Statistik – Einführung

Kategoriale Daten
Kapitel 10

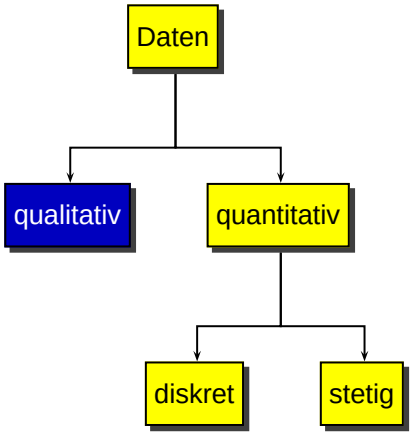
Statistik – WU Wien

Gerhard Derflinger · Michael Hauser · Jörg Lenneis · Josef Leydold ·
Günter Tirler · Rosmarie Wakolbinger

Lernziele

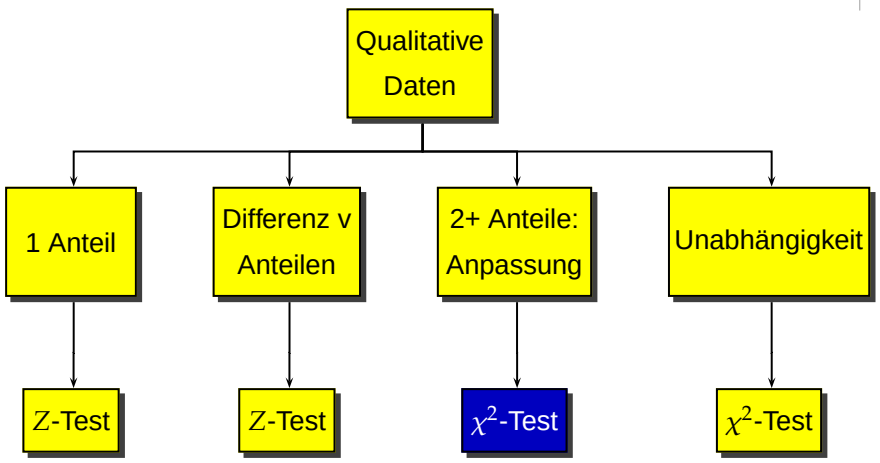
1. Erklären den χ^2-Test auf Anteile. (Anpassungstest)
2. Erklären den χ^2-Test auf Unabhängigkeit.

Datentypen



- Qualitative Zufallsvariable haben Ausprägungen, die klassifizieren (kategorisieren).
  - Z.B.: Geschlecht (♀, ♂)
- Sie zeigen die Häufigkeit in jeder Kategorie.
- Nominal- oder Ordinalskala.
- Beispiele:
  - Wohnen Sie in Wien?
  - Besitzen Sie ein Handy?

Hypothesen über kategoriale Daten



Chi-Quadrat (χ^2) Anpassungstest

1. Testet Hypothesen auf Anteile. (Alternativ-Hypothese „$\neq$“)
   - Beispiel:
     $H_0 : \theta_1 = 0.30, \theta_2 = 0.25, \theta_3 = 0.45$
     $H_A : \theta_1 \neq 0.30$ oder $\theta_2 \neq 0.25$ oder $\theta_3 \neq 0.45$.
2. Variable mit zwei oder mehreren Merkmalsausprägungen.
3. Voraussetzungen:
   - Multinomial-Verteilung
   - Jede erwartete Häufigkeit ≥ 5.
4. Verwendet die Ein-Weg-Kontingenztafel.

Multinomial-Verteilung

- n identische, unabhängige Versuche.
- k mögliche Versuchsausgänge.
- Wahrscheinlichkeit für Ausgang A_i konstant: $P(A_i) = \theta_i$.
- Zufallsvariable zählt die Erfolge für die einzelnen Versuchsausgänge, A_i. Bezeichnung dafür: n_i.
- Beispiel: Frage 100 Personen (n), welchen sie von 3 Kandidaten (k) wählen würden.

Ein-Weg-Kontingenztafel

Zeigt die Anzahl der Beobachtungen in k unabhängigen Kategorien.

Kandidat			
Müller	Huber	Mayer	Summe
35	20	45	100

1. **Hypothesen:**
   - $H_0 : \theta_1 = \theta_{1,0}, \quad \theta_2 = \theta_{2,0}, \quad \ldots \quad \theta_k = \theta_{k,0}$ (Behauptete Anteile)
   - H_A: Mindestens ein θ_i ist von $\theta_{i,0}$ verschieden.

2. **Teststatistik:** χ^2-verteilt mit $(k-1)$ Freiheitsgraden

$$\chi^2 = \sum \frac{(n_i^o - n_i^e)^2}{n_i^e}$$

   n_i^o ... be**o**bachtete Häufigkeit der i-ten Kategorie
   n_i^e ... **e**rwartete Häufigkeit der i-ten Kategorie: $n_i^e = n \cdot \theta_{i,0}$

3. Anzahl der Freiheitsgrade: $\nu = k - 1$
   (Anzahl der Kategorien -1)

χ^2-Anpassungstest // Grundidee

- Erwartete Anzahl ist die Häufigkeit, die unter der H_0 eintritt.
- Vergleiche die beobachtete Anzahl mit der erwarteten Anzahl.
- Der Abstand wird gemessen durch die Quadrate der Abweichungen (der Häufigkeiten) relativ zur (dividiert durch) erwarteten Anzahl.
- Je größer dieser Abstand ist, desto unwahrscheinlicher ist die Null-Hypothese.
  (H_0 wird für große χ^2-Werte abgelehnt.)

χ^2-Anpassungstest // Beispiel

Sie sind Leiter einer Personalabteilung. Sie wollen die Akzeptanz von drei verschiedenen Methoden zur Leistungsevaluierung unter ihren Mitarbeitern testen.
Von 180 befragten Angestellten halten 63 Methode 1 am besten geeignet, 45 Methode 2 und 72 Methode 3.

Gibt es bei einem Signifikanzniveau von 5% einen Unterschied in der Akzeptanz?

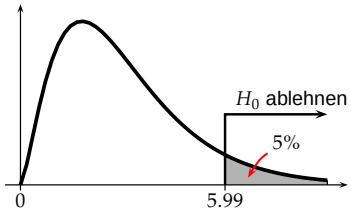
$H_0: \theta_1 = \theta_2 = \theta_3 = \frac{1}{3}$

H_A: Mind. 1 Anteil anders

$\alpha = 0.05$

$n_1^o = 63$, $n_2^o = 45$, $n_3^o = 72$
$\nu = 3 - 1 = 2$

Kritischer Wert: $\chi^2_{2,0.95} = 5.991$



Teststatistik:

χ^2-Anpassungstest // Lösung

	Methode 1	Methode 2	Methode 3	Summe
beobachtet:	63	45	72	180
erwartet:	60	60	60	180
	$n\,\theta_1 =$	$n\,\theta_2 =$	$n\,\theta_3 =$	
	$180 \cdot \frac{1}{3}$	$180 \cdot \frac{1}{3}$	$180 \cdot \frac{1}{3}$	

$$\chi^2 = \sum \frac{(n_i^o - n_i^e)^2}{n_i^e} = \frac{(63-60)^2}{60} + \frac{(45-60)^2}{60} + \frac{(72-60)^2}{60} = \frac{378}{60} = 6.3$$

χ^2-Anpassungstest // Lösung

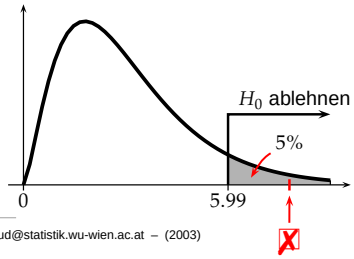
$H_0: \theta_1 = \theta_2 = \theta_3 = \frac{1}{3}$

H_A: 1 Methode anders

$\alpha = 0.05$

$n_1^o = 63$, $n_2^o = 45$, $n_3^o = 72$
$\nu = 3 - 1 = 2$

Kritischer Wert: $\chi^2_{2,0.95} = 5.991$



Teststatistik:

$\chi^2 = 6.3$

Entscheidung:

H_0 wird abgelehnt.

Interpretation:

Es gibt Evidenz, dass die drei Bewertungsmethoden unterschiedlich wahrgenommen werden.

Chi-Quadrat (χ^2) Unabhängigkeitstest

Chi-Quadrat (χ^2) Unabhängigkeitstest

1. Zeigt, ob eine Beziehung zwischen zwei qualitativen Variablen besteht.
   - Basiert auf einer Stichprobe.
   - Zeigt nicht Kausalität an!
2. Voraussetzungen:
   - Multinomial-Verteilung
   - Jede erwartete Häufigkeit ≥ 5.
3. Verwendet die (Zwei-Weg-)Kontingenztafel.

Kontingenztafel

Zeigt die zweidimensionale (gemeinsame) Verteilung der absoluten Häufigkeiten von zwei qualitativen Variablen (Wohnart und Wohnort) aus einer Stichprobe.

		Wohnort		
		Stadt	Land	Gesamt
Typ	Wohnung	63	49	112
	Haus	15	33	48
	Gesamt	78	82	160

1. **Hypothesen:**
   - H_0: Variable sind unabhängig
   - H_A: Variable sind nicht unabhängig (stehen in Beziehung)

2. **Teststatistik:** χ^2-verteilt mit ν Freiheitsgraden

$$\chi^2 = \sum \frac{(n^o_{ik} - n^e_{ik})^2}{n^e_{ik}}$$

n^o_{ik} ... be**o**bachtete Häufigkeit in Reihe i und Spalte k
n^e_{ik} ... **e**rwartete Häufigkeit in Reihe i und Spalte k

3. Anzahl der Freiheitsgrade:
   $\nu = (\text{Anzahl Reihen} - 1) \cdot (\text{Anzahl Spalten} - 1)$

Erwartete Beobachtungen

1. Statistisch unabhängig heißt, dass die Wahrscheinlichkeiten für jede Zelle gleich dem Produkt der Randwahrscheinlichkeiten ist.

2. Berechne Randwahrscheinlichkeiten.

3. Erwartete Anzahl der Beobachtungen ist das Produkt der Randwahrscheinlichkeiten mal Stichprobengröße:
   $n^e_{ik} = \frac{n_{i\star}}{n} \frac{n_{\star k}}{n} n$

$$n^e_{ik} = \frac{n_{i\star}\, n_{\star k}}{n}$$

$n_{i\star}$ ... Summe der Häufigkeiten über die i-te Reihe
$n_{\star k}$ ... Summe der Häufigkeiten über die k-te Spalte

Erwartete Beobachtungen // Beispiel

		Wohnort: Stadt	Land	Gesamt
Typ	Wohnung	63	49	112
	Haus	15	33	48
	Gesamt	78	82	160

$$n^e_{11} = (\frac{n_{1\star}}{n})(\frac{n_{\star 1}}{n})n = \frac{n_{1\star} \cdot n_{\star 1}}{n} = \frac{112 \cdot 78}{160}$$

Randwahrscheinlichkeiten

		Wohnort				
		Stadt		Land		Gesamt
		beob.	erw.	beob.	erw.	
Typ	Wohnung	63	54.6	49	57.4	112
	Haus	15	23.4	33	24.6	48
	Gesamt	78	78	82	82	160

χ^2-Unabhängigkeitstest // Beispiel

Sind sind Marktforscher in den USA. Sie befragen 286 Konsumenten (Zufallsstichprobe), ob sie *Diet Pepsi* und/oder *Diet Coke* kaufen. Besteht eine Evidenz auf einen Zusammenhang? (Signifikanzniveau 5%)

Diet Coke	*Diet Pepsi*		
	Nein	Ja	Gesamt
Nein	84	32	116
Ja	48	122	170
Gesamt	132	154	286

χ^2-Unabhängigkeitstest // Lösung

H_0: Kauf unabhängig

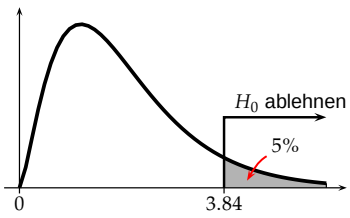
H_A: Kauf nicht unabhängig

$\alpha = 0.05$

$\nu = (2-1)(2-1) = 1$

Kritischer Wert: $\chi^2_{1,0.95} = 3.841$

Teststatistik:



Diet Pepsi

Diet Coke	Nein		Ja		Gesamt
	beob.	erw.	beob.	erw.	
Nein	84	53.5	32	62.5	116
Ja	48	78.5	122	91.5	170
Gesamt	132	132	154	154	286

$$\chi^2 = \frac{(84-53.5)^2}{53.5} + \frac{(32-62.5)^2}{62.5} + \frac{(48-78.5)^2}{78.5} + \frac{(122-91.5)^2}{91.5} = 54.29$$

χ^2-Unabhängigkeitstest // Lösung

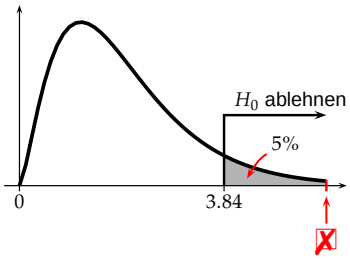
H_0: Kauf unabhängig

H_A: Kauf nicht unbhängig

$\alpha = 0.05$

$\nu = (2-1)(2-1) = 1$

Kritischer Wert: $\chi^2_{1,0.95} = 3.841$



Teststatistik:

$\chi^2 = 54.29$

Entscheidung:

H_0 wird abgelehnt.

Interpretation:

Es gibt Evidenz auf einen Zusammenhang.

Zusammenfassung

1. Erklärten den χ^2-Test auf Anteile. (Apassungstest)
2. Erklärten den χ^2-Test auf Unabhängigkeit.