Experimentelle Methoden der Teilchenphysik

Sommersemester 2011/2012

Albert-Ludwigs-Universität Freiburg

Prof. Markus Schumacher

Physikalisches Institut, Westbau, 2. OG Raum 008

Telefon 07621 203 7612

E-Mail: Markus.Schumacher@physik.uni-freiburg.de

Kapitel 15: Hypothesentests

http://terascale.physik.uni-freiburg.de/lehre/Sommersemester%202012

Statistische Hypothesentests: Einführung

Ziel: Vergleich der Beobachtung bzw. Auswertung der Messdaten einer Stichprobe mit Hypothesen
→ Entscheidung welche Hypothese bevorzugt wird, welche Hypothesen verworfen oder behalten werden

a) Vergleich mit Theorie
- Gauss-WDF + Annahme über Mittelwert, Varianz
- Güte der Anpassung von Parametern
- Hinweis auf neues Phänomen

b) Vergleich von zwei Stichproben
- Mittelwert, Varianz
- Form der Verteilung (selbe Grundgesamtheit)

c) Unterscheidung von Hypothesen
- Exponential- oder Gauss-WDF
- linearer oder quadratischer Zusammenhang zwischen Messpaaren
- Diskriminierung von Ereignisklassen (Signal oder Untergrund), Spam-Mail oder “gute Mail”, Teilchensorten: e, μ, π, γ….)

Statistische Hypothesentests: Einführung

Methode:

- Konstruktion einer Größe zur Quantifzierung der Übereinstimmung/ Diskrepanz mit Hypothesen → Testatistik t

- Quantifizierung der Übereinstimmung mittels Wahrscheinlichkeitsdichtefunktion f(t) für Teststatistik

- Entscheidung über Verwerfung der Hypothese oder Auswahl unter Alternativhypothesen

Grundbegriffe für statistische Tests: Hypothesen

Hypothese(n):
klare Aussage(n), die man falsifizieren bzw. unterscheiden kann

Fall (a): eine alllgemein anerkannte Hypothese, die falsifiziert werden soll
→ Nullhypothese H_0

Beispiele:
- Lebensdauer des Teilchens ist τ
- kein Anzeichen für neue Physik
- Theorie und Stichprobe (2 Stichroben) stimmen überein in Mittelwert oder Form der Verteilung

Bem.: oft Aussage (Nullhypothese) = Negation der Aussage, die einen interessiert

Bsp.: Suche nach neuem Teilchen → H_0: nur Untergrund
Lebensdauer $\tau \neq \tau_0$ → H_0: $\tau = \tau_0$

Grundbegriffe für statistische Tests: Hypothesen

Fall (b): mehrere Hypothesen, zwischen denen Unterschieden werden soll

"Standardannahme", Nullhypothese: H_0
Alternativhypothesen: $H_1, H_2, H_3, \ldots$

Beispiele: Lebensdauer $H_0: \tau = 2s$ $H_1: \tau = 1s$ $H_2: \tau > 2s$
Higgssuche: H_0: nur Untergrund H_1:Signal+Untergrund
Signatur im Detektor von: H_0 = Pion, H_1= Myon, H_3=Elektron

Arten von Hypothesen:
x sei ZV und $f(x;\lambda)$ Wahrscheinlichkeitsdichtefunktion
- einfach: wenn f(x) durch Hypothese vollständig fixiert
  entweder kein Parameter oder Parameter λ festgelegt

- zusammengesetzt: wenn mindestens einer der Parameter nicht bekannt ist
  $f(x;\lambda)$ mit λ unbekannt oder λ aus Intervall [a,b]

WDF für Hypothesen werden mit $f(x|H_0)$ und $f(x|H_1)$ bezeichnet

Grundbegriffe für statistische Tests: Teststatistik

Teststatistik $t(x_1, \dots x_n)$: Funktion der Stichprobenwerte $(x_1,..,x_n)$

zur Quantifzierung der Übereinstimmung mit Hypothesen H_i mit Ziel
- Verwerfung von H_0 (H_1)
- Unterscheidung von H_0, H_1, H_2

$t(x_1,..x_n)$ kann Vektor sein z.B. $t_i = x_i$ (nicht sehr nützlich)

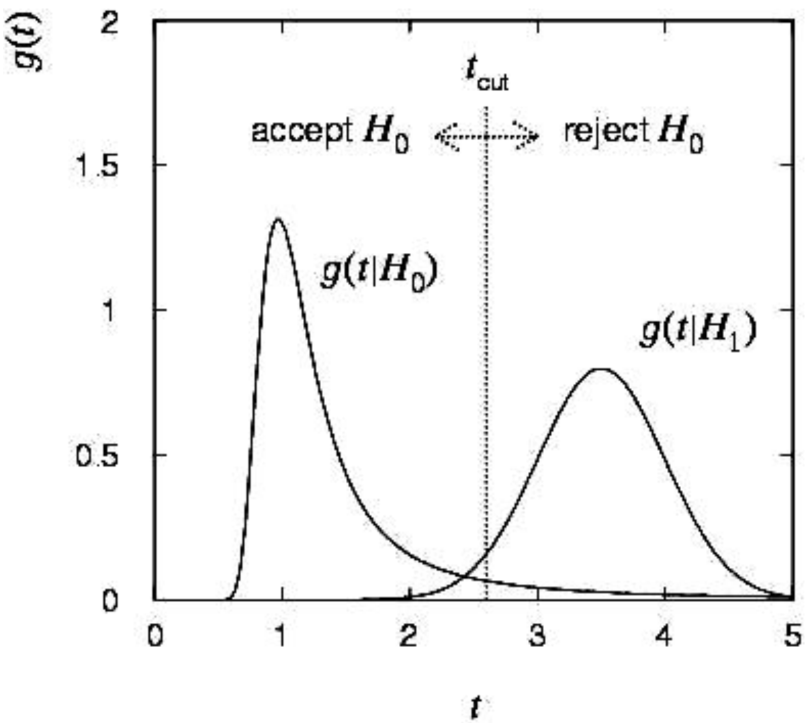
Ziel: Reduzierung der Dimension von t auf skalare Größe t
unter optimaler Ausnutzung der Information in Stichprobe (x_i)
bzgl. der Hypothesen H_i

Aufgaben: a) Definition/Auswahl von Teststatistik t
b) Bestimmung der WDFs für t unter den Hypothesen $f(t|H_i)$
c) Festlegung eines Kriteriums für Verwerfung/Unterscheidung der Hypothesen

Grundbegriffe für stat. Tests: Entscheidungsgrenze

Entscheidungsgrenze t_{cut} oder t_{krit}

$$t(x_1, \ldots, x_n) = t_{\mathrm{cut}}$$



Anname: wir können die WDFs für die Testatistik unter beiden Hypothesen ausrechnen $g(t|H_0),\ g(t|H_1),\ \ldots$

Festlegung der Entscheidungsgrenze (krtischer Wert) t_{krit} oder t_{cut}

kritische Region/Verwerfungregion: $t>t_{krit}$

Verwerfung von H_0, Nicht-Verwerfung (Akzeptanz) von H_1

Komplement zu kritischer Region/Akzeptanzregion: $t<t_{krit}$

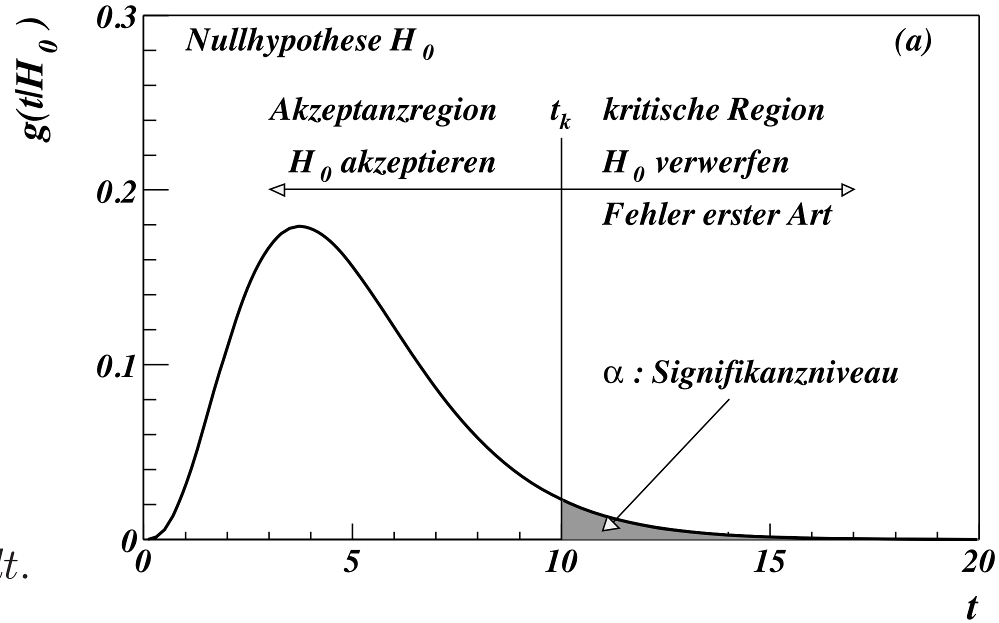
keine Verwerfung von H_0, Verwerfung von H_1

Grundbegriffe: Signifikanzniveau

kritische Region Ω_k

Akzeptanzregion $\Omega - \Omega_k$

$$\alpha = \int_{\Omega_k} g(\vec{t}|H_0)d\vec{t}. \qquad \alpha = \int_{t_k}^{\infty} g(t|H_0)dt.$$



α: Signifikanzniveau, Fehler erster Art
ist Wahrscheinlichkeit H_0 zu verwerfen, obwohl die Hypothese war ist

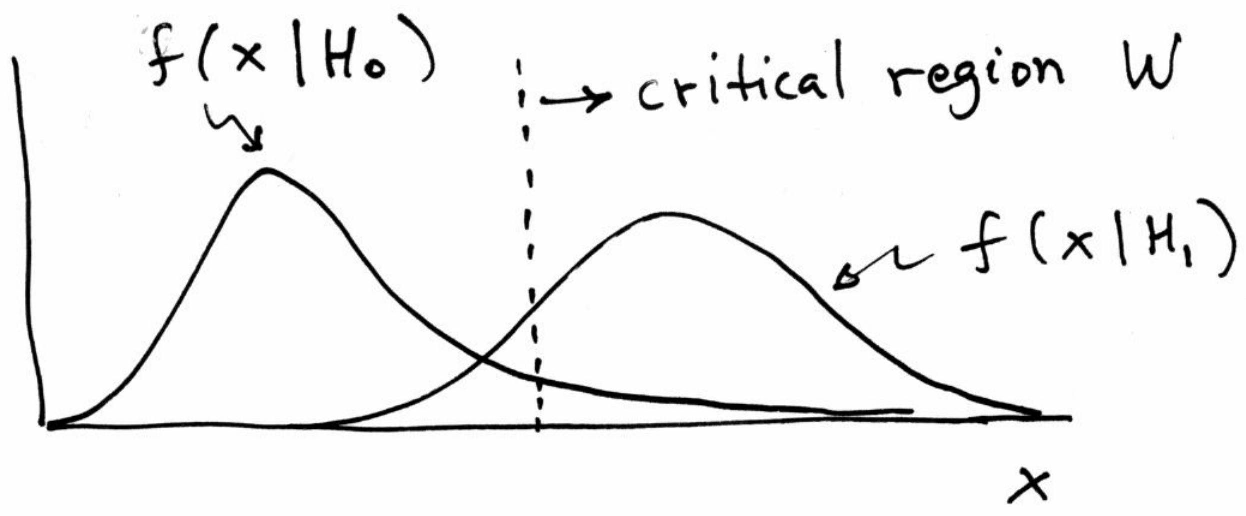
Bemerkung: α (=5%,10%,1%, $2.85\text{x}10^{-7}$) vor Experiment festlegen,
wenn H_0 verworfen/getestet werden soll
α ist keine Zufallsvariable sondern ein festgelegter Wert

Definition der Kritischen Region

Aber im allgmeinen gibt es unendliche viele Möglichkeiten kritische Regionen zu wählen, die alle das gleiches Signifinkanzniveau α besitzen.

Also muss die Wahl der kritischen Region für die Nullhypothese H_0 die Alternativhypothese H_1 berücksichtigen.

Ungefähr ausgedrückt heisst das Kriterium: wähle die kritische Region so, dass es eine kleine Wkt gibt dort eine Messung zu finden, wenn H_0 wahr ist, aber eine große, wenn H_1 ist wahr ist.



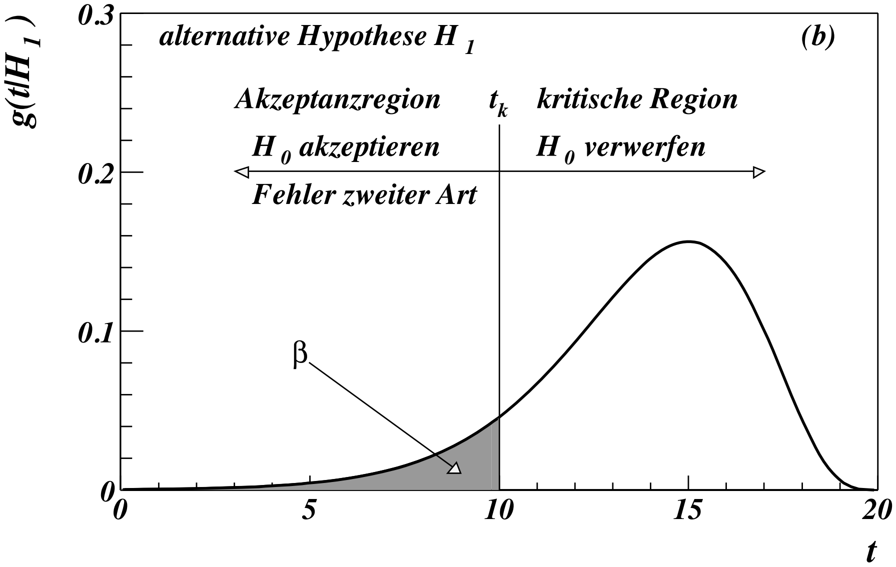
Grundbegriffe: Mächtigkeit

kritische Region Ω_k

Akzeptanzregion $\Omega - \Omega_k$

$$\beta = \int_{\Omega-\Omega_k} g(\vec{t}|H_1)d\vec{t}.$$

$$\beta = \int_{-\infty}^{t_k} g(t|H_1)dt.$$



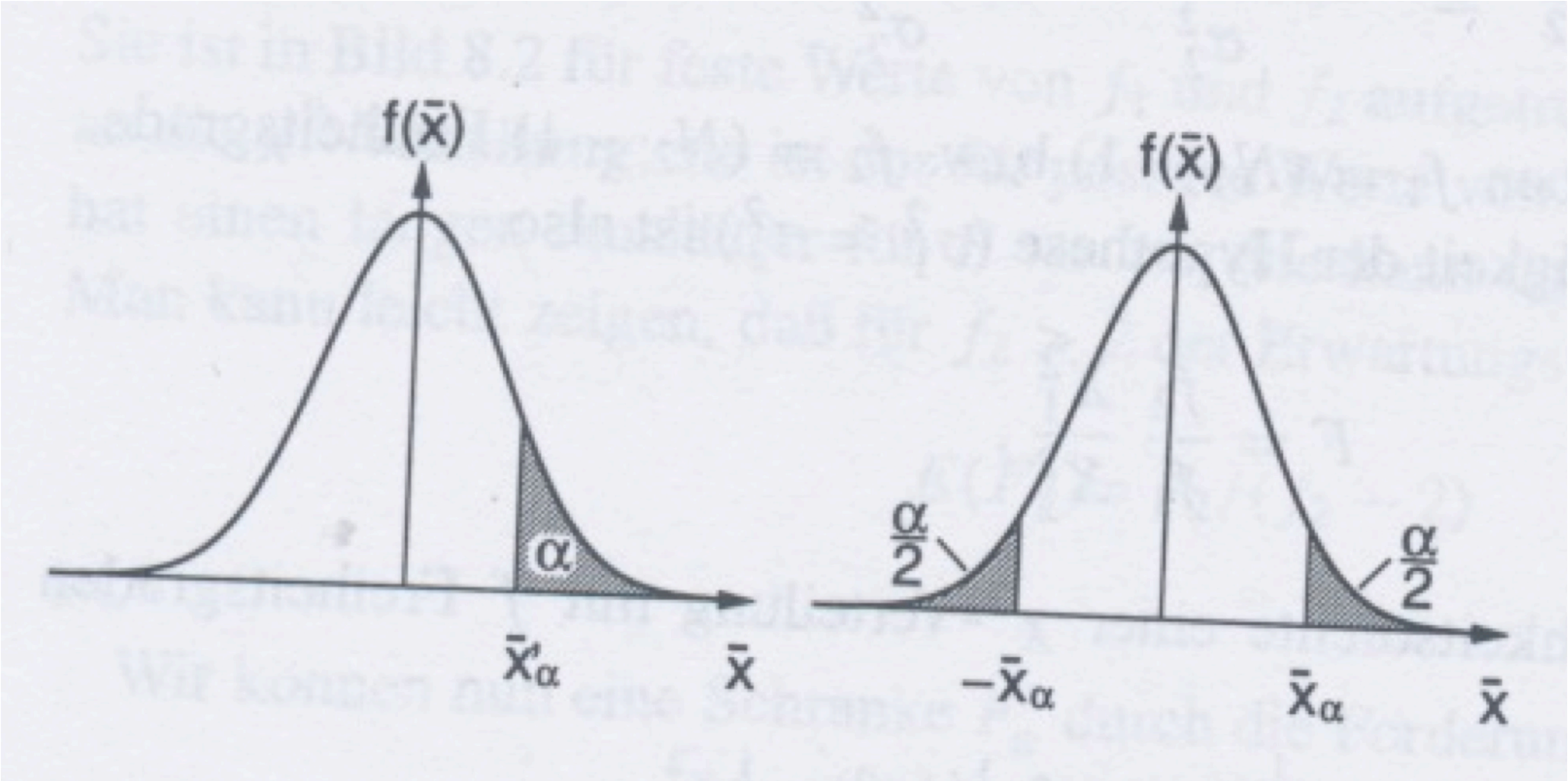
β: Fehler zweiter Art   1-β: Mächtigkeit des Tests

β   ist Wkt H_1 zu verwerfen obwohl die Hypothese war ist

1-β ist Wkt. H_1 zu akzeptieren, wenn Hypothese war ist

Zie: α,β minimieren (ideal = 0) und 1-β maximieren (ideal=1)
nicht gleichzeitig möglich → Kompromiss
Wahl der besten kritischen Region/Teststatistik für gegebenes α

Ein- und zweiseitige Tests



Je nach Problem sind Abweichungen in beide Richtungen interessant
→ dann zweiseitiger Test z.B. Toleranzen in der industriellen Produktion

verteile das Signifikanzniveau i.a. zur Hälfte auf Ausläufer nach oben u. unten

Eigenschaften von Hypothesentests

Gegeben Nullhypothese H_0 und Signifkanzniveau α

a) bester Test bzgl. Alternativhpothese H_1

maximale Mächtigkeit 1-ß unter H_1

b) gleichmäßig bester Test

maximale Mächtigkeit 1-ß unter allen Alternativhypothesen

c) unverzerrter Test

Mächtigkeit 1-ß > α für alle Alternativhpothesen

Illustratives Beispiel

Test der Hypothese, dass eine Gauss-WDF mit bekannter Varianz σ^2 den Mittelwert $\lambda=\lambda_0$ hat.

Stichprobe vom Umfang n (für Illustration =2): $x_1, x_2, \ldots$

Teststatistik: arithmetischer Mittelwert $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$

folgt Gauss-Verteilung mit Mittelwert λ und Varianz σ^2/n

$$f(\bar{x}; \lambda_0) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \lambda_0)^2\right)$$

Wahl von vier kritischen Regionen mit gleichem Signifikanzniveau α

$$\begin{array}{ll}
U_1 : \bar{x} < \lambda^{\mathrm{I}} \text{ und } \bar{x} > \lambda^{\mathrm{II}} & \text{mit } \int_{-\infty}^{\lambda^{\mathrm{I}}} f(\bar{x})\,\mathrm{d}\bar{x} = \int_{\lambda^{\mathrm{II}}}^{\infty} f(\bar{x})\,\mathrm{d}\bar{x} = \frac{1}{2}\alpha \;; \\
U_2 : \bar{x} > \lambda^{\mathrm{III}} & \text{mit } \int_{\lambda^{\mathrm{III}}}^{\infty} f(\bar{x})\,\mathrm{d}\bar{x} = \alpha \;; \\
U_3 : \bar{x} < \lambda^{\mathrm{IV}} & \text{mit } \int_{-\infty}^{\lambda^{\mathrm{IV}}} f(\bar{x})\,\mathrm{d}\bar{x} = \alpha \;; \\
U_4 : \lambda^{\mathrm{V}} \leq \bar{x} < \lambda^{\mathrm{VI}} & \text{mit } \int_{\lambda^{\mathrm{V}}}^{\lambda_0} f(\bar{x})\,\mathrm{d}\bar{x} = \int_{\lambda_0}^{\lambda^{\mathrm{VI}}} f(\bar{x})\,\mathrm{d}\bar{x} = \frac{1}{2}\alpha \;.
\end{array}$$

Illustratives Beispiel

Reihen: 4 verschiedene
kritische Regionen

linke Spalte:
kritische Regionen für n=2
im Stichprobenraum

mittlere Spalten:
WDFs für Teststatistik für H_0 und H_1

$\lambda = \lambda_1 = \lambda_0 + 1$

+ kritische Regionen

rechte Spalte:
Mächtigkeit für
n=2 und n=10

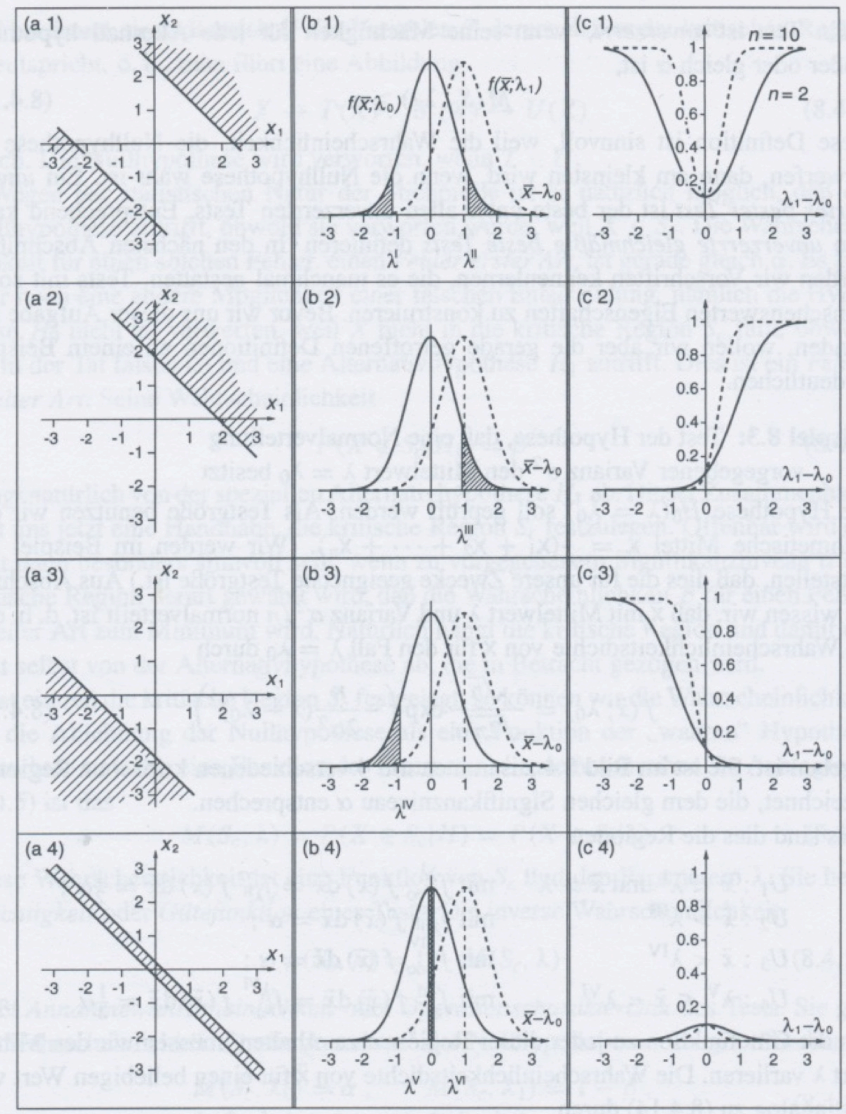


Bild 8.4: Kritische Region im Raum E (a), kritische Region der Testfunktion (b) und Gütefunktion (c) des Tests aus Beispiel 8.3.

Illustratives Beispiel

U1 ist unverzerrter Test

U2: mächtiger für $\lambda_1 > \lambda_0$

U3: mächtiger für $\lambda_1 < \lambda_0$

U4: kein guter Test
maximale Mächtigkeit für
$\lambda_1 = \lambda_0$

keiner von U1 bis U3
ist ein gleichmäßig bester Test

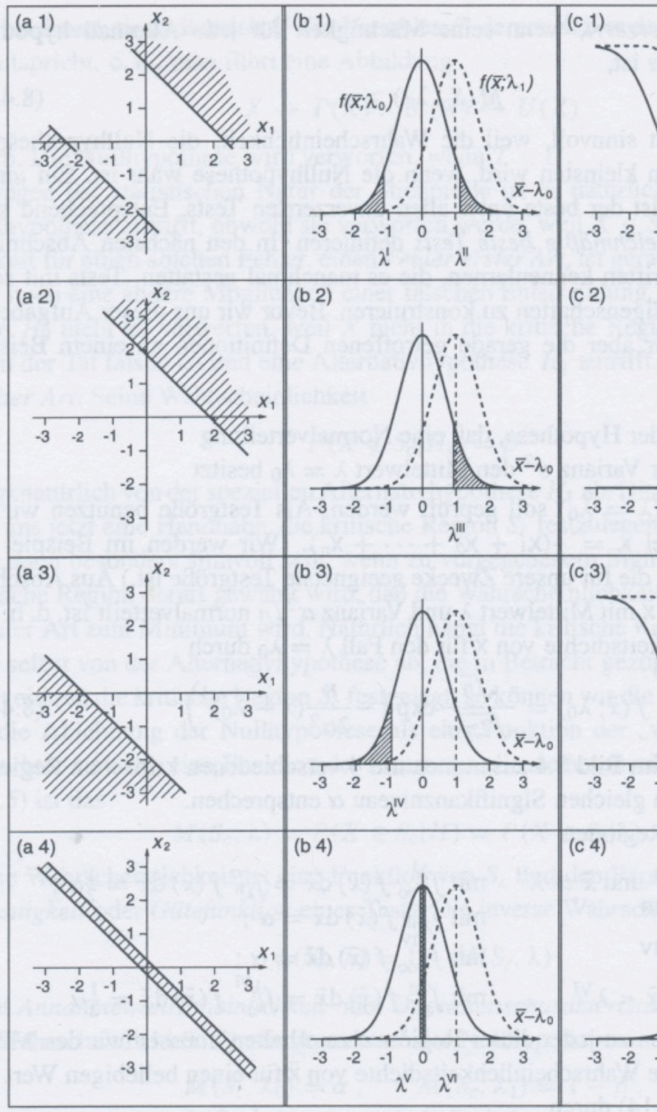
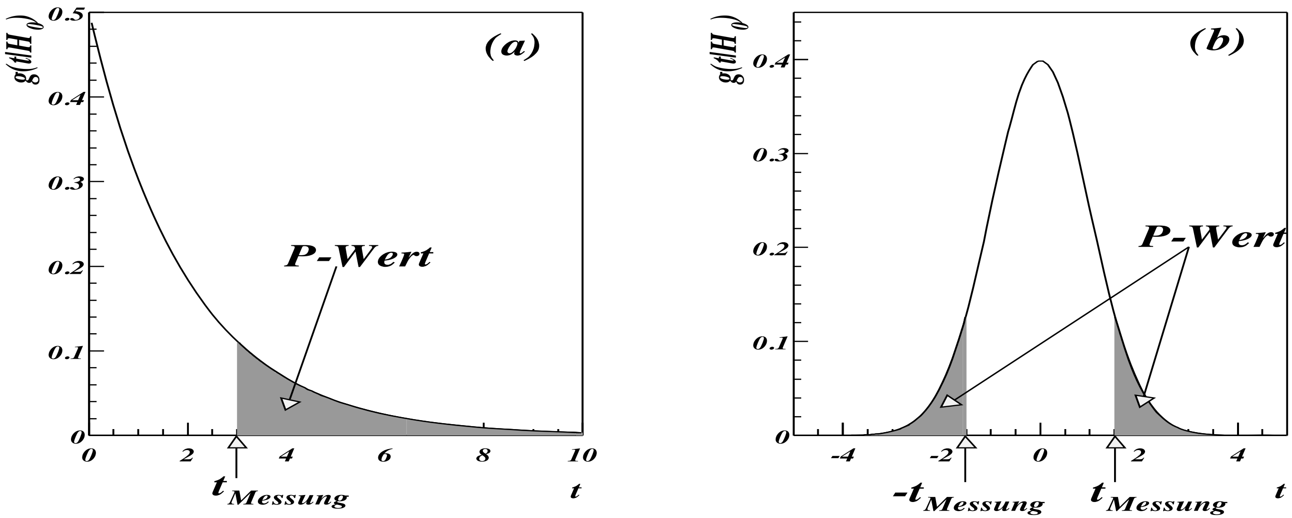


Bild 8.4: Kritische Region im Raum E (a), kritische Region der Testfunktion (b) und Gütefunktion (c) des Tests aus Beispiel 8.3.

Grundlegende Begriffe: P-Werte

P-Wert: Wahrscheinlichkeit eine Stichprobe zu beobachten,
die genauso verträglich oder weniger verträglich mit
der Nullhypothese wie die aktuelle Messung/Beobachtung



t=0 für perfekte Übereinstimmung zwischen Daten und H_0

links: einseitiger P-Wert   rechts: zweiseitiger P-Wert

Grundlegende Begriffe: Bemerkungen zum P-Wert

P-Wert ist Zufallsvariable (vgl.: Signifkanzniveau α vor Messung fixiert)

Wenn P-Wert = Signifkanzniveau α, dann gilt $t_{Messung} = t_{kritisch}$

P-Wert auch beobachtetes Signifkanzniveau genannt

1-P-Wert = Vetrauensniveau des Tests ("confidence level")
P-Wert = 5% dann ausgeschlossen mit 95% Vetrauensniveau

wenn P-Wert < Signifkanzniveau α, dann Hypothese H_0 verwerfen

Achtung vor Fehlinterpretationen:
P-Wert ist nicht Wkt., dass H_0 falsch ist
1-P-Wert ist nicht Wkt, dass H_0 wahr ist

Ein einfaches Beispiel

Bewertung der Fairness eines Würfels
Stichprobe aus n Münzwürfen
n_K= Anzahl des Auftretens von “Kopf”
$n_Z=n-n_K$ = Anzahl des Auftretens von “Zahl”

H_0: Münze ist fair, d.h. $p_K = 0{,}5$ $p_Z = 0{,}5$
Signifkanzniveau auf $\alpha = 5\%$ fixiert

Teststatistik: $t = n_K$ folgt einer Binomialverteilung

Annahme: n=20 (fix) , Beobachtung $n_K=17$ $E[n_K] = 10$

zweiseitiger P-Wert = $\Sigma_0^3\, f_{Bin}(n_K;20) + \Sigma_{17}^{20}\, f_{Bin}(n_K;20) = 0.26\%$

→ Nullhypothese “fairer Würfel” verwerfen

Achtung: Sei ZV $n_{Versuche}$ bis 3x “Zahl” und 3 x “Kopf” erscheint.
$n_{Versuche} = 20$ $P_{H0}=0.072\%$ “optional stopping”-Problem