Stochastik für Informatiker
χ^2-Tests

Hanno Gottschalk

July 3, 2023

Inhaltsverzeichnis der Vorlesung

Mehrfachvergleiche und Bonferronikorrektur 3 / 17

Mehrfachvergleiche

Was passiert, wenn man drei oder mehr Gruppen miteinander vergleicht?
Beispiel:

- Wir vergleichen den Milcheiweissgehalt in der Milch von Berg- (X), Vorgebirgs- (Y), und Talkühen (Z)
- Wir wollen zeigen, dass es Unterschiede gibt
- Die Nullhypothese lautet demnach $\mu_X = \mu_Y = \mu_Z$ zum Signifikanzniveau α
- Wir machen dann drei t-Tests: $H_0 : \mu_X = \mu_Y$, $H_0 : \mu_X = \mu_Z$, $H_0 : \mu_Y = \mu_Z$.

Approximative Berechnung des Signifikanzniveaus für den Einzeltest

Welches Signifikanzniveau müssen wir für die einzelnen Tests erreichen, um insgesamt eine signifikante Aussage zu erhalten?

Nehmen wir das Signifikanzniveau $\alpha = 0.05$ für jeden einzelnen Test. H_0 wird abgelehnt, falls ein Test zur Verwerfung von H_0 führt. Dann

$$P(H_0|H_0) = (1-\alpha)^3 < 1-\alpha \text{ also wird } H_0 \text{ zu oft abgelehnt}$$

Bonferronikorrektur

Diesen Sachverhalt berücksichtigt die *Bonferronikorrektur*
Falls k unabhängige Tests zum Signifikanzniveau α^* insgesamt ein Signifikanzniveau α ergeben sollen, dann

$$(1-\alpha^*)^k = 1-\alpha \Rightarrow \alpha^* \approx \alpha/k \text{ für } \alpha \text{ klein}$$

Eine genauere Behandlung erfolgt durch die *Siebformel von Poincaré-Sylvestre*

Publication Bias

Nehmen wir an, auf 1 publizierte Untersuchung aus einem Medizinischen Studienzentrum der Pharamafirma 'Schluckspecht' kommen 19 nicht publizierte.

Die publizierte Studie weist ein Signifikanzniveau von 5% auf.

Auch bei völliger Wirkungslosigkeit der Medikamente von Schluckspecht erwartet man im Schnitt alle 20 Mal ein $5\% = 1/20$-signifikantes Resultat...

χ^2-Tests 8 / 17

Problemstellung

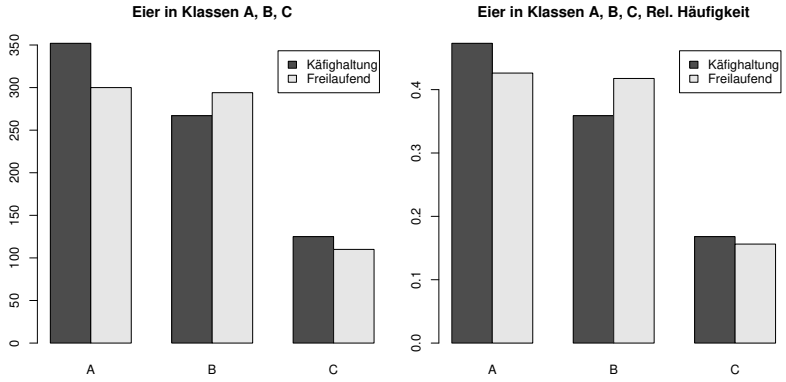
Was soll man machen, wenn man Zusammenhänge von Grössen bestimmen soll, die *nicht quantitativ* sind?

Beispiel:

- Hühner legen Eier der Güteklasse A, B oder C
- Sind Eier von freilaufenden Hühnern anders auf A, B, C verteilt, als die von Käfighühnern?

Eierklasse	A	B	C	Gesamt
Käfighaltung	352	267	125	744
Freilaufend	300	294	110	704
Gesamt	652	561	235	1448

Eierklassen — Daten



Sind die Unerschiede signifikant?

Wiederholung: Kontingenztabelle

Stelle zunächst eine *Kontingenztabelle der rel. Hkten* auf.

Eierklasse	A	B	C	Gesamt
Käfighaltung	0.243	0.184	0.086	0.513
Freilaufend	0.207	0.203	0.076	0.487
Gesamt	0.45	0.378	0.162	1

Allgemeiner Fall Kontingenztabelle

Allgemein betrachten wir ein Merkmal X mit den Ausprägungen $x_1, \ldots, x_r$ und ein Merkmal Y mit den Ausprägungen $y_1, \ldots y_s$

$$h_{i,j} = n_{i,j}/n\text{=rel. Häufigkeit, dass } X = x_i \text{ und } Y = y_j$$

Merkmal	y_1	y_2	$\ldots$	y_s	Gesamt
x_1	$h_{1,1}$	$h_{1,2}$	$\cdots$	$h_{1,s}$	h_1^X
x_2	$h_{2,1}$	$h_{2,2}$	$\cdots$	$h_{2,s}$	h_2^X
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
x_r	$h_{r,1}$	$h_{2,r}$	$\cdots$	$h_{r,s}$	h_r^X
Gesamt	h_1^Y	h_2^Y	$\cdots$	h_s^Y	1

Die Spalten "Gesamt" bilden die s ogenannten *Randverteilungen*

$\hat{\chi}^2$-Statistik und χ^2-Verteilung

Perfekte Unabhängigkeit: $h_{ij} = h_i^X h_j^Y$ für alle Paare i und j

$$\hat{\chi}^2 = n \sum_i \sum_j \frac{(h_{i,j} - h_i^X h_j^Y)^2}{h_i^X h_j^Y}, \quad n = \sum_i \sum_j n_{ij}$$

$\hat{\chi}^2$ heißt die *quadratische Kontingenz*. Sie "misst" die Abhängigkeit zweier Merkmale.

Bei perfekter Unabhängigkeit gilt $\hat{\chi}^2 = 0$ – dies ist jedoch wegen statistischer Schwankungen nicht zu erwarten

Verteilung der Quadratischen Kontingenz

Falls n genügend groß und $H_0 : X$ und Y sind *unabhängig*

$$\hat{\chi}^2 \sim \chi^2((r-1)(s-1)) \quad \text{(approximativ)} \tag{1}$$

Beachte: $\hat{\chi}^2$ ist zufallsabhängige Statistik...

$\chi^2(k)$ heisst die χ^2-Verteilung zu k Freiheitsgraden.

Testentscheidung im Unabhängigkeits / Homogenitätstest

χ^2-Homogenitäts/Unabhängigkeitstest zum Signifikanzniveau α

r Stichproben mit Umfang $n_1, \dots n_r$, $n = n_1 + \dots + n_r$ Gesamtstichprobenumfang

Y Merkmal mit Werten $\{y_1, \dots, y_s\}$

Nullhypothese H_0: Die Verteilung von Y hängt nicht ab vom Merkmal X

Stichprobenwerte: Kontingenztabelle relativer Hkt. (Gesamtstichpr.)

Merkmal X/Merkmal Y	y_1	y_2	$\dots$	y_s	Summe
x_1	$h_{1,1}$	$h_{1,2}$	$\cdots$	$h_{1,s}$	h_1^X
x_2	$h_{2,1}$	$h_{2,2}$	$\cdots$	$h_{2,s}$	h_2^X
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
x_r	$h_{r,1}$	$h_{2,r}$	$\cdots$	$h_{r,s}$	h_r^X
Summe	h_1^Y	h_2^Y	$\cdots$	h_s^Y	1

Testgröße: $K^2 = n \sum_i \sum_j \frac{(h_{i,j} - h_i^X h_j^Y)^2}{h_i^X h_j^Y}$

Testentscheidung: $K^2 \leq \chi^2_{1-\alpha}((r-1)(s-1))$

$\Rightarrow H_0$ annehmen, sonst ablehnen

(Hier $\hat{\chi}^2 = K^2$)

Fortsetzung Beispiel Eierklassen

- Haben $n = 1448$, $s = 3$, $r = 2$
- X= Haltungsart *(Käfig/Frei)*
- Y= Eierklasse *(A,B,C)*

Einsetzen für $\hat{\chi}^2$ liefert die Testentscheidung zum Signifikanzniveau $\alpha = 0.05$

$\hat{\chi}^2 = 5.3032 < \chi^2_{0.95}(2) = 5.9914$ ist richtig, daher H_0 beibehalten.

Der vom Computer berechnete p-Wert ist 7%, also wäre der Test auf Unterschiede signifikant zu $\alpha = 10\%$.

χ^2-Anpassungstest

Anstelle eines Merkmals X über zwei Gruppen $Y = y_1$ $Y = y_2$ kann man das Merkmal in Gruppe $Y = y_2$ auch durch eine vorgegebene Verteilung ersetzten. In diesem Fall spricht man vom χ^2-*Anpassungstest*.

$$H_0 : P(X = j) = p_j^0 > 0,\ j = 1, \ldots, k. \quad \text{Nullhypothese}$$

Teststatistik

$$\hat{\chi}^2 = n \sum_{j=1}^{k} \frac{(h_j - p_j^0)^2}{p_j^0} \sim \chi^2(k-1) \quad \text{asymptotisch.}$$

Testentscheidung $\hat{\chi}^2 \leq \chi^2_{1-\alpha}(k-1) \quad \Rightarrow \quad H_0$ annehmen, sonst H_1