Signifikant oder nicht?

Hypothesentests und Konfidenzintervalle

Oliver Ernst

Professur Numerische Mathematik, TU Chemnitz

Tag der Mathematik
21. März 2026

Mathematik!
TU Chemnitz

TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Stochastik im sächsischen Lehrplan (Gy)

Klassenstufe	Lernbereich	Stundenumfang
7	LB4 Vernetzung: Darstellen von Daten	4
8	LB2 Zufallsversuche	24
9	LB4 Auswerten von Daten	12
10	LB2 Diskrete Zufallsgrößen	16
11/12 Gk	LB4 Binomialverteilte Zufallsgrößen	18
	LB6 Beurteilende Statistik	12
	WB4 Bedingte Wahrscheinlichkeit	
11/12 Lk	LB4 Binomialverteilte Zufallsgrößen	32
	LB6 Normalverteilte Zufallsgrößen	12
	LB7 Beurteilende Statistik	12

Wozu? KMK-Beschluss vom 07.07.1972 i.d.F. vom 09.02.2012:

Der Unterricht in der gymnasialen Oberstufe vermittelt

- *eine vertiefte Allgemeinbildung,*
- *allgemeine Studierfähigkeit sowie*
- *wissenschaftspropädeutische Bildung.*

Plan

1. Wahrscheinlichkeit und Zufallsvariablen
2. Schätzen und Konfidenzintervalle
3. Testen und Signifikanz

Plan

- Wahrscheinlichkeiten (Zahlen zwischen 0 und 1) werden Mengen zugeschrieben.

- Wahrscheinlichkeiten (Zahlen zwischen 0 und 1) werden Mengen zugeschrieben.
- Wahrscheinlichkeitsraum: $(\Omega, \mathfrak{A}, \mathbf{P})$

  Ω : Menge von Elementarereignissen $\omega \in \Omega$

  $\mathfrak{A}$: System aus Teilmengen von Ω,

  geschlossen unter Vereinigung, Durchschnitt, Komplementbildung

  $\mathbf{P}$: Wahrscheinlichkeitsmaß, Abbildung $\mathbf{P}\colon \mathfrak{A} \to [0, 1]$.
- Ist Ω endlich oder abzählbar, kann man als $\mathfrak{A}$ die Potenzmenge von Ω nehmen. Andernfalls (z.B. $\Omega = \mathbb{R}$) muss $\mathfrak{A}$ kleiner gewählt werden.

- Wahrscheinlichkeiten (Zahlen zwischen 0 und 1) werden Mengen zugeschrieben.
- Wahrscheinlichkeitsraum: $(\Omega, \mathfrak{A}, \mathbf{P})$

  Ω : Menge von Elementarereignissen $\omega \in \Omega$

  $\mathfrak{A}$: System aus Teilmengen von Ω,

  geschlossen unter Vereinigung, Durchschnitt, Komplementbildung

  $\mathbf{P}$: Wahrscheinlichkeitsmaß, Abbildung $\mathbf{P}\colon \mathfrak{A} \to [0,1]$.
- Ist Ω endlich oder abzählbar, kann man als $\mathfrak{A}$ die Potenzmenge von Ω nehmen. Andernfalls (z.B. $\Omega = \mathbb{R}$) muss $\mathfrak{A}$ kleiner gewählt werden.

- Beispiel diskrete Gleichverteilung: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, $\mathbf{P}(\{\omega_i\}) = \frac{1}{n}\ \forall i$.
  Münzwurf : $\Omega = \{\text{Kopf}, \text{Zahl}\}$, $\mathbf{P}(\{\text{Kopf}\}) = \mathbf{P}(\{\text{Zahl}\}) = \frac{1}{2}$.
  Laplace-Wahrscheinlichkeit:
$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$
- Beispiel Bernoulli-Verteilung $\mathrm{B}(p)$: $\Omega = \{\omega_1, \omega_2\}$
$$\mathbf{P}(\omega_1) = p \in [0,1], \quad \mathbf{P}(\omega_2) = 1 - p.$$

- Zufallsvariable (Zufallsgröße): Abbildung $X : \Omega \to \mathbb{N}$

$$M \subset \mathbb{N}: \qquad \mathbf{P}(M) = \mathbf{P}(X \in M) = \mathbf{P}\big(\{\omega \in \Omega : X(\omega) \in M\}\big).$$

Beispiel: X = Anzahl „Erfolge“ bei n-stufiger Bernoulli-Kette

$$M = \{k\} \qquad \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Binomialverteilung $\mathsf{B}(n, p)$, $n \in \mathbb{N}$, $p \in [0, 1]$.

Zähldichte

$$p_X : \{0, \ldots, n\} \to [0, 1], \qquad p_X(k) = \mathbf{P}(X = k).$$

- Zufallsvariable (Zufallsgröße): Abbildung $X : \Omega \to \mathbb{N}$

$$M \subset \mathbb{N}: \qquad \mathbf{P}(M) = \mathbf{P}(X \in M) = \mathbf{P}\big(\{\omega \in \Omega : X(\omega) \in M\}\big).$$

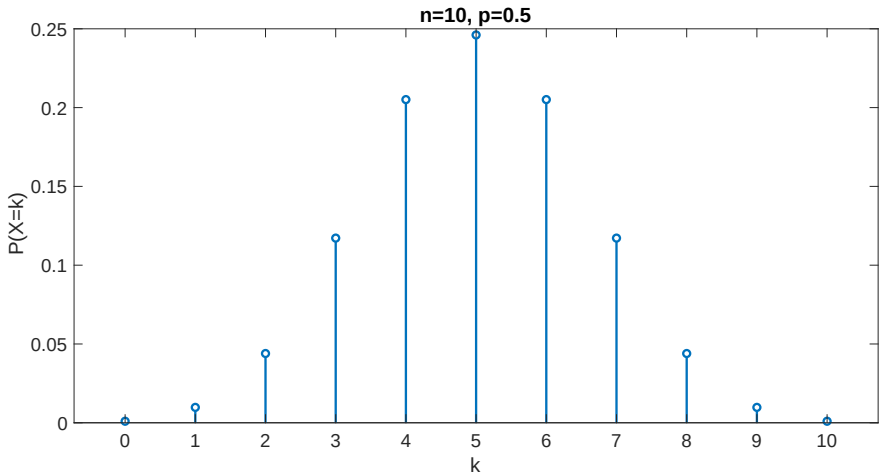
Beispiel: X = Anzahl „Erfolge“ bei n-stufiger Bernoulli-Kette

$$M = \{k\} \qquad \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Binomialverteilung $\mathrm{B}(n, p)$, $n \in \mathbb{N}$, $p \in [0, 1]$.

Zähldichte

$$p_X : \{0, \ldots, n\} \to [0, 1], \qquad p_X(k) = \mathbf{P}(X = k).$$



- Zufallsvariable (Zufallsgröße): Abbildung $X : \Omega \to \mathbb{N}$

$$M \subset \mathbb{N}: \qquad \mathbf{P}(M) = \mathbf{P}(X \in M) = \mathbf{P}\big(\{\omega \in \Omega : X(\omega) \in M\}\big).$$

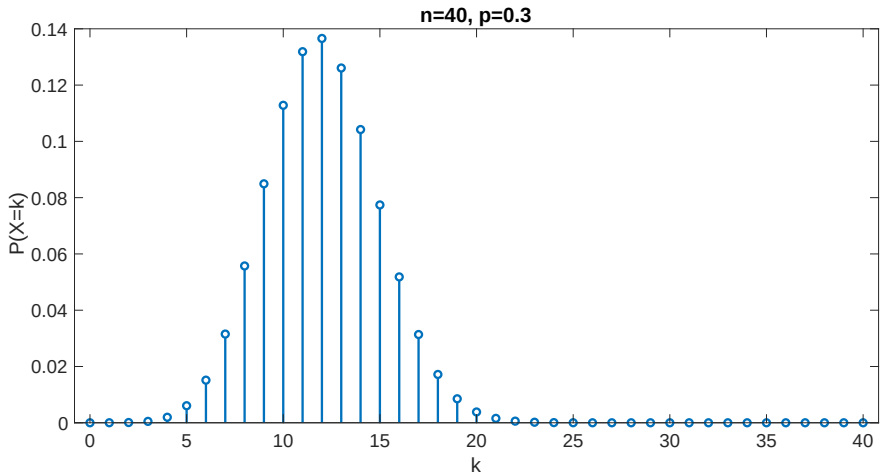
Beispiel: X = Anzahl „Erfolge“ bei n-stufiger Bernoulli-Kette

$$M = \{k\} \qquad \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Binomialverteilung $\mathrm{B}(n, p)$, $n \in \mathbb{N}$, $p \in [0, 1]$.

Zähldichte

$$p_X : \{0, \ldots, n\} \to [0, 1], \qquad p_X(k) = \mathbf{P}(X = k).$$



- Bei Zufallsvariablen $X\colon \Omega \to \mathbb{R}$ (Bildbereich überabzählbar) muss das zulässige Mengensystem (σ-algebra) $\mathfrak{A}$ kleiner als die Potenzmenge gewählt werden.
- Genügt: Intervalle $[a,b]$, $(a,b]$, $[a,b)$, (a,b), $a,b\in\mathbb{R}$, $a\leqslant b$, bzw. unbeschränkte Intervalle sind messbar.
- Deren Wahrscheinlichkeiten lassen sich mittels Verteilungsfunktion $F_X\colon \mathbb{R}\to[0,1]$ ausdrücken:

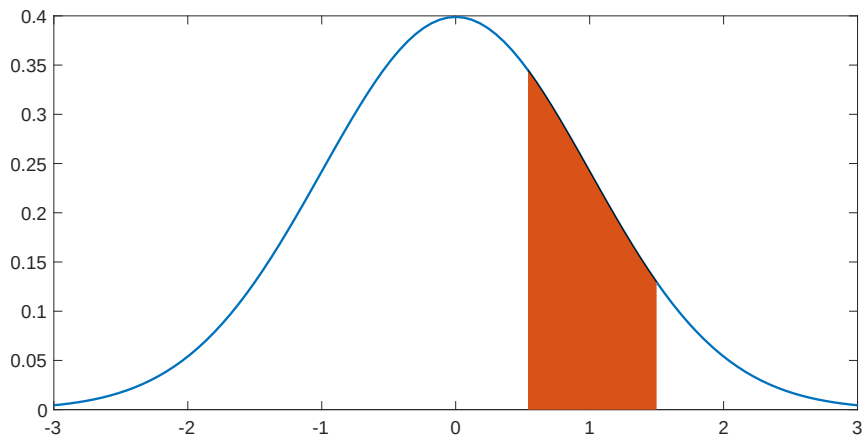
$$\mathbf{P}(X\in[a,b]) = F_X(b) - F_X(a).$$

- Absolutstetige Zufallsvariablen besitzen eine Dichtefunktion $p_X\colon \mathbb{R}\to\mathbb{R}_0^+$ sodass

$$\mathbf{P}(X\in[a,b]) = F_X(b) - F_X(a) = \int_a^b p_X(t)\,\mathrm{d}t \quad .$$

Beispiel Standardnormalverteilung

$$\mathbf{P}(\tfrac{1}{2}\leqslant X\leqslant \tfrac{3}{2}) = \frac{1}{\sqrt{2\pi}}\int_{1/2}^{3/2} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$



- Erwartungswert einer Zufallsvariable X

diskret
$$\mathbf{E}[X] = \sum_{k=1}^{K} k \cdot \mathbf{P}(X = k)$$

kontinuierlich
$$\mathbf{E}[X] = \int_{\mathbb{R}} t \cdot p_X(t)\,\mathrm{d}t$$

Beispiele:

$X \sim \mathrm{B}(p)$ (Bernoulli)
$$\mathbf{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

$X \sim \mathrm{B}(n,p)$ (binomial)
$$\mathbf{E}[X] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = np$$

$X \sim \mathrm{N}(\mu, \sigma^2)$ (normal)
$$\mathbf{E}[X] = \int_{-\infty}^{\infty} t \cdot \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{\frac{(t-\mu)^2}{2\sigma^2}}\,\mathrm{d}t = \mu$$

- Varianz einer Zufallsvariable X

$$\mathbf{Var}\, X = \mathbf{E}\left[(X - \mathbf{E}\,[X])^2\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}\,[X]^2 \geqslant 0$$

Beispiele:

$X \sim \mathrm{B}(p)$ (Bernoulli) $\quad \mathbf{Var}\, X = \mathbf{E}\left[(X-p)^2\right] = 1\cdot(1-p)^2 + 0\cdot(0-p)^2 = p(1-p)$

$X \sim \mathrm{B}(n,p)$ (binomial) $\quad \mathbf{Var}\, X = np(1-p)$

$X \sim \mathrm{N}(\mu,\sigma^2)$ (normal) $\quad \mathbf{Var}\, X = \sigma^2$

- Standardabweichung einer Zufallsvariable X: $\sigma := \sqrt{\mathbf{Var}\, X}$.

Entscheidend: Ist $X \sim \mathsf{N}(\mu, \sigma^2)$, so besitzt die Zufallsvariable

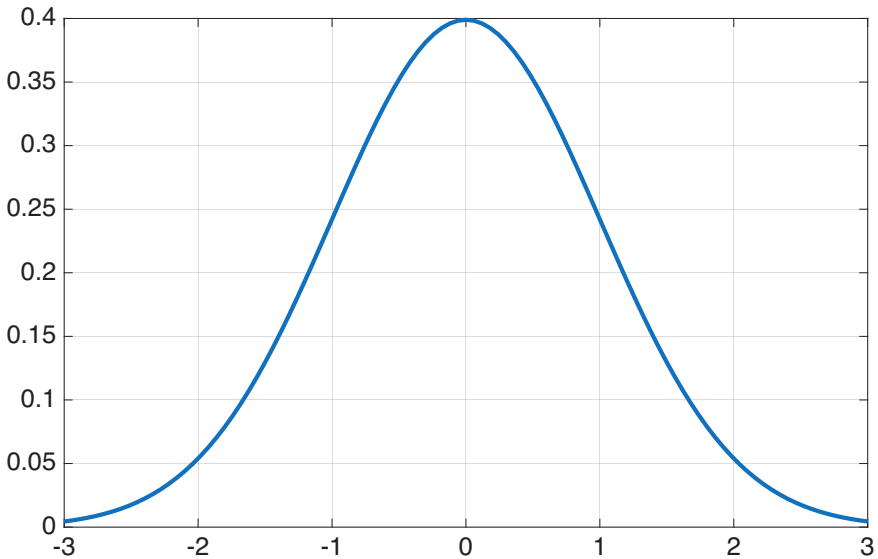
$$Z := \frac{X - \mu}{\sigma} \qquad \text{die Verteilung} \qquad Z \sim \mathsf{N}(0, 1).$$

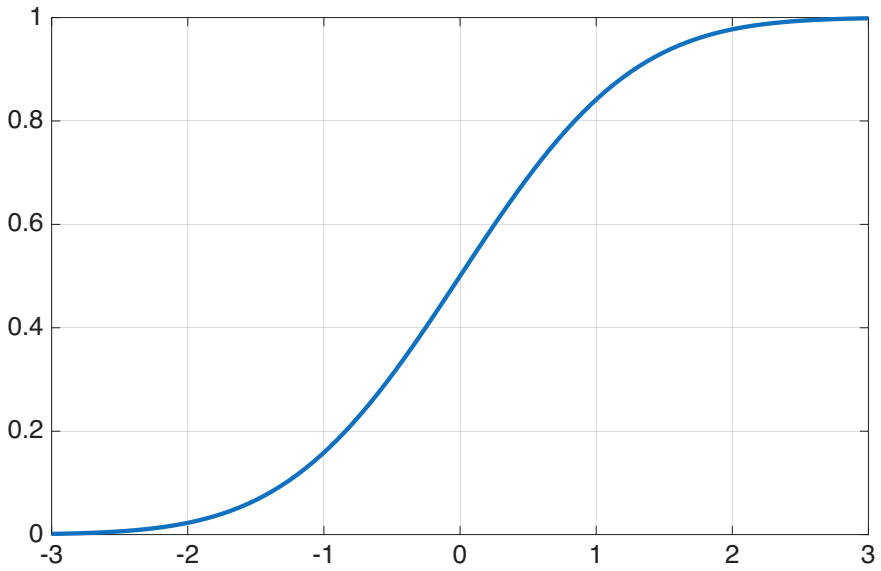
Entscheidend: Ist $X \sim \mathsf{N}(\mu, \sigma^2)$, so besitzt die Zufallsvariable

$$Z := \frac{X - \mu}{\sigma} \qquad \text{die Verteilung} \qquad Z \sim \mathsf{N}(0, 1).$$

Dichte und Verteilungsfunktion:

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}, \qquad \Phi(t) = \mathbf{P}(Z \leqslant t) = \int_{-\infty}^{t} \varphi(t)\,\mathrm{d}t.$$



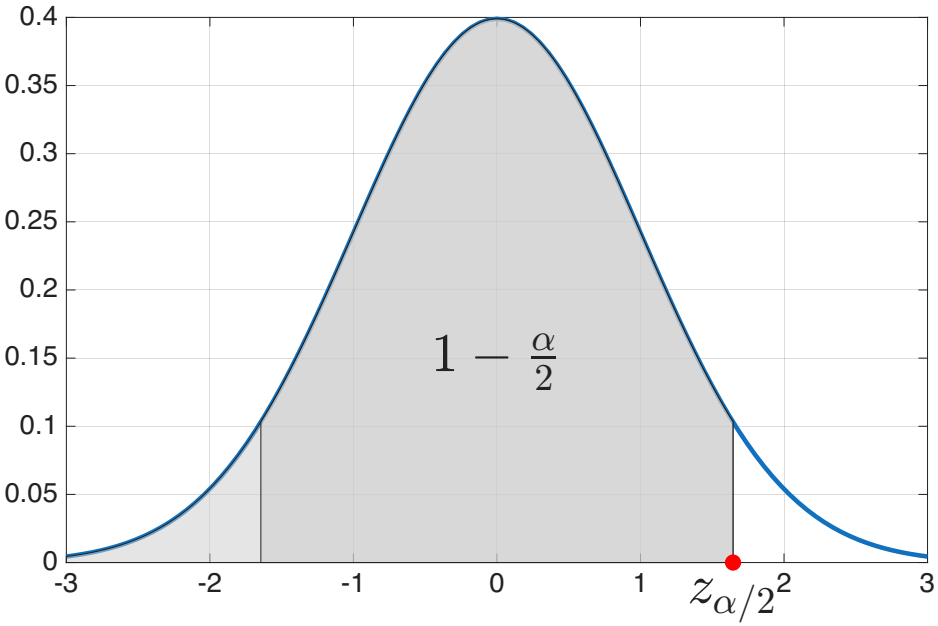


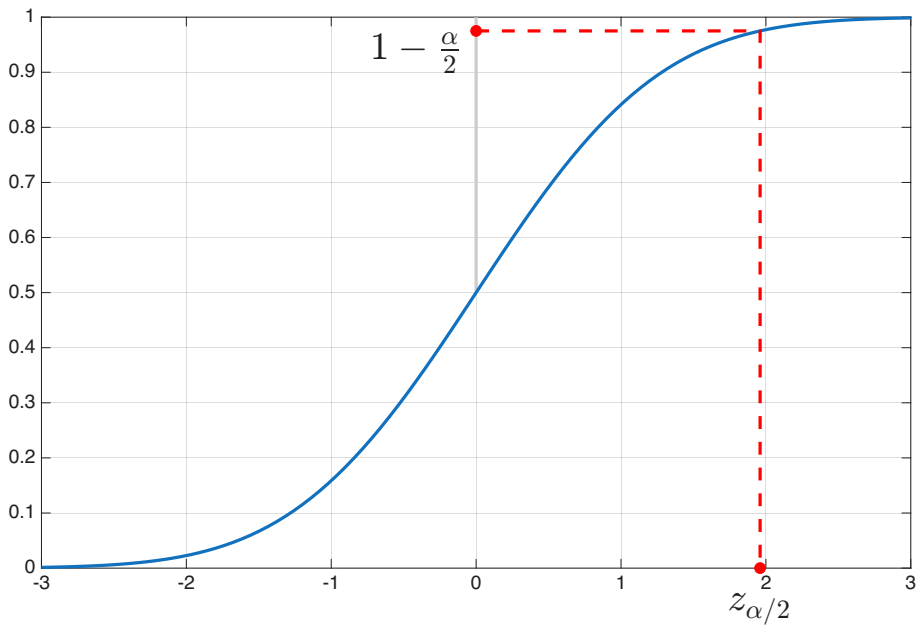
Wo liegt die Wahrscheinlichkeitsmasse?

Für $0 < \alpha < 1$ definiert man das Quantil $z_{\alpha/2}$ durch

$$\mathbf{P}(Z \leqslant z_{\alpha/2}) = \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2} \qquad \text{oder} \qquad z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Für $\alpha = 0,05$:





Plan

Schätzen und Konfidenzintervalle

Was möchte man schätzen?

- Ist die W-Verteilung einer Zufallsvariable X bekannt, etwa
  - $X \sim \mathrm{B}(p)$ (Bernoulli),
  - $X \sim \mathrm{B}(n, p)$ (binomial) oder
  - $X \sim \mathrm{N}(\mu, \sigma^2)$ (normal),

  so ergibt sich daraus die Wahrscheinlichkeit dafür, dass X in einem Intervall liegt.

Schätzen und Konfidenzintervalle

Was möchte man schätzen?

- Ist die W-Verteilung einer Zufallsvariable X bekannt, etwa
  - $X \sim \mathrm{B}(p)$ (Bernoulli),
  - $X \sim \mathrm{B}(n, p)$ (binomial) oder
  - $X \sim \mathrm{N}(\mu, \sigma^2)$ (normal),

  so ergibt sich daraus die Wahrscheinlichkeit dafür, dass X in einem Intervall liegt.
- Hat man hingegen Daten vorliegen, von denen man glaubt, sie entstammten einer bestimmten Verteilung, mit unbekannten Parametern, so liegt es nahe, diese Parameter aus den Daten zu gewinnen.

Schätzen und Konfidenzintervalle

Was möchte man schätzen?

- Ist die W-Verteilung einer Zufallsvariable X bekannt, etwa
  - $X \sim \mathrm{B}(p)$ (Bernoulli),
  - $X \sim \mathrm{B}(n, p)$ (binomial) oder
  - $X \sim \mathrm{N}(\mu, \sigma^2)$ (normal),

  so ergibt sich daraus die Wahrscheinlichkeit dafür, dass X in einem Intervall liegt.
- Hat man hingegen Daten vorliegen, von denen man glaubt, sie entstammten einer bestimmten Verteilung, mit unbekannten Parametern, so liegt es nahe, diese Parameter aus den Daten zu gewinnen.
- Beispiele:
  + Aus dem Ergebnis von n Münzwürfen $p = \mathbf{P}(\{\text{Kopf}\})$ rekonstruieren.
  + Aus n Messungen von Werkstückgewichten einer Produktionsserie Schwankungsbreite und Zentrum bestimmen.

Annahme: $X_1, \ldots, X_n$ i.i.d. Stichprobe, $X_j \sim \mathsf{N}(\mu, \sigma^2)$.

- Schätzer für Erwartungswert $\mu = \mathbf{E}[X]$: Stichprobenmittel

$$\hat{\mu} = \hat{\mu}(\omega) := \frac{X_1 + X_2 + \cdots + X_n}{n} = f(X_1, \ldots, X_n).$$

- $\hat{\mu}$ selbst Zufallsvariable mit ihrer eigenen Verteilung (Schätzverteilung). In diesem Fall ist dies ebenfalls eine Normalverteilung.
- Erwartungswert

$$\mathbf{E}[\hat{\mu}] = \mathbf{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n}\left(\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]\right) = \frac{n \cdot \mu}{n} = \mu.$$

  Dieser Schätzer ist damit erwartungstreu.

- Varianz

$$\mathbf{Var}\,\hat{\mu} = \mathbf{E}\left[(\hat{\mu} - \mu)^2\right] = \mathbf{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)\right)^2\right] = \cdots = \frac{\sigma^2}{n}.$$

- Fazit: $\hat{\mu} \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$

Gegeben: Realisierung $(x_1, \ldots, x_n)$ einer i.i.d.-normalverteilten Stichprobe $(X_1, \ldots, X_n)$
Schätzer liefert $\mu \approx \hat{\mu} = (x_1 + \cdots + x_n)/n$.

Frage: Wie stark differiert $\hat{\mu}$ von μ?

Gegeben: Realisierung $(x_1, \ldots, x_n)$ einer i.i.d.-normalverteilten Stichprobe $(X_1, \ldots, X_n)$
Schätzer liefert $\mu \approx \hat{\mu} = (x_1 + \cdots + x_n)/n$.
Frage: Wie stark differiert $\hat{\mu}$ von μ?

Wir kennen die Schätzverteilung $\hat{\mu} \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$. Damit ist $Z := \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$.

Für gegebenes $\alpha \in (0,1)$ seien $a \leqslant b$ so bestimmt, dass

$$1 - \alpha = \mathbf{P}\left(a \leqslant Z \leqslant b\right) = \Phi(b) - \Phi(a)$$

Da die Verteilung von Z symmetrisch zu Null ist, wählen wir $a = -b$, was zusammen mit $\Phi(-t) = 1 - \Phi(t)$ auf die Forderung

$$1 - \alpha = \Phi(b) - \Phi(-b) = \Phi(b) - 1 + \Phi(b) = 2\Phi(b) - 1$$

führt, oder

$$\Phi(b) = 1 - \frac{\alpha}{2} \qquad \text{d.h.} \qquad \boxed{b = z_{\alpha/2}}.$$

Gegeben: Realisierung $(x_1, \ldots, x_n)$ einer i.i.d.-normalverteilten Stichprobe $(X_1, \ldots, X_n)$
Schätzer liefert $\mu \approx \hat{\mu} = (x_1 + \cdots + x_n)/n$.
Frage: Wie stark differiert $\hat{\mu}$ von μ?

Wir kennen die Schätzverteilung $\hat{\mu} \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$. Damit ist $Z := \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$.

Für gegebenes $\alpha \in (0,1)$ seien $a \leqslant b$ so bestimmt, dass

$$1 - \alpha = \mathbf{P}\left(a \leqslant Z \leqslant b\right) = \Phi(b) - \Phi(a)$$

Somit ist wegen

$$a \leqslant Z \leqslant b \quad \Leftrightarrow \quad -z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2} \quad \Leftrightarrow \quad -z_{\alpha/2} \leqslant \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \leqslant z_{\alpha/2}$$

$$\Leftrightarrow \quad \hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

schließlich

$$1 - \alpha = \mathbf{P}\left(\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Wir erhalten also zum Konfidenzniveau $1-\alpha$ für den Verteilungsparameter μ einer normalverteilten Stichprobe vom Umfang n bei bekannter Varianz σ^2 das Konfidenzintervall

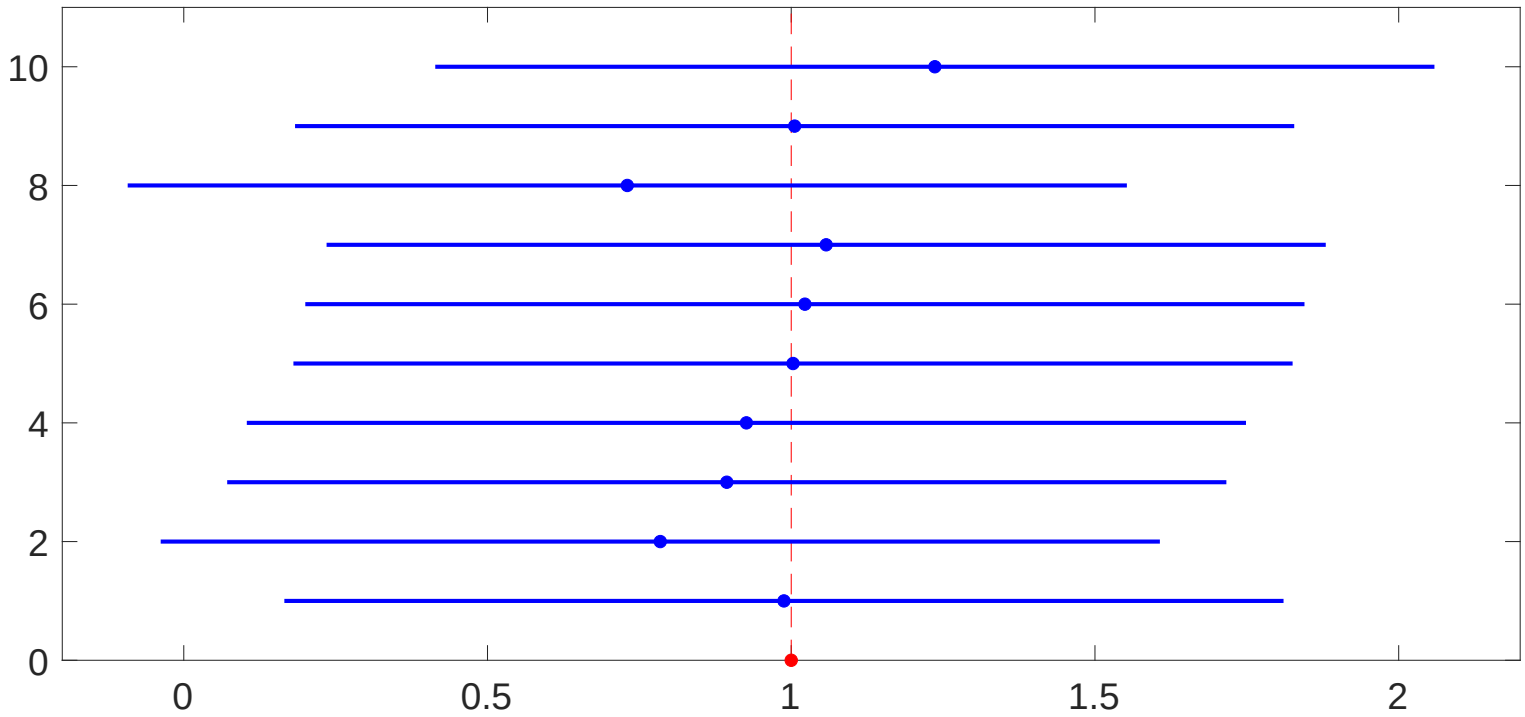
$$\left[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Wir erhalten also zum Konfidenzniveau $1-\alpha$ für den Verteilungsparameter μ einer normalverteilten Stichprobe vom Umfang n bei bekannter Varianz σ^2 das Konfidenzintervall

$$\left[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

- Ausgehend von einem festen Intervall $[a, b]$ für die Zufallsvariable Z ist dies nun ein zufälliges Intervall für die feste Größe μ.
- Die Intervallänge hängt (hier) nicht von $\hat{\mu}$, also auch nicht von der Stichprobe ab. Größere Stichproben führen zu engeren Intervallen.
- Analog leitet man einseitige Konfidenzintervalle her.
- Bei unbekannter Varianz σ^2 ändert sich die Schätzverteilung in eine t-Verteilung, welche auch vom Stichprobenumfang n abhängt.

Konfidenzintervalle für $\hat{\mu}$ für 10 Stichproben vom Umfang $n = 8$ aus $X \sim \mathsf{N}(1, \frac{1}{4})$



Ist $X \sim \mathsf{N}(\mu, \sigma^2)$ so ist $Z := \frac{X-\mu}{\sigma} \sim \mathsf{N}(0,1)$ und somit

$$\mathbf{P}(\mu - \ \sigma \leqslant X \leqslant \mu + \ \sigma) = \mathbf{P}(-1 \leqslant Z \leqslant 1) = 2\Phi(1) - 1 \approx 0.6827$$
$$\mathbf{P}(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma) = \mathbf{P}(-2 \leqslant Z \leqslant 2) = 2\Phi(2) - 1 \approx 0.9545$$
$$\mathbf{P}(\mu - 3\sigma \leqslant X \leqslant \mu + 3\sigma) = \mathbf{P}(-3 \leqslant Z \leqslant 3) = 2\Phi(3) - 1 \approx 0.9973$$

Ist $X \sim \mathsf{N}(\mu, \sigma^2)$ so ist $Z := \frac{X-\mu}{\sigma} \sim \mathsf{N}(0,1)$ und somit

$$\begin{aligned}
\mathbf{P}(\mu - \ \ \sigma \leqslant X \leqslant \mu + \ \ \sigma) &= \mathbf{P}(-1 \leqslant Z \leqslant 1) = 2\Phi(1) - 1 \approx 0.6827 \\
\mathbf{P}(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma) &= \mathbf{P}(-2 \leqslant Z \leqslant 2) = 2\Phi(2) - 1 \approx 0.9545 \\
\mathbf{P}(\mu - 3\sigma \leqslant X \leqslant \mu + 3\sigma) &= \mathbf{P}(-3 \leqslant Z \leqslant 3) = 2\Phi(3) - 1 \approx 0.9973
\end{aligned}$$

Runde Konfidenzniveaus $1-\alpha$ entsprechen krummen σ-Anteilen:

$$1 - \alpha = 2\Phi(z) - 1 \ \Leftrightarrow \ \Phi(z) = 1 - \frac{\alpha}{2} \ \Leftrightarrow \ z = \Phi^{-1}\left(\frac{\alpha}{2}\right) = z_{\alpha/2}$$

$$\mathbf{P}(\mu - z_{\alpha/2}\sigma \leqslant X \leqslant \mu + z_{\alpha/2}\sigma) = \mathbf{P}(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}) = \begin{cases} 0.9 & \alpha = 0.1 \\ 0.95 & \alpha = 0.05 \\ 0.99 & \alpha = 0.01 \end{cases}$$

$$z_{\alpha/2} = \begin{cases} 1.6449 & \alpha = 0.1 \\ 1.9600 & \alpha = 0.05 \\ 2.5758 & \alpha = 0.01 \end{cases}$$

Für den Erwartungswert μ einer normalverteilten Zufallsvariable wurde das 95% Konfidenzintervall $[0.1, 0.4]$ ermittelt. Welche der folgenden Aussagen treffen zu?

1. Mit Wahrscheinlichkeit 95% liegt μ in $[0.1, 0.4]$.
2. Die Wahrscheinlichkeit, dass $\mu > 0$ beträgt 95%.
3. Die Wahrscheinlichkeit, dass $\mu = 0$ beträgt weniger als 5%.
4. Bei wiederholter Durchführung der Schätzung mit gleich großen Stichproben derselben Verteilung würde μ in 95% der Fälle in $[0.1, 0.4]$ liegen.
5. 95% der Stichprobenwerte liegen in $[0.1, 0.4]$.

Lösung: Keine Antwort ist richtig.
Richtig wäre z.B.: Bei wiederholter Durchführung der Schätzung mit gleich großen Stichproben derselben Verteilung würden die Konfidenzintervalle in 95% der Fälle den wahren Erwartungswert μ enthalten.

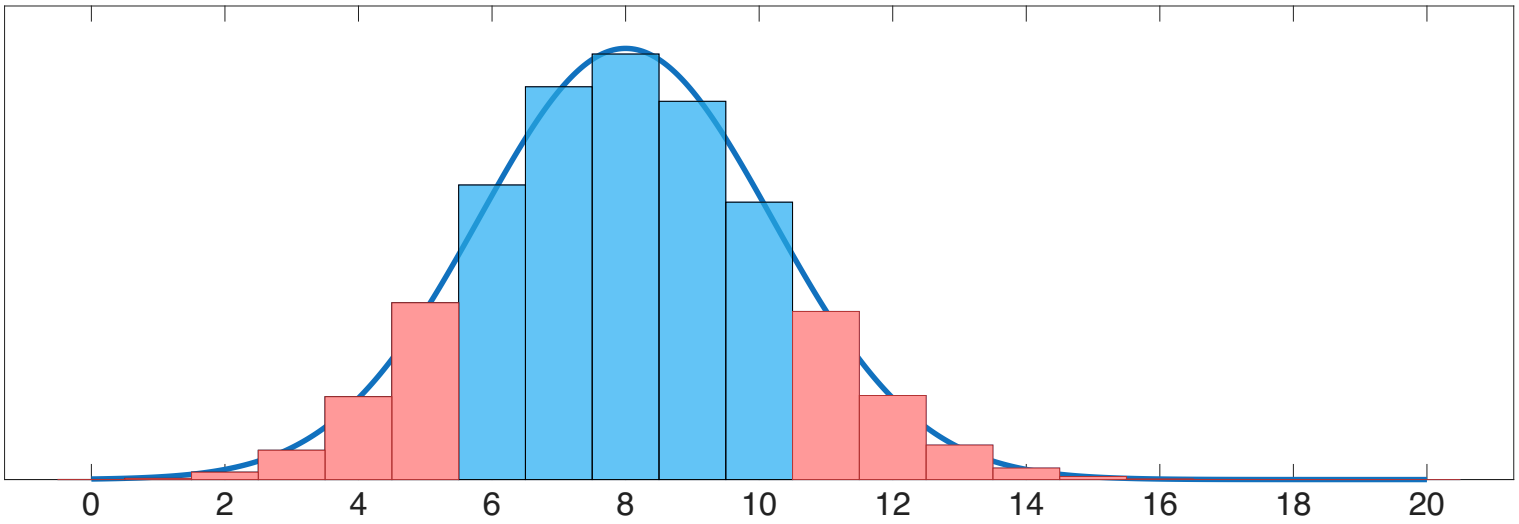
1. Dies ist die häufigste Fehlvorstellung. μ ist ein fester unbekannter Wert, keine Zufallsvariable. Die Wahrscheinlichkeitsaussage bezieht sich auf das Verfahren, nicht auf das konkrete Intervall. Sobald $[0.1, 0.4]$ berechnet ist, liegt μ entweder darin oder nicht – mit Wahrscheinlichkeit 0 oder 1, nicht 95%.
2. Aus demselben Grund: μ ist fest, also ist $\mathbf{P}(\mu > 0)$ entweder 0 oder 1. Allerdings kann man korrekt sagen, dass der Test $H_0 : \mu = 0$ auf Niveau $\alpha = 5\%$ verworfen wird, da 0 nicht im 95%-KI liegt. Aber das ist eine Aussage über das Testverfahren, nicht über die Wahrscheinlichkeit von $\mu > 0$.
3. Wieder dieselbe Fehlvorstellung, diesmal in frequentistischer Verkleidung. Außerdem ist $\mathbf{P}(\mu = 0)$ für einen stetigen Parameter ohnehin 0 — aber das ist nicht der eigentliche Punkt. Der eigentliche Punkt ist, dass dem deterministischen μ überhaupt keine Wahrscheinlichkeit zugeordnet werden kann.

Lösung: Keine Antwort ist richtig.
Richtig wäre z.B.: Bei wiederholter Durchführung der Schätzung mit gleich großen Stichproben derselben Verteilung würden die Konfidenzintervalle in 95% der Fälle den wahren Erwartungswert μ enthalten.

5. Falsch — aber subtil. Dies ist die häufigste fast richtige Antwort, und es lohnt sich, sie sorgfältig zu diskutieren. Das Intervall $[0.1, 0.4]$ ist fest – es ist das konkret beobachtete Intervall. Bei wiederholter Stichprobenziehung entstehen neue Intervalle, nicht dasselbe. Die korrekte Aussage wäre: 95% der so konstruierten Intervalle werden μ enthalten – aber diese Intervalle sind alle verschieden. $[0.1, 0.4]$ selbst wird in wiederholten Stichproben nicht reproduziert.
6. Dies verwechselt zwei grundlegend verschiedene Konzepte: Das Konfidenzintervall ist eine Aussage über die Lage des unbekannten Parameters μ – also des Verteilungsmittelpunkts – nicht über die Streuung der einzelnen Beobachtungen. Ein Intervall, das 95% der Datenwerte enthält, heißt Prognoseintervall und ist grundsätzlich breiter als das KI, da es zusätzlich zur Unsicherheit über μ auch die Streuung der Einzelbeobachtungen berücksichtigen muss. Bei großem Stichprobenumfang n wird das KI für μ sehr schmal — die Lage des Mittelpunkts ist dann gut bestimmt – während das Prognoseintervall stets eine Breite in der Größenordnung von σ behält, unabhängig von n. Die Verwechslung dieser beiden Intervalltypen ist besonders folgenreich in der medizinischen Praxis: ein sehr schmales KI für den mittleren Blutdruckeffekt eines Medikaments bedeutet nicht, dass das Medikament bei 95% der Patienten wirksam sein wird.

In sächsischen Schulbüchern werden Konfidenzintervalle nicht anhand der Normalverteilung eingeführt, sondern für die Binomialverteilung $\mathrm{B}(n, p)$.

- Gegeben ist eine Stichprobe aus n Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit $p \in [0, 1]$, deren Erfolgsanzahl X binomialverteilt ist, d.h. $X \sim \mathrm{B}(n, p)$.
- Der zu schätzende Parameter ist hier p, als Schätzer bietet sich an $\hat{p} = X/n$.



In sächsischen Schulbüchern werden Konfidenzintervalle nicht anhand der Normalverteilung eingeführt, sondern für die Binomialverteilung $\mathsf{B}(n,p)$.

- Gegeben ist eine Stichprobe aus n Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit $p \in [0,1]$, deren Erfolgsanzahl X binomialverteilt ist, d.h. $X \sim \mathsf{B}(n,p)$.
- Der zu schätzende Parameter ist hier p, als Schätzer bietet sich an $\hat{p} = X/n$.

Dies bringt eine Reihe von Verständnisproblemen mit sich:

- Beschränktes Parameterintervall $p \in [0,1]$.
- $\hat{p} \sim \frac{1}{n}\mathsf{B}(n,p)$, d.h. nimmt die Werte $0, \frac{1}{n}, \frac{2}{n} \ldots, 1$ an mit Wahrscheinlichkeit

$$\mathbf{P}\left(\hat{p} = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{1-k} \qquad k = 0, \ldots, n.$$

  Somit ist $\mathbf{E}\left[\hat{p}\right] = p$ und $\mathbf{Var}(\hat{p}) = p(1-p)/n$, die Varianz der Schätzverteilung hängt also vom zu schätzenden Parameter ab.
- $\hat{\mu}$ konnten wir überführen in eine Prüfgröße Z mit einfacher, vom zu schätzenden Parameter unabhängiger Verteilung (existiert hier nicht).

Man kann hier also gar kein sauberes Konfidenzintervall herleiten, sondern muss sich immer mit einer Approximation begnügen. Die gängigste ist nach Abraham Wald benannt.

- Man beginnt damit, die (diskrete) Binomialverteilung $\mathsf{B}(n,p)$ zu approximieren durch die (kontinuierliche) Normalverteilung $\mathsf{N}(np, np(1-p))$. Als Kriterium, wann dies zulässig ist, wird oft angegeben $np \geqslant 5$ und $n(1-p) \geqslant 5$. (Dies hängt aber vom (unbekannten) p ab!).
- Unter dieser Approximation ergibt sich die Schätzverteilung $\hat{p} \sim \mathsf{N}(p, p(1-p)/n))$. Auch diese hängt vom unbekannten p ab. Erst wenn man hier p durch das berechnete $\hat{p}$ ersetzt (eine weitere Approximation!) ergibt sich die Prüfgröße

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim \mathsf{N}(0,1).$$

- Ab hier erhält man auf dieselbe Weise wir für μ bei der Normalverteilung das Konfidenzintervall zum Niveau $1-\alpha$ für p als

$$W = W(\alpha, \hat{p}, n) := \left[\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right].$$

Genauigkeit:

- Approximation $\mathsf{B}(n,p) \approx \mathsf{N}(np, np(1-p))$ nur hinreichend gut, falls sowohl np als auch $n(1-p)$ hinreichend groß, ohne Kenntnis von p nicht verifizierbar.
- Ersetzen von p durch $\hat{p}$ in Varianz in der Schätzverteilung eine weitere Approximation.
- Resultierendes Konfidenzintervall (KI) kann durchaus über $(0,1)$ hinausragen.
- KI liegt symmetrisch um $\hat{p}$, die wahre Schätzverteilung von $\hat{p}$ ist jedoch für $p \neq \frac{1}{2}$ unsymmetrisch, eine weitere Approximation.

Didaktisch:

- In keinem der mir zugänglichen Schulbücher[1] werden KI hergeleitet (wie es im Fall von $\hat{\mu}$ bei der Normalverteilung sehr einfach möglich ist).
- Stattdessen werden KI nur für die Binomialverteilung angegeben (Wald-Intervall), aber ohne Herleitung, und ohne Erwähnung der zugrundeliegenden Approximationen. Hier müssen Kochrezepte auswendig gelernt werden.
- Oft wird die Normalverteilung erst danach eingeführt.

[1] EdM, Fundamente der Mathematik, Lambacher–Schweitzer

Wie ungenau ist das Wald-Intervall?[2] Man möchte $\mathbf{P}(p \in W(\alpha, \hat{p}, n)) = 1 - \alpha$.

[2] L. D. Brown, T. T. Cai und A. DasGupta. "Interval Estimation for a Binomial Proportion". In: *Statistical Science* 16.2 (2001), S. 101–133. URL: http://www.jstor.org/stable/2676784.

Wie ungenau ist das Wald-Intervall?[2] Man möchte $\mathbf{P}(p \in W(\alpha, \hat{p}, n)) = 1 - \alpha$.

Abdeckungswahrscheinlichkeit (coverage)

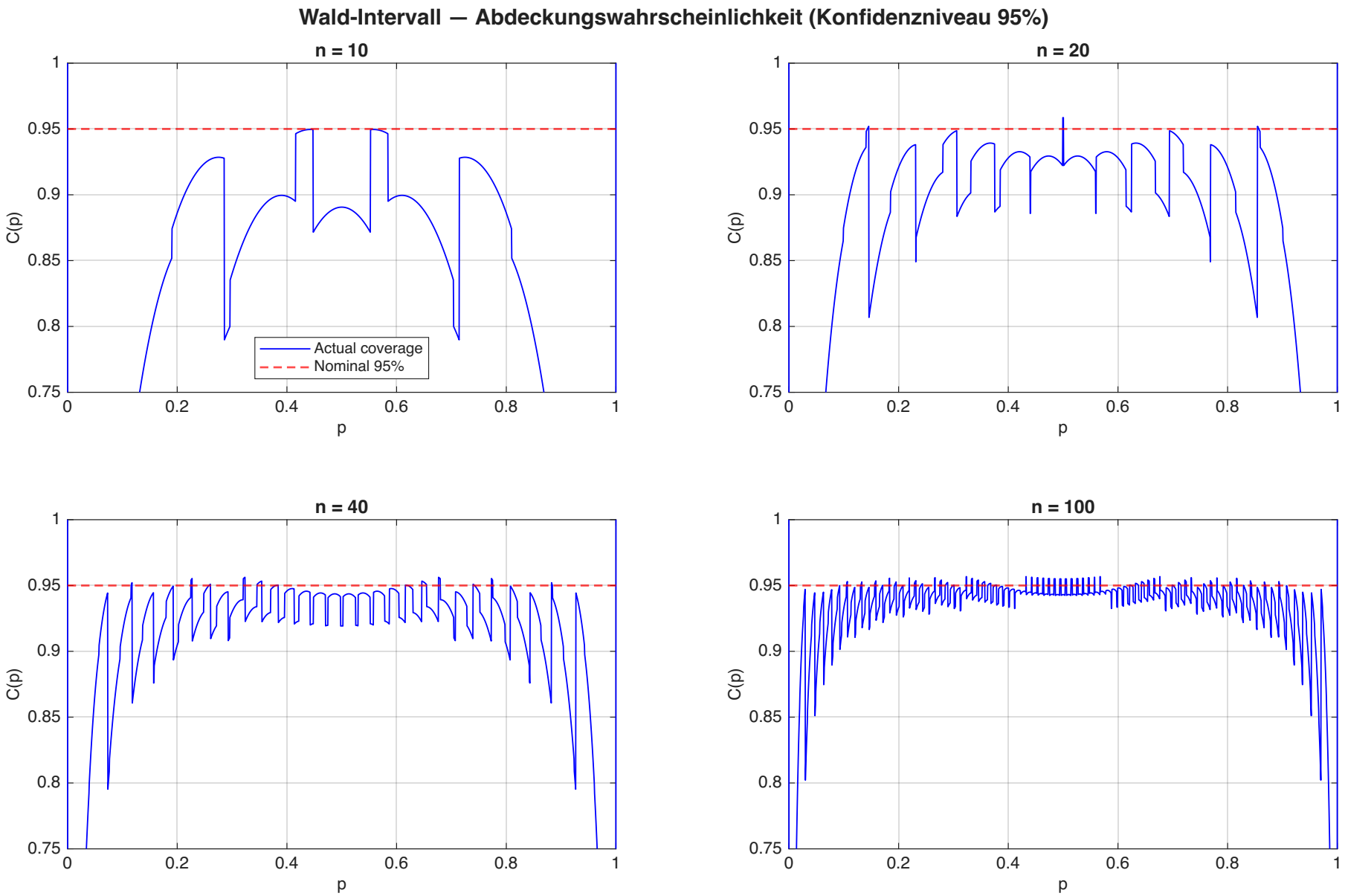
$$C(p) := \mathbf{P}\left(p \in W(\alpha, \hat{p}, n)\right) = \sum_{k \in A(p)} \binom{n}{k} p^k (1-p)^k, \qquad p \in [0, 1],$$

wobei $A(p) := \{k : p \in W(\alpha, \hat{p}, n) \text{ falls } X = k\}$.

Konsequenz:

- $C(p)$ ist ein Polynom in p, somit glatt, aber $A(p)$ ändert sich unstetig mit p (Sprünge).
- Mit wachsendem n werden die Sprünge lediglich dichter.
- Aus dem zentralen Grenzwertsatz ergibt sich $C(p) \to 1 - \alpha$ für $n \to \infty$, aber nicht gleichmäßig in p. Für jedes n gibt es also p-Werte, bei denen die Abdeckung deutlich unter $1 - \alpha$ liegt.
- Jede Menge Wald-Alternativen (Wilson, Clopper–Pearson, Agresti & Coull, Jeffreys, . . .)

[2] L. D. Brown, T. T. Cai und A. DasGupta. "Interval Estimation for a Binomial Proportion". In: *Statistical Science* 16.2 (2001), S. 101–133. URL: http://www.jstor.org/stable/2676784.


Wald-Intervall — Abdeckungswahrscheinlichkeit (Konfidenzniveau 95%)
n = 10
n = 20
n = 40
n = 100
C(p)
p
Actual coverage
Nominal 95%
1
0.95
0.9
0.85
0.8
0.75
0
0.2
0.4
0.6
0.8
1

Plan

Testen und Signifikanz

Zwei Schulen

In vielen Lehrbüchern (auch den sächsischen Schulbüchern) werden Hypothesentests als Mischmasch zwischen den zwei höchst unterschiedlichen Zugängen von einerseits Ronald A. Fisher und andererseits Jerzey Neyman und Egon Pearson dargestellt.

> *„Der Ausdruck ‚Fehler zweiter Art', obwohl scheinbar ein harmloses Stück Fachjargon, ist nützlich als Hinweis auf die Art der geistigen Verwirrung, in der er geprägt wurde."*
>
> *R. A. Fisher (1955),[3] p. 73, zitiert in Neyman (1961),[4] p. 151*

> *„Eine Unterscheidung ohne Unterschied wurde von gewissen Autoren eingeführt... Es zeugt von großem Vertrauen in die Unwissenheit der Studierenden, einen solchen Anspruch vorzubringen."*
>
> *[R. A. Fisher 1956, p. 141, zitiert in Neyman 1961, p. 151]*

> *„Sir Ronald Fishers Nachkriegspolemiken stellen ein ungewöhnliches Phänomen dar, das in der Geschichte der Wissenschaft wahrscheinlich seinesgleichen sucht."*
>
> *[J. Neyman 1961, p. 152]*

---

[3] R. A. Fisher. "Statistical Methods and Scientific Induction". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 17.1 (1955), S. 69–78. URL: http://www.jstor.org/stable/2983785.

[4] J. Neyman. "Silver Jubilee of My Dispute with Fisher". In: *Journal of the Operations Research Society of Japan* 3.4 (1961), S. 145–154.

- Für Zufallsvariable $X \sim \mathsf{N}(\mu, 1)$, μ unbekannt, soll die Nullhypothese

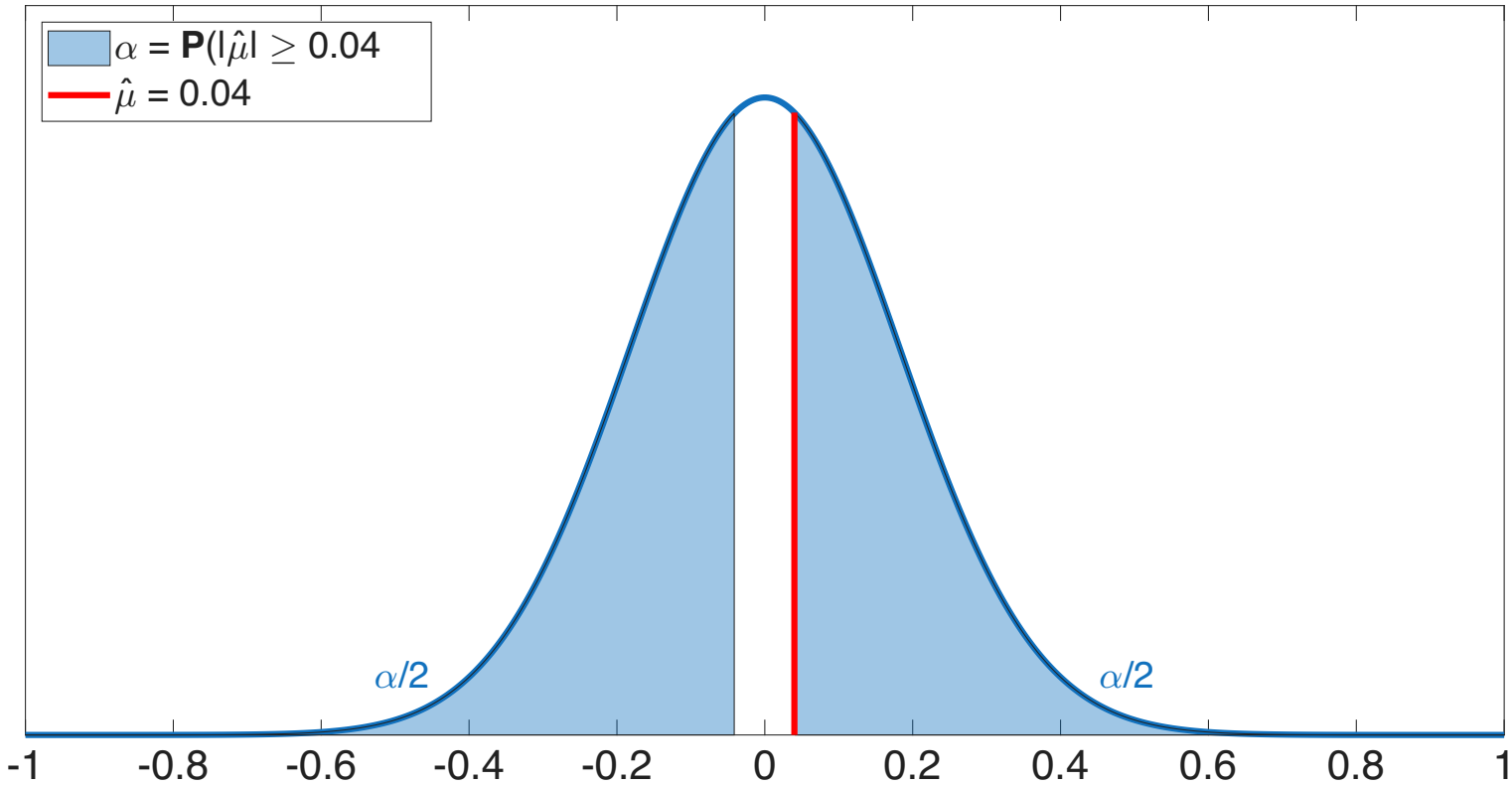
$$H_0 : \mu = 0$$

  anhand einer Stichprobe vom Umfang n verworfen werden (oder nicht).
- Als Testgröße ziehen wir $\hat{\mu}$ heran.
- Wir wissen bereits $\hat{\mu} \sim \mathsf{N}(\mu, \sigma^2/n) = \mathsf{N}(0, 1/n)$.
- Im Fall, dass H_0 zutrifft, bedeutet dies $\hat{\mu} \sim \mathsf{N}(0, 1/30)$.
- Für $n = 30$ und $\hat{\mu} = 0.04$: mit welcher Wahrscheinlichkeit α hat $|\hat{\mu}|$ einen Abstand von 0 von mindestens 0.04?

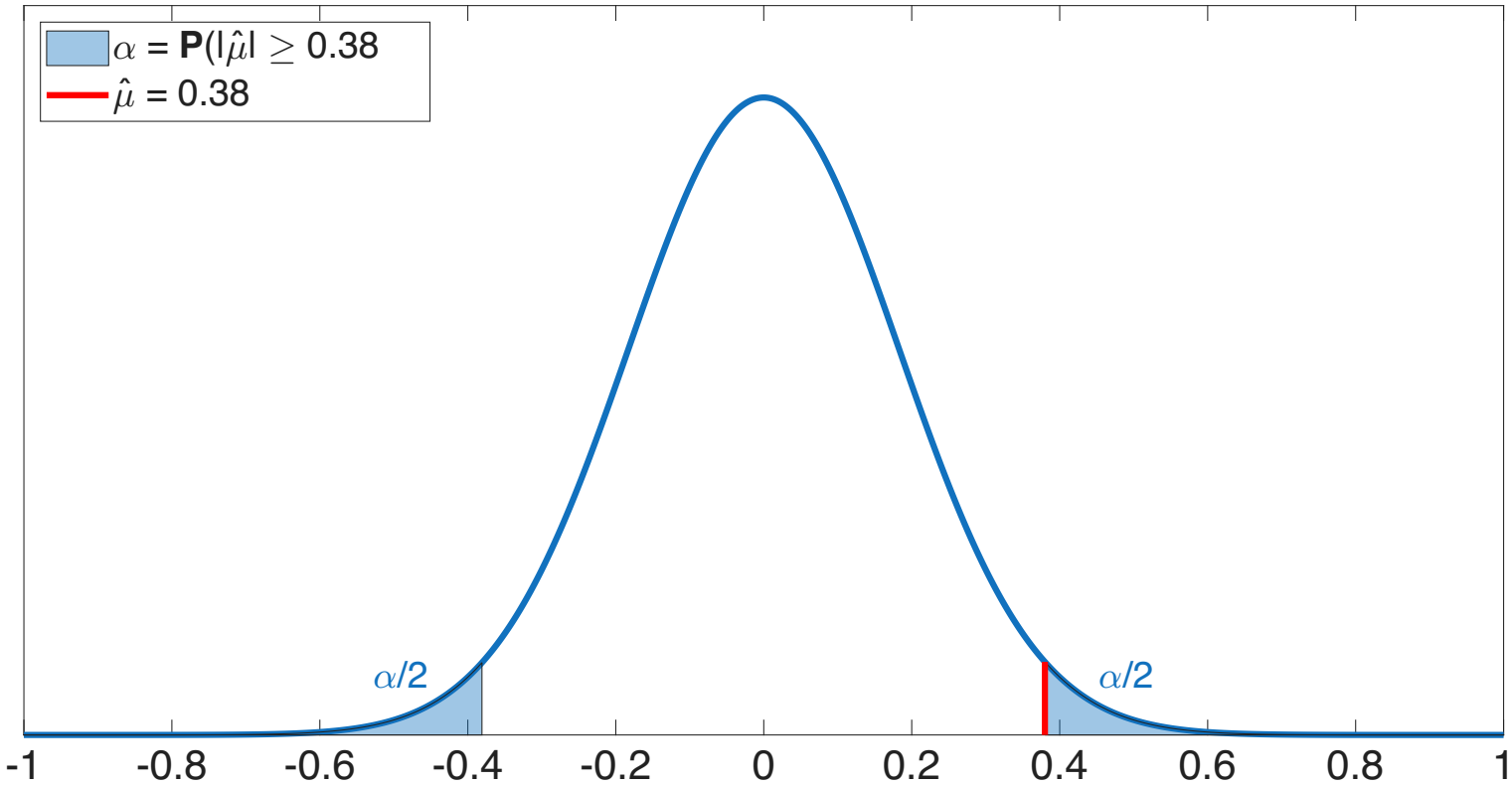
$$\begin{aligned}
\mathbf{P}(|\hat{\mu}| \geqslant 0.04) &= \mathbf{P}(\hat{\mu} \leqslant -0.04) + \mathbf{P}(\hat{\mu} \geqslant 0.04) \\
&= \mathbf{P}(\sqrt{30}\hat{\mu} \leqslant -\sqrt{30} \cdot 0.04) + 1 - \mathbf{P}(\sqrt{30}\hat{\mu} \leqslant \sqrt{30} \cdot 0.04) \\
&= 2\left[1 - \Phi(0.04\sqrt{30})\right] \approx 0.8266.
\end{aligned}$$

- H_0 wird man kaum verwerfen können.

$$n = 30, \quad \hat{\mu} = 0.04, \quad \mathbf{P}(|\hat{\mu}| \geqslant 0.04) \approx 0.83$$



$$n = 30, \quad \hat{\mu} = 0.38, \quad \mathbf{P}(|\hat{\mu}| \geqslant 0.38) \approx 0.037$$



Sie testen eine Nullhypothese H_0 anhand einer Stichprobe. Es stellt sich heraus, dass das Ergebnis der Stichprobe im Verwerfungsbereich von H_0 liegt, und zwar auf einem Signifikanzniveau von 5%. Welche der folgenden Aussagen lassen sich aus diesem Ergebnis folgern?

1. Es ist bewiesen, dass H_0 falsch ist.
2. Die Wahrscheinlichkeit, dass H_0 gilt, beträgt höchstens 5%.
3. Es ist bewiesen, dass die Verneinung von H_0 gilt.
4. Man kann die Wahrscheinlichkeit berechnen, dass die Verneinung von H_0 gilt.
5. Entscheidet man sich nun, H_0 abzulehnen, dann beträgt die Wahrscheinlichkeit höchstens 5%, dass man sich dabei irrt.
6. Wenn man H_0 sehr oft testet, dann erhält man in etwa 5% der Fälle ein signifikantes Ergebnis.

Testen und Signifikanz

Interpretation

Lösung: Keine der Antworten ist richtig.

1. Ein signifikantes Ergebnis beweist nichts. Selbst wenn H_0 wahr ist, landet das Ergebnis mit Wahrscheinlichkeit $\alpha = 5\%$ im Verwerfungsbereich – das ist die Definition des Signifikanzniveaus. Ein einzelnes signifikantes Ergebnis könnte genau dieser Fall sein.
2. Dies ist die häufigste und folgenreichste Fehlvorstellung. Der p-Wert ist $\mathbf{P}(\text{Daten}|H_0)$ — die Wahrscheinlichkeit, ein so extremes oder extremeres Ergebnis zu beobachten, wenn H_0 wahr ist. Er ist ausdrücklich nicht $\mathbf{P}(H_0|\text{Daten})$ – die Wahrscheinlichkeit, dass H_0 gilt, gegeben die beobachteten Daten. Diese Umkehrung der Konditionierung ist logisch unzulässig ohne Kenntnis einer A-Priori-Wahrscheinlichkeit für H_0 – also ohne ein Bayes'sches Modell.
3. Aus denselben Gründen wie (1). Darüberhinaus: im Fisherschen Rahmen gibt es gar keine explizit formulierte Gegenhypothese; das Verwerfen von H_0 bedeutet lediglich, dass die Daten unter H_0 unwahrscheinlich sind, nicht dass eine bestimmte Alternative bestätigt wird.
4. Im frequentistischen Rahmen – sowohl bei Fisher als auch bei Neyman–Pearson – sind Hypothesen keine Zufallsvariablen und haben keine Wahrscheinlichkeiten. Eine solche Aussage wäre nur im Bayes'schen Rahmen möglich, und selbst dort nur mit einer explizit gewählten A-Priori-Verteilung.

Lösung: Keine der Antworten ist richtig.

5. Dies ist die subtilste Fehlvorstellung der Liste und vermutlich diejenige, die am häufigsten von statistisch halbwegs gebildeten Anwendern gehalten wird. Das Signifikanzniveau α ist die Wahrscheinlichkeit, H_0 fälschlicherweise zu verwerfen wenn H_0 wahr ist – also $\mathbf{P}(\text{Verwerfung}|H_0)$ und nicht die Wahrscheinlichkeit, dass die getroffene Entscheidung falsch ist – also nicht $\mathbf{P}(H_0|\text{Verwerfung})$. Letzteres hängt von der Prävalenz wahrer Null -hypothesen in dem betreffenden Forschungsfeld ab und kann erheblich größer als α sein.
6. Die am schwersten zu widerlegende falsche Option: sie wäre korrekt unter der Zusatzbedingung, dass H_0 in allen Fällen wahr ist. Ohne diese Bedingung ist sie falsch: die Rate signifikanter Ergebnisse hängt sowohl von α als auch von der Trennschärfe der Tests und der Prävalenz tatsächlich falscher Nullhypothesen ab. Ist H_0 in vielen Fällen tatsächlich falsch und die Trennschärfe hoch, wird man weit mehr als 5% signifikante Ergebnisse erhalten. Option (6) beschreibt also nur den uninteressanten Spezialfall, in dem alle Nullhypothesen wahr sind.

Fisher

- Es geht nur um H_0 und die Frage: geben die vorliegenden Daten (Beobachtungen) hinreichend Anlass, diese in Zweifel zu ziehen/zu widerlegen.
- Da man H_0 nur verwerfen (nicht aber annehmen) kann, ist keine Alternativhypothese erforderlich, und ebensowenig Konzepte wie Fehler 1./2. Art, Trennschärfe (power).

Neyman–Pearson

- Man kann nicht etwas verwerfen ohne anzugeben, was man stattdessen zu akzeptieren bereit ist. Dies macht eine Alternativhypothese H_1 erforderlich, unter welcher die Verteilung der Teststatistik bekannt ist. Man hat

$$\alpha = \mathbf{P}(H_0 \text{ wird verworfen } \mid H_0) \qquad \text{(Fehler 1. Art)}$$
$$\beta = \mathbf{P}(H_0 \text{ wird nicht verworfen} \mid H_1) \qquad \text{(Fehler 2. Art)}$$
$$1-\beta : \text{Trennschärfe (power)}$$

Fisher

- Der *p*-Wert wird bestimmt nach Vorliegen der Beobachtungen.
- Stetiges Maß der „Beweislast“ gegen H_0: je kleiner p, desto inkonsistenter die Beobachtungen mit H_0.
- Fisher sah nie einen festen Schwellwert (5%) vor, sondern überließ dies dem Urteilsvermögen des Ausführenden.

Neyman–Pearson

- α : asymptotische Rate für Fehler 1. Art (falsch positiv), welche man bereit ist zu tolerieren bei wiederholter Erhebung der Daten.
- Wird vor Datenerhebung als Teil der Versuchsplanung festgelegt.
- Entscheidung für H_0/H_1 nach Vergleich der Prüfgröße mit kritischem Bereich; es wird kein *p*-Wert berechnet.

Mischmasch

- Durchführung nach Neyman–Pearson und Interpretation nach Fisher.
- $\alpha = 5\%$ a priori festgelegt (NP), danach *p*-Wert bestimmen und als Evidenz gegen H_0 interpretieren (F).

Fisher

- Hypothesentest Werkzeug für induktives Schließen; bestehende Hypothesen werden anhand neuer Daten überprüft.
- „Was sagen uns die Daten?“ (Wissenschaftliche Frage)

Neyman–Pearson

- Hypothesentest Werkzeug für induktives Verhalten: Entscheiden für H_0 oder H_1 je nach Prüfgröße sichert bei beliebig ofter Wiederholung Fehlerraten α bzw. β.
- „Was soll man tun angesichts einer vor Datenerhebung festgelegten Entscheidungsregel“. (Regel, etwa bei Qualitätskontrolle)
- Fisher lehnte dies ab.

Kann man die Nullhypothese annehmen?

Fisher

- **Nein**, allenfalls Verwerfen.
  Hypothesen vorläufig, können stets durch Daten falsifiziert werden.

Neyman–Pearson

- **Ja** in dem Sinne, dass man so handelt als sei H_0 wahr.

Kann man die Nullhypothese annehmen?

Fisher

- **Nein**, allenfalls Verwerfen.
  Hypothesen vorläufig, können stets durch Daten falsifiziert werden.

Neyman–Pearson

- **Ja** in dem Sinne, dass man so handelt als sei H_0 wahr.

Man hätte gerne, dass α bzw. p die Wahrscheinlichkeit angibt, dass H_0 bzw. H_1 wahr ist. Dies liefern **beide** Methoden aber **nicht**.

Gegeben: X normalverteilt, Varianz σ^2, Stichprobenumfang n, Stichprobenmittel $\hat{\mu}$
Gesucht: Test für unbekannten Erwartungswert μ

Gegeben: X normalverteilt, Varianz σ^2, Stichprobenumfang n, Stichprobenmittel $\hat{\mu}$
Gesucht: Test für unbekannten Erwartungswert μ

Unter $H_0 : \mu = \mu_0$ gilt

$$\hat{\mu} \sim \mathsf{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \quad \Leftrightarrow \quad Z := \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1).$$

Gegeben: X normalverteilt, Varianz σ^2, Stichprobenumfang n, Stichprobenmittel $\hat{\mu}$
Gesucht: Test für unbekannten Erwartungswert μ

Unter $H_0 : \mu = \mu_0$ gilt

$$\hat{\mu} \sim \mathsf{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \quad \Leftrightarrow \quad Z := \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1).$$

Unter der einfachen Alternativhypothese $H_1 : \mu = \mu_1$ $(\mu_1 > \mu_0)$ gilt

$$Z \sim \mathsf{N}(\delta, 1) \qquad \text{mit} \qquad \delta = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \quad \text{(standardisierte Effektgröße)} .$$

Gegeben: X normalverteilt, Varianz σ^2, Stichprobenumfang n, Stichprobenmittel $\hat{\mu}$
Gesucht: Test für unbekannten Erwartungswert μ

Unter $H_0 : \mu = \mu_0$ gilt

$$\hat{\mu} \sim \mathsf{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \quad \Leftrightarrow \quad Z := \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1).$$

Unter der einfachen Alternativhypothese $H_1 : \mu = \mu_1$ $(\mu_1 > \mu_0)$ gilt

$$Z \sim \mathsf{N}(\delta, 1) \qquad \text{mit} \qquad \delta = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \quad \text{(standardisierte Effektgröße)} .$$

Fehlerarten nach NP:

	H_0 wahr	H_1 wahr
H_0 nicht verwerfen	✓	✗ 2. Art (β)
H_0 verwerfen	✗ 1. Art (α)	✓

Gegeben: X normalverteilt, Varianz σ^2, Stichprobenumfang n, Stichprobenmittel $\hat{\mu}$
Gesucht: Test für unbekannten Erwartungswert μ

Unter $H_0 : \mu = \mu_0$ gilt

$$\hat{\mu} \sim \mathsf{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \quad \Leftrightarrow \quad Z := \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1).$$

Unter der einfachen Alternativhypothese $H_1 : \mu = \mu_1$ $(\mu_1 > \mu_0)$ gilt

$$Z \sim \mathsf{N}(\delta, 1) \qquad \text{mit} \qquad \delta = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \quad \text{(standardisierte Effektgröße)} .$$

Fehlerarten nach NP:

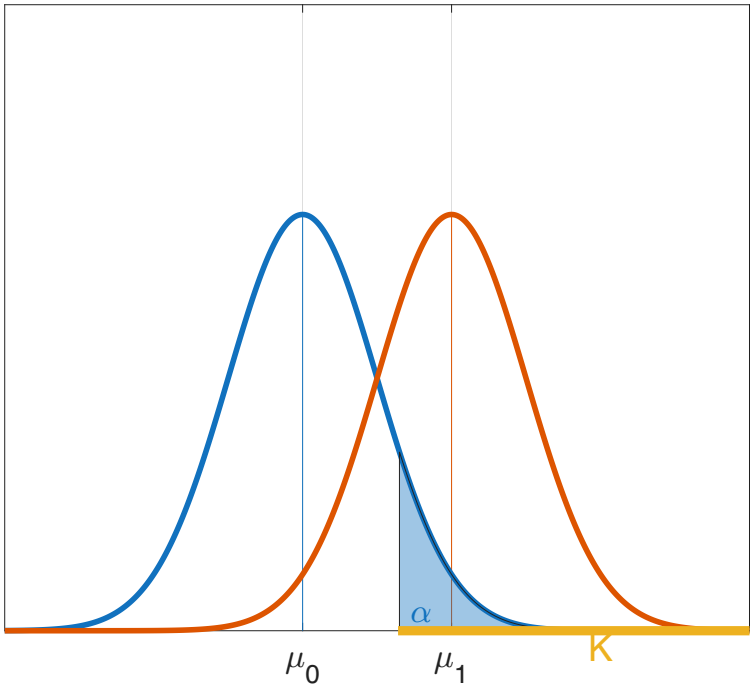
	H_0 wahr	H_1 wahr
H_0 nicht verwerfen	✓	✗ 2. Art (β)
H_0 verwerfen	✗ 1. Art (α)	✓

Frage: Wie Verwerfungsbereich K wählen, damit $\mathbf{P}(Z \in K | H_0) = \alpha$?

Frage: Wie Verwerfungsbereich K wählen, damit $\mathbf{P}(Z \in K | H_0) = \alpha$?

Dies ist genau dann der Fall, wenn

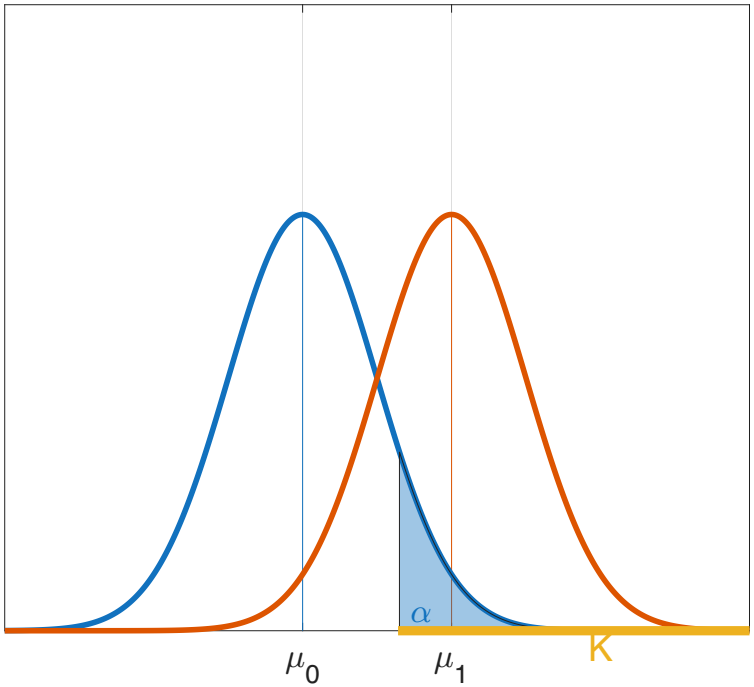
$$Z \in [z_\alpha, \infty) \quad \Leftrightarrow \quad \hat{\mu} \in \left[\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}, \infty \right) .$$



Frage: Wie Verwerfungsbereich K wählen, damit $\mathbf{P}(Z \in K | H_0) = \alpha$?

Dies ist genau dann der Fall, wenn

$$Z \in [z_\alpha, \infty) \quad \Leftrightarrow \quad \hat{\mu} \in \left[\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}, \infty \right) .$$



Die Trennschärfe (power) ist dann gegeben durch

$$1 - \beta = \mathbf{P}(\hat{\mu} \in K | H_1) = \mathbf{P}\left(\hat{\mu} \geqslant \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1 \right)$$

Wir stellen um

$$1-\beta = \mathbf{P}\left(\hat{\mu} \geqslant \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \,\middle|\, \mu = \mu_1\right) = \mathbf{P}\left(\frac{\hat{\mu}-\mu_1}{\sigma/\sqrt{n}} \geqslant \frac{\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}} \,\middle|\, \mu = \mu_1\right),$$

beachten

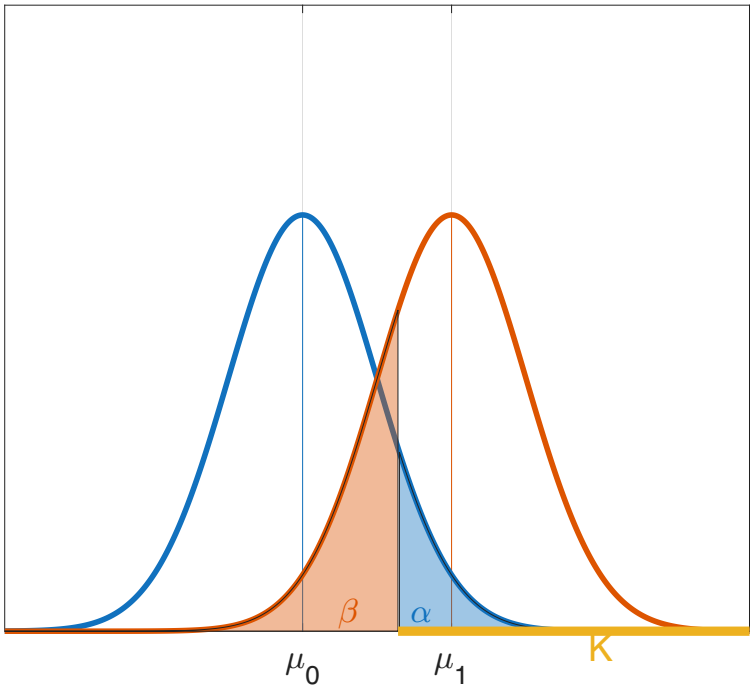
$$\frac{\hat{\mu}-\mu_1}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \qquad \text{sowie} \qquad \frac{\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}} = z_\alpha - \delta, \quad \delta = \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}},$$

und erhalten so

$$\boxed{1-\beta = \mathbf{P}(Z \geqslant z_\alpha - \delta) = 1 - \Phi(z_\alpha - \delta).}$$

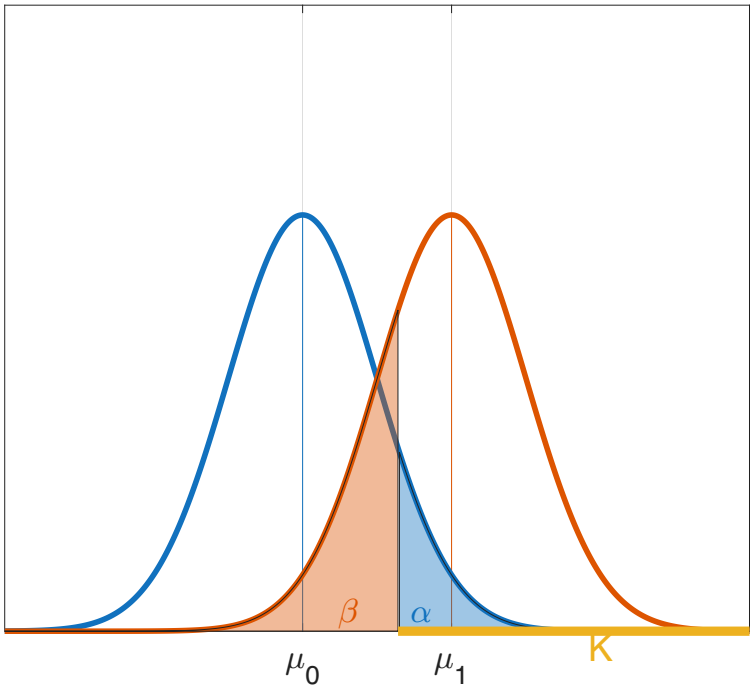
Fehler 2. Art:

$$\beta = \mathbf{P}(Z \notin K | H_1) = \Phi(z_\alpha - \delta)$$



Fehler 2. Art:

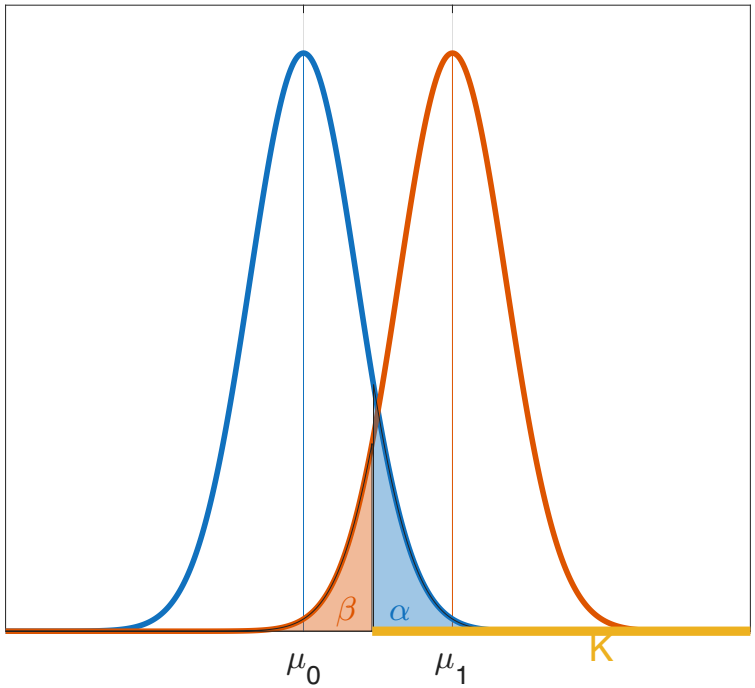
$$\beta = \mathbf{P}(Z \notin K | H_1) = \Phi(z_\alpha - \delta)$$



Wie kann man β verringern bei unverändertem α?

Fehler 2. Art:

$$\beta = \mathbf{P}(Z \notin K | H_1) = \Phi(z_\alpha - \delta)$$



Wie kann man β verringern bei unverändertem α?
Wir erinnern uns:

$$\hat{\mu} \sim \mathsf{N}\left(\mu_i, \frac{\sigma^2}{n}\right)$$

und erhöhen den Stichprobenumfang n.

Oft kennt man μ_1 nicht und formuliert

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu > \mu_0 .$$

Oft kennt man μ_1 nicht und formuliert

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu > \mu_0.$$

- H_1 ist nun zusammengesetzt und umfasst eine ganze Schar von Verteilungen, parametrisiert durch μ.

Oft kennt man μ_1 nicht und formuliert

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu > \mu_0.$$

- H_1 ist nun zusammengesetzt und umfasst eine ganze Schar von Verteilungen, parametrisiert durch μ.
- Verwerfungsbereich $K = [z_\alpha, \infty)$ hängt nun von μ ab (kein festes μ_1 mehr).

Oft kennt man μ_1 nicht und formuliert

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu > \mu_0.$$

- H_1 ist nun zusammengesetzt und umfasst eine ganze Schar von Verteilungen, parametrisiert durch μ.
- Verwerfungsbereich $K = [z_\alpha, \infty)$ hängt nun von μ ab (kein festes μ_1 mehr).
- Die Trennschärfe als Funktion des (wahren aber unbekannten) μ nennt man Gütefunktion (power function)

$$G(\mu) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right), \qquad \mu > \mu_0.$$

  - $G(\mu_0) = \alpha$ (Nullhypothese trifft zu)
  - G ist streng monoton wachgsend in μ
  - $G(\mu) \to 1$ für $\mu \to \infty$ (weit auseinenderliegende Alternativen werden sicher entdeckt).

Oft kennt man μ_1 nicht und formuliert

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu > \mu_0.$$

- H_1 ist nun zusammengesetzt und umfasst eine ganze Schar von Verteilungen, parametrisiert durch μ.
- Verwerfungsbereich $K = [z_\alpha, \infty)$ hängt nun von μ ab (kein festes μ_1 mehr).
- Die Trennschärfe als Funktion des (wahren aber unbekannten) μ nennt man Gütefunktion (power function)

$$G(\mu) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right), \qquad \mu > \mu_0.$$

  - $G(\mu_0) = \alpha$ (Nullhypothese trifft zu)
  - G ist streng monoton wachgsend in μ
  - $G(\mu) \to 1$ für $\mu \to \infty$ (weit auseinenderliegende Alternativen werden sicher entdeckt).
- Die Funktion $OC(\mu) := 1 - G(\mu)$ wird als Operationscharakteristik bezeichnet.
- Unter allen Tests mit Niveau α liefert dieser für jedes $\mu > \mu_0$ die maximale Trennschärfe (gleichmäßig bester Test, UMP).

Bezug zu Konfidenzintervallen: Bezeichnet

- $K_\alpha(\mu_0)$ den Verwerfungsbereich beim Testen von $H_0 : \mu = \mu_0$ gegen $H_1 : \mu > \mu_0$,
- $\mathrm{CI}_{1-\alpha}(\hat{\mu})$ das (einseitige) Konfidenzintervall der Schätzung $\hat{\mu}$,
- beide für gleiches α,
- so gilt

$$\hat{\mu} \in K_\alpha(\mu_0) \quad \Leftrightarrow \quad \mu_0 \notin \mathrm{CI}_{1-\alpha}(\hat{\mu}).$$

Warum verschleiert die Binomialverteilung $\mathrm{B}(n,p)$ diese Struktur?

- α nicht direkt einstellbar: da $\mathrm{B}(n,p)$ diskret nimmt die Verteilung der Prüfgröße nur endlich viele Werte an. Tatsächlicher Fehler 1. Art ist

$$\alpha^* = \mathbf{P}(K \mid H_0) \leqslant \alpha.$$

  Nominelles Niveau α, tatsächliches Niveau $\alpha^* \leqslant \alpha$.
  Man kann α also nicht im Voraus wählen.
- Die Gütefunktion $G(p)$ oszilliert wie bei den CI die Überdeckungswahrscheinlichkeit.
- Schreibt man $H_1 : p > p_0$ ohne konkretes p_1 zu nennen, bleibt die Trennschärfe undefiniert und die Planung des Stichprobenumfangs unmöglich.
- Die Dualität zwischen Test und Konfidenzintervall gilt nicht mehr.

In einem Wissenschaftsgebiet werden viele Hypothesen getestet. Erfahrungsgemäß sind etwa 10% der dort aufgestellten Nullhypothesen tatsächlich falsch. Die Tests werden auf dem Niveau $\alpha = 5\%$ durchgeführt und besitzen eine Trennschärfe von $1 - \beta = 40\%$.

Sie erhalten ein signifikantes Ergebnis. Wie groß ist die Wahrscheinlichkeit, dass Sie sich irren, wenn Sie H_0 verwerfen?

1. 5%
2. Etwa 50%
3. Weniger als 5%
4. Das lässt sich aus diesen Angaben nicht berechnen.

In einem Wissenschaftsgebiet werden viele Hypothesen getestet. Erfahrungsgemäß sind etwa 10% der dort aufgestellten Nullhypothesen tatsächlich falsch. Die Tests werden auf dem Niveau $\alpha = 5\%$ durchgeführt und besitzen eine Trennschärfe von $1 - \beta = 40\%$.

Sie erhalten ein signifikantes Ergebnis. Wie groß ist die Wahrscheinlichkeit, dass Sie sich irren, wenn Sie H_0 verwerfen?

1. 5%
2. Etwa 50%
3. Weniger als 5%
4. Das lässt sich aus diesen Angaben nicht berechnen.

Lösung: Antwort 2 ist richtig: etwa 50% .

Das Signifikanzniveau ist definiert als

$$\alpha = \mathbf{P}(\text{signifikantes Ergebnis} \mid H_0).$$

Die Irrtumswahrscheinlichkeit bei gegebenem signifikantem Ergebnis ihrerseits

$$\mathbf{P}(H_0 \mid \text{signifikantes Ergebnis})$$

Bezeichne ferner

$$\begin{aligned} \pi = 10\% &: \text{Anteil tatsächlich falscher Nullhypothesen} \\ \alpha = 5\% &: \text{Signifikanzniveau} \\ 1 = \beta = 40\% &: \text{Trennschärfe} \\ N &: \text{gegebene (große) Anzahl getesteter Hypothesen} \end{aligned}$$

Lösung: Antwort 2 ist richtig: etwa 50% .

Davon sind

$$\pi N = 0.1N : \text{tatsächlich wahr}$$
$$(1-\pi)N = 0.9N : \text{tatsächlich falsch}$$

Die signifikanten Ergebnisse setzen sich zusammen aus

- richtig positiv (H_0 falsch und signifikant) $\pi N \cdot (1-\beta) = 0.04N$
- falsch positiv (H_0 wahr obwohl signifikant) $(1-\pi)N \cdot \alpha = 0.045N$
- Gesamtzahl signifikanter Ergebnisse : $0.085N$, davon $0.045N$ falsch positiv

Somit

$$\mathbf{P}(H_0|\text{signifikant}) \approx \frac{0.045N}{0.085N} \approx 0.53.$$

- John Ioannidis (2005): In vielen Forschungsfeldern ist die Mehrheit der publizierten signifikanten Funde falsch[5] – nicht wegen Betrugs, sondern als strukturelle Konsequenz der Art und Weise, wie Hypothesentests in der Praxis durchgeführt werden.
- Positiver Vorhersagewert (positive predictive value)

$$\mathrm{PPV} = \mathbf{P}(H_0 \text{ falsch} \mid \text{signifikantes Ergebnis}) = \frac{\pi(1-\beta)}{\pi(1-\beta)+(1-\pi)\alpha}$$

- In vielen Wissenschaftsgebieten: π niedrig, Trennschärfe gering, $\alpha = 5\%$ zu hoch.
- Die Praxis des *p*-Hackings[6] verschärft dies noch.
- Eine Studie in der Psychologie 10 Jahre später bestätigte dieses Ergebnis[7].

[5] J. P. A. Ioannidis. "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2.8 (2005), e124. DOI: 10.1371/journal.pmed.0020124.

[6] A. M. Stefan und F. D. Schönbrodt. "Big little lies: a compendium and simulation of p-hacking strategies". In: *Royal Society Open Science* 10 (2023), S. 220346. DOI: 10.1098/rsos.220346.

[7] O. S. Collaboration. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015), aac4716. DOI: 10.1126/science.aac4716.

Fazit

- ☞ Schätzer, Konfidenzintervalle und Hypothesentests beruhen auf wenigen grundlegenden Ideen und Konzepten, die eng miteinander zusammenhängen.
- ☞ Die Beschränkung auf diskrete Verteilungen (im sächsischen Lehrplan) bringt wenige konzeptuelle Vorteile, verdeckt aber den „roten Faden“ etwas.
- ☞ Anregung: stetige Verteilungen sind eine gute Gelegenheit, den Hauptsatz der Integralrechnung zu üben.
- ☞ Korrekte Interpretation setzt Grundverständnis voraus. Dies gilt insbesondere für alle, die ein WiMINT-Studium anstreben
- ☞ In Fragen der Wahrscheinlichkeitsrechnung und Statistik ist die eigene Intuition oft trügerisch, daher ist die Mathematik ein unverzichtbarer Leitfaden.

Fazit

- ☞ Schätzer, Konfidenzintervalle und Hypothesentests beruhen auf wenigen grundlegenden Ideen und Konzepten, die eng miteinander zusammenhängen.
- ☞ Die Beschränkung auf diskrete Verteilungen (im sächsischen Lehrplan) bringt wenige konzeptuelle Vorteile, verdeckt aber den „roten Faden“ etwas.
- ☞ Anregung: stetige Verteilungen sind eine gute Gelegenheit, den Hauptsatz der Integralrechnung zu üben.
- ☞ Korrekte Interpretation setzt Grundverständnis voraus. Dies gilt insbesondere für alle, die ein WiMINT-Studium anstreben
- ☞ In Fragen der Wahrscheinlichkeitsrechnung und Statistik ist die eigene Intuition oft trügerisch, daher ist die Mathematik ein unverzichtbarer Leitfaden.

Vielen Dank

für’s Zuhören ! ☺

Literatur

Beck-Bornholdt, H.-P. und H.-H. Dubben. *Der Hund, der Eier legt*. Rowohlt, 2011.

Brown, L. D., T. T. Cai und A. DasGupta. "Interval Estimation for a Binomial Proportion". In: *Statistical Science* 16.2 (2001), S. 101–133. URL: http://www.jstor.org/stable/2676784.

Collaboration, O. S. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015), aac4716. DOI: 10.1126/science.aac4716.

Fisher, R. A. "Statistical Methods and Scientific Induction". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 17.1 (1955), S. 69–78. URL: http://www.jstor.org/stable/2983785.

Ioannidis, J. P. A. "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2.8 (2005), e124. DOI: 10.1371/journal.pmed.0020124.

Neyman, J. "Silver Jubilee of My Dispute with Fisher". In: *Journal of the Operations Research Society of Japan* 3.4 (1961), S. 145–154.

Randow, G. von. *Das Ziegenproblem: Denken in Wahrscheinlichkeiten*. Rowohlt, 2011.

Rosenhouse, J. *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brainteaser*. Oxford University Press, 2009.

Stefan, A. M. und F. D. Schönbrodt. "Big little lies: a compendium and simulation of p-hacking strategies". In: *Royal Society Open Science* 10 (2023), S. 220346. DOI: 10.1098/rsos.220346.

Stoyan, D. "Statistische Tests in Gymnasiallehrbüchern". In: *Stochastik in der Schule* 31 (2011), S. 28–32.

Vehling, R. *Beurteilende Statistik – wie kann eine Umsetzung in der Sek II auf Grundlage einer reflektierten Theorie sinnstiftend und alltagstauglich gelingen?* Vortrag Lehrerfortbildung TU Dresden. 2025. URL: https://github.com/RVeh/Hypothesentest.

Klarstellung irreführender statistischer Behauptungen in den Medien.

Blog „Unstatistik des Monats“ (Harding-Zentrum, Berlin)

HARDING-ZENTRUM FÜR RISIKOKOMPETENZ

Podcast „More or Less“ (BBC)