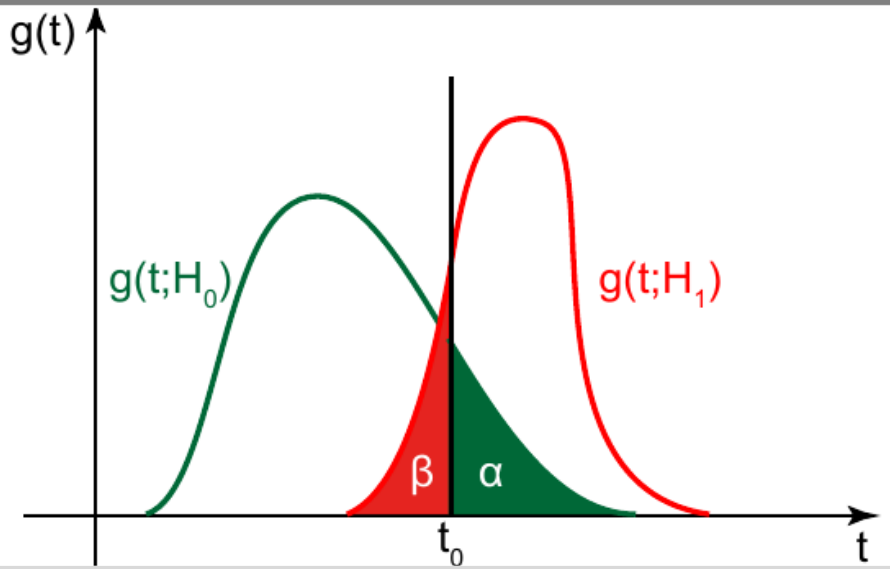
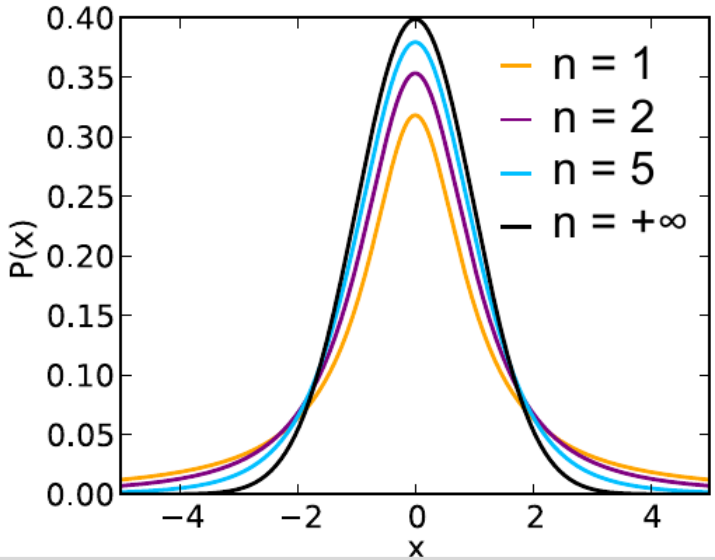
Karlsruher Institut für Technologie

VL06: Rechnernutzung in der Physik
Testen von Hypothesen

Günter Quast
Fakultät für Physik
Institut für Experimentelle Teilchenphysik

WS 2023/24





Block 1: Statistik und Datenanalyse

Themen:

31.10. Grundlagen der Statistik (Wiederholung CgDA);
python & friends, jupyter-Tutorials

07.11. Monte Carlo-Methode als numerisches Hilfsmittel
MC-Tools in PhyPraKit

14.11. Parameterschätzung mit der Maximum-Likelihood-Methode
Hilfsfunktionen in PhyPraKit & kafe2

21.11. Maximum Likelihood (2) und numerische Optimierung
Ensemble-Tests

28.11. Hypothesentests

Übungsblatt

Das **vierte Übungsblatt**
ist on-line !

Rechnernutzung in der Physik

Institut für Experimentelle Teilchenphysik
Institut für Theoretische Teilchenphysik
Prof. G. Quast, Prof. M. Steinhauser
Dr. A. Mildenberger, Dr. Th. Chwalek
Ilias Seite zum Kurs
WS 2023/24 – Blatt 03
Abgabe: Montag 4.12.2022 bzw. Dienstag 5.12.2022

Themen:

- Simulated Annealing
- Anpassung einer Geraden an Daten mit komplexen Unsicherheiten
- Vergleich von zwei Verteilungen durch Hypothesentest

ggf. Problem mit nicht installiertem Paket *PhyPraKit*:

im Terminal des Jupyter-Notebooks eingeben: pip install --user PhyPraKit

Zusammenfassung VL05: Modellanpassung

Anpassung mit mehreren, korrelierten Parametern mit Profile-Likelihood und ggf. Wahrscheinlichkeitskonturen von Paaren von Parametern

minimieren für jeden wert von t

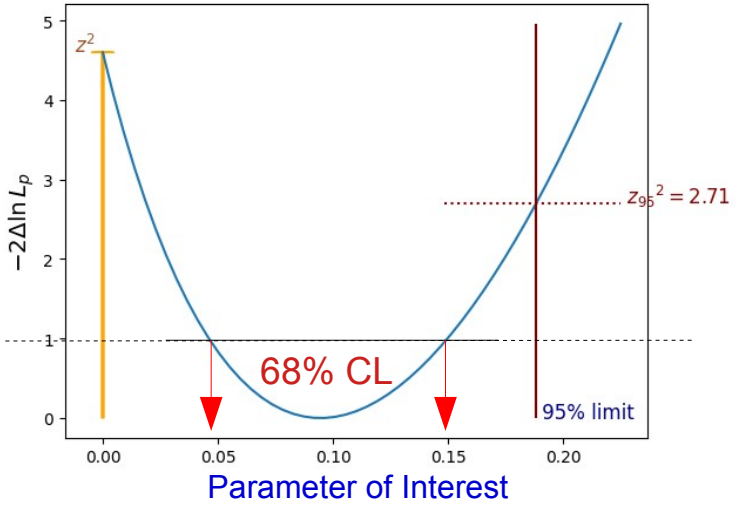
Profile Likelihood $\mathcal{L}_{\text{prof}}(t) = \mathcal{L}(t, \hat{r}(t))$

-ln $\mathcal{L}$ für jeden Wert der interessierenden Parameter t bezüglich der „Störparameter“ r optimieren.

- berücksichtigt Korrelationen der Parameter
- numerisch aufwändig, daher selten implementiert

(als Option „MINOS“ in MINUIT, iminuit)

Unsicherheiten aus Profile-Liklihood:



Übersicht: Methoden zur Bestimmung der Unsicherheiten

Lineare Probleme mit gaußförmigen Unsicherheiten:

- analytische Lösung möglich (χ^2-Minimierung)
- Position des Minimums gegeben durch Linearkombination der Messwerte: $\hat{\mathbf{a}} = (\mathbf{A}^\mathbf{T}\mathbf{V}^{-1}\mathbf{A})^{-1}\,\mathbf{A}^\mathbf{T}\mathbf{V}^{-1}\,\mathbf{y}$
- Varianz der Parameterschätzung durch Fehlerfortpflanzung:

$$V(\hat{\mathbf{a}}) = \left(\mathbf{A}^\mathbf{T}\mathbf{V}^{-1}\mathbf{A}\right)^{-1}$$

Nicht- lineare Probleme oder andere als gauß-förmige Unsicherheiten:

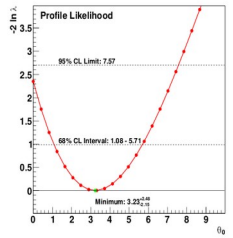
- Likelihood-Analyse:
  Ausnutzen der Cramer-Rao-Frechet-Grenze:

2. Ableitungen am Minimum: $(\hat{V}[\hat{\mathbf{a}}]^{-1})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial a_i\, \partial a_j}\right|_{\mathbf{a}=\hat{\mathbf{a}}}$

Nicht-lineare Probleme mit nicht-parabolischer Likelihood am Minimum:
(für Grenzfall großer Stichproben)

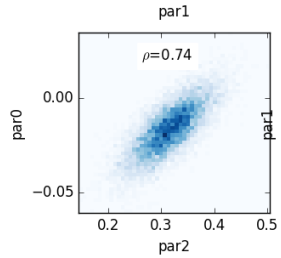
- Scan der (Profil-) Likelihood in der Nähe des Minimums

$$\ln(\mathcal{L}(\hat{a})) - \ln(\mathcal{L}(a)) = \frac{n^2}{2} \;\Rightarrow\; |a - \hat{a}| = n\sigma$$



Bei Unklarheit oder sehr kleinen Stichproben:

Monte-Carlo-Studie: Anpassung an viele der Genauigkeit und Verteilung der Daten entsprechende Stichproben, Verteilungen der Parameter studieren.



Achtung: nur die letzten beiden Methoden liefern „Konfidenz-Intervalle“ (z. B. 1σ ≙ 68%)

Numerische Optimierung

Kernaufgabe der numerischen Optimierung:

Funktionsminimierung

Klassen von Methoden:

- Monte-Carlo-Methoden, z. B. Simuliertes Abkühlen
- Heuristische direkte Suchverfahren, z. B. Simplex-Algorithmus
- Abstiegsverfahren erster Ordnung: Gradientenverfahren
- Abstiegsverfahren zweiter Ordnung: Newton-Verfahren

Viele verbesserte Varianten: adaptive Schrittweite, Impuls, ...

Sehr dynamisches Feld mit ständig verbesserten Werkzeugen,
besonders auch wegen Anwendungen im Bereich Machine Learning

Einführung Hypothesentest

Entscheidungsfindung mit Statistik

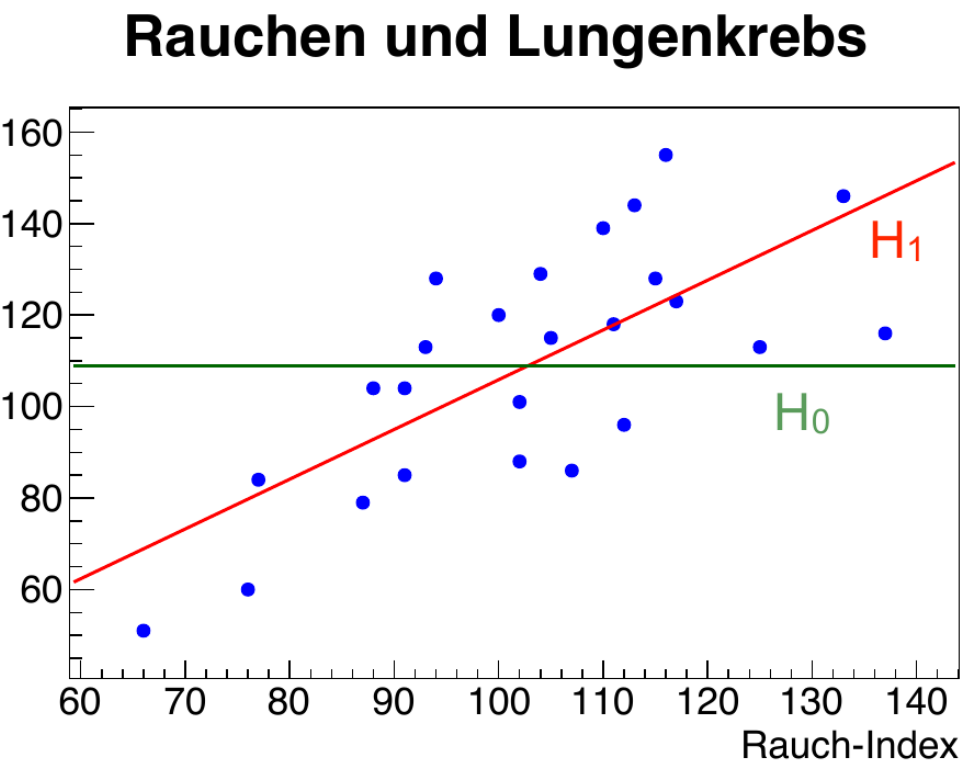
Das Ziel einer jeden Datenanalyse ist die
Beantwortung von Fragen und das **Treffen von Entscheidungen:**

- Gibt es einen Unterschied bei Prüfungsergebnisse verschiedener Gruppen (Geschlecht, Jahrgang, Studienfach, Dozent, Kurskonzept, …) ?
- Wirkt ein Medikament ?
- Soll ich morgen einen Regenschirm mitnehmen ?
- Soll ich Aktien der Firma *KleinWeich* kaufen ?
- Gibt es in den Daten eines Experiments einen Hinweis auf einen neuen Effekt (kalte Fusion, Higgs-Boson, Supersymmetrie, …)
- Biete ich Kunden, die rote Schuhe kaufen, lieber grüne oder rote Hüte an ?
- Ist der Online-Kunde ein potenzieller Betrüger ?
- …

ganz allgemein: **Hypothesentest**

Vergleich von (empirischen) Daten mit verschiedenen **Hypothesen** $\mathcal{H}i$

Beispiel: Verursacht Rauchen Lungenkrebs ?



Daten aus: *Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales*, 1970-1972, Her Majesty's Stationery Office, London, 1978.

25 Personengruppen

- Rauchindex 100, wenn Zahl der Zigaretten pro Tag dem Durchschnitt von allen Männern desselben Alters entspricht
- Lungenkrebs-Index: 100, wenn Zahl der Lungenkrebstoten dem Durchschnitt entspricht

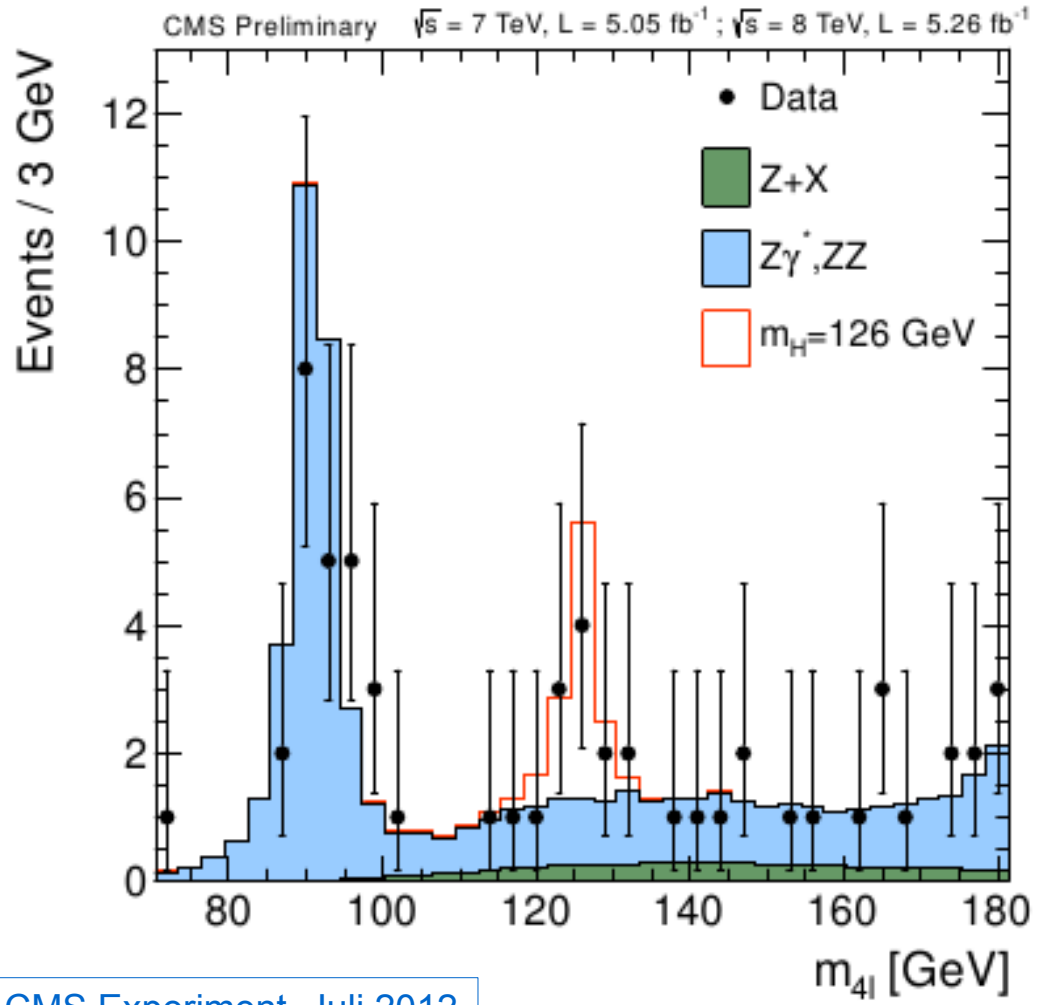
Korrelations-Koeffizient: 0.716

Frage:

Kann „Null-Hypothese“ H_0 (=kein Zusammenhang) verworfen werden ?

Beispiel 2: Entdeckung des Higgs-Bosons

Häufigkeitsverteilung der Massen von vier Myonen



CMS Experiment, Juli 2012

Frage:
Ist das eingezeichnete (rote)
Signal statistisch signifikant ?

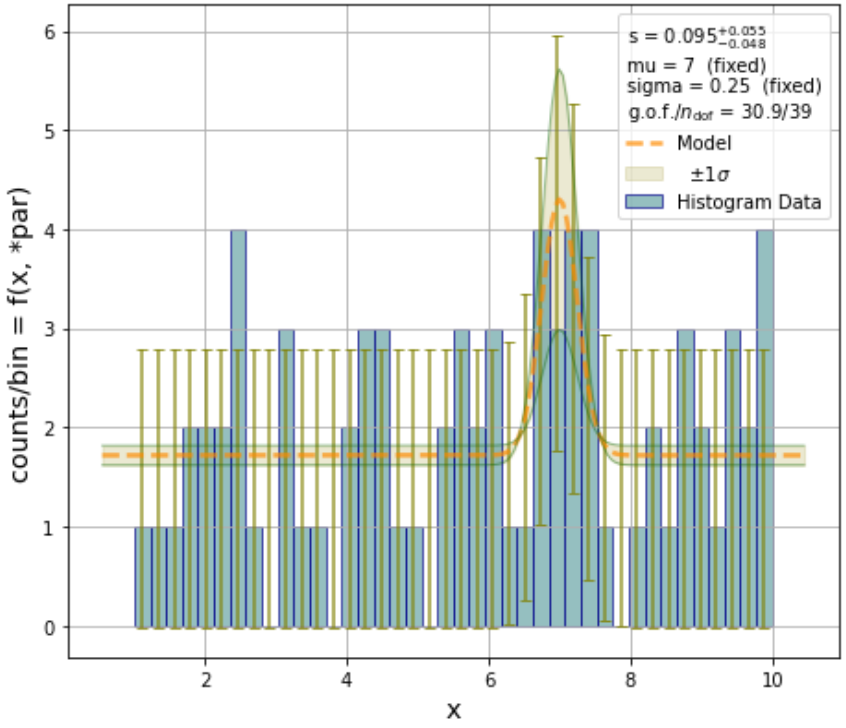
zwischen 121.5 und 130.5 GeV:
- 9 Ereignisse beobachtet
- ohne Signal ~3 erwartet ?

Frage,
Mit welcher Wahrscheinlichkeit würden bei einer Erwartung von 3 Ereignissen tatsächlich 9 oder mehr beobachtet ?

Ist diese Wahrscheinlichkeit „klein“ (kleiner als eine vor der Messung festgelegte Grenze), so handelt es sich um ein „neues Signal“.

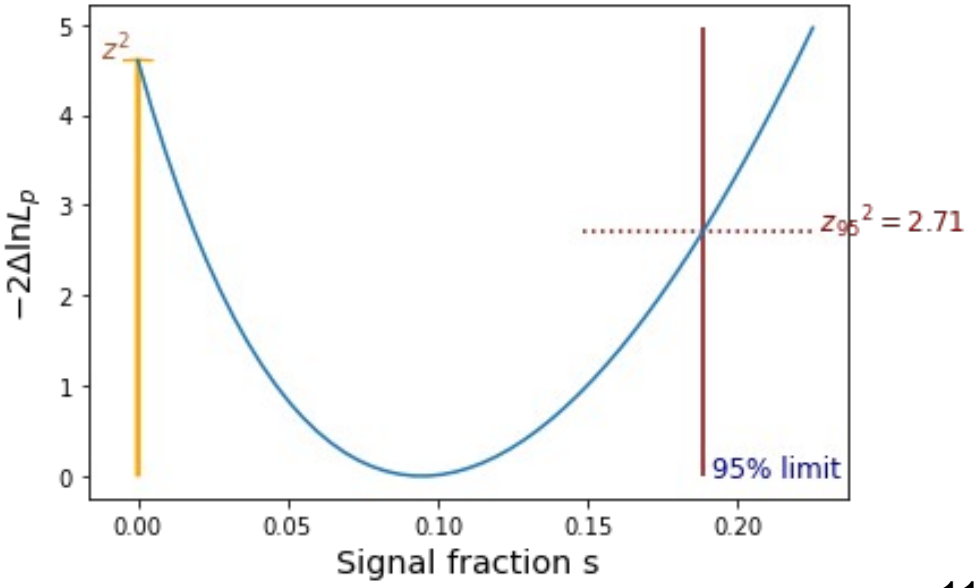
Beispiel 3: Künstliches Signal aus VL05b

Histogramm-Fit mit **PhyPraKit.hFit** und **Analyse der Profile-Likelihood**



Gaußförmiges Signal auf flachem Untergrund
mit Signalanteil s als Parameter

Ist das ein signifikantes Signal ?



== Analysis of profile likelihood
Significance, z-value: 2.14
p-value of Null-hypothesis: 0.016
95% limit: s < 0.188

Grundlagen Hypothesentest

Hypothesentest allgemein

= Vergleich einer Stichprobe aus Daten mit mehreren Hypothesen $\mathcal{H}_i$

Hypothese formuliert als PDF einer Zufallsvariablen *x*

- **Einfache Hypothese**

  spezifiziert Wahrscheinlichkeitsdichte vollständig

  PDF $f(x \mid \mathcal{H}(\lambda_i))$, alle λ_i bekannt

  z.B. „Daten folgen einer Poisson-Verteilung mit v=3,5“

- **Zusammengesetzte Hypothese**

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

  PDF $f(x \mid \mathcal{H}(\lambda_i ; \Theta_j))$, λ_i bekannt, Θ_j aus Daten bestimmt

  z.B. „Daten folgen einer Gaußverteilung mit bekanntem Mittelwert, aber unbekannter Standardabweichung

Hypothesentest: Prinzip

Zu testende Hypothese: **Null-Hypothese $\mathcal{H}_0$**

(„Alles beim Alten“, „Standardmodell“)

Andere Hypothesen: **Alternativhypothese(n): $\mathcal{H}_1$, $\mathcal{H}_2$, …**

(z.B. Abweichung von der Norm, neuer Effekt, …)

für konkreten Test muss diese explizit formuliert werden:

- „Daten folgen einer Poisson-Verteilung mit v=3,0“ (nicht 3,5)
- „Meßwerte folgen einer Poisson-Verteilung mit v < 7,0
- „Meßwerte folgen einer Gaußverteilung mit μ=1“ (nicht 0)
- „Daten sind nicht Poisson-verteilt“
  (schwieriger, da unendlich viele Alternativen möglich sind)

Typisches Ergebnis:

Verwerfen einer oder mehrerer Hypothesen,

aber Null-Hypothese kann nie bewiesen werden
denn es könnte eine (geringfügig) bessere Alternative geben !

Hypothesentest: Prüfgröße

Startpunkt: zufällige Stichprobe

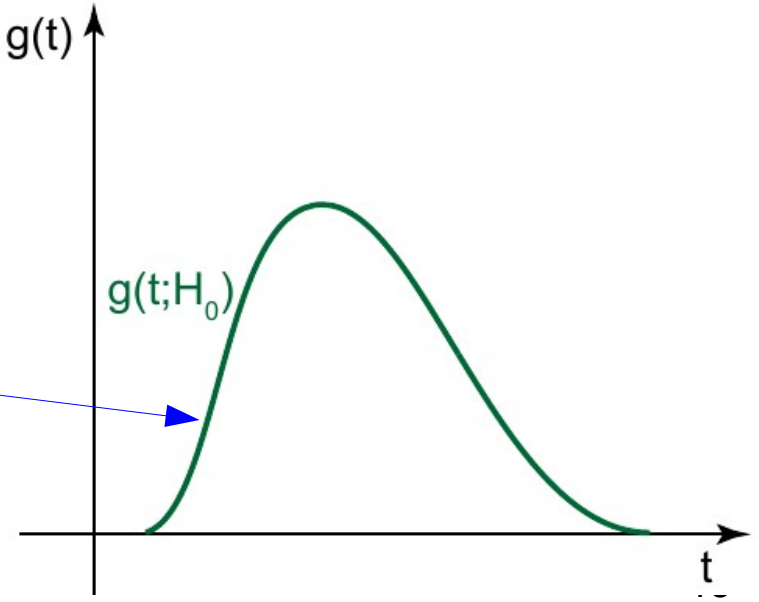
$\mathbf{x} = (x_1, \ldots , x_n)$, d. h. **die empirischen Daten**

Schritt 1: Definition einer
Prüfgröße t(**x**) (engl. „test statistic“)
zur bestmöglichen Unterscheidung
der Hypothesen $\mathcal{H}_i$

– t(**x**) im Prinzip bel. Funktion von **x**,
z.B. Mittelwert, Likelihood $\mathcal{L}(\mathcal{H}_i \mid \mathbf{x})$, ...

– idealerweise ist $t(\mathbf{x})$ eine skalare Größe

t(**x**) ist Zufallsvariable mit
Wahrscheinlichkeitsdichten $g(t| \mathcal{H}_i)$



Hypothesentest: Signifikanzniveau

Schritt 2: Festlegung eines Kriteriums zum **Verwerfen** der Nullhypothese (**vor** der Messung!)

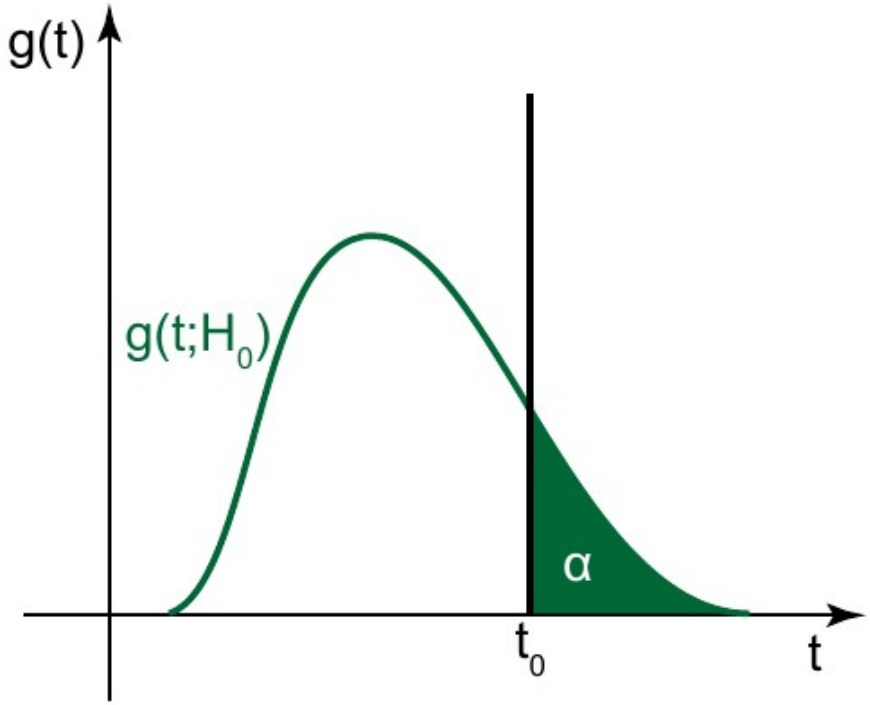
kritischer Wert t_0 :

$$\int_{t_0}^{\infty} g(t|\mathcal{H}_0)dt = \alpha$$

α: Signifikanzniveau

Bedeutung:

Auch wenn $\mathcal{H}_0$ gilt, ist im Bruchteil α aller Fälle $t > t_0$

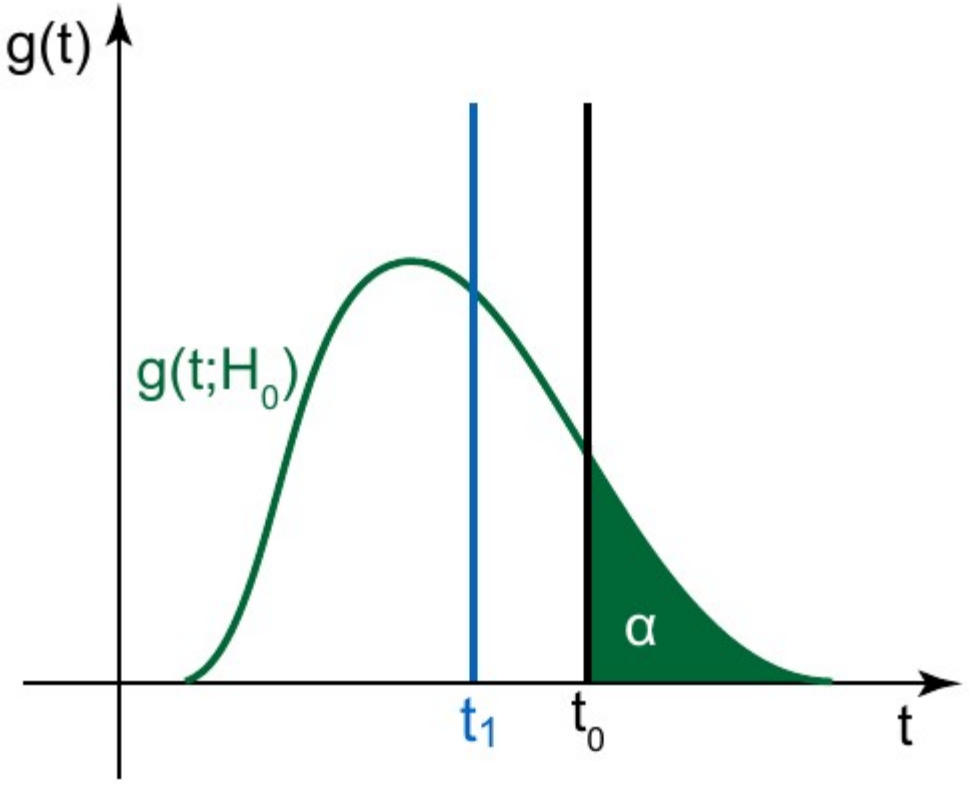


Hypothesentest: Messung

Schritt 3:
Messung liefert $t = t_1$

Berechnung des **p-Werts** = Wahrscheinlicheit für $t \geq t_1$

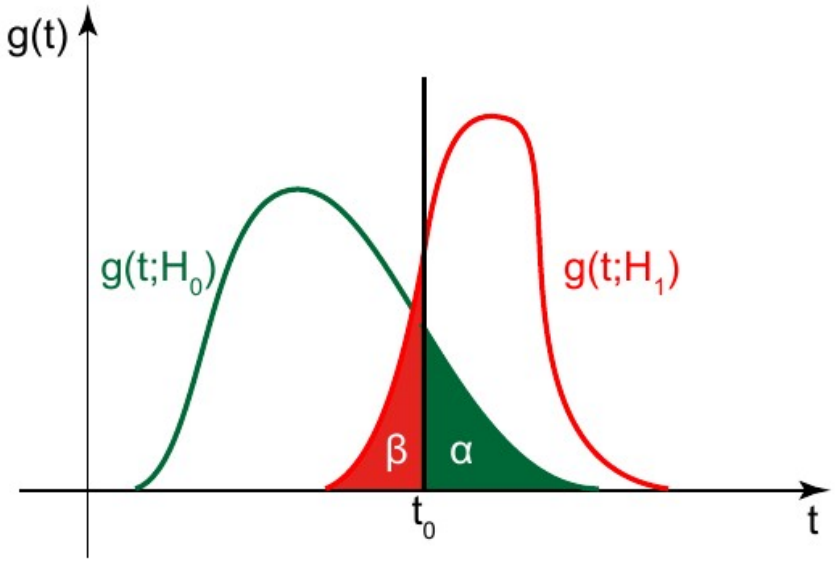
$$p = P(t \geq t_1 | \mathcal{H}_0) = \int_{t_1}^{\infty} g(t|\mathcal{H}_0)\, dt$$



Schritt 4:
Entscheidung:
Nullhypothese verwerfen, falls **p-Wert < α**

In diesem Beispiel: Nullhypothese wird nicht verworfen, weil p > α

Hypothesentest: Fehlertypen



Fehler 1. Art:
wahre Nullhypothese wird verworfen: Fläche α

Fehler 2. Art:
falsche Nullhypothese wird akzeptiert: Fläche β

1-β nennt man die Teststärke, auch Trennschärfe, Mächtigkeit (engl. power)

Beispiele Fehler 1. Art (auch „false positive“):
- Krankheit bei Gesundem diagnostiziert
- falsche Entdeckung eines neuen Teilchens
- ehrlichen Kunden als potentiellen Betrüger eingestuft

Beispiele Fehler 2. Art (auch „false negative“):
- echte Krankheit nicht erkannt
- neues Teilchen nicht gefunden, obwohl in Daten vorhanden
- Betrüger nicht erkannt und Ware auf Rechnung ausgeliefert

Wahl des Signifikanzniveaus hängt auch davon ab, welcher Fehler als schlimmer erachtet wird:
- *Wiss. Ruhm vs. Lächerlichkeit*
- *falsche vs. unterlassene Behandlung*
- *Kunden oder Ware verloren ?*

Hypothesentest: Signifikanz und p-Wert

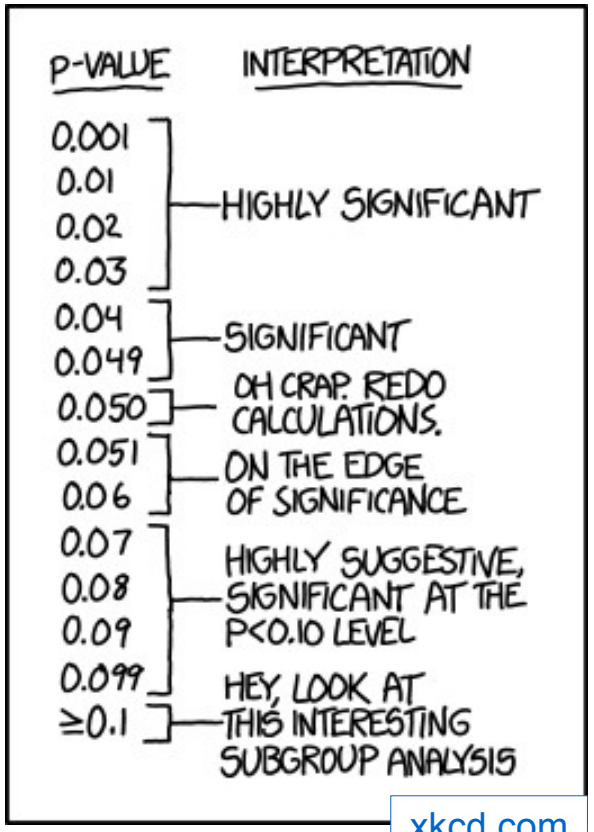
Unterscheidung Signifikanz und p-Wert:

- **α** = Wahrscheinlichkeit für Fehler 1. Art
  (festgelegt vor der Messung !)
- **p** = Wahrscheinlichkeit, dass Werte für die Prüfgröße $t \geq t_1$ gemessen würden, wenn die Nullhypothese wahr ist
  (nach der Messung von t_1 !)

Häufige Missverständnisse:

p-Wert ist nicht die Wahrscheinlichkeit, dass die Null-Hypohthese wahr oder falsch ist !

p-Wert ist auch nicht die Wahrscheinlichkeit, dass Messung „nur eine Fluktuation“ ist.



Es ist nicht unüblich, mit noch viel kleineren Signifikanzniveaus zu arbeiten:
z.B. Teilchenphysik ~10^{-7}
„Extraordinary Claims require Extraordinary Significance“

Einseitiger oder zweiseitiger Test ?

Die **„kritische Region“** zum **Verwerfen der Null-Hypothese** hängt von der Null-Hypothese selbst ab:

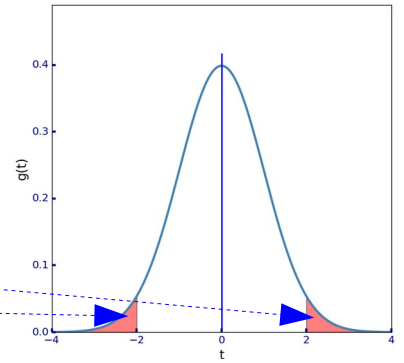
- Nullhypothese: $\bar{x}_1 = \bar{x}_2$

  kritische Region symmetrisch

  Wird verworfen, wenn mit hoher Signifikanz

  $\bar{x}_1 > \bar{x}_2$

  oder $\bar{x}_2 > \bar{x}_1$

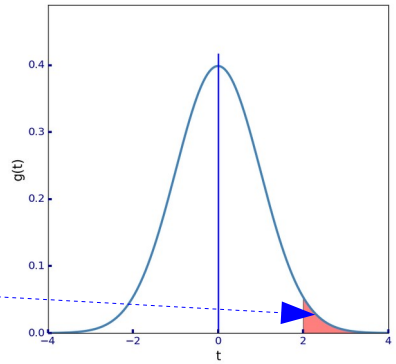


- Nullhypothese: $\bar{x}_2$ nicht größer $\bar{x}_1$

  kritische Region rechts

  (rechtsseitiger Test)

  verwerfen wenn mit hoher Signifikanz $\bar{x}_2 > \bar{x}_1$

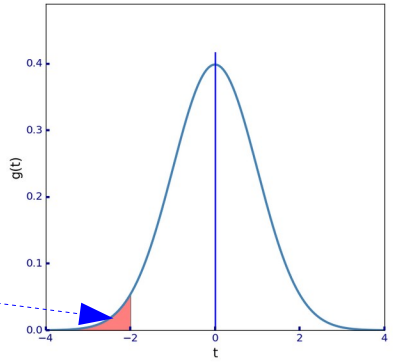


- Nullhypothese: $\bar{x}_1$ nicht größer $\bar{x}_2$

  kritische Region links

  (linksseitiger Test)

  verwerfen, wenn mit hoher Signifikanz $\bar{x}_1 > \bar{x}_2$



Beispiele Hypothesentests

- Binomial
- Poisson
- χ^2
- Kolmogorov-Smirnov
- Student'scher t-Test
- Fisher F-Test

Beispiel Münzwurf

Sie haben nach 20-maligem Wurf einer Münze 15 mal Kopf und 5 mal Zahl erhalten.
Wie kompatibel ist die Hypothese $p(\text{Kopf}) = 0.5$ mit diesem Ausgang der Messreihe?

Grundlage für die statistische Analyse ist die Binomialverteilung:

$$B(k;n,p) = \binom{n}{k} p^k \left(1-p\right)^{n-k}$$

$$\mu = np = 10 \qquad \sigma^2 = np(1-p) = 5$$

Berechnung des *p*-Werts als Summe der Wahrscheinlichkeiten für das Auftreten von Werten k ≥15 mit p=0.5:

$$\mathrm{p} = \sum_{\mathrm{k} \in \{15,16,17,18,19,20\}} B(k;n=20,p=0.5) = 0.0207$$

- bei einem Signifikanz-Niveau von α=5% würde man also die Hypothese, dass die Münze in Ordnung ist, verwerfen
- bei eine Signifikanzniveau von α=1% wäre sie noch akzeptabel.

aber Achtung:

Bei einer Münze ist $H_0 : p(\text{Kopf}) = 0.5$ eine sehr valide Annahme. Überlegen Sie sich genau, wann sie diese Hypothese ins Wanken bringen möchten !

Beispiel2: Binomial mit „Untergrund“

Ist eine medizinische Behandlung effektiv ?

60 % „spontane Heilung“, 100 Patienten behandelt →

- Nullhypothese: (nur) ≤60% der Patienten geheilt
  (also keine positive Wirkung der Behandlung)
- Alternative: die Behandlung wirkt
- Prüfgröße: Zahl der geheilten Patienten

Entscheidung über Wirksamkeit mit 5% Signifikanz

Grundlage ist wieder die Binomial-Verteilung

Zahlen sind „groß“, erlaube mir eine Gauß'sche Näherung

$$\sigma = \sqrt{Np(1-p)} = \sqrt{100 \cdot 0.6 \cdot 0.4} = 4.9$$

einseitiges 5%-Quantil der Gauß-Verteilung liegt bei μ+1.64 σ

→

mehr als 60 + 1.64 · 4,9 Patienten = **69 Patienten** müssten geheilt werden, um die Wirksamkeit (statistisch) zu belegen.

[kennen wir schon: χ^2 – Test]

$$S_{\min} = \chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i, \{\hat{p}\})\right)^2}{{\sigma_i}^2}$$

N Messungen
k Parameter

S_{min}, die **„gewichtete Summe der Residuenquadrate“**
am Minimum bzgl. der Parameter {p},
folgt bei Gauß-förmig verteilten Fehlern σ_i
einer χ^2-Verteilung mit $n_f = N-k$ Freiheitsgraden.

Erwartungswert: $\langle\chi^2\rangle = n_f$ oder $\langle\chi^2/n_f\rangle = 1$

Die **χ^2-Wahrscheinlichkeit**

$$\chi^2_{\text{prob}} = \int_{S_{\min}}^{\infty} \chi^2(s, n_f)\, ds = 1 - \int_{0}^{S_{\min}} \chi^2(s, n_f)\, ds$$

dient zur Quantifizierung der Qualität einer Anpassung

Aussage, mit welcher Wahrscheinlichkeit ein größerer Wert von χ^2 am Minimum als der tatsächlich beobachtete zu erwarten wäre.

[Erinnerung: Qualität der Anpassung aus Likelihood]

Durch geeignete „Normierung“ kann Qualitätsinformation auch aus der Likelihood gewonnen werden: **Likelihood-Verhältnis**

der beobachteten Daten und (geeigneter) Referenzdaten.

Beispiel 1: **Likelihood der Gaußverteilung**

$$-\ln\mathcal{L}(\vec{a}|\vec{y}) = \frac{1}{2}\sum_i \left(\frac{y_i - f(x_i,\vec{a})}{\sigma_i}\right)^2 + \sum_i \ln(\sqrt{2\pi}\sigma_i)$$

Referenz: $y_i = f(x_i,\hat{\vec{a}})$ „fully saturated model“

$$\Rightarrow \text{-ln}\left(\frac{\mathcal{L}(\hat{\vec{a}})}{\mathcal{L}_{\text{ref}}(\hat{\vec{a}})}\right) = \sum_i \underbrace{\left(\frac{y_i - f(x_i,\hat{\vec{a}})}{\sigma_i}\right)^2}_{\chi^2} + \underbrace{\sum_i \ln(\sqrt{2\pi}\sigma_i) - \left(0 + \sum_i \ln(\sqrt{2\pi}\sigma_i)\right)}_{0} = \chi^2(\hat{\vec{a}})$$

Beispiel 2: **Likelihood der Poisson-Verteilung**

$$\text{Poiss}(n_i;\mu_i) = \frac{{\mu_i}^{n_i}}{n_i\,!}\,\mathrm{e}^{-\mu_i}$$

$$-\ln\mathcal{L}_i = \sum_i -N_i\ln\mu_i + \mu_i + \ln N_i\,!$$

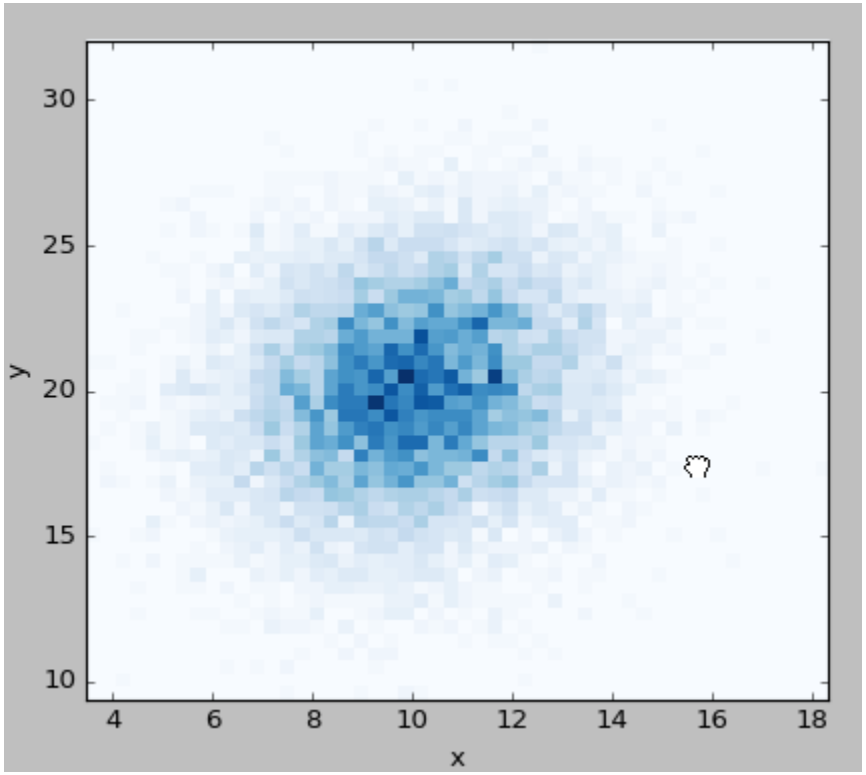
Referenz: $\mu_i = N_i,\ -\ln\mathcal{L}_{\text{sat}} = \sum_i -N_i\ln N_i + N_i + \ln N_i!$

$$\Rightarrow \quad \ln\frac{\mathcal{L}(\hat{\vec{a}})}{\mathcal{L}_{\text{ref}}(\hat{\vec{a}})} = \sum_i N_i\ln\frac{N_i}{\mu_i} + \mu_i - N_i =: \text{gof}_{\text{Poisson}}$$

gof = goodness of fit

Beispiel: χ^2 – Test auf Unabhängigkeit

? Sind die Variablen x und y unabhängig ?



Erinnerung:
für unabhängige Variable ist die Verteilung *f(x,y)* gegeben durch das Produkt der Randverteilungen:

$$f(x,y) = f_x(x) \cdot f_y(y)$$

Im Fall eines 2-dimensionalen Histogramms: Randverteilungen sind die Histogramme von *x* u. *y*

Daraus lässt sich ein

Test auf Unabhängigkeit konstruieren:

Nullhypothese:

$$n_{ij} = N_{\text{tot}}\, p_{x\,i} \cdot p_{y\,j} =: n_{\text{exp}}$$

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n_{\text{exp}})^2}{n_{\text{exp}}}$$ folgt unter Annahme der Nullhypothese einer χ^2-Verteilung mit $N_{\text{tot}} - b_x - b_y$ Freiheitsgraden

$b_{x,y}$: Zahl der Bins in *x* und *y*

→ p-value of chi2-independence test: 2.1%

Script PhyPraKit/chi2_indep.py

x und y sind also wohl nicht unabhängig !

„Run"-Test

Beispiel: $\chi^2 = 12$ für 12 Bins → χ^2-Test ok

Aber: Daten offensichtlich nicht linear

Komplementär: der Run-Test:

Aufteilung der Datenpunkte in Gruppen:
oberhalb (above) und unterhalb (below),
r = Anzahl der Gruppen

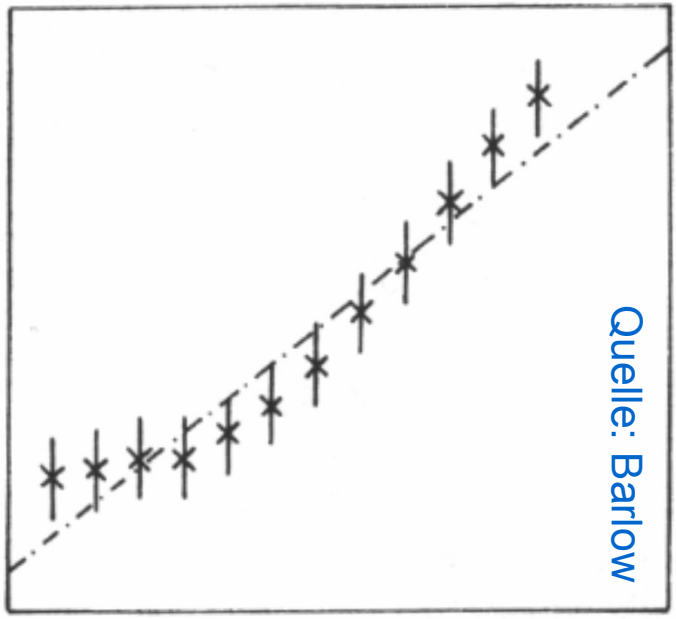


Fig. 8.3. A straight line through twelve data points.

Struktur der Abweichungen hier:

AAABBBBBBAAA

für A = above, B = below → nur 3 Gruppen oder sog. „Runs"

r ist eine Zufallszahl mit bekannter PDF und Momenten:

$$E[r] = 1 + \frac{2n_a n_b}{n_a + n_b} \qquad V[r] = \frac{(E[r]-1)(E[r]-2)]}{n_a + n_b - 1}$$

dieses Beispiel: E[r] = 7, σ(r) = 1.65 → Δr = 2.4σ, p(einseitig) = 1%

→ schlechte Übereinstimmung mit dem Modell !

Kolmogorov-Smirnov (KS) Test

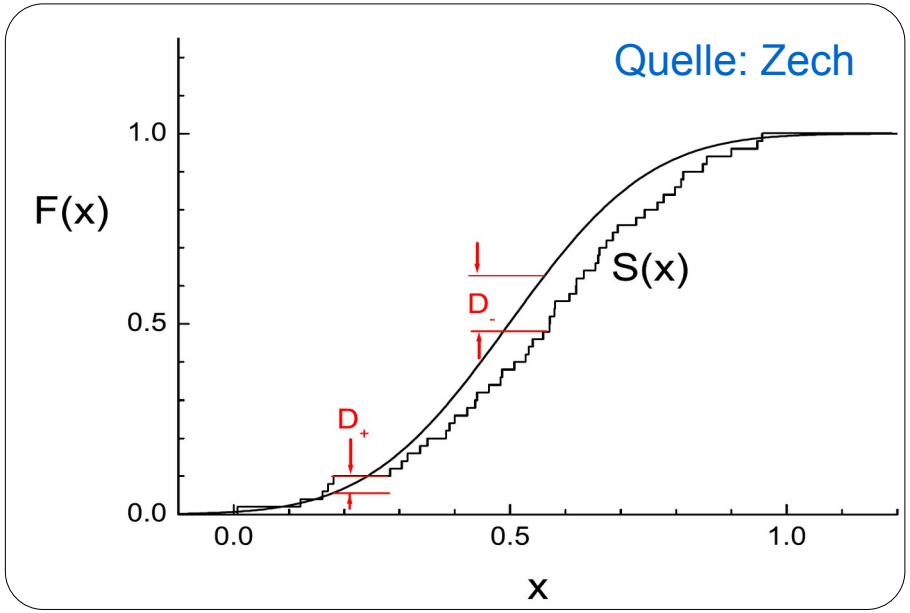
auf Gleichheit zweier Verteilungen

weitere Alternative zu χ^2-Test für (ungebinnte) kleine Datenproben

Folgt Stichprobe einer bestimmten (kontinuierlichen) PDF ?

Idee: Vergleich der kumulativen Verteilungsfunktionen (CDF)

- Daten der Größe nach sortieren, kumulierte Verteilung, normiert mit 1/N, auftragen → empirische CDF:
  $CDF_e(x)$ = (Anzahl Werte < x) / N
- Vergleich mit CDF
  $F(x) = \int_{-\infty}^{x} f(x')\,dx'$



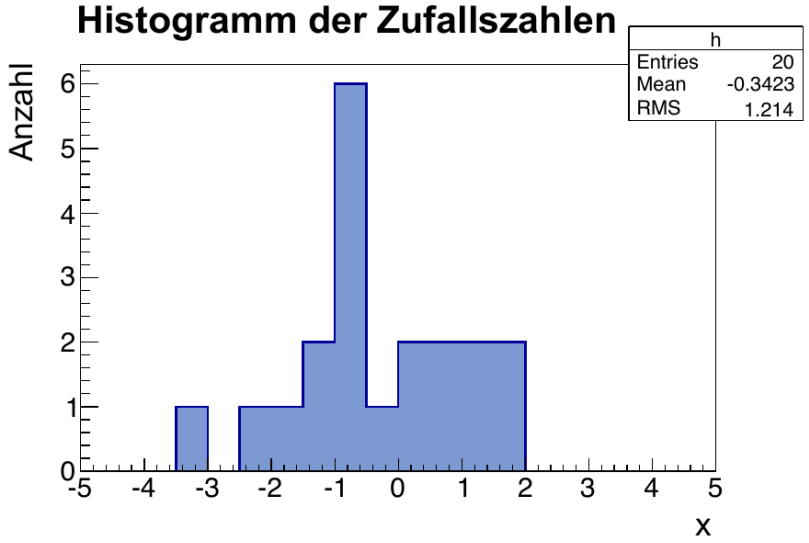
Prüfgöße: KS-Abstgand,
definiert durch maximale Abweichung
$t = \sqrt{N}\ \max|CDF_e(x) - F(X)|$

z.B. p = (1%, 5%,10%,20%) für t =(1.63, 1.36, 1.22, 1.07)
Gilt nur, wenn f(x) nicht an die Daten angepasst wurde

p-Werte für KS-Test sind tabelliert bzw. in *ROOT* oder *Scipy* verfügbar;
alternativ: Toy-MC ...

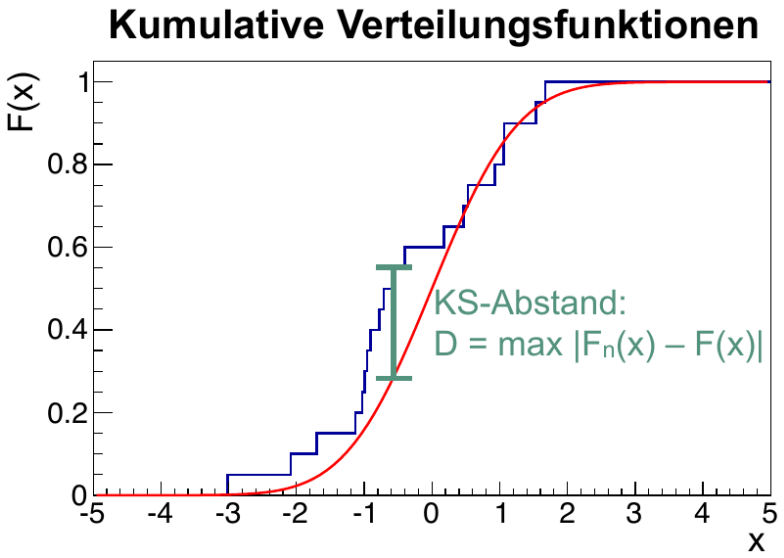
Beispiel KS Test

20 Zufallszahlen aus Gauß-Verteilung
(Binning nur zu Darstellungszwecken)



Vergleich der empirischen CDF mit Verteilungsfunktion der Gaußverteilung

→ KS-Abstand 0.271
→ **p-Wert 10.6%**



Alternativen:
Anderson-Darling- oder
Cramer-von Mises-Test

Mittelwert einer Stichprobe als Prüfgröße

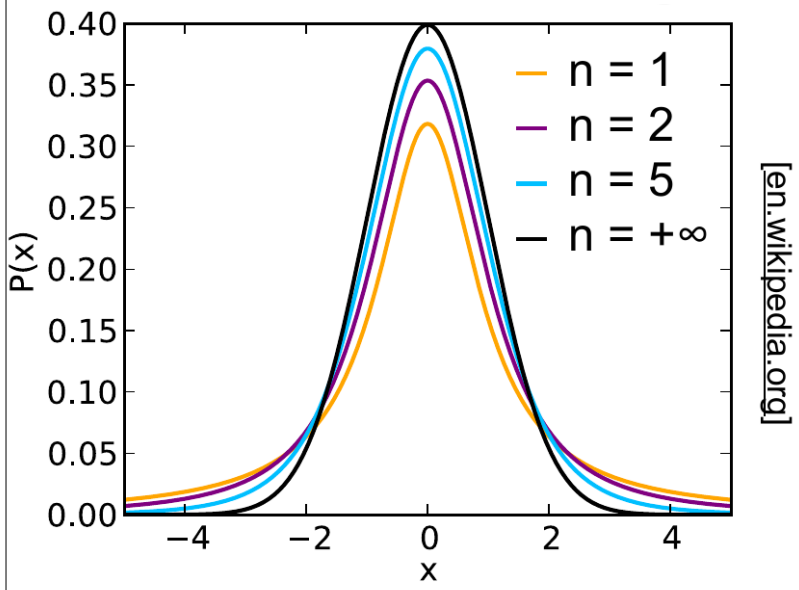
Student'sche t-Verteilung

n „standard-normalverteilte“ Zufallszahlen $u_i = \frac{x_i - \mu}{\sigma}$ mit Mittelwert $\bar{u} = \frac{\bar{x} - \mu}{\sigma}$

wenn σ nicht bekannt ist, nutzt man die Stichprobenvarianz $\hat{\sigma}$

die normierte Größe $t = \frac{\bar{x} - \mu}{\hat{\sigma}\sqrt{n}}$ folgt dann **nicht** der Gaußverteilung,

sondern der **Student'schen t-Verteilung** für n-1 Freiheitsgrade



n = 1: Cauchy-Verteilung
n = ∞: Normalverteilung

Insbesondere für kleine n viel größere Ausläufer als Gauß-Verteilung !

1908 vom Guiness-Mitarbeiter W.S. Gosset unter dem Pseudonym „Student“ veröffentlicht.

erste Anwendung: Qualitätssicherung bei Bier der Marke Guinness

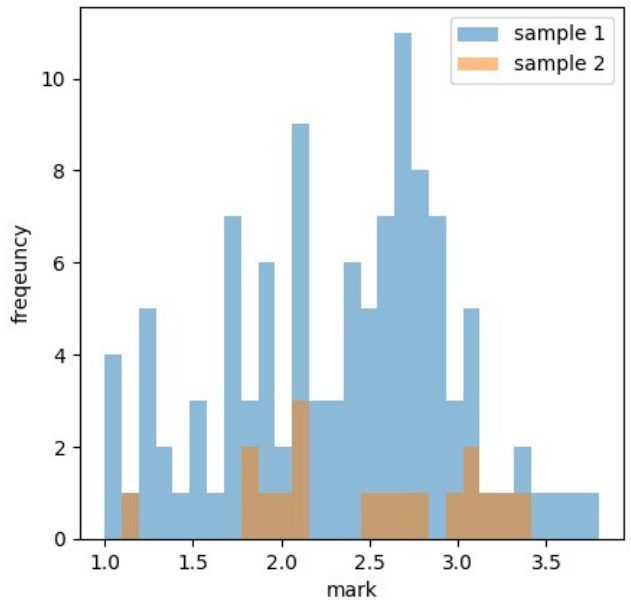
Mittelwert einer Stichprobe als Prüfgröße → **t-Verteilung verwenden !**

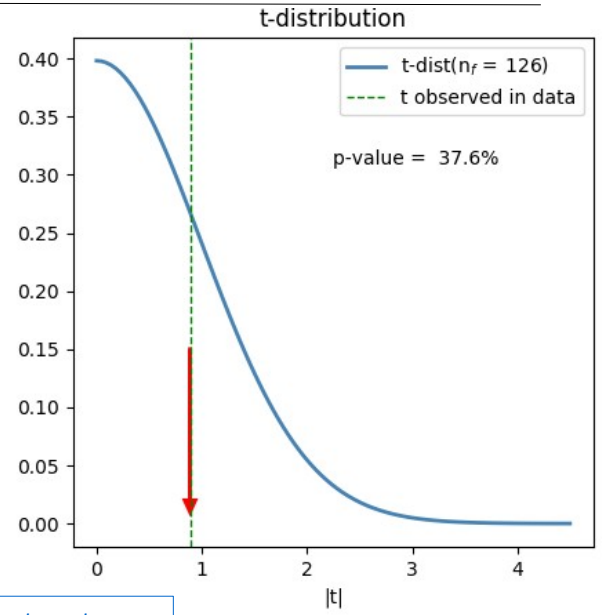
Anwendung: Student'sche *t*-Verteilung

Häufige Fragestellung: haben zwei unabhängige Grundgesamtheiten (mit der gleichen Varianz) den selben Mittelwert ?

- Nullhypothese: $\bar{x}_1 = \bar{x}_2$: *t*-verteilte Prüfgröße mit $n_f = n_1 + n_2 - 2$ Freiheitsgraden

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sigma_{12}\sqrt{1/n_1 + 1/n_2}} \quad \text{mit} \quad \sigma_{12} = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$





Script `t-test.py`

Student'scher t-Test: ***t* = 0.888**

p -Wert aus t-Verteilung ***p* = 37,6%**

Nullhypothese wird nicht verworfen
d. h. Unterschied nicht signifikant !

Welch-Test als Alternative zum Test empirischer Daten auf gleichen Mittelwert bei ungleichen Varianzen.

Test auf gleiche Varianz

zweier als gaußförmig angenommenen Verteilungen

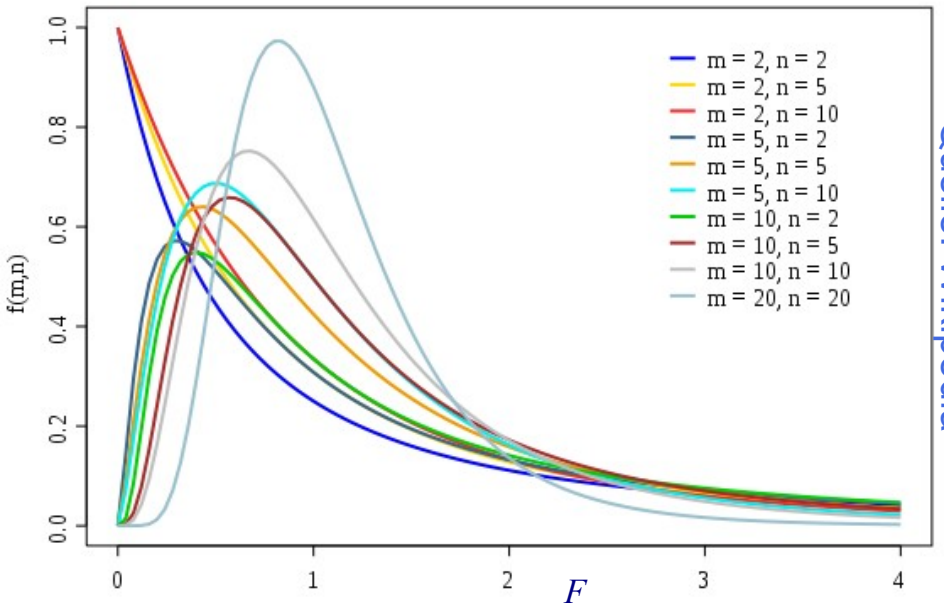
F-Verteilung mit $F = \frac{\hat{V}_1}{\hat{V}_2}$ als Testgröße $\hat{V} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2$

Für $\sigma_1 = \sigma_2$ ist F das Verhältnis zweier χ^2-Verteilungen mit $f_1 = N_1 - 1$ bzw. $f_2 = N_2 - 1$ Freiheitsgraden.

F-Verteilung:

$$f(F; f_1, f_2) = \frac{\Gamma((f_1 + f_2)/2)}{\Gamma(f_1/2)\,\Gamma(f_2/2)} \sqrt{f_1^{f_1} f_2^{f_2}} \frac{F^{\frac{f_2}{2}-1}}{(f_1 F + f_2)^{\frac{f_1+f_2}{2}}}$$

Konvention: $\hat{V}_1 \geq \hat{V}_2$ d. h. $F \geq 1$



Quelle: Wikipedia

Verteilungsdichte der F-Verteilung, f(F;n,m)

Für große N ist $z := \frac{1}{2} \log F$ gaußverteilt mit Mittelwert $\frac{1}{2}(1/f_2 - 1/f_1)$ und Varianz $\frac{1}{2}(1/f_2 + 1/f_1)$

Bester Hypothesentest

Likelihood-Verhältnis

Optimale Wahl der Prüfgröße

Neyman – Pearson Lemma: **Bester Test für einfache Hypothesen :**
(d.h. ohne anzupassende (Stör-)Parameter)

$$r(x) = \frac{\mathcal{L}(\mathcal{H}_0)}{\mathcal{L}(\mathcal{H}_1)} = \frac{\prod_{i=1}^{n} f(x_i|\mathcal{H}_0)}{\prod_{i=1}^{n} f(x_i|\mathcal{H}_1)} \leq \eta(\alpha)$$

äquivalent: $q(x) = -2 \ln r(x) = 2\left(\ln f(x_i|\mathcal{H}_1) - \ln f(x_i|\mathcal{H}_0)\right)$

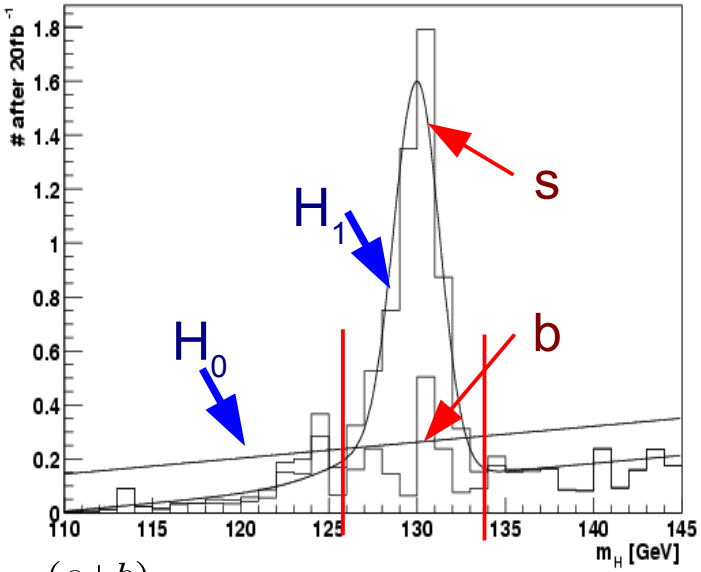
Problem: exakte Likelihood oft unbekannt.
Möglichkeiten:
– (plausiblen) Ansatz für funktionale Form verwenden
– Monte Carlo – Simulation

Im **Grenzfall großer Stichproben** gibt es **asymptotische Verteilungen** für ***r(x)*** auch für zusammengesetzte Hypothesen (also mit freien, aus den Daten zu bestimmenden Parametern) **„Wilks'sches Theorem“**

Beispiel: Likelihood-Quotient bei Zählexperiment

Suche nach kleinem Signal (s) über Untergrund (b)

- Verteilungsdichte: n(m) = s(m) + b(m)
  m: Parameter, z.B. Teilchenmasse
- Bei Suchen verwenden wir häufig:
  $\mu * s(m) + b(m)$
  μ: Signalstärke, d.h. H_0: μ=0, H_1: μ>0



Bilde Quotient der bzgl. H_0 und H_1 maximierten Likelihoods

$$P_0(n;b) = \frac{1}{n!} b^n \mathrm{e}^{-b} \;,\; P_1(n;s+b) = \frac{1}{n!}(s+b)^n \mathrm{e}^{-(s+b)}$$

$$2\ln(P_1/P_0) = 2\left(n\ln\left(1+\frac{s}{b}\right) - s\right)$$

- **Einfaches Verfahren:** Abzählen von Ereignissen in Signalregion
  Beobachtung von n Ereignissen (b hier als bekannt angenommen)

$\Rightarrow \hat{b} =: b, \hat{s} = n - \hat{b} =: s$, d.h. $2\ln T_{S+B} = 2(b+s)\ln(1+\frac{s}{b}) - 2s$

Differenz in Log-Likelihood zwischen n=b und n=s+b ist ein Maß für die **Signifikanz z** der Signalbeobachtung (z: „Zahl der Sigmas“)

$$z = \sqrt{T_{S+B}} = s/\sqrt{b} + \mathcal{O}((s/b)^2)$$

Verallgemeinerung: Profile-Likelihood Quotient

Übergang zu kontinuierlichem Parameter μ: Signalstärke (μ=0: H_0 ; μ=1: H_1)

Dividiere Likelihood $L(x|\mu,\Theta_\mu)$ durch das globale Minimum der Likelihood → näherungsweise Unabhängigkeit von Θ_μ (Abdeckung ok, sofern $\Theta \approx \Theta_\mu$)

Prüfgröße

Signalstärke

Optimierte Störparameter für vorgegebenes μ

$$q_\mu = \frac{\mathcal{L}(x_{\text{Daten}}|\mu, \hat{\Theta}_\mu)}{\mathcal{L}(x_{\text{Daten}}|\hat{\mu}, \hat{\Theta})} \;,\; 0 \leq \hat{\mu} \leq \mu$$

Bedingung erzwingt μ ≥ 0 und eine einseitige Grenze

Verteilungen der Daten

best-fit-Wert aller Parameter

^ := Parameterwerte, die Likelihood maximieren

*Bestimmung der **Verteilung von q_μ**, **$g(q_\mu|\mu)$**, für Untergrund (μ=0) bzw. Signal-Hypothese (μ≠0), durch simulierte Pseudo-Experimente oder asymptotische Formeln im Grenzfall großer Datensätze.*

Das Bootstrapping-Verfahren

Bootstrapping-Verfahren als besondere MC-Methode

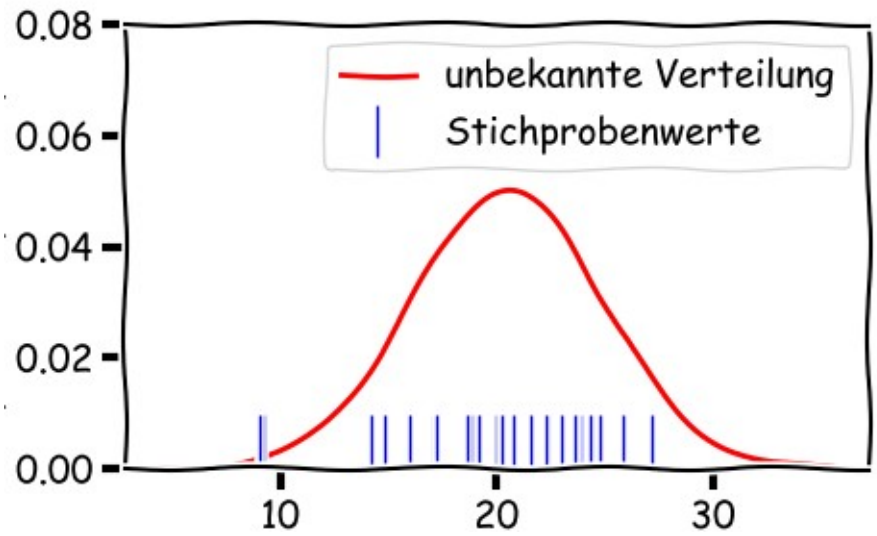
Gesucht: Funktionen $T(\{x_i\})$ einer Stichprobe $\{x_i\}$

Für T = Mittelwert oder T = Varianz gibt es allg. Formeln.

Für Median, Quantile, Ergebnis einer Parameteranpassung etc. gibt es allerdings keine allg. analytischen Formeln.

Ausweg „Toy-MC“: Ziehen vieler vergleichbarer Stichproben k aus angenommener Verteilung der Grundgesamtheit - jeweils T_k bestimmen, Varianz von T aus Varianz der T_k.

??? Bestimmung der **statistischen Unsicherheit von T bei unbekannter Verteilungsdichte** ???



Sich an den Schuhriemen (oder, wie Münchhausen) an den eigenen Haaren aus dem Sumpf ziehen

→ **Bootstrapping**

(B. Efron, 1979)

Bootstrapping: die Idee

Bootstrapping ist eine Resampling-Methode

Ziehen vieler Stichproben aus der beobachteten Stichprobe **mit Zurücklegen**

manche Werte der ursprünglichen Stichprobe kommen gar nicht vor, andere mehrmals – das ist genau so gewollt !

Die Methode besteht also darin, weitere Stichproben zu gewinnen, die so aussehen, wie die Daten auch hätten aussehen können !

Damit können statistische Unsicherheiten ohne Zusatzannahmen aus den beobachten Daten bestimmt werden.

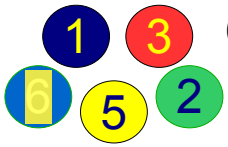
Beispiele solcher „zweifelhaften“ Zusatzannahmen:

- Punkteverteilung bei Prüfungen ist gaußverteilt
- Klausurnoten sind gaußverteilt
- Zahl der Beobachtungen eines seltenen Phänomens ist so groß, dass von einer Gaußverteilung ausgegangen werden kann.
- Gauß ist immer gut, wenn man sonst nichts weiß

Hauptanwendung:

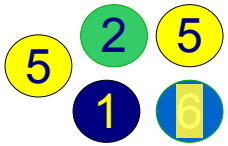
Parameterfreie Hypothesentests in Wirtschafts- und Sozialwissenschaften

Beispiel Bootstrapping



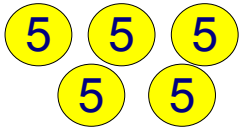
Origialdaten

Beispiel einer neu erzeugten Bootstrap-Probe



ein weiteres Beispiel

• • •



auch möglich, aber extrem selten

Bei geringem Stichprobenumfang
könnte man alle möglichen Kombinationen untersuchen,

Bei großen Stichproben wählt man eine sehr große Anzahl zufällig ausgewählter neuer Proben ***durch Ziehen mit Zurücklegen*** aus den Originaldaten.

Beispiel Bootstrapping in Python

Bootstrapping ist sehr einfach zu realisieren:

Original-Stichprobe $\{x_i\}$, i =1, …, I der Länge l :
ziehe zufällig l Indizes $k \in \{1, \ldots, l\}$
$\{x_k\}$ ist weitere Stichprobe

typischerweise werden einige Tausend solcher Resamplings erzeugt und statistisch ausgewertet.

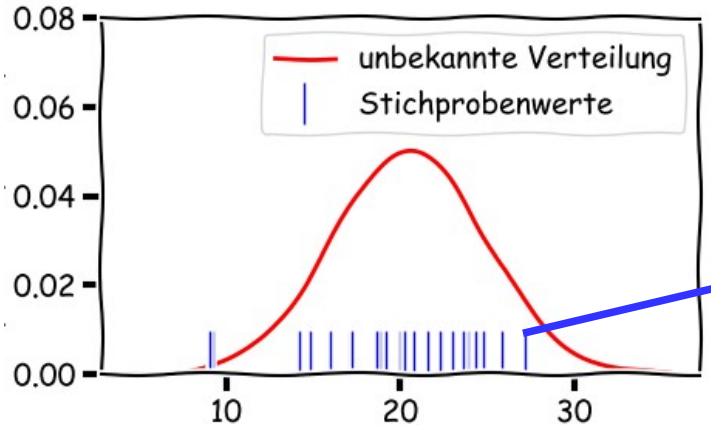
Python:

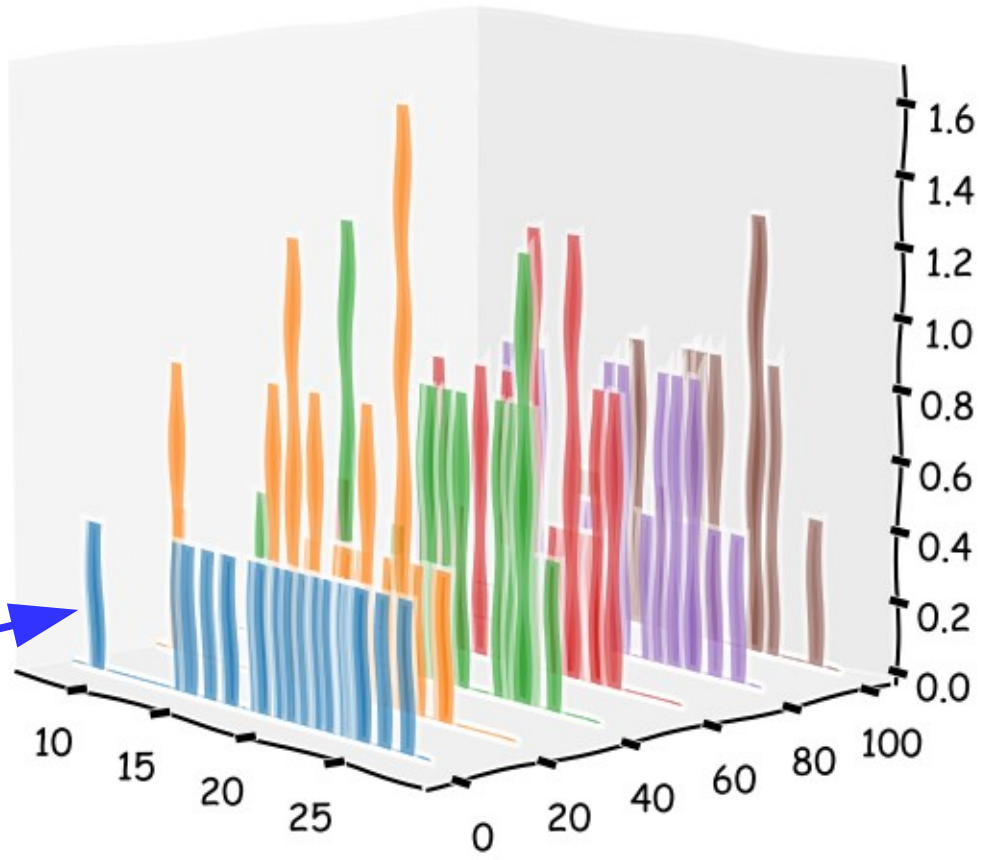
```
t_bs = numpy.random.choice(t, size=n)
```

erzeugt aus der beobachteten Strichprobe t durch Ziehen mit Zurücklegen („resampling with replacement“) eine neue Stichprobenvariante

Bootstrapping am Beispiel

Skript
bootstrapIllu.py



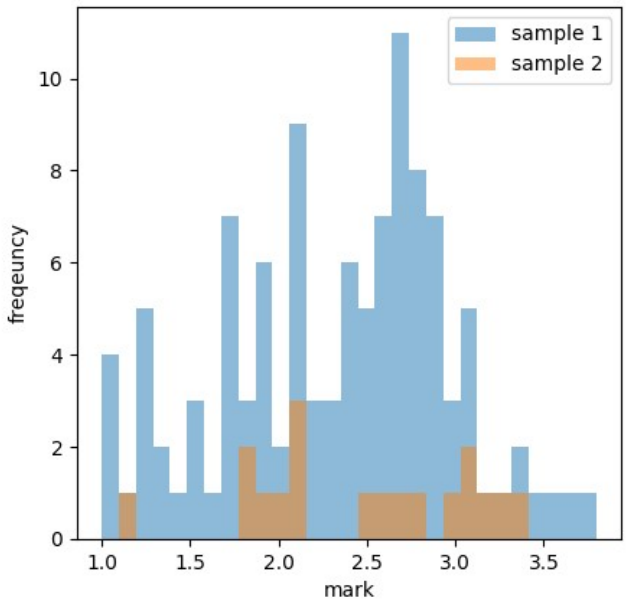


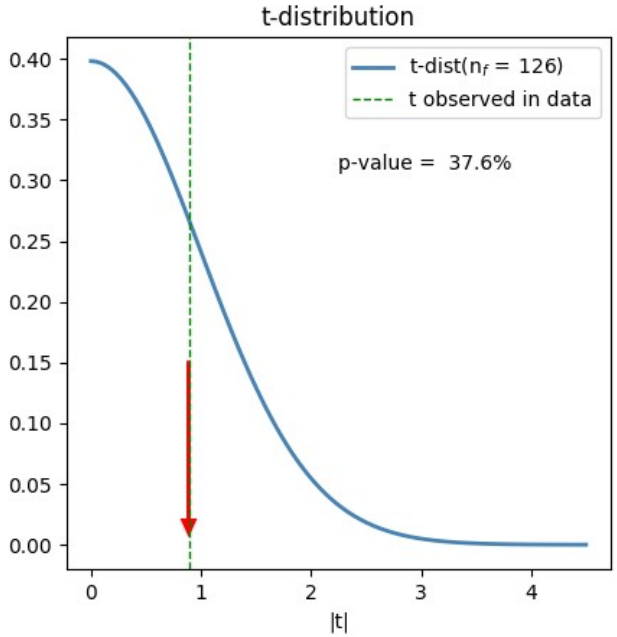
Beispiel von eben:

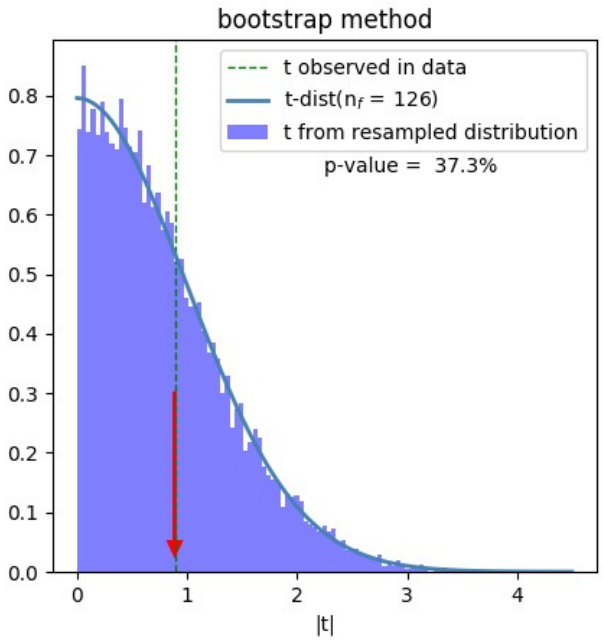
fünf aus der vorderen Stichprobe durch Resampling gewonnene Varianten der Stichprobe

Problem: Ausläufer der Verteilung sind meist unterrepräsentiert !

Test auf Gleichheit mit Bootstrapping







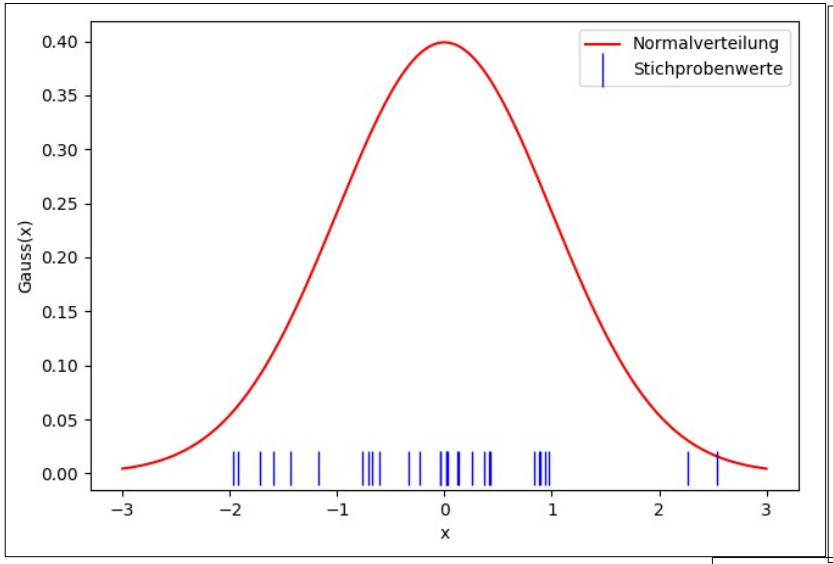
Student'scher t-test

und

Bootstrapping
liefern hier
identische Ergebnisse
!

Skript
t-test.py

Beispiel: Konfidenzintervalle für Mittelwert



Beispiel:

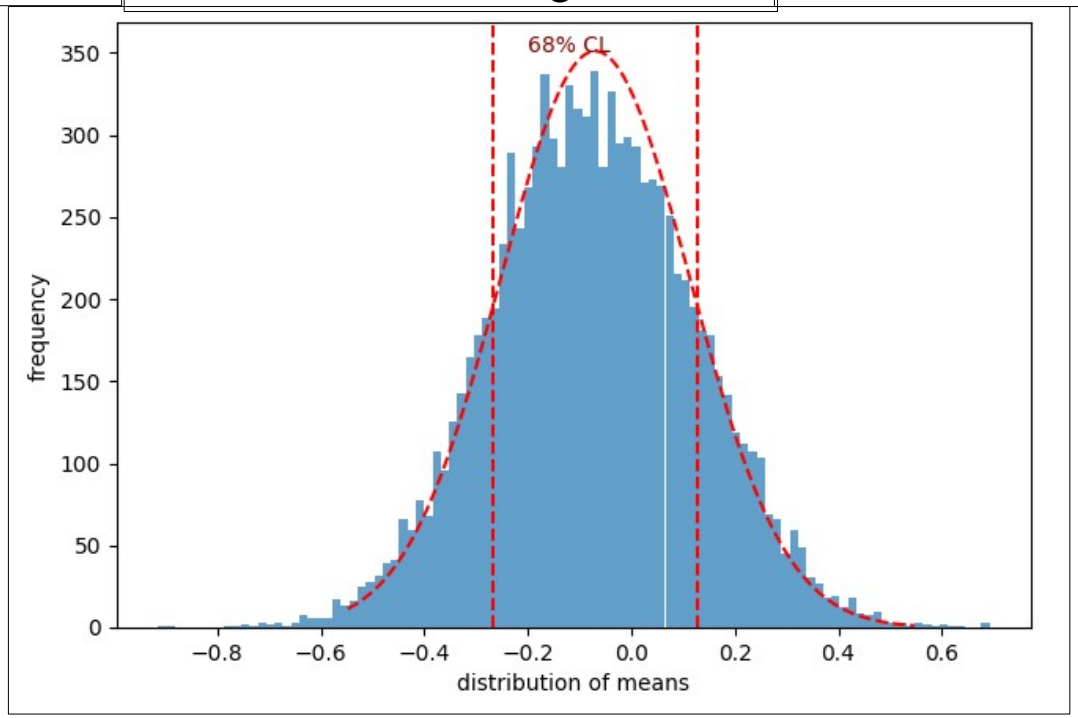
30 gaußverteilte Zahlen

10 000 Bootstrap-Samples

Vergleich:

Bootstrap-Verteilung mit Gauß-Verteilung um den Mittelwert der Orignaldaten

`bootstrap.py`



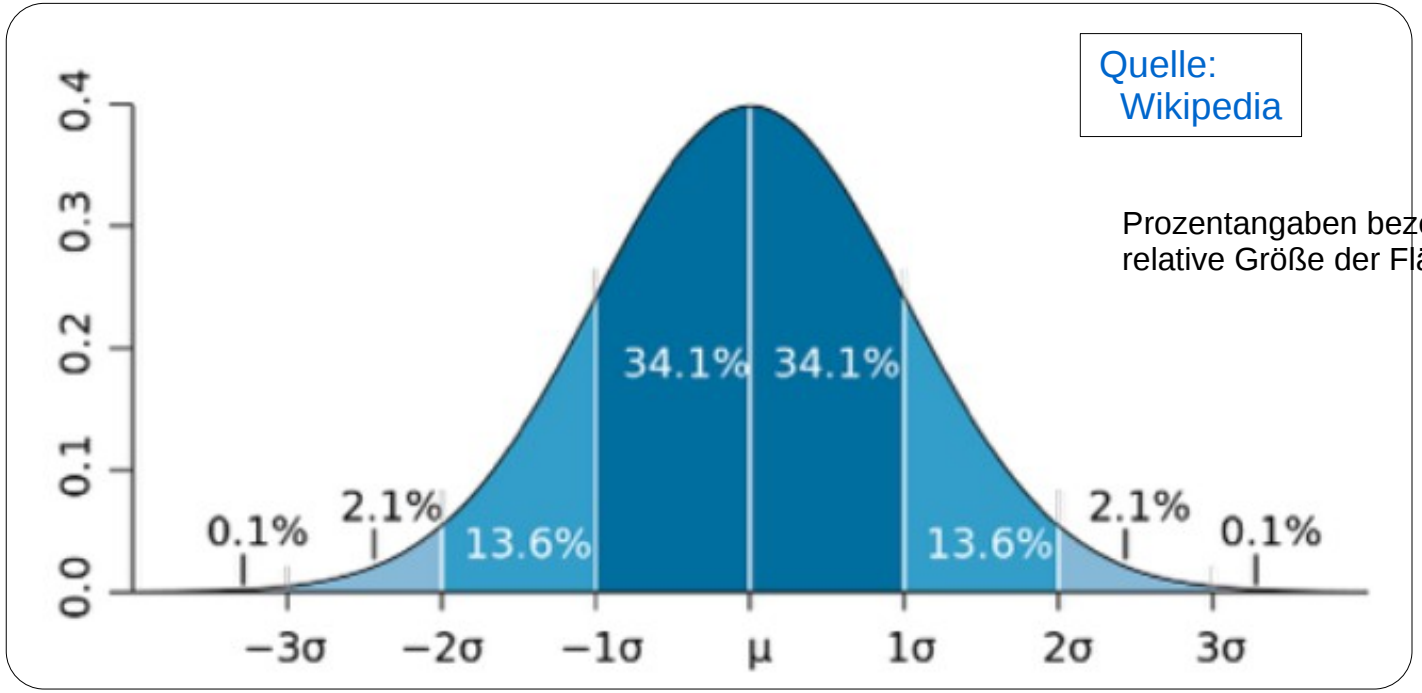
Fazit:

Bootstrapping reproduziert das mit der „klassischen" Methode bestimmte Konfidenzintervall $\mu \pm 1\sigma$ *(68%).*

Konfidenzintervalle

Konfidenz-Intervalle

- Intervall-Schätzung: Bestimme Intervall, in dem der wahre Wert mit vorgegebener Wahrscheinlichkeit liegt.
- Wahl einer vorgegebenen Wahrscheinlichkeit als Konfidenz-Niveau, auch „Vertrauensniveau“ (engl. Confidence Level „CL“)
- Üblich sind CL = 68.3%, 90% oder 95%



Konfidenz-Intervalle: zwei Sichtweisen

Frequentistische Sichtweise:

Es existiert ein fester wahrer Wert *a.*

„Das Konfidenz-Intervall zu einem Konfidenzniveau CL=p% deckt den wahren Wert a in p% aller Fälle ab“

- Wahrscheinlichkeitsbegriff bezogen auf Häufigkeiten (z.B. QM)
- Wahrscheinlichkeitsaussagen über Intervalle, nicht den wahren Wert
- „Inversion“ durch Neyman-Konstruktion, Abdeckung „by design“

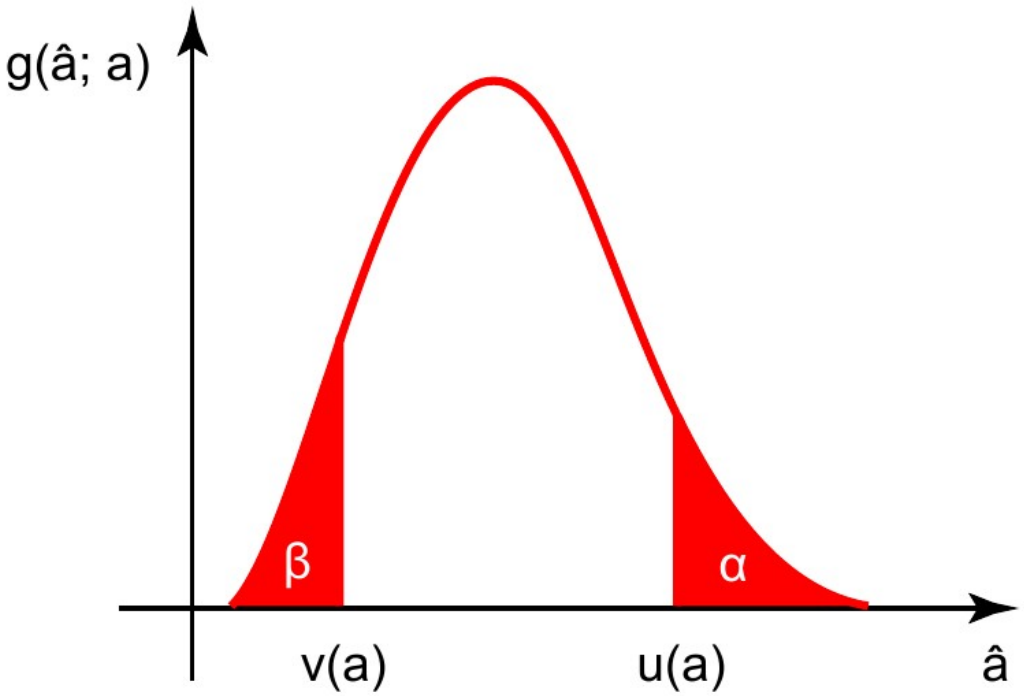
Bayes‘sche Sichtweise:
Der wahre Wert a hängt von Voraussetzungen ab. Wahrer Wert (im engeren Sinn) existiert nicht.
Konstruiere Wahrscheinlichkeitsdichte für unbekannten Wert a:

$$f(a|\hat{a}) \propto \mathcal{L}(\hat{a}|a) \cdot \pi(a)$$

- Benötigt Likelihood Funktion L und Prior π
- Abdeckung muss explizit überprüft werden (z.B. mit toy MC)

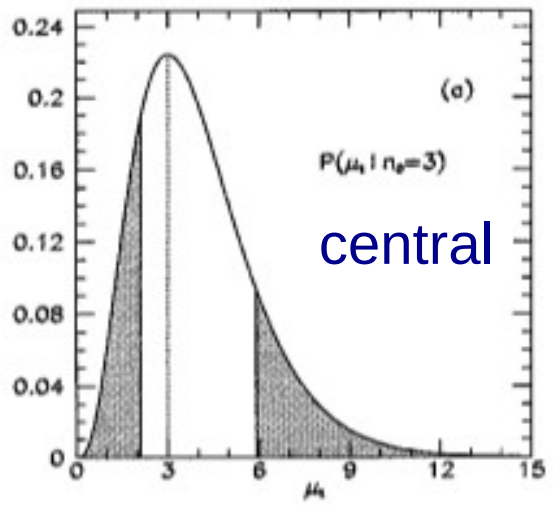
Konfidenz- bzw. Credibility-Intervalle

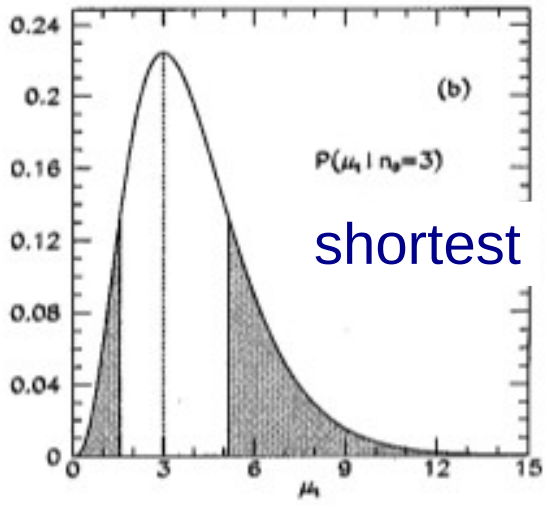
Im Nicht-Gauß-Fall sind verschiedene Intervall-Typen gebräuchlich:

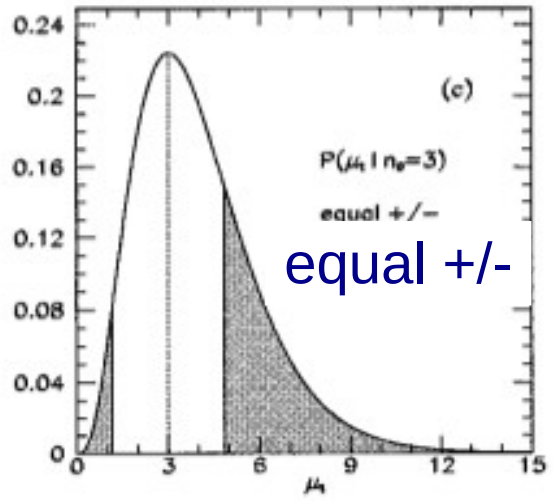


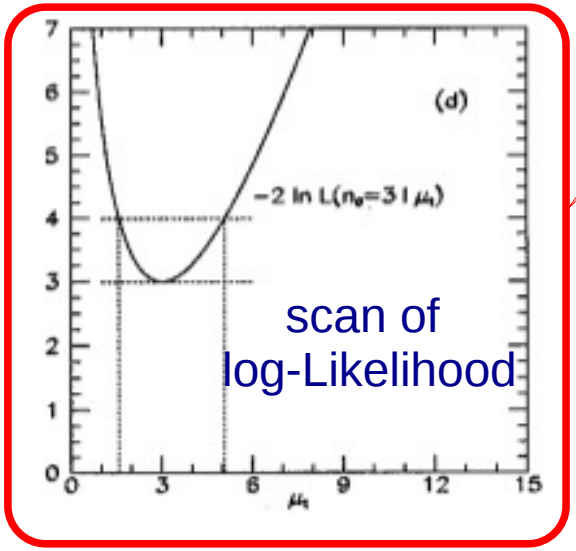
- α = β [= 16% für 68%-Konfidenzintervall] (wie z.B. auf voriger Seite
- kürzest-mögliches Intervall (in der Regel α ≠ β)
- andere, z.B. „likelihood-ordered“ („Feldman-Cousins unified approac

Unterschiedliche Intervall-Typen









Mein Favorit

FIG. 4. More 68% C.L. intervals in Poisson case with $n_0 = 3$. Bayesian (uniform prior) using Eqn. 16 with subsidiary conditions (a) Eqn. 17, (b) minimum width, and (c) $\hat{\mu} - \mu_1 = \mu_2 - \hat{\mu}$. (d) Likelihood ratio method, Eqn. 9.

Aus: R. Cousins, „Why isn't every Physicist a Bayesian?“

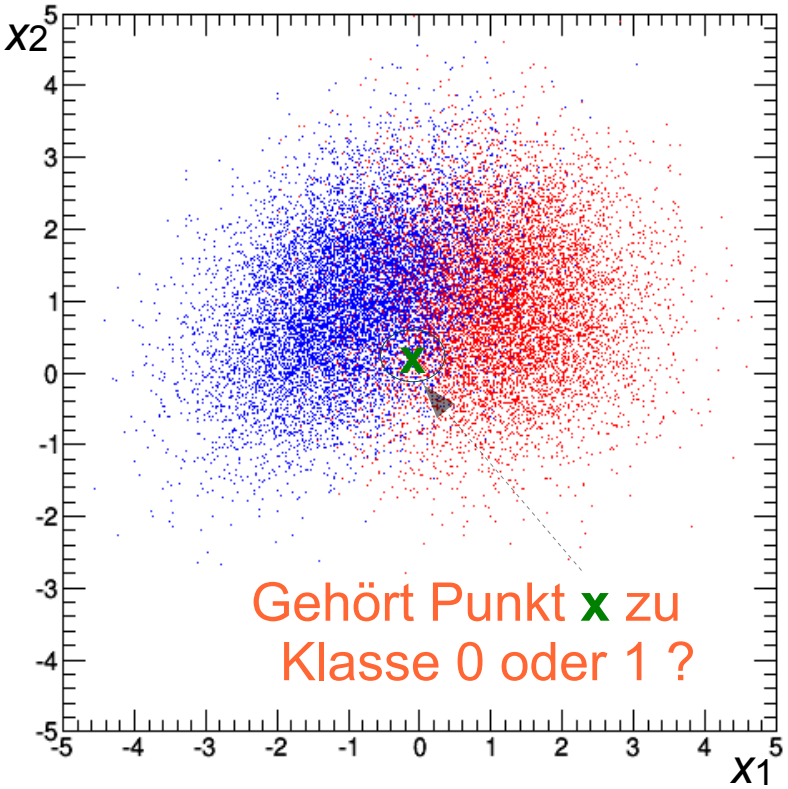
Ausblick

Ausblick: Klassifizierung als Hypothesentest

Gehört ein Ereignis zu einer von zwei oder mehreren Klassen ?

– Zufallsereignis beschrieben durch n Zufallsvariable $x_1, \ldots, x_n$

– Klasse k beschrieben durch PDF $f_k(x_1, \ldots, x_n)$



Klassifizierungsprobleme in hochdimensionalen Variablenräumen sind häufig:

– ist der Buchstabe ein „a“ ?

– ist das Teilchen ein Elektron oder Myon ?

– ist der Kunde ein potentieller Betrüger ?

– Signal oder Untergrund ?

– ist die E-Mail Spam ?

- …

Typische **„MVA“-Methoden** zur Behandlung, allg. Methoden des „Machine Learning“:

– künstliche neuronale Netze

– verstärkte Entscheidungsbäume

– Support-Vektoren

– … (s. VL Datenanalyse im Master)

Suche nach neuen Phänomenen

Auch die **Suche nach neuen Phänomenen**, z.B. in der (Teilchen-)Physik, **ist** ein **Hypothesentest**:

Frage: Ist Beobachtung verträglich mit der bekannten Physik?

- wenn ja, Ausschlussgrenze auf neues Phänomen bestimmen
  (geht nur, wenn die Alternativhypothese genau festgelegt ist – Aufg. der Theor. Physik)
- wenn nein, Signifikanz der Abweichung spezifizieren (als p-Wert)

Häufig führt man „Zählexperimente“ zur Suche nach einem neuen Signal durch.
- relevant für die Beobachtung von *n* Ereignissen ist die Poisson-Verteilung $P(n;\mu) = \frac{\mu^n}{n!}\mathrm{e}^{-\mu}\,;\ \sigma = \sqrt{\mu}$

- Nullhypothese: $\mu = \mu_0$
- Alternative: $\mu = \mu_0 + \mu_1$; μ_1: **Beitrag durch neuen Effekt**

Messung: Beobachtung von n_{obs} Ereignissen,
aufteilen in (erwarteten) Untergrund $b = \mu_0$ und Signal $s = n_{obs} - b$
***p*-Wert:** Wahrscheinlichkeit $n \geq n_{obs}$ falls Nullpyhothese wahr:

$$p = \mathrm{Prb}(n \geq n_{\mathrm{obs}}) = \sum_{k=n_{\mathrm{obs}}}^{\infty} P(k;b) = 1 - \sum_{k=0}^{n_{\mathrm{obs}}-1} P(k;b)$$

$$\text{Zahlenbeispiel}: n_{\mathrm{obs}} = 5,\ b = 0.5 \ \rightarrow\ \mathrm{Prb}(n \geq 5) = 1.7 \cdot 10^{-5}$$

Beispiel: Entdeckung des Higgs-Bosons

Bestimmung von Grenzen aus p-Werten:

- Untergrundhypothese

$$p_b := \mathrm{Prob}(\mu > \mu^{\mathrm{obs}} | b)$$

- Signalhypothese

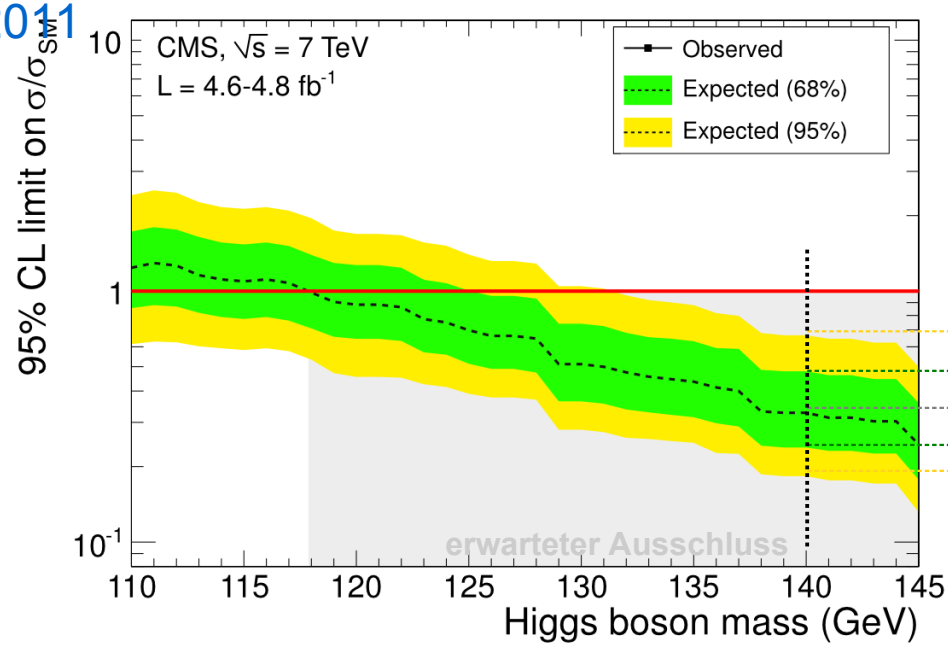
$$p_{s+b} := \mathrm{Prob}(\mu > \mu^{\mathrm{obs}} | s+b)$$

- *b* muss sehr gut bekannt / modelliert sein
  - *wenn man den „Normalfall" nicht kennt, kann man nicht nach Abweichungen suchen!*

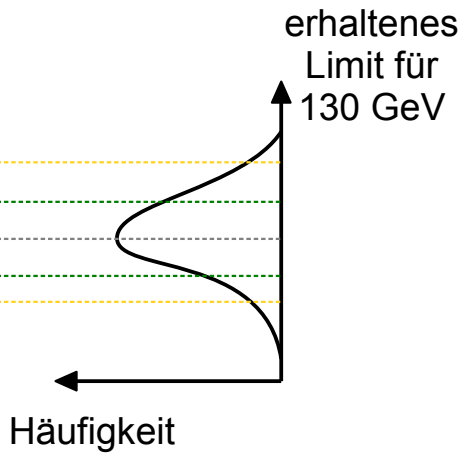
Übliche Art der Darstellung:

Signalgröße *s*, den man mit Signifikanzniveau von 95% ausschließen kann

Beispiel: Suche nach dem Higgs-Boson am LHC, Stand 2011



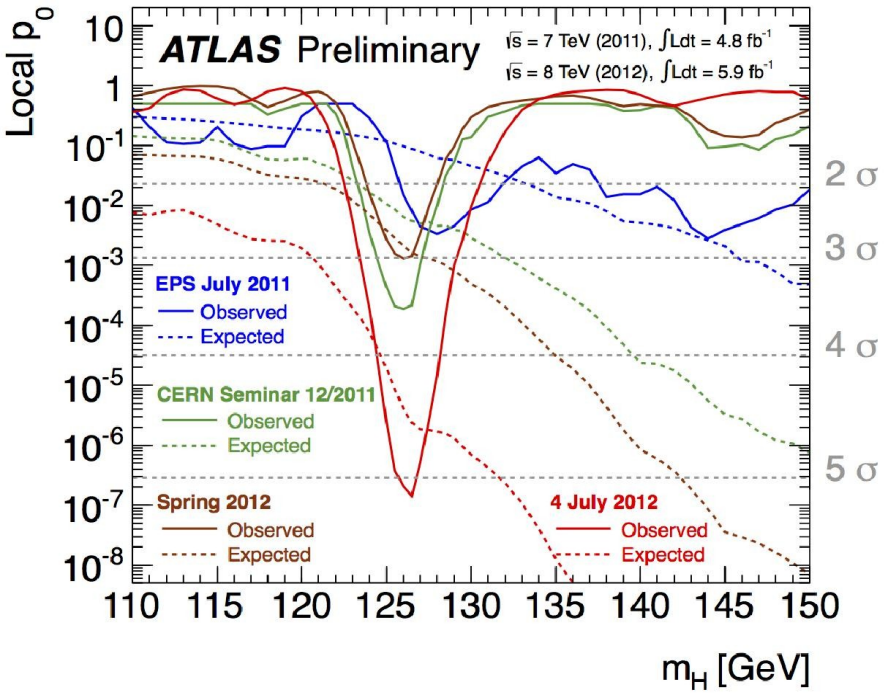
Hypothesentest für verschiedene H-Massen



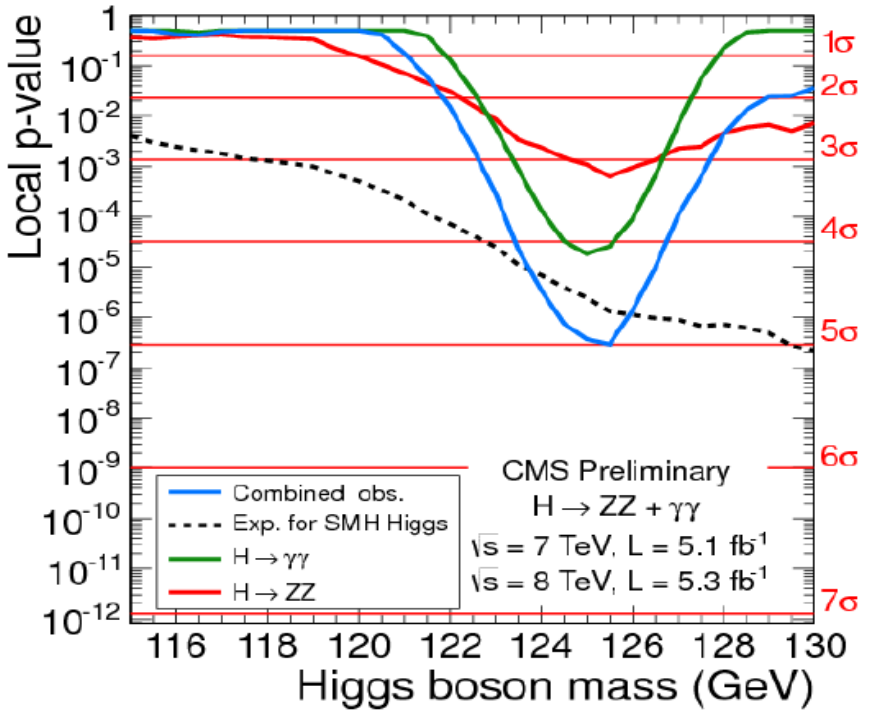
Higgs-Entdeckung: statistische Analyse

Bestimmung der Signalsignifikanz durch Vergleich mit der Untergrund-Hypothese und Bestimmung des „lokalen *p*-Werts“:

S_{ATLAS} = 5.9 σ (*) (publizierte Ergebnisse) **S_{CMS} = 5.0σ** (*)



Zeitliche Entwicklung der Signal-Signifikanz
Vorläufige Ergebnisse vom 4. Juli 2012
der ATLAS-Kollaboration am LHC



Signal-Signifikanz
vorläufige Ergebnisse vom 4. Juli 2012
der CMS-Kollaboration am LHC

(*) Im Jargon der Teilchenphysiker entspricht die Angabe ***n σ*** dem entsprechenden Quantil der Gaußverteilung

Vielfältig kombinierbar als

- eigenständiges Wahlfach, Teil eines Ergänzungs- oder Schwerpunktfachs.

Auch gut zum „Mastervorzug“ geeignet.

Moderne Methoden der Datenanalyse
im Sommersemester 2024

The **curriculum** includes

- Repetition of fundamental concepts
- Folding and Unfolding
- Hypothesis tests
- Applications of Neyman Pearson Lemma, Wilks’ Theorem and asymptotic formulae
- Confidence intervals and limits
- Machine learning in particle physics
  - Multilayer perceptron
  - Loss function & gradient descent
  - Neural network learning
  - Mapping of tasks into neural networks

Ebenfalls zu empfehlen:

Schlüsselkompetenzkurs

„Collaborative Software Design“

Vorlesung nächste Woche: Einstieg Block 2

- Rechnerarchitekturen, Parallelisierung und Vektorisierung
- Kollaborative Softwareentwicklung