

# Statistik

## Konfidenzintervall für relative Häufigkeiten

Alexander Hirner

### Konfidenzintervall für die relative Häufigkeit (für den Parameter $p$ einer BV)

Bei der Schätzung von Parametern ist es oft gewünscht, für den geschätzten Parameter  $\theta$  Intervalle  $[a, b]$  anzugeben, innerhalb derer der gesuchte “wahre” Populationsparameter mit einer gewissen Irrtumswahrscheinlichkeit (per Konvention verwendet man hier üblicherweise wieder 5% bzw. 1%) liegt, d.h. damit mit einer Wahrscheinlichkeit von 95% bzw. 99% die Aussage  $a \leq \theta \leq b$  gilt. Es ist möglich, solche “Konfidenzintervalle” für den Mittelwert, die Varianz, etc. zu bestimmen. Wir betrachten hier nur das Konfidenzintervall für die relative Häufigkeit.

Wir gehen von einer Befragung von  $n$  Personen zur Zustimmung bzw. Ablehnung zu einem gewissen Thema aus, wobei  $k$  Personen zustimmen (und folglich  $n-k$  ablehnen). Die relative Häufigkeit der Zustimmung beträgt  $r=k/n$ . Wir wollen Konfidenzintervalle angeben, innerhalb dieser sich die “wahre Zustimmungsrate” (in der Population) befindet.

#### **i** Konfidenzintervall für die relative Häufigkeit

Für ein 95%iges Konfidenzintervall gilt für die Intervallgrenzen  $p_1$  und  $p_2$ :

$$p_{1,2} = r \pm 1.96 \cdot \sqrt{\frac{r(1-r)}{n-1}}$$

Für ein 99%iges Konfidenzintervall gilt für die Intervallgrenzen  $p_1$  und  $p_2$ :

$$p_{1,2} = r \pm 2.58 \cdot \sqrt{\frac{r(1-r)}{n-1}}$$

**Interpretation** Diese Konfidenzintervalle ermöglichen uns Aussagen der Form “Mit 95 (99) prozentiger Wahrscheinlichkeit liegt der wahre Anteil in der Bevölkerung zwischen  $p_1$  und  $p_2$ ”.

Anmerkungen:

- Konfidenzintervalle sind uns bei der linearen Regression begegnet (KI für die Parameter der Regressionsgeraden, diese basieren allerdings auf einer sog. t-Verteilung.)
- Sie kennen diese Berechnungen sicherlich von den Hochrechnungen bei Wahlen.
- Wenn Sie statt 5% (d.h., berechnet wird ein 95%iges Konfidenzintervall) ein  $\alpha$  von 1% wählen (berechnet wird dann ein 99%iges KI) wird dieses natürlich **breiter**
- Mit diesen Überlegungen könnte man Konfidenzintervalle für beliebige Werte von  $\alpha$  erstellen. Gebräuchlich sind allerdings 5% bzw. 1%.

## Woher kommen die Zahlen in der Formel?

Falls die Stichprobe hinreichend groß ist (als Faustregel gilt hier, dass  $n \cdot p \geq 5$  und  $n \cdot (1 - p) \geq 5$  gelten soll) kann eine **Binomialverteilung** durch die **Normalverteilung** angenähert werden (dies vereinfacht die Rechnung beträchtlich).

Die Werte **1.96** (bzw. **-1.96**) sind genau jene z-Werte, die “nach rechts” (bzw. nach links) unter der Dichte der Standardnormalverteilung jeweils 2.5 % (also die Hälfte von  $\alpha=5\%$ ) “abschneiden”.

Die Werte **2.58** (bzw. **-2.58**) sind wiederum genau jene z-Werte, die “nach rechts” (bzw. nach links) unter der Dichte der Standardnormalverteilung jeweils 0.5 % (also die Hälfte von  $\alpha=1\%$ ) “abschneiden”.

### Beispiel

#### Angabe

Ein Meinungsforschungsinstitut erhebt bei  $n=200$  Bürgern die Zustimmung zu einer Wahlrechtsreform. Lediglich 78 Bürger stimmen dieser Reform zu. Vorausgesetzt, es handelt sich um eine repräsentative Stichprobe - was kann über die “wahre Zustimmungsrate” in der Bevölkerung gesagt werden? Berechne

- ein 95%iges KI für die Zustimmungsrate
- ein 99%iges KI für die Zustimmungsrate

#### Lösung

##### KI der Breite 95%

$$r = 78/200 = 0.39$$

$$p_{1,2} = 0.39 \pm 1.96 \cdot \sqrt{\frac{0.39 \cdot 0.61}{199}} = [0.3222, 0.4578]$$

Mit 95%iger Wahrscheinlichkeit liegt die wahre Zustimmungsrate in der Bevölkerung zwischen 32.22% und 45.78%.

##### KI der Breite 99%

$$r = 78/200 = 0.39$$

$$p_{1,2} = 0.39 \pm 2.58 \cdot \sqrt{\frac{0.39 \cdot 0.61}{199}} = [0.3008, 0.4792]$$

Mit 99%iger Wahrscheinlichkeit liegt die wahre Zustimmungsrate in der Bevölkerung zwischen 30.08% und 47.92%.