Statistische Methoden der Datenanalyse

Wintersemester 2011/2012

Albert-Ludwigs-Universität Freiburg

Dr. Stan Lai und Prof. Markus Schumacher

Physikalisches Institut Westbau 2 OG Raum 008

Telefonnummer 07621 203 8408 (SL) / 7612 (MS)

E-Mail: Stan.Lai@physik.uni-freiburg.de

Markus.Schumacher@physik.uni-freiburg.de

http://terascale.physik.uni-freiburg.de/lehre/ws_1213/statmethoden_ws1213

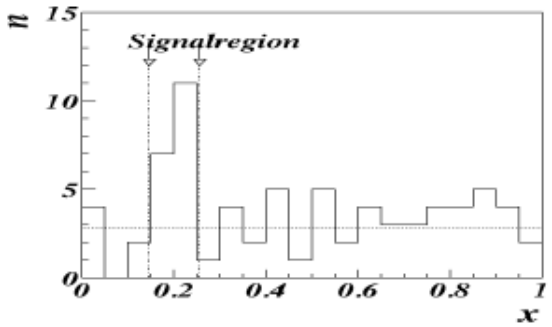
Die Signifkanz eines beobachteten Signals

H_0: Nur Untergundprozesse

H_1: Signal+Untergrund

Ziel: Zurückweisung der Nullhypothese

Zunächst: nur Zählexperiment



Annahme: wir beobachten *n* Ereignisse; diese können bestehen aus :

n_b Ereignisse aus bekannten Prozessen (Untergrund b)
n_s Ereignisse aus neuem Prozess (Signal s)

Wenn n_s, n_b Poisson-verteilte ZV mit Mittelwerten s,b sind, dann ist $n = n_s + n_b$ ebenfalls Poissonverteilt mit Mittelwert = $s + b$:

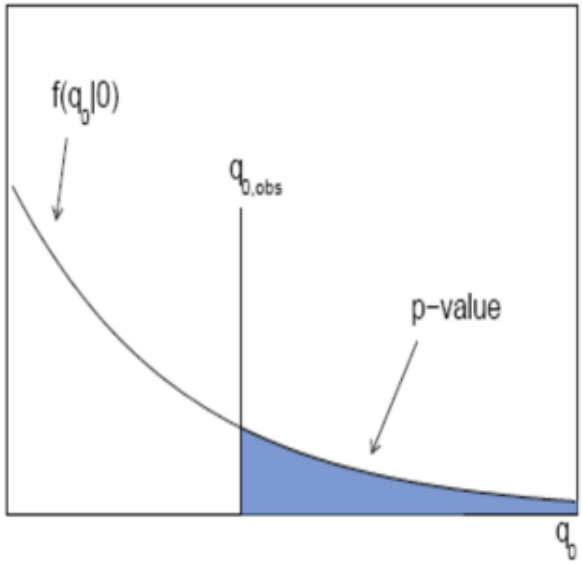
$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

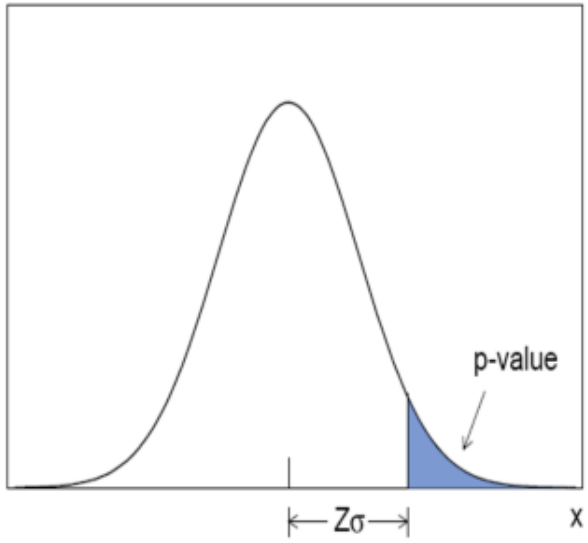
Sei $b = 0.5$, und wir beobachten $n_{obs} = 5$. Sollten wir eine Evidenz für Entdeckung verkünden? Der P-Wert für die Hypothese $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$

P-Wert und Signifikanz

Oft wird die Signifikanz Z definiert, als die Anzahl der Standardbweichungen, die eine Gaussverteilte Zufallsvariable in ein Richtung fluktuieren muss, um den selben P-Wert zu liefern





$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

`1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$

`TMath::NormQuantile`

Konvention für Entdeckung: Eine Signifikanz von 5 entspricht P = 2.87×10^{-7}

Likelihoodverhältnis aus Neyman-Pearson-Lemma

Die Likelihood n zu beobachten unter H_0 (s=0,b) ist: $L_b = \frac{b^n}{n!} e^{-b}$

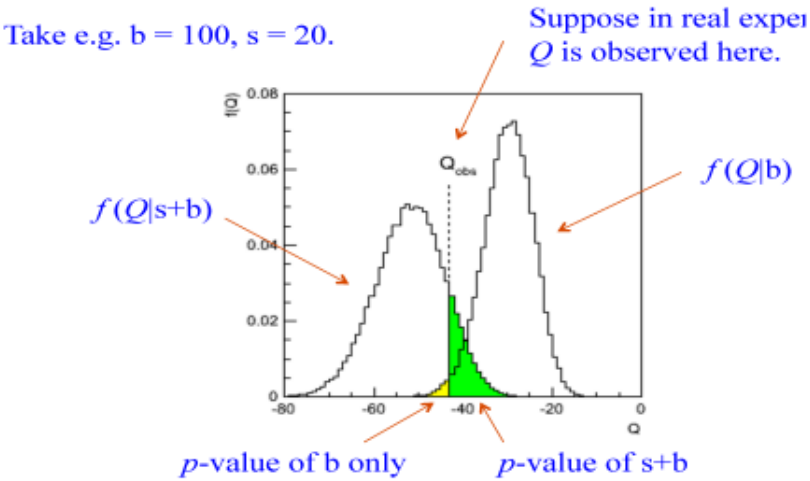
Die Likelihood n zu beobachten unter H_1 (s,b) ist: $L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)}$

Dies sind einfache Hypothesen. Das Neyman-Pearson-Lemma sagt uns, dass die optimale Testastatistik geben ist durch:

$\frac{L_{s+b}}{L_b}$ oder monotone Funktionen $\ln \frac{L_{s+b}}{L_b} = n \ln\left(1 + \frac{s}{b}\right) - s$

Likelihoodverhältnis ist monotone Funktion von n. WDF ist also ebenfalls Poissonverteilung.
Zählrate ist optimale Teststatistik.

Oft verwendet: $Q = -2 \ln \frac{L_{s+b}}{L_b}$



Likelihoodquotient/Profillikelihood

Eben: Signalrate fixiert auch unter Alternativhypothese.
Nun: Finde beste Signalanzahl unter H_1 über ML-Methode

Teststatistik: $\lambda(s) = \frac{L(s)}{L(\hat{s})}$

Zähler: Likelihood unter H_0 (s fixiert, für Entdeckung s=0)
Nenner: Likelihood unter H_1 (s geschätzt aus Daten)

$$\hat{s} = n - b$$

Teststatistik für Entdeckung (s=0 im Zähler):

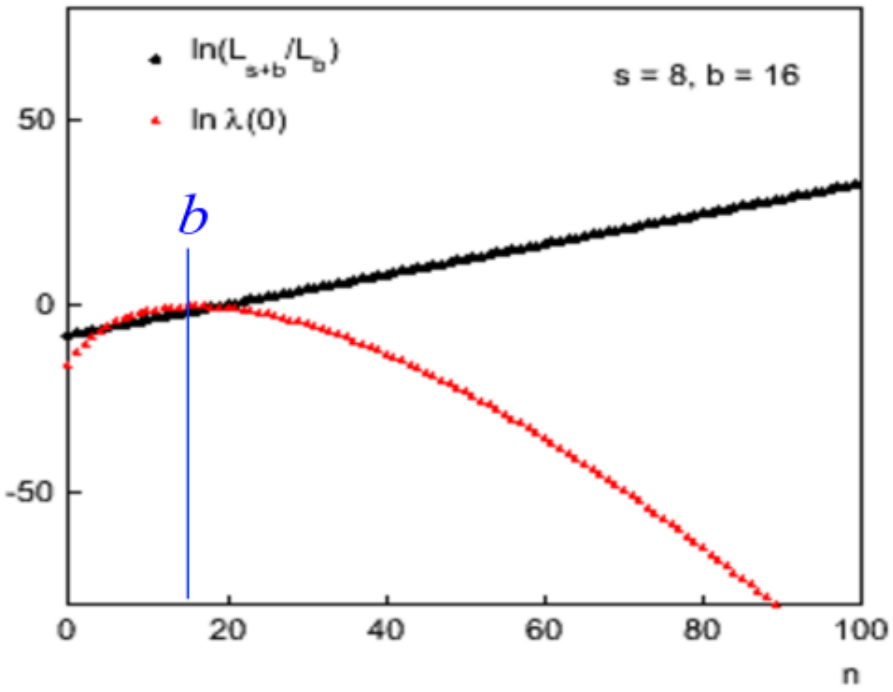
$$\ln \lambda(0) = n \ln(b) - b - n \ln n + n$$

λ aus [0,1]: 1 (0) gute (schlechte) Übereinstimmung mit H_0
$\ln\lambda$ aus [0, -unendlich]: 0 gute Übereinstimmung mit H_0

Vergleich der Teststatistiken

aus Neyman-Pearson-Lemma (einfache Hypothesen):

$$\ln \frac{L_{s+b}}{L_b} = n \ln \left(1 + \frac{s}{b}\right) - s$$



aus Likelihoodquotient (zusammengesetze Alternativehyp. H_1)

$$\ln \lambda(0) = n \ln(b) - b - n \ln n + n$$

wenn wir als Hinweis auf Signal (Abweichung von H_0) nur n>b betrachten

dann sind beide monoton in *n* und führen auf gleichen Test

ln λ(0) hat gute Eigenschaften für nicht zu kleine Ereigniszahlen und erlaubt Berücksichtigung syst. Fehler

Unsicherheiten auf den Untergrund

meist ist der Untergrund nicht genau bekannt → syst. Unsicherhiet

wie wird diese in Berechnung von P-Wert berücksichtigt?

a) Bayesianische Methode über Marginalisierung
b) Profillikelihood als Störparameter

a) Annahme: Gaussfehler auf Untergrund (b_0 nomineller Wert, σ_b Fehler)

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_0)^2/2\sigma_b^2}$$

Bestimmung der WDF für Q $\quad Q = -2\ln\frac{L_{s+b}}{L_b}$

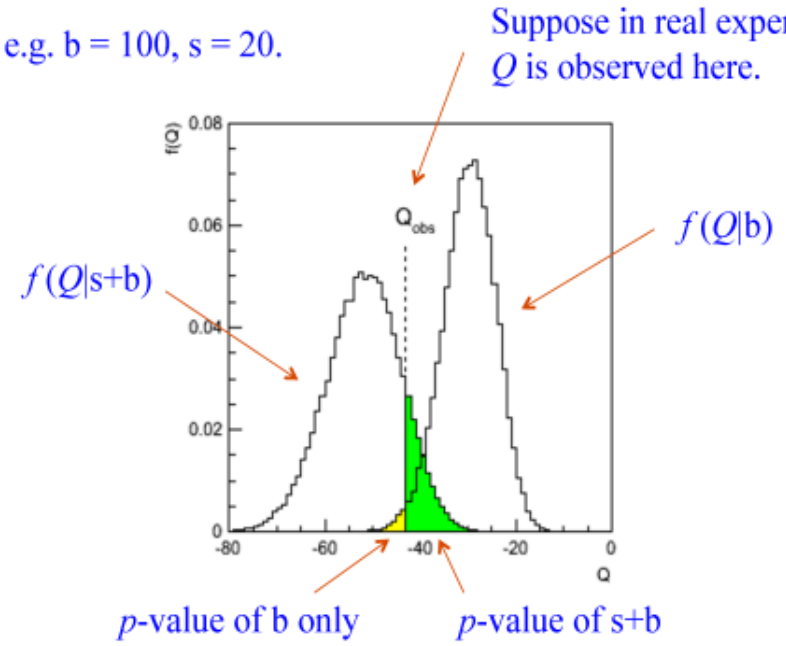
mit MC-Simulationen: würfele b entsprechend Gauss und bestimme Q:

$$f(Q) = \int f(Q|b)\pi(b)\,db$$

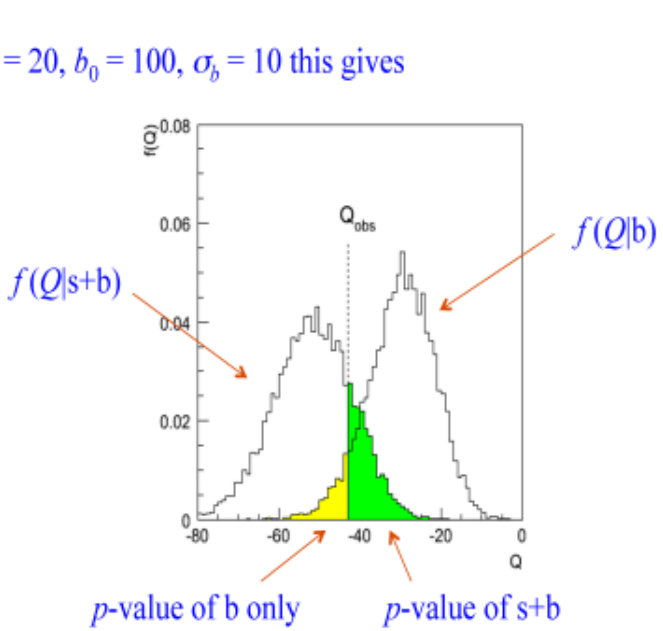
Unsicherheiten auf den Untergrund: Bayesianisch

Einfluss des systematischen Fehlers auf WDF für Teststatistik Q

ohne systematische Unsicherheit



mit systematischer Unsicherheit



Verbreiterung der WDFs → schlechtere Trennkraft
→ größere P-Wert, kleinere Signifikanz

Unsicherheiten auf den Untergrund: Frequentistisch

Annahme: Kontrollmessung m für Untergrund τb in Kontrolldatensatz
Skalierungsfaktor τ bekannt ($\tau >> 1$)

Messung von n in Signalregion folgt: $n \sim \text{Poisson}(s+b)$

Messung von m in Kontrollregion folgt: $m \sim \text{Poisson}(\tau b)$

Gemeinsame Likelihoodfunktion: $$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Teststatistik = Profillikelihood (Störparameter b)

$$\lambda(s) = \frac{L(s,\hat{\hat{b}})}{L(\hat{s},\hat{b})}$$

Zähler: ML-Schätzer für b unter H_0

Nenner: ML-Schätzer für s und b

-2ln $\lambda(0)$ folgt Chi-Quadrat-WDF mit 1 Freiheitsgrad (aus Satz von Wilks)
für nicht zu kleine Ereigniszahlen, d.h. WDF ist approximativ bekannt

Profilelikelihoodteststatistic für Entdeckung

H_0: nur Untergrund → $\mu=0$ H_1: Signal + Untergrund

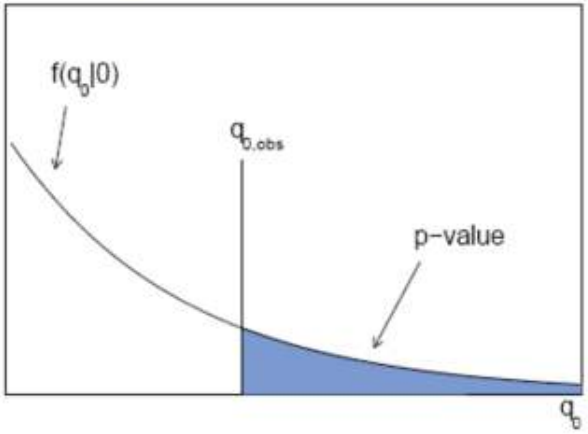
Teststatistik q_0:

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

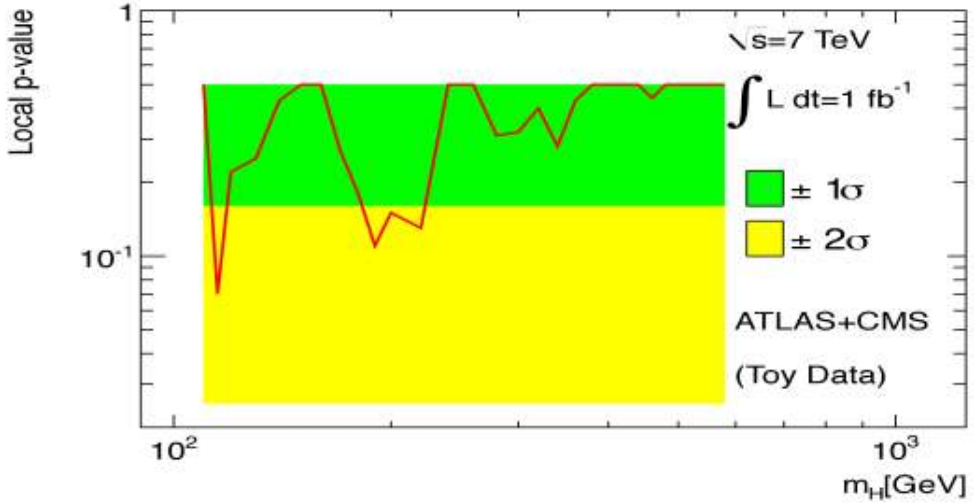
$\lambda(0)$ zwischen 0: H_1 artig und 1: H_0 artig

→ q_0 zwischen 0 und unendlich
0: H_0 artig >> 0 H_1 artig

Einseitiger Test, nur positive Signalstärke = Abweichung von H_0



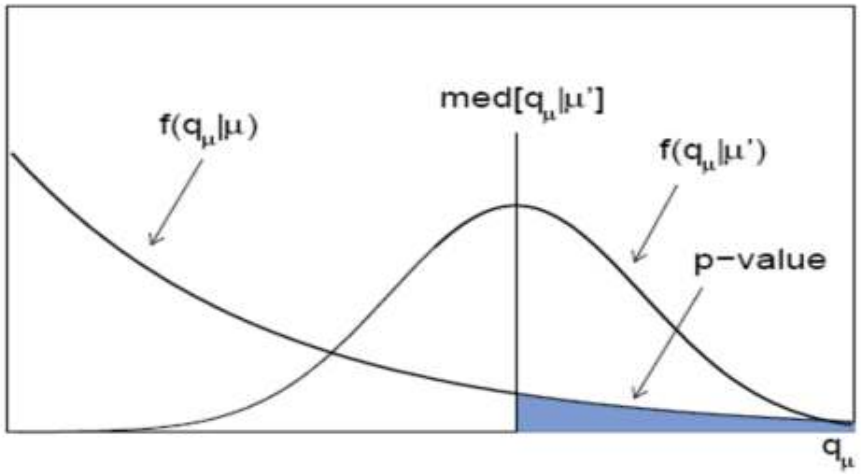
$$p_0 = \int_{q_{0,obs}}^{\infty} f(q_0|0)\, dq_0$$



Erwartete Signifikanz/Sensitivität des Experimentes

Bestimme den Median der Teststatistik unter Alternativhypothese μ'

Berechne den Entsprechende P-Wert



Wir brauchen WDFs:
Näherungsweise aus Sätzen von Wilks und Wald

für Bewertung der Daten → WDF für Teststatistik unter Nullhypothese

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2} \qquad Z = \Phi^{-1}(1-p_0) = \sqrt{q_0}$$

für Bestimmung der Sensitivität → WDF für Teststatistik unter Alternativhyp.

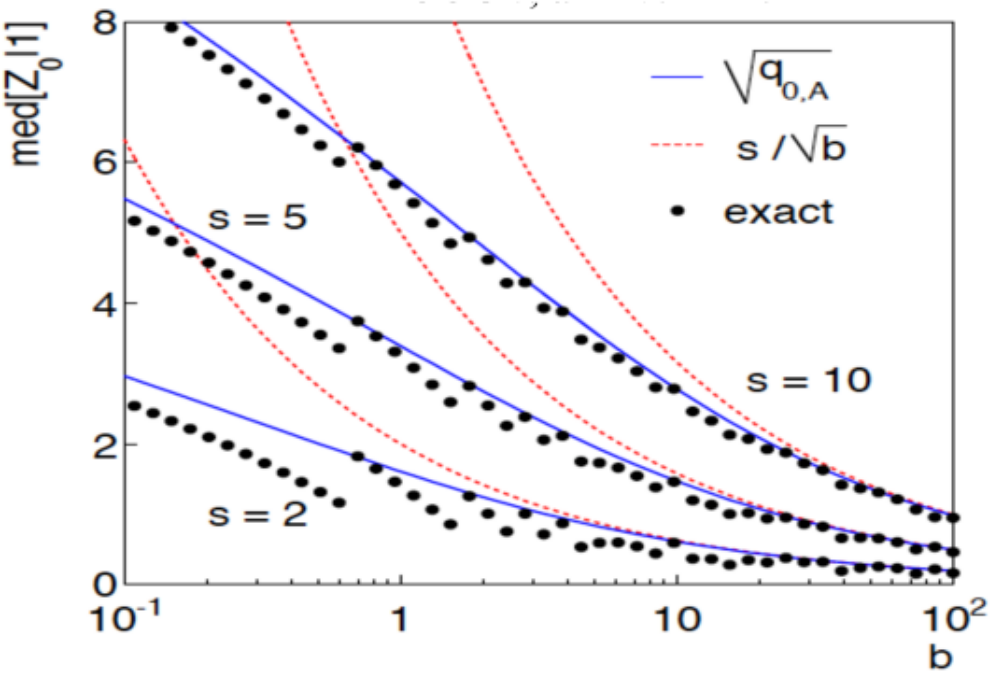
$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

Qualität der Näherung für Zählexperiment

Für Sensitivität: ersetze n durch s+b (in Literatur Asimov-Daten genannt)

Gaussnäherung: $Z_0 = \Phi^{-1}(1-p_0) = \frac{x_{\mathrm{obs}} - b}{\sqrt{b}} \quad \mathrm{median}[Z_0|s+b] = \frac{s}{\sqrt{b}}$

Wald+Wilks-Näherung: $Z_0 \approx \sqrt{q_0} = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \quad \mathrm{median}[Z_0|s+b] \approx \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}$



exakte Werte von MC zeigt „Sprünge“ wegen diskreter Natur der Daten

Asimov-Näherung gut für gr. Bereich an s und b

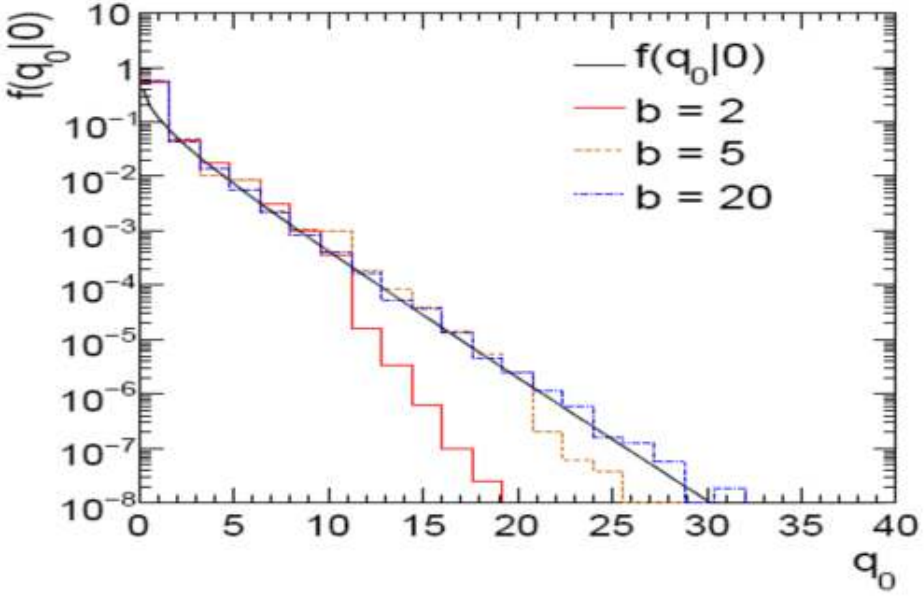
s/sqrt(b) nur gut für s<<b

Qualität der Näherung für Zählexperiment

$$n \sim \text{Poisson}(\mu s + b)$$
$$m \sim \text{Poisson}(\tau b)$$

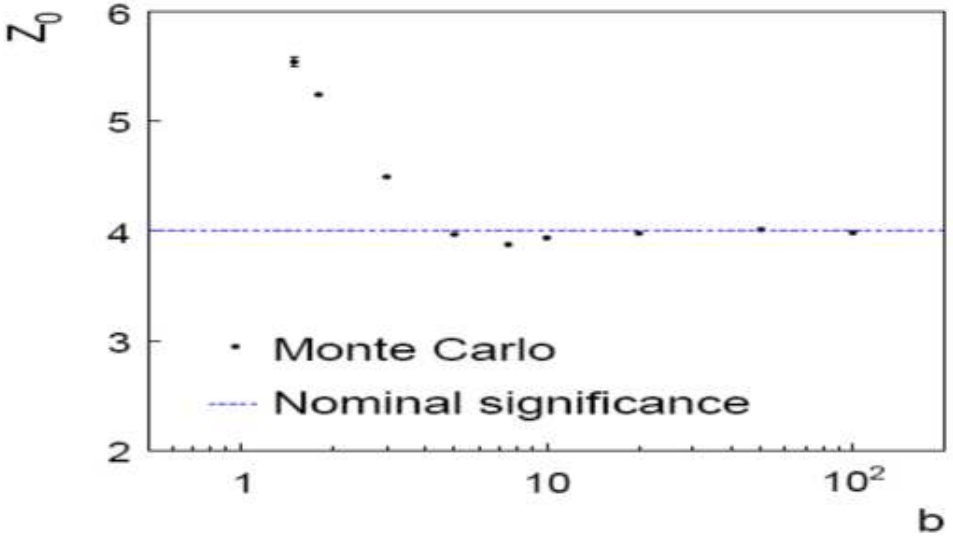
Here take $\tau = 1$.

Asymptotische Formel der WDF gute Näherung bis 5σ bereits für b ~20



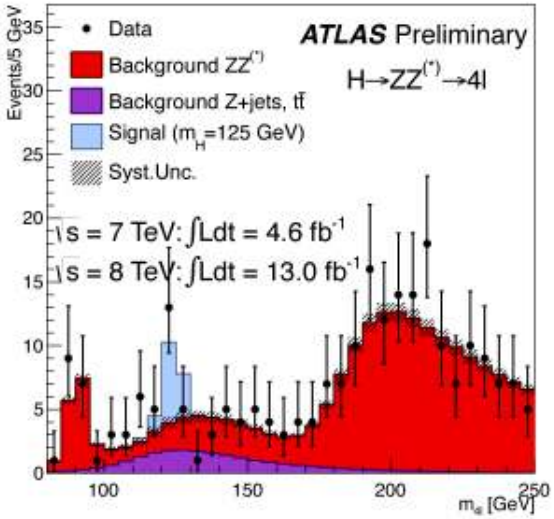
Für sehr kleine b unterschätzt asymptotische Formel die Signifikanz.

Dann leichte Überschätzung bevor es gegen MC konvergiert



Mehr als Zählen: Form von Verteilungen

Oft Signal = lokale Überhöhung im Spektrum
a) definiere eine Massenfenster
→ nicht optimal
b) nutze gesamtes Massenspektrum
= Form der Verteilungen $f(_x)$



Bilde Likelihoodfunktionen für beide Hypothesen

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^{n} \left(\pi_s f(\mathrm{x}_i|\mathrm{s}) + \pi_b f(\mathrm{x}_i|\mathrm{b})\right) \qquad L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^{n} f(\mathbf{x}_i|\mathrm{b})$$

Neyman-Pearson-Teststatistik (bei LEP verwendet):

$$Q = -2\ln\frac{L_{s+b}}{L_b} = -s + \sum_{i=1}^{n} \ln\left(1 + \frac{s}{b}\frac{f(\mathbf{x}_i|\mathrm{s})}{f(\mathbf{x}_i|\mathrm{b})}\right)$$

Mehr als Zählen: Profillikelihoodteststatstik

Beobachtetes Massenspektrum: $\mathbf{n} = (n_1, \ldots, n_N)$

Erwartetes Massenspektrum in jedem Bin: $E[n_i] = \mu s_i + b_i$

μ beschriebt die Signalstärke m=0 nur Untergrund

$$s_i = s_{\text{tot}} \int_{\text{bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx \qquad b_i = b_{\text{tot}} \int_{\text{bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx$$

Messung in Kontrollregion: $\mathbf{m} = (m_1, \ldots, m_M)$

Erwartung in Kontrollregion abhängig von Störparameteren θ

$$E[m_i] = u_i(\boldsymbol{\theta}) \qquad (\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}})$$

Profillikelihoodteststatstik (2)

Gesamtlikelihoodfunktion abhängig von μ und Störparametern θ

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

Teststatistik:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

μ fixiert unter H_0

$\hat{\hat{\boldsymbol{\theta}}}$ Maximum Likelihood-Schätzer unter H_0

$\hat{\mu}, \hat{\boldsymbol{\theta}}$ Maximum Likelihood-Schätzer unter H_1

im Gegensatz zu Neyman-Pearson-Teststatistik wird Signalstärke μ unter Alternativhypothese nicht fixiert.

Der „Look Elsewhere Effect“ (LLE)

Bisher: lokaler P-Wert/ Signifikanz= Wkt. einen Überschuss an einer speziellen Stelle im Massenspektrum zu beobacxhten, da wir M_H in der Alternativhypothese fixiert haben

$$t_{\text{fix}} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)} \qquad p_{\text{fix}} = \int_{t_{\text{fix,obs}}}^{\infty} f(t_{\text{fix}}|0) dt_{\text{fix}} \qquad Z_{\text{fix}} = \Phi^{-1}(1 - p_{\text{fix}})$$

Nun: Wkt. einen solchen Überschuss irgendwo im Massenspektrum zu finden.
→ Verwende Testmasse als einen Störparameter in neuer Teststatistik

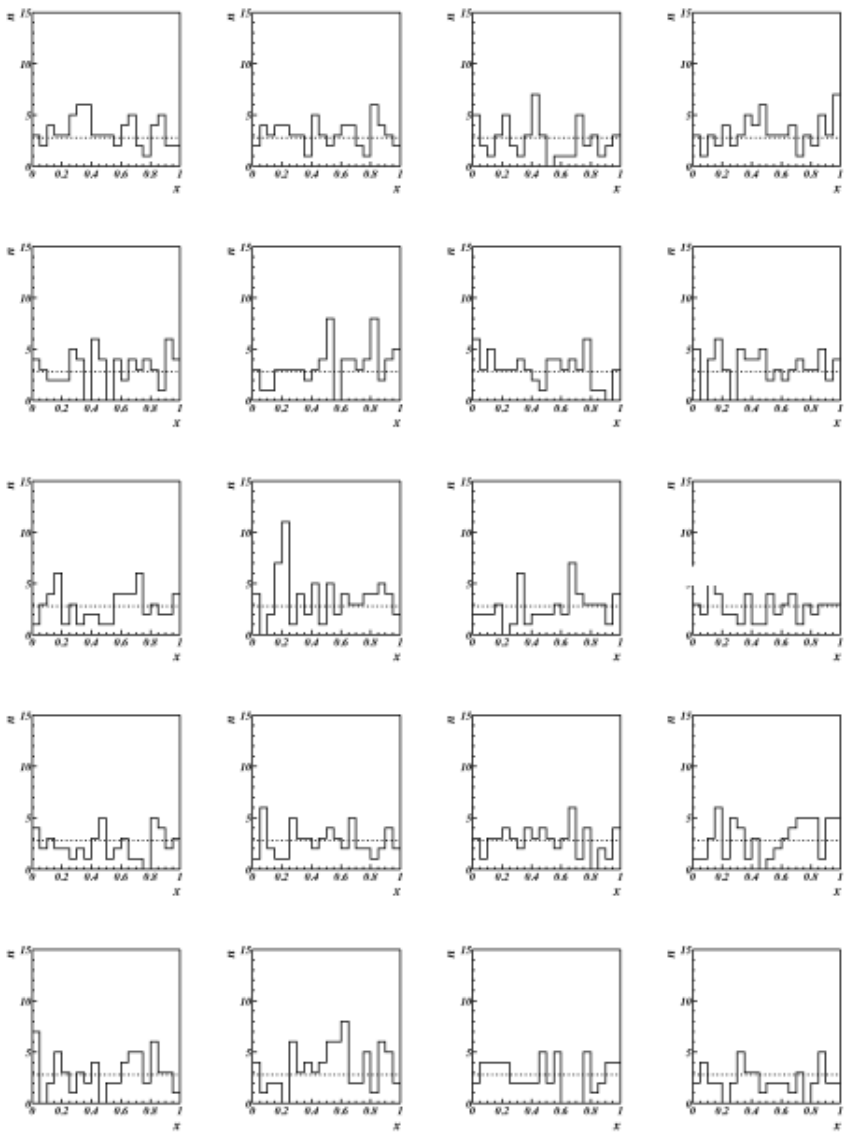
$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})} \qquad p_{\text{float}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}}|0) dt_{\text{float}}$$

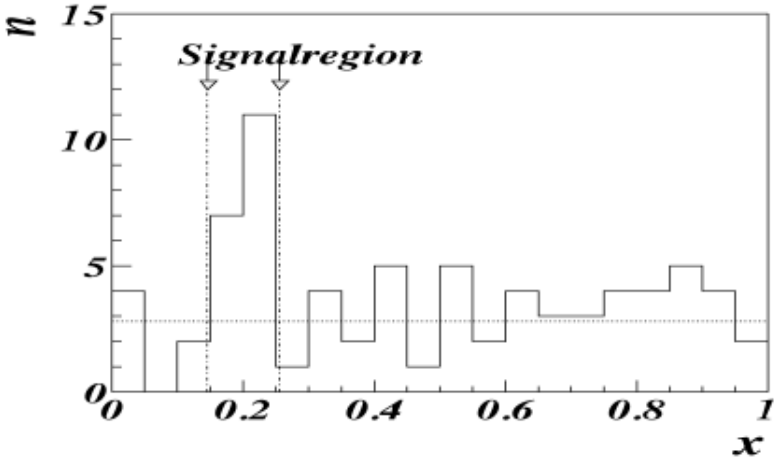
p_{float} als wird globaler P-Wert genannt. Berechnung sehr aufwendig. Viele MC Exp.

$$F_{\text{trials}} \equiv \frac{p_{\text{float}}}{p_{\text{fix}}}$$

Versuchsfaktor „Trial Factor“ ~ Anzahl unabhängiger Suchregionen im Massenspektrum
Kann approximativ mit wenig MC-Exp bestimmt werden.

Der„Look Elsewhere Effect“ (LLE)





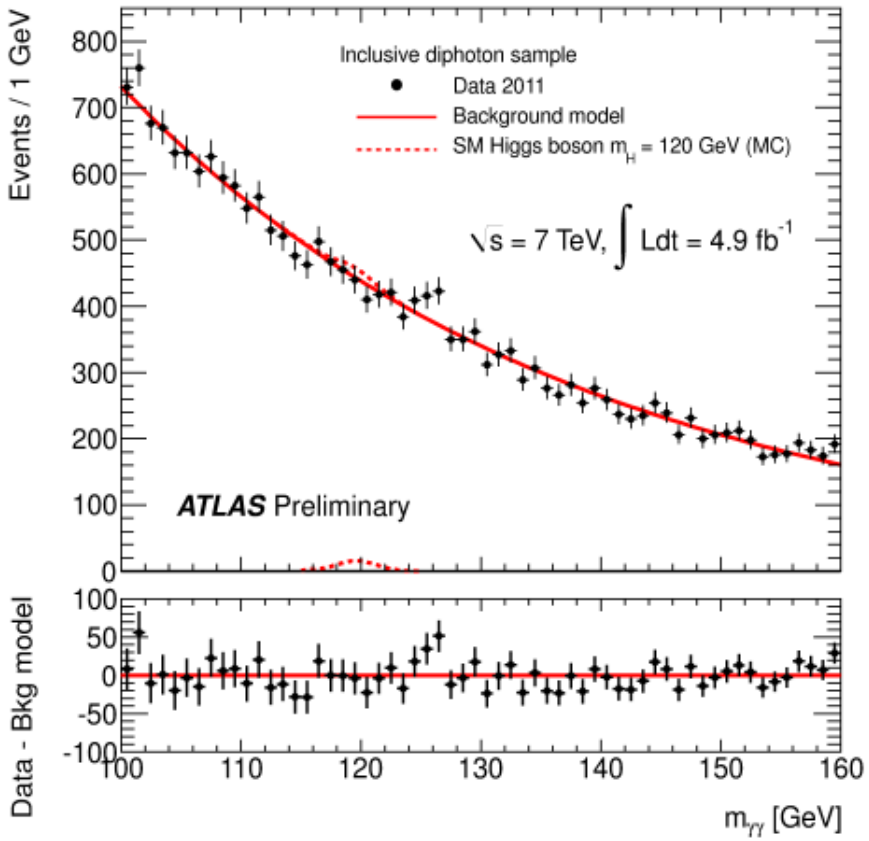
$\nu_b = 2 \times 2.8 = 5.6.$ $n_{Daten} = 18$

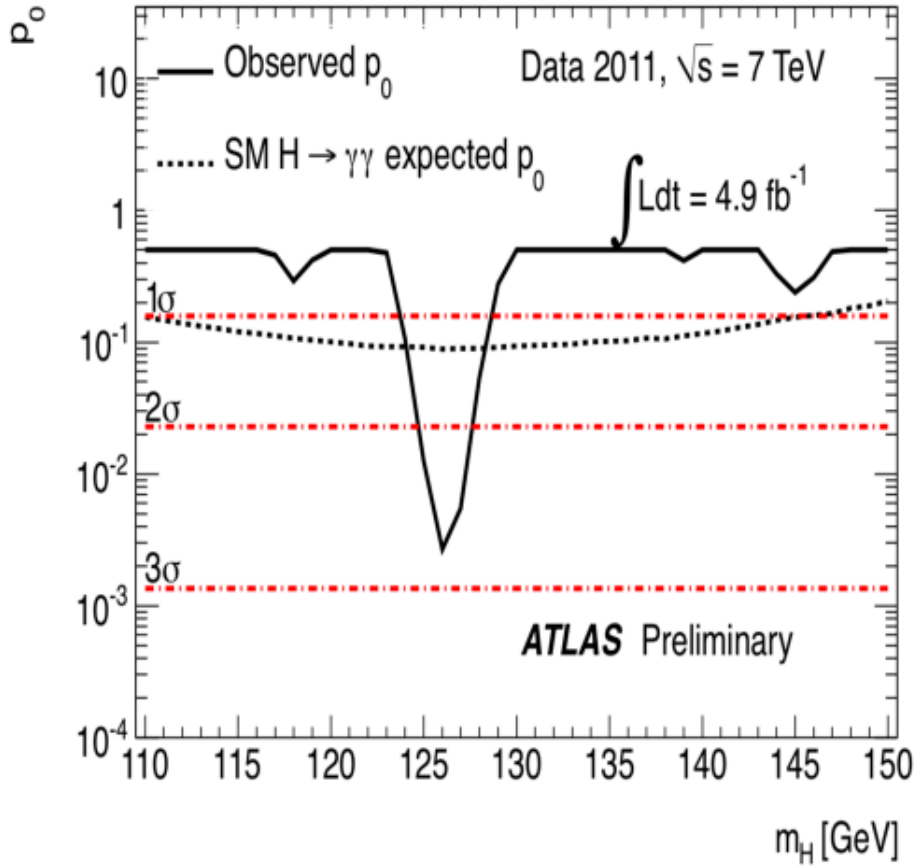
lokaler P-Wert 2.4×10^{-5}

40 Bins,
20 Experimente=
Histogramme
Signalbreite = 2 Bins
→ Versuchsfaktor“ ~400

globalerP-Wert $20 \times 20 \times 2.4 \times 10^{-5} \approx 1\%.$

Ein Beispiel: H→2 γ in ATLAS





Maximale Abweichung von der "Nur-Untergrundhypothese" bei m_H~126GeV:
Lokaler p_0-Wert 0.27% (2.8σ) Global p_0-Wert (inklusive LEE) ~7% (1.5σ)
Erwartete Signifkanz für ein SM Higgs: ~ 1.4σ bei 126 GeV