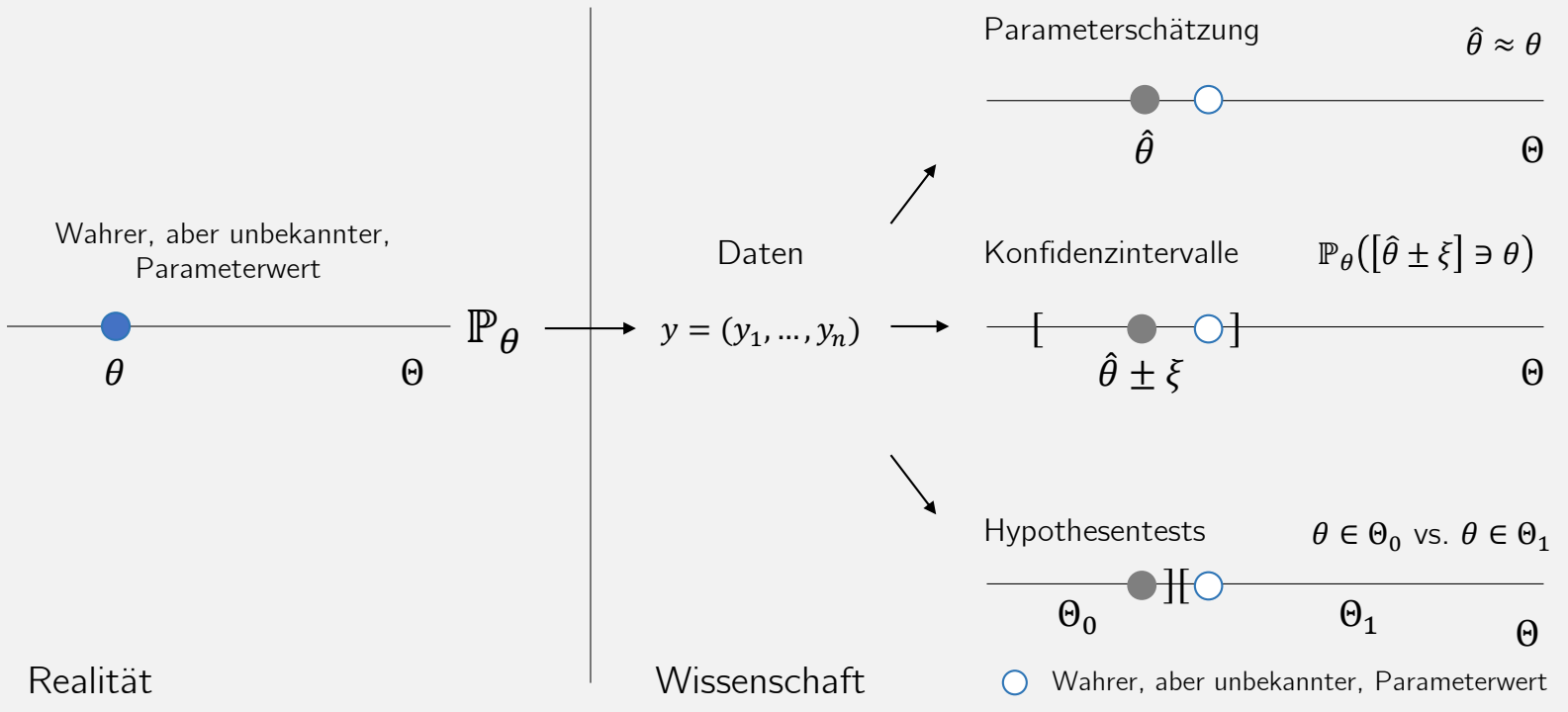
Wahrscheinlichkeitstheorie und Frequentistische Inferenz

BSc Psychologie WiSe 2021/22

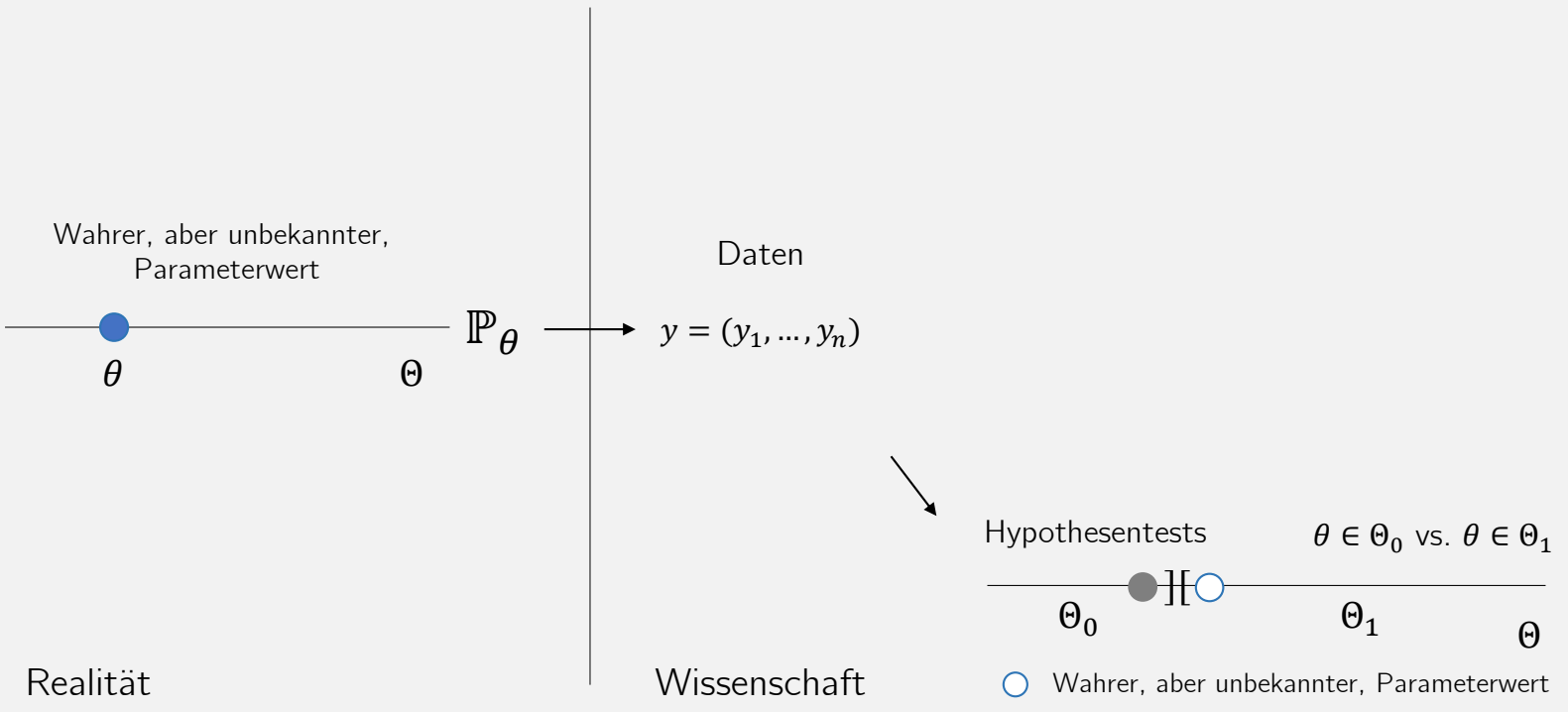
Prof. Dr. Dirk Ostwald

(12) Hypothesentests

Modell und Standardprobleme Frequentistischer Inferenz



Modell und Standardprobleme Frequentistischer Inferenz



Standardannahmen Frequentistischer Inferenz

$\mathcal{M}$ sei ein statistisches Modell mit Stichprobe $\upsilon_1, ..., \upsilon_n \sim p_\theta$. **Es wird angenommen, dass ein konkret vorliegender Datensatz $y = (y_1, ..., y_n) \in \mathbb{R}^n$ eine der möglichen Realisierungen von $\upsilon_1, ..., \upsilon_n \sim p_\theta$ ist.** Aus Frequentistischer Sicht kann man eine Studie unter gleichen Umständen unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

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

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

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

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

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

Um die Qualität statistischer Methoden zu beurteilen betrachtet die Frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme von $\upsilon_1, ..., \upsilon_n \sim p_\theta$. Was zum Beispiel ist die Verteilung der $\bar{y}_n^{(1)}$, $\bar{y}_n^{(2)}$, $\bar{y}_n^{(3)}$, $\bar{y}_n^{(4)}$, ... also die Verteilung der Zufallsvariable $\bar{\upsilon}$?

Wenn eine statistische Methode im Sinne der Frequentistischen Standardannahmen "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 $y_1, \ldots, y_n$ vorliegen und nimmt an, dass es sich dabei um die Realisation einer Stichprobe handelt, zum Beispiel von $\upsilon_1, \ldots, \upsilon_n \sim N(\mu, \sigma^2)$.

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.

Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

Selbstkontrollfragen

Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

Selbstkontrollfragen

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.

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.

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.

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 $\phi(\upsilon)$.

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.

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.

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.

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

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

Bemerkung

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

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*.

Die untenstehende Graphik gibt eine Übersicht.

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

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 $\upsilon_1, \ldots, \upsilon_n \sim p_\theta$.
- Es gilt $\mathbb{P}_\theta(\phi = 1) = \mathbb{P}_\theta(\upsilon \in K) = \mathbb{P}_\theta(\gamma \in A)$
- Für jedes $\theta \in \Theta$ liefert q_ϕ die Wahrscheinlichkeit, dass H_0 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

$$q_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = 1 \text{ für } \theta \in \Theta_1. \tag{15}$$

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

Intuition zur Testkonstruktion

Im Idealfall hätte man einen Test ϕ mit

$$q_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = 1 \text{ für } \theta \in \Theta_1. \tag{16}$$

$\Rightarrow$ 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

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0. \tag{17}$$

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

Unter allen Tests und statistischen Modellen, die Ungleichung (17) 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 (oft auch als *Signifikanzlevel* bezeichnet) und des Testumfangs α (auch als *effektives Niveau* bezeichnet).

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

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

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0 \tag{18}$$

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

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) \in [0, 1] \tag{19}$$

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

Typ I Fehlerwahrscheinlichkeit vs. Testumfang vs. Signifikanzlevel

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

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \{\theta_0\}} q_\phi(\theta) = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi = 1). \tag{20}$$

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

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi = 1). \tag{21}$$

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

$$\alpha = \max_{\theta \in \Theta_0} q_\phi(\theta) \leq \alpha_0 \tag{22}$$

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

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

Zur Wahl von Nullhypothese und Alternativhypothese

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 wissenschaftichen Anwendung ist es Standard, die falsche Konfirmation der eigenen Theorie als einen schwerwiegenderen Fehler als die falsche Ablehnung der eigenen Theorie zu werten.

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.

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), Horvath et al. (2021))

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 (wenn auch vermutlich nicht immer ganz verstanden) und sollte deshalb im Rahmen eines wissenschaftlichen Studiums wie der Psychologie intellektuell durchdrungen werden.

Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

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.

Summe Seite 1:

PEARSON PsychCorp

Post-BDI

⇒ Pre-Post BDI Score Reduktion

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

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

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \tag{23}$$

zugrunde. Dabei wird die Pre-Post BDI Reduktion υ_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

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2). \tag{24}$$

$\Rightarrow$ Wir seien nun am Testen einer statistischen Hypothese hinsichtlich μ interessiert.

$\Rightarrow$ Das betrachtete Anwendungsszenario ist dann ein Beispiel für einen **Eintstichproben-T-Test**.

Mögliche Hypothesen im Einstichproben-Test-Szenario

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

Einstichproben-T-Test

Hier betrachtete Hypothese des Einstichproben-Test-Szenarios

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

Gliederung

(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

Einstichproben-T-Test | (1) Statistischem Modell und Testhypothesen

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

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, ..., n \tag{25}$$

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. Wie unten (erneut) gezeigt, ist dieses Modell äquivalent zum Normalverteilungsmodell

$$\upsilon = \upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2), \tag{26}$$

also der Annahme unabhängig und identisch normalverteilter Stichprobenvariablen mit Erwartungswertparameter μ und Varianzparameter σ^2. Für ein μ_0 betrachten wir die einfache **Nullhypothese** und die zusammengesetzte **Alternativhypothese**

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R} \setminus \{\mu_0\}, \tag{27}$$

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.

Beweis der Äquivalenz der Modellformen

Die Äquivalenz beider Modellformen folgt direkt aus der Transformation normalverteilter Zufallsvariablen durch linear-affine Funktionen (cf. (8) Transformationen der Normalverteilung). Speziell gilt im vorliegenden Fall für $\varepsilon_i \sim N(0, \sigma^2)$, dass

$$\upsilon_i = f(\varepsilon_i) \text{ mit } f : \mathbb{R} \to \mathbb{R}, \epsilon_i \mapsto f(\epsilon_i) := \epsilon_i + \mu. \tag{28}$$

Mit den WDF Transformationstheorem bei linear-affinen Abbildungen folgt dann

$$\begin{aligned} p_{\upsilon_i}(y_i) &= \frac{1}{|1|} p_{\varepsilon_i} \left(\frac{y_i - \mu}{1} \right) \\ &= N\left(y_i - \mu; 0, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu - 0)^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu)^2 \right) \\ &= N(y_i; \mu, \sigma^2), \end{aligned} \tag{29}$$

also $\upsilon_i \sim N(\mu, \sigma^2)$.

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

$$T := \sqrt{n} \left(\frac{\bar{\upsilon} - \mu_0}{S} \right). \tag{30}$$

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:

$$\sqrt{\text{Stichprobengröße}} \left(\frac{\text{Signal}}{\text{Rauschen}} \right) = \sqrt{n} \left(\frac{\bar{\upsilon} - \mu_0}{S} \right) \tag{31}$$

- 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:
  - $T = 1 \Leftrightarrow \sqrt{n}(\bar{\upsilon} - \mu_0) = 1S$
  - $T = 2 \Leftrightarrow \sqrt{n}(\bar{\upsilon} - \mu_0) = 2S$

Definition (Nichtzentrale t-Zufallsvariable)

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

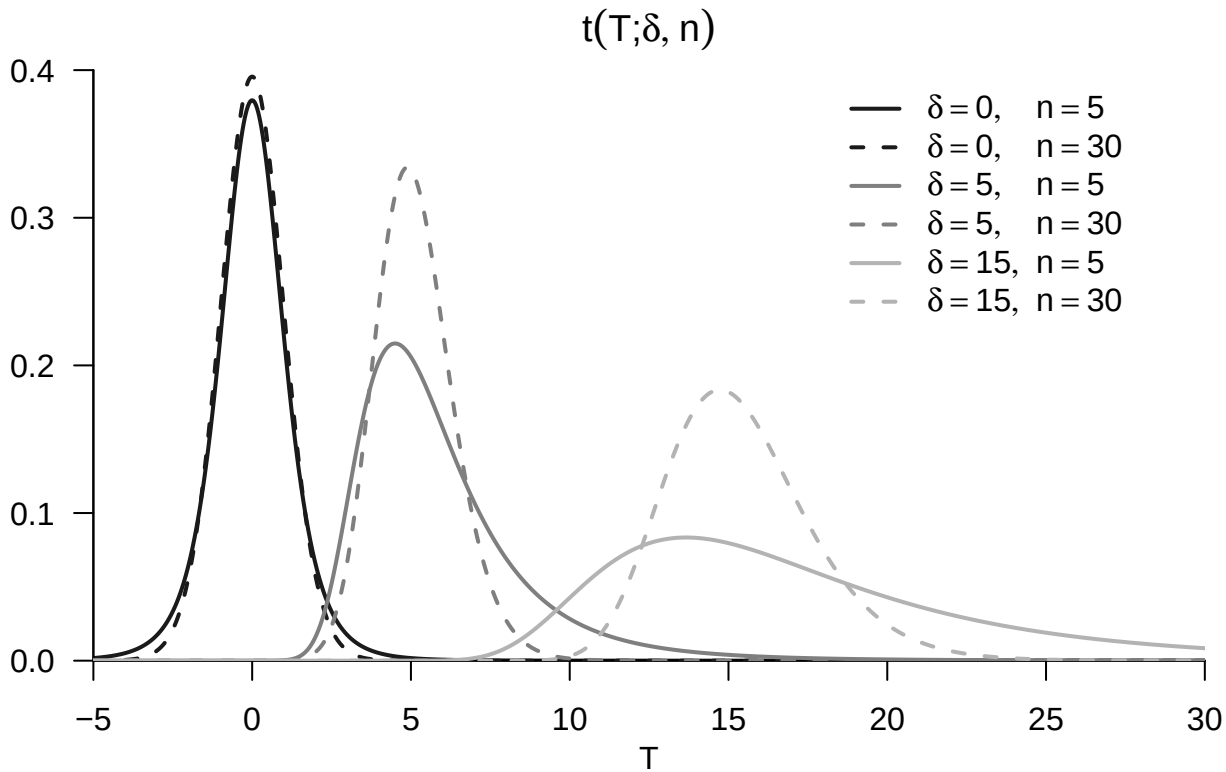
$$p : \mathbb{R} \to \mathbb{R}_{>0}, t \mapsto p(t) := \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n}{2}\right) (n\pi)^{\frac{1}{2}}} \times \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left(-\frac{\tau}{2}\right) \exp\left(-\frac{1}{2}\left(t\left(\frac{\tau}{n}\right)^{\frac{1}{2}} - \delta\right)^2\right) d\tau. \quad (32)$$

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)

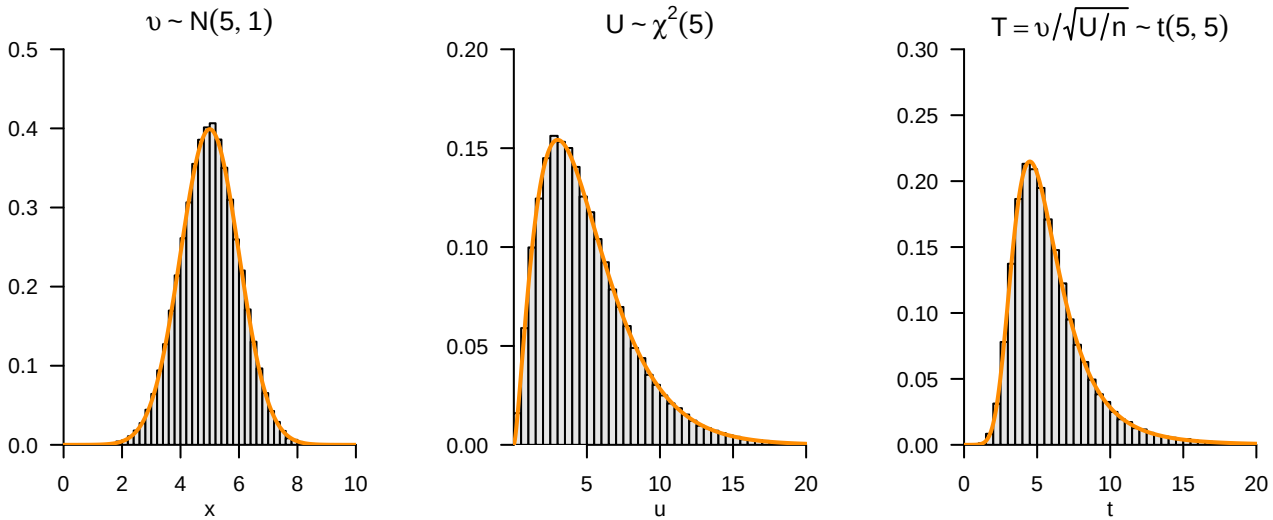
$\upsilon \sim N(\mu, 1)$ sei eine normalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2 Zufallsvariable mit Freiheitsgradparameter n, und υ und U seien unabhängige Zufallsvariablen. Dann ist die Zufallsvariable

$$T := \frac{\upsilon}{\sqrt{U/n}} \tag{33}$$

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

Bemerkung

- Wir verzichten auf einen Beweis.



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

$$T := \sqrt{n} \left(\frac{\bar{\upsilon} - \mu_0}{S} \right) \tag{34}$$

eine nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter

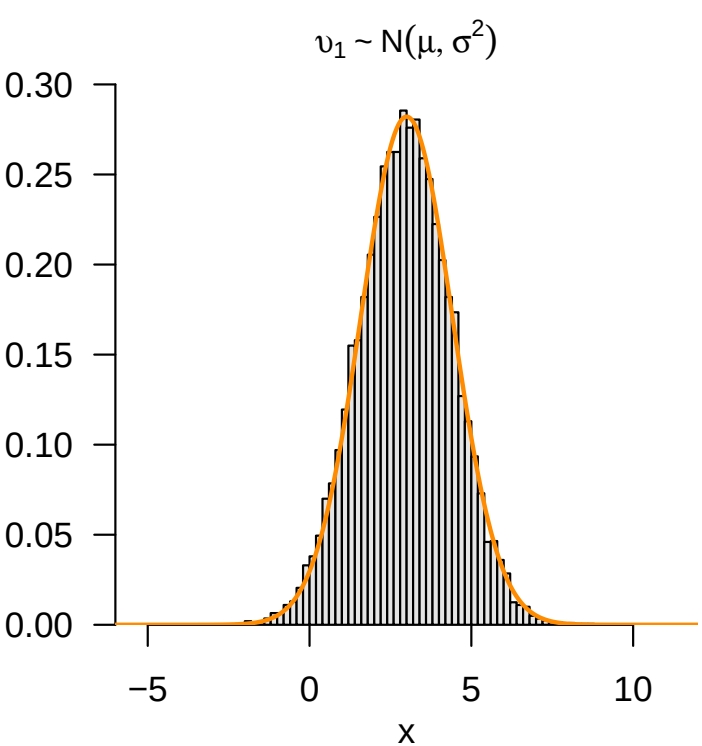
$$d = \sqrt{n} \left(\frac{\mu - \mu_0}{\sigma} \right) \tag{35}$$

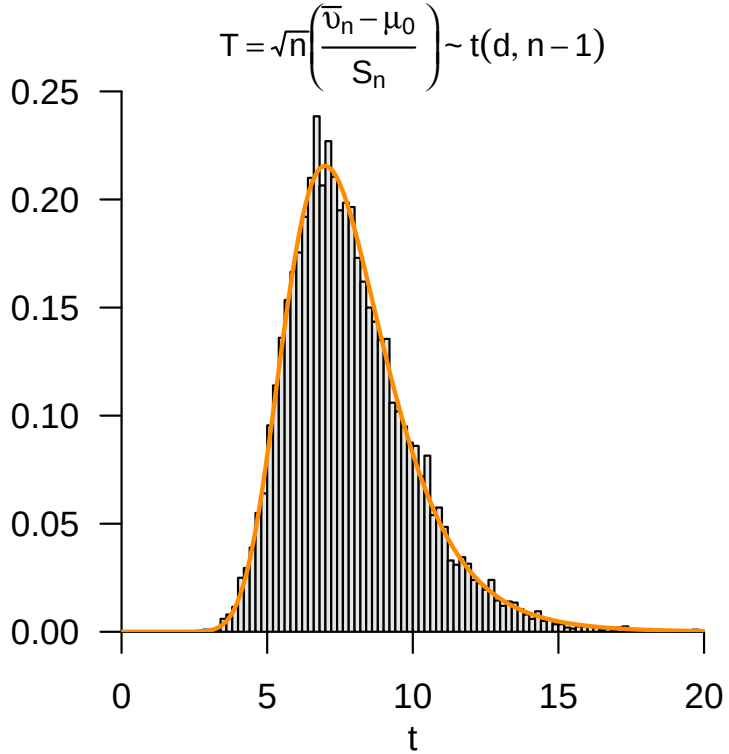
und Freiheitsgradparameter $n - 1$, es gilt also $T \sim t(d, n - 1)$

Bemerkung

- Wir verzichten auf einen Beweis.

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





Vor dem Hintergrund des statistischen Modells des Einstichproben-T-Tests betrachten wir die einfache Nullhypothese und die zusammengesetzte Alternativhypothese

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R} \setminus \{\mu_0\}, \tag{36}$$

respektive, sowie die oben definierte T-Teststatistik

$$T := \sqrt{n} \left(\frac{\bar{v} - \mu_0}{S} \right). \tag{37}$$

Wir definieren nun den zweiseitigen Einstichproben-T-Test

$$\phi(v) := 1_{\{|T| \geq k\}} = \begin{cases} 1 & |T| \geq k \\ 0 & |T| < k \end{cases}. \tag{38}$$

Theorem (Testgütefunktion)

ϕ sei der im obend definierte Test. Dann ist die Testgütefunktion von ϕ gegeben durch

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1) \tag{39}$$

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

$$d_\mu := \sqrt{n}\left(\frac{\mu - \mu_0}{\sigma}\right) \tag{40}$$

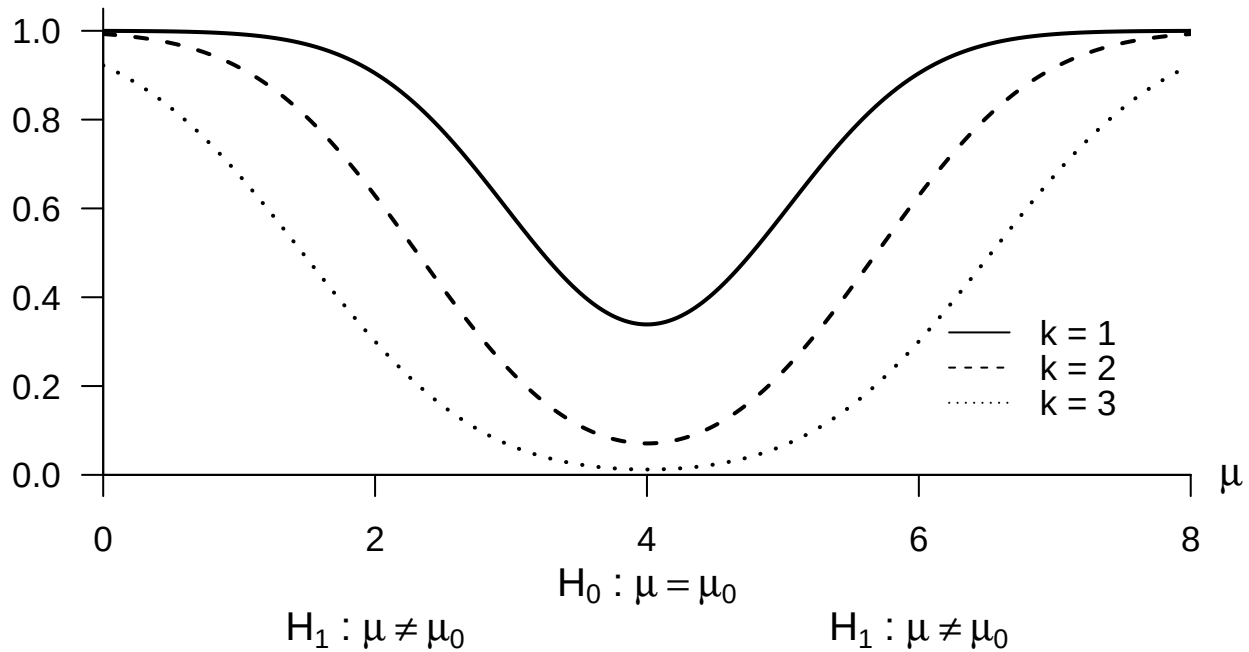
und Freiheitsgradparameter $n-1$ bezeichnet.

Bemerkungen

- Wir visualisieren die Testgütefunktion unten in Abhängigkeit von k.

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

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



Beweis

Die Testgütefunktion des betrachteten Test im vorliegenden Testszenario ist definiert als

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1). \tag{41}$$

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

$$T := \sqrt{n}\left(\frac{\bar{v} - \mu_0}{S}\right) \tag{42}$$

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

$$d_\mu := \sqrt{n}\left(\frac{\mu - \mu_0}{\sigma}\right) \tag{43}$$

verteilt ist. Der Ablehnungsbereich des zweiseitigen T-Tests ergibt sich zu

$$A =]-\infty, -k] \cup]k, \infty[. \tag{44}$$

Beweis (fortgeführt)

Mit diesem Ablehungsbereich ergibt sich dann

$$\begin{aligned} q_\phi(\mu) &= \mathbb{P}_\mu(\phi = 1) \\ &= \mathbb{P}_\mu\left(T \in]-\infty, -k] \cup]k, \infty[\right) \\ &= \mathbb{P}_\mu\left(T \in]-\infty, -k]\right) + \mathbb{P}_\mu\left(T \in [k, \infty[\right) \\ &= \mathbb{P}_\mu(T \leq -k) + \mathbb{P}_\mu(T \geq k) \\ &= \mathbb{P}_\mu(T \leq -k) + (1 - \mathbb{P}_\mu(T \leq k)) \\ &= 1 - \mathbb{P}_\mu(T \leq k) + \mathbb{P}_\mu(T \leq -k) \\ &= 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1), \end{aligned} \tag{45}$$

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

□

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

$$k_{\alpha_0} := \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right), \tag{46}$$

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

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

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k; d_{\mu_0}, n-1) + \Psi(-k; d_{\mu_0}, n-1) \\ &= 1 - \Psi(k; 0, n-1) + \Psi(-k; 0, n-1) \\ &= 1 - \Psi(k; n-1) + \Psi(-k; n-1), \end{aligned} \tag{47}$$

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.

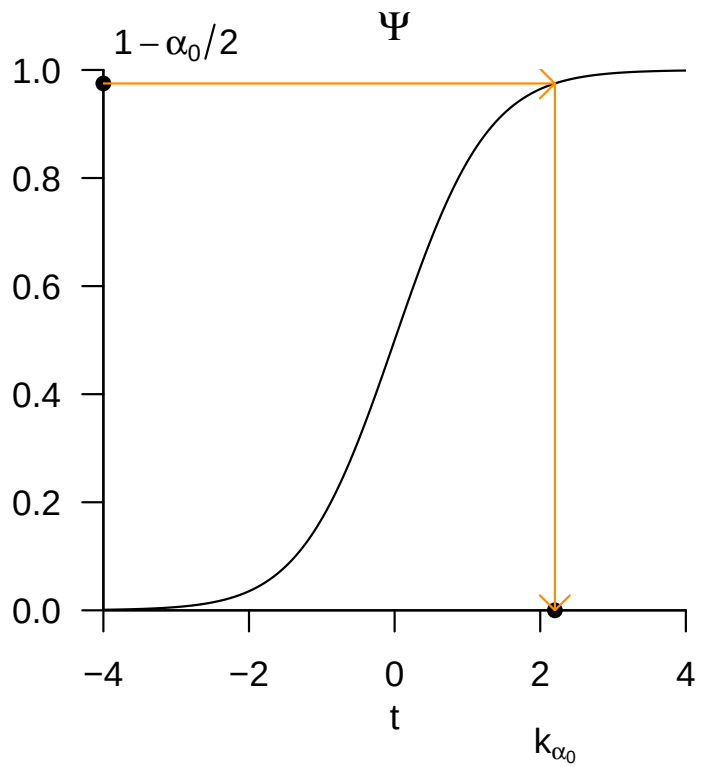
Beweis (fortgeführt)

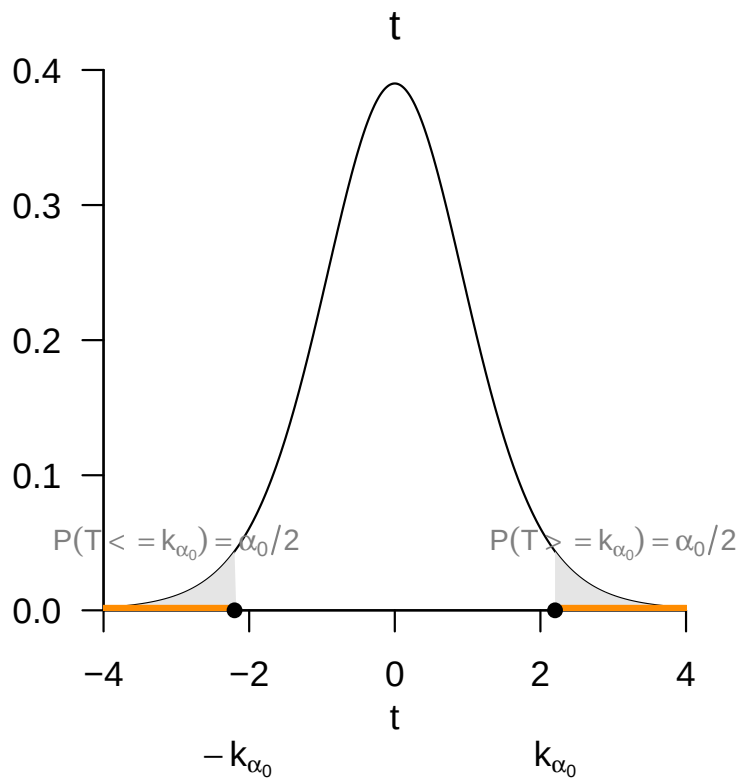
Sei nun also $k := k_{\alpha_0}$. Dann gilt

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k_{\alpha_0}; n-1) + \Psi(-k_{\alpha_0}; n-1) \\ &= 1 - \Psi(k_{\alpha_0}; n-1) + (1 - \Psi(k_{\alpha_0}; n-1) \\ &= 2(1 - \Psi(k_{\alpha_0}; n-1)) \\ &= 2\left(1 - \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}, n-1\right), n-1\right)\right) \\ &= 2\left(1 - 1 + \alpha_0/2\right) \\ &= \alpha_0, \end{aligned} \tag{48}$$

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





Praktisches Vorgehen

- 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

$$t := \sqrt{n} \left(\frac{\bar{y} - \mu_0}{s} \right) \tag{49}$$

- 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.

Simulation des praktischen Vorgehens

```
# Modellparameter
n          = 12                                          # Anzahl der Datenpunkte
mu         = 0                                           # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr     = 2                                           # wahrer, aber unbekannter, Varianzmatrixparameter

# Testparameter
mu_0       = 0                                           # H_0 Hypothesenparameter, hier \mu = \mu_0
alpha_0    = 0.05                                        # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2,n-1)                          # kritischer Wert

# Simulation der Testumfangkontrolle
set.seed(1)                                              # Random number generator seed
nsim       = 1e6                                         # 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)                  # T-Teststatistik
    if(abs(Tee) > k_alpha_0){                            # Test 1_{|t| >= k_alpha_0}
        phi[j] = 1                                       # Ablehnen von H_0
    } else {
        phi[j] = 0                                       # Nicht Ablehnen von H_0
    }
}

# Ausgabe
cat("Kritischer Wert              =", k_alpha_0,
    "\nGeschätzter Testumfang alpha =", mean(phi))
```

```
> Kritischer Wert              = 2.2
> Geschätzter Testumfang alpha = 0.0498
```

Wir betrachten die Testgütefunktion

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k_{\alpha_0}; d_\mu, n-1) + \Psi(-k_{\alpha_0}; d_\mu, n-1) \tag{50}$$

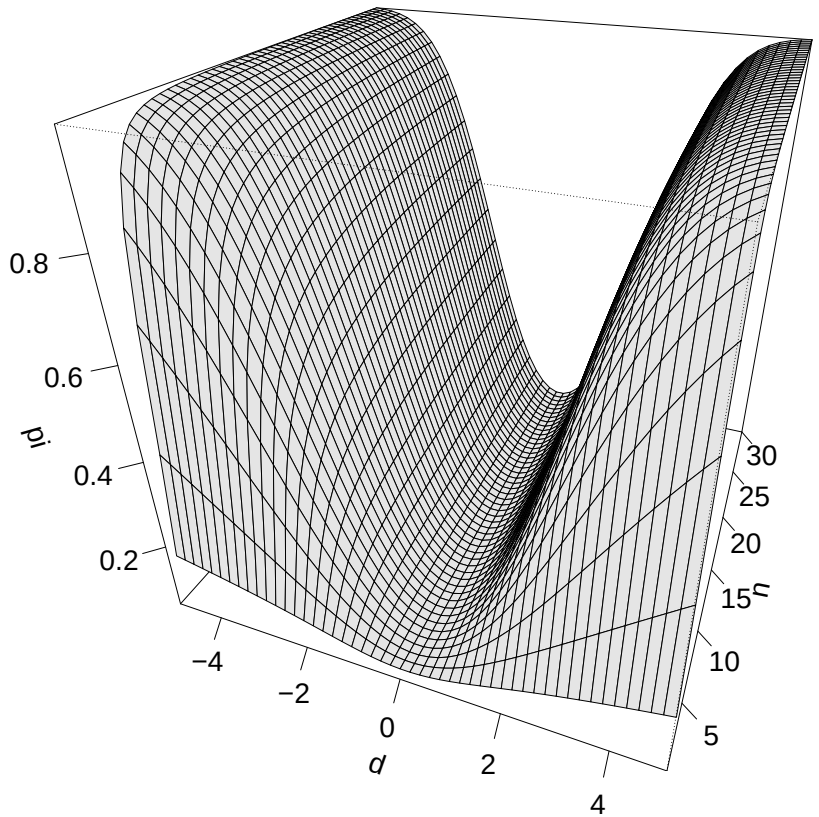
bei kontrolliertem Testumfang, also für $k_{\alpha_0} := \Psi^{-1}(1-\alpha_0/2; n-1)$ mit festem α_0 als Funktion des Nichtzentralitätsparameters und des Stichprobenumfangs. Namentlich hängt hier k_{α_0} auch von n ab.

Es ergibt sich die bivariate reellwertige Funktion

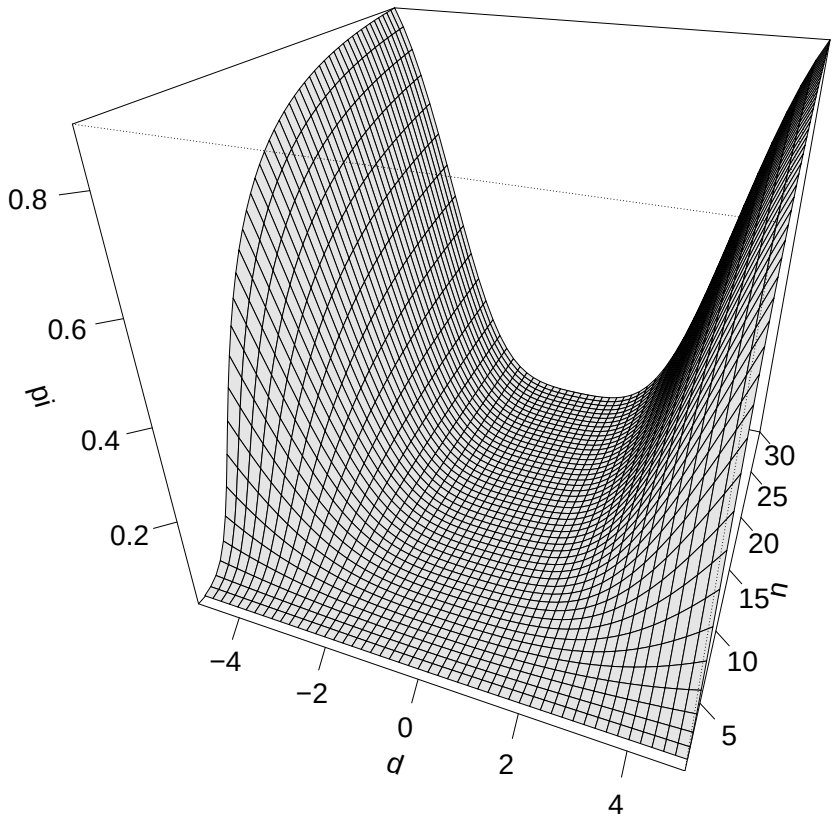
$$\pi : \mathbb{R} \times \mathbb{N} \to [0,1], (d,n) \mapsto \pi(d,n) := 1 - \Psi(k_{\alpha_0}; d, n-1) + \Psi(-k_{\alpha_0}; d, n-1) \tag{51}$$

Bei festgelegten α_0 hängt die Powerfunktion des zweiseitigen T-Tests mit einfacher Nullhypothese also vom unbekannten Wert d und von der Stichprobengröße n ab. Wir visualisieren diese Abhängigkeiten untenstehend.

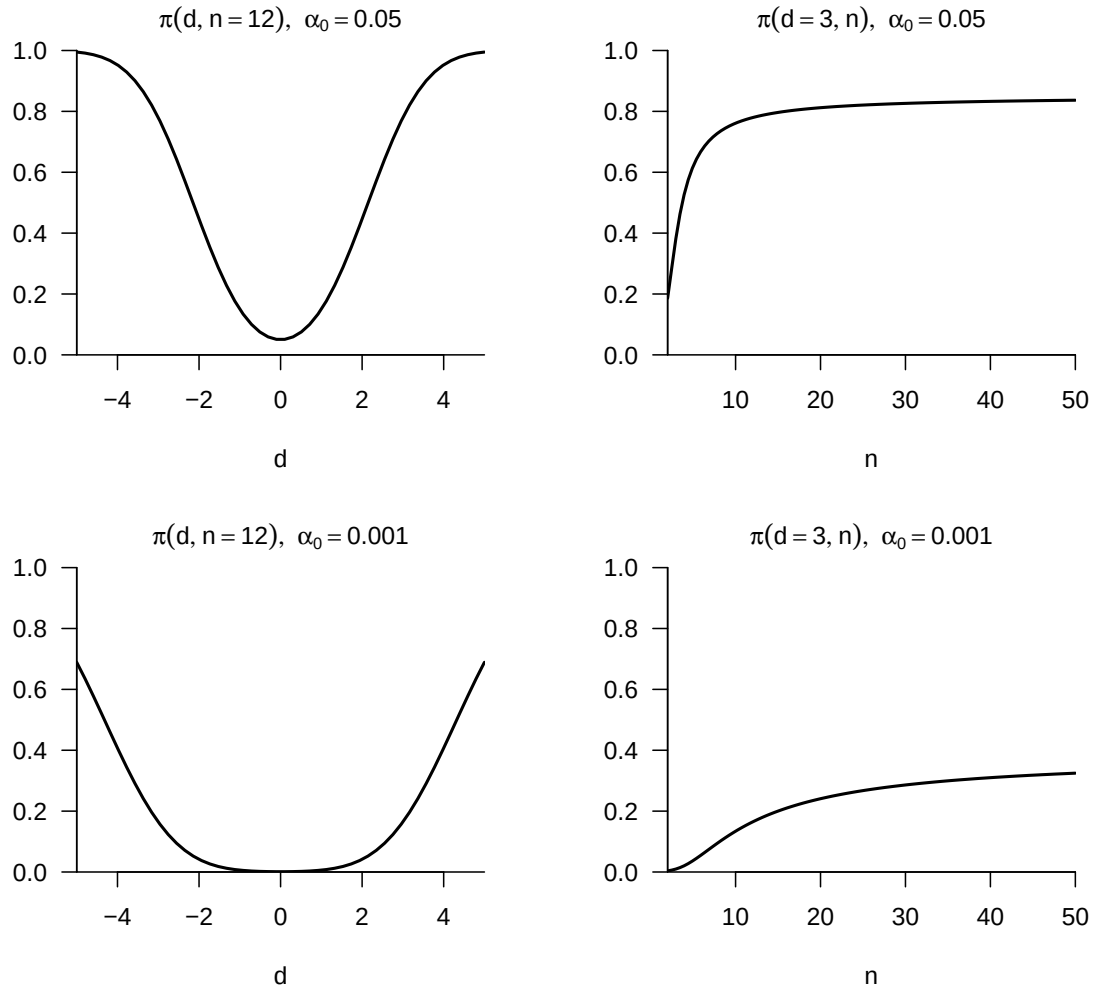
Powerfunktion für $\alpha_0 = 0.05$



Powerfunktion für $\alpha_0 = 0.001$



Powerfunktionen für $\mu_0 = 0$



Praktisches Vorgehen

Mit größerem n steigt die Powerfunktion des Tests an

- Ein großer Stichprobenumfang ist besser als ein kleiner Stichprobenumfang.
- Kosten für die Erhöhung des Stichprobenumfangs werden aber nicht berücksichtigt.

$\Rightarrow$ Die Theorie statistischer Hypothesentests ist nicht besonders lebensnah.

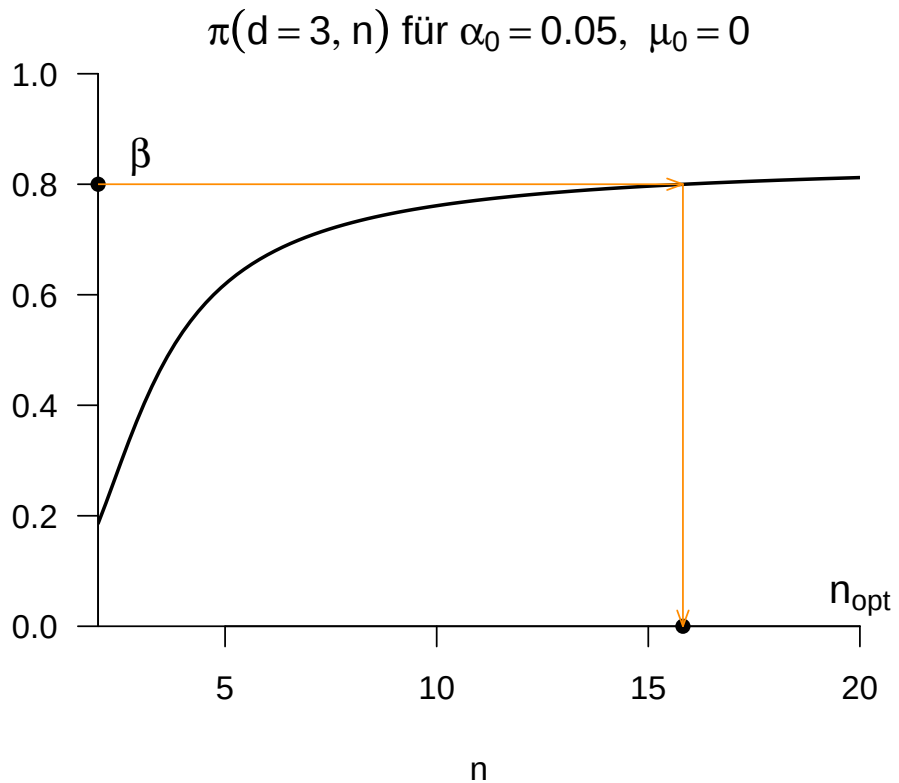
Die Powerfunktion hängt vom wahren, aber unbekannten, Parameterwert $d = \sqrt{n}(\mu - \mu_0)/\sigma$ ab.

$\Rightarrow$ Wenn man d schon kennen würde, würde man den Test nicht durchführen.

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.

Praktisches Vorgehen



Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

Selbstkontrollfragen

Motivation

- 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.

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,
$$\alpha_0 \geq 2\mathbb{P}(T \geq |t|). \tag{52}$$
- 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
$$\text{p-Wert} = 2\mathbb{P}(T \geq |t|) = 2(1 - \Psi(|t|; n-1)). \tag{53}$$
- 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.

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

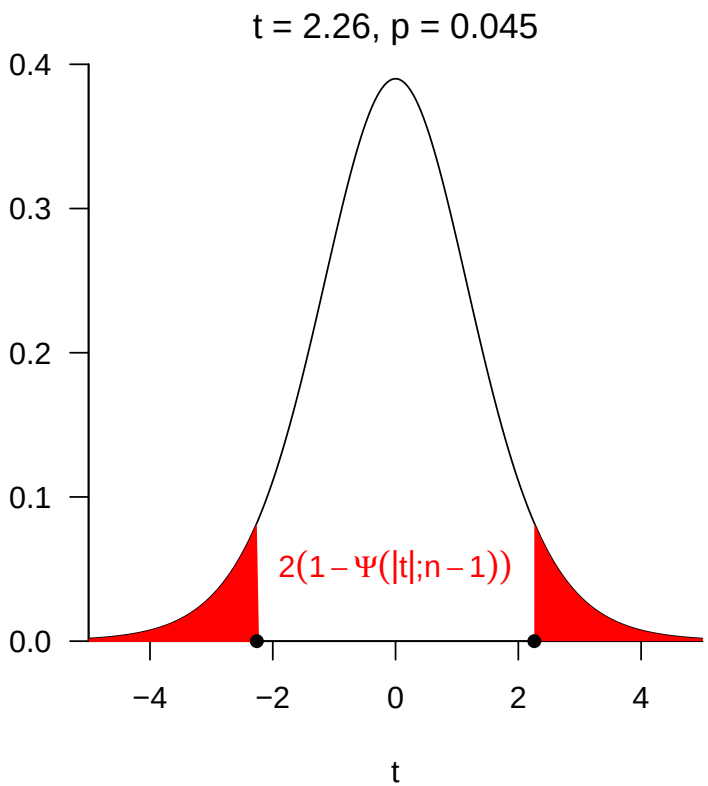
- Es bleibt zu zeigen, dass gilt

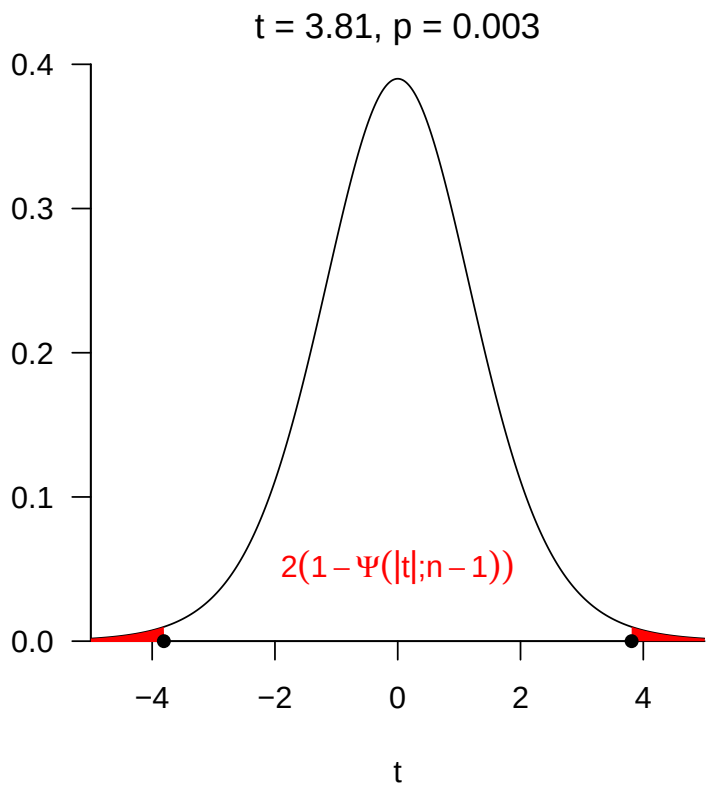
$$|t| \geq \Psi^{-1}(1 - \frac{\alpha_0}{2}; n-1) \Leftrightarrow \alpha_0 \geq 2\mathbb{P}(T \geq |t|) \tag{54}$$

- Dies aber folgt aus

$$\begin{aligned}
& |t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right) \\
\Leftrightarrow\ & \Psi(|t|; n-1) \geq \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right); n-1\right) \\
\Leftrightarrow\ & \Psi(|t|; n-1) \geq 1 - \frac{\alpha_0}{2} \\
\Leftrightarrow\ & \mathbb{P}(T \leq |t|) \geq 1 - \frac{\alpha_0}{2} \\
\Leftrightarrow\ & \frac{\alpha_0}{2} \geq 1 - \mathbb{P}(T \leq |t|) \\
\Leftrightarrow\ & \frac{\alpha_0}{2} \geq \mathbb{P}(T \geq |t|) \\
\Leftrightarrow\ & \alpha_0 \geq 2\mathbb{P}(T \geq |t|).
\end{aligned} \tag{55}$$

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





p-Werte

Bemerkungen

- p-Werte spiegeln 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, ihre uninformierte Benutzung ist aber auch sehr umstritten.

→ The American Statistician (2019) Statistical Inference in the 21st Century: A World Beyond p < 0.05

- 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 Möglichkeit ein Signal-zu-Rauschen Verhältnis zu quantifizieren.
- p-Werte sind eine Möglichkeit Unsicherheit zu quantifizieren.

Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

Selbstkontrollfragen

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

$$\phi_\theta : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := \begin{cases} 0, & [G_u(y), G_o(y)] \ni \theta_0 \\ 1, & [G_u(y), G_o(y)] \not\ni \theta_0 \end{cases} \tag{56}$$

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

$$\Theta_0 := \{\theta_0\} \text{ und } \Theta_1 := \Theta \setminus \{\theta_0\}. \tag{57}$$

Beweis

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

$$\alpha_0 = \alpha = \mathbb{P}_{\theta_0}(\phi(\upsilon) = 1) = \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \not\ni \theta) = 1 - \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \ni \theta) = 1 - \delta. \tag{58}$$

Bemerkung

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

Beispiel (Konstruktion eines Hypothesentests aus einem Konfidenzintervall)

Wir haben bereits gesehen, dass für eine Stichprobe $\upsilon = \upsilon_1, \ldots, \upsilon_n \sim N(\mu, \sigma^2)$, $\delta \in]0, 1[$ und

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right) \tag{59}$$

ein δ-Konfidenzintervall durch

$$\kappa := \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right]. \tag{60}$$

definiert ist. Mit der Dualität von Konfidenzintervallen und Hypothesentests können wir also folgenden Test für die Hypothesen $\Theta_0 = \{\mu_0\}$ und $\Theta_1 = \mathbb{R} \setminus \mu_0$ definieren:

$$\phi : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi(y) := \begin{cases} 0, & \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \ni \mu_0 \\ 1, & \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \not\ni \mu_0 \end{cases} \tag{61}$$

Dann gilt

$$\begin{aligned} \mathbb{P}_{\mu_0}\left(\phi(\upsilon) = 1\right) &= 1 - \mathbb{P}_{\mu_0}\left(\phi(\upsilon) = 0\right) \\ &= 1 - \mathbb{P}_{\mu_0}\left(\left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \ni \mu_0\right) \\ &= 1 - \delta. \end{aligned} \tag{62}$$

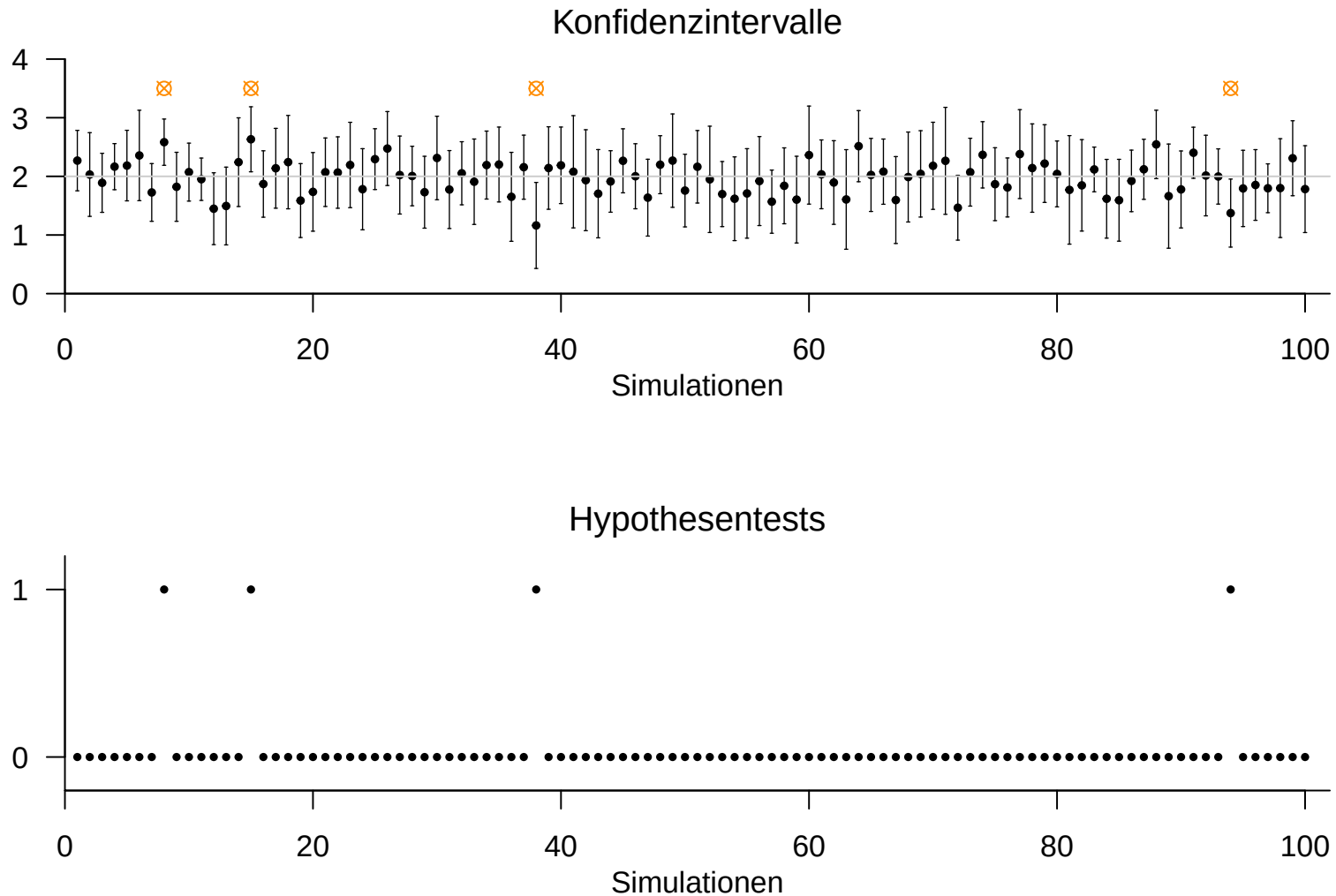
und wir haben gezeigt, dass ϕ ein Test vom Signifikanzlevel $\alpha_0 = 1 - \delta$ ist.

Simulation der Dualität von Konfidenzintervallen und Hypothesentests

```
# Modellformulierung
n       = 12                                        # Stichprobengröße
mu      = 2                                         # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 1                                         # wahrer, aber unbekannter, Varianzparameter

# Konfidenzintervallparameter und Testparameter
delta   = 0.95                                      # Konfidenzbedingung
t_delta = qt((1+delta)/2, n-1)                      # \Psi^{-1}((\delta + 1)/2, n-1)
mu_0    = mu                                        # Nullhypothesenparameter

# Simulationen
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
  # Stichprobenrealisation und Konfidezintervallevaluation
  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 KI Grenze
  kappa[i,2] = y_bar[i] + (s[i]/sqrt(n))*t_delta    # obere KI Grenze

  # Überdeckungs- und Testevaluation
  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

# Ausgabe
cat(   "Geschätztes Konfidenzniveau =", mean(kfn),
     "\nGeschätzter Testumfang      =", mean(phi))
```

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

Simulation der Dualität von Konfidenzintervallen und Hypothesentests



Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

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

Wir legen für die BDI Score Reduktion υ_i der $i$ten von n Patient:innen das Modell

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \tag{63}$$

zugrunde.

Wir erklären die BDI Reduktion υ_i der $i$ten Patient:in also mithilfe einer über die Gruppe von Patient:innen identischen BDI Score Reduktion μ als Effekt der Therapieintervention und einer Patient:innen-spezifischen normalverteilten BDI Score Reduktionsabweichung ε_i, die sich aus sehr vielen additiven Prozessen zusammensetzt, für die wir also eine Normalverteilungsannahme treffen, und deren Varianz wir mit σ^2 parameterisieren.

Vor dem Hintergrund dieses Modells evaluieren wir die drei Standardprobleme der Frequentistischen Inferenz für den Effekt der Therapieintervention μ:

(1) Was ist unserer Schätzung für den Effekt μ der Therapie auf die BDI Score Reduktion?
(2) Welches 95%-Konfidenzintervall ist mit dieser Schätzung von μ assoziiert?
(3) Entscheiden wir uns sinnvoller Weise für die Nullhypothese $\mu = 0$?

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
fname = file.path(getwd(), "12_Hypothesentests.csv")
D     = read.table(fname, sep = ",", header = T)
```

i	BDI.Reduktion
1	-1
2	3
3	-2
4	9
5	3
6	-2
7	4
8	5
9	5
10	1
11	9
12	4

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
# Einlesen und Auswahl der Daten
fname     = file.path(getwd(), "12_Hypothesentests.csv")  # Dateiname
D         = read.table(fname, sep = ",", header = T)      # Dataframe
y         = D$BDI.Reduktion                               # Datenrealisation
n         = length(y)                                     # Anzahl Datenpunkte

# Parameterschätzung
y_bar      = mean(y)                                      # Stichprobenmittel
mu_hat     = y_bar                                        # Unverzerrte Maximum-Likelihood Schätzung

# Konfidenzintervallevaluation
delta    = 0.95                                           # Konfidenzlevel
t_delta  = qt((1+delta)/2,n-1)                            # \Psi^-1((\delta + 1)/2, n-1)
G_u      = y_bar - (sd(y)/sqrt(n))*t_delta                # untere KI Grenze
G_o      = y_bar + (sd(y)/sqrt(n))*t_delta                # obere  KI Grenze

# Testevaluation
mu_0      = 0                                             # H_0 Hypothesenparameter, hier \mu = \mu_0
alpha_0   = 0.05                                          # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2,n-1)                           # kritischer Wert
Tee       = sqrt(n)*((y_bar - mu_0)/sd(y))                # T-Teststatistik
if(abs(Tee) > k_alpha_0){phi = 1} else {phi = 0}          # Test 1_{|t| >= k_alpha_0}

# p-Wert Evaluation
p         = 2*(1 - pt(Tee,n-1))                           # p-Wert
```

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
# Ausgabe
cat("Parameterschätzwert    =", mu_hat,
    "\n95%-Konfidenzintervall =", G_u, G_o,
    "\nSignifikanzlevel       =", alpha_0,
    "\nKritischer Wert         =", k_alpha_0,
    "\nTeststatistik           =", Tee,
    "\nTestwert                =", phi,
    "\np-Wert                  =", p)
```

```
> Parameterschätzwert    = 3.17
> 95%-Konfidenzintervall = 0.807 5.53
> Signifikanzlevel       = 0.05
> Kritischer Wert        = 2.2
> Teststatistik          = 2.95
> Testwert               = 1
> p-Wert                 = 0.0131
```

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Frequentistische Inferenz mit R's `t.test()` Funktion

```
t.test(y)          # Anwendung der in Einheiten (1) bis (12) entwickelten Theorie
```

```
>
>   One Sample t-test
>
> data:  y
> t = 3, df = 11, p-value = 0.01
> alternative hypothesis: true mean is not equal to 0
> 95 percent confidence interval:
>  0.807 5.526
> sample estimates:
> mean of x
>      3.17
```

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

(1) Was ist unserer Schätzung für den wahren, aber unbekannten, Effekt μ der Therapie?

- Unsere Schätzung für den Therapieeffekt ist eine BDI Reduktion von $\hat{\mu} = 3.17$.

(2) Welches 95%-Konfidenzintervall ist mit dieser Schätzung von μ assoziiert?

- Das 95%-Konfidenzintervall für den Therapieeffekt ist eine BDI Reduktion $[0.81, 5.53]$.

(3) Entscheiden wir uns sinnvoller Weise für die Nullhypothese $\mu = 0$?

- Bei einem Signifikanzlevel von $\alpha_0 = 0.05$ lehnen wir die Nullhypothese $\mu = 0$ ab $(p = 0.01)$.

Aus datenwisschenschaftlicher Sicht sind Ergebnis und Unsicherheit dieses Szenarios im Frequentistischen Sinne nun quantifiziert. Die Interpretation des Ergebnisses inklusive des Mehrwerts der Therapie bei einer geschätzten BDI Score Reduktion von ≈ 3 obliegt der Klinischen Psychologie.

Grundlegende Definitionen

Einstichproben-T-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Anwendungsbeispiel

Selbstkontrollfragen

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.

References

Horvath, Lilla, Stanley Colcombe, Michael Milham, Shruti Ray, Philipp Schwartenbeck, and Dirk Ostwald. 2021. "Human Belief State-Based Exploration and Exploitation in an Information-Selective Symmetric Reversal Bandit Task." *Computational Brain & Behavior*, August. https://doi.org/10.1007/s42113-021-00112-3.

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.