Chi² Test und Kontingenzkoeffizient

Für nominalskalierte Daten: - diese haben unterschiedliche Ausprägung,
- aber keine natürliche Reihenfolge

1. Chi² Test – Test nominalskalierter Daten

Vergleich von beobachteten mit erwarteten Häufigkeiten:
Für Stichproben und mehrere Datenreihen
Testzweck: geprüft wird, ob die beobachteten Daten aus einer erwarteten, bzw. vermuteten Verteilung stammen

1.1 Test auf bestimmte Verteilung für eine Stichprobe

- damit wird geprüft, ob die beobachteten Daten einer bestimmten Verteilung folgen (Normalverteilung, Poisson, parameterfreie Verteilung, etc)

Fragestellung: - Weichen die beobachteten Häufigkeiten B_i einer Stichprobe signifikant von den erwarteten Häufigkeiten E_i einer vermuteten Verteilung ab?
Voraussetzung: - keine, nominalskaliere Daten genügen

Vorgangsweise:

(1) Berechne zu den beobachteten Häufigkeiten B_i die erwarteten Häufigkeiten E_i einer vermuteten Verteilung und χ^2

H_0: Es gibt keine Abweichung zwischen beobachteter und erwarteter Verteilung

(2) Berechne: $\chi^2_{Vers} = \sum_{i=1}^{n} \frac{(B_i - E_i)^2}{E_i} = \left(\sum_{i=1}^{n} \frac{(B_i)^2}{E_i}\right) - N$

n Anzahl Merkmalsklassen
N Stichprobenumfang

(3) Entnehme der χ^2 Tabelle den Wert $\chi^2_{Tab}(FG; \alpha)$

FG = n - 1 - α
α Signifikanzniveau
a Anzahl, der aus den Daten geschätzten Parameter

(4) Teststatistik: $\chi^2_{Vers} \leq \chi^2_{Tab} \Longrightarrow$ akzeptiere H_0
$\chi^2_{Vers} > \chi^2_{Tab} \Longrightarrow$ verwerfe H_0 auf dem Signifikanzniveau α

Bsp.: - Test auf Normalverteilung,
- Kreuzungsversuch von Drosophila mit normalen und braunen Augen. Ist in der 2. Filialgeneration das Spaltungsverhältnis 3: 1? Test auf Verteilung 3:1 in F2

1.2. Homogenitätstest für den Vergleich der Häufigkeitsverteilung mehrerer Stichproben

– prüft, ob die jeweiligen beobachteten Häufigkeitsverteilungen Stichproben aus einer Grundgesamtheit sind, die bezüglich des untersuchten Merkmals gleiche Verteilungen aufweisen. Homogenität ist z.B. Voraussetzung für das Zusammenfassen einer Versuchsserie zu einer Stichprobe. Darstellung der Daten in einer sogenannten Kontingenztafel.

Fragestellung: Gibt es signifikante Unterschiede zwischen den Verteilungen in den r Stichproben? (Inhomogenität des Materials ?)

Voraussetzung: nominalskalierte Daten genügen

Vorgangsweise:

(1) r Stichproben mit c Merkmalsausprägungen, deren beobachtete Häufigkeiten B_{ij} in eine Kontingenztafel eingetragen werden.
Berechnung der Zeilen und Spaltensummen:

		Merkmalsausprägungen					
		1	2	3	**… j …**	c	**Σ**
Stichprobe	1	B_{11}	B_{12}	B_{13}	… B_{1j} …	B_{1c}	$\mathbf{Z_1}$
	2	B_{21}	B_{22}	B_{23}	… B_{2j} …	B_{2c}	$\mathbf{Z_2}$
	3	B_{31}	B_{32}	B_{33}	… B_{3j} …	B_{3c}	$\mathbf{Z_3}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	B_{i1}	B_{i2}	B_{i3}	… B_{ij} …	B_{ic}	$\mathbf{Z_i}$
	r	B_{r1}	B_{r2}	B_{r3}	… B_{rj} …	Br_c	$\mathbf{Z_r}$
	Σ	$\mathbf{Z_1}$	$\mathbf{Z_2}$	$\mathbf{Z_3}$	$\mathbf{Z_j}$	$\mathbf{Z_c}$	**N**

Mit:

$B_{i,j}$	Beobachtete Häufigkeiten
$Z_i = \sum_{j=1}^{c} B_{i,j}$	Die i-te Zeilensumme der Kontingenztafel (i fest, j variabel)
$S_j = \sum_{i=1}^{r} B_{i,j}$	Die j-te Spaltensumme der Kontingenztafel (i variabel, j fest)
$N = \sum_{i,j} B_{i,j}$	Stichprobenumfang
c, r	Anzahl Merkmalsausprägungen, Anzahl der Stichproben
i (bzw j)	Der Laufindex von 1 bis r (bzw. von 1 bis c)

(2) die erwarteten Häufigkeiten werden berechnet:

$E_{i,j} = Z_i \cdot S_j \cdot \frac{1}{N}$	Erwartete Häufigkeiten

(3) Berechnung von χ^2 (Chi-Quadrat)

$$\chi^2 = \sum_{i,j} \frac{(B_{ij} - E_{ij})^2}{E_{ij}} = \left(\sum_{i,j} \frac{(B_{ij})^2}{E_{ij}} \right) - N$$

(4) Entnehme der χ^2 Tabelle den Wert $\chi^2_{Tab}(FG;\alpha)$

FG = (c-1).(r-1)
α Signifikanzniveau

(5) Teststatistik: $\chi^2_{Vers} \leq \chi^2_{Tab} \Longrightarrow$ akzeptiere H_0 (Homogenität der Verteilungen)
$\chi^2_{Vers} > \chi^2_{Tab} \Longrightarrow$ verwerfe H_0 auf dem Signifikanzniveau α: mindestens eine Stichprobe weicht ab

Die Merkmalsklassen sind so zusammenzufassen, dass alle $E_{ij} > 1$ sind.

Bsp. (aus Köhler, Biostatistik). Sind Augenfarbe und Haarfarbe unabhängig voneinander?
Es liegen 3 Stichproben vor: Blau- Grün- und Braunäugige. Zu jeder Stichprobe ist die Häufigkeitsverteilung des Merkmals Haarfarbe gegeben; Das Merkmal hat 4 verschiedene Ausprägungen, c = 4. Mit dem Homogenitätstest soll geklärt werden, ob die 3 Stichproben als eine gemeinsame Stichprobe behandelt werden können, ob sie also homogen sind.

Häufigkeiten und Kontingenztabellen,

N Anzahl der Individuen
X_1, X_2 unterschiedliche Merkmalsausprägungen Haarfarbe und Augenfarbe
r Anzahl verschiedener Merkmalsausprägungen bei X_1, in diesem Fall Anzahl der unterschiedlichen Stichproben
c Anzahl verschiedener Merkmalsausprägungen bei X_2

Kontingenztafel mit r x c Feldern: - Eintragung der beobachteten Häufigkeiten, Berechnung der Spalten und Zeilensummen und der Randverteilungen = relative Spalten- und Zeilenhäufigkeit.

	j	1	2	3	Σ	
i	Haare \ Augen	Blau	Braun	Grün	**Zi**	Randvertlg, rel. Zeilenhfg
1	Blond	42	1	6	**49**	$\frac{49}{128}$=0,38
2	Braun	12	5	22	**39**	$\frac{39}{128}$=0,31
3	Schwarz	0	26	2	**28**	$\frac{28}{128}$=0,22
4	rot	8	4	0	**12**	$\frac{12}{128}$=0,09
Σ	**Sj**	**62**	**36**	**30**	**128**	
	Randverteilung Rel. Spaltenhfgkt.	$\frac{62}{128}$=0,48	$\frac{36}{128}$=0,28	$\frac{30}{128}$=0,24	r=4 c=3 N= 128	Zeilenzahl Spaltenzahl Stichprobengröße

Randverteilungen S_j/N; Z_i/N

Erwartete Häufigkeiten: $E_{i,j} = \frac{Z_i}{N} \cdot \frac{S_j}{N} \cdot N = Z_i \cdot S_j \cdot \frac{1}{N}$

(i,j)	(1,1)	(2,1)	(3,1)	(4,1)	(1,2)	(2,2)	(3,2)	(4,2)	(1,3)	(2,3)	(3,3)	(4,3)	$\sum$
$B_{i,j}$	42	12	0	8	1	5	26	4	6	22	2	0	128
$E_{i,j}$	23,73	18,89	13,56	5,81	13,78	10,97	7,88	3,38	11,48	9,14	6,56	2,81	128
$\frac{B_{ij}{}^2}{E_{ij}}$	74,34	7,62	0,00	11,02	0,07	2,28	85,79	4,73	3,14	52,95	0,61	0,00	242,55

Berechnung von χ^2 (Chi-Quadrat)

$$\chi^2 = \sum_{i,j} \frac{(B_{ij} - E_{ij})^2}{E_{ij}} = \left(\sum_{i,j} \frac{(B_{ij})^2}{E_{ij}} \right) - N$$

χ^2 = 242,55 – 128 = 114,55, daraus folgt:

Teststatistik: Aus Tabelle $\chi^2_{Tab(6;0,05)} = 12{,}59$

Entscheidung: $\chi^2_{Vers} > \chi^2_{Tab}$; **114,55 > 12,59,**

die Hypothese H_0 wird daher verworfen, die einzelnen Stichproben sind nicht homogen und können daher nicht als eine gemeinsame Stichprobe behandelt werden

2. Pearson'scher Kontingenzkoeffizient

- Kontingenz oder Assoziation ersetzt Korrelation intervallskalierter Daten.
Der Pearson'sche Kontingenzkoeffizient ist ein Maß für Kontingenz oder Assoziation, analog dem Korrelationskoeffizient als Maß für die Korrelation intervallskalierter Daten.

Definition und Berechnung des Kontingenzkoeffizienten C nach Pearson:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Eigenschaften von C: $C \in [0,1)$

C ist nach oben hin begrenzt mit $C \in \left[0, \sqrt{\frac{k-1}{k}}\right]$, k = min(r,c)

Korrigierter Kontingenzkoeffizient: - Normierung eliminiert den Einfluß von k, der minimalen Dimension der Kontingenztafel

$$C_{korr} = \sqrt{\frac{k}{k-1}} \cdot C = \sqrt{\frac{k}{k-1}} \cdot \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Im obigen Beispiel über den Zusammenhang Haar- und Augenfarbe ist demnach der Kontingenzkoeffizient C:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{114{,}55^2}{114{,}55 + 128}} = 0{,}69$$

k = min (r,c) = min (3,4) = 3

und der korrigierte Kontingenzkoeffizient C_{korr}:

$$C_{korr} = \sqrt{\frac{k}{k-1}} \cdot C = \sqrt{\frac{3}{3-1}} \cdot_{0{,}69} = 0{,}84$$

Literatur

Biostatistik. Köhler, Springer Verlag 2012
Intuitive Biostatistics, Harvey Motulsky, Oxford University Press, 1995