2.3 Kontingenztafeln und Chi-Quadrat-Test

Die Voraussetzungen an die Daten in diesem Kapitel sind dieselben, wie im vorangegangenen Kapitel, nur dass die Stichprobe hier aus Realisierungen von kategoriellen Zufallsvariablen besteht. D.h. wir gehen von zwei kategoriellen Zufallsvariablen X und Y aus, wobei X die Werte (Kategorien) $a_1, a_2, \ldots, a_r$ und Y die Werte $b_1, b_2, \ldots, b_s$ annehmen kann. Das Datenniveau muss hier dann sogar nur nominal sein.

Fasst man eine Datenzeile in einem Wertepaar zusammen, so kann man die (absolute) Häufigkeit n_{ij} bestimmen, mit der das Wertepaar (a_i, b_j) in der Stichprobe aufgetreten ist. Diese Häufigkeiten, die die Grundlage für den Chi-Quadrat-Test auf Unabhängigkeit bilden, kann man in einer so genannten Kontingenztafel (oder Kreuztabelle) darstellen:

		Y			
		b_1	**b_2**	**b_s**	**Zeilensumme**
X	**a_1**	**n_{11}**	**n_{12}**	**N_{1s}**	**$n_{1.}$**
	a_2	**n_{21}**	**n_{22}**	**N_{2s}**	**$n_{2.}$**
	.				
	.				
	a_r	**n_{r1}**	**n_{r2}**	**n_{rs}**	**$n_{r.}$**
Spaltensumme		**$n_{.1}$**	**$n_{.2}$**	**$n_{.s}$**	**N**

Die Zeilen- und Spaltensummen ergeben sich hierbei aus

$$n_{i.} = \sum_{j=1}^{s} n_{ij} \quad \text{und} \quad n_{.j} = \sum_{i=1}^{r} n_{ij} .$$

Betrachtet man nun die Häufigkeiten n_{ij} als Realisierungen der Zufallsvariablen N_{ij}, die als Komponenten eines Zufallsvektors $(N_{11},N_{12},\ldots,N_{1s},N_{21,\ldots,}N_{rs})^T$ mit einer r·s-dimensionalen Multinomialverteilung aufgefasst werden, so ergeben sich für die Wahrscheinlichkeiten $P(X = a_i) = p_i$, $P(Y = b_j) = q_j$ und $P(X = a_i, X = b_j) = p_{ij}$ die folgenden Maximum-Likelihood-Schätzer (was die entsprechenden relativen Häufigkeiten sind):

$$\hat{p}_i = \frac{n_{i.}}{n},\ \hat{q}_j = \frac{n_{.j}}{n} \text{ und } \hat{p}_{ij} = \frac{n_{ij}}{n}\ .$$

Bemerkung: Die von uns in allen Beispielen verwendeten Zahlenwerte entstammen keiner realen Studie, sondern sind für das jeweilige Beispiel ad hoc gewählt worden.

Um die Grundidee des Chi-Quadrat-Tests (oder χ^2-Tests) besser verstehen zu können, kommen wir nun zu drei Beispielen:

Beispiel a:
Mit Hilfe einer Kreuztabelle soll ein möglicher Zusammenhang zwischen den Antworten auf die Frage A und den Antworten auf die Frage B untersucht werden, die beide mit "ja" oder "nein" beantwortet werden konnten. Es liegt eine Stichprobe vom Umfang n = 45 vor.

		Frage B		
		ja	**nein**	**Zeilensumme**
Frage A	**Ja**	**9**	**6**	**15**
	nein	**5**	**15**	**20**
Spaltensumme		**14**	**21**	**45**

In der obigen Tabelle ist beispielsweise zu sehen, dass 5 Personen die Frage A mit „nein“ und die Frage B mit „ja“ beantwortet.

Beispiel b:
Es soll ein möglicher Zusammenhang untersucht werden zwischen den Behandlungsmethoden A, B und C eines Arztes und dem Heilungserfolg. Es liegt eine Stichprobe vom Umfang n = 320 vor.

		Heilungserfolg		
		Erfolg	**kein Erfolg**	**Zeilensumme**
Beh.	**A**	**98**	**22**	**120**
	B	**46**	**44**	**90**
	C	**40**	**70**	**110**
Spaltensumme		**184**	**136**	**320**

Beispiel c:
Es wurden n = 36 Personen befragt, ob sie Raucher seien, bzw. ob sie regelmäßig Alkohol konsumieren. Es soll festgestellt werden, ob die Variable “Rauchverhalten” mit den beiden Kategorien “Raucher” und “Nichtraucher” von der Variablen “Trinkverhalten” mit den beiden Kategorien “Abstinenzler” und “Alkoholkonsument” unabhängig ist.

Es waren in dieser Stichprobe 9 Alkoholkonsumenten und 27 Abstinenzler. Festgestellt wurden 12 Raucher (und daher 36 – 12 = 24 Nichtraucher). Unter den 12 Rauchern waren 8 Abstinenzler (und daher 12 - 8 = 4 Alkoholkonsumenten). Bei den 24 Nichtrauchern waren 19 Abstinenzler und 5 Alkoholkonsumenten.

Die vollständige Tabelle mit den zugehörigen Randsummen sehen Sie hier:

	Raucher	Nichtraucher	Zeilensumme
Alkohol-Konsument	**4**	**5**	**9**
Abstinenzler	**8**	**19**	**27**
Spaltensumme	**12**	**24**	**36**

Wenn nun Rauchverhalten und Trinkverhalten unabhängig von einander wären, so gäbe es unter den Abstinenzlern in dieser Stichprobe (das sind 27/36 = 3/4 aller Personen) einen ebenso großen Anteil an Rauchern, wie in der Gesamtstichprobe, wo 12/36 = 1/3 aller Personen rauchen. Es müssten daher 1/3 von 3/4 aller 36 Personen (das sind absolut $1/3 \cdot 3/4 \cdot 36 = 9$ Personen) rauchende Abstinenzler sein. Diese Anzahl wäre bei Unabhängigkeit der beiden Variablen zu erwarten. Tatsächlich festgestellt wurden 8 rauchende Abstinenzler, was in diesem Beispiel keine große Abweichung darstellt. Entsprechende Überlegungen wären allerdings nun für die anderen Zelleninhalte durchzuführen.

Wären die summierten (quadratischen und bezüglich der erwarteten Häufigkeiten relativierten) Abweichungen aller Zelleninhalte der konkreten (aus den Beobachtungen aufgestellten) Kreuztabelle von den bei Unabhängigkeit erwarteten Werten "groß", so spräche das gegen eine Unabhängigkeit. Diese Abweichungen könnten allerdings zufallsbedingt groß sein, da es sich um eine Stichprobe handelt.

Der Chi-Quadrat-Test auf Unabhängigkeit stellt ein statistisches Verfahren dar, das aus diesen Abweichungen eine Prüfgröße berechnet, anhand derer die Hypothese der Unabhängigkeit für die Gesamtpopulation bzw. Grundgesamtheit, die hinter der Stichprobe steht, überprüft werden kann. Der Name "Chi-Quadrat-Test" rührt von der Tatsache her, dass die Prüfgröße unter der Hypothese der Unabhängigkeit eine Realisierung einer Chi-Quadrat-verteilten zufälligen Größe ist.

Bevor wir nun auf Details eingehen, ist es hier ist zunächst wichtig, sich die oben gezeigte Vorgehensweise bei der Berechnung der erwarteten Zelleninhalte klar zu machen. Wir wollen dies nun etwas mathematischer fassen:

Zwei kategorielle Zufallsvariablen X und Y sind genau dann stochastisch unabhängig, wenn

$$P(X = a_i, Y = b_j) = P(X = a_i) \cdot P(Y = b_j)$$

für alle i = 1, 2, …, r und j = 1, 2, …, s gilt. Das heißt in Worten: Die Wahrscheinlichkeit, dass die Zufallsvariable X den Wert a_i annimmt und zudem die Zufallsvariable Y den Wert b_j, ist gleich dem Produkt der Wahrscheinlichkeit, dass die Zufallsvariable X den Wert a_i annimmt, mit der Wahrscheinlichkeit, dass die Zufallsvariable Y den Wert b_j annimmt.

Somit muss bei Unabhängigkeit der beiden Zufallsvariablen X und Y für die relativen Häufigkeiten als Schätzungen für die Wahrscheinlichkeiten gelten:

$$\frac{n_{ij}}{n} \approx \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}$$

Dementsprechend gilt für die bei Unabhängigkeit erwarteten absoluten Häufigkeiten:

$$n_{ij} \approx \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} \cdot n = \frac{n_{i.} \cdot n_{.j}}{n}$$

Vergleichen Sie dies mit der anfangs gezeigten Berechnung der erwarteten absoluten Häufigkeiten im Beispiel des Rauch- und Trinkverhaltens.

Bei Unabhängigkeit der beiden Zufallsvariablen müssten die beobachteten Häufigkeiten n_{ij} in der Nähe der erwarteten Häufigkeiten

$$n'_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

liegen. Falls diese Abweichungen zu groß sind, so spricht dies gegen die Unabhängigkeit.

Aus den beobachteten und (den bei Unabhängigkeit) erwarteten Häufigkeiten lässt sich nun eine Prüfgröße z berechnen, die als Realisierung einer Zufallsvariablen Z betrachtet werden kann, welche unter der Unabhängigkeitshypothese H_0 asymptotisch χ^2 -verteilt ist, mit (r-1)·(s-1) Freiheitsgraden. Diese Prüfgröße errechnet sich aus den bezüglich der erwarteten Häufigkeiten relativierten Differenzenquadraten zwischen beobachteten (n_{ij}) und erwarteten (n'_{ij}) Häufigkeiten, welche dann aufsummiert werden:

$$z = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Gehen wir nun von den Hypothesen

H_0: X und Y sind unabhängig

gegen

H_1: X und Y sind abhängig

aus, so kann die Nullhypothese H_0 der stochastischen Unabhängigkeit auf dem Signifikanzniveau α verworfen werden, falls z größer als das (oder gleich dem) $(1-\alpha)$-Quantil der χ^2 -Verteilung mit $(r-1)\cdot(s-1)$ Freiheitsgraden ist, also wenn gilt:

Ist F die Verteilungsfunktion der χ^2-Verteilung mit $(r-1)\cdot(s-1)$ Freiheitsgraden, dann ist H_0 zu verwerfen für

$$z \geq F^{-1}(1 - \alpha) \quad \text{oder} \quad \text{p-Wert} = 1 - F(z) \leq \alpha \;.$$

1 - F(z) wird als (approximativer) p-Wert im Output ausgegeben. Ist dieser kleiner als das (oder gleich dem) gewählte Signifikanzniveau α, so kann die Nullhypothese verworfen werden. In diesem Fall wurde dann ein signifikanter Zusammenhang zwischen den Zufallsvariablen X und Y nachgewiesen, da die Nullhypothese H_0 zugunsten der Alternativhypothese H_1 verworfen werden kann.

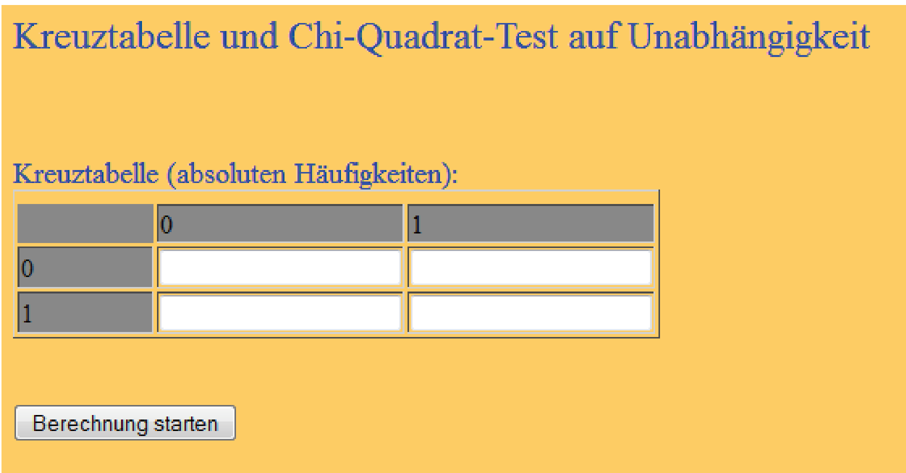
Wir geben die Werte der drei Beispiele bei Statistikpaket.de ein. Hier haben wir zwei Möglichkeiten. Wir beginnen mit Beispiel a und erklären, wie man die Daten im System eintragen kann. Die Daten für die Beispiele b und c können dann analog eingegeben werden.

Die Stichprobe zum Beispiel a könnte wie folgt eingeben werden:

v1 v2
ja ja
ja ja
……..
ja ja (insgesamt 9-mal ja, ja)
ja nein
ja nein
……..
ja nein (insgesamt 6-mal ja, nein)
……..
……..

Den Output erhalten Sie, wenn Sie auf den Button →**Zusammenhänge untersuchen** klicken**,** wobei Sie zuvor die Variablen v1 und v2 unter diesem Menüpunkt auswählen müssen. Danach können Sie →**Kreuztabelle und Chi-Quadrat-Test** wählen. Alternativ könnte man auch nach der Auswahl von →**Zusammenhänge untersuchen** auf den Link „Kreuztabelle selbst eingeben“ klicken und die Kreuztabelle direkt eingeben.

Oder Sie gehen auf die Seite www.Statistikpaket.de/Kreuztabelle/Kreuztabelle.html (hierhin gelangt man auch über die Startseite von www.Statistikpaket.de, wenn man unter „Untersuchung von Zusammenhänge“ den Punkt „Kreuztabelle und Chi-Quadrat-Test auf Unabhängigkeit“ wählt). Danach können Sie im Text „hier kann die Kreuztabelle direkt eingegeben werden“ auf den Linke „hier“ klicken. Dann müssen Sie im Beispiel a bei Spalten und Zeilenanzahl 2 eingeben und kommen dann zum folgenden Fenster:



Hier können Sie die Kreuztabelle direkt eingeben und gelangen dann zum Output, wenn Sie →**Berechnung starten** wählen.

Output des Beispiels a:

Kreuztabelle / Chi-Quadrat-Test auf Unabhängigkeit

Kreuztabelle mit absoluten Häufigkeiten:

	0	1	Summe
0	9	6	15
1	5	15	20
Summe	14	21	35

Die bei Unabhängigkeit erwarteten absoluten Häufigkeiten:

	0	1
0	6	9
1	8	12

Prüfgröße für den Chi-Quadrat-Test auf Unabhängigkeit:
4.375

Approximativer p-Wert (Freiheitsgrade Chi-Quadrat-Verteilung: 1):
0.0365

Auf einem Signifikanzniveau von 5% könnte die Nullhypothese der Unabhängigkeit verworfen werden, denn p-Wert $\approx 0{,}0365 \leq 0{,}05$.

Output im Beispiel b:

Kreuztabelle / Chi-Quadrat-Test auf Unabhängigkeit

Kreuztabelle mit absoluten Häufigkeiten:

	0	1	Summe
0	98	22	120
1	46	44	90
2	40	70	110
Summe	184	136	320

Die bei Unabhängigkeit erwarteten absoluten Häufigkeiten:

	0	1
0	69	51
1	51.75	38.25
2	63.25	46.75

Prüfgröße für den Chi-Quadrat-Test auf Unabhängigkeit:
50.291146761735

Approximativer p-Wert (Freiheitsgrade Chi-Quadrat-Verteilung: 2):
0

Wie zu sehen ist, kann im Beispielen b die Hypothese der Unabhängigkeit auf einem Signifikanzniveau von $\alpha = 1\%$ verworfen werden, da der p-Wert $\leq 0{,}01$ ist. Gegebenenfalls könnte hier sogar auf mit kleinerem Signifikanzniveau verworfen werden, wobei allerdings zu bedenken ist, dass das Signifikanzniveau bei „sauberer“ Arbeitsweise stets vor Beginn eines Tests festgelegt werden muss.

Output im Beispiel c:

Kreuztabelle / Chi-Quadrat-Test auf Unabhängigkeit

Kreuztabelle mit absoluten Häufigkeiten:

	0	1	Summe
0	4	5	9
1	8	19	27
Summe	12	24	36

Die bei Unabhängigkeit erwarteten absoluten Häufigkeiten:

	0	1
0	3	6
1	9	18

Prüfgröße für den Chi-Quadrat-Test auf Unabhängigkeit:
0.66666666666667

Approximativer p-Wert (Freiheitsgrade Chi-Quadrat-Verteilung: 1):
0.4142
In mindestens einer Zelle der Kreuztabelle ist die absolute Häufigkeit gering als 5!

Im Beispiel c kann auf keinem gängigen Signifikanzniveau die Nullhypothese verworfen werden. Hier ist zudem zu beachten, dass es in einer Zelle der Kreuztabelle eine absolute Häufigkeit von weniger als 5 gibt, was problematisch ist, da die asymptotische Verteilung verwendet wird!

Umsetzung des Beispiels a mit SAS:

```
data dat1;
input x y;
datalines;
0 0
0 0
0 0
0 0
0 0
0 0
0 0
0 0
0 0
0 1
0 1
0 1
0 1
0 1
0 1
1 0
1 0
1 0
1 0
1 0
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
1 1
run;
```

```
proc freq data = dat1;
table x*y / Chisq;
run;
```

Oder:

```
data dat1;
input anzahl x y;
datalines;
9  0 0
6  0 1
5  1 0
15 1 1
run;

proc freq data=dat1 order = data;
   tables x*y / chisq;
   weight anzahl;
run;
```

SAS-Output zur Prozedur FREQ:

Die Prozedur FREQ

Häufigkeit
Prozent
Row Pct
Col Pct

Tabelle von x nach y

x	**y**		
	0	**1**	**Summe**
0	9	6	15
	25.71	17.14	42.86
	60.00	40.00	
	64.29	28.57	
1	5	15	20
	14.29	42.86	57.14
	25.00	75.00	
	35.71	71.43	
Summe	14	21	35
	40.00	60.00	100.00

Statistiken für Tabelle von x nach y.

Statistik	DF	Wert	Prob
Chi-Quadrat	1	4.3750	0.0365
Likelihood-Ratio Chi-Quadrat	1	4.4271	0.0354
Kontinuitätskorr. Chi-Quad.	1	3.0382	0.0813
Mantel-Haenszel Chi-Quadrat	1	4.2500	0.0393
Phi-Koeffizient		0.3536	
Kontingenzkoeffizient		0.3333	
Cramers V		0.3536	

Exakter Test von Fisher	
Zelle (1,1) Häufigkeit (F)	9
Linksseitige Pr <= F	0.9930
Rechtsseitige Pr >= F	0.0404
Tabellenwahrscheinlichkeit (P)	0.0334
Zweiseitige Pr <= P	0.0796

Stichprobengröße = 35