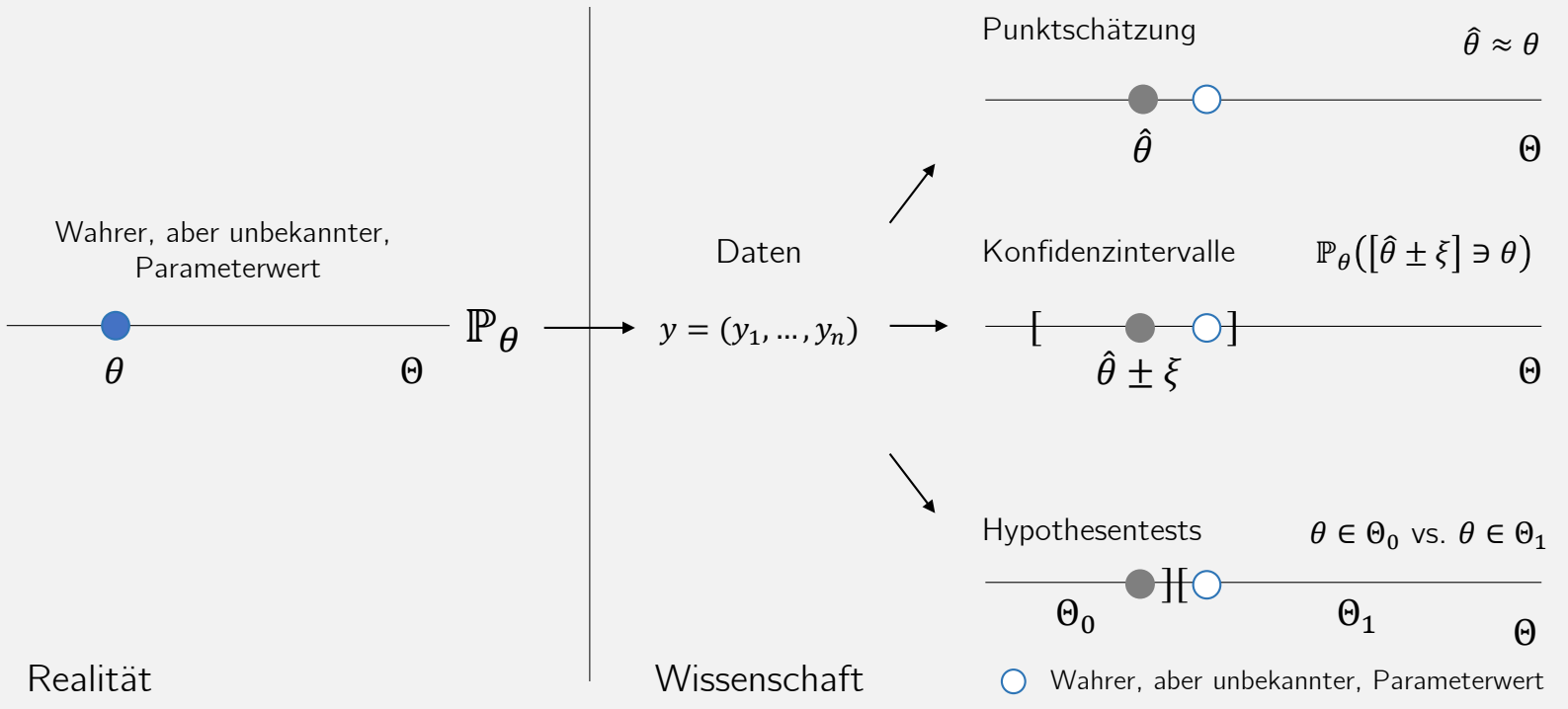
Wahrscheinlichkeitstheorie und Frequentistische Inferenz

BSc Psychologie WiSe 2023/24

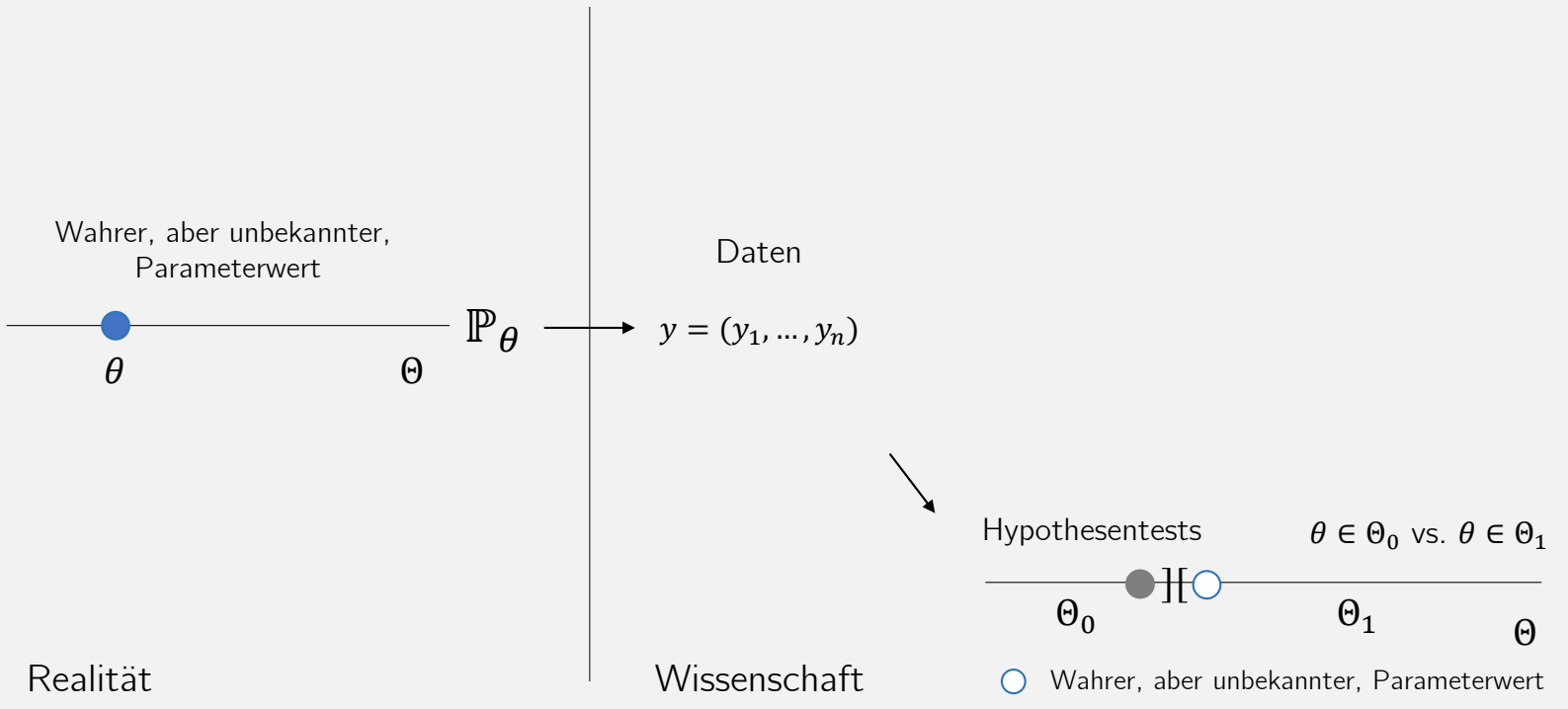
Prof. Dr. Dirk Ostwald

(11) Hypothesentests

Standardproblemstellungen und Standardannahme Frequentistischer Inferenz



Standardproblemstellungen und Standardannahme Frequentistischer Inferenz



Standardannahme Frequentistischer Inferenz

$\mathcal{M}$ sei ein Frequentistisches Inferenzmodell mit Zufallsvektor v. Es wird angenommen, dass ein Datensatz $y \in \mathbb{R}^n$ eine der möglichen Realisierungen von v ist.

Aus Frequentistischer Sicht kann man eine Studie unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

Datensatz (1) : $y^{(1)} = \left(y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)}\right)$ mit $\bar{y}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(1)}$

Datensatz (2) : $y^{(2)} = \left(y_1^{(2)}, y_2^{(2)}, ..., y_n^{(2)}\right)$ mit $\bar{y}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(2)}$

Datensatz (3) : $y^{(3)} = \left(y_1^{(3)}, y_2^{(3)}, ..., y_n^{(3)}\right)$ mit $\bar{y}^{(3)} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(3)}$

Datensatz (4) : $y^{(4)} = \left(y_1^{(4)}, y_2^{(4)}, ..., y_n^{(4)}\right)$ mit $\bar{y}^{(4)} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(4)}$

Datensatz (5) : $y^{(5)} = ...$

Um die Qualität ihrer Methoden zu beurteilen betrachtet die Frequentistische Inferenz deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken. Was zum Beispiel ist die Verteilung der $\bar{y}^{(1)}$, $\bar{y}^{(2)}$, $\bar{y}^{(3)}$, $\bar{y}^{(4)}$, ... also die Verteilung der Zufallsvariable $\bar{v}_n$?

Wenn eine Methode im Sinne der Frequentistischen Standardannahme "gut" ist, dann heißt das also, dass sie bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden Datensatzes, kann sie auch "schlecht" sein.

Grundlegende Logik Frequentistischer 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 Stichprobenstandardabweichung s und dem Stichprobenumfang n normalisierte Stichprobenmittel $\bar{y}$, $\sqrt{n}\frac{\bar{y}}{s}$.

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 dabei natürlich Frequentistisch zu verstehen, also als idealisierte relative Häufigkeit, über viele viele Stichprobenrealisationen.

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 in einem konkret vorliegenden Fall nun wirklich das Nullmodel oder ein anderes Modell den Datensatz generiert hat. Man kann aber durch die entsprechende Konstruktion der Prozedur sicherstellen, dass man im langfristigen Mittel sinnvolle Entscheidungen trifft, wenn die Annahmen der Prozedur zutreffen und man die Prozedur sehr oft wiederholt.

Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Definition (Testhypothesen und Testszenario)

Gegeben sei ein Frequentistisches Inferenzmodell mit Stichprobe υ, Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ. Weiterhin sei $\{\Theta_0, \Theta_1\}$ eine Partition des Parameterraums, so dass

$$\Theta = \Theta_0 \cup \Theta_1 \text{ und } \Theta_0 \cap \Theta_1 = \emptyset. \tag{1}$$

Dann ist eine *Testhypothese* 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* und
- $\theta \in \Theta_1$ als *Alternativhypothese*

bezeichnet. Der Einfachheit halber bezeichnet man auch Θ_0 und Θ_1 direkt als Nullhypothese und Alternativhypothese, respektive. Die Einheit aus Frequentistischem Inferenzmodell und Testhypothesen wird als *Testszenario* bezeichnet.

Definition (Einfache und zusammengesetzte Testhypothesen)

Für die Testhypothesen Θ_i mit $i = 0, 1$ 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.

Definition (Einseitige und zweiseitige Testhypothesen)

Gegeben sei ein Testszenario mit eindimensionalem Parameteraum $\Theta := \mathbb{R}$ und es sei $\theta_0 \in \Theta$.

- Dann werden zusammengesetzte Nullhypothesen der Form $\Theta_0 :=]-\infty, \theta_0]$ oder $\Theta_0 := [\theta_0, \infty[$ *einseitige Nullhypothesen* genannt und auch in der Form $H_0 : \theta \leq \theta_0$ bzw. $H_0 : \theta \geq \theta_0$ geschrieben. Die entsprechenden Alternativhypothesen haben dabei die Form $\Theta_1 :=]\theta_0, \infty[$ bzw. $\Theta_1 :=]-\infty, \theta_0[$, auch geschrieben als $H_1 : \theta > \theta_0$ bzw. $H_1 : \theta < \theta_0$.
- Bei einer einfachen Nullhypothese der Form $\Theta_0 := \{\theta_0\}$, auch geschrieben als $H_0 : \theta = \theta_0$, wird die Alternativhypothese $\Theta_1 := \Theta \setminus \{\theta_0\}$, auch geschrieben als $H_1 : \theta \neq \theta_0$, *zweiseitige Alternativhypothese* genannt.

Definition (Test)

Gegeben sei ein Testszenario. Dann ist ein *Test* eine Abbildung ϕ aus dem Ergebnisraum der Stichprobe $\mathcal{Y}$ in die Menge $\{0, 1\}$, also

$$\phi : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi(y), \tag{2}$$

wobei

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

repräsentieren.

Bemerkung

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

Definition (Standardtest)

Gegeben sei ein Testszenario. Dann ist ein *Standardtest* definiert als die Verkettung einer *Teststatistik*

$$\gamma : \mathcal{Y} \to \mathbb{R} \tag{3}$$

und einer *Entscheidungsregel*

$$\delta : \mathbb{R} \to \{0, 1\}. \tag{4}$$

Ein Standardtest kann also geschrieben werden als

$$\phi := \delta \circ \gamma : \mathcal{Y} \to \{0, 1\}. \tag{5}$$

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.

Definition (Kritischer Bereich)

Gegeben sei ein Testszenario und ein Test ϕ. Dann heißt die Untermenge K des Ergebnisraums $\mathcal{Y}$ der Stichprobe υ, für die der Test den Wert 1 annimmt, *kritischer Bereich* des Tests, formal

$$K := \{y \in \mathcal{Y} | \phi(y) = 1\} \subset \mathcal{Y}. \tag{6}$$

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.

Definition (Ablehnungsbereich)

Gegeben sei ein Testszenario und ein Standardtest ϕ mit Teststatistik γ. Die Untermenge A des Ergebnisraums der Teststatistik, für die der Test den Wert 1 annimmt, *Ablehnungsbereich des Tests*, formal

$$A := \{\gamma(y) \in \mathbb{R} | \phi(y) = 1\} \subset \mathbb{R}. \tag{7}$$

Bemerkungen

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

Definition (Kritischer Wert-basierte Tests)

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

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

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := 1_{\{\gamma(y) \geq k\}} = \begin{cases} 1 & \gamma(y) \geq k \\ 0 & \gamma(y) < k \end{cases} \tag{8}$$

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

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := 1_{\{|\gamma(y)| \geq k\}} = \begin{cases} 1 & |\gamma(y)| \geq k \\ 0 & |\gamma(y)| < k \end{cases} \tag{9}$$

Bemerkung

- Wir betrachten in der Folge nur kritischer Wert-basierte Tests.

Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Definition (Richtige Testentscheidungen und Testfehler)

Gegeben seien ein Testszenario und ein Test.

- Dann gibt es mit dem Nichtablehnen der Nullhypothese $\phi(y) = 0$, wenn die Nullhypothese $\theta \in \Theta_0$ zutrifft und dem Ablehnen der Nullhypothese $\phi(y) = 1$, wenn die Alternativhypothese $\theta \in \Theta_1$ zutrifft zwei Formen der *richtigen Testentscheidung*.
- Ebenso gibt es zwei Arten von *Testfehlern*: Das Ablehnen der Nullhypothese $\phi(y) = 1$, wenn die Nullhypothese $\theta \in \Theta_0$ zutrifft, heißt *Typ I Fehler* und das Nichtablehnen der Nullhypothese, wenn die Alternativhypothese $\theta \in \Theta_1$ zutrifft, heißt *Typ II Fehler*.

Bemerkung

Testentscheidung

Wahrer, aber unbekannter, Parameter	$\phi(v) = 0$	$\phi(v) = 1$
$\theta \in \Theta_0$	Richtige Entscheidung	Typ I Fehler
$\theta \in \Theta_1$	Typ II Fehler	Richtige Entscheidung

Definition (Testgütefunktion)

Gegeben sei ein Testszenario und ein Test ϕ. Dass ist die *Testgütefunktion* von ϕ definiert als

$$q_\phi : \Theta \to [0,1], \theta \mapsto q_\phi(\theta) := \mathbb{P}_\theta(\phi(\upsilon) = 1). \tag{10}$$

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

Bemerkungen

- $\mathbb{P}_\theta$ bezeichnet die Verteilung von ϕ unter der Annahme $\upsilon \sim \mathbb{P}_\theta$.
- Es gilt $\mathbb{P}_\theta(\phi(\upsilon) = 1) = \mathbb{P}_\theta(\upsilon \in K) = \mathbb{P}_\theta(\gamma \in A)$
- Für jedes $\theta \in \Theta$ liefert q_ϕ die Wahrscheinlichkeit, dass die Nullhypothese durch ϕ abgelehnt wird.
- Bei Poweranalysen betrachtet man q_ϕ als Funktion aller Testszenario und Testparameter.
- Ändert sich ϕ, z.B. weil sich der kritische Wert von ϕ ändert, dann ändert sich $q_\phi(\theta)$.
- Im Idealfall hätte man einen Test ϕ mit

$$q_\phi(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 1 \text{ für } \theta \in \Theta_1. \tag{11}$$

- Die Testentscheidung eines solchen ϕ wäre mit Wahrscheinlichkeit 1 richtig.

Intuition zur Testkonstruktion

Im Idealfall hätte man einen Test ϕ mit

$$q_\phi(\theta) = \mathbb{P}_\theta(\phi(v) = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi(v) = 1) = 1 \text{ für } \theta \in \Theta_1. \tag{12}$$

$\Rightarrow$ 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 $\theta \in \Theta_0$.

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.

Intuition zur Testkonstruktion

Die populärste Methode, eine Balance zwischen zwischen kleinen Werten von q für $\theta \in \Theta_0$ und großen Werten von q für $\theta \in \Theta_1$ zu finden, ist in einem ersten Schritt ein $\alpha_0 \in [0, 1]$ zu wählen und sicher zu stellen, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0. \tag{13}$$

Eine konventionelle Wahl für sein solches α_0 ist zum Beispiel $\alpha_0 := 0.05$.

Unter allen Tests und statistischen Modellen, die Ungleichung (13) 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.

Dieses Vorgehen ist nicht alternativlos, man kann zum Beispiel auch lineare Kombinationen verschiedener Fehlerwahrscheinlichkeiten minimieren. Es ist aber das in der Anwendung populärste Vorgehen. Wir werden uns deshalb in der Folge auf dieses Vorgehen beschränken.

Das beschriebene Vorgehen motiviert die folgenden Definitionen der Begriffe des Level-α_0-Tests, des *Signifikanzlevels* α_0 (manchmal auch als *nominales Niveau* bezeichnet) und des Testumfangs α (manchmal auch als *effektives Niveau* bezeichnet).

Definition (Level-α_0-Test, Signifikanzlevel α_0, Testumfang α)

Gegeben seien ein Testszenario, ein Test ϕ, seine Testgütefunktion q_ϕ und ein $\alpha_0 \in [0,1]$. ϕ heißt ein *Level-α_0-Test*, wenn gilt, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0. \tag{14}$$

Wenn ϕ ein Level-α_0-Test ist, nennt man den Wert α_0 auch das *Signifikanzlevel* des Tests. Weiterhin heißt die Zahl

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) \in [0,1] \tag{15}$$

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 $\alpha = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi(v) = 1)$.

Typ I Fehlerwahrscheinlichkeit vs. Testumfang vs. Signifikanzlevel

Bei einfacher Θ_0 ist der Testumfang gleich der Wahrscheinlichkeit eines Typ I Fehlers

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \{\theta_0\}} q_\phi(\theta) = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi(\upsilon) = 1). \tag{16}$$

Bei zusammengesetzter Θ_0 gibt es je nach Wert von $\theta \in \Theta_0$ verschiedene Wahrscheinlichkeiten für einen Typ I Fehler. Die größte dieser Wahrscheinlichkeiten ist der Testumfang

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi(\upsilon) = 1). \tag{17}$$

Ein Test hat Signifikanzlevel α_0, wenn der Testumfang kleiner oder gleich α_0 ist.

$$\alpha = \max_{\theta \in \Theta_0} q_\phi(\theta) \leq \alpha_0 \tag{18}$$

Ein Test, bei dem das Signifikanzlevel größer als der Testumfang ist, heißt *konservativ*.

Ein Test, bei dem das Signifikanzlevel gleich dem Testumfang ist, heißt *exakt*.

Ein Test, bei dem das Signifikanzlevel kleiner als dem Testumfang ist, heißt *liberal*.

Motivation des p-Wert Begriffs

Es werde ein zweiseitiger kritischer Wert-basierter Test durchgeführt.

- H_0 wird abgelehnt, wenn $|\gamma(\upsilon)| \geq k$.

Nehmen wir an, es werde $\gamma(\upsilon) = k + \frac{k}{100}$ beobachtet.

- Das Testergebnis lautet $\phi(\upsilon) = 1 \Leftrightarrow$ Ablehnen der Nullhypothese

Nehmen wir an, es werde $\gamma(\upsilon) = k + 100k$ beobachtet.

- Das Testergebnis lautet $\phi(\upsilon) = 1 \Leftrightarrow$ Ablehnen der Nullhypothese

Der Bericht des Testergebnis allein supprimiert potentiell interessante Information.

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

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.

Bemerkung

- Eine Intuition zu dieser Definition ergibt sich im Rahmen des Einstichproben-T-Tests. p-Werte spiegeln dort die Antwort auf die intuitive Frage wie wahrscheinlich es im Frequentistischen Sinne wäre, den beobachteten oder einen extremeren Wert der Teststatistik unter der Annahme eines Nullmodels zu observieren.
- p-Werte sind ist extrem populär, aber auch sehr umstritten.
- p-Werte werden, wie Hypothesentestergebnisse generell, leider oft überinterpretiert.
- Es gibt basierend auf dem Gesagten keinen Grund dies anzunehmen, trotzdem vorsorglich:
  - p-Werte quantifizieren nicht die Wahrscheinlichkeit, dass die Nullhypothese wahr ist.
  - Aufgrund von $p < 0.05$ sollte man nicht glauben, dass ein Effekt existiert.
  - Aufgrund von $p > 0.05$ sollte man nicht glauben, dass ein Effekt nicht existiert.
- p-Werte sind eine der vielen Möglichkeit ein Signal-zu-Rauschen Verhältnis zu quantifizieren.
- p-Werte sind eine der vielen Möglichkeiten, Unsicherheit zu quantifizieren.

(vgl. Wasserstein, Schirm, and Lazar (2019))

Kommentar zum Frequentistischen Hypothesentesten in der Wissenschaft

Frequentistisches Hypothesentesten ist als Entscheidungsproblem ohne klar und explizit definierte Entscheidungsnutzenfunktion formuliert und deshalb recht mühselig zu analysieren und zu studieren. Es gibt sehr viel zugänglichere Theorien zu Entscheidungen unter Unsicherheit (vgl. Pratt, Raiffa, and Schlaifer (1995), Puterman (2005), Ostwald, Starke, and Hertwig (2015))

Oberflächlich betrachtet liefern Hypothesentests einfache binäre Aussagen der Form "Die Hypothese (Theorie) ist gegeben die Evidenz abzulehnen oder zu akzeptieren". Solche Aussagen sind im Entscheidungskontext hilfreich, denn es muss etwas passieren, also eine Entscheidung getroffen werden. In der Wissenschaft, also der menschlichen Kommunikationsstruktur über die Beschaffenheit der Welt, muss aber nichts final entschieden, sondern nur das Maß an Unsicherheit über den gerade vorherrschenden Theoriestand quantifiziert und kommuniziert werden. Generell sollten Fragestellungen der Grundlagenwissenschaften deshalb gerade nicht als Entscheidungsprobleme formuliert werden.

Trotz landläufiger Meinung das Bayesianische Herangehensweisen wie Positive Predictive Values oder Bayes Factors hier irgendwie besser sind, ist dem nicht so, so lange die mit einer gewissen Modellpräferenz assoziierte Unsicherheit nicht klar mitkommuniziert wird.

Und trotz alledem ist Frequentistisches Hypothesentesten in der Wissenschaftscommunity weiterhin sehr populär und sollte deshalb im Rahmen eines wissenschaftlichen Studiums wie der Psychologie intellektuell durchdrungen werden.

Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Einstichproben-T-Test

(1) Anwendungsszenario

(2) Frequentistisches Inferenzmodell

(3) Testhypothesen

(4) Definition der Teststatistik

(5) Verteilung der Teststatistik

(6) Testdefinition

(7) Testgütefunktion

(8) Testumfangkontrolle

(9) p-Wert

(10) Powerfunktion

(11) Praktische Durchführung

Anwendungsszenario eines Einstichproben-T-Tests

Eine Stichprobe (Gruppe) randomisierter experimenteller Einheiten.

Annahme der unabhängigen und identischen Normalverteilung $N(\mu, \sigma^2)$ der Datenpunkte.

μ und σ^2 unbekannt.

Absicht der Inferenz hinsichtlich einer Nullhypothese und einer Alternativhypothese

Anwendungsbeispiele

Pre-Post-Psychotherapie BDI Differenzanalyse einer Gruppe von Patient:innen

- $\mu \neq \mu_0 := 0 \quad \Rightarrow$ Evidenz für Depressionsymptomatikveränderung

Gruppenanalysen mit Wechsler Adult Intelligence Scale

- $\mu \neq \mu_0 := 100 \Rightarrow$ Evidenz für über- oder unterdurchschnittliche WAIS Performanz

Gruppenanalysen in der funktionellen Kernspintomographie

- $\mu > \mu_0 := 0 \quad \Rightarrow$ Evidenz für regionale Gehirnaktivierung

Hypothesen- und Testszenarien bei Einstichproben-T-Tests

Einfache Nullhypothese, einfache Alternativhypothese $H_0 : \mu = \mu_0, H_1 : \mu = \mu_1$

- Theoretisch wichtiges Szenario (Neymann-Pearson Lemma)
- Praktische Relevanz eher gering

Einfache Nullhypothese, zusammengesetzte Alternativhypothese $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$

- Zweiseitiger Einstichproben-T-Test mit ungerichteter Hypothese
- Ungerichtete Fragestellung nach einem Unterschied

Zusammengesetzte Nullhypothese/Alternativhypothese $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$

- Einseitiger Einstichproben-T-Test mit gerichteter Hypothese
- Gerichtete Fragestellung nach einem positiven Unterschied

Zusammengesetzte Nullhypothese/Alternativhypothese $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$

- Gerichtete Fragestellung nach einem negativen Unterschied
- Qualitativ äquivalente Theorie zum umgekehrten Fall

Hier betrachtetes Hypothesen- und Testszenario eines Einstichproben-T-Tests

Einfache Nullhypothese, einfache Alternativhypothese $H_0 : \mu = \mu_0, H_1 : \mu = \mu_1$

- Theoretisch wichtiges Szenario (Neymann-Pearson Lemma)
- Praktische Relevanz eher gering

Einfache Nullhypothese, zusammengesetzte Alternativhypothese $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$

- Zweiseitiger Einstichproben-T-Test mit ungerichteter Hypothese
- Ungerichtete Fragestellung nach einem Unterschied

Zusammengesetzte Nullhypothese/Alternativhypothese $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$

- Einseitiger Einstichproben-T-Test mit gerichteter Hypothese
- Gerichtete Fragestellung nach einem positiven Unterschied

Zusammengesetzte Nullhypothese/Alternativhypothese $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$

- Gerichtete Fragestellung nach einem negativen Unterschied
- Qualitativ äquivalente Theorie zum umgekehrten Fall

Definition (Frequentistisches Inferenzmodell des Einstichproben-T-Tests)

Das Frequentistische Inferenzmodell des Einstichproben-T-Tests ist gegeben durch das Normalverteilungsmodell

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \tag{19}$$

Bemerkung

- Wir erinnern daran, dass aus generativer Sicht das Normalverteilungsmodell dem Modell

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, ..., n \tag{20}$$

  entspricht.

Definition (Testhypothesen des Einstichproben-T-Tests)

Gegeben sei das Frequentistische Inferenzmodell des Einstichproben-T-Tests

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \tag{21}$$

und es sei $\Theta := \mathbb{R}$ der Parameterunteraum des Parameters von Interesse μ. Dann sind für den *Nullhypothesenparameterwert* $\mu_0 \in \mathbb{R}$ die *einfache Nullhypothese* und die *zusammengesetzte Alternativhypothese* des Einstichproben-T-Tests gegeben durch

$$\Theta_0 := \{\mu_0\} \Leftrightarrow H_0 : \mu = \mu_0 \text{ und } \Theta_1 := \mathbb{R} \setminus \{\mu_0\} \Leftrightarrow H_1 : \mu \neq \mu_0. \tag{22}$$

Bemerkung

- μ ist der wahre, aber unbekannte, Parameter, μ_0 ist der Nullhypothesenparameter.

Definition (Einstichproben-T-Test-Statistik)

Gegeben sei das Testszenario eines Einstichproben-T-Tests mit Stichprobe $\upsilon_1, ..., \upsilon_n$, Stichprobenmittel $\bar{\upsilon}$, Stichprobenstandardabweichung S und Nullhypothesenparameter μ_0. Dann ist die *Einstichproben-T-Test-Statistik* definiert als

$$T := \sqrt{n}\frac{\bar{\upsilon} - \mu_0}{S}. \tag{23}$$

Bemerkungen

- Wir kürzen den Begriff *Einstichproben-T-Test-Statistik* z.T. auch mit ETT-Statistik ab.
- Im Gegensatz zur T-Konfidenzintervallstatistik muss bei der ETT-Statistik nicht $\mu_0 = \mu$ gelten.
- Intuitiv kann die ETT-Statistik als mit dem Stichprobenumfang (Evidenz) gewichtetes Verhältnis von Signal (sytematischer Variabilität) zu Rauschen (unsystematischer Variabilität) verstanden werden:

$$\sqrt{\text{Stichprobenumfang}}\left(\frac{\text{Signal}}{\text{Rauschen}}\right) = \sqrt{n}\frac{\bar{\upsilon} - \mu_0}{S} \tag{24}$$

- Die ETT-Statistik ist eine skalare Deskription des Effekt vs. Variabilität Verhältnisses eines Datensatzes.
- In der ETT-Statistik 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 Einstichproben-T-Test-Statistik)

Gegeben sei das Testszenario eines Einstichproben-T-Tests mit Stichprobe $v_1, ..., v_n$, Stichprobenmittel $\bar{v}$, Stichprobenstandardabweichung S, Nullhypothesenparameter μ_0 und Einstichproben-T-Test-Statistik definiert als

$$T := \sqrt{n} \frac{\bar{v} - \mu_0}{S}. \tag{25}$$

Dann ist T eine nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter

$$d = \sqrt{n} \frac{\mu - \mu_0}{\sigma} \tag{26}$$

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

Bemerkungen

- Wir verzichten auf einen Beweis

Definition (Nichtzentrale t-Zufallsvariable)

T sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

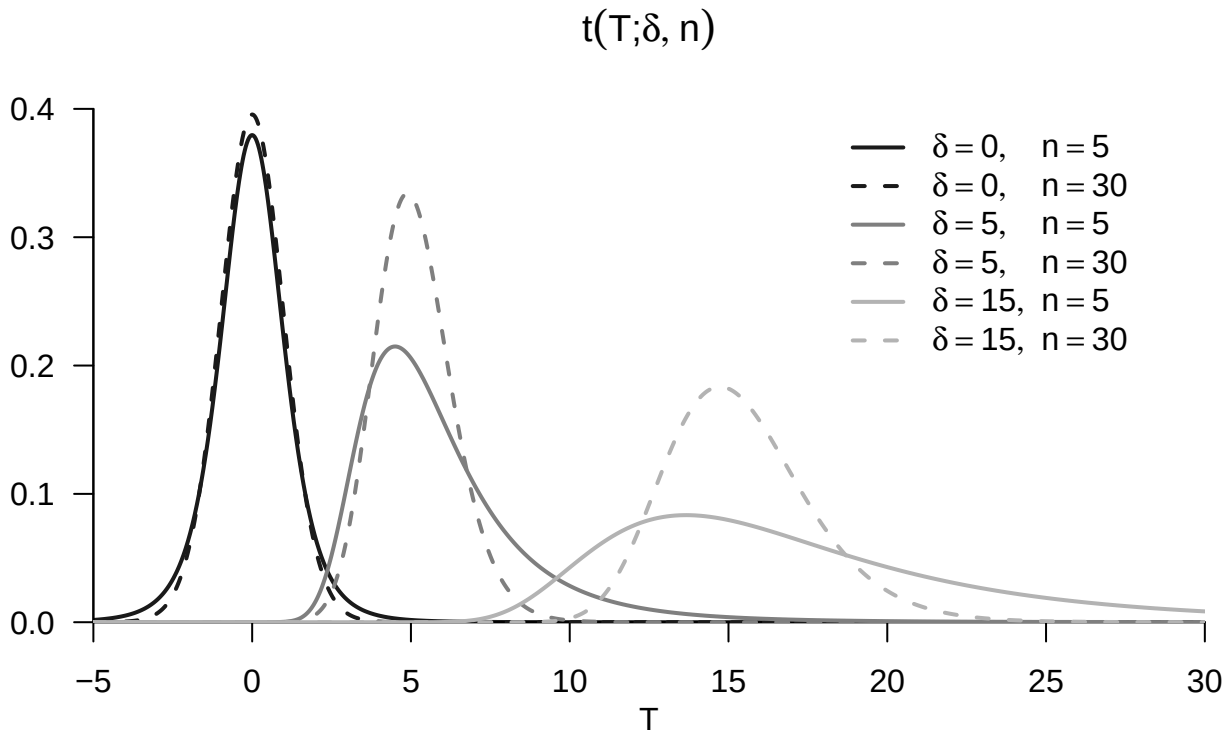
$$p : \mathbb{R} \to \mathbb{R}_{>0}, t \mapsto p(t) := \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n}{2}\right) (n\pi)^{\frac{1}{2}}} \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left(-\frac{\tau}{2}\right) \exp\left(-\frac{1}{2}\left(t\left(\frac{\tau}{n}\right)^{\frac{1}{2}} - \delta\right)^2\right) d\tau. \quad (27)$$

Dann sagen wir, dass T einer nichtzentralen t-Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n unterliegt und nennen T eine *nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n*. Wir kürzen dies mit $t(\delta, n)$ ab. Die WDF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $t(T; \delta, n)$. Die KVF und inverse KVF einer nichtzentralen t-Zufallsvariable bezeichnen wir mit $\Psi(\cdot; \delta, n)$ und $\Psi^{-1}(\cdot; \delta, n)$, respektive.

Bemerkungen

- Eine nichtzentrale t-Zufallsvariable mit $\delta = 0$ ist eine t-Zufallsvariable.
- Es gilt also $t(T; 0, n) = t(T; n)$.
- Weiterhin gelten $\Psi(T; 0, n) = \Psi(T; n)$ und $\Psi^{-1}(T; 0, n) = \Psi^{-1}(T; n)$
- Die funktionale Form der WDF findet sich zum Beispiel in Lehmann (1986), Seite 254, Gl. (80).

Wahrscheinlichkeitsdichtefunktionen nichtzentraler t-Verteilungen



Theorem (Nichtzentrale T-Transformation)

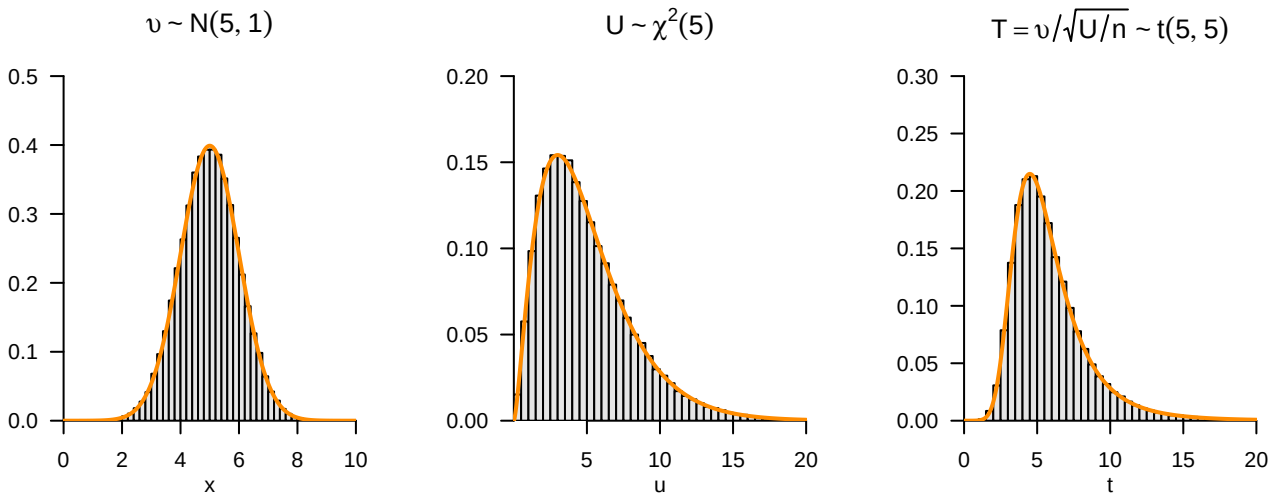
$\upsilon \sim N(\mu, 1)$ sei eine normalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2 Zufallsvariable mit Freiheitsgradparameter n, und υ und U seien unabhängige Zufallsvariablen. Dann ist die Zufallsvariable

$$T := \frac{\upsilon}{\sqrt{U/n}} \tag{28}$$

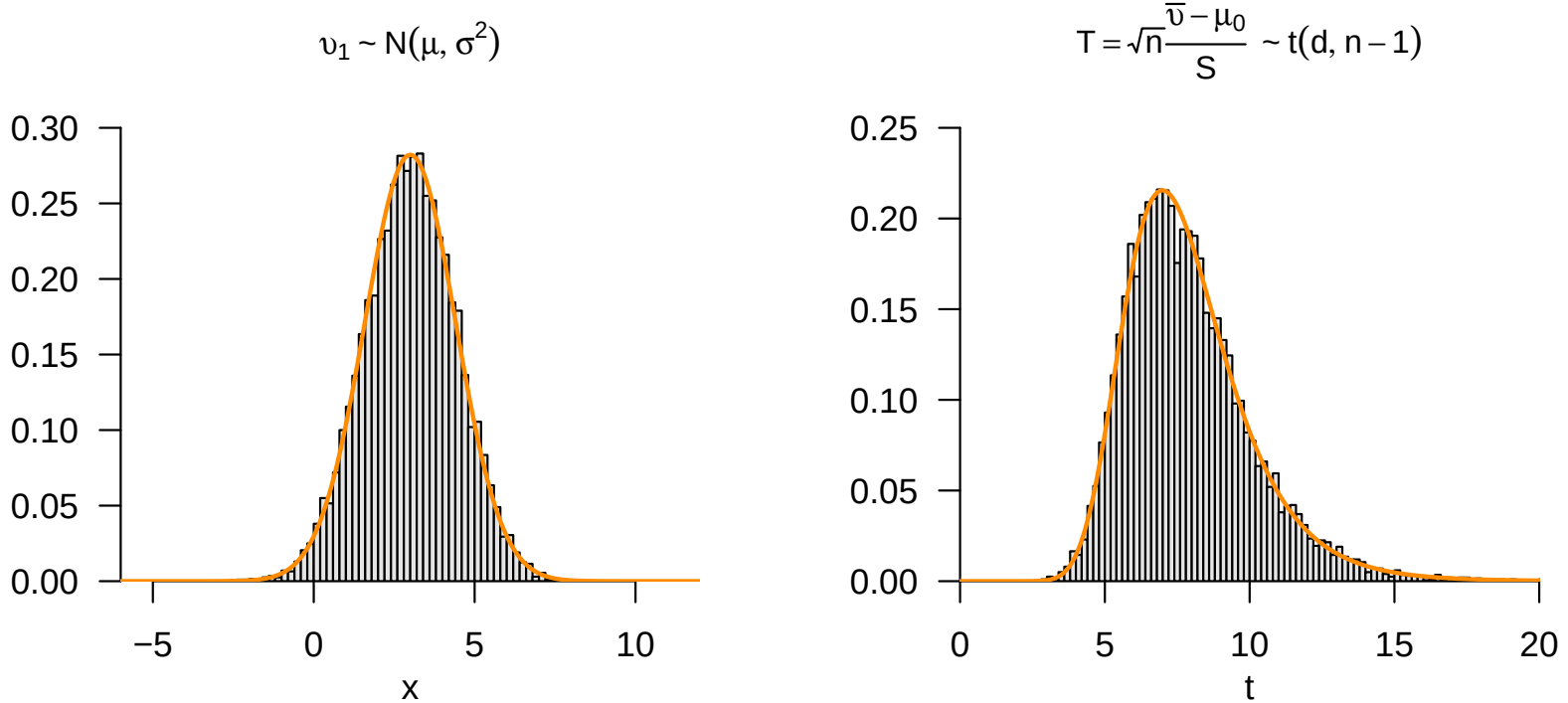
eine nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter μ und Freiheitsgradparameter n, also $T \sim t(\mu, n)$.

Bemerkung

- Wir verzichten auf einen Beweis.



Einstichproben-T-Test-Statistik bei $n = 12, \mu = 3, \sigma^2 = 2, \mu_0 = 0$



Definition (Zweiseitiger Einstichproben-T-Test)

Gegeben seien das Frequentistische Inferenzmodell des Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese und T bezeichne die Einstichproben-T-Test-Statistik mit Werten $t \in \mathbb{R}$. Dann ist der *zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese* definiert als der zweiseitige kritische Wert-basierte Test

$$\phi : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi(y) := 1_{\{|t| \geq k\}} = \begin{cases} 1 & |t| \geq k \\ 0 & |t| < k \end{cases}. \tag{29}$$

Theorem (Testgütefunktion des Einstichproben-T-Test)

ϕ sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. Dann ist die Testgütefunktion von ϕ gegeben durch

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1), \tag{30}$$

wobei $\Psi(\cdot; d_\mu, n-1)$ die KVF der nichtzentralen t-Verteilung mit Nichtzentralitätsparameter

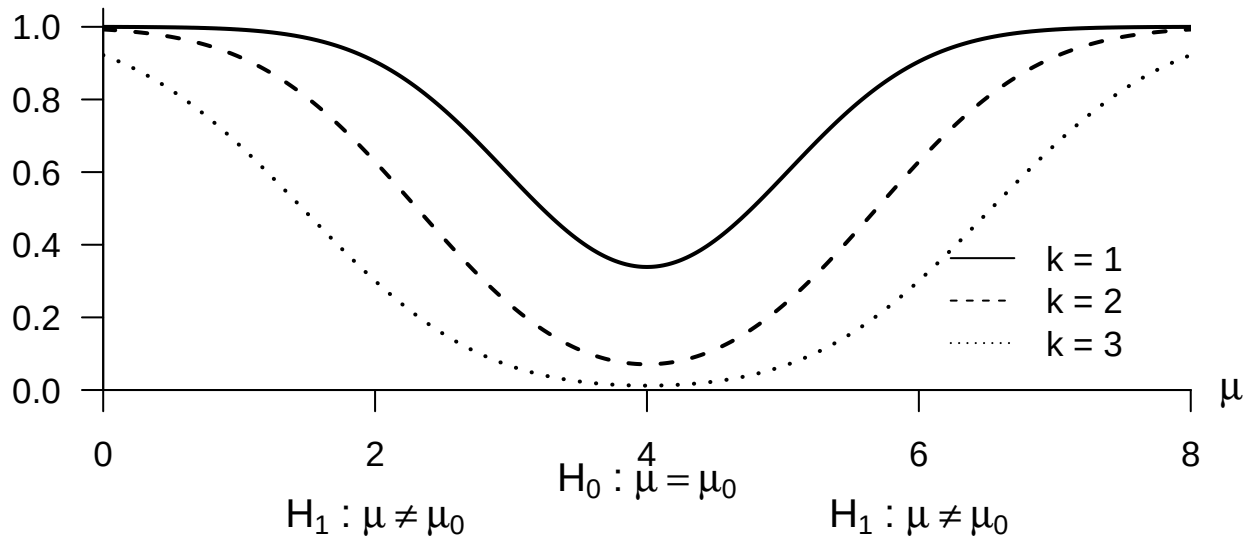
$$d_\mu := \sqrt{n}\frac{\mu - \mu_0}{\sigma} \tag{31}$$

und Freiheitsgradparameter $n-1$ bezeichnet.

Bemerkungen

- Wir visualisieren die Testgütefunktion unten in Abhängigkeit von k.

$q_\phi(\mu) = \mathbb{P}_\mu(\phi(v) = 1)$ für $\sigma^2 = 9, \mu_0 = 4, n = 12$ und $k = 1, 2, 3$



Beweis

Die Testgütefunktion des betrachteten Test im vorliegenden Testszenario ist definiert als

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1). \tag{32}$$

Da die Wahrscheinlichkeiten für $\phi = 1$ und dafür, dass die zugehörige Teststatistik im Ablehnungsbereich des Tests liegt gleich sind, benötigen wir also zunächst die Verteilung der Teststatistik. Wir haben oben bereits gesehen, dass die Einstichproben-T-Test-Statistik

$$T := \sqrt{n}\frac{\bar{v} - \mu_0}{S} \tag{33}$$

unter der Annahme $v_1, ..., v_n \sim N(\mu, \sigma^2)$ anhand einer nichtzentralen t-Verteilung $t(d_\mu, n-1)$ mit Nichtzentralitätsparameter

$$d_\mu := \sqrt{n}\frac{\mu - \mu_0}{\sigma} \tag{34}$$

verteilt ist. Der Ablehnungsbereich des zweiseitigen Einstichproben-T-Tests ist

$$A =]-\infty, -k] \cup]k, \infty[. \tag{35}$$

Beweis (fortgeführt)

Mit diesem Ablehungsbereich ergibt sich dann

$$\begin{aligned} q_\phi(\mu) &= \mathbb{P}_\mu(\phi = 1) \\ &= \mathbb{P}_\mu\left(T \in]-\infty, -k] \cup]k, \infty[\right) \\ &= \mathbb{P}_\mu\left(T \in]-\infty, -k]\right) + \mathbb{P}_\mu\left(T \in [k, \infty[\right) \\ &= \mathbb{P}_\mu(T \leq -k) + \mathbb{P}_\mu(T \geq k) \\ &= \mathbb{P}_\mu(T \leq -k) + (1 - \mathbb{P}_\mu(T \leq k)) \\ &= 1 - \mathbb{P}_\mu(T \leq k) + \mathbb{P}_\mu(T \leq -k) \\ &= 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1), \end{aligned} \tag{36}$$

wobei $\Psi(\cdot; d_\mu, n-1)$ die KVF der nichtzentralen T-Verteilung mit Nichtzentralitätsparameter d_μ und Freiheitsgradparameter $n-1$ bezeichnet.

□

Theorem (Testumfangkontrolle für den Einstichproben-T-Test)

ϕ sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. Dann ist ϕ ein Level-α_0-Test mit Testumfang α_0, wenn der kritische Wert definiert ist durch

$$k_{\alpha_0} := \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right), \tag{37}$$

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

Bemerkung

- In **R** kann Ψ^{-1} mit der Funktion `qt()` ausgewertet werden.

Beweis

Damit der betrachtete Test ein Level-α_0-Test ist, muss bekanntlich $q_\phi(\mu) \leq \alpha_0$ für alle $\mu \in \{\mu_0\}$, also hier $q_\phi(\mu_0) \leq \alpha_0$, gelten. Weiterhin ist der Testumfang des betrachteten Tests durch $\alpha = \max_{\mu \in \{\mu_0\}} q_\phi(\mu)$, also hier durch $\alpha = q_\phi(\mu_0)$ gegeben. Wir müssen also zeigen, dass die Wahl von k_{α_0} garantiert, dass ϕ ein Level-α_0-Test mit Testumfang α_0 ist. Dazu merken wir zunächst an, dass für $\mu = \mu_0$ gilt, dass

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k; d_{\mu_0}, n-1) + \Psi(-k; d_{\mu_0}, n-1) \\ &= 1 - \Psi(k; 0, n-1) + \Psi(-k; 0, n-1) \\ &= 1 - \Psi(k; n-1) + \Psi(-k; n-1), \end{aligned} \tag{38}$$

wobei $\Psi(\cdot; d, n-1)$ und $\Psi(\cdot; n-1)$ die KVF der nichtzentralen t-Verteilung mit Nichtzentralitätsparameter d und Freiheitsgradparameter $n-1$ sowie der t-Verteilung mit Freiheitsgradparameter $n-1$, respektive, bezeichnen.

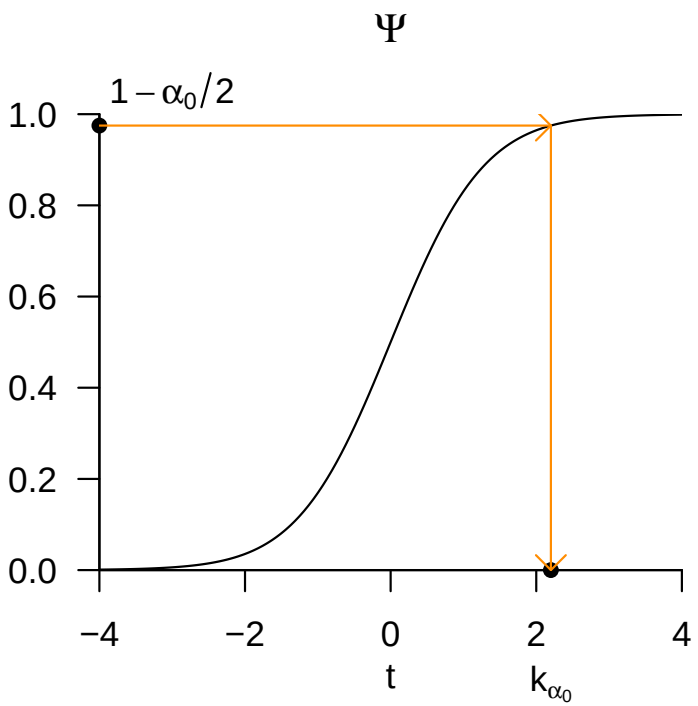
Beweis (fortgeführt)

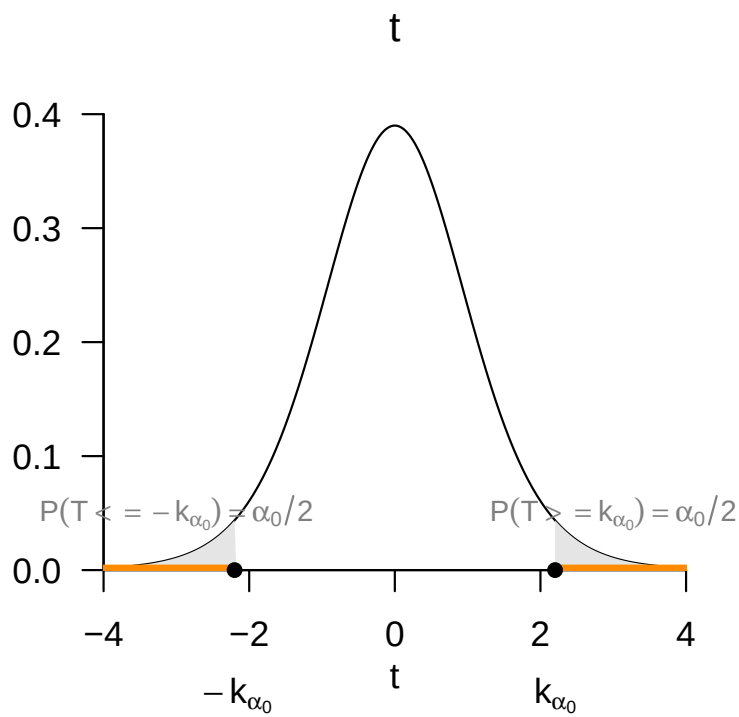
Sei nun also $k := k_{\alpha_0}$. Dann gilt

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k_{\alpha_0}; n-1) + \Psi(-k_{\alpha_0}; n-1) \\ &= 1 - \Psi(k_{\alpha_0}; n-1) + (1 - \Psi(k_{\alpha_0}; n-1) \\ &= 2(1 - \Psi(k_{\alpha_0}; n-1)) \\ &= 2\left(1 - \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}, n-1\right), n-1\right)\right) \\ &= 2\left(1 - 1 + \frac{\alpha_0}{2}\right) \\ &= \alpha_0, \end{aligned} \tag{39}$$

wobei die zweite Gleichung mit der Symmetrie der t-Verteilung folgt. Es folgt also direkt, dass bei der Wahl von $k = k_{\alpha_0}$, $q_\phi(\mu_0) \leq \alpha_0$ ist und der betrachtete Test somit ein Level-α_0-Test ist. Weiterhin folgt direkt, dass der Testumfang des betrachteten Tests bei der Wahl von $k = k_{\alpha_0}$ gleich α_0 ist.

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





Simulation

```
n         = 12                                          # Anzahl der Datenpunkte
mu        = 0                                           # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr    = 2                                           # wahrer, aber unbekannter, Varianzparameter
mu_0      = 0                                           # Nullhypothesenparameter, hier \mu = \mu_0
alpha_0   = 0.05                                        # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2,n-1)                         # Kritischer Wert
set.seed(1)                                             # Random number generator seed
nsim      = 1e5                                         # Anzahl Simulationen
phi       = rep(NaN,nsim)                               # Testentscheidungsarray
for(j in 1:nsim){                                       # Simulationsiterationen
    y      = rnorm(n,mu,sigsqr)                         # \ups_i \sim N(\mu,\Sigma), i = 1,...,n
    y_bar  = mean(y)                                    # Stichprobenmittel
    s      = sd(y)                                      # Stichprobenstandardabweichung
    Tee    = sqrt(n)*((y_bar - mu_0)/s)                 # Einstichproben-T-Test-Statistik
    if(abs(Tee) >= k_alpha_0){                          # Test 1_{\vert t \vert >= k_alpha_0}
        phi[j] = 1                                      # Ablehnen der Nullhypothese
    } else {
        phi[j] = 0                                      # Nichablehnen der Nullhypothese
    }
}
cat("Kritischer Wert              =", k_alpha_0,
    "\nGeschätzter Testumfang alpha =", mean(phi))
```

```
Kritischer Wert              = 2.200985
Geschätzter Testumfang alpha = 0.0493
```

Theorem (p-Wert)

Gegeben sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese und t sei ein Wert der Einstichproben-T-Test-Statistik T. Dann gilt

$$\text{p-Wert} = 2(1 - \Psi(|t|; n-1)) \tag{40}$$

wobei $\Psi(\cdot; n-1)$ die KVF der t-Verteilung mit Freiheitsgradparameter $n-1$ bezeichnet.

Bemerkung

- In **R** kann Ψ mit der Funktion `pt()` ausgewertet werden.

Einstichproben-T-Test | (9) p-Wert

Beweis

Per Definition ist der p-Wert das kleinste Signifikanzlevel α_0 bei dem für den betrachteten Test die Nullhypothese basierend auf dem Wert von t abgelehnt werden würde. Im vorliegenden Fall würde die Nullhypothese für jedes α_0 mit

$$|t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right) \tag{41}$$

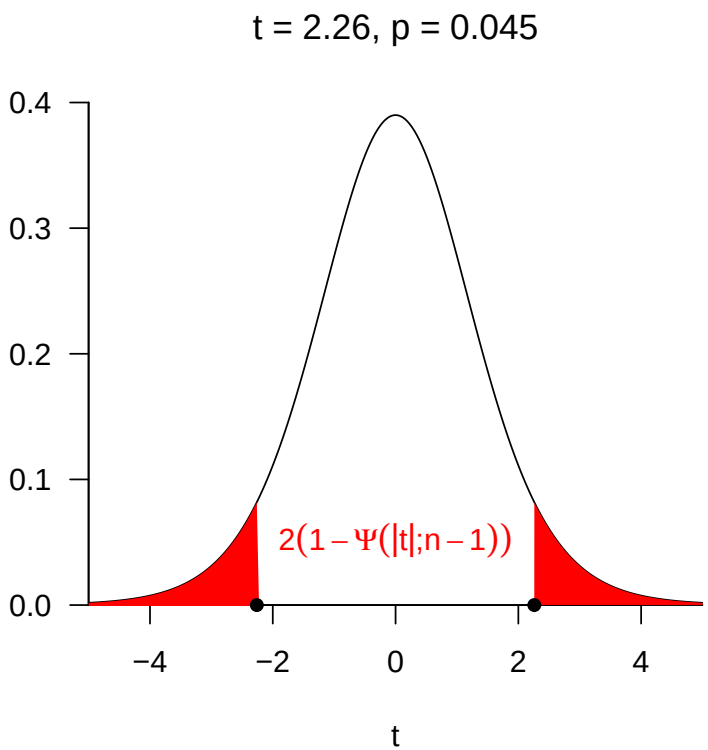
abgelehnt werden. Für diese α_0 gilt, dass

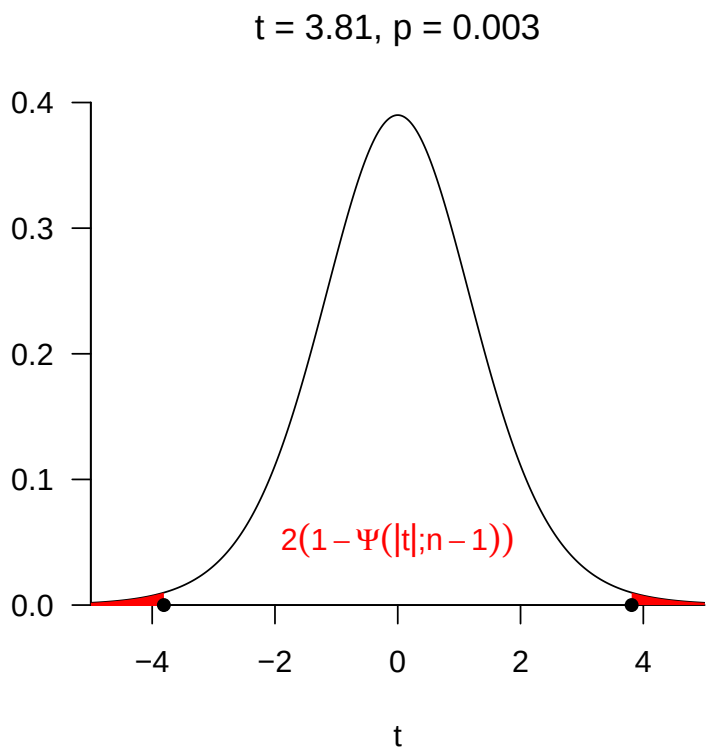
$$\alpha_0 \geq 2(1 - \Psi(|t|; n-1)), \tag{42}$$

denn

$$\begin{aligned}
& |t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right) \\
\Leftrightarrow\ & \Psi(|t|; n-1) \geq \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right); n-1\right) \\
\Leftrightarrow\ & \Psi(|t|; n-1) \geq 1 - \frac{\alpha_0}{2} \\
\Leftrightarrow\ & \mathbb{P}(T \leq |t|) \geq 1 - \frac{\alpha_0}{2} \\
\Leftrightarrow\ & \frac{\alpha_0}{2} \geq 1 - \mathbb{P}(T \leq |t|) \\
\Leftrightarrow\ & \frac{\alpha_0}{2} \geq \mathbb{P}(T \geq |t|) \\
\Leftrightarrow\ & \alpha_0 \geq 2\mathbb{P}(T \geq |t|) \\
\Leftrightarrow\ & \alpha_0 \geq 2(1 - \Psi(|t|; n-1)).
\end{aligned} \tag{43}$$

Das kleinste $\alpha_0 \in [0, 1]$ mit $\alpha_0 \geq 2\mathbb{P}(T \geq |t|)$ ist dann entsprechend $\alpha_0 = 2(1 - \Psi(|t|; n-1))$.


t = 2.26, p = 0.045
0.4
0.3
0.2
0.1
0.0
2(1 – Ψ(|t|;n – 1))
−4
−2
0
2
4
t


t = 3.81, p = 0.003
0.4
0.3
0.2
0.1
0.0
2(1 – Ψ(|t|;n – 1))
−4
−2
0
2
4
t

Definition (Powerfunktion des Einstichproben-T-Tests)

Gegeben sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. Dann ist die *Powerfunktion* des Tests gegeben durch

$$\pi : \mathbb{R} \times \mathbb{N} \to [0,1], (d,n) \mapsto \pi(d,n) := 1 - \Psi(k_{\alpha_0}; d, n-1) + \Psi(-k_{\alpha_0}; d, n-1) \tag{44}$$

Bemerkungen

- Wir betrachten hier lediglich die Testgütefunktion

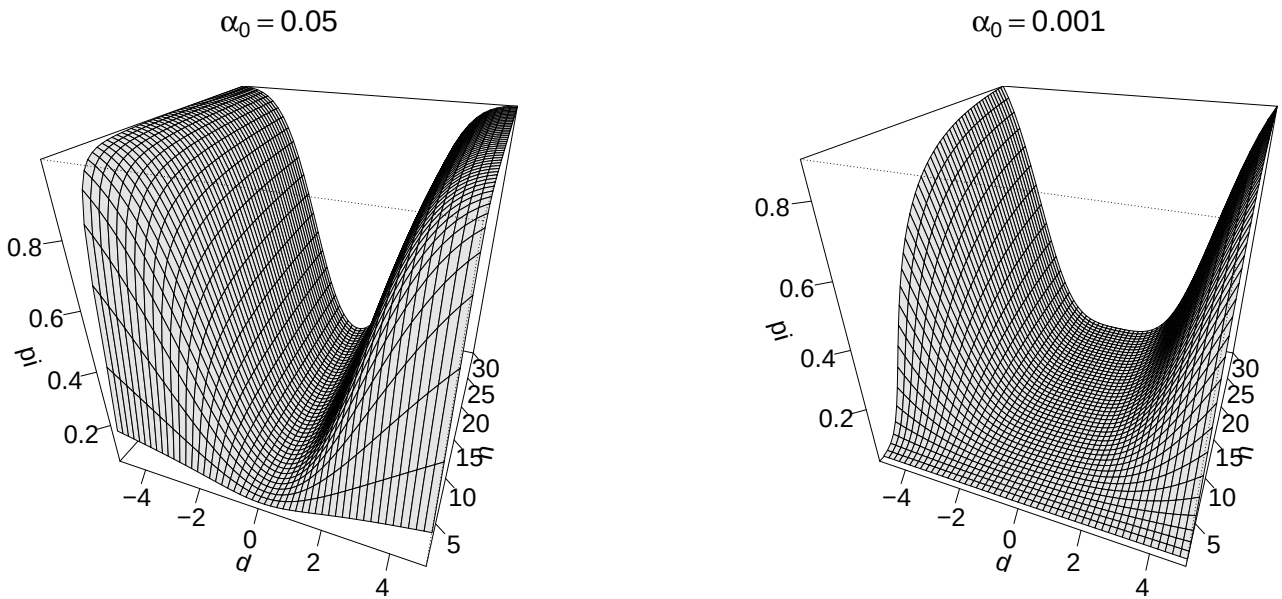
$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k_{\alpha_0}; d_\mu, n-1) + \Psi(-k_{\alpha_0}; d_\mu, n-1) \tag{45}$$

  bei kontrolliertem Testumfang, also für

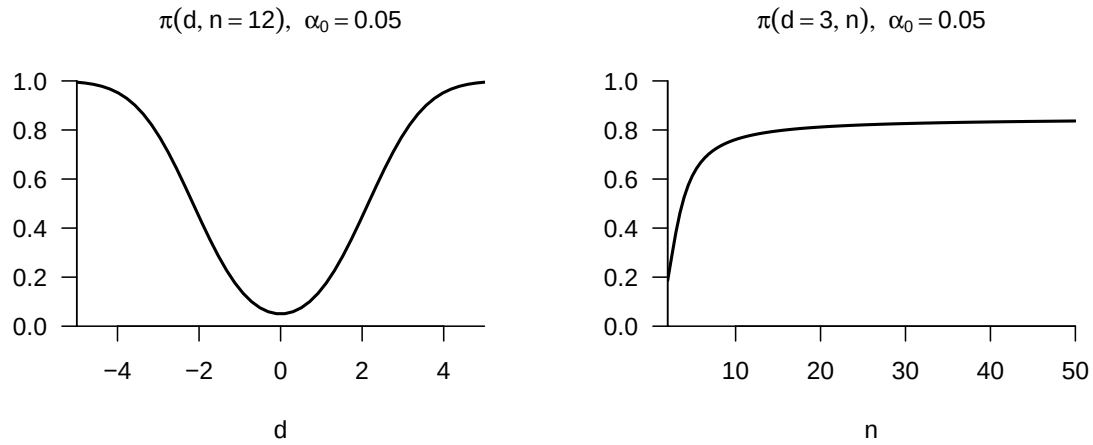
$$k_{\alpha_0} := \Psi^{-1}(1 - \alpha_0/2; n-1) \tag{46}$$

  mit festem α_0 als Funktion des Nichtzentralitätsparameters und des Stichprobenumfangs.

Powerfunktionen des Einstichproben-T-Tests



Powerfunktionsschnitte des Einstichproben-T-Tests



Man nimmt an, dass ein vorliegender univariater Datensatz $y_1, ..., y_n$ eine Realisierung des Frequentistischen Inferenzmodells $v_1, ..., v_n \sim N(\mu, \sigma^2)$ des Einstichproben-T-Tests mit wahren, aber unbekannten, Parameter μ und $\sigma^2 > 0$ ist.

Man nimmt ferner an, dass man entscheiden muss ob für einen gewählten Nullhypothesenparameter μ_0 eher die Nullhypothese $H_0 : \mu = \mu_0$ oder die Alternativhypothese $H_1 : \mu \neq \mu_0$ zutrifft.

Um den Testumfang über viele Wiederholungen dieser Testprozedur zu kontrollieren, wählt ein Signifikanzlevel α_0 und bestimmt den zugehörigen kritischen Wert k_{α_0}, so dass zum Beispiel bei einem Stichprobenumfang von $n = 12$ und der Wahl von $\alpha_0 := 0.05$ ein kritischer Wert von $k_{0.05} = 2.20$ gewählt wird.

Anhand des Stichprobenumfangs n, des Nullhypothesenparameters μ_0, des Stichprobenmittels $\bar{v}$ und der Stichprobenstandardabweichung s berechnet man sodann den Wert der Einstichproben-T-Test-Statistik durch

$$t := \sqrt{n} \frac{\bar{y} - \mu_0}{s}. \tag{47}$$

Wenn dieses für den vorliegenden Datensatz so bestimmte t größer als k_{α_0} ist oder wenn t kleiner als $-k_{\alpha_0}$ ist, lehnt man die Nullhypothese ab, andernfalls lehnt man sie nicht ab. Die oben entwickelte Theorie garantiert dann, dass man im langfristigen Mittel in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.

Weiterhin bestimmt man basierend auf dem vorliegenden Wert der Einstichproben-T-Test-Statistik den zugehörigen p-Wert durch

$$\text{p-Wert} = 2(1 - \Psi(|t|; n-1)). \tag{48}$$

Folgender **R** Code demonstriert dieses Vorgehen bei Annahme eines vorliegenden Datenvektors `y` der Länge `n`.

```
n           = length(y)                                  # Stichprobenumfang
mu_0        = 0                                          # Nullhypothesenparameter
alpha_0     = 0.05                                       # Signifikanzlevel
k_alpha_0   = qt(1-alpha_0/2,n-1)                        # kritischer Wert
Tee         = sqrt(n)*((mean(y) - mu_0)/sd(y))           # Einstichproben-T-Test-Statistik
if(abs(Tee) >= k_alpha_0){phi = 1} else {phi = 0}        # Testauswertung
p = 2*(1 - pt(Tee,n-1))                                  # p-Wert Evaluation
```

Will man im Rahmen einer Studienplanung eine Poweranalyse zur Optimierung des Stichprobenumfangs im vorliegenden Testszenario durchführen, so gilt natürlich zunächst grundsätzlich, dass mit steigendem Stichprobenumfang die Powerfunktion des Tests ansteigt. Vor dem Gesichtspunkt der Power des Tests ist ein größerer Stichprobenumfang also immer besser als ein kleinerer Stichprobenumfang.

Allerdings bleiben dabei mögliche Kosten für die Erhöhung des Stichprobenumfangs, wie zum Beispiel mögliche Risiken für die Studienteilnehmer:innen, unberücksichtigt. Weiterhin ist der Wert, den die Powerfunktion bei einem gewählten Stichprobenumfang immer von den wahren, aber unbekannten, Parameterwerten μ und σ, die in den Wert des Nichtzentralitätsparameters d einfließen, abhängig. Würde man diese Werte in einem gegebenen Anwendungskontext schon sehr genau kennen, so würde man vermutlich keine Studie durchführen wollen.

Generell wird im Rahmen der Studienplanung deshalb folgendes Vorgehen favorisiert. Zunächst entscheided man sich für ein Signifikanzlevel α_0 zur Kontrolle des Testumfangs und evaluiert die Powerfunktion. Man überlegt sich dann einen Nichtzentralitätswert d^*, den man mit einer Power von mindestens β detektieren möchte, wobei ein typischer konventioneller $\beta = 0.8$ ist. Man wertet dann die für einen Powerfunktionswert

$$\pi(d = d^*, n) = \beta \tag{49}$$

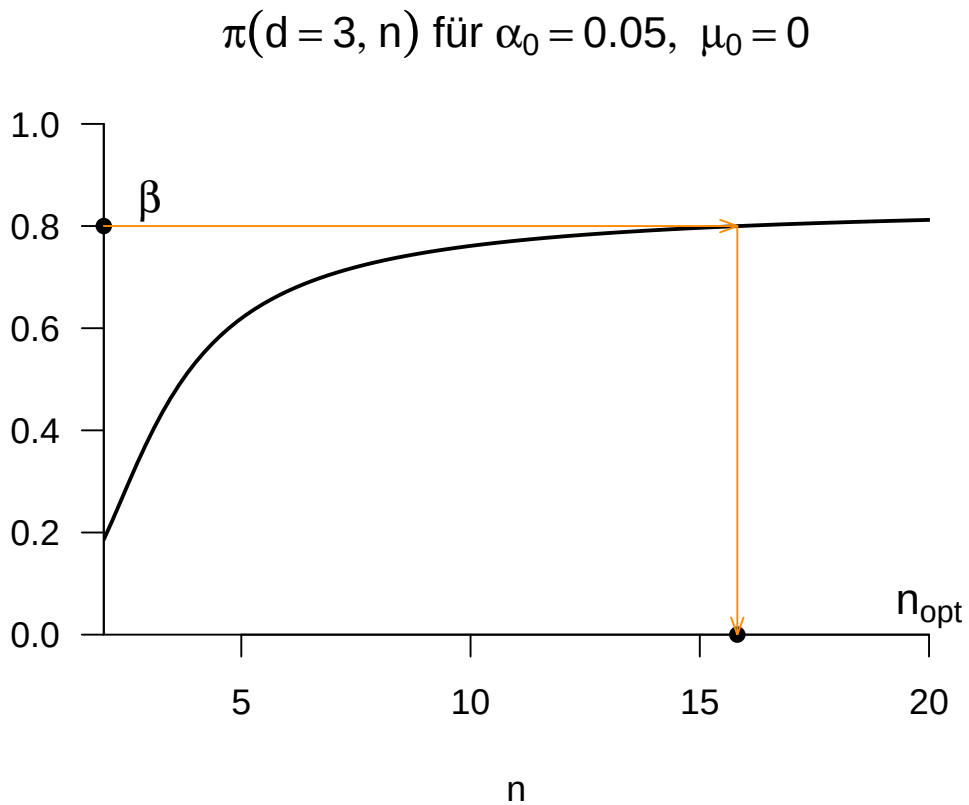
nötige Stichprobengröße aus.

Aufgrund der Monotonie der Powerfunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese im Bereich nicht-negativer Nichtzentralitätsparameter ist dann gewährleistet, dass die Power des Tests für Nichtzentralitätsparameter, die größer als d^* sind, größer oder gleich β sind. Folgender **R** Code implementiert dieses Vorgehen zur Optimierung des Stichprobenumfangs

```
# Powerfunktionsbasierte Stichprobenumfangsoptimierung
alpha_0   = 0.05                                                    # Signifikanzlevel
beta      = 0.8                                                     # gewünschter Powerfunktionswert
d_stern   = 3                                                       # fester Nichtzentralitätsparameter
n_min     = 2                                                       # minimal betrachteter Stichprobenumfang
n_max     = 20                                                      # maximal betrachteter Stichprobenumfang
n_res     = 1e2                                                     # Auflösung des Stichprobenumfangraums
n         = seq(n_min,n_max, len = n_res)                           # Stichprobenumfangraum
k_alpha_0 = qt(1-alpha_0/2, n-1)                                    # kritische Werte
pi_n      = 1-pt(k_alpha_0, n-1, d_stern)+pt(-k_alpha_0, n-1, d_stern)  # Powerfunktion
i         = 1                                                       # Indexinitialisierung
n_min     = NaN                                                     # minimales n Initialisierung
while(pi_n[i] < beta){                                              # Solange \pi(d*,n) < \beta
    n_min = n[i]                                                    # Aufnahme des minimal nötigen ns
    i     = i + 1}                                                  # und Erhöhung des Indexes
cat("Minimal nötiges n =", ceiling(n_min))                          # Ausgabe
```

```
Minimal nötiges n = 16
```

Powerfunktionsbasierte Stichprobenumfangsoptimierung



Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

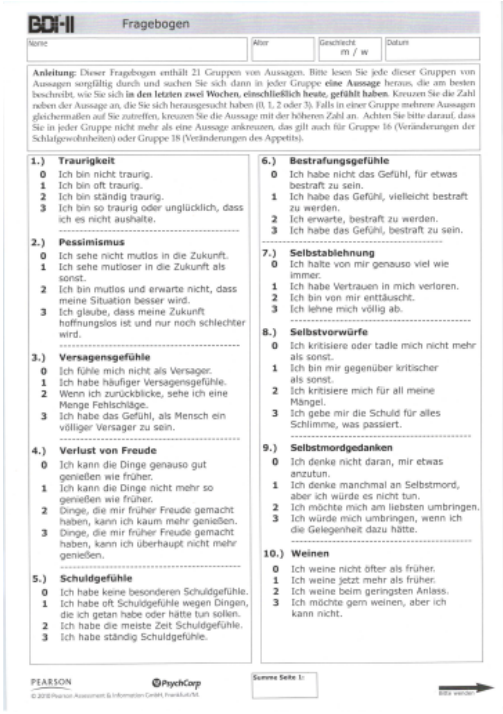
Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Pre-BDI

n = 12



BDI-II Fragebogen

Name | Alter | Geschlecht m / w | Datum

Anleitung: Dieser Fragebogen enthält 21 Gruppen von Aussagen. Bitte lesen Sie jede dieser Gruppen von Aussagen sorgfältig durch und suchen Sie sich dann in jeder Gruppe **eine Aussage** heraus, die am besten beschreibt, wie Sie sich **in den letzten zwei Wochen, einschließlich heute, gefühlt haben**. Kreuzen Sie die Zahl neben der Aussage an, die Sie sich herausgesucht haben (0, 1, 2 oder 3). Falls in einer Gruppe mehrere Aussagen gleichermaßen auf Sie zutreffen, kreuzen Sie die Aussage mit der höheren Zahl an. Achten Sie bitte darauf, dass Sie in jeder Gruppe nicht mehr als eine Aussage ankreuzen, das gilt auch für Gruppe 16 (Veränderungen der Schlafgewohnheiten) oder Gruppe 18 (Veränderungen des Appetits).

1.) Traurigkeit
0 Ich bin nicht traurig.
1 Ich bin oft traurig.
2 Ich bin ständig traurig.
3 Ich bin so traurig oder unglücklich, dass ich es nicht aushalte.

2.) Pessimismus
0 Ich sehe nicht mutlos in die Zukunft.
1 Ich sehe mutloser in die Zukunft als sonst.
2 Ich bin mutlos und erwarte nicht, dass meine Situation besser wird.
3 Ich glaube, dass meine Zukunft hoffnungslos ist und nur noch schlechter wird.

3.) Versagensgefühle
0 Ich fühle mich nicht als Versager.
1 Ich habe häufiger Versagensgefühle.
2 Wenn ich zurückblicke, sehe ich eine Menge Fehlschläge.
3 Ich habe das Gefühl, als Mensch ein völliger Versager zu sein.

4.) Verlust von Freude
0 Ich kann die Dinge genauso gut genießen wie früher.
1 Ich kann die Dinge nicht mehr so genießen wie früher.
2 Dinge, die mir früher Freude gemacht haben, kann ich kaum mehr genießen.
3 Dinge, die mir früher Freude gemacht haben, kann ich überhaupt nicht mehr genießen.

5.) Schuldgefühle
0 Ich habe keine besonderen Schuldgefühle.
1 Ich habe oft Schuldgefühle wegen Dingen, die ich getan habe oder hätte tun sollen.
2 Ich habe die meiste Zeit Schuldgefühle.
3 Ich habe ständig Schuldgefühle.

6.) Bestrafungsgefühle
0 Ich habe nicht das Gefühl, für etwas bestraft zu sein.
1 Ich habe das Gefühl, vielleicht bestraft zu werden.
2 Ich erwarte, bestraft zu werden.
3 Ich habe das Gefühl, bestraft zu sein.

7.) Selbstablehnung
0 Ich halte von mir genauso viel wie immer.
1 Ich habe Vertrauen in mich verloren.
2 Ich bin von mir enttäuscht.
3 Ich lehne mich völlig ab.

8.) Selbstvorwürfe
0 Ich kritisiere oder tadle mich nicht mehr als sonst.
1 Ich bin mir gegenüber kritischer als sonst.
2 Ich kritisiere mich für all meine Mängel.
3 Ich gebe mir die Schuld für alles Schlimme, was passiert.

9.) Selbstmordgedanken
0 Ich denke nicht daran, mir etwas anzutun.
1 Ich denke manchmal an Selbstmord, aber ich würde es nicht tun.
2 Ich möchte mich am liebsten umbringen.
3 Ich würde mich umbringen, wenn ich die Gelegenheit dazu hätte.

10.) Weinen
0 Ich weine nicht öfter als früher.
1 Ich weine jetzt mehr als früher.
2 Ich weine beim geringsten Anlass.
3 Ich möchte gern weinen, aber ich kann nicht.

PEARSON PsychCorp

Summe Seite 1:

Post-BDI

⇒ Pre-Post BDI Score Reduktion

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

dBDI
-1
3
-2
9
3
-2
4
5
5
1
9
4

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Für die Pre-Post BDI Score Reduktion v_i der $i$ten von n Patient:innen legen wir das Modell

$$v_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \tag{50}$$

zugrunde. Dabei wird die Pre-Post BDI Reduktion v_i der $i$ten Patient:in also mithilfe einer über die Gruppe von Patient:innen identischen Pre-Post BDI Score Reduktion $\mu \in \mathbb{R}$ und einer Patient:innen-spezifischen normalverteilten Pre-Post BDI Score Reduktionsabweichung ε_i erklärt

Wie gezeigt ist dieses Modell äquivalent zum Normalverteilungsmodell

$$v_1, ..., v_n \sim N(\mu, \sigma^2). \tag{51}$$

Die Standardproblemstellungen der Frequentistischen Inferenz führen dann auf folgende Fragen:

(1) Was sind sinnvolle Tipps für die wahren, aber unbekannten, Parameterwerte μ und σ^2?
(2) Wie kann im Sinne einer Intervallschätzung eine möglichst sichere Schätzung von μ gelingen?
(3) Entscheiden wir uns sinnvollerweise für die Hypothese, dass gilt $\mu \neq 0$?

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
D         = read.csv("./11_Daten/11_Hypothesentests.csv")   # Datensatzeinlesen
y         = D$dBDI                                          # Datenauswahl
n         = length(y)                                       # Stichprobenumfang
mu_hat    = mean(y)                                         # Erwartungswertparameterschätzer
delta     = 0.95                                            # Konfidenzlevel
t_delta   = qt((1+delta)/2,n-1)                             # \Psi^-1((\delta + 1)/2, n-1)
G_u       = mean(y) - (sd(y)/sqrt(n))*t_delta               # untere Konfidenzintervallgrenze
G_o       = mean(y) + (sd(y)/sqrt(n))*t_delta               # obere  Konfidenzintervallgrenze
mu_0      = 0                                               # Nullhypothesenparameter, hier \mu = \mu_0
alpha_0   = 0.05                                            # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2,n-1)                             # kritischer Wert
Tee       = sqrt(n)*((mean(y) - mu_0)/sd(y))                # T-Teststatistik
if(abs(Tee) >= k_alpha_0){phi = 1} else {phi = 0}           # Test 1_{\vert t \vert >= k_alpha_0}
p         = 2*(1 - pt(Tee,n-1))                             # p-Wert
cat("Parameterschätzwert    =", mu_hat,                     # Ausgabe
    "\n95%-Konfidenzintervall =", G_u, G_o,
    "\nSignifikanzlevel       =", alpha_0,
    "\nKritischer Wert        =", k_alpha_0,
    "\nTeststatistik          =", Tee,
    "\nTestwert               =", phi,
    "\np-Wert                 =", p)
```

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
Parameterschätzwert    = 3.166667
95%-Konfidenzintervall = 0.8074098 5.525923
Signifikanzlevel       = 0.05
Kritischer Wert        = 2.200985
Teststatistik          = 2.95423
Testwert               = 1
p-Wert                 = 0.01310986
```

Im vorliegenden Fall würde man die Nullhypothese bei einem Signifikanzlevel von $\alpha_0 = 0.05$ ablehnen. Ob die Nullhypothese zutrifft oder nicht bleibt wie der wahre, aber unbekannte, Wert von μ unbekannt. Im langfristigen Mittel würde man, basierend auf den zugrundegelegten Annahmen, die Nullhypothese nur in 5 von 100 Fällen fälschlicherweise ablehnen.

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Durchführung mit der **R** Funktion `t.test()`

```
D         = read.csv("./11_Daten/11_Hypothesentests.csv")   # Datensatzeinlesen
y         = D$dBDI                                          # Datenauswahl
t.test(y)                                                   # Einstichproben-T-Test
```

```
    One Sample t-test

data:  y
t = 2.9542, df = 11, p-value = 0.01311
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.8074098 5.5259235
sample estimates:
mean of x
 3.166667
```

Im vorliegenden Fall würde man die Nullhypothese bei einem Signifikanzlevel von $\alpha_0 = 0.05$ ablehnen. Ob die Nullhypothese zutrifft oder nicht bleibt wie der wahre, aber unbekannte, Wert von μ unbekannt. Im langfristigen Mittel würde man, basierend auf den zugrundegelegten Annahmen, die Nullhypothese nur in 5 von 100 Fällen fälschlicherweise ablehnen.

Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Theorem (Dualität von Konfidenzintervallen und Hypothesentests)

Es sei υ die Stichprobe eines Frequentistischen Inferenzmodells mit Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ. Weiterhin sei für ein $\delta \in]0,1[$ mit $[G_u(\upsilon), G_o(\upsilon)]$ ein δ-Konfidenzintervall für den wahren, aber unbekannten, Parameter $\theta \in \Theta$ definiert. Dann gilt, dass der Hypothesentest definiert durch

$$\phi_\theta : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := \begin{cases} 0, & [G_u(y), G_o(y)] \ni \theta_0 \\ 1, & [G_u(y), G_o(y)] \not\ni \theta_0 \end{cases} \tag{52}$$

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

$$\Theta_0 := \{\theta_0\} \text{ und } \Theta_1 := \Theta \setminus \{\theta_0\}. \tag{53}$$

Beweis

Aufgrund der einfachen Nullhypothese und somit $\alpha_0 = \alpha$ folgt

$$\alpha_0 = \alpha = \mathbb{P}_{\theta_0}(\phi(\upsilon) = 1) = \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \not\ni \theta_0) = 1 - \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \ni \theta_0) = 1 - \delta. \tag{54}$$

□

Bemerkung

- Mit δ-Konfidenzintervallen kann man also Hypothesentests mit Signifikanzlevel $\alpha_0 = 1 - \delta$ konstruieren.

Theorem (Dualität von Konfidenzintervall und Einstichproben-T-Test)

Gegeben sei das Normalverteilungsmodell und es sei

$$\kappa := \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right]. \tag{55}$$

das mithilfe von

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right) \tag{56}$$

in (9) Konfidenzintervalle definierte δ-Konfidenzintervall für den Erwartungswertparameter. Dann ist der Test

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := \begin{cases} 0, & \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \ni \mu_0 \\ 1, & \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \not\ni \mu_0 \end{cases} \tag{57}$$

ein Test der einfachen Nullhypothese $H_0 : \mu = \mu_0$ und der zusammengesetzten Alternativhypothese $H_1 := \mu \neq \mu_0$ mit Signifikanzlevel $\alpha_0 = 1 - \delta$.

Beweis

Es gilt

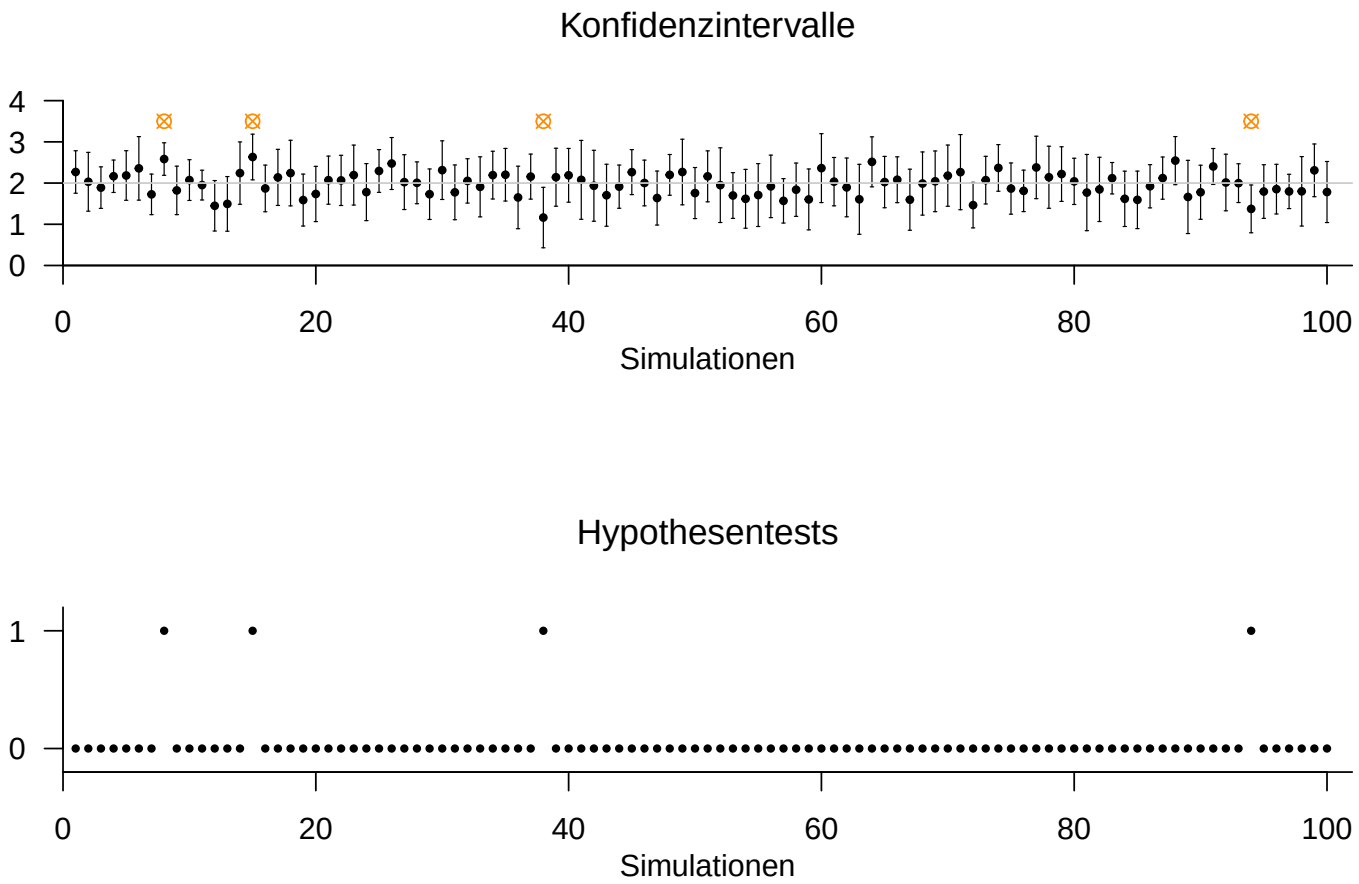
$$\mathbb{P}_{\mu_0}\left(\phi(\upsilon) = 1\right) = 1 - \mathbb{P}_{\mu_0}\left(\phi(\upsilon) = 0\right) = 1 - \mathbb{P}_{\mu_0}\left(\left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \ni \mu_0\right) = 1 - \delta. \tag{58}$$

Simulation der Dualität von Konfidenzintervall und Einstichproben-T-Test

```
n       = 12                                    # Stichprobenumfang
mu      = 2                                     # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 1                                     # wahrer, aber unbekannter, Varianzparameter
delta   = 0.95                                  # Konfidenzlevel
t_delta = qt((1+delta)/2, n-1)                  # \Psi^{-1}((\delta + 1)/2, n-1)
mu_0    = mu                                    # Nullhypothesenparameter bei Zutreffen von H_0
set.seed(1)                                     # random number generator seed
ns      = 1e2                                   # Anzahl Simulationen
y_bar   = rep(NaN,ns)                           # Stichprobenmittelarray
s       = rep(NaN,ns)                           # Stichprobenstandardabweichungarray
kappa   = matrix(rep(NaN,2*ns), ncol = 2)       # Konfidenzintervallarray
kfn     = rep(NaN,ns)                           # Überdeckungsindikatorarray
phi     = rep(NaN,ns)                           # Testarray
for(i in 1:ns){                                 # Simulationsiterationen
  y          = rnorm(n,mu_0,sqrt(sigsqr))       # Stichprobenrealisierung
  y_bar[i]   = mean(y)                          # Stichprobenmittel
  s[i]       = sd(y)                            # Stichprobenstandardabweichung
  kappa[i,1] = y_bar[i] - (s[i]/sqrt(n))*t_delta # untere Konfidenzintervallgrenze
  kappa[i,2] = y_bar[i] + (s[i]/sqrt(n))*t_delta # obere Konfidenzintervallgrenze
  if(kappa[i,1] <= mu_0 & mu_0 <= kappa[i,2]){
      kfn[i] = 1} else{kfn[i] = 0}              # Überdeckungsindikatorevaluation
  if(kappa[i,1] <= mu_0 & mu_0 <= kappa[i,2]){
      phi[i] = 0} else{phi[i] = 1}}             # Testevaluation
cat(   "Geschätztes Konfidenzniveau =", mean(kfn),   # Ausgabe
     "\nGeschätzter Testumfang      =", mean(phi))
```

```
Geschätztes Konfidenzniveau = 0.96
Geschätzter Testumfang      = 0.04
```

Simulation der Dualität von Konfidenzintervall und Einstichproben-T-Test



Testhypothesen und Tests

Testgütekriterien und Testkonstruktion

Einstichproben-T-Test

Anwendungsbeispiel

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie die grundlegende Logik Frequentistischer Hypothesentests.
2. Geben Sie die Definition der Begriffe der Testhypothesen und des Testszenario wieder.
3. Geben Sie die Definition der Begriffe der einfachen und zusammengesetzten Testhypothesen wieder.
4. Geben Sie die Definition der Begriffe einseitigen und zweiseitigen Testhypothesen wieder.
5. Geben Sie die Definition des Begriff des Tests wieder.
6. Geben Sie die Definition des Begriffs des Standardtests wieder.
7. Geben Sie die Definition des Begriffs des kritischen Bereichs wieder.
8. Geben Sie die Definition des Begriffs des Ablehungsbereichs wieder
9. Geben Sie die Definition des Begriffs des kritischen Wert-basierten Tests wieder.
10. Geben Sie die Definition der Begriffe der richtigen Testentscheidungen und der Testfehler wieder.
11. Geben Sie die Definition des Begriffs der Testgütefunktion wieder.
12. Erläutern Sie die Bedeutung der Testgütefunktion im Rahmen der Konstruktion von Hypothesentests.
13. Geben Sie die Definition der Begriffe des Level-α_0-Tests und des Signifikanzlevels α_0 wieder
14. Geben Sie die Definition des Begriffs des Testumfangs α wieder.
15. Geben Sie die Definition des Begriffs des p-Werts wieder.

Selbstkontrollfragen - Lösungen

1. Siehe Grundlegende Logik Frequentistischer Hypothesentests auf Folie 6
2. Siehe Definition (Testhypothesen und Testszenario).
3. Siehe Definition (Einfache und zusammengesetzte Testhypothesen).
4. Siehe Definition (Einseitige und zweiseitige Testhypothesen).
5. Siehe Definition (Test).
6. Siehe Definition (Standardtest).
7. Siehe Definition (Kritischer Bereich).
8. Siehe Definition (Ablehungsbereich).
9. Siehe Definition (Kritischer Wert-basierte Tests).
10. Siehe Definition (Richtigen Testentscheidungen und Testfehler).
11. Siehe Definition (Testgütefunktion).
12. Siehe Intuition zur Testkonstruktion auf Folien 20 und 21.
13. Siehe Definition (Level-α_0-Test, Signifikanzlevel α_0, Testumfang α).
14. Siehe Definition (Level-α_0-Test, Signifikanzlevel α_0, Testumfang α).
15. Siehe Definition (p-Wert).

Referenzen

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. Wiley Series in Probability and Statistics.

Ostwald, Dirk, Ludger Starke, and Ralph Hertwig. 2015. "A Normative Inference Approach for Optimal Sample Sizes in Decisions from Experience." *Frontiers in Psychology* 6 (September). https://doi.org/10.3389/fpsyg.2015.01342.

Pratt, John, Howard Raiffa, and Robert Schlaifer. 1995. *Statistical Decision Theory*. MIT Press.

Puterman, Martin. 2005. *Markov Decision Processes*. Wiley-Interscience.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World Beyond ' $p < 0.05$'." *The American Statistician* 73 (sup1): 1–19. https://doi.org/10.1080/00031305.2019.1583913.