ldots, b\}$.

Um Tests für dieses Testproblem entwickeln zu können, nimmt man - wie schon beim exakten Test von Fisher für die Vierfeldertafel - an, dass die Randhäufigkeiten fest sind. Als Modell wählt man daher eine Multinomialverteilung $M_{n;(\pi_{ij})}$ mit $a \cdot b$ verschiedenen Ausprägungen und festen Summe der Klassenwahrscheinlichkeiten $\pi_{i\bullet} = n_{i\bullet}/n$ und $\pi_{\bullet j} = n_{\bullet j}/n$. Unter der Nullhypothese gilt dann für die Erfolgswahrscheinlichkeit von Zelle i, j: $\pi_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n^2}$.
Beim exakten Test von Fisher berechnet man für jede Tabelle, die aufgrund der Randhäufigkeiten möglich ist, die Wahrscheinlichkeit ihres Auftretens unter der Nullhypothese. Anschließend summiert man über alle Wahrscheinlichkeiten auf, die kleiner oder gleich der Wahrscheinlichkeit der beobachteten Tabelle sind, und erhält als Summe den p-Wert. Da dies bei großen Tafeln oder bei großen Anzahlen sehr viele Tabellen sein können, ist enorm viel Rechenarbeit notwendig. Daher kommt auch heute noch ein approximatives Verfahren um Einsatz:
Wählt man nämlich die gleiche Teststatistik wie beim Testen der Erfolgswahrscheinlichkeit im Multinomialmodell ohne feste Randhäufigkeiten, also

$$T((n_{ij})) = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(n_{ij} - n\pi_{ij})^2}{n\pi_{ij}} = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(n_{ij} - n_{i\bullet} n_{\bullet j}/n)^2}{n_{i\bullet} n_{\bullet j}/n},$$

so kann man zeigen, dass T als Funktion von $(a-1)(b-1)$ unabhängigen Zellhäufigkeiten geschrieben werden kann.[1] Es verwundert daher nicht, dass man zeigen kann, dass T asymptotisch $\chi^2_{(a-1)(b-1)}$ verteilt ist.[2]
Der χ^2-Test auf stochastische Unabhängigkeit lautet also:

$$\varphi(n_{11}, ..., n_{ab}) = \begin{cases} 1 \\ 0 \end{cases} \text{ für } T((n_{ij})) \begin{cases} > \\ \leq \end{cases} (1-\alpha)\text{-Qu. der } \chi^2_{(a-1)(b-1)}\text{-Vtlg.}$$

Beispiel Die (fiktiven) Ergebnisse einer Untersuchung von Bandscheibenvorfällen bei Krankenpflegern, Lehrern und Maurern sind in folgender Tabelle zusammengefasst:

	Krankenpfleger	Lehrer	Maurer	Summe
Bandscheibenvorfall	20	20	10	50
Kein Bandscheibenvorfall	10	20	20	50
Summe	30	40	30	100

Da hier sowohl Dimension als auch Inhalt der Kontigenztabelle nicht zu groß ist, können wir den exakten Test von Fisher mit dem R-Kommando

[1] Intuitiv kann man sich das klar machen, in dem man schaut, wie viele der Zellen in der Kontingenztafel bei gegebenen Randsummen frei wählbar sind. Dies sind gerade $(a-1)(b-1)$.

[2] Eine Daumenregel der Statistik sagt, dass man so viele Freiheitsgrade abziehen muss, wie Parameter der Multinomialverteilung festgelegt sind. Bei der $M_{n;p}$-Verteilung ohne Randbedingungen ist dies einer, da $p_k = 1 - \sum_{i=1}^{k} p_i$, daher kommt man auf $k-1$-Freiheitsgrade (siehe Abschnitt 10). In unserem Fall sind von den ab vielen Parametern der $M_{n;\pi}$-Verteilung wegen der $(a-1)$ gegebenen Zeilenhäufigkeiten, den $(b-1)$ gegebenen Spaltenhäufigkeiten und der gegebenen Gesamthäufigkeit (die letzte Zeilen-/Spaltenhäufigkeit ist keine zusätzliche Einschränkung, da diese Bedingungen bereits wegen der gegebenen Gesamthäufigkeit erfüllt sind) nur $ab-(a-1+b-1+1) = ab-a-b+1 = (a-1)(b-1)$ Parameter frei wählbar.

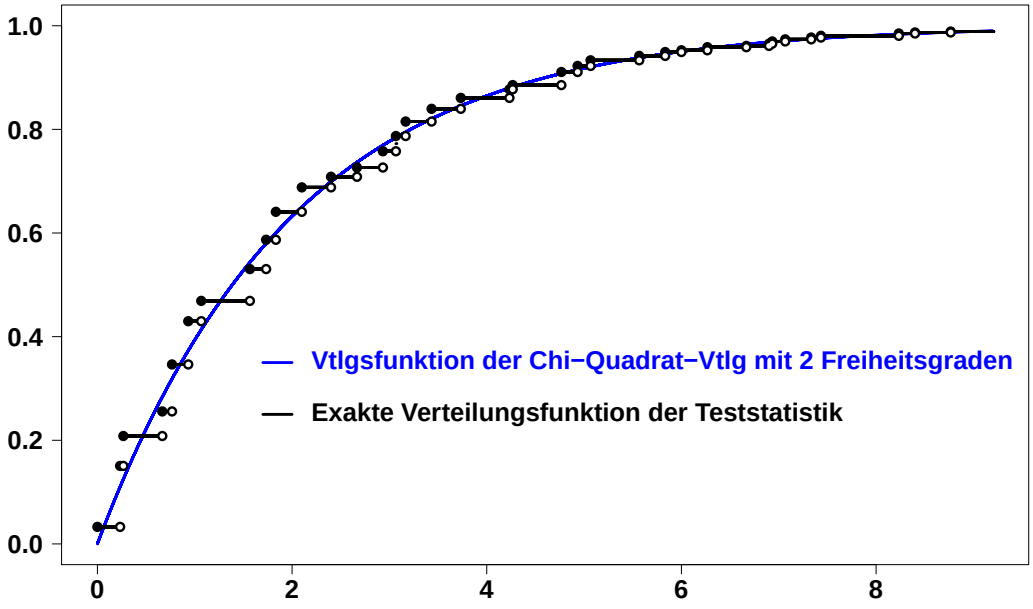
`fisher.test(x=matrix(c(20,10,20,20,10,20),nrow=2))` durchführen.
Man erhält einen exakt berechneten p-Wert von 0.04141, so dass die Nullhypothese der stochastischen Unabhängigkeit zwischen Beruf und dem Auftreten eines Bandscheibenvorfalls auf einem Niveau von 5% abgelehnt werden kann.
Alternativ führen wir nun noch den χ^2-Test auf stochastische Unabhängigkeit durch. Der Wert der Teststatistik ergibt sich hier zu:

$$T((n_{ij})) = \frac{(n_{11}-15)^2}{15} + \frac{(n_{12}-20)^2}{20} + \ldots + \frac{(n_{23}-15)^2}{15} \approx 6.67.$$

Da das 95%-Quantil der $\chi^2_{(2-1)(3-1)} = \chi^2_2$-Verteilung `qchisq(0.95, 2)`=5,99 beträgt, kann die Nullhypothese auch hier signifikant abgelehnt werden.
Alternativ sieht man dies auch, wenn man den R-Befehl
`chisq.test(x=matrix(c(20,10,20,20,10,20),nrow=2))` verwendet. Dieser liefert als Ausgabe sowohl die oben berechnete Teststatistik als auch den p-Wert des Testproblems: `X-squared = 6.6667, df = 2, p-value = 0.03567`.
Die erwarteten Zellhäufigkeiten erfüllen zwar alle den Tipp für die Verwendung der χ^2-Approximation, dennoch stellt sich die Frage, wie gut die Approximation ist.
Per Option des `chisq.test`-Kommandos können wir den p-Wert auch per Simulation ermitteln:

```
chisq.test(x=matrix(c(20,10,20,20,10,20),nrow=2),
           simulate.p.value = TRUE, B=20000000)
```

Diese liefert einen p-Wert in Höhe von ca. 0.04149 und damit schon eine gewisse Abweichung zum approximativ berechneten p-Wert.
Die folgende Grafik zeigt die Unterschiede zwischen approximierter und exakter Verteilungsfunktion der Teststatistik des χ^2-Test auf stochastische Unabhängigkeit.
Interessanterweise liefern in diesem Beispiel der exakte Test von Fisher und der χ^2-Test auf stochastische Unabhängigkeit unter Verwendung der Simulation nahezu identische p-Werte, obwohl sie auf unterschiedlichen Teststatistiken basieren. Das muss nicht so sein, wie wir in Übungsaufgabe 2.3 sehen werden.



14 χ^2-Homogenitätstest

Als letzte Anwendung der χ^2-Tests erweitern wir den Zweistichproben-Binomialtest aus Abschnitt **??** auf den Zweistichproben-Multinomialtest bzw. χ^2-Homogenitätstest.[3] Gegeben seien dazu die stochastisch unabhängigen und $\mathbb{R}^k$-wertigen Zufallsvektoren $X_{11}, \ldots, X_{1n_1}$ sowie $X_{21}, \ldots, X_{2n_2}$. Die Zufallsvektoren aus der ersten Stichprobe seien M_{1,p_1}-verteilt, die aus der zweiten Stichprobe M_{1,p_2}-verteilt. Wir wollen prüfen, ob sich die Vektoren p_1 und p_2 signifikant unterscheiden. Das Testproblem lautet also:

Nullhypothese: $p_1 = p_2$, d.h. $p_{1j} = p_{2j}$ für alle $j = 1, \ldots, k$

Alternative: $p_1 \neq p_2$, d.h. $p_{1j} \neq p_{2j}$ für mindestens ein $j \in \{1, \ldots, k\}$.

Wir fassen unsere Beobachtungen in einer $2 \times k$ - Feldertafel zusammen.

	Klasse 1	...	Klasse k	Σ
Population 1	n_{11}	...	n_{1k}	$n_1 := n_{1\bullet}$
Population 2	n_{21}	...	n_{2k}	$n_2 := n_{2\bullet}$
Σ	$n_{\bullet 1}$	...	$n_{\bullet k}$	n

Zur Berechnung der Teststatistik unterscheiden wir zwei Fälle:

1. Fall: $n_1 = n_2$, d.h. die beiden Stichprobenumfänge sind gleich lang. Dann schätzen wir unter der Nullhypothese für die Klasse j

$$\widehat{p}_{1j} = \widehat{p}_{2j} = \widehat{p}_j = \frac{n_{1j} + n_{2j}}{n}$$

und erhalten die Teststatistik

$$T((n_{ij})) = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n}{2} \cdot \frac{n_{1j}+n_{2j}}{n}\right)^2}{\frac{n_{1j}+n_{2j}}{2}} = \sum_{j=1}^{k} \frac{(n_{1j} - n_{2j})^2}{n_{1j} + n_{2j}}.$$

2. Fall: $n_1 \neq n_2$, d.h. die beiden Stichprobenumfänge besitzen unterschiedliche Länge. Die Schätzung der Wahrscheinlichkeit für Population i und Klasse j lauten unter der Nullhypothese

$$\widehat{p}_{ij} = \frac{n_i}{n} (n_{1j} + n_{2j}) = \frac{n_i \cdot n_{\bullet j}}{n}$$

und die Teststatistik lautet

$$T((n_{ij})) = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_i \cdot n_{\bullet j}}{n}\right)^2}{\frac{n_i \cdot n_{\bullet j}}{n}}.$$

Da die Formel für die Teststatistik im ersten Fall ein Spezialfall der Formel aus dem zweiten Fall ist, lautet der Test in beiden Fällen

$$\varphi(n_{11}, ..., n_{22}) = \begin{cases} 1 & \text{für } T((n_{ij})) \begin{cases} > \\ \leq \end{cases} (1-\alpha)\text{-Qu. der } \chi^2_{k-1}\text{-Vtlg.} \\ 0 \end{cases}$$

Bemerkungen:

[3] Am Ende des Abschnitts werden wir sehen, dass der Zweistichprobentest unproblematisch auf mehr als zwei Stichproben erweitert werden kann.

(a) Die Teststatistik ist identisch mit der Teststatistik des χ^2-Unabhängigkeitstests für die Situation, dass es bei einem der beiden Merkmale nur zwei Ausprägungen gibt. Der χ^2-Test zum Vergleich zweier Multinomialverteilungen ist also ein Spezialfall des χ^2-Unabhängigkeitstests. Dies ist sofort einsichtig, wenn man die Nullhypothese so interpretiert, dass das Merkmal *Klasse* unabhängig vom Merkmal *Population* ist.

(b) Analog kann man auch mehr als zwei Stichproben auf Gleichheit ihrer Multinomialverteilungen testen („Homogenitätstest"). Der χ^2-Test zum Vergleich mehrerer Multinomialverteilungen ist identisch mit dem χ^2-Unabhängigkeitstests, wenn man die Nullhypothese so interpretiert, dass das Merkmal *Klasse* unabhängig vom Merkmal *Population* ist.

15 Korrelationstests

Für zwei metrisch skalierte Zufallsvariablen X und Y haben wir in Abschnitt IV.8 die Korrelation $\text{Corr}[X, Y]$ eingeführt, um den linearen Zusammenhang der beiden Größen in einer Maßzahl zusammenzufassen. Sie wird durch die Stichprobenkorrelation r geschätzt (siehe DA 1).
In Abschnitt IV.8 haben wir auch gesehen, dass stochastisch unabhängige Größen unkorreliert sind. Wenn die beiden Größen einer gemeinsamen Normalverteilung genügen[4], gilt auch die Umkehrung, also, dass unkorrelierte Größen stochastisch unabhängig sind.
Für diesen Spezialfall formulieren wir einen Test, der in der Praxis häufig auch für metrisch skalierte Größen, die nicht gemeinsam normalverteilt sind, verwendet wird. Alternativen sind in den Bemerkungen am Ende des Abschnitts zu finden.
Für gemeinsam normalverteilte Zufallsvariablen $(X_i, Y_i)_{i=1,\dots,n}$ können die Nullhypothesen:
$H_0 : \text{Corr}[X, Y] = 0$, $H_0 : \text{Corr}[X, Y] \geq 0$ und $H_0 : \text{Corr}[X, Y] \leq 0$
mit der Prüfgröße

$$T(X, Y) = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

getestet werden. Dabei ist r der Stichprobenkorrleationskoeffizient zwischen X und Y. Die Prüfgröße ist unter der Nullhypothese t_{n-2}-verteilt, weswegen die Ablehungsbereiche wie folgt bestimmt werden können: $|T| > (1 - \alpha/2)$-Quantil der t_{n-2}-Verteilung, $T < \alpha$-Quantil der t_{n-2}-Verteilung und $T > (1 - \alpha)$-Quantil der t_{n-2}-Verteilung.
Dieser Test wird „Korrelationstest nach Pearson" genannt und in R mit dem Kommando **cor.test** durchgeführt.
Als Effektstärke verwendet Cohen einfach der Stichprobenkorrelationskoeffizienten, da dieser bereits dimensionslos und unabhängig von der Wahl der Einheiten ist. Cohen schreibt einem Korrelationskoeffizienten von 0.1 einen schwachen Effekt zu, einem Wert von 0.3 einem mittleren, moderaten Effekt und ein Wert von 0.5 einem starken Effekt - genau wie bei der Effektstärke w bei den Chiquadrat-Tests. Diese Schranken sind mehr oder weniger willkürlich gewählt und sollten wieder nur als ganz groben Richtschnur verwendet werden. Andere Autoren schlagen andere Kategorisierung vor, wie z.B. Gignac und Szodorai,[5] die bei 0.1 von einen schwachen, bei 0.2 von einem mittleren und bereits bei 0.3 einem starken Effekt sprechen.
Bei gegebenem Effekt kann die Teststärke in R mit der Funktion `pwr.r.test` des bereits bekannten Pakets `pwr` berechnet werden.
Beispiel (aus Fahrmeier et al.):
In der folgenden Tabelle ist für die Jahre von 1975 bis 1994 die Prognose des Wirtschaftswachstums vom Sachverständigenrat sowie das reale Wachstum aufgelistet:

[4]Die sogenannten bivariate Normalverteilung haben wir nicht besprochen. Die Dichte dieser Verteilung hat als Definitionsbereich die Ebene und sieht aus wie eine (evtl. deformierte) Glocke.

[5]Gignac, G.E. und Szodorai, E.T. (2016). Effect size guidelines for individual differences researchers. Personality and Individual Differences, 102, 74-78.

Jahr	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Prognose	2.0	4.5	4.5	3.5	3.75	2.75	0.5	0.5	1.0	2.5
reales Wachstum	-3.6	5.6	2.4	3.4	4.4	1.8	-0.3	-1.2	1.2	2.6

Jahr	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Prognose	3.0	3.0	2.0	1.5	2.5	3.0	3.5	2.5	0.0	0.0
reales Wachstum	2.5	2.5	1.7	3.4	4.0	4.6	3.4	1.5	-1.9	2.3

Es soll nun signifikant nachgewiesen werden, dass eine positive Korrelation zwischen Prognose und realem Wachstum vorliegt.

```
prognose <- c(2.0,4.5,4.5,3.5,3.75,2.75,0.5,0.5,1.0,2.5,3.0,3.0,2.0,1.5,2.5,3.0,
              3.5,2.5,0.0,0.0)
wachstum <- c(-3.6,5.6,2.4,3.4,4.4,1.8,-0.3,-1.2,1.2,2.6,2.5,2.5,1.7,3.4,4.0,
              4.6,3.4,1.5,-1.9,2.3)
cor.test( prognose, wachstum, alternative = "greater" )

##
##  Pearson's product-moment correlation
##
## data:  prognose and wachstum
## t = 3.5, df = 18, p-value = 0.0012
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.3452 1.0000
## sample estimates:
##    cor
## 0.6404
```

Da der p-Wert bei 0.12% liegt, ist der Nachweis einer positiven Korrelation hochsignifikant. Eine derart hohe Korrelation in Höhe der Stichprobenkorrelation von 0.64 hat bei einem Signifikanzniveau von 5% auch eine hohe Teststärke:

```
library(pwr)
pwr.r.test(n=20, r=0.64, sig.level = 0.05, alternative = "greater")

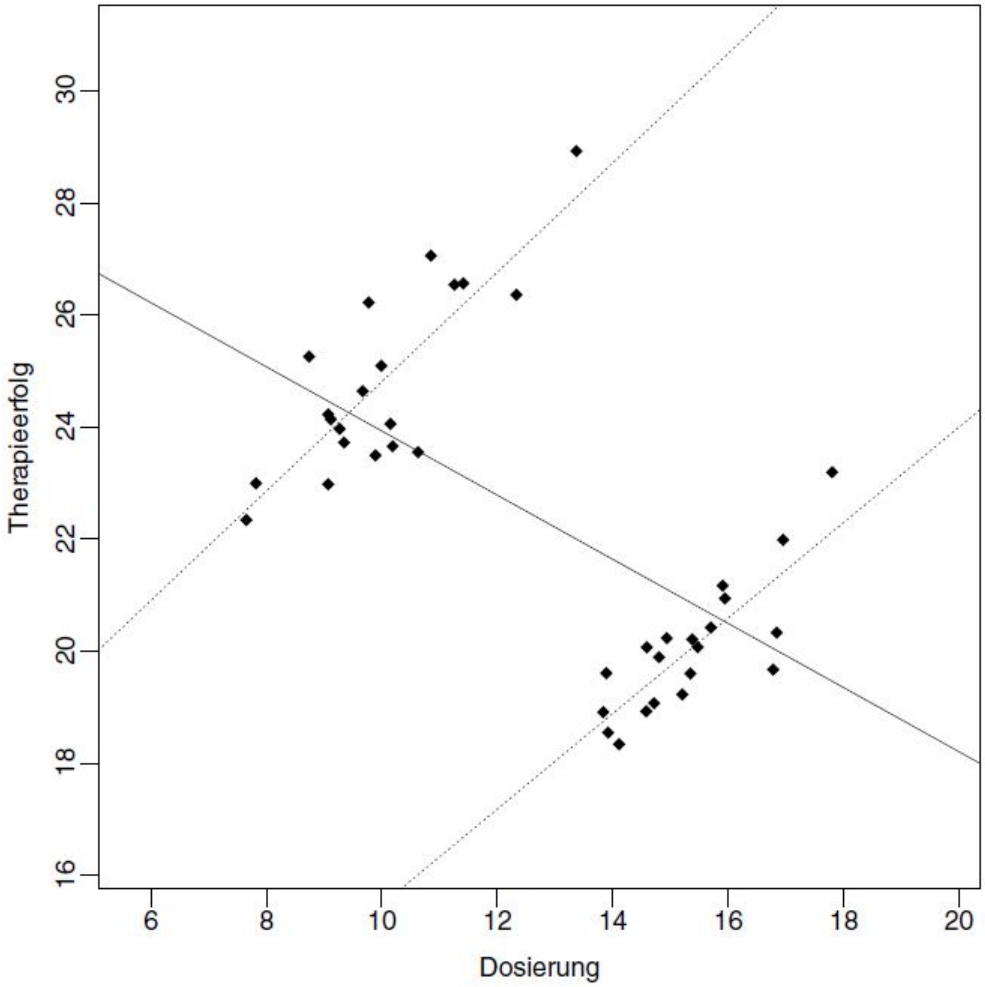
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 20
##               r = 0.64
##       sig.level = 0.05
##           power = 0.9399
##     alternative = greater

#
```

Bemerkungen:

- Achtung! Die Korrelation misst lediglich den lineare Zusammenhang. Zeichnen Sie immer auch ein Streudiagramm, wenn Sie den Zusammenhang zwischen zwei metrischen Merkmalen analysieren. Folgende Schockbilder stammen aus Fahrmeir et al. und sollten einen festen Platz in Ihrem Gedächtnis finden:





Negative Korrelation zwischen Dosierung und Therapieerfolg, obwohl die Therapieerfolg sowohl bei stark Erkrankten wie auch bei schwach Erkrankten mit der Dosierung ansteigt.

- Mit dem Paket `MVN` können einige Tests auf die bivariate Normalverteilung durchgeführt werden. Zu beachten ist allerdings, dass die Teststärke bei kleinen und mittlerer Stichprobengrößen gering ist. In unserem Beispiel testen wir die Gleichverteilung auf dem Einheitsquadrat, deren Dichte wie ein Würfel aussieht gegen die glockenförmige bivariate Normalverteilung. Erst bei einem Stichprobenumfang von 100 wird der (sehr große) Unterschied von den beiden populärsten Tests, dem Henze-Zirkler Test und dem Royston Test signifikant erkannt.

```r
set.seed(1)
zufaelligePunkteimQuadrat10 <- matrix( c(runif(20)), ncol=2 )
zufaelligePunkteimQuadrat30 <- matrix( c(runif(60)), ncol=2 )
zufaelligePunkteimQuadrat100 <- matrix( c(runif(200)), ncol=2 )
library(MVN)
mvn(zufaelligePunkteimQuadrat10, mvnTest = "hz")$multivariateNormality

##            Test     HZ p value MVN
## 1 Henze-Zirkler 0.3452  0.4053 YES

mvn(zufaelligePunkteimQuadrat10, mvnTest = "royston")$multivariateNormality

##      Test      H p value MVN
## 1 Royston 0.8118  0.6664 YES

mvn(zufaelligePunkteimQuadrat30, mvnTest = "hz")$multivariateNormality

##            Test     HZ p value MVN
## 1 Henze-Zirkler 0.6348  0.1489 YES

mvn(zufaelligePunkteimQuadrat30, mvnTest = "royston")$multivariateNormality

##      Test    H p value MVN
## 1 Royston 5.61  0.0605 YES

mvn(zufaelligePunkteimQuadrat100, mvnTest = "hz")$multivariateNormality

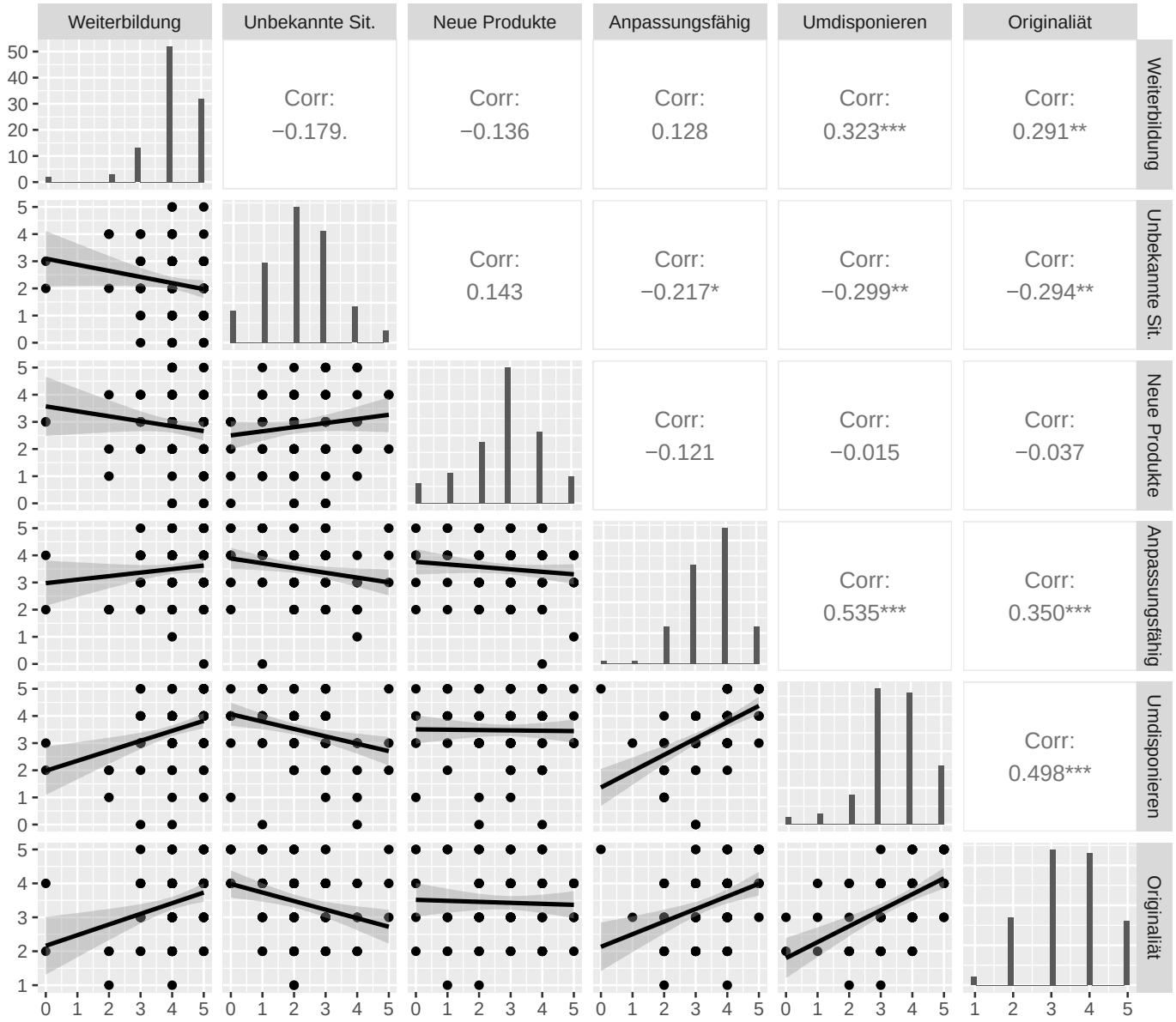
##            Test    HZ  p value MVN
## 1 Henze-Zirkler 1.441 0.002581  NO

mvn(zufaelligePunkteimQuadrat100, mvnTest = "royston")$multivariateNormality

##      Test     H   p value MVN
## 1 Royston 19.52 5.769e-05  NO
```

- Mit dem Kommando `cor.test` können auch die Korrelationstests von Spearman für die Rangkorrelationen und von Kendall durchgeführt werden, die keine bivariate Normalverteilung der Daten voraussetzen (siehe auch Übung 2.5). Beim Korrelationstests von Spearman wird vorausgesetzt, dass die Abstände zwischen zwei Rängen als gleich groß betrachtet werden können, beim Korrelationstest von Kendall werden lediglich ordinal skalierte Daten vorausgesetzt.
- Mit dem Befehl `ggpairs` aus dem Paket `GGally` werden die Korrelationen für alle Kombinationen von Merkmalen berechnet und der Korrelationstest nach Pearson durchgeführt. Im folgenden Beispiel sieht man die Anwendung des Befehls auf die Auswertung einer Umfrage unter 100 Personen zum Thema „Zusammenhang zwischen Offenheit und der Bereitschaft zu

Mobile Payment“. Dargestellt ist eine Auswahl von sechs Items zur Offenheit.



Man erkennt, dass zwei Items invers kodiert sind und eine negative Korrelation zu den anderen Items aufweisen. Die Zeichen hinter der Korrelation bedeuten:

.	p-Wert kleiner als 10%	*	p-Wert kleiner als 5%
**	p-Wert kleiner als 1%	***	p-Wert kleiner als 0,1%

In der Übungsaufgabe 2.5 werden die Ergebnisse für eine Kombination von Items nachgerechnet.

Ü 2, ABC 6