Zusammenhang zwischen zwei kategoriellen Variablen

1 2×2-Kreuztabelle, Kontingenztafel

X	Y: 1	Y: 2	Total
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Verschiedene relative Häufigkeiten:

Gemeinsame Häufigkeit: der Anteil der Beobachtungen mit $X = 1$ und $Y = 1$ an der Gesamtzahl Beobachtungen.

Randhäufigkeit, marginale Häufigkeit: der Anteil der Beobachtungen mit $X = 1$ an der Gesamtzahl Beobachtungen.

Bedingte Häufigkeit: der Anteil der Beobachtungen mit $Y = 1$ unter den Beobachtungen mit $X = 1$.

Besteht ein Zusammenhang zwischen X und Y? Vergleiche die bedingten Häufigkeiten miteinander.

2 Chiquadrat-Test auf Unabhängigkeit

Gemeinsame Verteilung von X und Y: $p_{ij} = P(X = i, Y = j)$ mit $i = 1, 2, j = 1, 2$.
Randverteilungen: $p_{i.} = P(X = i) = \sum_j p_{ij}$ und $p_{.j} = P(Y = j) = \sum_i p_{ij}$.

X	Y: 1	Y: 2	Total
1	p_{11}	p_{12}	$p_{1.}$
2	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

Wenn X und Y unabhängig sind, so gilt:
$P(X = i, Y = j) = P(X = i)P(Y = j)$, d. h. $p_{ij} = p_{i.} \cdot p_{.j}$

Null- und Alternativhypothese:
$H_0 : p_{ij} = p_{i.} \cdot p_{.j}$ kein Zusammenhang zwischen X und Y
$H_A : p_{ij} \neq p_{i.} \cdot p_{.j}$ es gibt einen Zusammenhang zwischen X und Y

Erwartete Häufigkeiten unter H_0:

X	Y: 1	Y: 2	Total
1	$\frac{n_{1.}n_{.1}}{n_{..}}$	$\frac{n_{1.}n_{.2}}{n_{..}}$	$n_{1.}$
2	$\frac{n_{2.}n_{.1}}{n_{..}}$	$\frac{n_{2.}n_{.2}}{n_{..}}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Die Testgrösse basiert auf den Abweichungen zwischen beobachteten (O_k) und erwarteten Häufigkeiten (E_k) in den einzelnen Zellen k: $X^2 = \sum_{k=1}^{4} \frac{(O_k - E_k)^2}{E_k} = \sum_{ij} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}}$.

X^2 ist unter H_0 genähert χ^2-verteilt mit einem Freiheitsgrad. Näherung ist gut, wenn $n_{..} \geq 30$ und alle **erwarteten** Häufigkeiten ≥ 5.
Der χ^2-Test ist ein einseitiger Test! Mit $\alpha = 5\%$ wird H_0 verworfen, wenn $X^2 > 3.84$.

3 $r \times s$ -Kontingenztafeln

Die kategoriellen Variablen X und Y haben r und s Ausprägungen.
Nullhypothese H_0: kein Zusammenhang zwischen X und Y
Testgrösse basiert wieder auf den Abweichungen zwischen beobachteten (O_k) und erwarteten Häufigkeiten (E_k): $X^2 = \sum_{k=1}^{rs} \frac{(O_k - E_k)^2}{E_k}$.
X^2 hat unter H_0 genähert eine χ^2-Verteilung mit $(r-1)(s-1)$ Freiheitsgraden.
Näherung ist gut, falls $n_{..} \geq 30$, die meisten erwarteten Häufigkeiten ≥ 4 und höchstens 20% aller erwarteten Häufigkeiten zwischen 1 und 4 sind.

4 Bemerkungen zum Chiquadrat–Test

Immer Absolutzahlen verwenden, keine Prozentzahlen oder sonst irgendwie standardisierte Zahlen nehmen.
Beobachtungen müssen unabhängig sein. Bei z. B. gepaarten Daten ist *McNemar's Test* statt dem χ^2- Test durchzuführen.
Für Kategorien mit einer Rangordnung gibt es den Chiquadrattest auf Trend.
Bei zu kleinen Anzahlen ist *Fisher's exakter Test* durchzuführen.

5 Chiquadrat–Anpassungstest

Stimmt eine beobachtete Häufigkeitsverteilung mit einer theoretischen Verteilung überein?
Die Testgrösse $X^2 = \sum_{k=1}^{r} \frac{(O_k - E_k)^2}{E_k}$ ist χ^2 -verteilt mit ν Freiheitsgraden, wobei
$\nu =$ Anz. Klassen $-1-$ Anz. geschätzter Parameter.