4.3 Methoden zur Evaluierung von Tests

Bei der Entscheidung H_0 als wahr oder als falsch zu klassifizieren können Fehler gemacht werden. Tests vergleicht man miteinander bzgl. ihrer Wahrscheinlichkeit Fehler zu begehen.

Fragen:

- Wie können diese Wahrscheinlichkeiten kontrolliert werden?
- Welcher Test hat die geringste Fehlerwahrscheinlichkeit?

4.3.1 Fehlerwahrscheinlichkeiten und die Power Funktion

Wählt man zwischen $H_0 : \theta \in \Theta_0$ und $H_1 : \theta \in \Theta_0^c$ können generell 4 Situationen eintreten:

Wahrheit	Entscheidung H_0	Entscheidung H_1
H_0	ok	Type I Error
H_1	Type II Error	ok

Bezeichne R den kritischen Bereich.

Ist $\theta \in \Theta_0$: $P_\theta(\text{Type I Error}) = P_\theta(\mathbf{X} \in R)$

Ist $\theta \in \Theta_0^c$: $P_\theta(\text{Type II Error}) = P_\theta(\mathbf{X} \notin R) = 1 - P_\theta(\mathbf{X} \in R)$

Also gilt:

$$P_\theta(\mathbf{X} \in R) = \begin{cases} \text{Wahrscheinlichkeit für Type I Error falls } \theta \in \Theta_0 \\ 1 - \text{Wahrscheinlichkeit für Type II Error falls } \theta \in \Theta_0^c \end{cases}$$

Definition 4.3.1: Die **Power Funktion** (Macht-Funktion) eines Hypothesentests mit Verwerfungsbereich R ist die Funktion in θ definiert durch

$$\beta(\theta) = P_\theta\left(\mathbf{X} \in R\right) .$$

Die **ideale** Power Funktion wäre:

$$\beta(\theta) = \begin{cases} 0 & \text{für } \theta \in \Theta_0 \\ 1 & \text{für } \theta \in \Theta_0^c \end{cases}$$

Bis auf sehr wenige Ausnahmen ist diese aber **nie** erreichbar. Ein guter Test hat Power nahe Null für die meisten $\theta \in \Theta_0$ und nahe Eins für die meisten $\theta \in \Theta_0^c$.

Beispiel 4.3.1: Sei $X \sim$ Binomial$(5, \theta)$. Teste $H_0 : \theta \leq 1/2$ gegen $H_1 : \theta > 1/2$.

(1) Ein Test verwirft H_0 falls nur Erfolge beobachtet sind $(R = \{5\})$ mit Power

$$\beta_1(\theta) = P_\theta(X \in R) = P_\theta(X = 5) = \binom{5}{5}\theta^5(1-\theta)^{5-5} = \theta^5\,.$$

• Für $\theta \in [0, 1/2]$ gilt $\beta_1(\theta) \leq (1/2)^5 = 0.03125$, die Wahrscheinlichkeit eines Type I Errors ist sehr gering!

• Wegen $\beta_1(\theta) < 1/2 \Leftrightarrow \theta^5 < 1/2 \Leftrightarrow \theta < 0.87$ ist die Wahrscheinlichkeit eines Type II Errors, $1 - \beta_1(\theta)$, für $\theta \in (1/2, 0.87)$ größer $1/2$. Sie ist nur kleiner $1/2$ für $\theta > 0.87$.

(2) Ein alternativer Test verwirft falls $X \in \{3, 4, 5\}$. Dieser hat Power

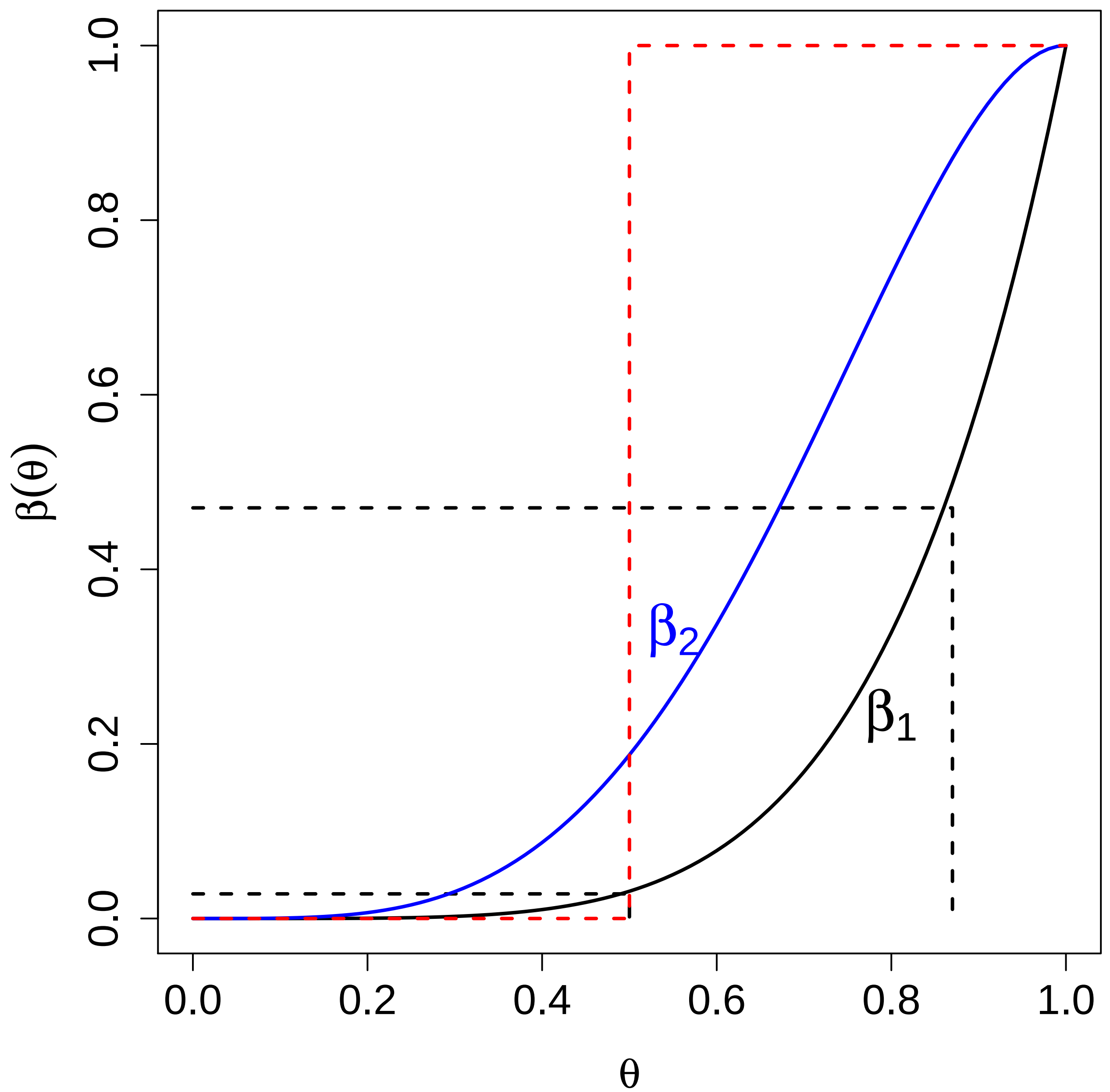
$$\beta_2(\theta) = P_\theta(X \in \{3, 4, 5\}) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta)^1 + \beta_1(\theta)\,.$$

Somit gilt: $\beta_2(\theta) > \beta_1(\theta)$ für alle $\theta \in (0, 1)$.

Also hat Test 2 die geringere Type II Error Wahrscheinlichkeit, da $\beta_2(\theta) > \beta_1(\theta)$ für $\theta > 1/2$.

Jedoch hat dieser eine größere Type I Error Wahrscheinlichkeit, da $\beta_2(\theta) > \beta_1(\theta)$ für $\theta \leq 1/2$.

Wir müssen uns entscheiden, welche Error Struktur, $\beta_1(\theta)$ oder $\beta_2(\theta)$, besser für das vorliegende Problem anwendbar ist.


1.0
0.8
0.6
0.4
0.2
0.0
β(θ)
β2
β1
0.0
0.2
0.4
0.6
0.8
1.0
θ

Beispiel 4.3.2: Seien $X_1, \ldots, X_n$ iid Normal(θ, σ^2), σ^2 bekannt. Teste

$$H_0 : \theta \leq \theta_0 \qquad \text{gegen} \qquad H_1 : \theta > \theta_0 \,.$$

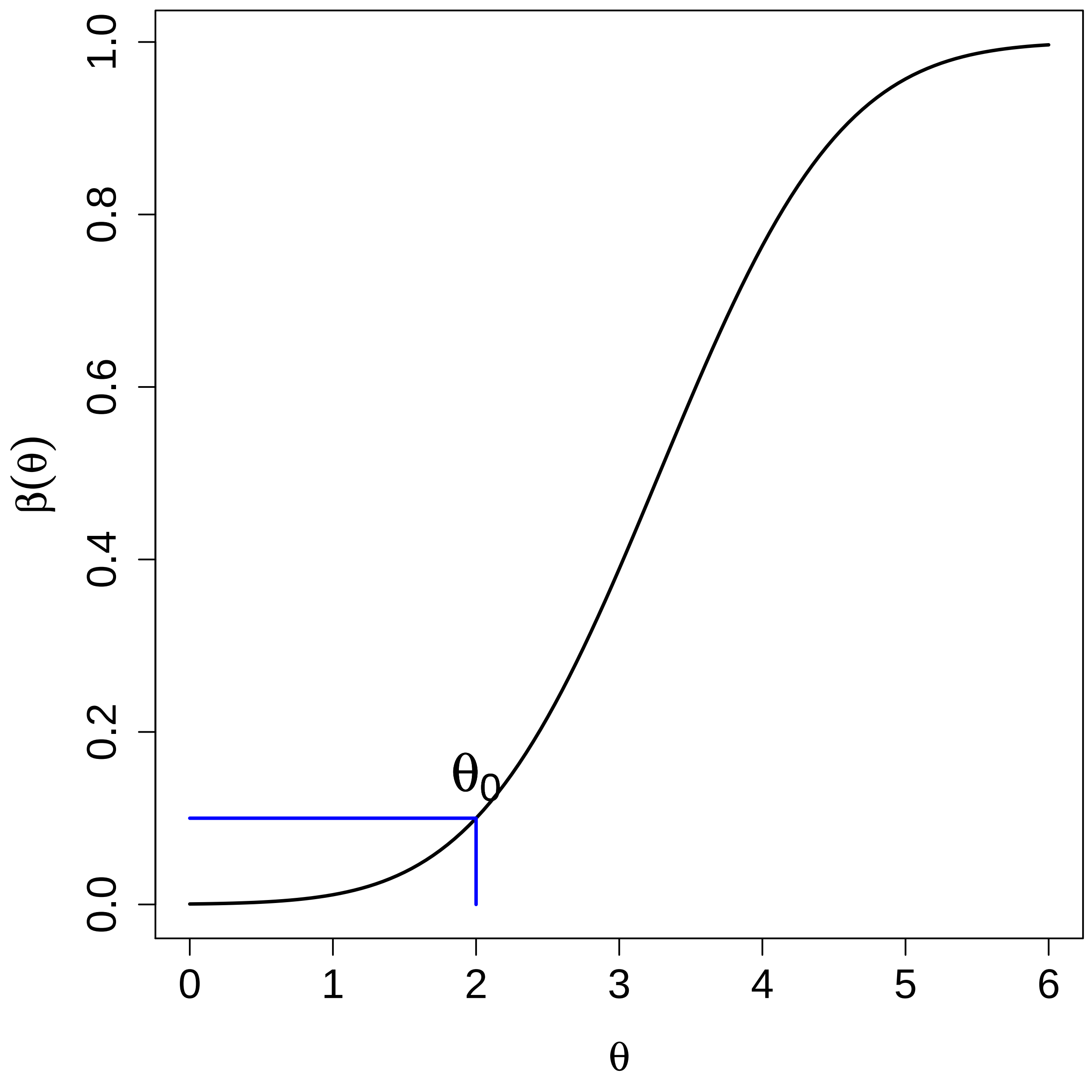
Der LRT verwirft H_0 falls (vgl. Bsp.4.2.3)

$$\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > c$$

mit beliebiger Konstanten $c > 0$. Dieser Test hat Power

$$\begin{aligned}
\beta(\theta) &= P_\theta\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) = P_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} > c\right) \\
&= P_\theta\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \qquad \text{mit } Z \sim \text{Normal}(0,1)
\end{aligned}$$

mit $\beta(\theta_0) = \alpha$ falls $P_{\theta_0}(Z > c) =: \alpha$. Für $\alpha = 0.10$ resultiert $c = 1.28$.



1.0
0.8
0.6
0.4
0.2
0.0
β(θ)
θ0
0
1
2
3
4
5
6
θ

Also ist $\beta(\theta)$ monoton wachsend in θ mit

$$\lim_{\theta\to-\infty} \beta(\theta) = 0\,, \qquad \lim_{\theta\to+\infty} \beta(\theta) = 1\,.$$

Bemerke, dass für **alle** $\theta \leq \theta_0$ gilt: $\beta(\theta) \leq \beta(\theta_0)$. Daher ist die maximale Wahrscheinlichkeit eines Type I Errors gleich $\beta(\theta_0)$.

Falls wir maximale Type I Error Wahrscheinlichkeit von 0.1 wollen und eine Type II Error Wahrscheinlichkeit von maximal 0.2 für alle $\theta \geq \theta_0 + \sigma$, dann müssen c und n entsprechend gewählt werden.

Da $\beta(\theta)$ streng monoton wachsend in θ ist, sind die Forderungen erfüllt für

$$\beta(\theta_0) = 0.10 \qquad \text{und} \qquad \beta(\theta_0 + \sigma) \geq 0.80\,.$$

Mit der Wahl $c = 1.28$ haben wir (unabhängig von n) erreicht, dass

$$\beta(\theta_0) = P_{\theta_0}(Z > 1.28) = 0.1\,.$$

Wir müssen nun n so wählen, dass

$$\beta(\theta_0 + \sigma) = P_\theta\left(Z > 1.28 + \frac{\theta_0 - (\theta_0 + \sigma)}{\sigma/\sqrt{n}}\right) = P_\theta\left(Z > 1.28 - \sqrt{n}\right) \geq 0.8\,.$$

Setzen wir $1.28 - \sqrt{n} = -0.84$, da $P(Z > -0.84) = 0.8$ gilt, so folgt $\sqrt{n} = 2.12$ und damit $n = 4.49$. Also wählen wir schlussendlich $c = 1.28$ und $n \geq 5$.

Für festes n ist es nicht möglich, beide Fehler beliebig klein zu machen. Wir nehmen dann Tests, die die Type I Error Wahrscheinlichkeit auf einem bestimmten Niveau kontrollieren. Aus dieser Klasse von Tests nehmen wir dann jenen Test mit der kleinsten Type II Error Wahrscheinlichkeit.

Definition 4.3.2: Für $0 \le \alpha \le 1$ nennt man den Test mit Power $\beta(\theta)$ einen **size** (Größe) α Test, falls

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha \,.$$

Definition 4.3.3: Für $0 \le \alpha \le 1$ nennt man den Test mit Power $\beta(\theta)$ einen **level** (Niveau) α Test, falls

$$\sup_{\theta \in \Theta_0} \beta(\theta) \le \alpha \,.$$

Für gewöhnlich wird als level eines Tests $\alpha \in \{0.01, 0.05, 0.10\}$ gewählt. Fixiert man α, so kontrolliert man den Type I Error und nicht den Type II Error. Wird α fixiert, so sollten H_0 und H_1 derart spezifiziert sein, dass der wichtigere Fehler der Type I Error ist.

Hat man als Statistiker bereits eine Vermutung und möchte diese nicht aussagen bevor nicht die Daten diese unterstützen, dann sollte die Alternativ-Hypothese diese Vermutung beinhalten (**Research Hypothesis**).

Beispiel 4.3.3: (Size of LRT) Für gewöhnlich wird ein size α LRT so konstruiert (c derart gewählt), dass

$$\sup_{\theta \in \Theta_0} P_\theta(\lambda(\mathbf{X}) \leq c) = \alpha\,.$$

Im Beispiel 4.2.1 ($X \sim$ Normal$(\theta, 1)$, $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$) ist $\Theta_0 = \{\theta_0\}$ und

$$\sqrt{n}(\overline{X} - \theta_0) \overset{H_0}{\sim} \text{Normal}(0,1) \qquad \text{für } \theta = \theta_0 \text{ (unter } H_0\text{)}.$$

Der LRT verwirft H_0, falls $|\overline{X} - \theta_0| \geq c$ und hat Power in θ_0

$$\begin{aligned} \beta(\theta_0) &= P_{\theta_0}(|\overline{X} - \theta_0| \geq c) = 1 - P_{\theta_0}(-c \leq \overline{X} - \theta_0 \leq c) \\ &= 1 - P_{\theta_0}(-c\sqrt{n} \leq Z_0 \leq c\sqrt{n}) \end{aligned}$$

mit $Z_0 = (\overline{X} - \theta_0)\sqrt{n} \overset{H_0}{\sim}$ Normal$(0,1)$.

Nimmt man $c = z_{1-\alpha/2}/\sqrt{n} = -z_{\alpha/2}/\sqrt{n}$, so folgt

$$\beta(\theta_0) = 1 - P_{\theta_0}(z_{\alpha/2} \leq Z_0 \leq z_{1-\alpha/2}) = \alpha\,.$$

Der size α LRT verwirft H_0 falls

$$\sqrt{n}|\overline{X} - \theta_0| \geq z_{1-\alpha/2}\,.$$

Beispiel 4.3.4: (Fortsetzung von Beispiel 4.2.2) Seien $X_1, \ldots, X_n$ iid Exponential$(1, \theta)$, d.h. mit Dichte

$$f(x|\theta) = \exp(-(x-\theta))I_{[\theta,\infty)}(x)\,, \qquad \theta \in \mathbb{R}$$

Wir testen $H_0 : \theta \leq \theta_0$ gegen $H_1 : \theta > \theta_0$. Der LRT verwirft falls $X_{(1)} - \theta_0 \geq c$.

Dieser Test hat Power

$$\beta(\theta) = P_\theta(\text{verwerfe } H_0) = P_\theta(X_{(1)} - \theta \geq c + \theta_0 - \theta) = P_\theta(W \geq c + \theta_0 - \theta)$$

mit $W = X_{(1)} - \theta$. Für die Statistik W folgt

$$P(W > w) = P(X_1 - \theta > w, \ldots, X_n - \theta > w) = \prod_{i=1}^{n} P(X_i - \theta > w)\,.$$

Da $Y_i = X_i - \theta \overset{iid}{\sim}$ Exponential$(1, 0)$, also Dichte $f(y) = \exp(-y)I_{[0,\infty)}(y)$ bzw. Verteilungsfunktion $F_Y(y) = (1 - \exp(-y))I_{[0,\infty)}(y)$ haben, resultiert weiters

$$P(W > w) = (1 - F_Y(w))^n = (\exp(-w))^n = \exp\left(-\frac{w}{1/n}\right)\,,$$

also exponentialverteilt auf $[0, \infty)$ mit Erwartung $1/n$.

Somit ist

$$\beta(\theta) = P_\theta(W \geq c + \theta_0 - \theta) = \exp(-n(c + \theta_0 - \theta))$$

und $\beta(\theta)$ ist **monoton wachsend** in $\theta \in \mathbb{R}$. Damit ist die size dieses LRT gleich

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0)\,.$$

Wir setzen $\alpha = \beta(\theta_0) = \exp(-nc)$ und erhalten somit den Wert c für einen size α Test. Für diesen folgt

$$c = -\tfrac{1}{n} \log \alpha\,.$$

Für den Fall $n = 100$ und $\alpha = 0.05$ liefert dies beispielsweise $c = 0.03$.

Definition 4.3.4: Ein Test mit Power Funktion $\beta(\theta)$ heißt **unbiased** (unverzerrt), falls

$$\beta(\theta_1) \geq \beta(\theta_0)\,,$$

wobei $\theta_0 \in \Theta_0$ und $\theta_1 \in \Theta_0^c$.

Beispiel 4.3.5: (siehe dazu Beispiel 4.3.2) Seien $X_1, \ldots, X_n$ iid Normal(θ, σ^2) mit σ^2 bekannt. Der LRT für $H_0 : \theta \leq \theta_0$ gegen $H_1 : \theta > \theta_0$ hat Power

$$\beta(\theta) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

mit $Z \sim$ Normal$(0, 1)$. Da $\beta(\theta)$ für festes θ_0 monoton wachsend in θ ist, folgt unter $H_1 : \theta > \theta_0$

$$\beta(\theta) > \beta(\theta_0) = \max_{t \leq \theta_0} \beta(t)$$

und der Test ist **unbiased**.

4.3.2 Beste Tests (Most Powerful Tests)

Ein level α Test kontrolliert den Type I Error. Wir suchen aus dieser Klasse von Tests jenen mit geringstem Type II Error (größtes $\beta(\theta)$ unter H_1).

Definition 4.3.5: Sei $\mathcal{C}$ eine Klasse von Tests für $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_0^c$. Ein Test in $\mathcal{C}$ mit Power Funktion $\beta(\theta)$ ist ein **Uniformly Most Powerful (UMP)** (gleichmäßig bester) Klasse $\mathcal{C}$ Test, falls

$$\beta(\theta) \geq \beta'(\theta) \qquad \text{für alle } \theta \in \Theta_0^c$$

und für alle Power Funktionen $\beta'(\theta)$ zu anderen Tests in der Klasse $\mathcal{C}$.

Wir betrachten im Folgenden die Klasse $\mathcal{C}$ der level α Tests und wir sprechen vom UMP level α Test.

Das folgende Neyman-Pearson Lemma erklärt, wie man UMP level α Tests findet.

Motivation: Seien $f(\mathbf{x}|\theta_0)$ und $f(\mathbf{x}|\theta_1)$ zwei Wahrscheinlichkeitsfunktionen und teste $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$ (beides seien **einfache Hypothesen**).

Der beste level α Test (oder der beste level α Verwerfungsbereich) ist derart, dass

$$P_{\theta_0}(\mathbf{X} \in R) = \sum_{\mathbf{x} \in R} f(\mathbf{x}|\theta_0) \leq \alpha \qquad \text{und} \qquad P_{\theta_1}(\mathbf{X} \in R) = \sum_{\mathbf{x} \in R} f(\mathbf{x}|\theta_1) \ \text{ groß.}$$

Welche $\mathbf{x}$ sollen R ausmachen?

Wir wollen $f(\mathbf{x}|\theta_0)$ klein und $f(\mathbf{x}|\theta_1)$ groß für alle $\mathbf{x} \in R$.

Gib $\mathbf{x}$ in R, wenn $r(\mathbf{x}) = f(\mathbf{x}|\theta_1)/f(\mathbf{x}|\theta_0)$ ausreichend groß!

Beispiel 4.3.6: Sei $f(x|\theta_0)$ die Binomial$(4, 1/4)$ und $f(x|\theta_1)$ die Binomial$(4, 3/4)$ Wahrscheinlichkeitsfunktion. Teste $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$.

x	0	1	2	3	4
$f(x\|\theta_1)$	$(1/4)^4$	$4(3/4)(1/4)^3$	$6(3/4)^2(1/4)^2$	$4(3/4)^3(1/4)$	$(3/4)^4$
$f(x\|\theta_0)$	$(3/4)^4$	$4(1/4)(3/4)^3$	$6(1/4)^2(3/4)^2$	$4(1/4)^3(3/4)$	$(1/4)^4$
$r(x)$	1/81	1/9	1	9	81

Unter H_0 haben die Verwerfungsbereiche folgende levels

R	$\{4\}$	$\{3, 4\}$	$\{2, 3, 4\}$
$\alpha = P_{\theta_0}(x \in R)$	0.0039	0.0508	0.2617

- Sei $\alpha = 0.10$. Der **beste level $\boldsymbol{\alpha}$** Test ist einer der H_0 verwirft falls $x \in \{3, 4\}$.
- Die **size** dieses Tests ist aber *nur* 0.0508.
- Wie bekomme man einen Test mit exakter size von $\alpha = 0.10$?

- Ein Test mit exakter size 0.10 ist folgender:

  $x \in \{3, 4\} \Longrightarrow$ verwerfe H_0

  $x \in \{2\} \Longrightarrow$ verwerfe H_0 mit Wahrscheinlichkeit 0.233 (Randomisieren)

  $x \in \{0, 1\} \Longrightarrow$ verwerfe H_0 nicht

Dafür gilt

$$\begin{aligned} P_{\theta_0}(H_0 \text{ verwerfen}) &= 1 \cdot P_{\theta_0}(X \in \{3, 4\}) + 0.233 \cdot P_{\theta_0}(X = 2) \\ &= 0.0508 + 0.233 \cdot 0.2109 = 0.10\,. \end{aligned}$$

Satz 4.3.1: (Neyman-Pearson Lemma) Wir testen die einfachen Hypothesen $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$, wobei die Dichte oder Wahrscheinlichkeitsfunktion $f(\mathbf{x}|\theta_i)$ zu θ_i $(i = 0, 1)$ gehört, und verwenden einen Test mit Verwerfungsbereich R, welcher folgenden Bedingungen genügt:

$$\begin{array}{lll} \mathbf{x} \in R & \text{falls} & f(\mathbf{x}|\theta_1) > k \cdot f(\mathbf{x}|\theta_0) \\ \mathbf{x} \in R^c & \text{falls} & f(\mathbf{x}|\theta_1) < k \cdot f(\mathbf{x}|\theta_0) \end{array} \tag{1}$$

für ein $k \geq 0$, und

$$P_{\theta_0}(\mathbf{X} \in R) = \alpha\,. \tag{2}$$

Dann gilt:

1. (Hinreichend) Ein Test der (1) und (2) erfüllt ist ein UMP level α Test.

2. (Notwendig) Falls ein Test existiert der (1) und (2) mit $k > 0$ genügt, so ist jeder UMP level α Test ein size α Test (genügt (2)) und jeder UMP level α Test genügt (1) bis auf einer Ausnahmemenge mit $P_{\theta_0}(\mathbf{X} \in A) = P_{\theta_1}(\mathbf{X} \in A) = 0$.

Beispiel 4.3.6 (Fortsetzung) Sei $f(x|\theta_0)$ die Binomial$(4, 1/4)$ und $f(x|\theta_1)$ die Binomial$(4, 3/4)$ Wahrscheinlichkeitsfunktion. Teste $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$.

x	0	1	2	3	4
$r(x)$	1/81	1/9	1	9	81

Fall 1: Betrachte den Neyman-Pearson (NP) Test der H_0 verwirft falls

$$f(x|\theta_1) > k \cdot f(x|\theta_0) \qquad \text{mit } k \in (9, 81).$$

Dies ist der Test der H_0 verwirft falls $x = 4$. Mit dem NP Lemma folgt, dass der Test ein UMP level α Test ist mit $\alpha = P_{\theta_0}(\text{verwerfe } H_0) = (1/4)^4 = 0.0039$.

Fall 2: Betrachte den NP Test der H_0 verwirft falls

$$f(x|\theta_1) > 9 \cdot f(x|\theta_0)\,.$$

Mit dem NP Lemma ist dies ein UMP level α Test wobei $\alpha = P_{\theta_0}(\text{verwerfe } H_0)$.

- Aber dies hängt davon ab, was wir machen wollen wenn $x = 3$.

(a) Falls wir H_0 nur dann verwerfen wenn $f(x|\theta_1) > 9 \cdot f(x|\theta_0)$, dann ist dieser Test derselbe wie in Fall 1 und der Test hat size 0.0039.

(b) Falls wir jedoch H_0 verwerfen wenn $f(x|\theta_1) \geq 9 \cdot f(x|\theta_0)$, so folgt mit dem NP Lemma, dass dies ein UMP level α Test ist mit $\alpha = P_{\theta_0}(X \in \{3, 4\}) = 0.0508$. Diesen Test erhält man auch für die Wahl $k \in (1, 9)$.

Für $k < 1/81$ oder $k > 81$ resultieren UMP Tests mit level $\alpha = 1$ oder $\alpha = 0$.

Korollar 4.3.1: Betrachte einen Test von $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$. Sei $T(\mathbf{X})$ eine suffiziente Statistik für θ und $g(t|\theta_i)$ ihre Dichte oder Wahrscheinlichkeitsfunktion unter H_i $(i = 0, 1)$. Ein Test basierend auf T mit Verwerfungsbereich S (ein Unterbereich des Stichprobenraums von T) ist ein UMP level α Test, falls

$$\begin{array}{lll} t \in S & \text{wenn} & g(t|\theta_1) > k \cdot g(t|\theta_0) \\ t \in S^c & \text{wenn} & g(t|\theta_1) < k \cdot g(t|\theta_0) \end{array}$$

für ein $k \geq 0$, und

$$P_{\theta_0}(T(\mathbf{X}) \in S) = \alpha\,.$$

Beweis: Anwendung des Faktorisierungssatz.

Beispiel 4.3.7: (UMP Normal Test) Seien $X_1, \ldots, X_n$ iid Normal(θ, σ^2), mit σ^2 bekannt. Teste $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$ mit $\theta_0 > \theta_1$.

$\overline{X}$ ist suffizient für θ. Wegen $\overline{X} \sim$ Normal$(\theta, \sigma^2/n)$ folgt als kritischer Bereich

$$
\begin{aligned}
g(\overline{x}|\theta_1) &> k \cdot g(\overline{x}|\theta_0) \\
(2\pi\sigma^2/n)^{-\frac{1}{2}} \exp\Big(-\frac{n}{2\sigma^2}(\overline{x}-\theta_1)^2\Big) &> k \cdot (2\pi\sigma^2/n)^{-\frac{1}{2}} \exp\Big(-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2\Big) \\
-\frac{n}{2\sigma^2}(\overline{x}-\theta_1)^2 &> \log k - \frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2 \\
\frac{n}{2\sigma^2}\left[(\overline{x}-\theta_0)^2 - (\overline{x}-\theta_1)^2\right] &> \log k \\
2\overline{x}(\theta_1-\theta_0) - (\theta_1^2-\theta_0^2) &> \frac{2\sigma^2}{n}\log k \\
\overline{x} &< \left[\frac{2\sigma^2}{n}\log k + (\theta_1^2-\theta_0^2)\right]\frac{1}{2(\theta_1-\theta_0)} = c
\end{aligned}
$$

Mit dem Korollar folgt, dass der Test der H_0 verwirft falls $\overline{x} < c$ und H_0 akzeptiert falls $\overline{x} > c$ ein UMP level α Test ist mit

$$\alpha = P_{\theta_0}(\overline{X} < c) = P_{\theta_0}\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} < \frac{c - \theta_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{c - \theta_0}{\sigma/\sqrt{n}}\right) .$$

Will man einen UMP level $\alpha = 0.05$ Test, ergibt sich

$$\frac{c - \theta_0}{\sigma/\sqrt{n}} = z_{0.05} \Longleftrightarrow c = \theta_0 + z_{0.05}\sigma/\sqrt{n} = \theta_0 - z_{0.95}\sigma/\sqrt{n}$$

Bemerkung:

- Konstruktion des kritischen Bereichs wie im Korollar 4.3.1 gefordert.
- Die spezielle Wahl von c sichert, dass auch $P_{\theta_0}(H_0 \text{ verwerfen}) = \alpha$ gilt.

Hypothesen wie H_0 oder H_1 im Neyman-Pearson Lemma, die nur eine mögliche Verteilung für $\mathbf{X}$ spezifizieren, nennt man **einfache Hypothesen**. In der Praxis spezifiziert man jedoch Hypothesen mit mehr als nur eine mögliche Verteilung. Diese werden **zusammengesetzte Hypothesen** genannt.

In der Definition eines UMP Tests wird verlangt, dass dieser der beste Test ist für jedes $\theta \in \Theta_0^c$. Daher kann das Neyman-Pearson Lemma verwendet werden um einen UMP Test für zusammengesetzte Hypothesen zu finden.

Hypothesen, die aussagen dass ein eindimensionaler Parameter groß ist, wie z.B. $H : \theta \geq \theta_0$, oder klein ist, z.B. $H : \theta \leq \theta_0$, nennt man **einseitig**. Hypothesen, die behaupten dass ein Parameter entweder groß oder klein ist, d.h. $H : \theta \neq \theta_0$ nennt man **zweiseitig**.

Eine große Klasse von Problemen die UMP level α tests erlauben umfassen einseitige Hypothesen und Dichte- oder Wahrscheinlichkeitsfunktionen mit der monotonen Likelihood Quotienten Eigenschaft.

Definition 4.3.6: Eine Familie von Dichten oder Wahrscheinlichkeitsfunktionen $\{g(t|\theta) : \theta \in \Theta\}$ für eine univariate Zufallsvariable T hat einen **monotonen Likelihood Quotienten (MLR)**, falls für jedes $\theta_2 > \theta_1$ gilt, dass der Quotient $g(t|\theta_2)/g(t|\theta_1)$ auf dem Bereich $\{t : g(t|\theta_1) > 0 \text{ oder } g(t|\theta_2) > 0\}$ eine monotone (steigende oder fallende) Funktion in t ist.

Beispiel 4.3.8: Sei $T \sim$ Normal$(\theta, 1)$. Für $\theta_2 > \theta_1$ folgt

$$\frac{g(t|\theta_2)}{g(t|\theta_1)} = \frac{(2\pi)^{-1/2}\exp(-(t-\theta_2)^2/2)}{(2\pi)^{-1/2}\exp(-(t-\theta_1)^2/2)} = \exp(t(\theta_2-\theta_1))\exp(-\tfrac{1}{2}(\theta_2^2-\theta_1^2))$$

was wachsend in $t \in \mathbb{R}$ ist. Also hat die Normalverteilung einen MLR.

Tatsache: Falls die Familie $\{f(x|\theta) : \theta \in \Theta\}$ einen steigenden MLR hat, dann ist deren Familie von Verteilungsfunktionen $\{F(x|\theta) : \theta \in \Theta\}$ **stochastisch steigend**, d.h. es gilt $F(x|\theta_2) \leq F(x|\theta_1)$ für alle x und für $\theta_2 > \theta_1$

Satz 4.3.2: (Karlin-Rubin) Wir testen $H_0 : \theta \leq \theta_0$ gegen $H_1 : \theta > \theta_0$. Sei die Statistik T suffizient für θ und habe die Familie von Dichten oder Wahrscheinlichkeitsfunktionen $\{g(t|\theta) : \theta \in \Theta\}$ von T einen MLR. Der Test, der H_0 genau dann verwirft wenn $T > t_0$ (t_0 beliebig), ist ein UMP level α Test, wobei $\alpha = P_{\theta_0}(T > t_0)$.

Bemerkung: Natürlich gilt analog für den Test von $H_0 : \theta \geq \theta_0$ gegen $H_1 : \theta < \theta_0$, der verwirft wenn $T < t_0$, auch der Karlin-Rubin Satz mit $\alpha = P_{\theta_0}(T < t_0)$.

Beispiel 4.3.7: (Fortsetzung) Seien $X_1, \ldots, X_n$ iid Normal(θ, σ^2), mit σ^2 bekannt.

(a) bereits gezeigt (mittels Korollar 4.3.1 zum NP Lemma für eine suffiziente Statistik), dass der Test auf $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$ mit $\theta_0 > \theta_1$ ein UMP level $\alpha = 0.05$ Test ist, falls er verwirft für $\overline{x} < \theta_0 - z_{0.95}\sigma/\sqrt{n}$.

(b) Bemerke $P_{\theta_0}(\overline{X} < \theta_0 - z_{0.95}\sigma/\sqrt{n}) = P_{\theta_0}((\overline{X} - \theta_0)/(\sigma/\sqrt{n}) < -z_{0.95}) = 0.05$.

(c) Wegen $\overline{X} \sim$ Normal$(\theta, \sigma^2/n)$ und weil die Normalverteilung mit σ^2 bekannt MLR hat (vgl. Beispiel 4.3.8), kann der Karlin-Rubin Satz angewendet werden. Damit folgt: Testet man $H_0 : \theta \geq \theta_0$ gegen $H_1 : \theta < \theta_0$ (vgl. mit $\theta_0 > \theta_1$), so ist der Test der H_0 verwirft falls $\overline{x} < \theta_0 - z_{1-\alpha}\sigma/\sqrt{n}$ ein UMP level α Test.

(d) Power Funktion:

$$\begin{aligned}\beta(\theta) &= P_\theta\left(\overline{X} < \theta_0 - z_{0.95}\frac{\sigma}{\sqrt{n}}\right) = P_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} < \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} - z_{0.95}\right) \\ &= P_\theta\left(Z < \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} - z_{0.95}\right)\end{aligned}$$

Die Power $\beta(\theta)$ ist somit monoton fallend in θ, womit für die Type I Error Wahrscheinlichkeit folgt

$$\sup_{\theta \geq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha\,.$$

Beispiel 4.3.9: (Existenz von UMP Tests) Seien $X_1, \ldots, X_n$ iid Normal(θ, σ^2), mit σ^2 bekannt. Teste $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$ (zweiseitige Hypothese).

Für gegebenes α ist ein level α Test für dieses Problem ein Test mit

$$P_{\theta_0}(\text{verwerfe } H_0) \leq \alpha \,.$$

Situation $\boldsymbol{\theta_1 < \theta_0}$: Zuvor wurde im Beispiel 4.3.7 gezeigt, dass der Test der H_0 verwirft falls

$$\overline{X} < \theta_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

die maximale Power in θ_1 hat. Wir nennen diesen **Test 1**.

Teil 2. des Neyman-Pearson Lemmas (Notwendigkeit) sagt, dass ein beliebiger anderer level α Test mit gleich großer Power in θ_1 wie Test 1, auch denselben Verwerfungsbereich wie Test 1 haben muss. Falls daher ein UMP level α Test **existiert**, muss es Test 1 sein, weil kein anderer Test die gleich große Power in θ_1 hat wie Test 1.

Situation $\boldsymbol{\theta_2 > \theta_0}$: Test 2 ist auch ein level α Test und verwirft H_0 falls

$$\overline{X} > \theta_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,.$$

Bezeichne $\beta_i(\theta)$ die Power zu Test i ($i = 1, 2$). Dann gilt

$$\begin{aligned}
\beta_2(\theta_2) &= P_{\theta_2}(\overline{X} > \theta_0 + z_{1-\alpha}\sigma/\sqrt{n}) \\
&= P_{\theta_2}\left(\frac{\overline{X}-\theta_2}{\sigma/\sqrt{n}} > \frac{\theta_0-\theta_2}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \\
&> P_{\theta_2}(Z > z_{1-\alpha}) = P_{\theta_2}(Z < -z_{1-\alpha}) \\
&> P_{\theta_2}\left(\frac{\overline{X}-\theta_2}{\sigma/\sqrt{n}} < \frac{\theta_0-\theta_2}{\sigma/\sqrt{n}} - z_{1-\alpha}\right) \\
&= P_{\theta_2}(\overline{X} < \theta_0 - z_{1-\alpha}\sigma/\sqrt{n}) \\
&= \beta_1(\theta_2)\,.
\end{aligned}$$

Somit ist $\beta_2(\theta_2) > \beta_1(\beta_2)$ und Test 1 kann kein UMP level α Test sein, da Test 2 größere Power in θ_2 hat. Also existiert kein UMP level α Test für dieses Problem.

Was ist nun der **richtige** level α Test für dieses Problem?

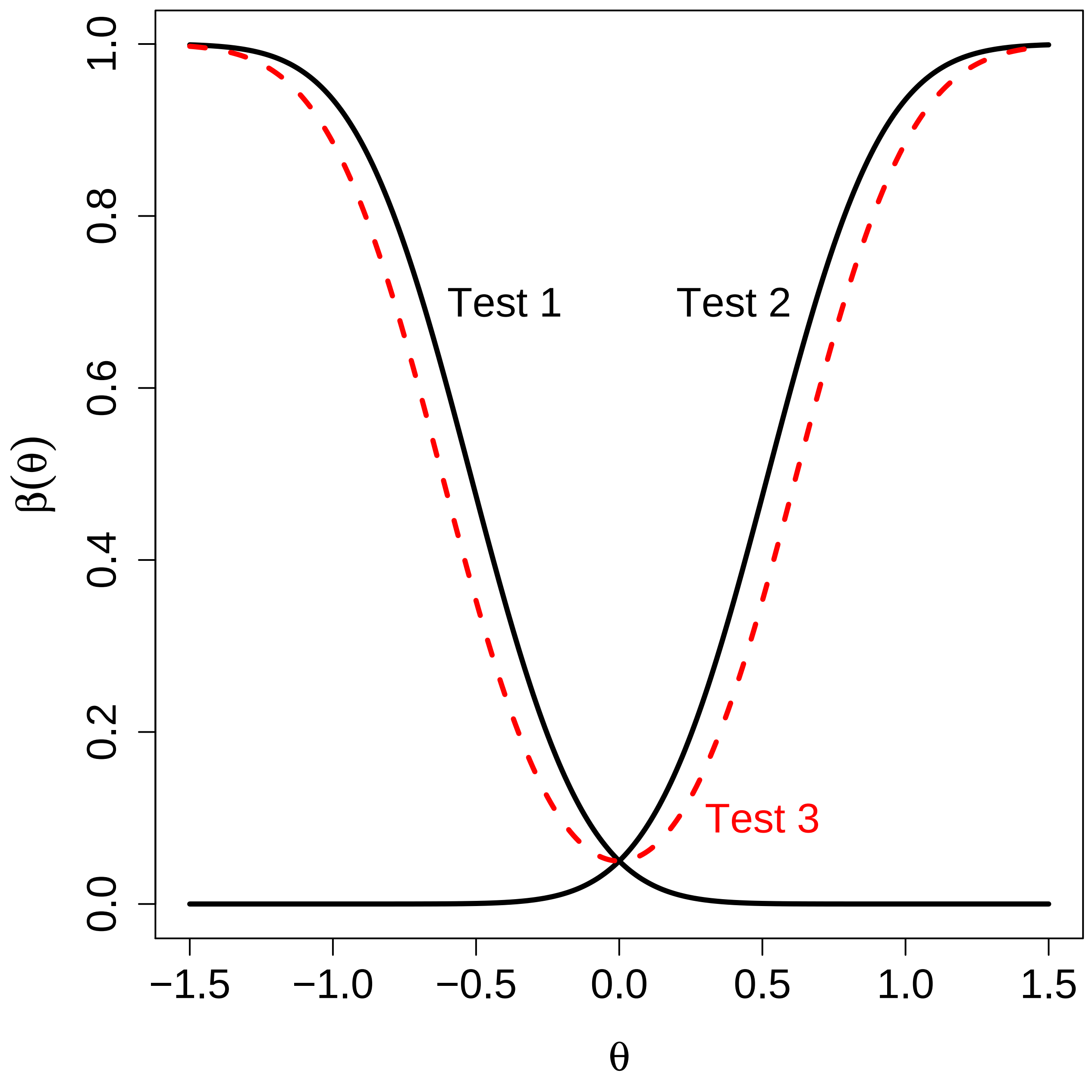
Test 3 verwirft $H_0 : \theta = \theta_0$ zu Gunsten von $H_1 : \theta \neq \theta_0$ falls

$$\overline{X} < \theta_0 - z_{1-\alpha/2}\sigma/\sqrt{n} \qquad \text{oder} \qquad \overline{X} > \theta_0 + z_{1-\alpha/2}\sigma/\sqrt{n}\,.$$

Dieser Test ist UMP in der Klasse aller **unbiased level $\boldsymbol{\alpha}$ Tests** (vgl. Definition 4.3.4).

Obwohl Test 1 und Test 2 größere Power als Test 3 für einige Parameterwerte haben, hat Test 3 größere Power in anderen Stellen. So ist z.B. $\beta_3(\theta_2) \approx 1$ während $\beta_1(\theta_2) \approx 0$.

Falls wir H_0 verwerfen wollen für große und kleine Werte von θ, dann hat Test 3 bessere *overall* Performance als Test 1 und Test 2.


Test 1
Test 2
Test 3
β(θ)
θ
0.0
0.2
0.4
0.6
0.8
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Für Tests liefern Computerprogramme keine logische Entscheidung sondern den p-Wert. Dieser ist die anhand der Stichprobe beobachtete Type I Error Rate.

Satz 4.3.3: (Probability Integral Transformation) Habe X stetige Verteilungsfunktion $F_X(x)$ und sei $Y = F_X(X)$. Dann ist Y gleichverteilt auf $(0, 1)$, d.h.

$$P(Y \le y) = y\,, \qquad 0 < y < 1\,.$$

Beweis:

$$\begin{aligned} P(Y \le y) &= P(F_X(X) \le y) = P(F_X^{-1}(F_X(X)) \le F_X^{-1}(y)) \\ &= P(X \le F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y\,. \end{aligned}$$

Bemerkung: Ist X diskret, so gilt: $P(Y \le y) \le y$, für $0 \le y \le 1$.

Definition 4.3.7: F_X ist **stochastisch größer** als F_Y, falls $F_X(t) \leq F_Y(t)$ für alle t gilt. Für $X \sim F_X$ und $Y \sim F_Y$ folgt $P(X \leq t) = F_X(t) \leq F_Y(t) = P(Y \leq t)$ und für alle t gilt

$$P(X > t) \geq P(Y > t)\,.$$

Nach dem Test wird Ergebnis mitgeteilt. Eine Möglichkeit ist es α und damit die Entscheidung bzgl. H_0 zu berichten. Alternativ kann p-Wert übermittelt werden.

Definition 4.3.8: Der **p-Wert** $p(\mathbf{X})$ ist eine Teststatistik mit $0 \leq p(\mathbf{x}) \leq 1$. Kleine Werte von $p(\mathbf{X})$ weisen auf die Richtigkeit von H_1 hin. Ein p-Wert ist **gültig**, falls für jedes $\theta \in \Theta_0$ und jedes $0 \leq \alpha \leq 1$ gilt

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha\,.$$

Ist $p(\mathbf{X})$ gültig, kann damit ein Level α Test konstruiert werden. Der Test, der H_0 genau dann verwirft wenn $p(\mathbf{X}) \leq \alpha$ ist ein Level α Test.
Wie kann nun ein gültiger p-Wert definiert werden?

Satz 4.3.4 Sei $W(\mathbf{X})$ eine Teststatistik. Große Werte von W sprechen gegen H_0. Definiere für einen beliebigen Stichprobenpunkt $\mathbf{x}$

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x}))\,.$$

Damit ist $p(\mathbf{X})$ ein gültiger p-Wert.

Beweis: Fixiere ein $\theta \in \Theta_0$. Sei dafür $F_\theta(w)$ die cdf von $-W(\mathbf{X})$. Definiere dafür

$$p_\theta(\mathbf{x}) = P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})) = P_\theta(-W(\mathbf{X}) \leq -W(\mathbf{x})) = F_\theta(-W(\mathbf{x}))\,.$$

Für dieses θ entspricht die ZV'e $p_\theta(\mathbf{X})$ dem $F_\theta(-W(\mathbf{X}))$. Mit Satz 4.3.3 (PIT) folgt, dass die Verteilung von $p_\theta(\mathbf{X})$ stochastisch größer oder gleich einer Uniform(0, 1) ist. D.h. für jedes $0 \leq \alpha \leq 1$ gilt $P_\theta(p_\theta(\mathbf{X}) \leq \alpha) \leq \alpha$.

Nun ist der p-Wert definiert über alle $\theta \in \Theta_0$, und es gilt dafür für jedes $\mathbf{x}$

$$p(\mathbf{x}) = \sup_{\theta' \in \Theta_0} p_{\theta'}(\mathbf{x}) \geq p_\theta(\mathbf{x}) \, ,$$

da der größte p-Wert für alle Elemente in Θ_0 zumindest so groß ist als für unseren Wert θ. Somit gilt auch für jedes $\theta \in \Theta_0$ und jedes $0 \leq \alpha \leq 1$

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq P_\theta(p_\theta(\mathbf{X}) \leq \alpha) \leq \alpha$$

und $p(\mathbf{X})$ ist daher ein gültiger p-Wert.

Beispiel 4.3.10: Seien $X_1, \ldots, X_n$ iid aus $N(\mu, \sigma^2)$. Teste H_0: $\mu = \mu_0$ gegen H_1: $\mu \neq \mu_0$.

Der LRT verwirft H_0 für große Werte von $W(\mathbf{X}) = |\overline{X} - \mu_0|/(S/\sqrt{n})$.

Für $\mu = \mu_0$ folgt $(\overline{X} - \mu_0)/(S/\sqrt{n})$ einer t_{n-1}-Verteilung, unabhängig von σ. Deshalb gilt hierfür

$$p(\mathbf{x}) = P_{\theta_0}(W(\mathbf{X}) \geq W(\mathbf{x})) = 2P\Big(T_{n-1} \geq (\overline{x} - \mu_0)/(s/\sqrt{n})\Big)\,.$$