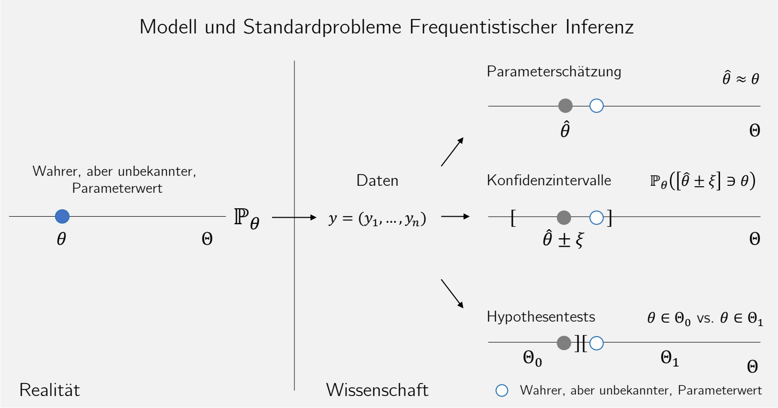
Tutorium

Wahrscheinlichketstheorie und Frequentistische Inferenz

BSc Psychologie WiSe 2022/23

Belinda Fleischmann

(12) Hypothesentests


Modell und Standardprobleme Frequentistischer Inferenz
Wahrer, aber unbekannter, Parameterwert
θ
Θ
$\mathbb{P}_\theta$
Daten
$y = (y_1, \dots, y_n)$
Parameterschätzung
$\hat{\theta} \approx \theta$
$\hat{\theta}$
Θ
Konfidenzintervalle
$\mathbb{P}_\theta([\hat{\theta} \pm \xi] \ni \theta)$
$\hat{\theta} \pm \xi$
Θ
Hypothesentests
$\theta \in \Theta_0$ vs. $\theta \in \Theta_1$
Θ_0
Θ_1
Θ
Realität
Wissenschaft
Wahrer, aber unbekannter, Parameterwert

Selbstkontrollfragen

1. Erläutern Sie die grundlegende Logik statistischer Hypothesentests.
2. Geben Sie die Definition statistischer Hypothesen und eines Testszenarios wieder.
3. Definieren Sie die Begriffe der einfachen und zusammengesetzten Hypothesen.
4. Definieren Sie die Begriffe der einseitigen und zweiseitigen Hypothesen.
5. Definieren Sie den Begriff des Tests.
6. Definieren Sie den Begriff des Standardtests.
7. Definieren Sie den Begriff des kritischen Bereichs eines Tests.
8. Definieren Sie den Begriff des Ablehungsbereichs eines Tests.
9. Definieren Sie den Begriff des kritischen Wert-basierten Tests.
10. Definieren Sie richtige Testentscheidungen, Typ I Fehler und Typ II Fehler.
11. Definieren Sie die Testgütefunktion.
12. Erläutern Sie die Bedeutung der Testgütefunktion im Rahmen der Konstruktion statistischer Tests.
13. Definieren Sie die Begriffe des Signifikanniveaus und des Level-α_0-Tests.
14. Definieren Sie den Begriff des Testumfangs.
15. Erläutern Sie die prinzipielle Strategie zur Wahl von Null- und Alternativhypothesen in der Wissenschaft.
16. Erläutern Sie zentrale Schritte zur Konstruktion eines Hypothesentests.

Selbstkontrollfragen

17. Formulieren Sie das statistische Modell eines Einstichproben-T-Tests.
18. Formulieren Sie die einfache Nullhypothese und zusammengesetzte Alternativhypothese dieses Tests.
19. Definieren Sie den zweiseitigen Einstichproben-T-Test (ZETT).
20. Skizzieren Sie qualitativ die Testgütefunktionen eines ZETTs für verschiedene kritische Werte.
21. Wie muss der kritische Wert eines ZETTs definiert sein, damit der Test ein Level-α_0-Test ist?
22. Skizzieren Sie qualitativ die Bestimmung des kritischen Wertes k_{α_0} bei einem zws Einstichproben-T-Test.
23. Erläutern Sie das praktische Vorgehen zur Durchführung eines ZETTs.
24. Von welchen Werten hängt die Powerfunktion eines ZETTs ab?
25. Skizzieren Sie qualitativ die Powerfunktion des ZETTs bei fester Stichprobengröße.
26. Skizzieren Sie qualitativ die Powerfunktion des ZETTs bei festem Erwartungswertparameter.
27. Erläutern Sie das favorisierte praktische Vorgehen zur Durchführung einer Poweranalyse.
28. Erläutern Sie die Motivation zur Auswertung von p-Werten.
29. Definieren Sie den Begriff des p-Werts.
30. Geben Sie das Theorem zur Dualität von Konfidenzintervallen und Hypothesentests wieder.
31. Erläutern Sie die Dualität von Konfidenzintervallen und Hypothesentests.

1. Erläutern Sie die grundlegende Logik statistischer Hypothesentests.

Man hat einen **Datensatz** $y_1, ..., y_n$ vorliegen und nimmt an, dass es sich dabei um die **Realisation einer Stichprobe** handelt, zum Beispiel von $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$.

Man berechnet basierend auf dem Datensatz eine **Teststatistik**, zum Beispiel das anhand der Stichprobenvarianz und der Stichprobengröße normalisierte Stichprobenmittel $\sqrt{n}\bar{y}_n / s_n$.

Man fragt sich, **wie wahrscheinlich** es wäre, den beobachteten oder einen extremeren Wert der Teststatistik **unter der Annahme eines *Nullmodels*** zu observieren. Dabei meint man mit *Nullmodell* intuitiv ein Wahrscheinlichkeitsverteilungsmodell bei dem kein "interessanter Effekt" vorliegt, also zum Beispiel $\mu = 0$ gilt. Die Wahrscheinlichkeit ist wie immer Frequentistisch zu verstehen, d.h. als idealisierte relative Häufigkeit, wenn man viele Stichprobenrealisationen des Nullmodels generieren würde.

Ist die betrachtete **Wahrscheinlichkeit dafür, den beobachteten oder einen extremeren Wert der Teststatistik unter Annahme des Nullmodells zu observieren groß**, so sagt man sich "Nunja, dann ist es wohl ganz plausibel, dass das Nullmodel die Daten generiert hat". Im Wissenschaftsjargon spricht man von einem **"nicht-signifikanten Ergebnis"**.

Ist die betrachtete **Wahrscheinlichkeit dafür, den beobachteten oder einen extremeren Wert der Teststatistik unter Annahme des Nullmodells zu observieren dagegen klein**, so sagt man sich "Aha, dann ist es wohl nicht so plausibel, dass das Nullmodel die Daten generiert hat". Im Wissenschaftsjargon spricht man von einem **"signifikanten Ergebnis"**.

Wie immer in der Frequentistischen Statistik weiß man nach Durchführung dieser Prozedur nicht, ob im vorliegenden Fall nun wirklich das Nullmodel oder ein anderes Modell die Daten generiert hat, sondern man weiß nur, wie oft man bei dieser Prozedur im Mittel richtig oder falsch liegen würde, wenn alle Annahmen zuträfen und man diese Prozedur sehr oft wiederholen würde.

2. Geben Sie die Definition statistischer Hypothesen und eines Testszenarios wieder.

Definition (Statistische Hypothesen und Testszenario)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei eine Stichprobe mit WMF oder WDF p_θ, $\mathcal{Y}$ sei der Ergebnisraum des Zufallsvektors $\upsilon := (\upsilon_1, ..., \upsilon_n)$, und Θ sei der Parameterraum des zugrundeliegenden statistischen Modells. Weiterhin sei $\{\Theta_0, \Theta_1\}$ eine Partition des Parameterraumes, so dass $\Theta = \Theta_0 \cup \Theta_1$ und $\Theta_0 \cap \Theta_1 = \emptyset$ gelten. Eine *statistische Hypothese* ist dann eine Aussage über den wahren, aber unbekannten, Parameterwert θ in Hinblick auf die Untermengen Θ_0 und Θ_1 des Parameterraums. Speziell werden die Aussagen

- $\theta \in \Theta_0$ als *Nullhypothese* H_0
- $\theta \in \Theta_1$ als *Alternativhypothese* H_1

bezeichnet. Die Einheit aus Stichprobe, Ergebnisraum, Parameterraum, und Hypothesen wird im Folgenden als *Testszenario* bezeichnet.

3. Definieren Sie die Begriffe der einfachen und zusammengesetzten Hypothesen.

Definition (Einfache und zusammengesetzte Hypothesen)

Für statistische Hypothesen $\Theta_i, i = 0, 1$ gilt:

- Enthält Θ_i nur ein einziges Element, so heißt Θ_i *einfach*.
- Enthält Θ_i mehr als ein Element, so heißt Θ_i *zusammengesetzt*.

Bemerkungen

- Die Nullhypothese $\Theta_0 = \{0\}$ ist ein Beispiel für eine einfache Hypothese.
- Bei einer einfachen Hypothese ist die Wahrscheinlichkeitsverteilung von υ genau festgelegt.
- Bei einer zusammengesetzten Hypothese ist nur die Verteilungsklasse von υ festgelegt.

4. Definieren Sie die Begriffe der einseitigen und zweiseitigen Hypothesen.

Definition (Einseitige und zweiseitige Hypothesen)

$\Theta := \mathbb{R}$ sei ein eindimensionaler Parameterraum und θ_0 sei ein Element von Θ. Dann werden zusammengesetzte Nullhypothesen der Form

$$\Theta_0 :=] - \infty, \theta_0] \text{ oder } \Theta_0 := [\theta_0, \infty[\tag{1}$$

einseitige Nullhypothesen genannt und auch in der Form

$$H_0 : \theta \leq \theta_0 \text{ oder } H_0 : \theta \geq \theta_0 \tag{2}$$

geschrieben. Die entsprechenden Alternativhypothesen haben dabei die Form

$$\Theta_1 :=]\theta_0, \infty[\text{ oder } \Theta_1 :=] - \infty, \theta_0[\text{ bzw. } H_1 : \theta > \theta_0 \text{ oder } H_1 : \theta < \theta_0. \tag{3}$$

Bei einer einfachen Nullhypothese der Form

$$\Theta_0 := \{\theta_0\} \text{ bzw. } H_0 : \theta = \theta_0 \tag{4}$$

wird die Alternativhypothese

$$\Theta_1 := \Theta \setminus \{\theta_0\} \text{ bzw. } H_1 : \theta \neq \theta_0 \Leftrightarrow \Theta_1 :=] - \infty, \theta_0[\cup]\theta_0, \infty[\tag{5}$$

zweiseitige Alternativhypothese genannt.

5. Definieren Sie den Begriff des Tests.

Definition (Test)

In einem Testszenario ist ein *Test* ϕ eine Abbildung aus dem Ergebnisraum $\mathcal{Y}$ nach $\{0, 1\}$,

$$\phi : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi(y). \tag{6}$$

Dabei repräsentiert

- $\phi(y) = 0$ den Vorgang des Nichtablehnens der Nullhypothese.
- $\phi(y) = 1$ den Vorgang des Ablehnens der Nullhypothese.

Bemerkung

- Weil y eine Realisation von υ ist, ist $\phi(y)$ eine Realisation von $\phi(\upsilon)$.

6. Definieren Sie den Begriff des Standardtests.

Definition (Standardtest)

Ein *Standardtest* ist definiert durch die Verkettung einer *Teststatistik*

$$\gamma : \mathcal{Y} \to \mathbb{R} \tag{7}$$

und einer *Entscheidungsregel*

$$\delta : \mathbb{R} \to \{0, 1\}. \tag{8}$$

Ein Standardtest kann also geschrieben werden als

$$\phi := \delta \circ \gamma : \mathcal{Y} \to \{0, 1\}. \tag{9}$$

Bemerkungen

- Weil y eine Realisation von υ ist, ist $\gamma(y) \in \mathbb{R}$ eine Realisation von $\gamma(\upsilon)$.
- Weil $\gamma(y)$ eine Realisation von $\gamma(\upsilon)$ ist, ist $(\delta \circ \gamma)(y)$ eine Realisation von $(\delta \circ \gamma)(\upsilon)$.
- Wir betrachten in der Folge nur Standardtests.

7. Definieren Sie den Begriff des kritischen Bereichs eines Tests.

Definition (Kritischer Bereich)

Die Untermenge K des Ergebnisraums des Zufallsvektors $\upsilon := (\upsilon_1, ..., \upsilon_n)$, für die ein Test den Wert 1 annimmt, heißt *kritischer Bereich* des Tests,

$$K := \{y \in \mathcal{Y} | \phi(y) = 1\} \subset \mathcal{Y}. \tag{10}$$

Bemerkungen

- Die Ereignisse $\{\phi(\upsilon) = 1\}$ und $\{\upsilon \in K\}$ sind äquivalent.
- Die Ereignisse $\{\phi(\upsilon) = 1\}$ und $\{\upsilon \in K\}$ haben die gleiche Wahrscheinlichkeit.

8. Definieren Sie den Begriff des Ablehungsbereichs eines Tests.

Definition (Ablehnungsbereich)

Die Untermenge A des Ergebnisraums einer Teststatistik, für die der Test den Wert 1 annimmt, heißt *Ablehnungsbereich* des Tests,

$$A := \{\gamma(y) \in \mathbb{R} | \phi(y) = 1\} \subset \mathbb{R}. \tag{11}$$

Bemerkungen

- Die Ereignisse $\{\phi(\upsilon) = 1\}$ und $\{\gamma(\upsilon) \in A\}$ sind äquivalent.
- Die Ereignisse $\{\phi(\upsilon) = 1\}$ und $\{\gamma(\upsilon) \in A\}$ haben die gleiche Wahrscheinlichkeit.

9. Definieren Sie den Begriff des kritischen Wert-basierten Tests.

Definition (Kritischer Wert-basierte Tests)

Ein *kritischer Wert-basierter Test* ist ein Standardtest, bei dem die Entscheidungsregel δ von einem kritischen Wert $k \in \mathbb{R}$ abhängt. Speziell ist

- ein *einseitiger kritischer Wert-basierter Test* von der Form

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := 1_{\{\gamma(y) \geq k\}} = \begin{cases} 1 & \gamma(y) \geq k \\ 0 & \gamma(y) < k \end{cases} \tag{12}$$

- ein *zweiseitiger kritischer Wert-basierter Test* von der Form

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := 1_{\{|\gamma(y)| \geq k\}} = \begin{cases} 1 & |\gamma(y)| \geq k \\ 0 & |\gamma(y)| < k \end{cases} \tag{13}$$

10. Definieren Sie richtige Testentscheidungen, Typ I Fehler und Typ II Fehler.

Definition (Richtige Testentscheidungen und Testfehler)

Das Nichtablehnen der Nullhypothese, wenn die Nullhypothese zutrifft werden, sowie das Ablehnen der Nullhypothese, wenn die Nullhypothese nicht zutrifft, werden *richtige Testentscheidungen* genannt. Es können weiterhin zwei Arten von Testfehlern auftreten: das Ablehnen der Nullhypothese, wenn die Nullhypothese zutrifft, heißt *Typ I Fehler*, das Nichtablehnen der Nullhypothese, wenn die Alternativhypothese zutrifft, heißt *Typ II Fehler*.

Testentscheidung

Wahrer Parameterwert	$\phi(y) = 0$	$\phi(y) = 1$
$\theta \in \Theta_0$	Richtige Entscheidung	Typ I Fehler
$\theta \in \Theta_1$	Typ II Fehler	Richtige Entscheidung

11. Definieren Sie die Testgütefunktion.

Definition (Testgütefunktion)

Für einen Test ϕ ist die *Testgütefunktion* definiert als

$$q_\phi : \Theta \to [0,1], \theta \mapsto q_\phi(\theta) := \mathbb{P}_\theta(\phi = 1). \tag{14}$$

Für $\theta \in \Theta_1$ heißt q_ϕ auch *Powerfunktion* oder *Trennschärfefunktion*.

Bemerkungen

- Wir verzichten hier und im Folgenden auf die explizite Notation der Abhängigkeit von ϕ von υ.
- $\mathbb{P}_\theta$ bezeichnet die Verteilung von ϕ unter der Annahme $\upsilon_1, ..., \upsilon_n \sim p_\theta$.
- Es gilt $\mathbb{P}_\theta(\phi = 1) = \mathbb{P}_\theta(\upsilon \in K) = \mathbb{P}_\theta(\gamma \in A)$
- Bei Poweranalysen betrachtet man q_ϕ als Funktion aller Testszenarien und Testparameter.
- Ändert sich ϕ, z.B. weil sich der kritische Wert von ϕ ändert, dann ändert sich $q_\phi(\theta)$.

12. Erläutern Sie die Bedeutung der Testgütefunktion im Rahmen der Konstruktion statistischer Tests.

Wir wollen statistische Tests so konstruieren, dass die **Wahrscheinlichkeit für Testfehler möglichst gering** ist.

- Mithilfe der Testgütefunktion können wir die **Wahrscheinlichkeiten für** $\phi = 1$ (Nullhypothese H_0 wird abgelehnt), und somit auch die Wahrscheinlichkeit für einen **Typ I Fehler** betrachten.
- Für jedes $\theta \in \Theta$ (wobei $\Theta = \Theta_0 \cup \Theta_1$) liefert q_ϕ die Wahrscheinlichkeit, dass H_0 durch ϕ abgelehnt wird.
- Im Idealfall hätte man einen Test ϕ mit

  $$q_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = 1 \text{ für } \theta \in \Theta_1.$$

  Die Testentscheidung eines solchen ϕ wäre mit Wahrscheinlichkeit 1 richtig.
- $\Rightarrow$ Gut sind kleine Werte von q_ϕ für $\theta \in \Theta_0$ und große Werte von q_ϕ für $\theta \in \Theta_1$.

Generell gibt es Abhängigkeiten zwischen den Werten von q_ϕ für $\theta \in \Theta_0$ und $\theta \in \Theta_1$:

- Sei zum Beispiel ϕ_a der Test definiert durch $\phi_a(y) := 0$ für alle $y \in \mathcal{Y}$, also der Test, der die Nullhypothese, unabhängig von den beobachteten Daten, *niemals ablehnt*. Für diesen Test gilt $q_{\phi_a}(\theta) = 0$ für $\theta \in \Theta_0$. Allerdings gilt für diesen Test auch $q_{\phi_a}(\theta) = 0$ für $\theta \in \Theta_1$.
- Andersherum sei ϕ_b der Test definiert durch $\phi_b(y) := 1$ für alle $y \in \mathcal{Y}$, also ein Test, der die Nullhypothese, unabhängig von den beobachteten Daten, *immer ablehnt*. Für diesen Test gilt $q_{\phi_b}(\theta) = 1$ für $\theta \in \Theta_1$. Allerdings gilt für diesen Test auch $q_{\phi_b}(\theta) = 1$ für $\theta \in \Theta_0$.

SKF 12. *Bedeutung der Testgütefunktion* (fortgeführt)

In der Konstruktion eines Tests muss also eine **angemessene Balance** zwischen

- kleinen Werten von q_ϕ für $\theta \in \Theta_0$ und
- großen Werten von q_ϕ für $\theta \in \Theta_1$ gefunden werden.

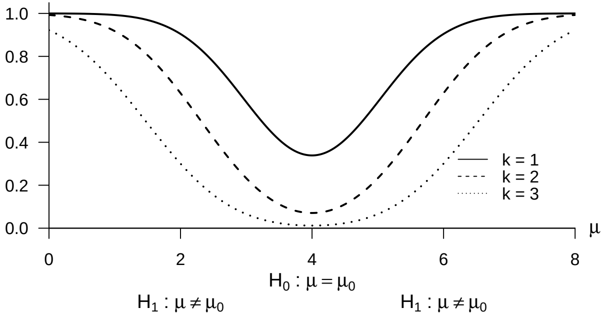
Die populärste Methode, eine solche Balance zu finden, ist in einem ersten Schritt ein $\alpha_0 \in [0, 1]$ zu wählen und sicher zu stellen, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0. \tag{15}$$

- Eine konventionelle Wahl für sein solches α_0 ist zum Beispiel $\alpha_0 := 0.05$.
- Unter allen Tests und statistischen Modellen, die Ungleichung (15) erfüllen, wird man dann einen Test oder ein statistisches Modell auswählen, so dass $q_\phi(\theta)$ für $\theta \in \Theta_1$ so groß wie möglich ist.

Testgütefunktion q_ϕ für $\sigma^2 = 9, \mu_0 = 4, n = 12$ und $k = 1, 2, 3$.

$$q_\phi(\mu) = \mathbb{P}_\mu(\phi = 1)$$



13. Definieren Sie die Begriffe des Signifikanniveaus und des Level-α_0-Tests.

Definition (Level-α_0-Test, Signifikanzlevel α_0)

q_ϕ sei die Testgütefunktion eines Tests ϕ und es sei $\alpha_0 \in [0,1]$. Dann heißt ein Test ϕ, für den gilt, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0 \tag{16}$$

ein *Level-α_0-Test* und man sagt, dass der Test das *Signifikanzlevel* α_0 hat.

14. Definieren Sie den Begriff des Testumfangs.

Definition (Testumfang α)

Es sei ϕ ein *Level-α_0-Test* mit *Signifikanzlevel* α_0. Die Zahl

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) \in [0, 1] \tag{17}$$

heißt der *Testumfang* von ϕ.

Bemerkungen

- α ist die größtmögliche Wahrscheinlichkeit für einen Typ I Fehler.
- Ein Test ist dann, und nur dann, ein Level-α_0-Test, wenn $\alpha \leq \alpha_0$ gilt.
- Bei einer einfachen Nullhypothese gilt für den Testumfang, dass $\alpha = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi = 1)$.

15. Erläutern Sie die prinzipielle Strategie zur Wahl von Null- und Alternativhypothesen in der Wissenschaft.

Das Vorgehen in der Testkonstruktion zunächst durch die Wahl eines Signifikanzlevels den Testumfang zu begrenzen und erst in einem zweiten Schritt dafür zu sorgen, dass die Wahrscheinlichkeit von $\phi = 1$ bei $\theta \in \Theta_1$ bei diesem Signifikanzlevel möglichst groß ist, induziert eine **Asymmetrie in der Behandlung von Null- und Alternativhypothese**. Implizit wichtet man mit diesem Vorgehen Typ I Fehler als schwerwiegender als Typ II Fehler.

Dies wiederum impliziert eine mögliche Strategie zur Festlegung von Null- und Alternativhypothese:

- Die **Nullhypothese** ist die Hypothese, hinsichtlich deren assoziierter Testentscheidung man eher keinen Fehler machen möchte bzw. deren Fehlerwahrscheinlichkeit man primär kontrollieren möchte.
- **In der wissenschaftlichen Anwendung ist es Standard, die falsche Konfirmation der eigenen Theorie als einen schwerwiegenderen Fehler als die falsche Ablehnung der eigenen Theorie zu werten.**
- $\Rightarrow$ Die **falsche Konfirmation der eigenen Theorie** sollte also ein **Typ I Fehler**, das falsche Ablehnen der eigenen Theorie ein Typ II Fehler sein.
- Damit die falsche Konfirmation der eigenen Theorie einen Typ I Fehler, also das Ablehnen von H_0 bei Zutreffen von H_0, darstellt, muss **die eigene Theorie** als **Alternativhypothese** aufgestellt werden.
- Die **Alternativhypothese fälschlichweise Abzulehnen** wird damit ein **Typ II Fehler**.

16. Erläutern Sie zentrale Schritte zur Konstruktion eines Hypothesentests.

(1) Statistisches Modell und Testhypothesen
(2) Definition und Analyse der Teststatistik
(3) Definition des Tests
(4) Analyse der Testgütefunktion
(5) Testumfangkontrolle
(6) Analyse der Powerfunktion

17. Formulieren Sie das statistische Modell eines Einstichproben-T-Tests.

Das **Statistische Modell des Einstichproben-T-Tests** ist definiert als

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, ..., n \tag{18}$$

wobei

- $\upsilon_i, i = 1, ..., n$ beobachtbare Zufallsvariablen,
- μ den wahren, aber unbekannten, Erwartungswertparameter der Stichprobenvariablen,
- $\varepsilon_i, i = 1, ..., n$ unabhängige normalverteilte nicht-beobachtbare Zufallsvariablen und
- $\sigma^2 > 0$ den Varianzparameter der ε_i

bezeichnen.

Dieses Modell ist äquivalent zum Normalverteilungsmodell

$$\upsilon = \upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2), \tag{19}$$

also der **Annahme unabhängig und identisch normalverteilter Stichprobenvariablen** mit Erwartungswertparameter μ und Varianzparameter σ^2.

18. Formulieren Sie die einfache Nullhypothese und zusammengesetzte Alternativhypothese dieses Tests.

Für ein μ_0 betrachten wir die einfache **Nullhypothese** und die zusammengesetzte **Alternativhypothese**

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R} \setminus \{\mu_0\}, \tag{20}$$

respektive.

Bezogen auf das Anwendungsbeispiel ist hier $\mu_0 := 0$ von Interesse:

- $H_0 : \mu = 0$ entspricht der Hypothese keines Effekts der Therapie auf die BDI Score Reduktion.
- $H_1 : \mu \neq 0$ entspricht der Hypothese eines Effekts der Therapie auf die BDI Score Reduktion.

19. Definieren Sie den zweiseitigen Einstichproben-T-Test (ZETT).

$$\phi(\upsilon) := 1_{\{|T| \geq k\}} = \begin{cases} 1 & |T| \geq k \\ 0 & |T| < k \end{cases}.$$

Definition (Einstichproben-T-Teststatistik)

$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ sei die Stichprobe eines Normalverteilungmodells, $\bar{\upsilon}$ bezeichne das Stichprobenmittel, S bezeichne die Stichprobenstandardabweichung und es sei $\mu_0 \in \mathbb{R}$. Dann ist die *T-Teststatistik* definiert als

$$T := \sqrt{n}\left(\frac{\bar{\upsilon} - \mu_0}{S}\right). \tag{21}$$

Bemerkungen

- Im Gegensatz zur T-Konfidenintervallstatistik muss bei der T-Teststatistik nicht $\mu_0 = \mu$ gelten.
- Intuitiv kann die T-Teststatistik als mit der Stichprobengröße (Evidenz) gewichtetes Verhältnis von Signal (sytematischer Variabilität) zu Rauschen (unsystematischer Variabilität) verstanden werden:

$$\sqrt{\text{Stichprobengröße}}\left(\frac{\text{Signal}}{\text{Rauschen}}\right) = \sqrt{n}\left(\frac{\bar{\upsilon} - \mu_0}{S}\right) \tag{22}$$

- Die T-Teststastitik ist eine skalare Deskription des Effekt vs. Variabilität Verhältnisses eines Datensatzes.
- In der T-Teststatistik wird die Effektgröße in Einheiten der Stichprobenstandardabweichung gemessen:
  - $T = 1 \Leftrightarrow \sqrt{n}(\bar{\upsilon} - \mu_0) = 1S$
  - $T = 2 \Leftrightarrow \sqrt{n}(\bar{\upsilon} - \mu_0) = 2S$

Theorem (Verteilung der T-Teststatistik)

$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ sei die Stichprobe eines Normalverteilungmodells, $\bar{\upsilon}$ sei das Stichprobenmittel, S sei die Stichprobenstandardabweichung, und es sei $\mu_0 \in \mathbb{R}$. Dann ist die T-Teststatistik

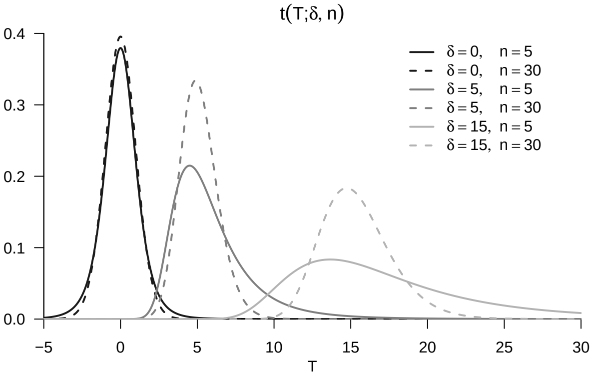
$$T := \sqrt{n}\left(\frac{\bar{\upsilon} - \mu_0}{S}\right) \tag{23}$$

eine nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter

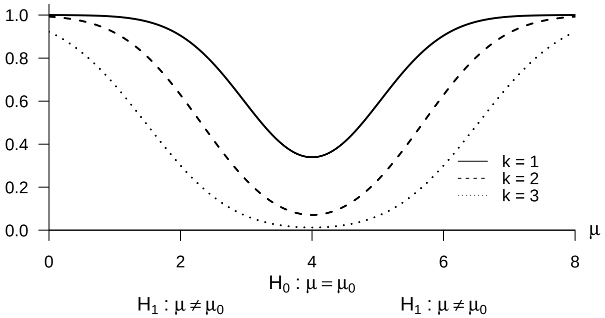
$$d = \sqrt{n}\left(\frac{\mu - \mu_0}{\sigma}\right) \tag{24}$$

und Freiheitsgradparameter $n - 1$, es gilt also $T \sim t(d, n - 1)$

Wahrscheinlichkeitsdichtefunktionen nichtzentraler t-Verteilungen



20. Skizzieren Sie qualitativ die Testgütefunktionen eines ZETTs für verschiedene kritische Werte.



21. Wie muss der kritische Wert eines ZETTs definiert sein, damit der Test ein Level-α_0-Test ist?

Theorem (Testumfangkontrolle)

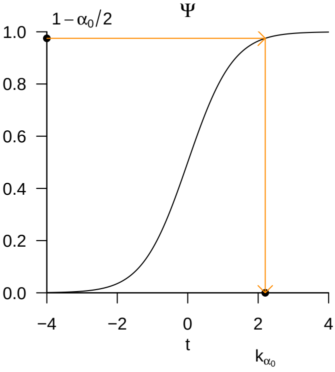
ϕ sei der oben definierte Test. Dann ist ϕ ein Level-α_0-Test mit Testumfang α_0, wenn der kritische Wert definiert ist durch

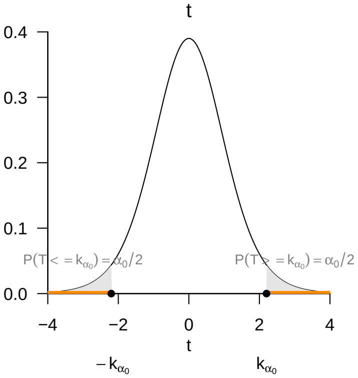
$$k_{\alpha_0} := \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right), \tag{25}$$

wobei $\Psi^{-1}(\cdot; n-1)$ die inverse KVF der t-Verteilung mit $n-1$ Freiheitsgraden ist.

22. Skizzieren Sie qualitativ die Bestimmung des kritischen Wertes k_{α_0} bei einem zws Einstichproben-T-Test.

Wahl von $k_{\alpha_0} := \Psi^{-1}(1 - \frac{\alpha_0}{2}; n-1)$ mit $n = 12$, $\alpha_0 := 0.05$ und Ablehnungsbereich





23. Erläutern Sie das praktische Vorgehen zur Durchführung eines ZETTs.

- Man nimmt an, dass ein vorliegender Datensatz $y_1, ..., y_n$ eine Realisation von $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ mit unbekannten Parametern μ und $\sigma^2 > 0$ ist.
- Man möchte entscheiden ob für ein $\mu_0 \in \mathbb{R}$ eher $H_0 : \mu = \mu_0$ oder $H_1 : \mu \neq \mu_0$ zutrifft.
- Man wählt ein Signifikanzlevel α_0 und bestimmt den zugehörigen Freiheitsgradparameter-abhängigen kritischen Wert k_{α_0}. Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05$ und $n = 12$, also Freiheitsgradparameter 11, dass $k_{0.05} = \Psi^{-1}(1 - 0.05/2; 11) \approx 2.20$ ist.
- Anhand von $n, \mu_0, \bar{\upsilon}$ und s_n berechnet man die Realisierung der T-Teststatistik

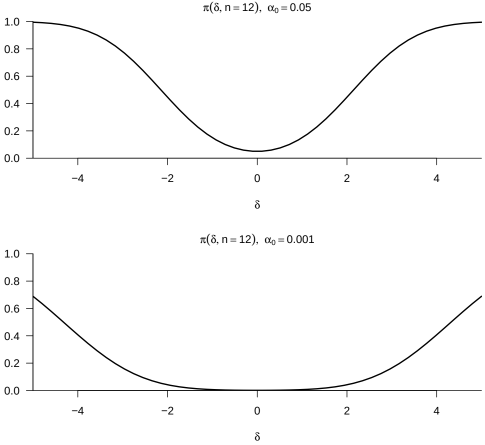
$$t := \sqrt{n}\left(\frac{\bar{y} - \mu_0}{s}\right) \tag{26}$$

- Wenn t größer-gleich k_{α_0} ist oder wenn t kleiner- gleich $-k_{\alpha_0}$ ist, lehnt man die Nullhypothese ab, andernfalls lehnt man sie nicht ab.
- Die oben entwickelte Theorie garantiert dann, dass man in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.

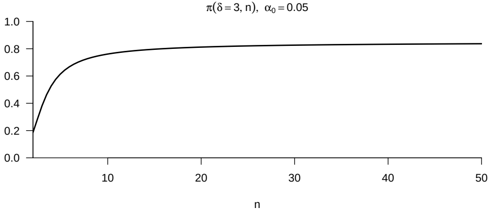
24. Von welchen Werten hängt die Powerfunktion eines ZETTs ab?

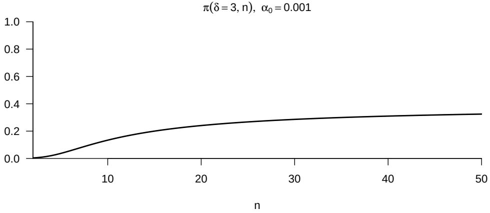
Die Powerfunktion eines zweiseitigen Einstichproben T-Tests hängt bei festgelegtem α_0 vom wahren, aber unbekannten Parameterwert $\delta = \sqrt{n}\frac{\mu - \mu_0}{\sigma}$ und von der Stichprobengröße n ab.

25. Skizzieren Sie qualitativ die Powerfunktion des ZETTs bei fester Stichprobengröße.



26. Skizzieren Sie qualitativ die Powerfunktion des ZETTs bei festem Erwartungswertparameter.





27. Erläutern Sie das favorisierte praktische Vorgehen zur Durchführung einer Poweranalyse.

Generell wird folgendes Vorgehen favorisiert

- Man legt das Signifikanzlevel α_0 fest und evaluiert die Powerfunktion.
- Man wählt einen Mindestparameterwert d^*, den man mit $\pi(d, n) = \beta$ detektieren möchte.
- Ein konventioneller Wert ist $\beta = 0.8$.
- Man liest die für $\pi(d = d^*, n) = \beta$ nötige Stichprobengröße n ab.

28. Erläutern Sie die Motivation zur Auswertung von p-Werten.

- Es werde ein zweiseitiger Einstichproben-T-Test mit $n = 12$ und $\alpha_0 = 0.05$ durchgeführt.
  - H_0 wird abgelehnt, wenn $|T| \geq 2.20$.
- Nehmen wir an, es werde $t = 2.26$ beobachtet.
  - Das Testergebnis lautet "H_0 Ablehnen".
- Nehmen wir an, es werde $t = 3.81$ beobachtet.
  - Das Testergebnis lautet "H_0 Ablehnen".
- Der alleinige Bericht des Testergebnis supprimiert interessante Information.

$\Rightarrow$ Neben der Testumfangkontrolle durch z.B. $\alpha_0 = 0.05$ ist es daher üblich, alle Werte von α_0 anzugeben, für die ein Level-α_0-Test zum Ablehnen von H_0 führen würde.

  - Bei $t = 2.26$ würde H_0 für jedes α_0 mit $2.26 \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; 11\right)$ abgelehnt werden.
  - Bei $t = 3.81$ würde H_0 für jedes α_0 mit $3.81 \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; 11\right)$ abgelehnt werden.

- Das kleinste Signifikanzlevel α_0, bei dem man H_0 basierend auf einem Wert der Teststatistik ablehnen würde, wird *p-Wert* des Wertes der Teststatistik genannt.

29. Definieren Sie den Begriff des p-Werts.

Definition (p-Wert)

ϕ sei ein kritischer Wert-basierter Test. Der *p-Wert* ist das kleinste Signifikanzlevel α_0, bei welchem man die Nullhypothese basierend auf einem vorliegendem Wert der Teststatistik ablehnen würde.

Beispiel (Zweiseitiger Einstichproben-T-Test mit einfacher Nullhypothese)

- Bei $T = t$ würde H_0 für jedes α_0 mit $|t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right)$ abgelehnt werden. Für diese α_0 gilt, wie unten gezeigt,

$$\alpha_0 \geq 2\mathbb{P}(T \geq |t|). \tag{27}$$

- Das kleinste $\alpha_0 \in [0, 1]$ mit $\alpha_0 \geq 2\mathbb{P}(T \geq |t|)$ ist dann $\alpha_0 = 2\mathbb{P}(T \geq |t|)$, also folgt

$$\text{p-Wert} = 2\mathbb{P}(T \geq |t|) = 2(1 - \Psi(|t|; n-1)). \tag{28}$$

- Zum Beispiel ist bei $n = 12$ für $T = 2.26$ der p-Wert 0.045, für $T = -2.26$ ist der p-Wert auch 0.045, für $T = 3.81$ ist der p-Wert 0.003 und für $T = -3.81$ ist der p-Wert auch 0.003.

30. Geben Sie das Theorem zur Dualität von Konfidenzintervallen und Hypothesentests wieder.

Theorem (Dualität von Konfidenzintervallen und Hypothesentest)

$\upsilon = \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei eine Stichprobe mit Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ. Weiterhin sei $[G_u(\upsilon), G_o(\upsilon)]$ ein δ-Konfidenzintervall für θ. Dann ist der Hypothesentest

$$\phi_\theta : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := \begin{cases} 0, & [G_u(y), G_o(y)] \ni \theta_0 \\ 1, & [G_u(y), G_o(y)] \not\ni \theta_0 \end{cases} \tag{29}$$

ein Test vom Signifikanzlevel $\alpha_0 = 1 - \delta$ für die Hypothesen

$$\Theta_0 := \{\theta_0\} \text{ und } \Theta_1 := \Theta \setminus \{\theta_0\}. \tag{30}$$

31. Erläutern Sie die Dualität von Konfidenzintervallen und Hypothesentests.

Mit δ-Konfidenzintervallen können Tests vom Signifikanzniveau $\alpha_0 = 1 - \delta$ konstruiert werden.

Dieser Test ergibt den Wert 0 (Nullhypothese ablehnen), wenn das Konfidenzintervall einen Nullhypothesenparameterwert einschließt und den Wert 1, wenn (Nullhypothese *nicht* ablehnen), wenn das Konfidenzintervall den Nullhypothesenparameterwert nicht überdeckt.