Einführung in Statistik

Peter Nejjar

July 9, 2025

Praxisbeispiel: Die Wirksamkeit zweier Medikamente

- Wir entwickeln **zwei neue Medikamente (A und B)** gegen eine seltene Krankheit.
- **Klinische Studie:** N Patienten werden zufällig in zwei Gruppen eingeteilt:
    - Gruppe A: n_1 Patienten erhalten Medikament A.
    - Gruppe B: n_2 Patienten erhalten Medikament B.

  (Natürlich gilt: $N = n_1 + n_2$)
- Nach der Behandlung: Hat sich die Symptomatik "verbessert" oder ist sie "unverändert"?

Beobachtete Daten (Beispiel: $n_1 = 10, n_2 = 10, N = 20$):

	Verbessert	**Unverändert**	**Summe**
Medikament A	8 (k_A)	2	**10 (n_1)**
Medikament B	4 (k_B)	6	**10 (n_2)**
Summe	**12 ($k_1 = k_A + k_B$)**	**8 (k_2)**	**20 (N)**

Hypothesen formulieren

- **Unsere Fragestellung:** Zeigen diese Daten einen *signifikanten Unterschied* in der Wirksamkeit zwischen Medikament A und Medikament B? Oder ist die Beobachtung rein zufällig?
- Um dies zu beantworten, formulieren wir zwei Hypothesen:
  1. **Nullhypothese (H_0):**
     - Die **Standardannahme** (kein Unterschied, kein Zusammenhang).
     - H_0: Es gibt **keinen Unterschied** in der Wirksamkeit zwischen Medikament A und Medikament B. Die beobachteten Unterschiede in der Tabelle sind **rein zufällig**.
  2. **Arbeitshypothese (H_1 oder H_A):**
     - Die **Behauptung**, die wir stützen wollen.
     - H_1: Es gibt einen **Unterschied** in der Wirksamkeit zwischen Medikament A und Medikament B. (z.B. Medikament A ist wirksamer oder es gibt einen generellen Unterschied).
- **Ziel:** Bewerten, wie plausibel H_0 angesichts unserer Daten ist. Wenn die Daten unter H_0 sehr unwahrscheinlich sind, verwerfen wir H_0 und stützen H_1.

Die Kernidee von Fishers exaktem Test

- Wir gehen davon aus, dass die **Randsummen** der Tabelle fix sind:
  - n_1 Patienten erhielten Medikament A.
  - n_2 Patienten erhielten Medikament B.
  - k_1 Patienten zeigten insgesamt eine Verbesserung.
  - k_2 Patienten zeigten insgesamt keine Verbesserung.
- **Unter der Annahme der Nullhypothese (H_0):**
  - Es gibt **keinen Unterschied** in der Wirksamkeit.
  - Die k_1 "Verbesserungen" sind zufällig auf die n_1 Patienten der Gruppe A und die n_2 Patienten der Gruppe B verteilt.
- Wir fragen: Wie wahrscheinlich ist es, genau unsere beobachtete Tabelle zu sehen, **wenn H_0 wahr wäre**?

Berechnung der Wahrscheinlichkeit (Beispiel)

Unter der Nullhypothese, dass beide Medikament gleich wirken, argumentieren wie folgt:

- die Zahl k_A sollte dann die Realisierung $X(\omega)$ einer hypergeometrisch verteilten Zufallsvariable sein,:
- $$\mathbb{P}(X = k) = \frac{\binom{k_1}{k}\binom{n_1+n_2-k_1}{n_1-k}}{\binom{n_1+n_2}{n_1}}$$
- Für unser Beispiel $k_A = 8$ erhalten wir $\mathbb{P}(X = k_A) = \frac{\binom{12}{8}\times\binom{8}{2}}{\binom{20}{10}}$

Berechnung der Wahrscheinlichkeit (Beispiel)

Unter der Nullhypothese, dass beide Medikament gleich wirken, argumentieren wie folgt:

- die Zahl k_A sollte dann die Realisierung $X(\omega)$ einer hypergeometrisch verteilten Zufallsvariable sein,:
- $$\mathbb{P}(X = k) = \frac{\binom{k_1}{k}\binom{n_1+n_2-k_1}{n_1-k}}{\binom{n_1+n_2}{n_1}}$$
- Für unser Beispiel $k_A = 8$ erhalten wir $\mathbb{P}(X = k_A) = \frac{\binom{12}{8}\times\binom{8}{2}}{\binom{20}{10}}$
- **Berechnung der Binomialkoeffizienten:**
  - $\binom{12}{8} = 495 \quad \binom{8}{2} = 28 \quad \binom{20}{12} = 184756$
- $$\mathbb{P}(X = k_A) = \frac{\binom{12}{8} \times \binom{8}{2}}{\binom{20}{10}} \approx 0.075$$

Berechnung der Wahrscheinlichkeit (Beispiel)

Unter der Nullhypothese, dass beide Medikament gleich wirken, argumentieren wie folgt:

- die Zahl k_A sollte dann die Realisierung $X(\omega)$ einer hypergeometrisch verteilten Zufallsvariable sein,:
- $$\mathbb{P}(X = k) = \frac{\binom{k_1}{k}\binom{n_1+n_2-k_1}{n_1-k}}{\binom{n_1+n_2}{n_1}}$$
- Für unser Beispiel $k_A = 8$ erhalten wir $\mathbb{P}(X = k_A) = \frac{\binom{12}{8}\times\binom{8}{2}}{\binom{20}{10}}$
- **Berechnung der Binomialkoeffizienten:**
  - $\binom{12}{8} = 495 \quad \binom{8}{2} = 28 \quad \binom{20}{12} = 184756$
- $$\mathbb{P}(X = k_A) = \frac{\binom{12}{8} \times \binom{8}{2}}{\binom{20}{10}} \approx 0.075$$
- Das bedeutet: Eine Wahrscheinlichkeit von ca. $7,5\%$ für diese Tabelle, **wenn die Medikamente gleich wirksam wären**.

P-Werte

- Wir betrachten auch **Ergebnisse** , die noch stärkere Hinweise auf einen Unterschied geben würden. Um zu entscheiden, ob der Wert k_A unter der Nullhypothese verdächtig klein oder groß ist, betrachten wir den linksseitigen p-Wert
  - $\mathbb{P}(X \leq k_A)$

  sowie den rechtsseitigen p-Wert:
  - $\mathbb{P}(X \geq k_A)$
- zweiseitiger p-Wert: $2 \min\{\mathbb{P}(X \geq k_A), \mathbb{P}(X \leq k_A)\}$

Interpretation des p-Wertes

- **Kleiner p-Wert (z.B. < 0.05):** Das beobachtete Ergebnis ist unter H_0 sehr unwahrscheinlich. $\rightarrow$ Wir haben **Grund, die Nullhypothese zu verwerfen** und die **Arbeitshypothese (H_1) zu stützen**.
- **Großer p-Wert (z.B. > 0.05):** Das beobachtete Ergebnis ist unter H_0 plausibel. $\rightarrow$ Wir haben **keinen ausreichenden Grund, die Nullhypothese zu verwerfen**.

Theoretische Herleitung: Die Urnen-Analogie

Die 2x2 Kontingenztafel:

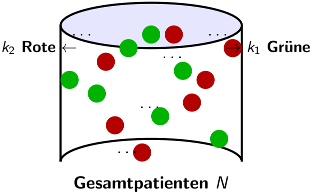
	V	U	Summe
Med. A	a	b	n_1
Med. B	c	d	n_2
Summe	k_1	k_2	N

Nullhypothese (H_0):

- Medikament A und B sind **gleich wirksam**. Die k_1 Verbesserungen sind **zufällig** auf die Gruppen verteilt.

Die Urnen-Analogie unter H_0:

- Stellen Sie sich N Kugeln (Patienten) in einer Urne vor.
- k_1 Kugeln sind **grün**(Verbesserung) .
- k_2 Kugeln sind **rot**(Unverändert) .
- Diese k_1 und k_2 sind fix.



Beispiel 2: Rauchen und Herz-Kreislauf-Erkrankungen

- Wir untersuchen den Zusammenhang zwischen **Rauchen** und einer **Herz-Kreislauf-Erkrankung (HKK)**.
- **Studiendesign:** 50 Personen werden nach Rauchstatus und HKK-Diagnose befragt.

Beobachtete Daten:

	HKK Ja	**HKK Nein**	**Summe**
Raucher	15	5	**20**
Nichtraucher	8	22	**30**
Summe	**23**	**27**	**50**

Hypothesen für Beispiel 2

- **Nullhypothese (H_0):**
  - Es gibt **keinen Zusammenhang** zwischen Rauchen und dem Auftreten der Herz-Kreislauf-Erkrankung.
  - Die HKK-Fälle sind **rein zufällig** auf Raucher und Nichtraucher verteilt.
- **Arbeitshypothese (H_1):**
  - Es gibt einen **Zusammenhang** zwischen Rauchen und der Herz-Kreislauf-Erkrankung.
  - (Z.B. Raucher sind häufiger von HKK betroffen.)

Kernidee & Berechnung für Beispiel 2

- **Kernidee:** Unter H_0 sind die 23 HKK-Fälle zufällig auf die 20 Raucher und 30 Nichtraucher verteilt (Randsummen fix).
- Wir berechnen, wie wahrscheinlich unsere beobachtete Tabelle (oder eine extremere) unter dieser Annahme ist.

Wahrscheinlichkeit unserer Tabelle ($a = 15, b = 5, c = 8, d = 22$):

$$\mathbb{P}(\text{unsere Tabelle}) = \frac{\binom{20}{15} \times \binom{30}{8}}{\binom{50}{23}} = \mathbf{0.00083988353} \tag{1}$$

$$\mathbb{P}(\text{unsereTabelle}) = \frac{\binom{20}{15} \times \binom{30}{8}}{\binom{50}{23}} = \mathbf{0.00083988353} \tag{2}$$

Hier haben wir die Logik etwas umgedreht: Vorher stellten wir uns vor, dass unter H_0 die n_1 Patienten aus Gruppe A sich so verhalten, wie als wären sie zufällig ohne Zurücklegen gezogen wurden. Im Raucherbeispiel würden wir uns dann vorstellen, dass die 20 Raucher zufällig ohne Zurücklegen aus den 50 Personen gezogen worden, und berechnen die Wahrscheinlichkeit, 15 Personen mit HKK zu ziehen: Das ergibt

$$\mathbb{P}(\text{unsereTabelle}) = \frac{\binom{23}{15} \times \binom{27}{5}}{\binom{50}{20}} \tag{3}$$

In (2) stellen wir uns jedoch vor, dass zufällig 23 Personen (mit HKK) gezogen wurden, und gucken, wie wahrscheinlich es unter H_0 ist, dass 15 Personen zur Gruppe der Raucher gehören. Allerdings ergeben (3),(2) dieselbe Wahrscheinlichkeit, d.h. beide Ansätze sind gleichwertig.

Einführung: Warum Schätzer?

- In der Statistik interessieren wir uns oft für **Parameter** einer **Grundgesamtheit (Population)**.
- Beispiele:
  - Durchschnittsgröße aller Erwachsenen in Deutschland (μ).
  - Varianz der Einkommen in einer Stadt (σ^2).
  - Anteil der Wähler, die Partei X unterstützen (p).
- Problem: Es ist meist **unmöglich**, die gesamte Population zu untersuchen.
- **Lösung:** Wir ziehen eine **Stichprobe** und nutzen die Informationen aus dieser Stichprobe, um Rückschlüsse auf die unbekannten Populationsparameter zu ziehen.
- **Schätzer** sind die mathematischen Werkzeuge, die uns dabei helfen.

Was ist ein Schätzer?

- Ein **Schätzer** (engl. *estimator*) ist eine **Funktion der Stichprobendaten**, die verwendet wird, um den Wert eines unbekannten Populationsparameters zu approximieren.
- Er wird oft mit einem "Dach" über dem Parameter-Symbol gekennzeichnet, z.B. $\hat{\theta}$ (gesprochen "Theta-Dach") für einen Parameter θ.
- Das **Ergebnis** der Anwendung eines Schätzers auf eine konkrete Stichprobe ist der **Schätzwert** (engl. *estimate*).

- **Parameter** der Population: Der wahre Mittelwert μ.
- **Schätzer** für μ: Der Stichprobenmittelwert $\bar{X}$.
- **Stichprobe:** $X_1, X_2, \ldots, X_n$
- **Schätzwert:** Der berechnete Wert $\bar{x}$ aus dieser konkreten Stichprobe.

Güte von Schätzern: Welche Schätzer sind "gut"?

- Wir wollen Schätzer, die uns möglichst genaue und zuverlässige Schätzwerte liefern.
- Wichtige Kriterien zur Bewertung von Schätzern:
  1. **Erwartungstreue (Unbiasedness):**
     - Ein Schätzer $\hat{\theta}$ ist erwartungstreu für θ, wenn sein **Erwartungswert** dem wahren Parameterwert entspricht: $E[\hat{\theta}] = \theta$.
     - Im Durchschnitt liegt der Schätzer richtig.
  2. **Effizienz (Efficiency):**
     - Ein Schätzer ist effizienter als ein anderer, wenn er eine **kleinere Varianz** hat (unter mehreren erwartungstreuen Schätzern).
     - Weniger Streuung um den wahren Parameterwert.
  3. **Konsistenz (Consistency):**
     - Wenn die Stichprobengröße n gegen unendlich geht, konvergiert der Schätzer zum wahren Parameterwert.
     - Mit mehr Daten wird der Schätzer immer genauer.

Schätzer für den Erwartungswert (μ)

- Der unbekannte Populationsparameter ist der Erwartungswert (Mittelwert) μ.
- Die Beobachtungen in unserer Stichprobe sind $X_1, X_2, \ldots, X_n$.
- Der am häufigsten verwendete Schätzer für μ ist der **Stichprobenmittelwert**:

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Warum ist der Stichprobenmittelwert erwartungstreu?

- Wir wollen zeigen: $\mathbb{E}[\bar{X}] = \mu$.
- $\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$
- Aufgrund der Linearität des Erwartungswertes:

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Schätzer für die Varianz (σ^2)

- Der unbekannte Populationsparameter ist die Varianz σ^2.
- Die Populationsvarianz ist definiert als: $\sigma^2 = E[(X-\mu)^2]$.
- Ein intuitiver Schätzer könnte sein:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Allerdings ist dieser Schätzer **nicht erwartungstreu**! Er unterschätzt die wahre Varianz systematisch.

Die Bessel-Korrektur: Der erwartungstreue Schätzer

- Der erwartungstreue Schätzer für die Populationsvarianz σ^2 ist die **korrigierte Stichprobenvarianz**:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Der Nenner $(n-1)$ statt n wird als **Bessel-Korrektur** bezeichnet.

Zusammenfassung und Wichtige Punkte

- **Schätzer** sind Funktionen der Stichprobendaten, um unbekannte Populationsparameter zu approximieren.
- Wir streben nach **guten Schätzern**:
  - **Erwartungstreu:** Der Schätzer liegt im Durchschnitt richtig ($E[\hat{\theta}] = \theta$).
  - **Effizient:** Geringe Streuung um den wahren Wert.
  - **Konsistent:** Wird mit mehr Daten immer genauer.
- **Schätzer für den Erwartungswert (μ):**
  - Der **Stichprobenmittelwert** $\bar{X}$ ist ein erwartungstreuer Schätzer für μ.
- **Schätzer für die Varianz (σ^2):**
  - Die **korrigierte Stichprobenvarianz** $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ ist ein erwartungstreuer Schätzer für σ^2.
  - Die **Bessel-Korrektur ($n - 1$)** ist entscheidend für die Erwartungstreue.