Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Analyse von Zusammenhängen

χ^2-Homogenitätstest

In Verallgemeinerung der bereits besprochenen Hypothesen wird im Folgenden untersucht, ob m empirische Multinomialverteilungen identisch sind

- Die Multinomialverteilungen können entweder unmittelbar durch ein kategoriales Merkmal definiert sein oder sich (wie oben dargestellt) durch Gruppierung für ein (quasi-)stetiges Merkmal ergeben

Hierzu betrachten wir m unabhängige Stichproben mit einem Umfang von jeweils n_j, $j = 1, \ldots, m$ (n_j ist hier fest, kann aber prinzipiell auch als Zufallsgröße aufgefasst werden)

Die Zufallsvariable X_{ji} mit $j = 1, \ldots, m$ und $i = 1, \ldots, n_j$ beschreibt dann das Ergebnis der i-ten Untersuchungseinheit in der j-ten Stichprobe

Wir nehmen an, dass die Menge $\{1, \ldots, r\}$ den Wertevorrat der Zufallsvariablen X_{ji} bildet, für alle $j = 1, \ldots, m$ und $i = 1, \ldots, n_j$

Zufallsvariablen

Mit Y_{jk} bezeichnen wir die Zufallsvariable, die für die Stichprobe $j = 1, \ldots, m$ die Anzahl der Fälle angibt, in denen der Wert $k \in \{1, \ldots, r\}$ auftritt, d. h.

$$Y_{jk} = \text{Anzahl der Fälle, in denen } X_{ji} = k$$

Dann gilt

$$\sum_{k=1}^{r} Y_{jk} = n_j$$

und

$$\sum_{j=1}^{m} \sum_{k=1}^{r} Y_{jk} = \sum_{j=1}^{m} n_j = n$$

Die Realisierung der Zufallsvariablen Y_{jk} wird mit h_{jk} bezeichnet

Kontingenztabelle

Damit ergibt sich ein Schema, das man als Kontingenztabelle interpretieren kann, wobei die Zeilen durch die einzelnen Stichproben gebildet werden

	1	2	$\cdots$	k	$\cdots$	r	
1	Y_{11}	Y_{12}	$\cdots$	Y_{1k}	$\cdots$	Y_{1r}	n_1
2	Y_{21}	Y_{22}	$\cdots$	Y_{2k}	$\cdots$	Y_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
j	Y_{j1}	Y_{j2}	$\cdots$	Y_{jk}	$\cdots$	Y_{jr}	n_j
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m	Y_{m1}	Y_{m2}	$\cdots$	Y_{mk}	$\cdots$	Y_{mr}	n_m
	$Y_{.1}$	$Y_{.2}$	$\cdots$	$Y_{.k}$	$\cdots$	$Y_{.r}$	n

Kontingenztabelle

Entsprechend hat man für die beobachteten absoluten Häufigkeiten die Kontingenztabelle mit den Stichprobenumfängen als Zeilensummen und der bereits in Statistik I eingeführten "Punkt-Notation" für die Spaltensummen

	1	2	$\cdots$	k	$\cdots$	r	
1	h_{11}	h_{12}	$\cdots$	h_{1k}	$\cdots$	h_{1r}	n_1
2	h_{21}	h_{22}	$\cdots$	h_{2k}	$\cdots$	h_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
j	h_{j1}	h_{j2}	$\cdots$	h_{jk}	$\cdots$	h_{jr}	n_j
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m	h_{m1}	h_{m2}	$\cdots$	h_{mk}	$\cdots$	h_{mr}	n_m
	$h_{.1}$	$h_{.2}$	$\cdots$	$h_{.k}$	$\cdots$	$h_{.r}$	n

Nullhypothese

Als Nullhypothese betrachten wir nun die Annahme der Homogenität der Verteilungen, d. h. die Verteilung des Merkmals ist über alle Stichproben dieselbe

Konkret formuliert man als Nullhypothese

$$H_0\colon\ P(X_{ji} = k) = p_k \text{ für alle } j = 1, \ldots, m,\ i = 1, \ldots, n_j \text{ und } k = 1, \ldots, r$$

Unter der Gültigkeit der Nullhypothese kann man die Stichproben zusammenfassen ("gepoolte Stichprobe") und erhält für alle $j = 1, \ldots, m$ und $i = 1, \ldots, n_j$ dann als Schätzfunktion für die Wahrscheinlichkeiten $P(X_{ji} = k)$

$$\hat{p}_k = \frac{h_{\cdot k}}{n}$$

Die Zufallsvariablen Y_{jk} sind für alle $j = 1, \ldots, m$ und $k = 1, \ldots, r$ binomialverteilt gemäß $B(n_j, p_k)$

Teststatistik

Unter der Nullhypothese gilt also für alle $j = 1, \ldots, m$ und $k = 1, \ldots, r$

$$E(Y_{jk}) = n_j \cdot p_k$$

Die den beobachteten Häufigkeiten h_{jk} gegenüber zu stellenden erwarteten Häufigkeiten werden aus den Daten dann über die Randhäufigkeiten geschätzt durch

$$n_j \cdot \hat{p}_k = \frac{n_j \cdot h_{\cdot k}}{n}$$

Die Pearsonsche Teststatistik berechnet man dann durch

$$\chi^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{\left(h_{jk} - \frac{n_j \, h_{\cdot k}}{n}\right)^2}{\frac{n_j \, h_{\cdot k}}{n}}$$

Festlegung der Anzahl der Freiheitsgrade

Insgesamt summiert man über alle $m \cdot r$ Zellen der betrachteten Kontingenztabelle, die Zellen sind aber nicht alle unabhängig voneinander

- Für alle Stichproben $j = 1, \ldots, m$ gilt

$$Y_{j1} + \ldots + Y_{jr} = n_j,$$

  sodass eine Spalte jeweils determiniert ist, wenn die Werte in den übrigen $r - 1$ Spalten bekannt sind

- Für die Berechnung der Pearsonschen Teststatistik setzen wir die Spaltensummen $h_{.k}$ für alle $k = 1, \ldots, r$ als bekannt voraus, sodass auch hier eine Zeile jeweils determiniert ist, wenn die Werte in den übrigen $m - 1$ Zeilen bekannt sind

Man hat also lediglich $(m - 1)(r - 1)$ Freiheitsgrade

χ^2-Homogenitätstest

Annahmen

- $X_{j1}, \ldots, X_{jn_j}$ unabhängig und jeweils identisch multinomialverteilt (Wertevorrat $\{1, \ldots, r\}$) für alle voneinander unabhängigen Stichproben $j = 1, \ldots, m$

Hypothesen

- H_0: $P(X_{ji} = k) = p_k$ für alle $j = 1, \ldots, m$, $i = 1, \ldots, n_j$ und $k = 1, \ldots, r$
- H_1: $P(X_{ji} = k) \neq P(X_{j'i} = k)$ für mindestens ein Paar $j \neq j'$ aus $\{1, \ldots, m\}$ und ein $k \in \{1, \ldots, r\}$

Teststatistik

$$\chi^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{(Y_{jk} - n_j\, p_k)^2}{n_j\, p_k}$$

mit Y_{jk} gleich der Anzahl der X_{ji} mit $X_{ji} = k$ $(i = 1, \ldots, n_j)$

χ^2-Homogenitätstest

Verteilung der Teststatistik unter H_0

- Für hinreichend großen Stichprobenumfang n (also für $n \to \infty$) approximativ Chi-Quadrat-verteilt mit $(m-1)(r-1)$ Freiheitsgraden
- Die Approximation ist hinreichend gut für
  - $n_j\, p_k \geq 1$ für alle $j = 1, \ldots, m$ und $k = 1, \ldots, r$
  - $n_j\, p_k \geq 5$ für mindestens 80% der Fälle

Ablehnungsbereich

$$\chi^2 > \chi^2_{1-\alpha}((m-1)(r-1))$$

mit dem $(1-\alpha)$-Quantil $\chi^2_{1-\alpha}((m-1)(r-1))$ der Chi-Quadrat-Verteilung mit $(m-1)(r-1)$ Freiheitsgraden

Beispiel

Es soll geprüft werden, ob die Notenverteilung (1 bis 5) in drei parallelen Kursen identisch ist (Homogenität)

Zunächst bestimmen wir die Stichprobenumfänge n_1, n_2 und n_3, um mit ca. 80% Power bei $\alpha = 0.05$ nachweisen zu können, dass in Kurs 1 weniger gute Noten resultieren als in den anderen Kursen

Als inhaltlich minimal relevante Abweichung der bedingten Wahrscheinlichkeiten in Kurs 1 setzen wir

	p_1	p_2	p_3	p_4	p_5
Kurs 1	.06	.10	.32	.32	.20
Kurs 2	.16	.22	.24	.22	.16
Kurs 3	.16	.22	.24	.22	.16

Beispiel

Damit wird folgende Powersimulation durchgeführt

```
n1 <- n2 <- n3 <- 100
pval <- replicate(2000, {
  x <- cbind(   # bedingte Wahrscheinlichkeiten
        rmultinom(1, size = n1, prob = c(.06, .10, .32, .32, .20)),
        rmultinom(1, size = n2, prob = c(.16, .22, .24, .22, .16)),
        rmultinom(1, size = n3, prob = c(.16, .22, .24, .22, .16))
  )
  chisq.test(x)$p.value
})
mean(pval < 0.05)
```

Beispiel

Anschließend werden die Noten von Studierenden in den drei Kursen erhoben

	Note					
	1	2	3	4	5	
Kurs 1	4	10	33	35	20	102
Kurs 2	8	20	28	27	16	99
Kurs 3	12	23	26	18	25	104
	24	53	87	80	61	305

Mit diesen Daten wird nun die Nullhypothese der Homogenität getestet ($\alpha = 0.05$)

Beispiel

Für die gegebenen Noten der Studierenden schätzen wir die Wahrscheinlichkeiten p_k mit $k = 1, \ldots, 5$ durch

$$\hat{p}_1 = \frac{h_{.1}}{n} = \frac{24}{305} = 0.0787$$

$$\hat{p}_2 = \frac{h_{.2}}{n} = \frac{53}{305} = 0.1738$$

$$\hat{p}_3 = \frac{h_{.3}}{n} = \frac{87}{305} = 0.2852$$

$$\hat{p}_4 = \frac{h_{.4}}{n} = \frac{80}{305} = 0.2623$$

$$\hat{p}_5 = \frac{h_{.5}}{n} = \frac{61}{305} = 0.2000$$

Beispiel

Als erwartete Häufigkeiten für die Note 1 in Kurs 1 berechnet man

$$n_1 \cdot \hat{p}_1 = \frac{n_1 \cdot h_{\cdot 1}}{n} = \frac{102 \cdot 24}{305} = 8.0262$$

Die gesamte Tabelle der erwarteten Häufigkeiten lautet dann
(Approximationsbedingungen sind erfüllt!)

	Note					
	1	2	3	4	5	
Kurs 1	8.0262	17.7246	29.0951	26.7541	20.4000	102
Kurs 2	7.7902	17.2033	28.2393	25.9672	19.8000	99
Kurs 3	8.1836	18.0721	29.6656	27.2787	20.8000	104
	24	53	87	80	61	305

Beispiel

Berechnung der Pearsonschen Teststatistik ergibt

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{m}\sum_{k=1}^{r}\frac{(h_{jk} - n_j\,\hat{p}_k)^2}{n_j\,\hat{p}_k}\\ &= \frac{(4-8.0262)^2}{8.0262} + \frac{(10-17.7246)^2}{17.7246} + \ldots + \frac{(25-20.8000)^2}{20.8000}\\ &= 17.27\end{aligned}$$

Mit $(m-1)(r-1) = (3-1)(5-1) = 8$ Freiheitsgraden bestimmt man als 0.95-Quantil den Wert $\chi^2_{0.95}(8) = 15.51$

Die Nullhypothese wird somit abgelehnt und die Verteilungen können nicht als homogen betrachtet werden

Homogenitätstest mit Funktion `chisq.test()`

Eingabe

```
tab <- read.table(text = "
   4   10   33   35   20
   8   20   28   27   16
  12   23   26   18   25
")
chisq.test(tab)
```

Als Ausgabe erhält man

```
Pearson's Chi-squared test

data:  tab
X-squared = 17.273, df = 8, p-value = 0.0274
```

Da der *p*-Wert kleiner als die vorgegebene Irrtumswahrscheinlichkeit α ist, wird die Nullhypothese abgelehnt und die Verteilungen können nicht als homogen betrachtet werden

χ^2-Unabhängigkeitstest

Den Ausgangspunkt bildet eine unabhängige Stichprobe von Paaren (X_i, Y_i) mit $i = 1, \ldots, n$

Beschreiben die Zufallsvariablen kategoriale oder kategorisierte Merkmale, dann kann man die Beobachtungen in Form einer Kontingenztabelle darstellen

Als Wertevorrat betrachten wir hierfür

$$X_i(\Omega) = \{1, \ldots, m\}$$

und

$$Y_i(\Omega) = \{1, \ldots, r\}$$

für alle $i = 1, \ldots, n$

Kontingenztabelle

Für die beobachteten absoluten Häufigkeiten erhält man dann die Kontingenztabelle

		y_i						
		1	2	$\cdots$	k	$\cdots$	r	
	1	h_{11}	h_{12}	$\cdots$	h_{1k}	$\cdots$	h_{1r}	$h_{1.}$
	2	h_{21}	h_{22}	$\cdots$	h_{2k}	$\cdots$	h_{2r}	$h_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	j	h_{j1}	h_{j2}	$\cdots$	h_{jk}	$\cdots$	h_{jr}	$h_{j.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	m	h_{m1}	h_{m2}	$\cdots$	h_{mk}	$\cdots$	h_{mr}	$h_{m.}$
		$h_{.1}$	$h_{.2}$	$\cdots$	$h_{.k}$	$\cdots$	$h_{.r}$	$h_{..} = n$

Beispiel

Wir betrachten den Zusammenhang zwischen Alter (Merkmal X) von Autofahrern und der Anzahl der Unfälle (Merkmal Y), in die sie verwickelt sind

j	1	2	3	4	5
x_j	18-30	31-40	41-50	51-60	>60

k	1	2	3	4
y_i	0	1	2	>2

Es ergibt sich folgende Kontingenztabelle

		y_i				
		0	1	2	> 2	
x_i	18-30	748	74	31	9	862
	31-40	821	60	25	10	916
	41-50	786	51	22	6	865
	51-60	720	66	16	5	807
	>60	672	50	15	7	744
		3747	301	109	37	4194

Nullhypothese

Die Hypothese

$$H_0\text{: Die Merkmale } X \text{ und } Y \text{ sind unabhängig}$$

nimmt für kategoriale Merkmale eine sehr einfache Form an

Man hat die Nullhypothese

$$H_0\colon \; P(X_i = j, Y_i = k) = P(X_i = j) \cdot P(Y_i = k)$$
$$\text{für alle } j, k \text{ und } i = 1, \ldots, n$$

Für die Kontingenztabelle bedeutet diese Hypothese, dass sich die gemeinsame Auftretenswahrscheinlichkeit in jeder einzelnen Zelle als Produkt der entsprechenden Randwahrscheinlichkeiten ergibt

Nullhypothese

Mit den Definitionen

$$p_{jk} = P(X_i = j, Y_i = k)$$
$$p_{j.} = P(X_i = j)$$
$$p_{.k} = P(Y_i = k)$$

lässt sich die Nullhypothese dann formulieren durch

$$H_0: \; p_{jk} = p_{j.} \cdot p_{.k} \text{ für alle } j, k$$

Zur Schätzung der unter der Nullhypothese für jede Zelle (j, k) erwarteten Wahrscheinlichkeit geht man von den als fest vorgegeben betrachteten Randsummen $h_{j.}, h_{.k}$ aus

Erwartete Wahrscheinlichkeiten

Die Randwahrscheinlichkeiten werden geschätzt durch

$$\hat{p}_{j.} = \frac{h_{j.}}{n}$$
$$\hat{p}_{.k} = \frac{h_{.k}}{n}$$

Die unter H_0 erwartete Wahrscheinlichkeit einer Beobachtung in Zelle (j, k) bestimmt man dann durch

$$\hat{p}_{jk} = \hat{p}_{j.} \cdot \hat{p}_{.k}$$

Die zur Beobachtung h_{jk} gehörige Zufallsvariable H_{jk} ist unter H_0 binomialverteilt mit

$$H_{jk} \sim B(n, p_{j.} \cdot p_{.k})$$

Erwartete Häufigkeiten

Die unter H_0 zu erwartende Häufigkeit $E(H_{jk}) = n \cdot p_{jk}$ für eine Beobachtung in Zelle (j, k) wird daher geschätzt durch

$$e_{jk} = n \cdot \hat{p}_{jk} = n \cdot \hat{p}_{j\cdot} \cdot \hat{p}_{\cdot k} = n \cdot \frac{h_{j\cdot}}{n} \cdot \frac{h_{\cdot k}}{n} = \frac{h_{j\cdot}\, h_{\cdot k}}{n}$$

		y_i: 1	2	$\cdots$	k	$\cdots$	r	
x_i	1	$\frac{h_{1\cdot}\, h_{\cdot 1}}{n}$	$\frac{h_{1\cdot}\, h_{\cdot 2}}{n}$	$\cdots$	$\frac{h_{1\cdot}\, h_{\cdot k}}{n}$	$\cdots$	$\frac{h_{1\cdot}\, h_{\cdot r}}{n}$	$h_{1\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	j	$\frac{h_{j\cdot}\, h_{\cdot 1}}{n}$	$\frac{h_{j\cdot}\, h_{\cdot 2}}{n}$	$\cdots$	$\frac{h_{j\cdot}\, h_{\cdot k}}{n}$	$\cdots$	$\frac{h_{j\cdot}\, h_{\cdot r}}{n}$	$h_{j\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	m	$\frac{h_{m\cdot}\, h_{\cdot 1}}{n}$	$\frac{h_{m\cdot}\, h_{\cdot 2}}{n}$	$\cdots$	$\frac{h_{m\cdot}\, h_{\cdot k}}{n}$	$\cdots$	$\frac{h_{m\cdot}\, h_{\cdot r}}{n}$	$h_{m\cdot}$
		$h_{\cdot 1}$	$h_{\cdot 2}$	$\cdots$	$h_{\cdot k}$	$\cdots$	$h_{\cdot r}$	n

Teststatistik

Man betrachtet nun wiederum die Diskrepanz zwischen den beobachteten Häufigkeiten h_{jk} und den unter H_0 zu erwartenden Häufigkeiten e_{jk} im Rahmen der Pearsonschen Teststatistik

$$\chi^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{(h_{jk} - e_{jk})^2}{e_{jk}}$$

Damit sind χ^2-Unabhängigkeitstest und χ^2-Homogenitätstest formal äquivalent

- Die formalen Berechnungen sind identisch, lediglich die Interpretation der Ergebnisse unterscheidet sich
- Beide Tests sind wiederum spezielle χ^2-Anpassungstests über die Zellen einer Kontingenztabelle, bei denen sich die unter der H_0 erwarteten Häufigkeiten aus Annahmen zur Homogenität bzw. Unabhängigkeit der unterliegenden Verteilungen ergeben

χ^2-Unabhängigkeitstest

Annahmen

- Unabhängige Paare (X_i, Y_i), $i = 1, \ldots, n$, mit $X_i \in \{1, \ldots, m\}$ und $Y_i \in \{1, \ldots, r\}$

Hypothesen

- H_0: $P(X_i = j, Y_i = k) = P(X_i = j) \cdot P(Y_i = k)$ für alle j, k und $i = 1, \ldots, n$
- H_1: $P(X_i = j, Y_i = k) \neq P(X_i = j) \cdot P(Y_i = k)$ für mindestens ein Paar j, k und $i = 1, \ldots, n$

Teststatistik

$$\chi^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{(H_{jk} - n\, p_{j.}\, p_{.k})^2}{n\, p_{j.}\, p_{.k}}$$

mit H_{jk} gleich der Anzahl der Fälle, in denen $X_i = j$ und $Y_i = k$, sowie $p_{j.} = P(X_i = j)$, $p_{.k} = P(Y_i = k)$

χ^2-Unabhängigkeitstest

Verteilung der Teststatistik unter H_0

- Für hinreichend großen Stichprobenumfang n (also für $n \to \infty$) approximativ Chi-Quadrat-verteilt mit $(m-1)(r-1)$ Freiheitsgraden
- Die Approximation ist hinreichend gut für
  - $n\, p_{j\cdot}\, p_{\cdot k} \geq 1$ für alle $k = 1, \ldots, r$
  - $n\, p_{j\cdot}\, p_{\cdot k} \geq 5$ für mindestens 80% der Fälle

Ablehnungsbereich

$$\chi^2 > \chi^2_{1-\alpha}((m-1)(r-1))$$

mit dem $(1-\alpha)$-Quantil $\chi^2_{1-\alpha}((m-1)(r-1))$ der Chi-Quadrat-Verteilung mit $(m-1)(r-1)$ Freiheitsgraden

Beispiel

Zunächst bestimmen wir den Stichprobenumfang n, um mit ca. 85% Power bei $\alpha = 0.1$ eine spezifizierte Häufung von Unfällen in der jüngsten und ältersten Gruppe nachweisen zu können

```
n <- 4000          # Anzahl der Paare
pval <- replicate(2000, {
  dat <- cbind(expand.grid(alter = 1:5, unfall = 1:4),
           x = rmultinom(1, size = n,
                    # Verbundwahrscheinlichkeiten
                    # 18-30 31-40 41-50 51-60 >60
             prob = c(.167, .192, .192, .192, .167,    #  0
                      .015, .010, .010, .010, .015,    #  1
                      .004, .004, .004, .004, .004,    #  2
                      .002, .002, .002, .002, .002)))  # >2
  tab <- xtabs(x ~ alter + unfall, data = dat)
  chisq.test(tab)$p.value
})
mean(pval < 0.1)
```

Beispiel

Mit den obigen Daten berechnet man für die erwarteten Häufigkeiten e_{jk}

		y_i				
		0	1	2	> 2	
x_i	18-30	770.13	61.87	22.40	7.60	862
	31-40	818.37	65.74	23.81	8.08	916
	41-50	772.81	62.08	22.48	7.63	865
	51-60	720.99	57.92	20.97	7.12	807
	$>$60	664.70	53.40	19.34	6.56	744
		3747	301	109	37	4194

Beispiel

Berechnung der Pearsonschen Teststatistik ergibt

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{m}\sum_{k=1}^{r}\frac{(h_{jk}-e_{jk})^2}{e_{jk}}\\ &= \frac{(748-770.13)^2}{770.13}+\frac{(74-61.87)^2}{61.87}+\ldots+\frac{(7-6.56)^2}{6.56}\\ &= 14.40\end{aligned}$$

Für vorgegebene Irrtumswahrscheinlichkeit $\alpha = 0.1$ ergibt sich mit $(m-1)(r-1) = (5-1)(4-1) = 12$ Freiheitsgraden das 0.9-Quantil $\chi^2_{0.9}(12) = 18.55$

Die Nullhypothese wird beibehalten, sodass mit den Daten kein Zusammenhang zwischen den beiden Merkmalen Alter und Anzahl der Unfälle nachgewiesen werden kann

Unabhängigkeitstest mit Funktion `chisq.test()`

Eingabe

```
tab <- read.table(text = "
  748  74  31  9
  821  60  25 10
  786  51  22  6
  720  66  16  5
  672  50  15  7
")
chisq.test(tab)
```

Ausgabe

```
Pearson's Chi-squared test

data:  tab
X-squared = 14.395, df = 12, p-value = 0.276
```

Korrelation bei normalverteilten Merkmalen

Den Ausgangspunkt bilden zwei gemeinsam normalverteilte Merkmale für die eine Stichprobe (X_i, Y_i) mit $i = 1, \ldots, n$ vorliegt

Für die Merkmale soll geprüft werden, ob sie unkorreliert sind, also

$$\rho_{XY} = 0$$

gilt, oder allgemeiner, ob die Korrelation einen bestimmten Wert

$$\rho_{XY} = \rho_0$$

annimmt

- Beachte: Zwei gemeinsam normalverteilte Zufallsvariablen sind unkorreliert genau dann, wenn sie stochastisch unabhängig sind (gilt nicht allgemein, s. Statistik I)

Teststatistik

In diesem Fall wird also mit der Unkorreliertheit auch die Unabhängigkeit der Merkmale überprüft

Das Testverfahren baut auf dem empirischen Produkt-Moment-Korrelationskoeffizienten auf

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Basierend auf dem Produkt-Moment-Korrelationskoeffizienten wird eine Teststatistik entwickelt, deren Verteilung unter der Nullhypothese bekannt ist

- Hierzu unterscheidet man den allgemeinen Fall ρ_{XY} von dem Spezialfall $\rho_{XY} = 0$

Korrelation bei normalverteilten Merkmalen, $\rho_{XY} = 0$

Annahmen

- Unabhängige gemeinsam normalverteilte Stichprobenvariablen (X_i, Y_i), $i = 1, \ldots, n$

Hypothesen

1. $H_0\colon\ \rho_{XY} = 0 \qquad H_1\colon\ \rho_{XY} \neq 0$
2. $H_0\colon\ \rho_{XY} \geq 0 \qquad H_1\colon\ \rho_{XY} < 0$
3. $H_0\colon\ \rho_{XY} \leq 0 \qquad H_1\colon\ \rho_{XY} > 0$

Teststatistik für $\rho_{XY} = 0$

$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}$$

Korrelation bei normalverteilten Merkmalen, $\rho_{XY} = 0$

Verteilung der Teststatistik unter $\rho_{XY} = 0$

- $T \sim t(n-2)$

Ablehnungsbereich

1. $|T| > t_{1-\alpha/2}(n-2)$
2. $T < -t_{1-\alpha}(n-2)$
3. $T > t_{1-\alpha}(n-2)$

mit dem p-Quantil $t_p(n-2)$ der t-Verteilung mit $df = n-2$ Freiheitsgraden

Korrelation bei normalverteilten Merkmalen, $\rho_{XY} = \rho_0$

Annahmen

- Unabhängige gemeinsam normalverteilte Stichprobenvariablen (X_i, Y_i), $i = 1, \ldots, n$

Hypothesen

1. $H_0\colon\ \rho_{XY} = \rho_0 \qquad H_1\colon\ \rho_{XY} \neq \rho_0$
2. $H_0\colon\ \rho_{XY} \geq \rho_0 \qquad H_1\colon\ \rho_{XY} < \rho_0$
3. $H_0\colon\ \rho_{XY} \leq \rho_0 \qquad H_1\colon\ \rho_{XY} > \rho_0$

Teststatistik (generelles ρ_0)

$$Z = \frac{1}{2}\left(\log\frac{1+r_{XY}}{1-r_{XY}} - \log\frac{1+\rho_0}{1-\rho_0}\right)\sqrt{n-3}$$

Korrelation bei normalverteilten Merkmalen, $\rho_{XY} = \rho_0$

Verteilung der Teststatistik unter $\rho_{XY} = \rho_0$

- Z für $n > 25$ approximativ standardnormalverteilt

Ablehnungsbereich

1. $|Z| > z_{1-\alpha/2}$
2. $Z < -z_{1-\alpha}$
3. $Z > z_{1-\alpha}$

mit dem p-Quantil z_p der $N(0,1)$-Verteilung

Beispiel

Der Zusammenhang zwischen der Prognose des Sachverständigenrates und dem tatsächlich beobachteten Wirtschaftswachstum soll geprüft werden

Zunächst bestimmen wir den Stichprobenumfang n, um mit ca. 80% Power bei $\alpha = 0.05$ nachweisen zu können, dass unter Annahme der Normalverteilung die Korrelation $\rho_{XY} = 0.6$ beträgt, H_0: $\rho_{XY} = 0$

```
n <- 20  # Anzahl der Jahre
s <- 2; r <- 0.6; sxy <- r*s*s
pval <- replicate(2000, {
  x <- MASS::mvrnorm(n, mu = c(2, 2),
                     Sigma = matrix(c(s^2, sxy,
                                      sxy, s^2), 2, 2))
  cor.test(x[, 1], x[, 2])$p.value
})
mean(pval < 0.05)
```

Beispiel

Anschließend betrachten wir die Prognose des Sachverständigenrates (Merkmal X) und das tatsächlich beobachtete Wirtschaftswachstum (Merkmal Y) in den Jahren 1975-1994

Jahr	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
x_i	2.0	4.5	4.5	3.5	3.8	2.8	0.5	0.5	1.0	2.5
y_i	-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
x_i	3.0	3.0	2.0	1.5	2.5	3.0	3.5	2.5	0.0	0.0
y_i	2.5	2.5	1.7	3.4	4.0	4.6	3.4	1.5	-1.9	2.3

Damit soll unter Annahme der Normalverteilung die Hypothese $\rho_{XY} = 0$ überprüft werden ($\alpha = 0.05$)

Beispiel

Mit den Schätzwerten

$$Var(x) = 1.8954, \quad Var(y) = 5.2413$$
$$Cov(x, y) = 2.0186, \quad r_{xy} = 0.6404$$

berechnet man für die Teststatistik

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2} = \frac{0.6404}{\sqrt{1 - 0.6404^2}} \sqrt{20-2} = 3.54$$

Als kritisches Quantil bestimmt man $t_{0.975}(18) = 2.10$, sodass die Nullhypothese abgelehnt wird

Die Daten sprechen also für einen Zusammenhang zwischen prognostiziertem und tatsächlichem Wirtschaftswachstum

Test des Korrelationskoeffizienten mit cor.test()

Eingabe

```
x <- 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)
y <- 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(x, y)
```

Ausgabe

```
    Pearson's product-moment correlation

data:  x and y
t = 3.5378, df = 18, p-value = 0.002351
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2761733 0.8438096
sample estimates:
      cor
0.640426
```