Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Bootstrap-Konfidenzintervalle

Bootstrap-Methoden sind Teil einer allgemeineren Klasse von Verfahren, die als *Resampling-Verfahren* bezeichnet werden

Als weitere Resampling-Verfahren werden wir später noch Randomisierungstests (oder Permutationstests) kennenlernen

Der Ausdruck "Bootstrap" nimmt Bezug auf die Lügengeschichte des Barons von Münchhausen, der behauptet sich am eigenen Schopf aus dem Sumpf gezogen zu haben

- In der englischsprachigen Version zieht er sich an seinem Stiefelriemen (bootstrap) heraus

Nonparametrisches Bootstrapping

Die Grundlage der Parameterschätzung bildet die Ziehung einer einfachen Zufallsstichprobe vom Umfang n aus einer Population, die durch ein stochastisches Modell (also eine Verteilungsfunktion) beschrieben wird

Ist die zugrunde liegende Verteilungsfunktion unbekannt, lassen sich hieraus keine Aussagen zur Verteilung der Stichprobenkennwerte (sampling distribution) ableiten (auch nicht zu deren Standardfehler)

Die grundlegende Idee ist nun die beobachtete empirische Verteilungsfunktion der Stichprobe als Approximation der unbekannten Verteilungsfunktion in der Population anzusehen

- Im Rahmen des *nonparametrischen Bootstrap* werden wiederholt Zufallsstichproben vom selben Umfang n aus den beobachteten Daten gezogen

Bootstrap-Stichprobe

Es werden wiederholt Bootstrap-Stichproben aus den beobachteten Daten (mit Zurücklegen) gezogen, d. h. jeder Datenpunkt hat zu jedem Zeitpunkt dieselbe Wahrscheinlichkeit in die Bootstrap-Stichprobe aufgenommen zu werden

Beispiel

Daten		Bootstrap-Stichproben
x_1, x_2, x_3, x_4, x_5	B_1	x_1, x_1, x_3, x_4, x_4
	B_2	x_1, x_2, x_4, x_4, x_5
	B_3	x_2, x_3, x_4, x_5, x_5
	$\vdots$	$\vdots$
	B_b	x_1, x_3, x_3, x_4, x_5

Bootstrap-Replikation

Nach dem "Plug-in-Prinzip" werden Kennwerte der Verteilung der Population bzw. der Stichprobe durch die entsprechenden Kennwerte der empirischen Verteilung geschätzt

Beispiel

		Bootstrap		
Daten	Kennwert		Stichproben	Kennwerte
x_1, x_2, x_3, x_4, x_5	$\bar{x}$	B_1	x_1, x_1, x_3, x_4, x_4	$\bar{x}_1$
		B_2	x_1, x_2, x_4, x_4, x_5	$\bar{x}_2$
		B_3	x_2, x_3, x_4, x_5, x_5	$\bar{x}_3$
		$\vdots$	$\vdots$	
		B_b	x_1, x_3, x_3, x_4, x_5	$\bar{x}_b$

Bootstrap-Verteilung

Die Schätzfunktion für den Standardfehler eines Kennwertes in der ursprünglichen Zufallsstichprobe bildet dann die Standardabweichung der Kennwerte aus den Bootstrap-Stichproben

Allgemein wird die Verteilung eines Kennwertes in der ursprünglichen Zufallsstichprobe geschätzt durch die Verteilung der entsprechenden Kennwerte der Bootstrap-Stichproben, woraus sich dann Konfidenzintervalle ableiten lassen

Beispiele

- Die Schätzfunktion für den Standardfehler des arithmetischen Mittels $\bar{X}$ bildet die Standardabweichung der Bootstrap-Mittelwerte $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_b$
- Die Verteilung des Kennwertes $\bar{X}$ wird geschätzt durch die Verteilung der Bootstrap-Kennwerte $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_b$

Beispiel: IQ-Test

Wir betrachten die Intelligenztestdaten aus einer Untersuchung von Willerman et al. (1991), `https://dasl.datadescription.com/datafile/iq-brain/`

Für je 20 Männer und Frauen wurde darin ein IQ-Wert (Wechsler Adult Intelligence Scale-Revised) erhoben

Wir wollen ein Bootstrap-Konfidenzintervall für den Erwartungswert der Verteilung der IQ-Werte bestimmen

Einlesen der Daten, wobei `na.omit()` die Fälle mit fehlenden Werten (`NA`) ausschließt

```
dat <- read.table("brainsize-iq.txt", header = TRUE)
iq  <- na.omit(dat$FSIQ)
```

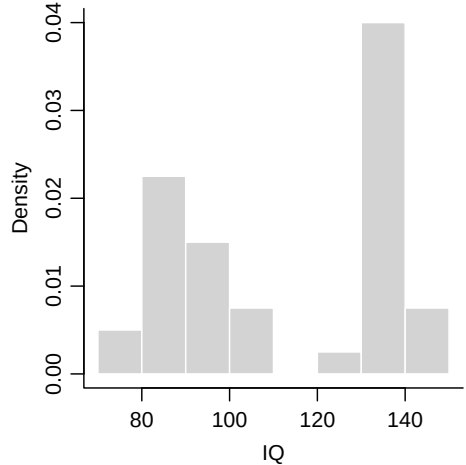
Beispiel: IQ-Test

Deskriptive Statistiken

```
mean(iq)
[1] 113.45

sd(iq)
[1] 24.08

hist(iq, freq = FALSE)
```



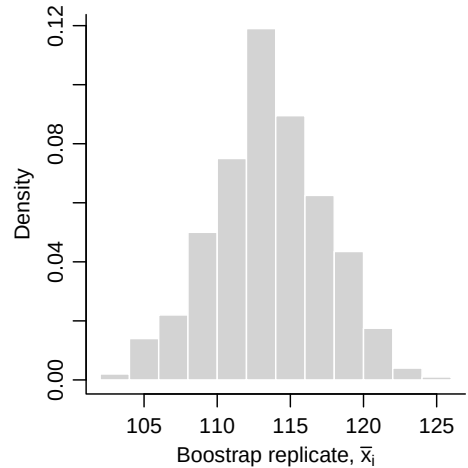
Beispiel: IQ-Test

Erzeugen der Bootstrap-Stichproben und Berechnen der jeweiligen Mittelwerte

```
## 1000 bootstrap replicates
size <- length(iq)
mean.b <- numeric(1000)
for(i in seq_along(mean.b)) {
  boot.sample <- sample(
      iq, size, replace = TRUE)
  mean.b[i] <- mean(boot.sample)
}
```

Bootstrap-Verteilung der Mittelwerte

```
hist(mean.b, freq = FALSE)
```



Perzentil-Intervall

Für ein *Bootstrap-Perzentil-Intervall* werden zu vorgegebener Irrtumswahrscheinlichkeit α die Intervallgrenzen durch die entsprechenden Quantile der Bootstrap-Verteilung bestimmt

- Grenzen für zweiseitiges Perzentil-Konfidenzintervall

```
alpha <- 0.05
quantile(mean.b, probs = c(alpha/2, 1 - alpha/2))
```

- Grenzen für einseitige Perzentil-Konfidenzintervalle

```
quantile(mean.b, probs = alpha)
quantile(mean.b, probs = 1 - alpha)
```

Bootstrap in R

R bietet verschiedene Pakete zur Durchführung der Bootstrap-Methode an (z. B. die Pakete `boot` und `bootstrap`)

Unter Verwendung des Pakets `boot` erfolgt die Berechnung der Bootstrap-Verteilung folgendermaßen

```
library(boot)

samplemean <- function(x, d) {     # Variable d erlaubt z.B.
  mean(x[d])                       # Definition verschiedener Gruppen
}

B <- boot(iq, samplemean, 1000)
hist(B$t, freq = FALSE)
```

Bootstrap in R

Hieraus lassen sich dann verschiedene Bootstrap-Konfidenzintervalle ableiten

```
boot.ci(B, conf = 0.95, type = c("perc", "bca"))
```

- Das Argument `type` definiert die Art des Konfidenzintervalls
  - `perc` Bootstrap-Perzentil-Intervall
  - `bca` Bootstrap-BC_a-Intervall (bias corrected and accelerated)
- Man kann zeigen, dass das BC_a-Konfidenzintervall präziser ist als das Perzentil-Konfidenzintervall (es kommt daher z. B. mit einer geringeren Anzahl von Bootstrap-Stichproben aus)

Statistisches Testen von Hypothesen

Einführung

Oft ist es von Interesse zu überprüfen, ob bestimmte Vermutungen über einen Parameter oder eine Verteilung in der Grundgesamtheit zutreffen

Die Überprüfung solcher Vermutungen über das Verhalten eines untersuchten Merkmals in der Population nennt man *statistischen Test*

Der Einsatz statistischer Tests zur Beantwortung von Fragestellungen hinsichtlich der unterliegenden Population erfordert

- die klare Operationalisierung der Vermutungen
- die Formulierung der Vermutungen als statistisches Testproblem (Spezifikation des daten-erzeugenden Modells)
- die Festlegung der inhaltlich minimal relevanten Abweichung von den Vermutungen (Relevanzbereich)
- eine hohe Wahrscheinlichkeit, diese Abweichung mit Hilfe von Daten nachweisen zu können (Teststärke)

Einführung

Die Durchführung statistischer Tests zur zufallskritischen Absicherung von Hypothesen ist obligatorisch für die Publikation empirischer Ergebnisse in der Psychologie

- Im Rahmen eines Begutachtungsverfahrens (peer review) werden die bei Fachzeitschriften eingereichten Manuskripte hinsichtlich der Einhaltung wissenschaftlich-methodischer Standards kritisch bewertet

Dies ist Ausdruck eines wohlverstandenen Konservativismus in der Wissenschaft; neue Ergebnisse sollen nur dann in den Wissenskanon aufgenommen werden, wenn diese durch ausreichende Evidenz gestützt werden (aber: P-Hacking, Replizierbarkeit?)

Die Zufallsbehaftetheit empirischer Beobachtungen macht eine Bewertung der Daten aus statistischer Sicht erforderlich, mit der dann die Schlussfolgerungen auf die unterliegende Grundgesamtheit abgesichert werden

Schlussfolgerungen sind aber nur in einem engen Rahmen möglich

Ansätze zum Testen von Hypothesen

Oft wird der Eindruck erweckt, die Statistik würde ein einheitliches Instrumentarium zur Durchführung statistischer Tests zur Verfügung stellen

Es existieren aber grundlegend voneinander verschiedene Ansätze

- Die Theorie des Signifikanztests nach Fisher
- Die Theorie des Hypothesentests nach Neyman-Pearson
- Die Theorie Bayesscher Hypothesentests

Dies wird in den meisten Statistikbüchern (insbesondere in solchen, die für Psychologen geschrieben wurden) nicht thematisiert

Gelehrt wird häufig ein Konglomerat, das Aspekte der Ansätze von Neyman-Pearson und Fisher vermischt, deren Unvereinbarkeit über Jahrzehnte Gegenstand heftigster Debatten zwischen den beiden Lagern war

Ansätze zum Testen von Hypothesen

Im Folgenden wird die Theorie der statistischen Hypothesentests nach Neyman-Pearson entwickelt

- Sie wird den formalen Rahmen für die zu besprechenden spezifischen statistischen Testentscheidungen bilden

Von diesem Ansatz abgrenzen werden wir

- die Theorie des Signifikanztests nach Fisher mit ihren zentralen Begriffen (z. B. dem sogenannten p-Wert als Signifikanzniveau)
- die Theorie der Bayesschen Hypothesentests, die einen gegenüber den frequentistischen Ansätzen von Neyman-Pearson und Fisher grundlegend verschiedenen Zugang wählt

Beispiel: Exakter Binomialtest

Zur Einführung der Begriffe betrachten wir zwei Beispiele, denen das gleiche Zufallsexperiment zugrunde liegt (Bernoulliexperiment)

- Wir interessieren uns für die Frage, ob eine bestimmte Münze fair ist, d. h. ob die Wahrscheinlichkeit für die Ereignisse Kopf und Zahl gleich groß ist
- Wir wollen feststellen, ob die Wahrscheinlichkeit der Geburt eines Jungen genauso groß ist, wie die eines Mädchens

Zur Beantwortung dieser Fragen führen wir das entsprechende Zufallsexperiment unabhängig voneinander wiederholt durch

- Die Münze wird n-fach wiederholt geworfen und die Anzahl der Fälle, in denen Kopf resultiert, wird registriert
- Es wird zufällig eine Geburtsklinik ausgewählt und für einen bestimmten Zeitraum (z. B. 48 Stunden) wird registriert, bei wie vielen der insgesamt n Geburten ein Junge geboren wird

Null- und Alternativhypothese

Den Zufallsexperimenten in beiden Beispielen unterliegt die Bernoulli-Verteilung, die von einem Parameter p abhängt, der die Auftretenswahrscheinlichkeit des betrachteten Ereignisses kennzeichnet

Die geäußerten Vermutungen über den Wert des Populationsparameters p werden im Rahmen des statistischen Testproblems als *statistische Hypothesen* formuliert

- In der *Nullhypothese* H_0 formulieren wir die Annahme

$$H_0: \ p = 0.5$$

- Die *Alternativhypothese* H_1 formuliert dann die gegenteilige Annahme

$$H_1: \ p \neq 0.5$$

Null- und Alternativhypothese

Die Frage, welche der Annahmen in der Null- bzw. in der Alternativhypothese zu formulieren ist, wird häufig fälschlich so beantwortet

- Die interessierende Forschungshypothese (also der Parameterwert, den man aus inhaltlichen Erwägungen erwarten würde) ist in der Alternativhypothese zu formulieren ("erwarteter Effekt")
- Diese Annahme muss sich dann gegen die Nullhypothese durchsetzen ("kein Effekt")

Richtig ist, dass die in der Nullhypothese formulierte Annahme die Grundlage für die Ableitung der konkreten Vorhersagen hinsichtlich des durchgeführten Zufallsexperiments bildet

- Die Annahme muss daher so spezifisch formuliert sein, dass diese Ableitungen überhaupt möglich sind (z. B. $p \neq 0.5$ scheidet damit von vornherein als Nullhypothese aus)

Null- und Alternativhypothese

Soll eine vorliegende Theorie statistisch getestet werden, muss diese als Annahme in der Nullhypothese formuliert werden

- Will man testen, ob die Ergebnisse eines Experiments zum Paarassoziationslernen durch das Alles-Oder-Nichts-Modell beschrieben werden können, so bildet dieses Modell als Nullhypothese die Grundlage für die empirischen Vorhersagen
- Im Rahmen eines deduktiven Vorgehens besteht dann die Möglichkeit, die Nullhypothese zu verwerfen und damit die Theorie zu falsifizieren (vgl. kritischer Rationalismus)

In der Psychologie gibt es kaum spezifische Theorien, die die Ableitung einer Nullhypothese erlauben

- Die Nullhypothese ist dann meist inhaltlich uninteressant
- Voraussetzung für gültige Schlussfolgerungen ist die Festlegung eines minimal relevanten Effekts, also der gerade noch interessierenden Abweichung von der Nullhypothese

Prüfgröße

Den empirischen Ausgangspunkt des statistischen Tests bildet eine einfache Zufallsstichprobe $X_1, \ldots, X_n$ vom Umfang n

- Im vorliegenden Fall sind die $X_1, \ldots, X_n$ binäre Zufallsvariablen (Indikatorvariablen)

Analog zur Situation bei der Parameterschätzung muss die Information aus der Stichprobe geeignet verdichtet werden

- Im Rahmen eines statistischen Tests nennt man diese zusammenfassende Größe *Teststatistik* oder *Prüfgröße*

In den betrachteten Beispielen bietet sich als Prüfgröße die relative Häufigkeit an, also

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

Stichprobenverteilung

Um die in den statistischen Hypothesen formulierten Vermutungen hinsichtlich des Parameters p zu prüfen, leitet man nun aus der Nullhypothese Voraussagen hinsichtlich der relevanten Prüfgröße ab

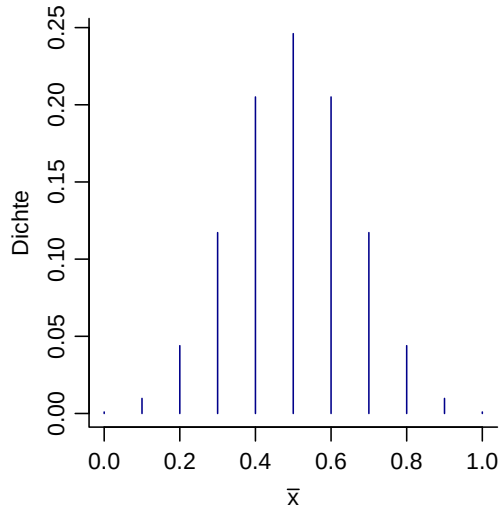
- Im vorliegenden Fall bedeutet das, dass die Summenvariable $\sum_{i=1}^{n} X_i$ binomialverteilt ist mit den Parametern n und $p = 0.5$, woraus sich unmittelbar die Verteilung $P(\bar{X} \mid p = 0.5)$ für die relative Häufigkeit als Prüfgröße ergibt

Aus Gründen der Einfachheit und Übersichtlichkeit werden wir in dieser Einführung $n = 10$ wählen

Gegeben die Annahme $p = 0.5$ der Nullhypothese ergibt sich für die Prüfgröße $\bar{X}$ folgende Verteilung ($\bar{x} = 0.0, 0.10, \ldots, 0.90, 1.0$)

Stichprobenverteilung

Verteilung der Prüfgröße $\bar{X}$ unter der Nullhypothese $p = 0.5$



Ablehnungsbereich

Für die Beibehaltung der Nullhypothese sprechen offensichtlich Werte $\bar{x}$, die nahe bei 0.5 liegen, während deutlich davon abweichende Werte gegen die Nullhypothese und für die Alternativhypothese sprechen

Wann ist die Abweichung zwischen der Prüfgröße und dem entsprechenden, theoretisch erwarteten Wert unter der Nullhypothese (hier dem Parameterwert $p = 0.5$) hinreichend groß, um das Verwerfen der Nullhypothese zu rechtfertigen?

- Der *Ablehnungsbereich* (also der Bereich der Werte der Prüfgröße, die zu einer Ablehnung der Nullhypothese führen) wird durch Vorgabe einer Irrtumswahrscheinlichkeit α definiert, die nicht überschritten werden soll
- Die Irrtumswahrscheinlichkeit α bezeichnet die Wahrscheinlichkeit, mit der die Nullhypothese verworfen wird, obwohl sie zutrifft (d. h. obwohl $p = 0.5$ gilt) und nimmt traditionell zumeist die Werte 0.1, 0.05 oder 0.01 an

Ablehnungsbereich

Für die Prüfgröße $\bar{X}$ bei $n = 10$ gilt unter der Nullhypothese

$$\left.\begin{aligned} P(\bar{X} = 0.0 \mid p = 0.5) &= 0.0009 \\ P(\bar{X} = 0.1 \mid p = 0.5) &= 0.0098 \end{aligned}\right\} \quad P(\bar{X} \in \{0.0, 0.1\} \mid p = 0.5) = 0.0107$$

$$P(\bar{X} = 0.2 \mid p = 0.5) = 0.044$$

$$\vdots$$

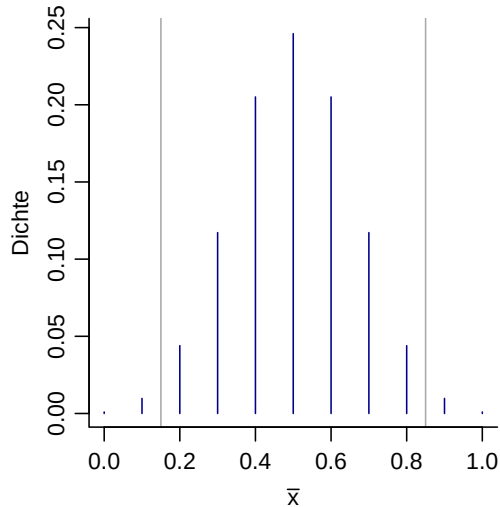
$$P(\bar{X} = 0.8 \mid p = 0.5) = 0.044$$

$$\left.\begin{aligned} P(\bar{X} = 0.9 \mid p = 0.5) &= 0.0098 \\ P(\bar{X} = 1.0 \mid p = 0.5) &= 0.0009 \end{aligned}\right\} \quad P(\bar{X} \in \{0.9, 1.0\} \mid p = 0.5) = 0.0107$$

Bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ wird die Nullhypothese also bei Vorliegen der Werte $\bar{x} \in \{0.0, 0.1, 0.9, 1.0\}$ verworfen

Ablehnungsbereich

Verteilung der Prüfgröße $\bar{X}$ unter der Nullhypothese $p = 0.5$ und Ablehnungsbereich des Tests



Fehler zweiter Art

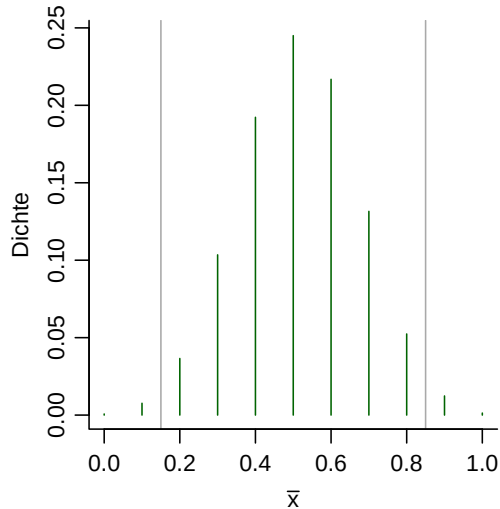
Erhält man einen der Werte $\bar{x} \in \{0.2, 0.3, \ldots, 0.7, 0.8\}$, so wird die *Nullhypothese beibehalten*

- Diese vorsichtige Sprechweise vermeidet es zu sagen, dass die Nullhypothese "angenommen" wird
- Dies hat seinen Grund darin, dass wir zwar über die Irrtumswahrscheinlichkeit α die Wahrscheinlichkeit kontrollieren, die Nullhypothese fälschlicherweise abzulehnen
- Die Wahrscheinlichkeit β des anderen möglichen Fehlers, nämlich sich für die Nullhypothese zu entscheiden, obwohl sie falsch ist, ist aber noch unbestimmt

Zur Bestimmung dieser Wahrscheinlichkeit legen wir als inhaltlich minimal relevante Abweichung einen Wert der Alternativhypothese fest und wählen in Hinblick auf das Beispiel der Jungengeburten den Parameter $p = 0.515$

Fehler zweiter Art

Verteilung der Prüfgröße $\bar{X}$ unter der Alternativhypothese
$p = 0.515$



Teststärke (Power)

Für die Wahrscheinlichkeit β einen Wert zu erhalten, der zur Beibehaltung der Nullhypothese führt, obwohl die Alternativhypothese zutrifft, ergibt sich

$$P(\bar{X} \in \{0.2, 0.3, \ldots, 0.7, 0.8\} \mid p = 0.515) = 0.978$$

Das zeigt, dass der durchgeführte Test nicht in der Lage ist, klar zwischen der Nullhypothese $p = 0.5$ und der Alternativhypothese $p = 0.515$ zu unterscheiden

- Diese Problematik ist Ausdruck einer mangelnden *Teststärke* (oder *Power*), die durch $1 - \beta$ definiert ist

Die mangelnde Teststärke ist durch den geringen Stichprobenumfang bedingt

- Für $n = 9000$ sinkt die Wahrscheinlichkeit β auf den Wert 0.19 bzw. steigt die Power auf den Wert 0.81

Einseitiger Test

Im Beispiel mit den Geburten ist es möglich, die Hypothesen nicht ungerichtet zu formulieren, sondern gezielt die Alternative zu testen, dass die Wahrscheinlichkeit für die Geburt eines Jungen > 0.5 ist

Dann formuliert man als Null- und Alternativhypothese

$$H_0: \ p \leq 0.5 \quad \text{bzw.} \quad H_1: \ p > 0.5$$

Bei der Berechnung der Verteilung für die Prüfgröße gegeben den Parameter unter der Nullhypothese setzt man den "ungünstigsten" Fall $p = 0.5$ ein, sodass sich dieselbe Verteilung wie im ungerichteten Fall ergibt

Der durch die Irrtumswahrscheinlichkeit α festgelegte Ablehnungsbereich liegt dann ausschließlich am "rechten Ende" dieser Verteilung

Statistischer Hypothesentest

Statistische Entscheidungen werden nach Neyman-Pearson als Entscheidungen zwischen zwei Hypothesen formalisiert

Jede der beiden Hypothesen ist dabei formal-logisch die Negation der jeweils anderen Hypothese

Im Rahmen der *Parametertests* (auch *parametrische Tests*) formulieren die Hypothesen einander ausschließende Annahmen über den Wert eines Populationsparameters

Man unterscheidet

- Nullhypothese H_0 (getestete Hypothese)
- Alternativhypothese H_1

Statistisches Entscheiden

Bei zwei möglichen Zuständen der Welt (H_0 trifft zu bzw. H_1 trifft zu) und zwei Entscheidungsalternativen (Entscheidung für H_0 bzw. H_1) ergibt sich folgendes Schema

Zustand der Welt	Entscheidung für H_0	Entscheidung für H_1
H_0 richtig	richtige Entscheidung	Fehler 1. Art (α-Fehler)
H_1 richtig	Fehler 2. Art (β-Fehler)	richtige Entscheidung

Fehler erster und zweiter Art

Der α-Fehler beschreibt das Ereignis, dass die Nullhypothese H_0 *fälschlicherweise abgelehnt* wird

- Entscheidung für die Alternativhypothese, obwohl die Nullhypothese zutrifft

Der β-Fehler beschreibt das Ereignis, dass die Nullhypothese H_0 *fälschlicherweise beibehalten* wird

- Entscheidung für die Nullhypothese, obwohl die Alternativhypothese zutrifft

Eine optimale statistische Entscheidung sollte nun versuchen, die Wahrscheinlichkeiten des α-Fehlers und des β-Fehlers zu minimieren

Üblicherweise wird die Entscheidung so getroffen, dass die Wahrscheinlichkeiten des α- und β-Fehlers bestimmte Werte nicht übersteigen, z. B. $\alpha = 0.05$, $\beta = 0.20$

Fehler erster und zweiter Art

Die Ablehnung der Nullhypothese ist über die Irrtumswahrscheinlichkeit α abgesichert, die auch als *Signifikanzniveau* bezeichnet wird

- Man spricht daher von einem *statistischen Test zum (Signifikanz-)Niveau* α

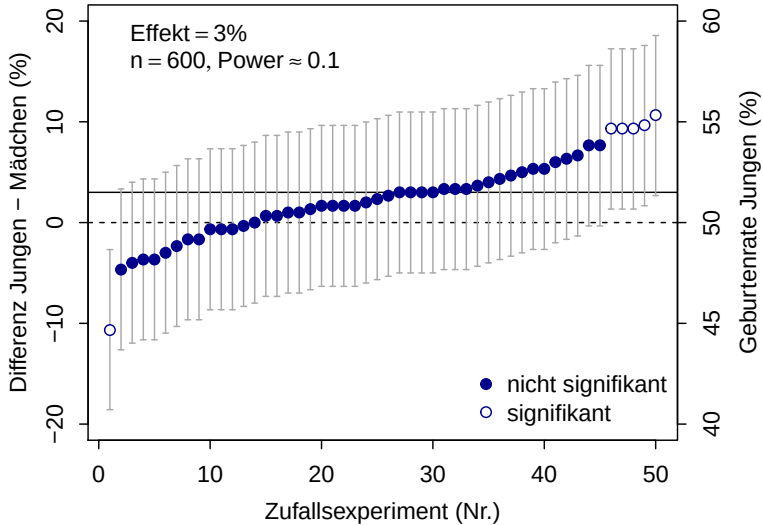
Bei Ablehnung der Nullhypothese sagt man

"Das Ergebnis ist statistisch signifikant (zum Niveau α*)"*

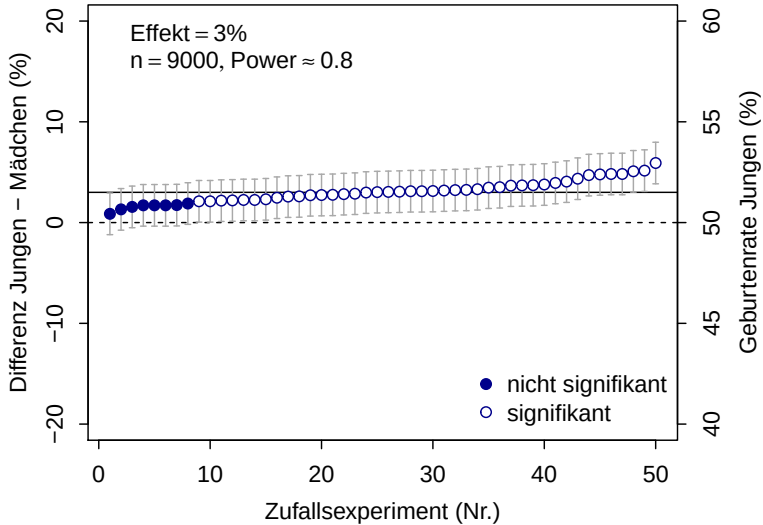
Der β-Fehler wird kontrolliert, indem eine bestimmte Power (z. B. 0.8 oder 0.95) nicht unterschritten wird

- Die Power steigt mit zunehmendem Stichprobenumfang, zunehmender Effektgröße und abnehmender Varianz der Beobachtungen
- Bei niedriger Power sind die Ergebnisse nicht signifikanter Tests wenig aussagekräftig, die Ergebnisse signifikanter Tests dagegen stark verzerrt

Beispiel Geburtenrate: Niedrige Power



Beispiel Geburtenrate: Hohe Power



Hypothesentests und statistisches Schließen

Mindestvoraussetzung für gültige Schlussfolgerungen aus Hypothesentests ist also eine hohe Teststärke (Power)

Dafür muss bei der Studienplanung die folgende Reihenfolge eingehalten werden

1. Festlegen des inhaltlich minimal relevanten Effekts bzw. der minimal relevanten Abweichung von der Nullhypothese
2. Per Powersimulation Bestimmung des Stichprobenumfangs n, mit dem der festgelegte Effekt gezeigt werden kann (s. Beispielsammlung)
3. Erhebung der geplanten Stichprobe und Testdurchführung

Ausserhalb dieses engen Rahmens haben inferenzstatistische Maße (wie p-Werte, Konfidenzintervalle) lediglich deskriptiven Charakter und erlauben keine Schlussfolgerungen bzgl. einer zugrunde liegenden Population