

5.7 Chi-Quadrat Tests für diskrete Daten

Gegeben seien Zählraten wie $X_i =$ Anzahl Fälle in der Klasse i . Wie testen wir, ob diese Daten mit einem Modell konsistent sein?

5.7.1 Beispiele

(1) Ist die Anzahl der Tore in einem Fußballspiel eine poissonverteilte ZV mit $\lambda = 3$? Daten aus 18 Spielen aus der 1. und 2. Bundesliga stehen zur Verfügung:

7, 9, 2, 1, 7, 3, 5, 3, 6 und 2, 3, 5, 3, 5, 2, 3, 0, 5

(2) Gibt es eine geschlechtsspezifische Abneigung gegen das Studium an der Mathematisch-Naturwissenschaftlichen Fakultät der Uni Augsburg? Im Sommersemester 1999 waren die Zahlen

	Frauen	Männer
MNF	433	693
Rest	5505	4641

Methode

Wir berechnen die unter dem Modell zu erwartenden Werten und vergleichen sie mit den beobachteten Werten.

Beispiel (1) Anzahl Tore

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Sei e_i die erwartete Anzahl Spiele mit i Toren unter den 18 Spielen und o_i die beobachtete Zahl:

Tore	0	1	2	3	4	5	6	≥ 7
o_i	1	1	3	5	0	4	1	3
e_i	0.89	2.69	4.03	4.03	3.02	1.81	0.91	0.60

Heuristische Idee. Die (quadrierten) Abweichungen

$$(o_i - e_i)^2$$

messen den Unterschied zum Modell für eine Klasse, aber sie sind nicht gut untereinander vergleichbar. Wir wollen nicht die (o_i, e_i) Paare

$$(100, 110) \quad \text{und} \quad (10, 20)$$

gleich bewerten. Deshalb arbeiten wir mit den relativen Abweichungen

$$\frac{(o_i - e_i)^2}{e_i}$$

Wir summieren sie über alle Klassen, und verwenden die Teststatistik

$$T = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Für H_0 stellt sich heraus, dass unter gewissen Annahmen

$$T \sim \chi_7^2$$

wo 7 die Anzahl der Klassen minus 1 ("Anzahl Freiheitsgrade") ist. Wir werden H_0 verwerfen, wenn T groß ist.

$$P(T > 14.1) = 0.05 \quad P(T > 18.5) = 0.01$$

Für die Daten finden wir $T = 16.79$, so dass wir H_0 zu einem Testniveau von 5% verwerfen können, aber nicht zu einem Niveau von 1%.

Der Durchschnitt der Tore pro Spiel war $71/18 = 3.944$. Wenn wir einen Test mit diesem λ ausführen, finden wir $T = 7.03$, aber wir müssen die Benutzung des Schätzers berücksichtigen. Da ein Parameter geschätzt worden ist, verlieren wir einen Freiheitsgrad und nehmen $T \sim \chi_6^2$ an.

$$P(T > 12.6) = 0.05$$

H_0 wird in diesem Fall akzeptiert.

Beispiel (2) Fakultäten

H_0 : kein Geschlechtseinfluß — der Anteil in MNF entspricht dem Gesamtanteil der Frauen an Studierenden.

Insgesamt gibt es 5938 Frauen und 5334 Männer mit einem Frauenanteil von $\frac{5938}{11272}$. Unter H_0 erwarten wir

$$\frac{5938}{11272} * 1126$$

Frauen in MNF. Die erwarteten Werte für die Tabelle sind

	Frauen	Männer
MNF	593.17	532.83
Rest	5344.83	4801.17

Aus demselben heuristischen Argument wie vorher nehmen wir als Teststatistik

$$X^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Approximativ gilt

$$X^2 \sim \chi_{(r-1)(c-1)}^2$$

wo r die Anzahl von Zeilen ("rows") sei und c die Anzahl von Spalten ("columns"). $r = 1, c = 1 \Rightarrow X^2 \sim \chi_1^2$

$$X^2 = 101.5 \Rightarrow H_0 \text{ verwerfen.}$$

5.7.2 (Zweidimensionale) Kontingenztafeln im allgemeinen

Gegeben seien zwei Klassifikationen der Daten, $i = 1, \dots, r$ und $j = 1, \dots, c$. Diese Information kann in einer $r \times c$ Tabelle mit Einträgen

$$n_{ij} = \text{Anzahl F\u00e4lle in } i \text{ und } j$$

zusammengefasst werden. Wir interessieren uns meistens f\u00fcr die Hypothese

H_0 : keine Assoziation zwischen den Klassifikationen.

Sei

$$p_{ij} = P(\text{Zeilenklasse} = i, \text{Spaltenklasse} = j)$$

$$p_{i.} = P(\text{Zeilenklasse} = i)$$

$$p_{.j} = P(\text{Spaltenklasse} = j)$$

$$\text{dann } H_0 \Rightarrow p_{ij} = p_{i.} p_{.j}$$

Wir schätzen $p_{i.}, p_{.j}$ mit

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \hat{p}_{.j} = \frac{n_{.j}}{n}$$

$$\Rightarrow e_{ij} = n \left(\frac{n_{i.}}{n} \right) \left(\frac{n_{.j}}{n} \right)$$

$$= \frac{(\text{Zeilensumme } i)(\text{ Spaltensumme } j)}{n}$$

Als Teststatistik haben wir

$$X^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Für $n \rightarrow \infty$ gilt $X^2 \sim \chi^2_{(r-1)(c-1)}$.

Als Faustregel (Cochran 1952!) hat man $e_{ij} \geq 5 \quad \forall ij$,
aber das ist sicherlich zu streng.

Freiheitsgrade

es sind rc Zellen (Kombinationen) rc

n wird festgelegt -1

$p_{i.}$ werden geschätzt $-(r - 1)$

$p_{.j}$ werden geschätzt $-(c - 1)$

Es bleiben $(r - 1)(c - 1)$

$$Y = X^2 \sim \chi_k^2$$

$$f(y) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} y^{k/2-1} e^{-y/2} \quad x \geq 0$$

$$E[Y] = k$$

$$V[Y] = 2k$$

Für $k = 1, 2$ hat die Dichte ein Maximum bei $y = 0$.

Für $k \geq 3$ ist $f(0) = 0$ und das Maximum liegt bei $y = k - 2$.

5.7.3 Interpretationen

In 1949 hatten Kieser und Schäfer aus *Who's Who* eine Tabelle für 1436 Frauen zusammengestellt, die all mindestens einmal verheiratet worden waren.

Ausbildung	Einmal Verheiratet	Mehrmals
College	550	61
Kein College	681	144

H_0 : keine Assoziation

$$X^2 = 16.0 \Rightarrow H_0 \text{ verwerfen, da } \chi_{1,0.99}^2 = 6.63$$

Heißt das, dass Frauen ohne Ausbildung heirateten öfters (17% v. 10%)?

Oder, dass Frauen die öfters heirateten eher nicht zum College gingen(30% v. 45%)?

Erstens haben wir nur die Hypothese von keiner Assoziation verworfen, ohne Kausalität zu besprechen. Zweitens wird es andere Faktoren geben.

Die Daten sind keine zufällige Stichprobe aus einer Grundgesamtheit, kann man trotzdem den Test durchführen und die Resultate verallgemeinern?

Ein komplexeres Beispiel — Intelligenz und Kleidung

Gilbey (Biometrika **8** 94) hat die Verteilung von 1725 Kindern besprochen, die nach Intelligenz

A: geistig mangelhaft

B: langsam und dumm

C. dumm

D: langsam aber intelligent

E: ziemlich intelligent

F: ausgesprochen fähig

G: sehr fähig

und nach Kleidung

sehr gut gekleidet

gut gekleidet

schlecht, aber passend

sehr schlecht

klassifiziert worden sind.

Was kann man daraus lesen?

Wie könnten die Daten graphisch dargestellt werden?

Könnte man Untergruppen testen oder Kategorien kombinieren?

Table 2. *Distribution of 1725 school children according to their standard of clothing and their intelligence: Kendall & Stuart (1967, p. 558) after Gilby.*

Standard of clothing	Intelligence class							Total
	A, B	C	D	E	F	G		
Very well clad	33	48	113	209	194	39	636	
Well clad	41	100	202	255	138	15	751	
Poor but passable	39	58	70	61	33	4	265	
Very badly clad	17	13	22	10	10	1	73	
Total	130	219	407	535	375	59	1725	

5.7.4 Warum die Chi-Quadrat Verteilung?

(1) Binomial und Chi-Quadrat Tests für 2 x 2 Kontingenztafeln

Als Beispiel nehmen wir die Erkältungsdaten, die Pauling berühmt gemacht hat:

279 Skifahrer in einer Doppelblindstudie (Ritzel [1961])

	Placebo	Ascorbic Acid	
Erkältet	31	17	48
Nein	109	122	231
	140	139	279

$$\chi^2 = 4.81$$

also, signifikant ($p < 0.05$), aber nicht sehr ($p > 0.01$).

Wir betrachten den allgemeinen Fall:

	G1	G2	
Ja	x_1	x_2	$x_1 + x_2$
Nein	$n_1 - x_1$	$n_2 - x_2$	$n_1 + n_2 - x_1 - x_2$
	n_1	n_2	$n_1 + n_2$

(a) Binomial Test

Wir vergleichen die Erfolgsquoten, p_1, p_2 für die zwei Gruppen, G1 und G2. Für groß n_1, n_2 sind die Schätzer $\hat{p}_j = \frac{x_j}{n_j}$ approximativ normalverteilt.

Sei $\{X_{ij}, i = 1, \dots, n_j \quad j = 1, 2\}$ u.i.v. Bernoulli ZV, d.h.

$$P(X_{ij} = 1) = p_j \quad P(X_{ij} = 0) = 1 - p_j$$

$$E[X_{ij}] = p_j \quad V[X_{ij}] = p_j(1 - p_j)$$

$$\Rightarrow \sum_{i=1}^{n_j} X_{ij} \sim B(n_j, p_j)$$

und für n_j groß gilt

$$\sum_{i=1}^{n_j} X_{ij} \sim N(n_j p_j, n_j p_j (1 - p_j))$$

bzw.

$$\hat{p}_j = \bar{X}_j = \sum_{i=1}^{n_j} X_{ij}/n_j \sim N\left(p_j, \frac{p_j(1 - p_j)}{n_j}\right)$$

$$\Rightarrow \hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}\right)$$

Unser H_0 wird $p_1 = p_2 = p$ sein. Für die Varianz setzen wir den Plugin-Schätzer für p ein

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{48}{279} = 17.2\% \text{ beim Pauling}$$

und betrachten die Teststatistik

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.2214 - 0.1223}{0.0452} = 2.193$$

$$T \sim N(0, 1) \Rightarrow T^2 \sim \chi_1^2$$

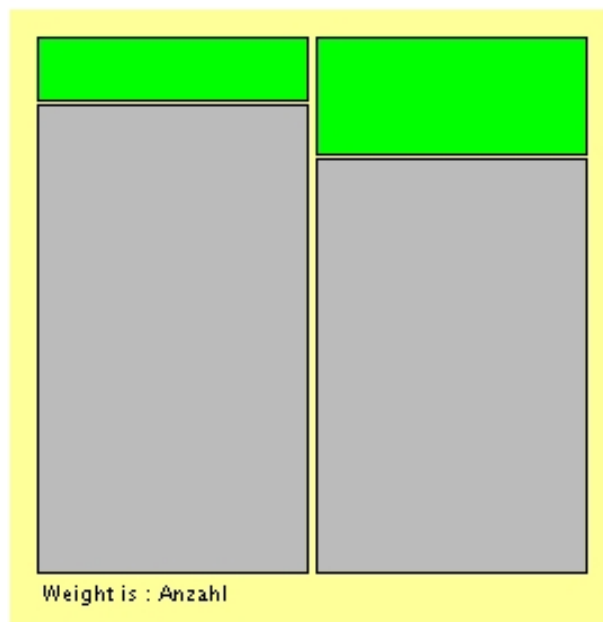
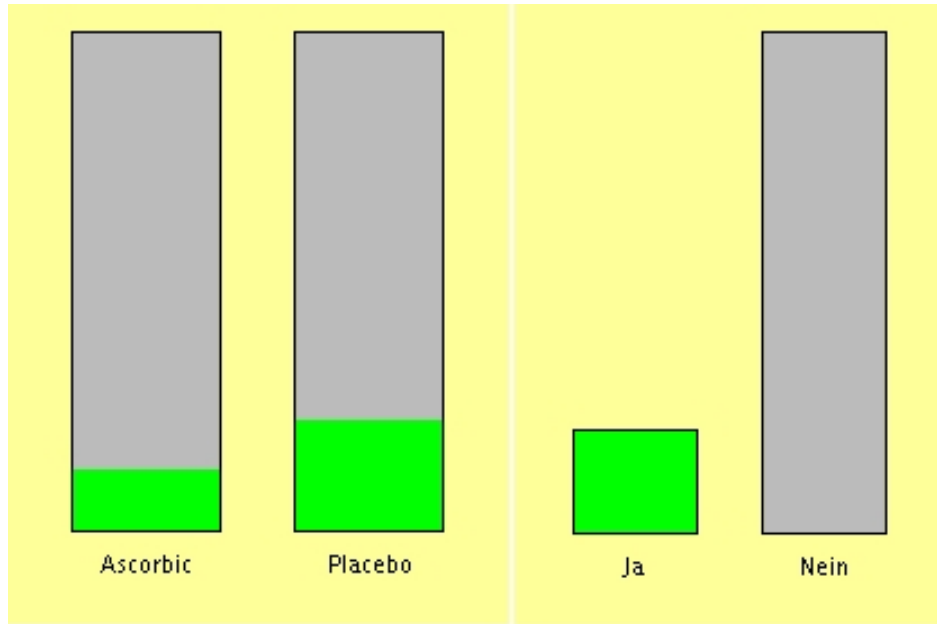
und wir können zeigen, dass

$$T^2 = \frac{(n_2x_1 - n_1x_2)^2(n_1 + n_2)}{n_1n_2(x_1 + x_2)(n_1 + n_2 - x_1 - x_2)}$$

(b) Chi-Quadrat Test

Für die angegebene 2x2 Kontingenztafel gilt

$$\begin{aligned} X^2 &= \left(x_1 - \frac{(x_1 + x_2)n_1}{n_1 + n_2}\right)^2 \left(\frac{1}{e_{11}} + \frac{1}{e_{21}} + \frac{1}{e_{12}} + \frac{1}{e_{22}}\right) \\ &= \frac{(n_2x_1 - n_1x_2)^2(n_1 + n_2)}{n_1n_2(x_1 + x_2)(n_1 + n_2 - x_1 - x_2)} \\ &= T^2 \end{aligned}$$



Erkältung gegen Ascorbic acid/Placebo
Oben: Spineplot und Säulendiagramm mit
Erkältungen selektiert
Unten: Mosaicplot mit Erkältungen selektiert

(2) Chi-Quadrat Tests für u. Poisson Verteilungen

Sei die Anzahl Beobachtungen in der Zelle i

$$X_i \quad \text{u.} \quad \sim P(\lambda_i)$$

Für λ_i groß gilt approximativ

$$X_i \sim N(\lambda_i, \lambda_i)$$

$$\Rightarrow \frac{X_i - \lambda_i}{\sqrt{\lambda_i}} \sim N(0, 1)$$

$$\Rightarrow \frac{(X_i - \lambda_i)^2}{\lambda_i} \sim \chi_1^2$$

$$\Rightarrow X^2 = \sum_{i=1}^m \frac{(X_i - \lambda_i)^2}{\lambda_i} \sim \chi_m^2$$

Aber im allgemeinen sind die Zellen nicht unabhängig und wir müssen die λ_i 's schätzen. Für jeden geschätzten Parameter nimmt man einen Freiheitsgrad weg. Der Beweis für die asymptotische Richtigkeit dieser Vorgehensweise ist aufwändig.

5.7.5 Beispiel — Tests von π

Gibt es Muster in der Dezimaldarstellung von π ?

1) Die Anzahl von Nullen testen

$$H_0 : P(0) = 0.1$$

$$H_A : P(0) \neq 0.1$$

Sei X_{0n} die Anzahl Nullen in den ersten n Stellen

$$X_{0n} \sim B(n, 0.1)$$

$$n \text{ groß} \Rightarrow X_{0n} \sim N(0.1n, 0.09n)$$

Ein Test des Niveaus $\alpha = 0.05$ hat in diesem Falle den Annahmebereich

$$0.1n \pm 1.96 * 0.3\sqrt{n}$$

z.B. $n = 1000000 \Rightarrow [99412, 100588]$ und in der Tat gibt es 99959 Nullen in den ersten Million Stellen.

Aber was ist mit den Zahlen 1, ..., 9?

2) Multinomial Test

Sei n_i die Anzahl von 'i', $i = 0, \dots, 9$

$$P(\{n_i\}) = \frac{n!}{\prod_{i=0}^9 n_i!} \prod_{i=0}^9 p_i^{n_i}$$

$$p_i = P(i)$$

$$H_0 : p_i = 0.1 \quad \forall i$$

$$H_A : p_i \neq 0.1 \text{ für einige } i$$

Annahmebereich B:

Alle Zustände (n_0, \dots, n_9) die unter H_0 eine Gesamtwahrscheinlichkeit $\geq 1 - \alpha$ haben und wo

$$P_{H_0}(\{n_i\} \in B) \geq P_{H_0}(\{n_i\} \notin B)$$

d.h. die Wahrscheinlichkeit von jedem Ereignis im Annahmebereich ist größer gleich die Wahrscheinlichkeit von jedem Ereignis im Ablehnungsbereich

3) Ein anderes Modell

Sei $n_i \sim P(np_i) \quad \forall i$ unabhängig von den anderen n_j 's und zuerst ohne der Einschränkung $\sum n_i = n$.

$$P(\{n_i\}) = \prod_{i=0}^9 e^{-np_i} \frac{(np_i)^{n_i}}{n_i!} = e^{-n} n^n \prod \frac{p_i^{n_i}}{n_i!}$$

Unter der Bedingung, daß $\sum n_i = n$, muß dann

$$P_p(\{n_i\} | \sum n_i = n) \propto \prod \frac{p_i^{n_i}}{n_i!}$$

Aus einem Vergleich mit dem Resultat für die Multinomiale Verteilung muß

$$P_p(\{n_i\} | \sum n_i = n) = n! \prod \frac{p_i^{n_i}}{n_i!}$$

Daraus schliessen wir, dass beide Ansätze zum selben Test führen. Aber der Zustandsraum ist zu groß. n identische Kugeln auf m verschiedene Zellen bedeutet

$$\binom{n + m - 1}{m - 1}$$

$$m = 10, n = 100 \Rightarrow \binom{109}{9} \gg 12^9$$

4) χ^2 Test

Für np_i groß gilt

$$n_i \sim N(np_i, np_i)$$
$$\Rightarrow \left(\frac{n_i - np_i}{\sqrt{np_i}} \right)^2 \sim \chi_1^2$$

Für $\{n_i\}$ unabhängig

$$X^2 = \sum_{i=0}^9 \frac{(n_i - np_i)^2}{np_i} \sim \chi_{10}^2$$

Unter der Einschränkung, dass $\sum n_i = n$, hat X^2 eine χ_9^2 Verteilung, weil wir einen Freiheitsgrad verloren haben.

$$\chi_{9,0.95}^2 = 16.92, \chi_{9,0.99}^2 = 21.67$$

Bei den Tests von Kanada war der höchste X^2 Wert 9.32.

Natürlich gibt es viele andere Tests auf die Zufälligkeit von Zahlenreihen — u.a. den Poker Test und den Lücken Test.

Frage: gibt es Muster in den ersten Dezimalstellen von e ?

5.7.6 Eine wichtige Eigenschaft von X^2

Für eine Kontingenztafel gilt

$$X^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

und unter H_0

$$= \sum_i \sum_j \frac{(n\hat{p}_{ij} - n\hat{p}_i\hat{p}_j)^2}{n\hat{p}_i\hat{p}_j}$$

$$= n \sum_i \sum_j \frac{(\hat{p}_{ij} - \hat{p}_i\hat{p}_j)^2}{\hat{p}_i\hat{p}_j}$$

d.h. der Wert von X^2 steigt direkt mit n für gegebene beobachtete Proportionen.

Wenn ein Resultat nicht signifikant ist, kann man leicht ausrechnen, wie groß die Stichprobe hätte sein müssen, um Signifikanz zu erreichen, gegeben die selben Proportionen.

5.7.7 Zusammenfassendes zum χ^2 Test

χ^2 Tests werden sehr oft verwendet. Man muss überprüfen, welche Annahmen gemacht worden sind, welche Nullhypothese getestet worden ist, und wieviele Tests gemacht worden sind. (Es gibt jede Menge Möglichkeiten, Kategorien einer Kontingenztafel zu kombinieren.)

Der p-Wert allein genügt nicht. Man sollte untersuchen, wo die Abweichungen von der Nullhypothese erscheinen. Dafür sind die χ^2 Beiträge der einzelnen Zellen aufschlussreich.

χ^2 Tests berücksichtigen nicht ordinale Struktur. Deshalb sind andere Testverfahren besser, wenn es um Anpassungstests geht. Trotzdem können χ^2 Tests als erster Versuch nützlich sein.

Nur große Werte von χ^2 werden als signifikant betrachtet, nicht kleine Werte, sogar wenn sie sehr unwahrscheinlich sind. ("Die beobachtete Werte sind zu gut.")