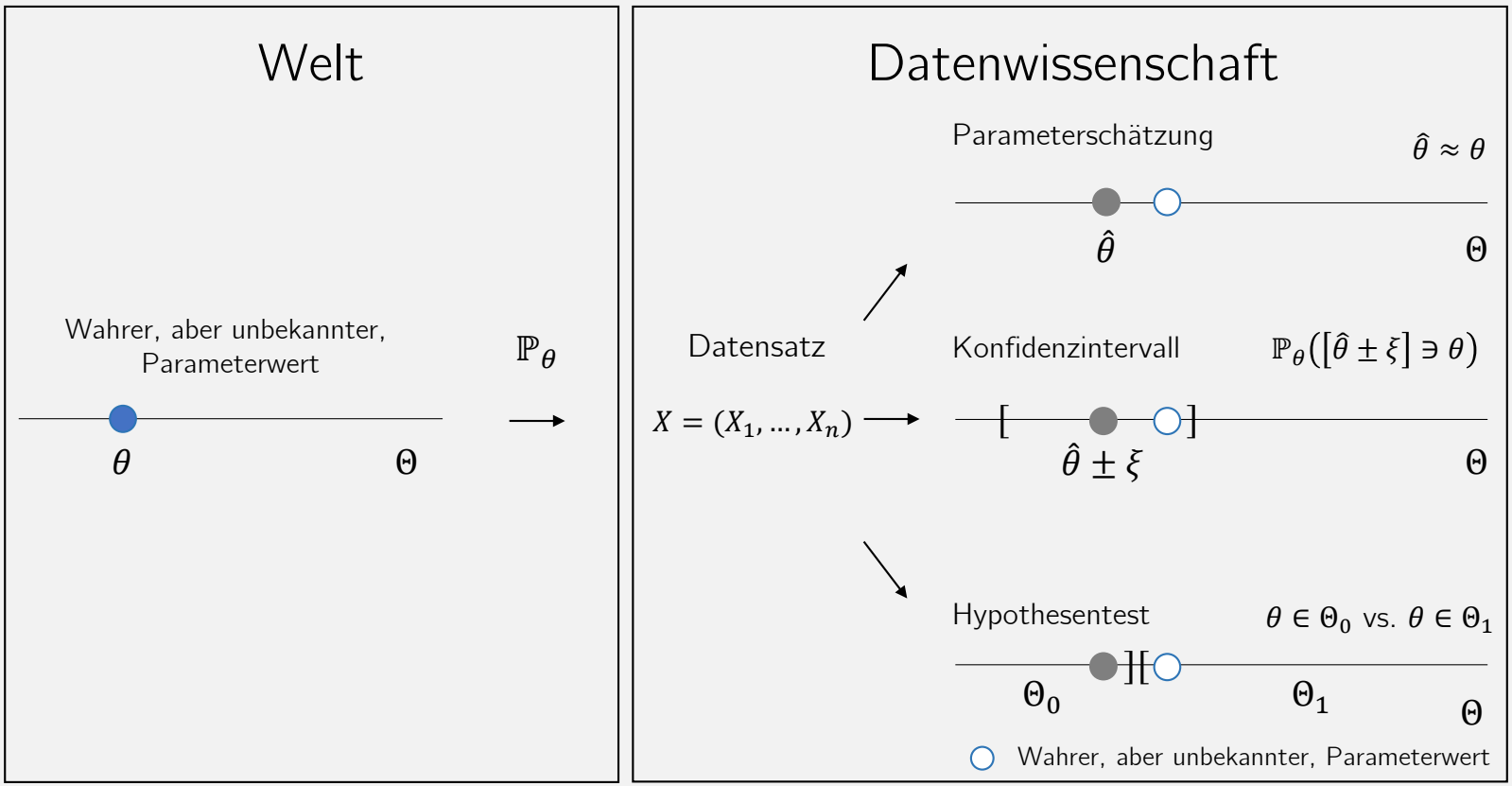
Wahrscheinlichkeitstheorie und Frequentistische Inferenz

BSc Psychologie WiSe 2021/22

Prof. Dr. Dirk Ostwald

(12) Hypothesentests

Modell und Standardprobleme der Frequentistischen Inferenz



Standardprobleme Frequentistischer Inferenz

(1) **Parameterschätzung**

Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für den wahren, aber unbekannten, Parameterwert (oder eine Funktion dessen) abzugeben, typischerweise basierend auf der Beobachtung einer Realisierung von $X_1, ..., X_n \sim p_\theta$.

(2) **Konfidenzintervalle**

Das Ziel der Bestimmung von Konfidenzintervallen ist es, basierend auf der Verteilung möglicher Parameterschätzwerte eine quantitative Aussage über die mit dem Schätzwert assoziierte Unsicherheit zu treffen.

(3) **Hypothesentests**

Das Ziel der Auswertung von Hypothesentests ist es, basierend auf der angenommenen Verteilung der Beobachtungen $X_1, ..., X_n$ in einer möglichst sinnvollen Form zu entscheiden, ob der wahre, aber unbekannte Parameterwert, in einer von zwei sich gegenseitig ausschließenden Untermengen des Parameterraumes, welche man als Hypothesen bezeichnet, liegt.

Standardannahmen Frequentistischer Inferenz

$\mathcal{M}$ sei ein statistisches Modell mit $X_1, \ldots, X_n \sim \mathbb{P}_\theta$. Es wird angenommen, dass ein vorliegender Datensatz eine der möglichen Realisierungen von $X_1, \ldots, X_n \sim \mathbb{P}_\theta$ ist. Aus frequentistischer Sicht kann man die Erhebung von Datensätzen unendlich oft wiederholen und zu jedem Datensatz Statistiken auswerten.

Datensatz (1) : $x^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right)$, Statistik (1): $S : \mathbb{R}^n \to \Sigma, x^{(1)} \mapsto S\left(x^{(1)}\right)$

Datensatz (2) : $x^{(2)} = \left(x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}\right)$, Statistik (2): $S : \mathbb{R}^n \to \Sigma, x^{(2)} \mapsto S\left(x^{(2)}\right)$

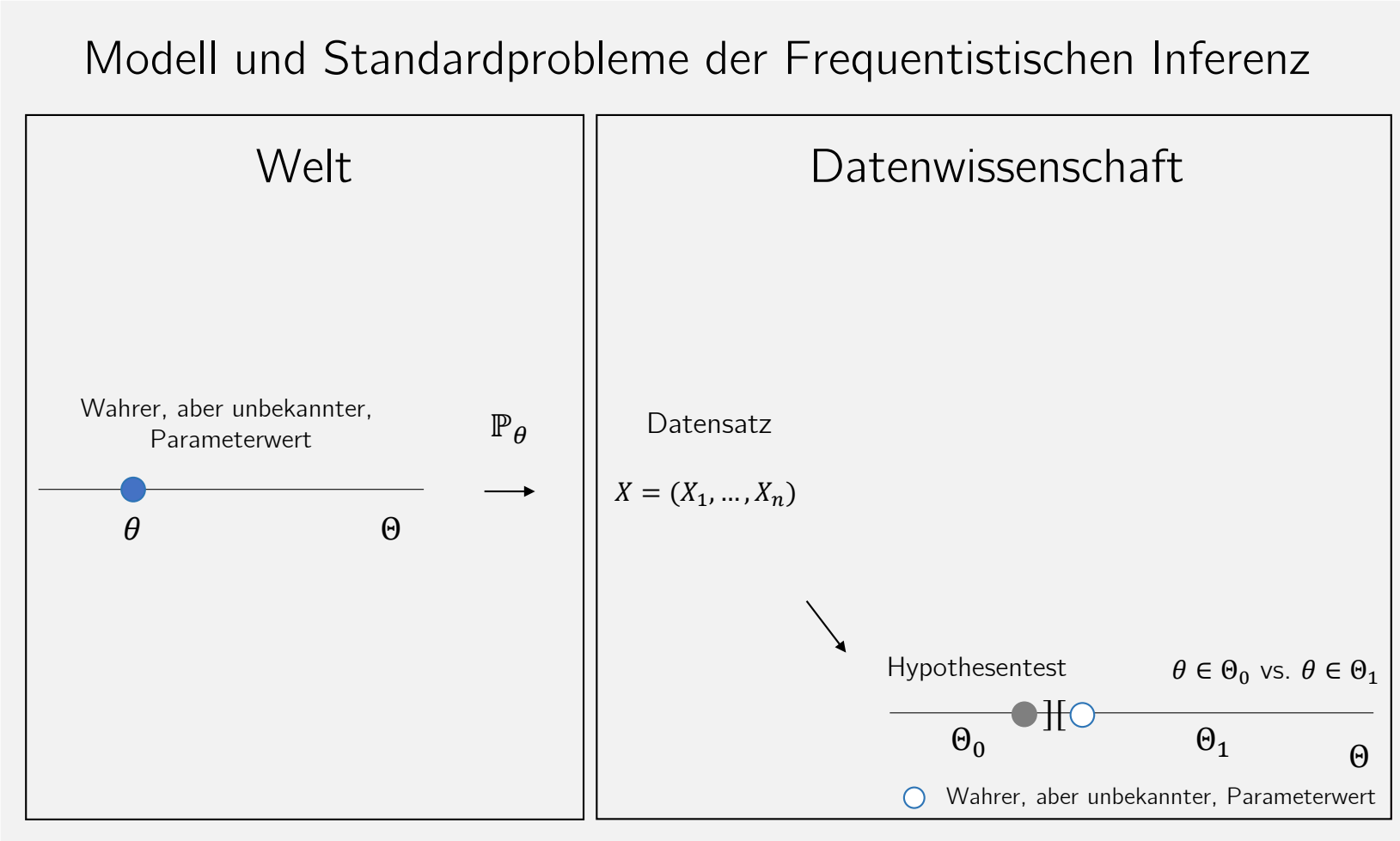
Datensatz (3) : $x^{(3)} = \left(x_1^{(3)}, x_2^{(3)}, \ldots, x_n^{(3)}\right)$, Statistik (3): $S : \mathbb{R}^n \to \Sigma, x^{(3)} \mapsto S\left(x^{(3)}\right)$

Datensatz (4) : $x^{(4)} = \left(x_1^{(4)}, x_2^{(4)}, \ldots, x_n^{(4)}\right)$, Statistik (4): $S : \mathbb{R}^n \to \Sigma, x^{(4)} \mapsto S\left(x^{(4)}\right)$

...

Um die Qualität statistischer Methoden zu beurteilen betrachtet die frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Statistiken und Schätzern unter der Annahme von $X_1, \ldots, X_n \sim p_\theta$.

Wenn eine statistische Methode im Sinne der frequentitischen Standardannahmen “gut” ist, dann heißt das also, dass sie bei häufiger Anwendung “im Mittel gut” ist. Im Einzelfall, also im realen Normalfall nur eines vorliegenden Datensatzes, kann sie auch “schlecht” sein.


Modell und Standardprobleme der Frequentistischen Inferenz
Welt
Wahrer, aber unbekannter, Parameterwert
$\mathbb{P}_\theta$
θ
Θ
Datenwissenschaft
Datensatz
$X = (X_1, \ldots, X_n)$
Hypothesentest
$\theta \in \Theta_0$ vs. $\theta \in \Theta_1$
Θ_0
Θ_1
Θ
Wahrer, aber unbekannter, Parameterwert

Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Grundlegende Logik statistischer Hypothesentests

Man hat einen Datensatz $x_1, \ldots, x_n$ vorliegen und nimmt an, dass es sich dabei um die Realisation einer Stichprobe handelt, zum Beispiel von $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

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

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.

Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Definition (Statistische Hypothesen und Testszenario)

$X_1, ..., X_n \sim p_\theta$ sei eine Stichprobe mit WMF oder WDF p_θ, $\mathcal{X}$ sei der Ergebnisraum des Zufallsvektors $X := (X_1, ..., X_n)$, 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 X genau festgelegt.
- Bei einer zusammengesetzten Hypothese ist nur die Verteilungsklasse von X 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 $\mathcal{X}$ nach $\{0, 1\}$,

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

Dabei repräsentiert

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

Bemerkung

- Weil die Stichprobe X ein Zufallsvektor ist, ist ein Test ϕ eine Zufallsvariable.

Definition (Standardtest)

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

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

und einer *Entscheidungsregel*

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

Ein Standardtest kann also geschrieben werden als

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

Bemerkungen

- Weil die Stichprobe X ein Zufallsvektor ist, sind sowohl γ als auch δ Zufallsvariablen.
- Wir betrachten in der Folge nur Standardtests.

Definition (Kritischer Bereich)

Die Untermenge K des Ergebnisraums des Zufallsvektors $X := (X_1, ..., X_n)$, für die ein Test den Wert 1 annimmt, heißt *kritischer Bereich* des Tests,

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

Bemerkungen

- Die Ereignisse $\phi(X) = 1$ und $X \in K$ sind äquivalent.
- Die Ereignisse $\phi(X) = 1$ und $X \in K$ 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,

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

Bemerkungen

- Die Ereignisse $\phi(X) = 1$ und $\gamma(X) \in A$ sind äquivalent.
- Die Ereignisse $\phi(X) = 1$ und $\gamma(X) \in A$ 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

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

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

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

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 $\phi(X) = 0$	Testentscheidung $\phi(X) = 1$
$\theta \in \Theta_0$	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

- $\mathbb{P}_\theta$ bezeichnet die Verteilung von ϕ unter der Annahme $X_1, ..., X_n \sim p_\theta$.
- Es gilt $\mathbb{P}_\theta(\phi = 1) = \mathbb{P}_\theta(X \in K) = \mathbb{P}_\theta(\gamma \in A)$
- 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 $\phi_a(X) := 0$ für alle $X \in \mathcal{X}$, 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 $\phi_b(X) := 1$ für alle $X \in \mathcal{X}$, 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 *Signifikanzniveau* 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 Signifikanzniveaus 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 Signifikanzniveau 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.

Dirk Ostwalds Meinung zum 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.

Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

(1) Statistisches Modell, Hypothesen, Teststatistik, Test

(2) Analyse der Testgütefunktion

(3) Testumfangkontrolle

(4) Analyse der Powerfunktion

Z-Test (1) Statistisches Modell, Hypothesen, Teststatistik, Test

Statistisches Modell

$X_1, ..., X_n \sim N(\mu, \sigma^2)$ sei die Stichprobe eines **Normalverteilungsmodells** mit **unbekanntem Erwartungswertparameter** μ und **bekanntem Varianzparameter** $\sigma^2 > 0$. Der Parameter dieses Modells ist $\theta = \mu$ und der Parameterraum dieses Modells ist $\Theta = \mathbb{R}$.

Testhypothesen, Teststatistik, Test

Für ein $\mu_0 \in \mathbb{R}$ betrachten wir die einfache Nullhypothese und die zusammengesetzte Alternativhypothese

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R} \setminus \{\mu_0\}, \tag{23}$$

respektive. Weiterhin betrachten wir die *Z-Teststatistik*

$$Z := \sqrt{n} \left(\frac{\bar{X}_n - \mu_0}{\sigma} \right) \tag{24}$$

und definieren den zweiseitigen kritischen Wert-basierten Test

$$\phi(X) := 1_{\{|Z| \geq k\}} = \begin{cases} 1 & |Z| \geq k \\ 0 & |Z| < k \end{cases}. \tag{25}$$

Theorem (Testgütefunktion bei zweiseitigem Z-Test, H_0 einfach)

Es sei $X_1, ..., X_n \sim N(\mu, \sigma^2)$ die Stichprobe eines Normalverteilungsmodells mit unbekanntem Erwartungswertparameter μ und bekanntem Varianzparameter $\sigma^2 > 0$. Für $\Theta := \mathbb{R}$ und ein $\mu_0 \in \mathbb{R}$ sei ein Testszenario durch

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R} \setminus \{\mu_0\}, \tag{26}$$

gegeben. Weiterhin sei mit der Z-Teststatistik

$$Z := \sqrt{n} \left(\frac{\bar{X}_n - \mu_0}{\sigma} \right) \tag{27}$$

in diesem Szenario der zweiseitige kritischer Wert-basierte Test

$$\phi(X) := 1_{\{|Z| \geq k\}} = \begin{cases} 1 & |Z| \geq k \\ 0 & |Z| < k \end{cases}. \tag{28}$$

definiert. Dann ist die Testgütefunktion von ϕ gegeben durch

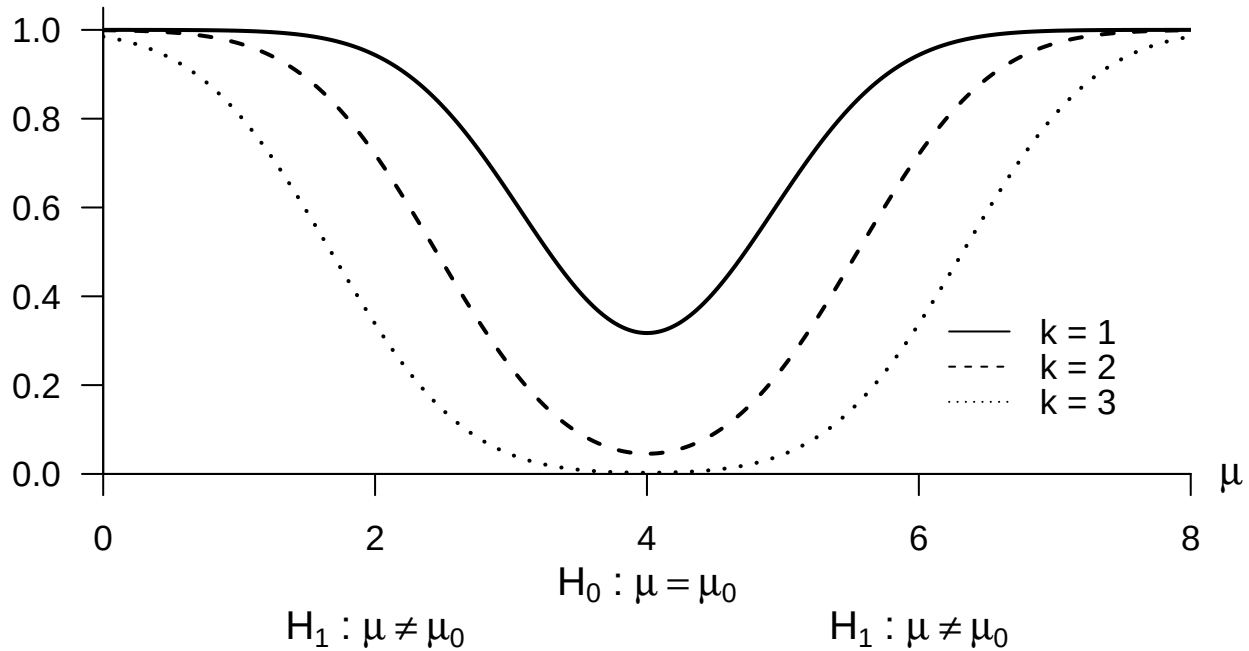
$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Phi\left(k - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0) \right) + \Phi\left(-k - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0) \right), \tag{29}$$

wobei Φ die KVF der Standardnormalverteilung bezeichnet.

Z-Test (2) Analyse der Testgütefunktion

Testgütefunktion q_ϕ für $\mu_0 = 4, n = 15, \sigma^2 = 9$ und drei zweiseitigen kritischen Wert-basierten Tests ϕ mit kritischen Werten $k = 1, 2, 3$ (vgl. DeGroot and Schervish (2012), Abbildung 9.1).

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



Z-Test (2) Analyse der Testgütefunktion

Beweis

Die Testgütefunktion des betrachteten Test im vorliegenden Testszenario ist definiert als

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1). \tag{30}$$

Wir sind also an der Wahrscheinlichkeit $\mathbb{P}_\mu(\phi = 1)$ mit $\mu \in \mathbb{R}$ interessiert, d.h. an der Wahrscheinlichkeit dafür, dass der betrachtete Test den Wert 1 annimmt, wenn der wahre, aber unbekannte, Parameterwert μ ist. Wir hatten oben festgehalten, dass die Wahrscheinlichkeiten dafür, dass ein Test den Wert 1 annimmt und dafür, dass die zugehörige Teststatistik im Ablehnungsbereich des Tests liegt, gleich sind.

Um $\mathbb{P}_\mu(\phi = 1)$ auswerten zu können, benötigen wir also zunächst einmal die Verteilung der Teststatistik für $\mu \in \mathbb{R}$. Sei also $\mu \in \mathbb{R}$. Dann betrachten wir im vorliegenden Testszenario die Stichprobe $X_1, ..., X_n \sim N(\mu, \sigma^2)$ mit unbekanntem Erwartungswertparameter μ und bekanntem Varianzparameter $\sigma^2 > 0$ und die Z-Teststatistik $Z := \sqrt{n}((\bar{X}_n - \mu_0)/\sigma))$. Man beachte, dass hier sowohl $\mu = \mu_0$ als auch $\mu \neq \mu_0$ zugelassen sind. Wir befinden uns also nicht nur im in (11) Konfidenzintervalle betrachteten Szenario $\mu = \mu_0$. Es gilt also zunächst, die Verteilung der Z-Teststatistik für ein generelles $\mu \in \mathbb{R}$ zu bestimmen.

Dazu folgen wir der in (11) Konfidenzintervalle etablierten Herangehensweise: Wir halten zunächst fest, dass mit der Mittelwertstransformation für unabhängig und identisch normalverteilte Zufallsvariablen gilt, dass $\bar{X}_n \sim N(\mu, \sigma^2/n)$ (cf. (8) Transformationen der Normalverteilung). Die Verteilung der Z-Teststatistik folgt dann mit dem WDF Transformationstheorem bei linear-affinen Abbildungen (cf. ibid.) durch Transformation von $\bar{X}_n$ mit der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \bar{x}_n \mapsto f(\bar{x}_n) := \frac{\sqrt{n}}{\sigma}\bar{x}_n - \frac{\sqrt{n}}{\sigma}\mu_0 \tag{31}$$

Z-Test (2) Analyse der Testgütefunktion

Beweis (fortgeführt)

Wir erhalten

$$
\begin{aligned}
p_Z(z) &= \frac{\sigma}{\sqrt{n}} N\left(\frac{\sigma}{\sqrt{n}}\left(z+\frac{\sqrt{n}}{\sigma}\mu_0\right); \mu, \frac{\sigma^2}{n}\right) \\
&= \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left(-\frac{1}{2\frac{\sigma^2}{n}}\left(\frac{\sigma}{\sqrt{n}}\left(z+\frac{\sqrt{n}}{\sigma}\mu_0\right)-\mu\right)^2\right) \\
&= \frac{\sigma}{\sqrt{n}} \frac{\sqrt{n}}{\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{n}{\sigma^2}\left(\frac{\sigma}{\sqrt{n}}z-(\mu-\mu_0)\right)^2\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{n}{\sigma^2}\left(\frac{\sigma^2}{n}z^2-2\frac{\sigma}{\sqrt{n}}z(\mu-\mu_0)+(\mu-\mu_0)^2)\right)\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}\frac{\sigma^2}{n}z^2-2\frac{n}{\sigma^2}\frac{\sigma}{\sqrt{n}}z(\mu-\mu_0)+\frac{n}{\sigma^2}(\mu-\mu_0)^2)\right)\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(z^2-2z\frac{\sigma}{\sqrt{n}}\frac{\sqrt{n}}{\sigma}\frac{\sqrt{n}}{\sigma}(\mu-\mu_0)+\left(\frac{\sqrt{n}}{\sigma}\right)^2(\mu-\mu_0)^2)\right)\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(z^2-2z\frac{\sqrt{n}}{\sigma}(\mu-\mu_0)+\left(\frac{\sqrt{n}}{\sigma}(\mu-\mu_0)\right)^2\right)\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(z-\frac{\sqrt{n}}{\sigma}(\mu-\mu_0)\right)^2\right)
\end{aligned}
\tag{32}
$$

Z-Test (2) Analyse der Testgütefunktion

Beweis (fortgeführt)

Wir haben also gefunden, dass für die Z-Teststatistik für beliebiges $\mu \in \mathbb{R}$ gilt, dass

$$Z \sim N\left(\frac{\sqrt{n}}{\sigma}(\mu - \mu_0), 1\right). \tag{33}$$

Für den Fall, dass $\mu_0 = \mu$ ist, die Nullhypothese also zutrifft, folgt $Z \sim N(0, 1)$. Als nächstes bestimmen wir den Ablehnungsbereich des betrachteten Tests. Per Definition gilt

$$A := \{Z \in \mathbb{R} | \phi(X) = 1\} \tag{34}$$

Es ergibt sich also

$$\begin{aligned}
A &= \{Z \in \mathbb{R} | \phi(X) = 1\} \\
&= \{Z \in \mathbb{R} | |Z| \geq k\} \\
&= \left\{ Z \in \mathbb{R} \middle| \begin{cases} Z \geq k, & Z \geq 0 \\ -Z \geq k, & Z < 0 \end{cases} \right\} \\
&= \left\{ Z \in \mathbb{R} \middle| \begin{cases} Z \geq k, & Z \geq 0 \\ Z \leq -k, & Z < 0 \end{cases} \right\} \\
&= \{Z \in \mathbb{R} | Z < 0, Z \leq -k\} \cup \{Z \in \mathbb{R} | Z > 0, Z \geq k\} \\
&=]-\infty, -k] \cup]k, \infty[.
\end{aligned} \tag{35}$$

Z-Test (2) Analyse der Testgütefunktion

Beweis (fortgeführt)

Mit diesem Ablehungsbereich des betrachteten Tests ergibt sich dann

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

Wir sind also an der Wahrscheinlichkeiten $\mathbb{P}_\mu(Z \leq -k)$ und $\mathbb{P}_\mu(Z \leq -k)$ interessiert. Dazu halten wir zunächst fest, dass für jede normalverteilte Zufallsvariable $X \sim N(m, s^2)$ gilt, dass

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(\frac{X - m}{s} \leq \frac{x - m}{s}\right) = \mathbb{P}\left(\xi \leq \frac{x - m}{s}\right) = \Phi\left(\frac{x - m}{s}\right) \tag{37}$$

gilt, wobei ξ hier die durch $(X - m)/s$ definierte Z-verteilte Zufallsvariable bezeichnet. Im vorliegenden Fall folgt mit $m = \frac{\sqrt{n}}{\sigma}(\mu - \mu_0), s = 1$
und $x = k$ bzw. $x = -k$ also

$$q_\phi(\mu) = 1 - \mathbb{P}_\mu(Z \leq k) + \mathbb{P}_\mu(Z \leq -k) = 1 - \Phi\left(k - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)\right) + \Phi\left(-k - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)\right)$$

und wir haben alles gezeigt. □

Theorem (Testumfangankontrolle bei zweiseitigem Z-Test, H_0 einfach)

Es sei $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ die Stichprobe eines Normalverteilungsmodells mit unbekanntem Erwartungswertparameter μ und bekanntem Varianzparameter $\sigma^2 > 0$. Für $\Theta := \mathbb{R}$ und ein $\mu_0 \in \mathbb{R}$ sei ein Testszenario durch

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R} \setminus \{\mu_0\}, \tag{38}$$

gegeben. Weiterhin sei mit der Z-Teststatistik

$$Z := \sqrt{n} \left(\frac{\bar{X}_n - \mu_0}{\sigma} \right) \tag{39}$$

in diesem Szenario der zweiseitige kritische Wert-basierte Test

$$\phi(X) := 1_{\{|Z| \geq k\}} = \begin{cases} 1 & |Z| \geq k \\ 0 & |Z| < k \end{cases}. \tag{40}$$

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

$$k := \Phi^{-1} \left(1 - \frac{\alpha_0}{2} \right) \tag{41}$$

Wir bezeichnen diesen kritischen Wert mit k_{α_0}.

Z-Test (3) Testumfangkontrolle

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

$$k_{\alpha_0} := \Phi^{-1}\left(1 - \alpha_0/2\right) \tag{42}$$

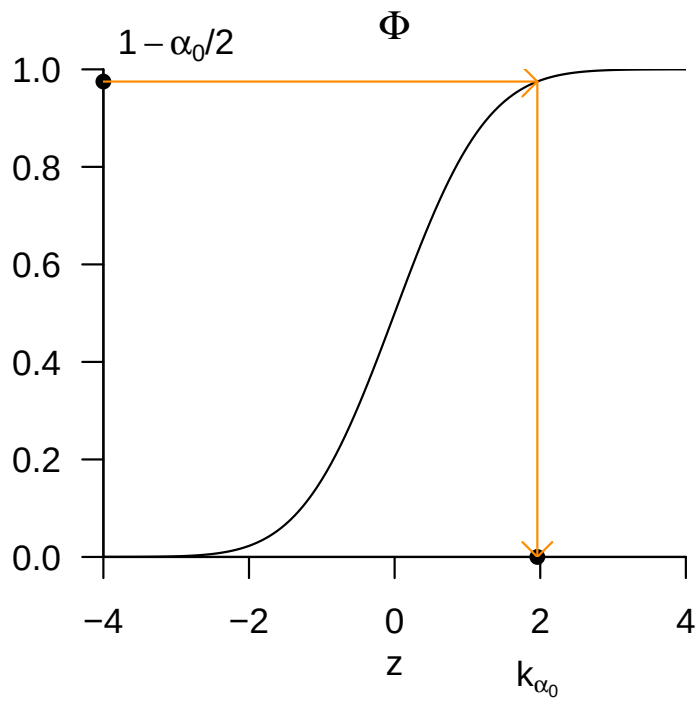
garantiert, dass ϕ ein Level-α_0-Test mit Testumfang α_0 ist. Sei also $k := k_{\alpha_0}$. Dann gilt

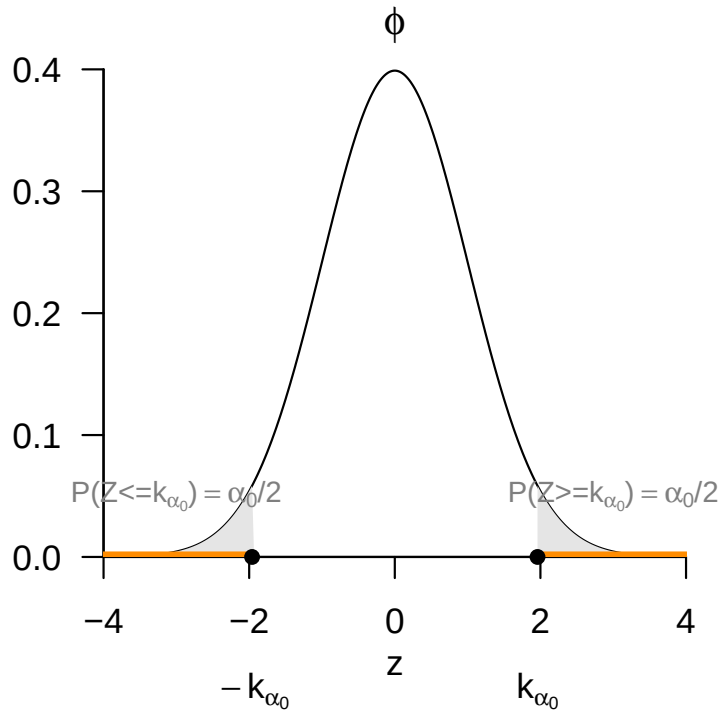
$$\begin{aligned}
q_\phi(\mu_0) &= 1 - \Phi(k_{\alpha_0}) + \Phi(-k_{\alpha_0}) \\
&= 1 - \Phi(k_{\alpha_0}) + (1 - \Phi(k_{\alpha_0})) \\
&= 2 - 2\Phi(k_{\alpha_0}) \\
&= 2(1 - \Phi(k_{\alpha_0})) \\
&= 2\left(1 - \Phi\left(\Phi^{-1}\left(1 - \alpha_0/2\right)\right)\right) \\
&= 2\left(1 - 1 + \alpha_0/2\right) \\
&= 2\alpha_0/2 \\
&= \alpha_0,
\end{aligned} \tag{43}$$

wobei die zweite Gleichung mit der Symmetrie der Standardnormalverteilung 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.

Z-Test (3) Testumfangkontrolle

Wahl von $k_{\alpha_0} := \Phi^{-1}(1 - \alpha_0/2)$ mit $\alpha_0 := 0.05$ und resultierender Ablehnungsbereich





Praktisches Vorgehen

- Man nimmt an, dass ein vorliegender Datensatz $x_1, ..., x_n$ eine Realisation von $X_1, ..., X_n \sim N(\mu, \sigma^2)$ mit unbekanntem μ und bekanntem Varianzparameter $\sigma^2 > 0$ 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 Signifikanzniveau α_0 und bestimmt den zugehörigen kritischen Wert k_{α_0}. Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05$, dass $k_{0.05} = \Phi^{-1}(1 - 0.05/2) \approx 1.96$ ist.
- Anhand von n, μ_0, σ^2 und $\bar{x}_n$ berechnet man die Realisierung der Z-Teststatistik

$$z := \sqrt{n} \left(\frac{\bar{x}_n - \mu_0}{\sigma} \right) \tag{44}$$

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

Z-Test (3) Testumfangkontrolle

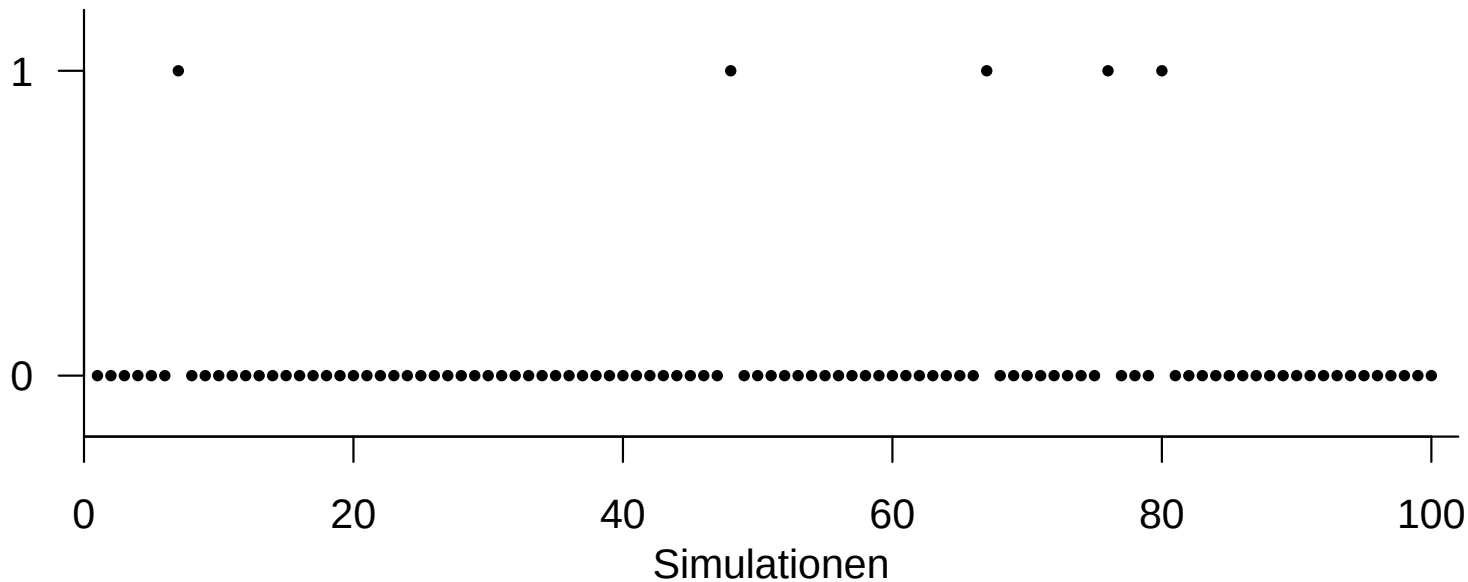
Simulation

```
# Modellspezifikation
mu        = 2                                   # w.a.u. Erwartungswertparameter
sigsqr    = 1                                   # bekannter Varianzparameter
n         = 12                                  # Stichprobengröße
mu_0      = mu                                  # Nullhypothese H_0
alpha_0   = 0.05                                # Signifikanzniveau
k_alpha_0 = qnorm(1 - (alpha_0/2))              # kritischer Wert

# Simulation
n_sim     = 1e4                                 # Anzahl von Stichprobenrealisierung
phi       = rep(NaN, n_sim)                     # Testergebnisarray
for(i in 1:n_sim){
  X         = rnorm(n, mu, sqrt(sigsqr))        # Stichprobenrealisation
  Z         = sqrt(n)*(mean(X) - mu_0)/sqrt(sigsqr)   # Teststatistik
  if(abs(Z) >= k_alpha_0){                      # Test 1_{|Z(X) >= k_alpha_0|}
      phi[i] = 1                                # Ablehnen von H_0
  } else {
      phi[i] = 0                                # Nicht Ablehnen von H_0
  }
}
cat("Geschätzter Testumfang alpha =", mean(phi))        # Ausgabe
```

```
> Geschätzter Testumfang alpha = 0.0517
```

Simulation

$\phi(x)$ für $\mu = 2, \sigma^2 = 2, n = 12, \mu_0 = 2, \alpha_0 = 0.05$



Z-Test (4) Analyse der Powerfunktion

Wir betrachten die Testgütefunktion

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Phi\left(k - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)\right) + \Phi\left(-k - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)\right) \quad (45)$$

bei bekanntem und festen Wert von $\sigma^2 > 0$, bei festgelegter Nullhypothese, also festem Wert von μ_0 und bei kontrolliertem Testumfang, also für $k := k_{\alpha_0} = \Phi^{-1}(1 - \alpha_0/2)$ mit festem α_0, als Funktion der übrigen Testszenarioparameter.

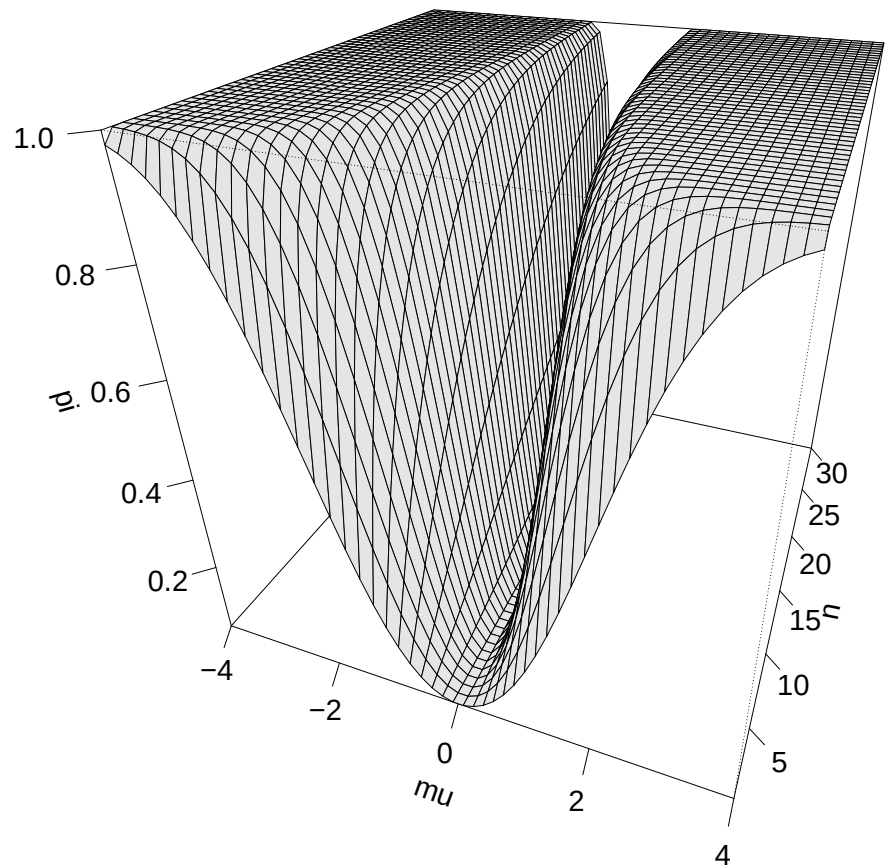
Es ergibt sich die bivariate reellwertige Funktion

$$\pi : \mathbb{R} \times \mathbb{N} \to [0,1], (\mu, n) \mapsto$$
$$\pi(\mu, n) := 1 - \Phi\left(k_{\alpha_0} - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)\right) + \Phi\left(-k_{\alpha_0} - \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)\right) \quad (46)$$

Bei festgelegten $\sigma^2 > 0, \mu_0$, k_{α_0} hängt für $\mu \notin \{\mu_0\}$ die Powerfunktion des zweiseitigen Z-Tests mit einfacher Nullhypothese also vom unbekannten Wert μ und von der Stichprobengröße n ab. Wir visualisieren diese Abhängigkeiten untenstehend.

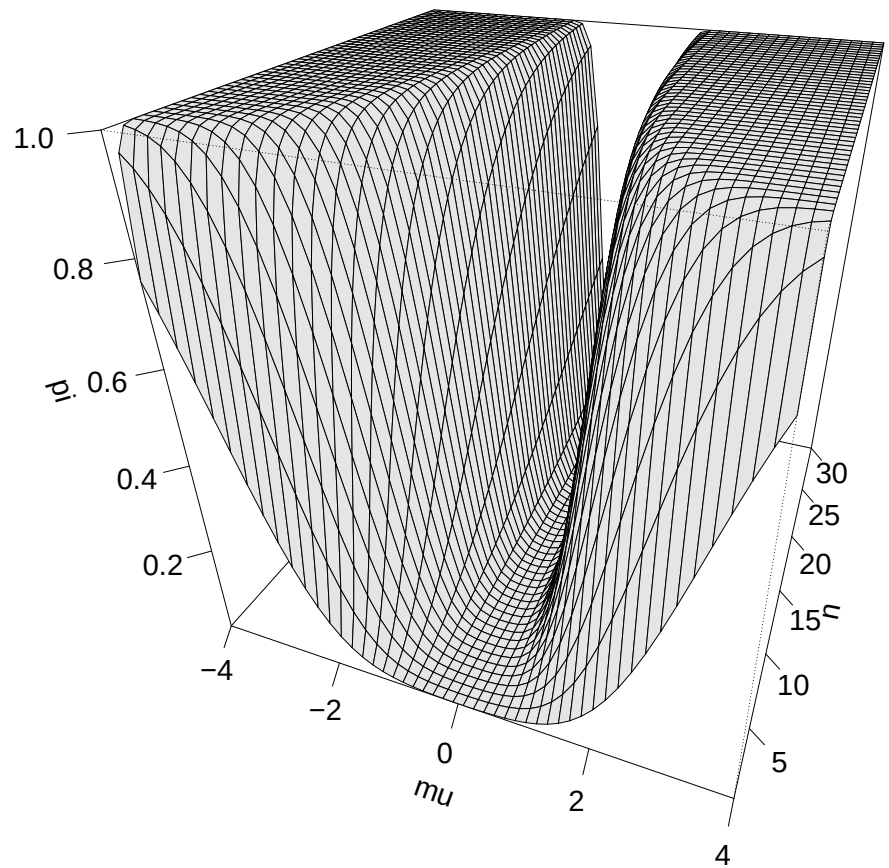
Z-Test (4) Analyse der Powerfunktion

Powerfunktion für $\mu_0 = 0$ und $\sigma^2 = 1$ bei $\alpha_0 = 0.05$



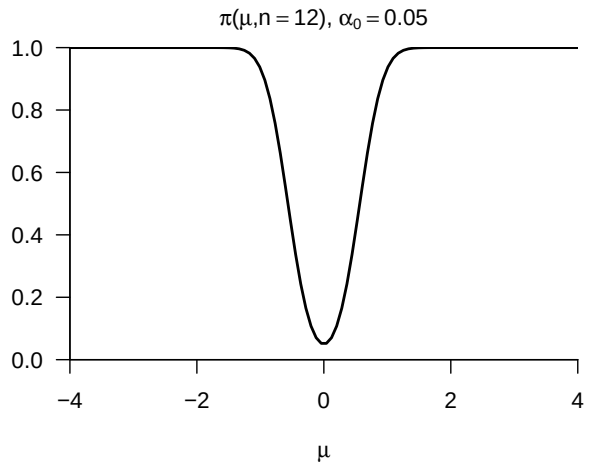
Z-Test (4) Analyse der Powerfunktion

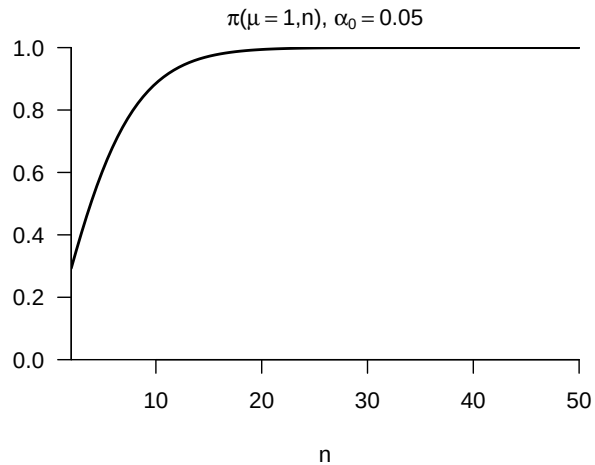
Powerfunktion für $\mu_0 = 0$ und $\sigma^2 = 1$ bei $\alpha_0 = 0.001$

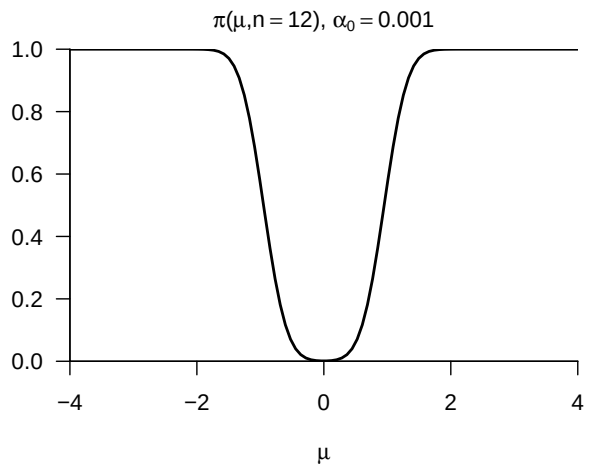


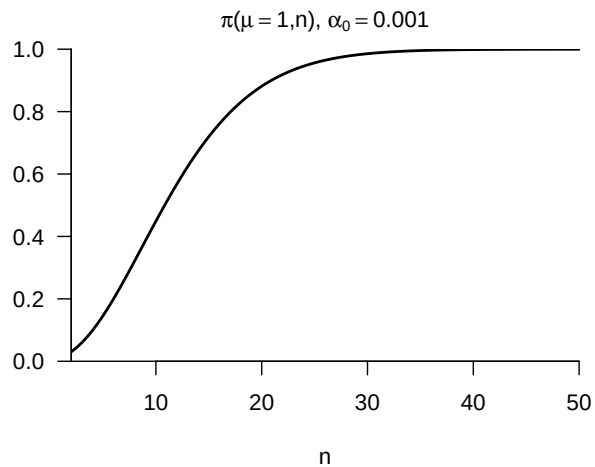
Z-Test (4) Analyse der Powerfunktion

Powerfunktionen für $\mu_0 = 0$ und $\sigma^2 = 1$









Z-Test (4) Analyse der Powerfunktion

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 μ ab.

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

Generell wird folgendes Vorgehen favorisiert

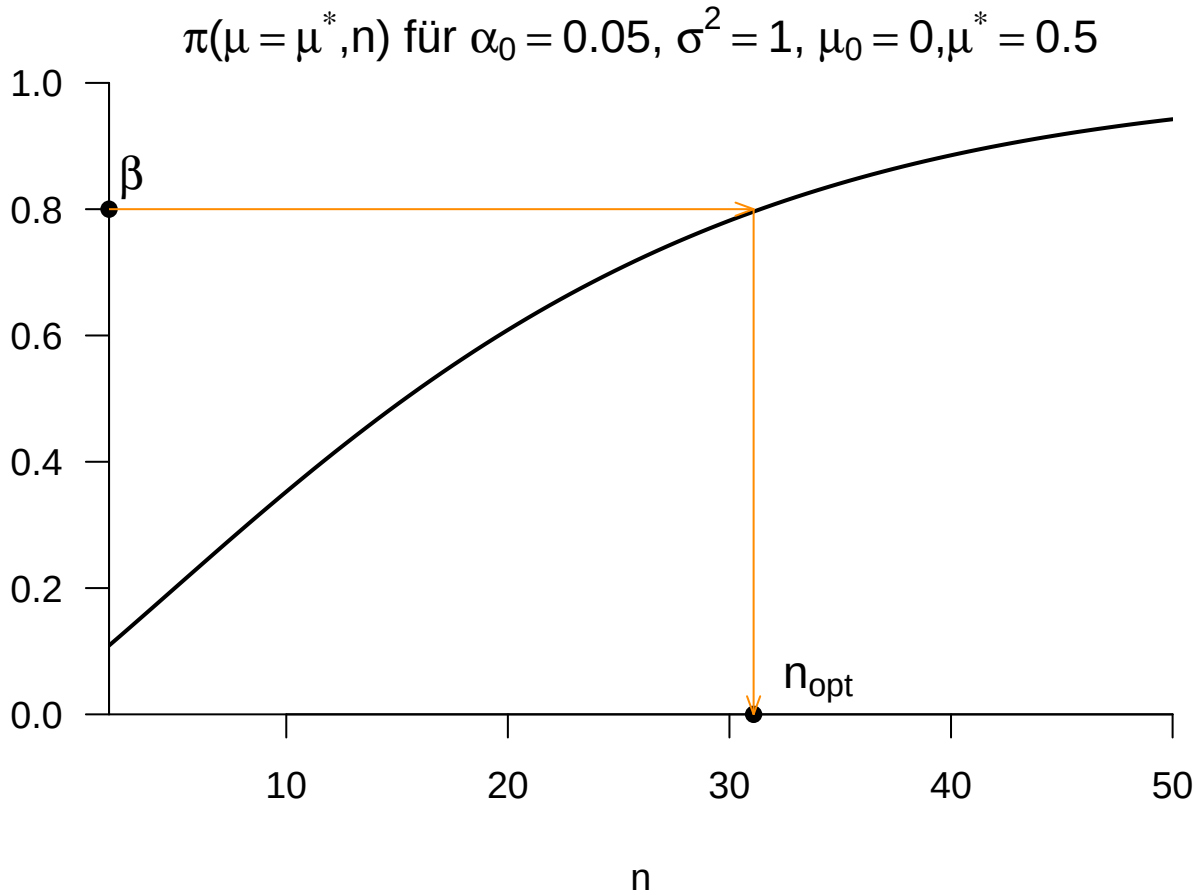
- Man legt das Signifikanzniveau α_0 fest und evaluiert die Powerfunktion.
- Man wählt einen Mindestparameterwert μ^*, den man mit $\pi(n, \mu) = \beta$ detektieren möchte.
- Ein konventioneller Wert ist $\beta = 0.8$.
- Man liest die für $\pi(n, \mu = \mu^*) = \beta$ nötige Stichprobengröße n ab.

Praktisches Vorgehen

```
# Szenariospezifikation
sigma     = 1                                                    # bekanntes \sigma
mu_0      = 0                                                    # einfache Nullhypothese
n_min     = 2                                                    # n Minimum
n_max     = 50                                                   # n Maximum
n_res     = 1e2                                                  # n Auflösung
n         = seq(n_min,n_max, len = n_res)                        # n Raum
alpha_0   = 0.05                                                 # Signifikanzniveau
k_alpha_0 = qnorm(1 - alpha_0/2)                                 # kritischer Wert

# Poweranalyse
mu_star   = .5                                                   # Mindestparameterwert
pi_n      = (1-pnorm( k_alpha_0-sqrt(n)/sigma*(mu_star-mu_0))    # Powerfunktion
              +pnorm(-k_alpha_0-sqrt(n)/sigma*(mu_star-mu_0)))
beta      = 0.8                                                  # gewünschter Powerfunktionswert
i         = 1                                                    # Indexinitialisierung
n_min     = NaN                                                  # minimales n Initialisierung
while(pi_n[i] < beta){                                           # Solange \pi(n,mu*) < \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 = 32
```

Z-Test (4) Analyse der Powerfunktion



Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Motivation

- Es werde ein zweiseitiger Z-Test mit Signifikanzlevel $\alpha_0 = 0.05$ durchgeführt.
  - H_0 wird abgelehnt, wenn $|Z| \geq \Phi^{-1}(1 - 0.05/2) \approx 1.96$.
- Nehmen wir an, es werde $z = 2.01$ beobachtet.
  - Das Testergebnis lautet "H_0 Ablehnen".
- Nehmen wir an, es werde $z = 3.81$ 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 $z = 2.01$ würde H_0 für jedes α_0 mit $2.01 \geq \Phi^{-1}(1 - \alpha_0/2)$ abgelehnt werden.
- Bei $z = 3.81$ würde H_0 für jedes α_0 mit $3.81 \geq \Phi^{-1}(1 - \alpha_0/2)$ abgelehnt werden.

- Das kleinste Signifikanzlevel α_0, bei dem man H_0 basierend auf einem Wert der Teststatistik ablehnen würde, wird *p-Wert* 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 Z-Test mit einfacher Nullhypothese)

- Bei $Z = z$ würde H_0 für jedes α_0 mit $|z| \geq \Phi^{-1}(1 - \alpha_0/2)$ abgelehnt werden. Für diese α_0 gilt, wie unten gezeigt,
$$\alpha_0 \geq 2\mathbb{P}(Z \geq |z|). \tag{47}$$
- Das kleinste $\alpha_0 \in [0, 1]$ mit $\alpha_0 \geq 2\mathbb{P}(Z \geq |z|)$ ist dann $\alpha_0 = 2\mathbb{P}(Z \geq |z|)$, also folgt
$$\text{p-Wert} = 2\mathbb{P}(Z \geq |z|) = 2(1 - \Phi(|z|)). \tag{48}$$
- Zum Beispiel ist für $Z = 2.01$ der p-Wert 0.04, für $Z = -2.01$ ist der p-Wert auch 0.04, für $Z = 3.81$ ist der p-Wert 0.0001, und für $Z = -3.81$ ist der p-Wert auch 0.0001

Beispiel (Zweiseitiger Z-Test mit einfacher Nullhypothese, fortgeführt)

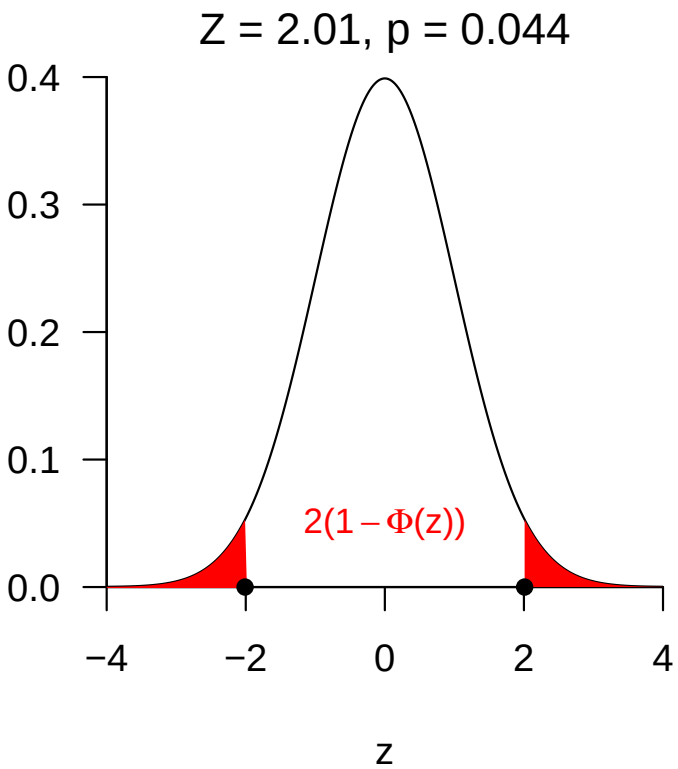
Es bleibt zu zeigen, dass gilt

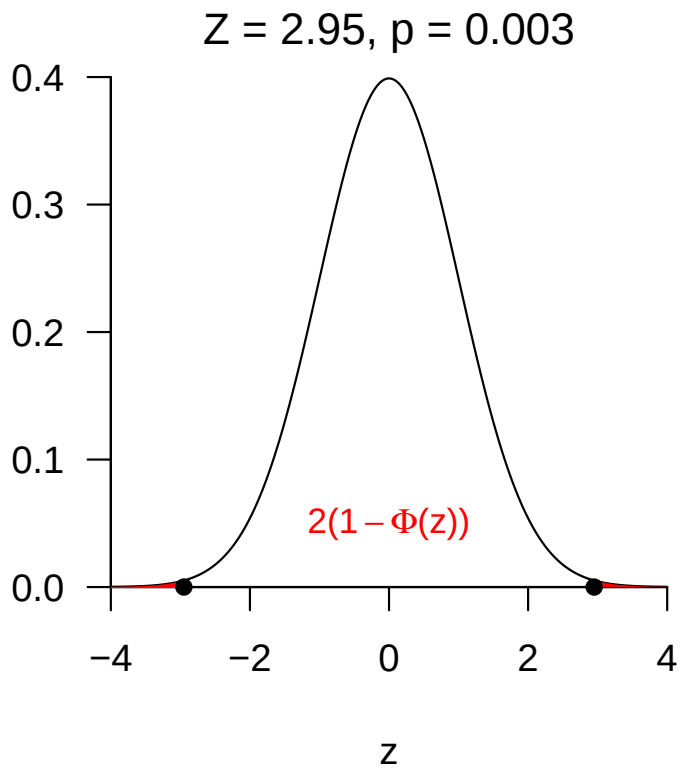
$$|z| \geq \Phi^{-1}(1-\alpha_0/2) \Leftrightarrow \alpha_0 \geq 2\mathbb{P}(Z \geq |z|) \tag{49}$$

Wir haben

$$\begin{aligned}
|z| &\geq \Phi^{-1}\left(1-\frac{\alpha_0}{2}\right) \\
\Leftrightarrow |z| &\geq \Phi^{-1}\left(1-\frac{\alpha_0}{2}\right) \\
\Leftrightarrow \Phi(|z|) &\geq \Phi\left(\Phi^{-1}\left(1-\frac{\alpha_0}{2}\right)\right) \\
\Leftrightarrow \Phi(|z|) &\geq 1-\frac{\alpha_0}{2} \\
\Leftrightarrow \mathbb{P}(Z \leq |z|) &\geq 1-\frac{\alpha_0}{2} \\
\Leftrightarrow \frac{\alpha_0}{2} &\geq 1-\mathbb{P}(Z \leq |z|) \\
\Leftrightarrow \frac{\alpha_0}{2} &\geq \mathbb{P}(Z \geq |z|) \\
\Leftrightarrow \alpha_0 &\geq 2\mathbb{P}(Z \geq |z|).
\end{aligned} \tag{50}$$

Beispiel (Zweiseitiger Z-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.

Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

Selbstkontrollfragen

Theorem (Dualität von Konfidenzintervallen und Hypothesentests I)

$X = (X_1, ..., X_n) \sim p_\theta$ sei eine Stichprobe mit Ergebnisraum $\mathcal{X}$ und Parameterraum Θ. Weiterhin sei $[b_l(X), b_u(X)]$ ein δ-Konfidenzintervall für θ.
Dann ist der Hypothesentest

$$\phi_\theta : \mathcal{X} \to \{0,1\}, x \mapsto \phi(x) := \begin{cases} 0, & [b_l(x), b_u(x)] \ni \theta \\ 1, & [b_l(x), b_u(x)] \not\ni \theta \end{cases} \tag{51}$$

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

$$\Theta_0 := \{\theta\} \text{ und } \Theta_1 := \Theta \setminus \{\theta\}. \tag{52}$$

Beweis

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

$$\alpha_0 = \alpha = \mathbb{P}_\theta(\phi(X) = 1) = \mathbb{P}_\theta([b_l(x), b_u(x)] \not\ni \theta) = 1 - \mathbb{P}_\theta([b_l(x), b_u(x)] \ni \theta) = 1 - \delta. \tag{53}$$

Bemerkungen

- θ **bezeichnet hier das Element der einfachen Nullhypothese**
- Mit δ-Konfidenzintervallen kann man also Tests vom Signifikanzniveau $\alpha_0 = 1 - \delta$ konstruieren.

Beispiel (Konstruktion eines Hypothesentests aus einem Konfidenzintervall)

Wir haben bereits gesehen, dass für eine Stichprobe $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$ mit unbekanntem Erwartungswertparameter μ_0 und bekanntem Varianzparameter $\sigma^2 > 0$, $\delta \in]0,1[$ und $z_\delta := \Phi^{-1}\left(\frac{1+\delta}{2}\right)$,

$$C_n := \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\delta, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\delta\right]. \tag{54}$$

ein δ-Konfidenzintervall definiert.

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:

$$\phi : \mathcal{X} \to \{0,1\}, x \mapsto \phi(x) := \begin{cases} 0, & \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\delta, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\delta\right] \ni \mu_0 \\ 1, & \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\delta, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\delta\right] \not\ni \mu_0 \end{cases} \tag{55}$$

Dann gilt

$$\begin{aligned} \mathbb{P}_{\mu_0}(\phi(X) = 1) &= 1 - \mathbb{P}_{\mu_0}(\phi(X) = 0) \\ &= 1 - \mathbb{P}_{\mu_0}\left(\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\delta, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\delta\right] \ni \mu_0\right) \\ &= 1 - \delta. \end{aligned} \tag{56}$$

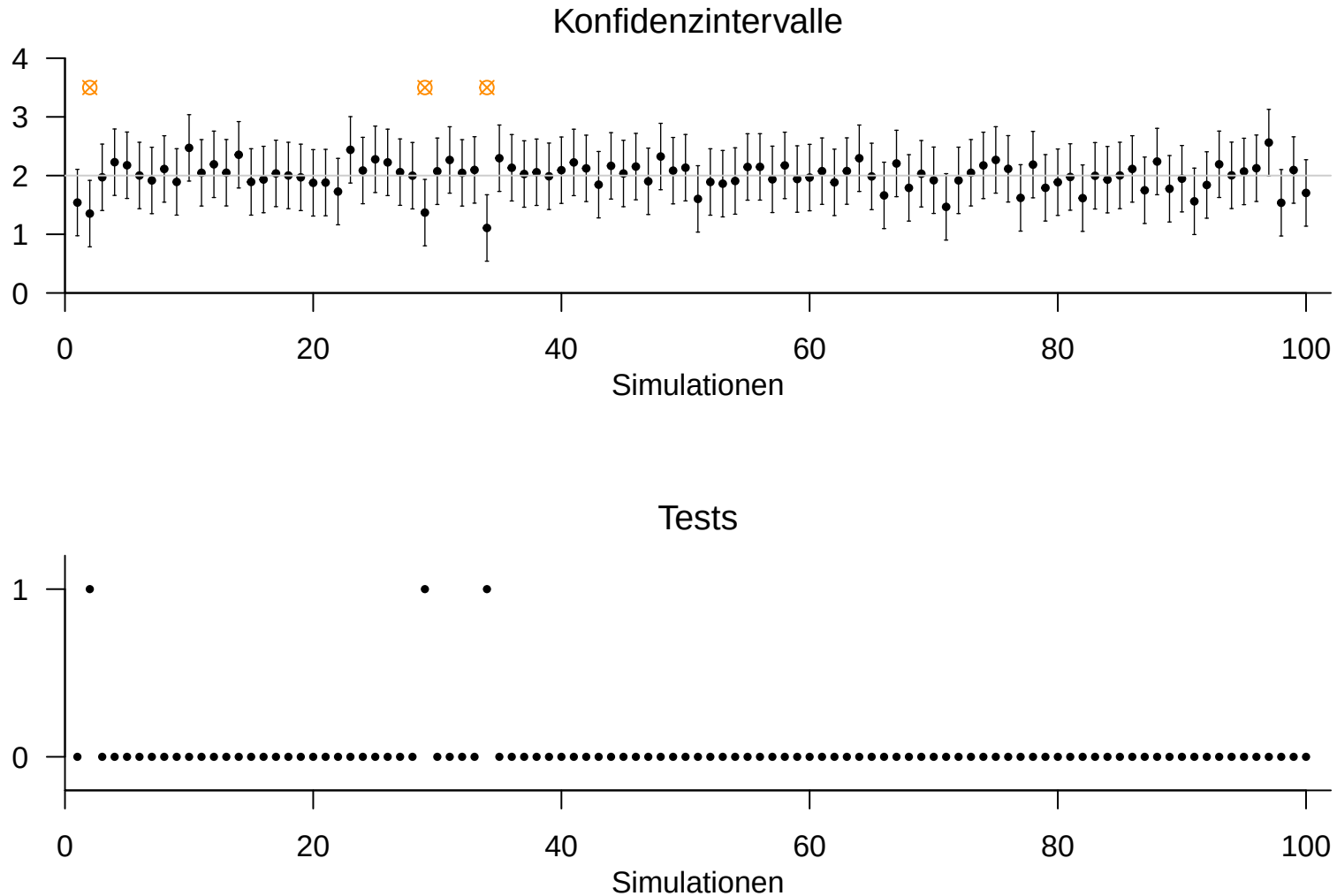
und wir haben gezeigt, dass ϕ ein Test vom Signifikanzniveau $\alpha_0 = 1 - \delta$ ist.

Simulation

```
# Modellformulierung
mu_0    = 2                                     # w.a.u. Erwartungswertparameter/Nullhypothese
sigma   = 1                                     # Standardabweichungsparameter
n       = 12                                    # Stichprobengröße
delta   = 0.95                                  # Konfidenzbedingung
phi_inv = qnorm((1+delta)/2)                    # \Phi^{-1}((\delta + 1)/2)

# Simulationssetup
ns      = 1e2                                   # Anzahl Simulationen
X_bar   = rep(NaN,ns)                           # Stichprobenmittelarray
C       = matrix(rep(NaN,2*ns), ncol = 2)       # Konfidenzintervallarray
kfn     = rep(NaN,ns)                           # Überdeckungsindikatorarray
phi     = rep(NaN,ns)                           # Testarray

# Simulationsiterationen
for(i in 1:ns){

  # Datengeneration
  X        = rnorm(n,mu_0,sigma)                # Stichprobenrealisierung

  # Konfidenzintervallevaluation
  X_bar[i] = mean(X)                            # Stichprobenmittel
  C[i,1]   = X_bar[i] - (sigma/sqrt(n))*phi_inv # untere KI Grenze
  C[i,2]   = X_bar[i] + (sigma/sqrt(n))*phi_inv # obere KI Grenze

  # Überdeckungsindikatorevaluation
  if(C[i,1] <= mu_0 & mu_0 <= C[i,2]){
      kfn[i] = 1} else{
      kfn[i] = 0}

  # Testevaluation
  if(C[i,1] <= mu_0 & mu_0 <= C[i,2]){
      phi[i] = 0} else{
      phi[i] = 1}
}
cat(   "Geschätztes Konfidenzniveau = ", mean(kfn),
     "\nGeschätzter Testumfang      = ", mean(phi))
```

```
> Geschätztes Konfidenzniveau =  0.97
> Geschätzter Testumfang      =  0.03
```

Simulation


Konfidenzintervalle
4
3
2
1
0
0
20
40
60
80
100
Simulationen
Tests
1
0
0
20
40
60
80
100
Simulationen

Theorem (Dualität von Konfidenzintervallen und Hypothesentests II)

Unter den gleichen Annahmen wie im vorherigen Theorem sei

$$\Phi := \{\phi_\theta(X) | \theta \in \Theta\} \tag{57}$$

eine Familie von Tests, so dass $\phi_\theta(X)$ ein Test mit Signifikanzniveau α_0 für die Hypothesen

$$\Theta_0 := \{\theta\} \text{ and } \Theta_1 := \Theta \setminus \{\theta\}. \tag{58}$$

sei.
Man nehme weiter an, dass die Menge

$$C_n := \{\theta \in \Theta | \phi_\theta(X) = 0\} \tag{59}$$

als $C_n = [b_l(X), b_u(X)]$ für entsprechend bestimmte $b_l(X)$ und $b_u(X)$ geschrieben werden kann.
Dann ist C_n ein $\delta := 1 - \alpha_0$ Konfidenzintervall für θ.

Beweis

Aufgrund der einfachen Nullhypothese und somit $\alpha_0 = \alpha$ folgt das Level $\delta = 1 - \alpha_0$ des Konfidenzintervalls aus

$$\delta = \mathbb{P}_\theta(C_n \ni \theta) = \mathbb{P}_\theta(\phi_\theta(X) = 0) = 1 - \mathbb{P}_\theta(\phi_\theta(X) = 1) = 1 - \alpha_0. \tag{60}$$

Bemerkung

- Mit Tests von Signifikanzniveau α_0 kann man also $1 - \alpha_0$-Konfidenzintervalle konstruieren.

Vorbemerkungen

Grundlegende Definitionen

Z-Test

p-Werte

Konfidenzintervalle und Hypothesentests

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. Nennen Sie vier Schritte in der Konstruktion eines Tests.

Selbstkontrollfragen

17. Definieren Sie das statistische Modell eines Z-Tests.
18. Definieren Sie die Hypothesen eine Z-Tests mit einfacher Nullhypothese und zweiseitiger Alternativhypothese.
19. Definieren Sie den zweiseitigen Z-Test.
20. Skizzieren Sie qualitativ Testgütefunktionen eines zweiseitigen Z-Tests für verschiedene kritische Werte.
21. Wie muss der kritische Wert eines zweiseitigen Z-Tests 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 zweiseitigen Z-Test.
23. Erläutern Sie das praktische Vorgehen zur Durchführung eines zweiseitigen Z-Tests.
24. Von welchen Werten hängt die Powerfunktion eines zweiseitigen Z-Tests ab?
25. Skizzieren Sie qualitativ die Powerfunktion des zweiseitigen Z-Tests bei fester Stichprobengröße.
26. Skizzieren Sie qualitativ die Powerfunktion des zweiseitigen Z-Tests 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 erste Theorem zur Dualität von Konfidenzintervallen und Hypothesentests wieder.
31. Geben Sie das zweite Theorem zur Dualität von Konfidenzintervallen und Hypothesentests wieder.
32. Erläutern Sie die Dualität von Konfidenzintervallen und Hypothesentests.

References

DeGroot, Morris H., and Mark J. Schervish. 2012. *Probability and Statistics*. 4th ed. Boston: Addison-Wesley.

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.

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.