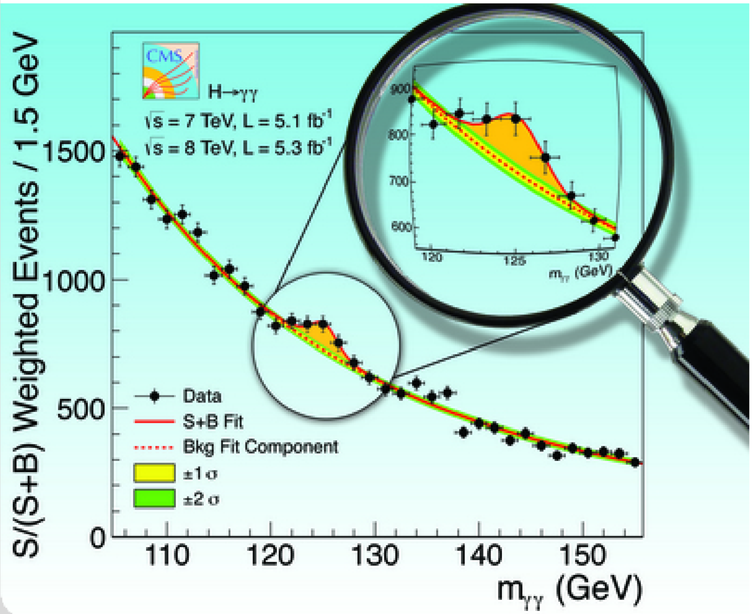
Karlsruher Institut für Technologie

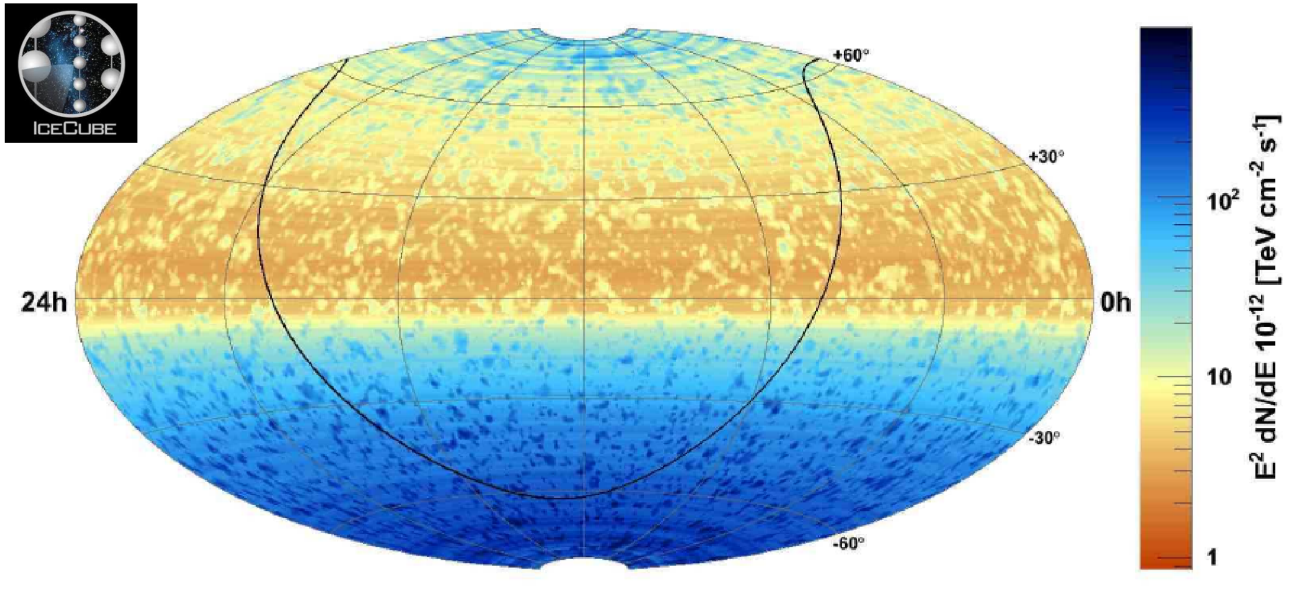
Moderne Methoden der Datenanalyse

Ralf Ulrich, **Andreas Meyer**

Institut für Kernphysik, Institut für Experimentelle Kernphysik

Master-Kurs SS 2017



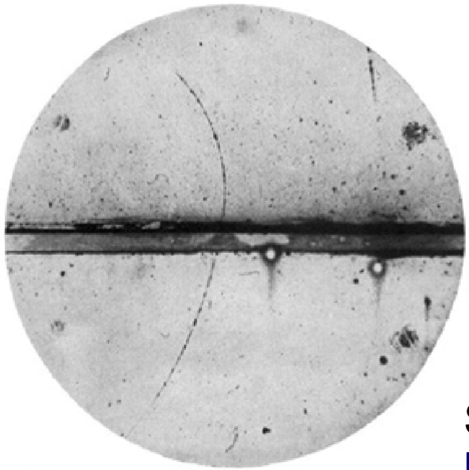


Themen, Volesungsprogramm

- Einführung: Überblick und grundlegende Konzepte (1)
- Einführung: Überblick und grundlegende Konzepte (2)
- Zufallszahlen und Monte-Carlo Methoden
- **Hypothesentests (1)** ← Heute
- **Parameterschätzung**
- **Parameterschätzung (Goodness-of-fit)**
- **Optimierungs- und Parametrisierungsmethoden**
- Konfidenzintervalle
- Hypothesentests (2)
- Klassifikation
- Klassifikation
- Klassifikation
- Messen und Entfalten
- Systematische Unsicherheiten

Fällen von Entscheidungen

- Beispiele:
  - Ist das beobachtete Teilchen ein Pion?
  - Ist die gemessene Zerfallszeitverteilung einer radioaktiven Substanz eine Exponentialverteilung?
  - Existiert „*neue Physik*"?



C. Anderson 1932

Welche Hypothesen könnten die Beobachtung im Prinzip erklären, und sind diese konsistent mit den Daten.

Science 76 (1932) 238
http://inspirehep.net/record/47921

- Statistische Formulierung in Form von Hypothesen: H_i

Wahrscheinlichkeitsdichten für Daten: $f(x|H_i)$

Hypothesen

Einfache Hypothese spezifiziert Wahrscheinlichkeitsdichte vollständig:

$f\,(x \mid H(\lambda_i)\,)$, alle λ_i bekannt

z.B. „Daten folgen einer Poisson-Verteilung mit ν=3,5"

Zusammengesetzte Hypothese spezifiziert WD bis auf einige, aus den Daten zu bestimmende Parameter:

$f\,(x \mid H(\lambda_i\,;\,\mu_j\,)\,)$, λ_i bekannt

z.B. „Daten folgen einer Gaussverteilung mit μ=0, aber unbekanntem σ"

- Zu testende Hypothese: **Null-Hypothese H_0** („Alles beim Alten", „Standardmodell")
- Andere Hypothesen: **Alternativhypothese(n): H_1, H_2, …** (neuer Effekt, neue Physik)

  für konkreten Test muss diese explizit formuliert werden:
  - „Daten folgen einer Poisson-Verteilung mit ν=3,0" (nicht 3,5)
  - „Meßwerte folgen einer Poisson-Verteilung mit ν < 7,0
  - „Meßwerte folgen einer Gaußverteilung mit μ=1" (nicht 0)
  - „Daten sind nicht Poisson-verteilt" (schwieriger, da im Prinzip unendlich viele …)

Frage: Ist die Hypothese gegenüber einer Alternativen zu Bevorzugen?

Antwort eines statistischen Tests kann JA oder NEIN
zu Gunsten von H_1 oder H_0 lauten

$\Rightarrow$ – 4 Fälle
– 2 Fehlermöglichkeiten
I. Verwerfen einer wahren Hypothese („false negative")
II. Akzeptieren einer falschen Hypothese („false positive")

Wahrscheinlichkeit für I. heißt **Signfikanz, α** des Tests

Definiere **„Kritische Region"**
für fälschliches Verwerfen von H_0 :

$$\alpha = P(x \in S_c \mid H_0) \quad \text{(Fehler I. Art)}$$

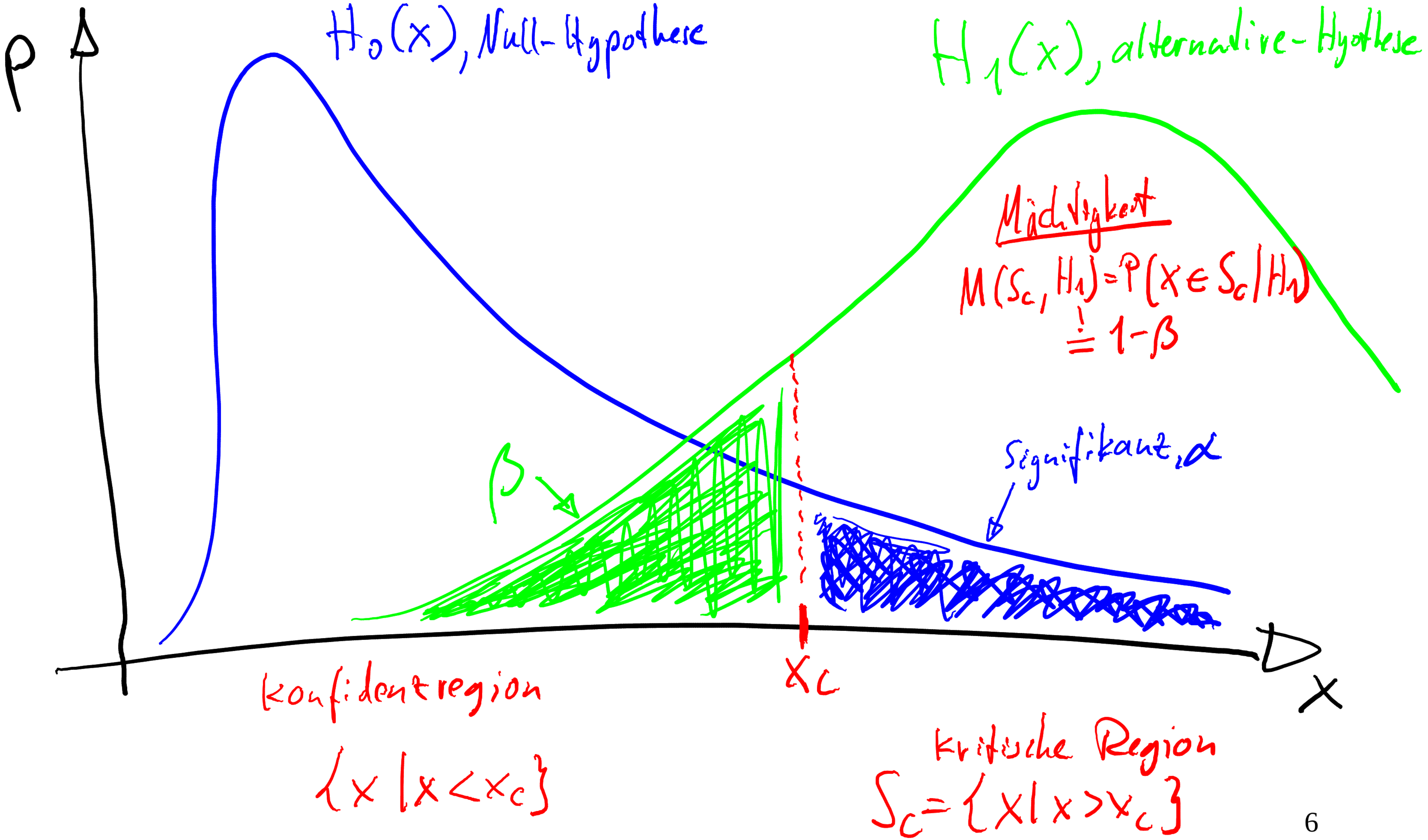
bzw. fälschliches Akzeptieren von H0

$$\beta = P(x \notin S_c \mid H_1) \quad \text{(Fehler II. Art)}$$

sehr häufig: S_c gegeben durch $\{ x \mid x>x_c \}$: einfacher „Schnitt" in x

(1 - β) heißt **Mächtigkeit** (auch „Power", Güte, Trennschärfe) des Tests
Wahrscheinlichkeit, keinen Fehler der zweiten Art zu begehen,
also eine falsche Null-Hypothese auch als solche zu erkennen

Entscheidung zwischen zwei Hypothesen



Mächtigkeit/Güte ist wesentlich wichtiger als Signifikanz

Bsp: Test mit beliebig hoher Signifikanz, aber Mächtigkeit 0

- behaupten Sie irgend etwas
- ziehen Sie eine Zufallszahl $0 < r < 1$;
- akzeptieren Sie Ihre Behauptung als wahr, wenn $r < \alpha$

da Ihre Teststatistik r nicht mit Ihrer Behauptung zusammenhängt, kann Sie auch nicht falsifiziert werden !

Signifikanz und Mächtigkeit bei Klassifizierung: Effizienz und Reinheit

Frage: Ausgewähltes Objekt ist vom Typ 0 oder 1
, (z.B. Pion oder Kaon, Positron oder Proton)

Entscheidung durch Vergleich von beobachteter Meßgröße t („Teststatistik“) mit Verteilungen $f(t|H0)$ und $f(t|H1)$.

Auch dies ist also statistisch gesehen ein Hypothesentest.

Effizienz: Anteil der richtig ausgewählten Objekte vom Typ 0
$\varepsilon = 1-\alpha$

Reinheit: Anteil der Objekte vom Typ 0 unter den ausgewählten

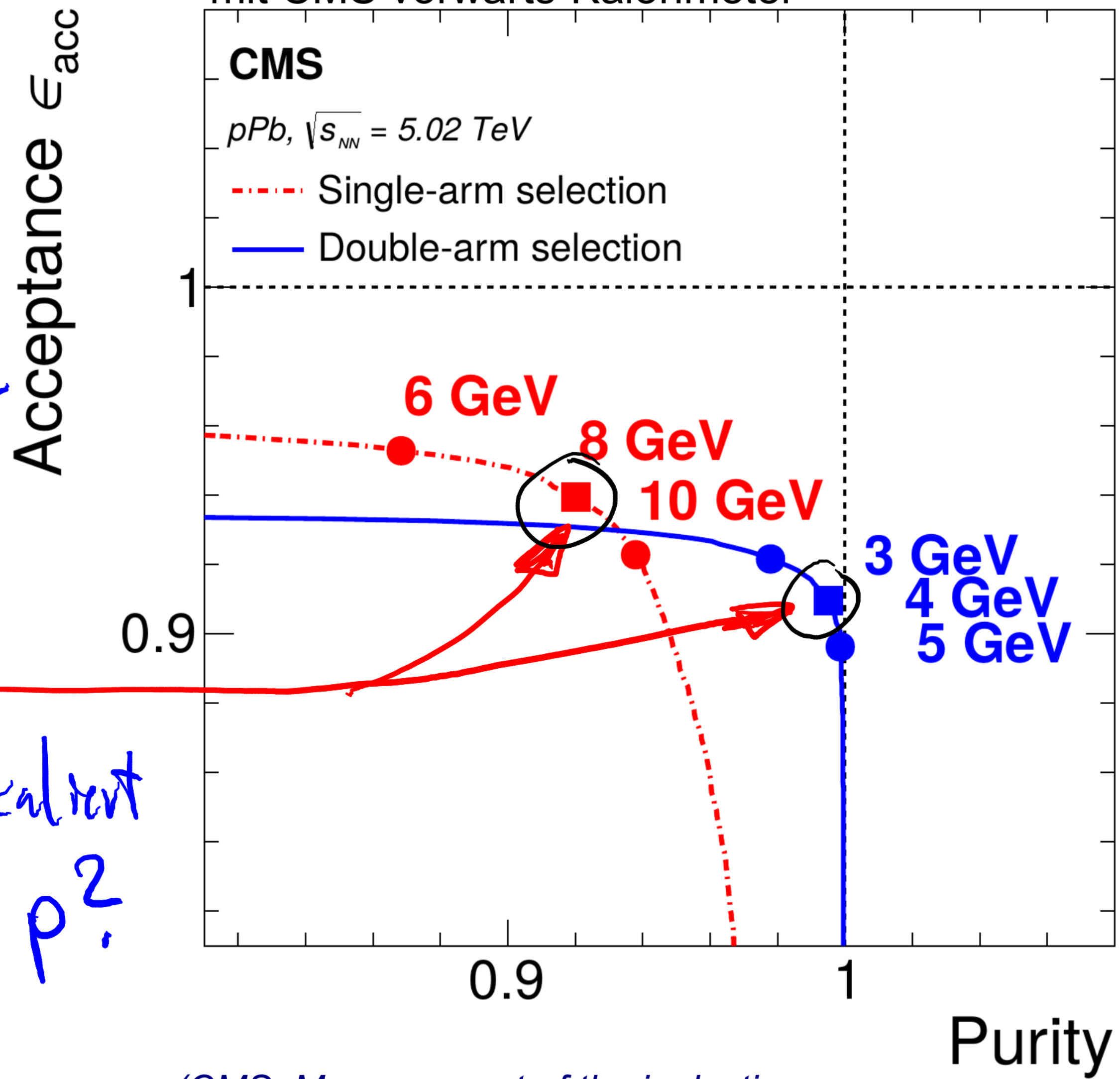
$p = (1 - \alpha) N_0 \; / \; [(1 - \alpha) N_0 + \beta N_1]$
hängt von den Anzahlen der Objekte von Typ 0 und 1, N_0 und N_1 ab

„ROC"-Kurve: **Effizienz und Reinheit**

(Receiver Operating Characteristics)

- Ein statistischer Test ist um so besser je näher er an den Punkt (1,1) kommt
- ... je größer die Fläche unter der ROC-Kurve ist
- Optimale Wahl des Arbeitspunktes je nach Aufgabe: Wie skaliert Gesamtfehler mit ε und p?

Beispiel: inelastische Ereignis-Selektion mit CMS vorwärts-Kalorimeter



(CMS; Measurement of the inelastic cross section in proton–lead collisions at 5.02 TeV; Physics Letters B 759 (2016) 641)

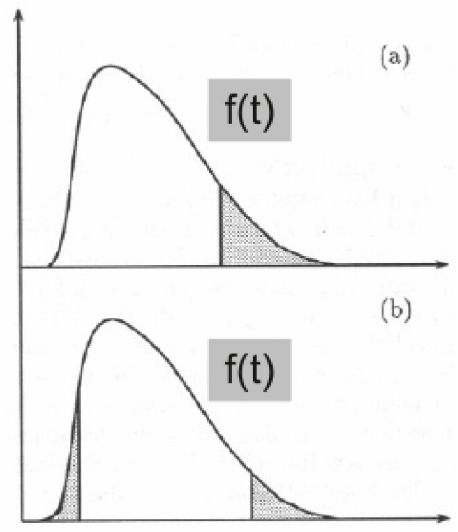
Hypothesentests in mehreren Dimensionen

Reduktion der Dimensionalität durch sog. **Teststatistik**

- $\mathbf{x} = (x_1, \dots x_n) \rightarrow t(\mathbf{x}),\ f(\mathbf{x}|H) \rightarrow g(t|H)$
  Beispiele: Mittelwert, Varianz, Likelihood-Fkt., …
  (Fischer-Diskriminante, künstliches neuronales Netz, ...)

→ Definition von Entscheidungskriterien anhand der Teststatistik

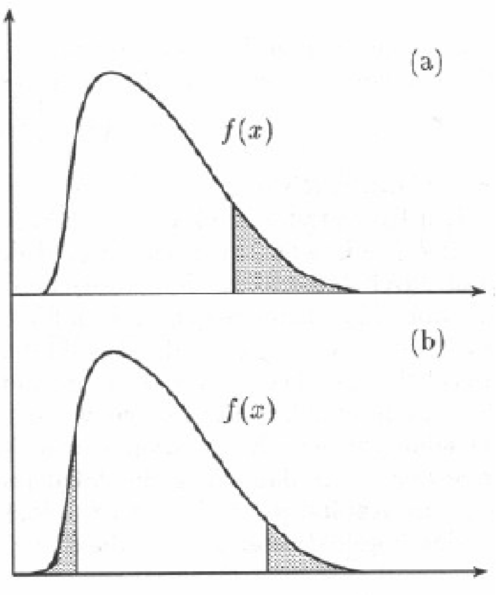
Festlegung des Signifikanzniveaus α und Bestimmung der kritischen Region $S_c(\alpha)$, d.h. des Ablehnungsbereichs der Nullhypothese.



Beispiele zur Wahl der kritischen Region

Vorgehensweise bei Hypothesentests

- Formulierung der Nullhypothese H_0 und der Alternativhypothese H_1
- Wahl einer geeigneten Testgröße oder Teststatistik
- Festlegung des Signifikanzniveaus α
- Bestimmung der kritischen Region $S_c(\alpha)$,
  d.h. des Ablehnungsbereichs der Nullhypothese
  Achtung: alles andere als eindeutig:
  einseitig, zweiseitig symmetrisch, ...



- erst dann: Ziehung der statistischen Stichprobe -
  d.h. Anschauen der Messdaten!
- Treffen der Testentscheidung und Interpretation:

  Liegt das Ergebnis der Stichprobe innerhalb S_c
  wird H0 verworfen, andernfalls angenommen.

Test einer Hypothese

- **Akzeptieren der Null-Hypothese führt zur Ablehnung der Alternativhypothese**
- **Akzeptieren der Alternativhypothese führt zur Ablehnung der Null-Hypothese**

aber: Null - Hypothese kann nicht bewiesen werden !

es könnte eine Alternativ-Hypothese existieren, die die Beobachtung noch besser beschreibt als die (gegenwärtige) Null-Hypothese!

Oft statt vorheriger Wahl von α :

Angabe von beobachtetem p-Wert („p-Value") bzgl. H_0

z.B. bei einfacher einseitiger kritischer Region: $\alpha_{\mathrm{obs}} = \int_{x_{\mathrm{obs}}}^{\infty} f(x)\, dx$

als Maß für die Übereinstimmung (oder auch Nicht-Übereinstimmung) einer Beobachtung mit der Null-Hypothese.

Allg. **p-Value**: Geringste Irrtumswahrscheinlichkeit α , deren Festlegung mit den beobachteten Stichprobenwerten gerade noch zum Verwerfen der Nullhypothese geführt hätte.

Test einer Hypothese:

Hauptprobleme:

1. Finden der optimalen Teststatistik – gibt es einen besten Test ?

2. Bestimmen der Verteilungsdichte der Teststatistik, f(t)
   Machmal analystisch möglich, oft mit Hilfe von Monte-Carlo

3. Festlegen der kritischen Region (bzw. der Konfidenzregion)

4. Interpretation des Ergebnisses:
   – Frequentist p-Value bzgl. P(Daten | Hypothese)
   oder
   – Bayes`sche Sichtweise ? P(Hypothese | Daten) benötigt Prior-Annahme

Beginnen zunächst mit einigen Beispielen

Binomial Test

$$\boxed{P(k;n) = \binom{n}{k} p^k (1-p)^{n-k} \quad ,k = 1,\ldots,n}$$

Beispiel:

- 15 Münzwürfe
- Daten: k = Anzahl Kopf
- Null-Hypothese: p = ½ (Münze ok)
- Alternative: p > ½ (Münze gezinkt)
- Gewählte Signifikanz: 10%

n=15

k	p-Value
15	0.003%
14	0.05%
13	0.3%
12	1.4%
11	4.2%
10	9.2%

Ein Experiment, Anzahl Kopf, k, als Teststatistik.

Mit einer Wahrscheinlichkeit von ca. 10% findet man bis zu 10 Mal Kopf.
Ab 11 mal Kopf würde man mit einer Signifikanz von 10% die Null-Hypothese verwerfen – die Münze also als gezinkt bezeichnen.

Effektivität einer med. Behandlung

Ist eine medizinische Behandlung effektiv?

100 Patienten, spontane Heilung 60 %

Entscheidung: 5 % Signifikanz-Niveau

H_0 : kein Effekt

Binomial-Verteilung: $\sigma = \sqrt{Np(1-p)} = \sqrt{100 \cdot 0{,}6 \cdot 0{,}4} = 4{,}90$

N „groß" → Gauß-Näherung

Einseitiger Test mit $\alpha = 5\%$: $1{,}64 \cdot \sigma$ (Gauß)

$\Rightarrow 60 + 1{,}64 \cdot 4{,}90 = 69$ Heilungen mindestens zur Verwerfung von H_0

Poisson Test

Sehr typisches Problem bei der Suche nach neuen Effekten:

Erwartung: Anzahl von μ_0 Ereignissen aus bekannten Quellen,

Hypothese: $P(k;\, \mu_0)$

Zusätzliche Anzahl μ_1 Ereignisse durch neuen Effekt,

Alternative: $P(k;\, \mu_1+\mu_0)$

Im Experiment beobachtete Zahl von Ereignissen, k, als Teststatistik.

Falls n groß → Gauß'sche Näherung

- $f(k|H_0) = \text{Gauß}(\mu = \boldsymbol{\mu_0},\ \sigma = \sqrt{\boldsymbol{\mu_0}}\ )$ bzw.

  $f(k|H_1) = \text{Gauß}(\mu = \boldsymbol{\mu_0}+\boldsymbol{\mu_1},\ \sigma = \sqrt{(\boldsymbol{\mu_0}+\boldsymbol{\mu_1})})$

- Einfaches Maß für die „Signifikanz" einer Beobachtung:

  „Signifikanz" $S = s / \sqrt{b}$, $s = n - \boldsymbol{\mu_0}$, $b = \boldsymbol{\mu_0}$

Beobachteter Signalüberschuss normiert auf den Fehler der Untergrunderwartung

Häufiges Problem:

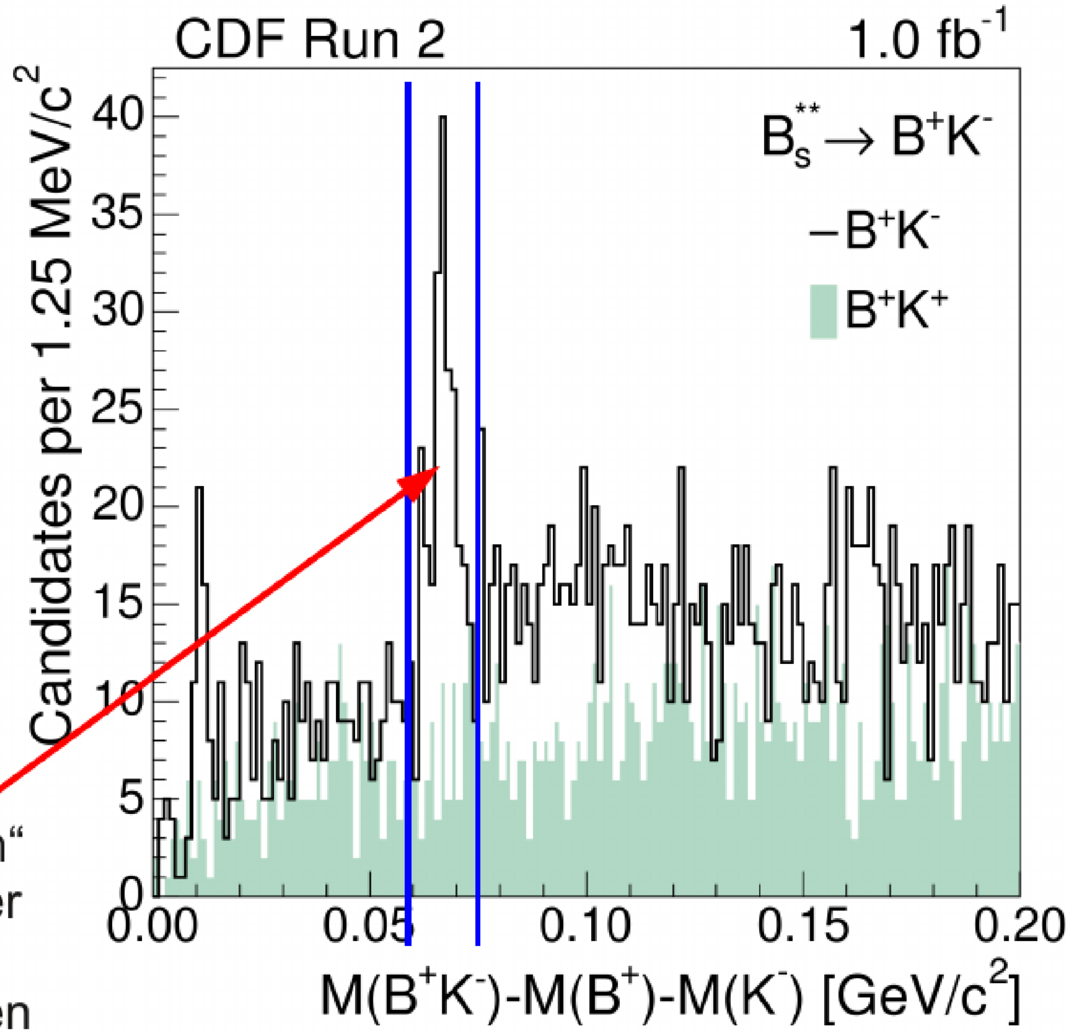
systematischer Fehler, $\boldsymbol{\Delta\mu_0}$, auf $\boldsymbol{\mu_0}$, erfordert spezielle Maßnahmen

Angeregte Bs Mesonen

- Grün unterlegte Verteilung modelliert Form des Untergrundes
- Histogramm enthält mögliche Signalereignisse.

Frage: wie signifikant sind die einzelnen Signale, wie viele gibt es überhaupt, oder sieht man nur statistische Fluktuationen?

Typische Vorgehensweise: Definition einer „Signalregion“ um einen Peak herum, in der die Zahl der beobachteten Ereignisse und der erwarteten Untergrundereignisse bestimmt werden, dann **Poisson-Test.**



„Neue Physik“ in 750GeV Di-Photonen

Gibt es ein Signal?

CMS + ATLAS haben in 2015-Daten je >3σ Signale bei 750 GeV beobachtet!

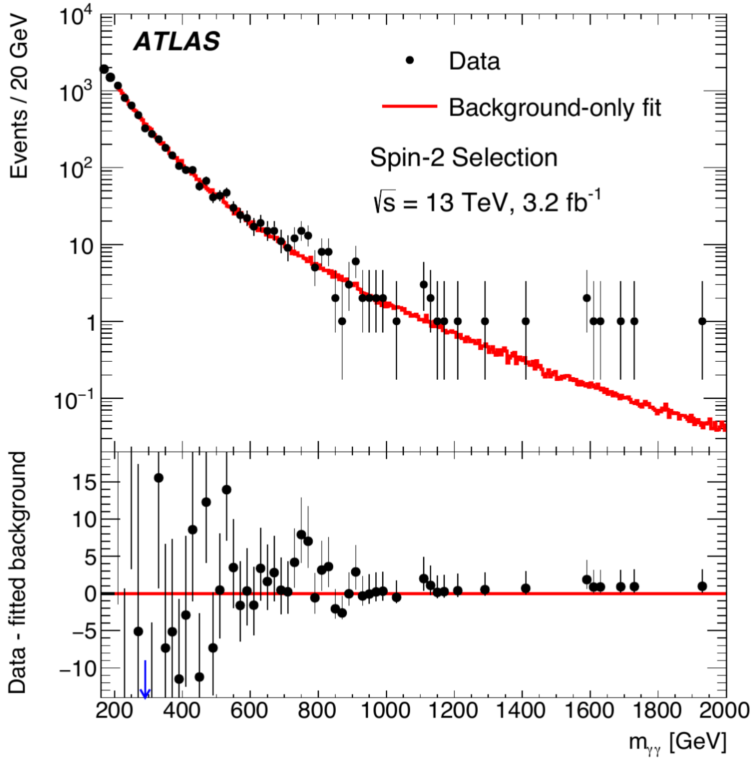
ATLAS (3,9σ)

$$N_{signal} = 15 = N_{obs} - N_{bg}$$

$$\sigma = \frac{N_{sig}}{3,9} = 3,8 = \sqrt{N_{bg}}$$

$$\Rightarrow N_{bg} \approx 14$$

$$H_0: N = N_{bg} \pm \sqrt{N_{bg}}$$ (Poisson / Gauß)



Zum Beispiel:
https://arxiv.org/pdf/1606.03833.pdf
https://inspirehep.net/record/1480039

(leider verschwand der Effekt in 2016-Daten)

„Signale“ und ihre Signifikanz

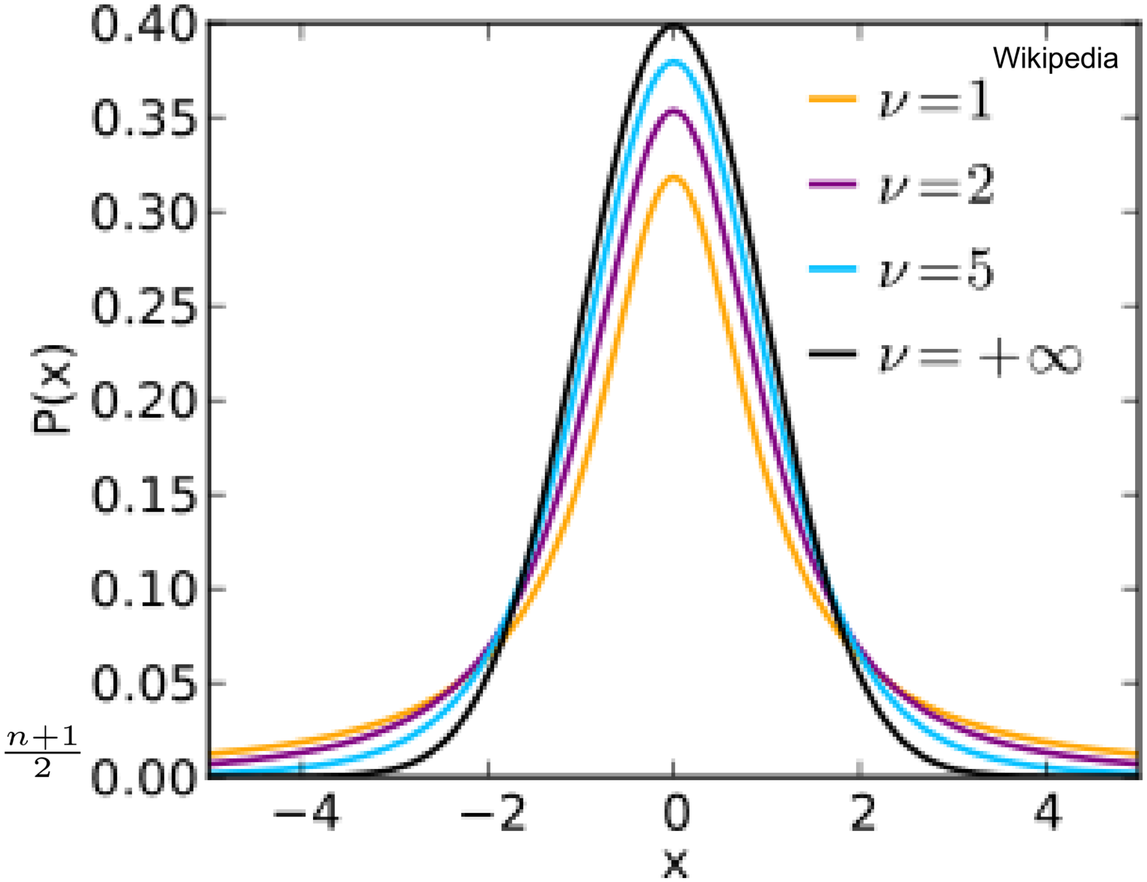
- Signifikanz nur richtig falls schon vor der Messung ein Signal vermutet wurde

- Signifikanz viel geringer falls:
  - sie irgendwo in der Verteilung auftreten kann
  - viele Verteilungen betrachtet werden
  - die Selektionskriterien (bewusst oder unbewusst) gewaehlt wurden um Signal zu erzeugen

- **Lösung z.B. mit Hilfe von**
  - **Blinden Analysen**
    - festlegen aller Kriterien bevor Daten angeschaut werden
  - **Pseudo-Experimente**
    - Künstliches Nachstellen der Daten (sehr oft!) um zu „messen“ wie wahrscheinlich die Beobachtung wirklich ist.

Student'sche t-Verteilung

- Betrachte Gauß-förmig verteile Zufallszahlen x_i mit Erwartungswert μ und Varianz σ^2
- Mittelwert $\overline{x}$ als Testgröße
- Varianz aus Daten: $\hat{\sigma}$
  → X^2-Summe
- $$t = \frac{\overline{x} - \mu}{\hat{\sigma}}$$
- folgt t-Verteilung
  mit n Freiheitsgraden



$$f_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

(n→ unendlich: Gauß, n=1→ Breit-Wiegner/Cauchy)

- t-Verteilung hat für kleiner n deutlich ausgeprägtere Schwänze als Gauß.

Test auf gleichen Mittelwert bei unbekannter Varianz

Zwei Gauß-verteilte Datensätze mit Mittelwerten μ_1, μ_2 und empirischen Varianzen $\hat{\sigma_1}^2, \hat{\sigma_2}^2$

$$u := \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma_1{}^2/N_1 + \sigma_2{}^2/N_2}}$$

ist standard-normalverteilt

$$\chi^2 := \frac{(N_1 - 1)\hat{\sigma_1}^2}{\sigma_1{}^2} + \frac{(N_2 - 1)\hat{\sigma_2}^2}{\sigma_2{}^2}$$

ist χ^2-Summe mit N_1+N_2-2 Freiheitsgraden

Verhältnis ist t-verteilt !

$$\text{mit } \sigma_1 = \sigma_2 \text{ und } \mu_1 = \mu_2 \text{ ergibt sich :}$$

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{1/N_1 + 1/N_2}} \text{ mit } S^2 = \frac{(N_1 - 1)\hat{\sigma_1}^2 + (N_2 - 1)\hat{\sigma_2}^2}{N_1 + N_2 - 2}$$

Diese Teststatistik aus zwei Datensätzen folgt einer t-Verteilung mit N_1+N_2-2 Freiheitsgraden (für die Nullhypotese: „Mittelwerte sind gleich“)

Beispiel, Student'sch t-Verteilung

- Punkte bei Klausur in zwei Gruppen
  - Gruppe 1: 39 18 8 22 24 29 22 22 27 28 23 48
  - Gruppe 2: 42 23 36 35 28 42 33

$\bar{x}_1 = 25{,}4$ $\quad \hat{\sigma}_1 = 10{,}9$ $\quad N_1 = 12$

$\bar{x}_2 = 35{,}6$ $\quad \hat{\sigma}_2 = 6{,}5$ $\quad N_2 = 7$

$H_0: \mu_1 = \mu_2$

$S^2 = 91{,}788$

$$\frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{10{,}2}{9{,}6 \cdot 0{,}48} = 2{,}21$$

95 % C.L. $(1-\alpha)$ für t-Verteilung mit 17 Freiheitsgraden ist 2,110

$\Rightarrow$ Gruppen sind unterschiedlich!

Test auf gleiche Varianz

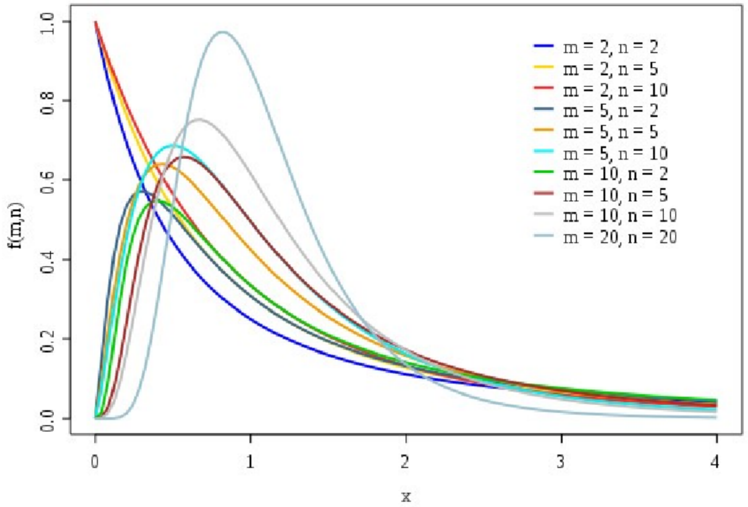
- Fischer-, F-Verteilung

$$f(F|m,n) = m^{\frac{m}{2}} n^{\frac{n}{2}} \cdot \frac{\Gamma(\frac{m}{2}+\frac{n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}$$

- Der Größe

$$F_{m,n} = \frac{\chi^2_m/m}{\chi^2_n/n}$$

zweier X^2-verteilter Zufallsvariablen mit m und n Freiheitsgraden



Verteilungsdichte der F-Verteilung, f(x;n,m)
Quelle: Wikipedia

- Für große N ist **z:= 0.5 log F**
  Gauß-verteilt mit Mittelwert 0.5(1/n-1/m) und Varianz 0.5(1/m + 1/n)

Bester Test, Optimale kritische Region

Bsp: Hypothese: Normalverteilung mit σ^2 und $\lambda=\lambda_0$

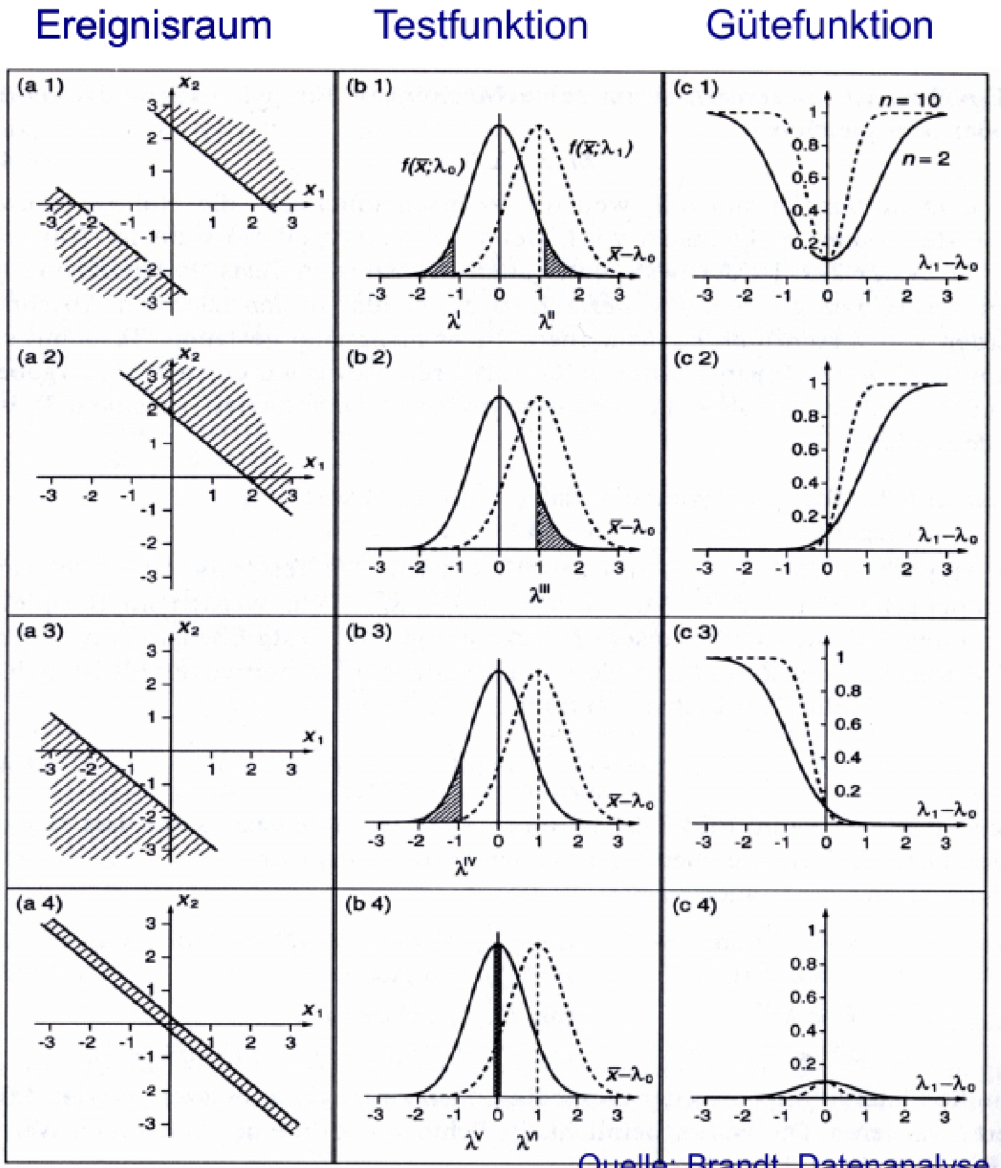
Bsp.: Test einer Gauß-Vert. auf Erwartungswert $\lambda_0=0$ für verschiedene kritische Regionen
Alternativ-Hypothese: Erwartungswert ist λ_1

Teststatistik: Mittelwert von Gauß-verteilten Zuf.-Zahlen x_i

Erwartungswert λ der Alternativ-Hypothese wird variiert, um Mächtigkeit (=Güte, 1-β) des jeweiligen Tests zu bestimmen.

(a1): $1-\beta \geq \alpha$ („unverzerrter Test")
(a2): am besten für $\lambda_1 \geq \lambda_0$
(a3): am besten für $\lambda_0 \geq \lambda_1$
(a4): Verwerfungswahrscheinlichkeit maximal, wenn H0 wahr ist
(Unsinn, nur zur Demonstration!)

⇒ **Es existiert kein „bester Test" für alle Alternativhypothesen!**



Quelle: Brandt, Datenanalyse

Bester Test einer einfachen Hypothese

- Was ist der (gleichmäßig) <u>beste Test</u> einer Hypothese $H_0(x)$ gegen Alternativhypothese $H_1(x)$

  $M(S_c, \boldsymbol{\lambda}_1) = 1 - \beta = \max$ (gleichmäßig: für <u>alle</u> Alternativhypothesen $\boldsymbol{\lambda}_i$)

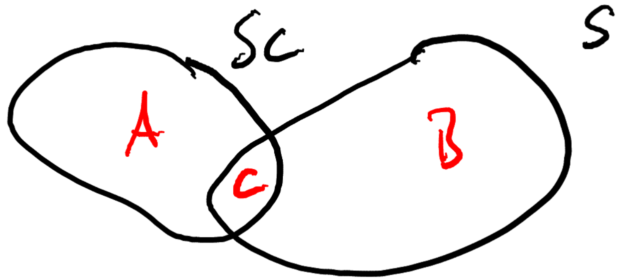
  mit $$\int_{S_c} f(X|H_0)\mathrm{d}X = P(X \in S_c|H_0) = \alpha$$

- **Satz von Neyman und Pearson:**

$$\frac{f(X|H_1)}{f(X|H_0)} \geq c(\alpha) \quad \Longleftrightarrow \quad X \in S_c$$

wobei c eine Konstante ist, die vom Signifikanzniveau abhängt.

Beweis (siehe Brandt)



$$P(x \in S \mid H_0) = P(x \in S_c \mid H_0) \overset{!}{=} \alpha$$

$$P(X \in A \mid H_0) = P(X \in S_c \mid H_0) - P(X \in C \mid H_0)$$
$$= P(X \in S \mid H_0) - P(x \in C \mid H_0) = P(X \in B \mid H_0)$$

$$\Rightarrow \begin{cases} \leq c \cdot P(x \in A \mid H_1) \\ > c \cdot P(x \in B \mid H_1) \end{cases}$$

$$M(S_c, H_1) = P(x \in S_c \mid H_1) \ldots \geq P(x \in S \mid H_1) = M(S, H_1) \quad \blacksquare$$

Beispiel, Edelsteinabbau (siehe Barlow)

$\rho_{opal} = 2{,}2\ g/cm^3$ $\quad\rho_{Quartz} = 2{,}6\ g/cm^3$

Testen kleiner Proben verschiedener Mienen

$\Delta\rho = 0{,}2\ g/cm^3$ Meßfehler

H_0: Opal Gauß mit $\mu = 2{,}2$ $\sigma = 0{,}2$

H_1: Quartz Gauß mit $\mu = 2{,}6$ $\sigma = 0{,}2$

$$c = \frac{H_1}{H_0} = \frac{e^{-(\rho-2{,}6)^2/2\Delta\rho^2}}{e^{-(\rho-2{,}2)^2/2\Delta\rho^2}} \propto e^{10\cdot\rho}$$

$\Rightarrow$ jeder einfache Schnitt $\rho > \rho_c$ definiert ein eindeutiges c und damit ein optimales $1-\beta$ für gegebenes α

z.B.:

$\alpha = 5\%$ $\quad\rho_c = 2{,}53$ $\quad\beta = 36\%$

gute Mienen verworfen

Quartz abgebaut

$\alpha = 10\%$ $\quad\rho_c = 2{,}46$ $\quad\beta = 24\%$

Gefahr von Verzerrungen, auch bei hoher Signifikanz

Beispiel:

- 20 Physiker führen (unabhängig voneinander) jeweils eine Messung durch
- Einer sieht eine Abweichung von der Erwartung um 2σ (Ausschluss der Null-Hypothese mit 5% Signifikanz)
- Der eine publiziert sein Ergebnis, die anderen nicht

➔ Bias der veröffentlichten Ergebnisse!

Publikation sollte nicht vom Ausgang des Tests abhängen

- ➢ Auch „negative" Resultate publizieren

Denn: LHC $\approx$ 10000 Physiker, $\mathcal{O}(10^4)$–$\mathcal{O}(10^5)$ Messungen/Verteilungen/Schnitte
$\Rightarrow$ auch extrem seltene Fluktuationen können signifikant werden!

Zusammenfassung

- Mit Hilfe von Hypothesen könne Daten verwendet werden um definierte Entscheidungen zu treffen
- Es kann dabei immer zu Fehlern kommen: Fehler 1. und 2. Art.
- Ein Hypothese kann nie „bewiesen“ werden, sondern immer nur Alternativ-Hypothesen verworfen werden (d.h. das Standardmodell kann nie bewiesen werden)
- Die Signifikanz von Tests ist oft schwer zu beurteilen: das Verwenden von blinden Analysen, oder Pseudo-Experimenten ist sehr empfohlen.

- Gleichmäßig beste Tests können mit Hilfe des Neyman-Pearson Lemmas definiert werden.
- Auch nicht-Signifikante Effekte sollten publik gemacht werden.