Einführung in die Statistik und Datenanalyse (4)

Martin Völkl
martin.voelkl@uni-tuebingen.de

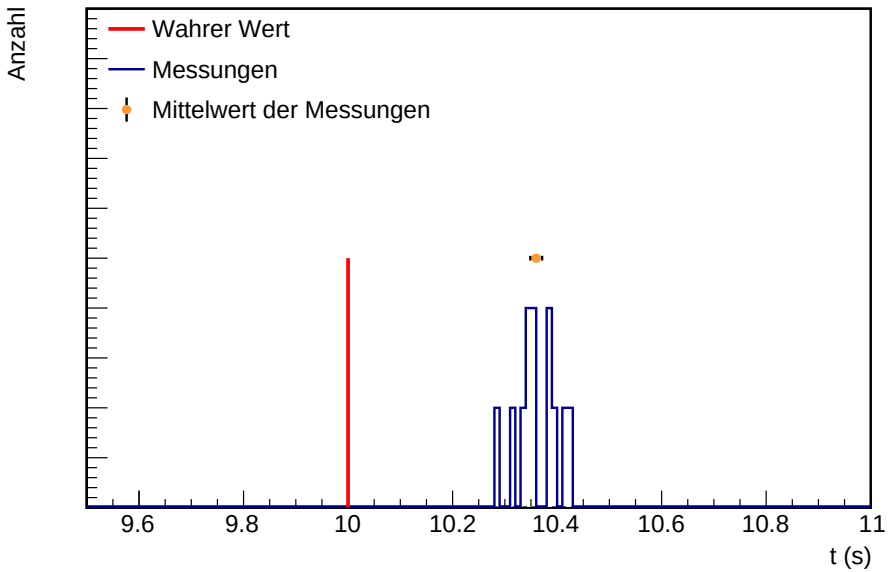
Universität Tübingen
2018-03-20

Aufbau der Vorlesung

1. Einführung und systematische Unsicherheiten
2. Werkzeugkiste
3. Wahrscheinlichkeit und Unsicherheit
4. **Frequentistische Methoden**
   - Schätzfunktionen
   - Maximum-Likelihood
   - Regressionsanalyse
   - p-Wert
   - Vertrauensintervalle
5. Bayesianische Methoden

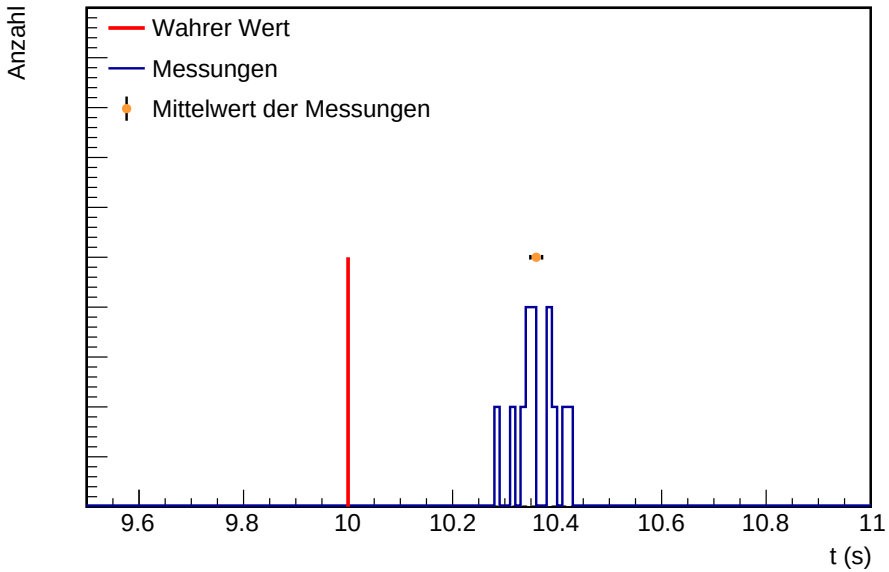
Wie gut ist der Mittelwert?

- Beispiel: Zeitpunkt eines Ereignisses per Stoppuhr bestimmen
- Messwert Beeinflusst durch Reaktionszeit und Fluktuationen
- Anwendung von $\sigma \approx \sqrt{\frac{\sum(x_i - \bar{x})^2}{N(N-1)}}$ irreführend
- Wie damit umgehen?



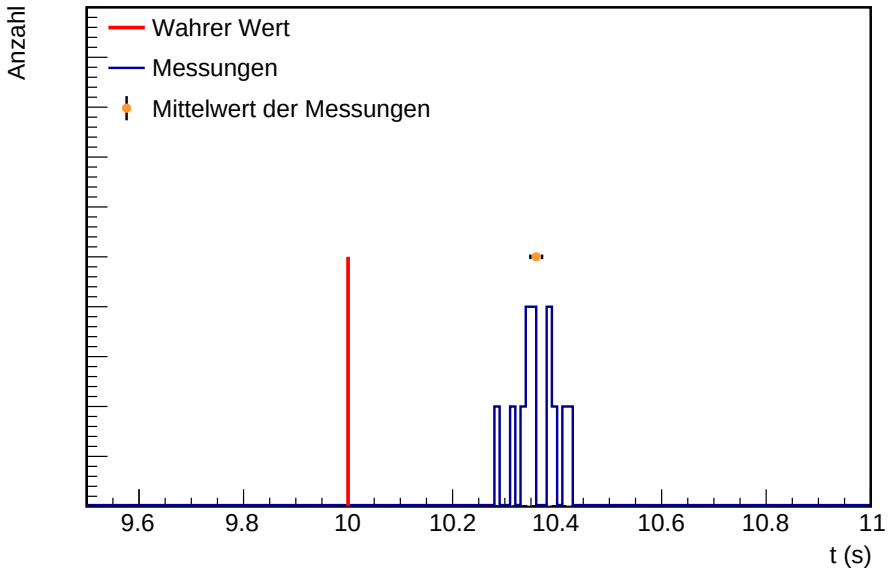
Schätzfunktionen

- Schätzfunktion (*estimator*) ist eine Funktion der Daten, die einen Schätzwert zurückgibt
- Bezeichnet durch ^-Symbol
- $\hat{\lambda} = \frac{\sum x_i}{N}$, der Mittelwert der Messwerte ist Schätzfunktion für den Parameter der Poissonverteilung



Schätzfunktionen – Charakterisierung

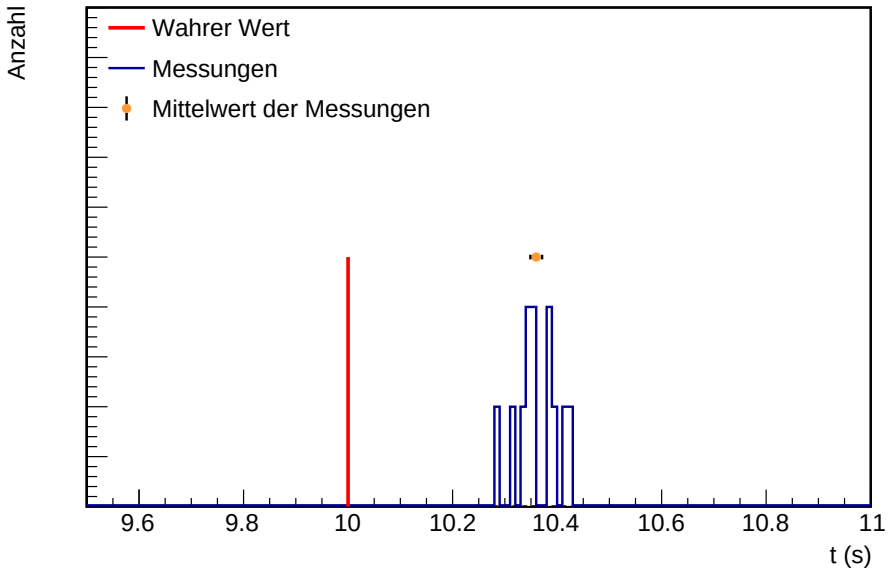
1. *Konsistenz:* Bei mehr und mehr Daten, Konvergenz zu wahrem Wert
2. *Erwartungstreue:* (*Bias*) Mittelwert von wiederholten Messungen Konvergiert gegen wahren Wert
3. *Effizienz/Varianz:* Geschwindigkeit der Konvergenz bei mehr Daten
4. *Robustheit:* Falls einige Annahmen nur ungefähr gelten, ist das Ergebnis immer noch gut



- Konsistenz typischerweise erfüllt
- Kompromiss was die Anderen Eigenschaften angeht

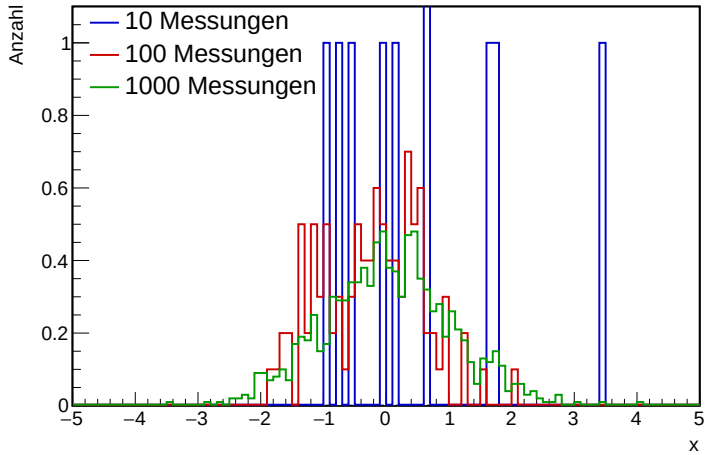
Beispiel – Mittelwert

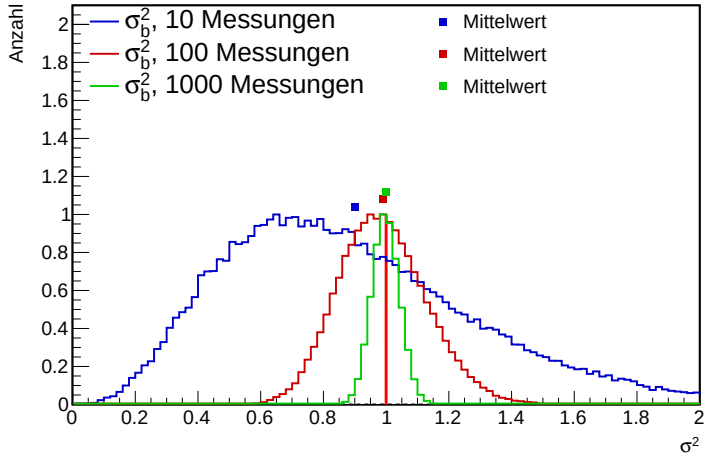
- Für Konsistenz: Kontrollmessungen für Reaktionszeit
- Definiere korrigierten Mittelwert als Schätzfunktion (Mittelwert - Reaktionszeit)
- Sonderfall bei Mittelwert: Konsistent und Erwartungstreue äquivalent
- Wenn möglich: Korrigiere Bias



Beispiel – Varianz

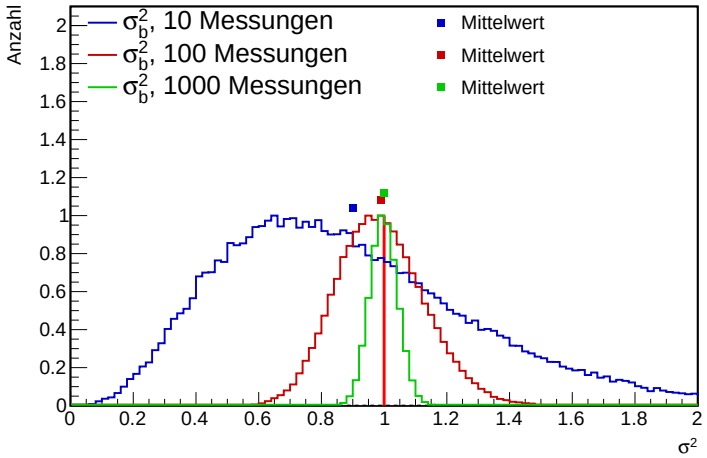
- Gesucht ist Varianz einer (Normal-)Verteilung
- Idee: $\hat{\sigma}_b^2 = \frac{\sum(x_i - \bar{x})^2}{N}$
- Schätzfunktion ist konsistent, hat aber einen Bias

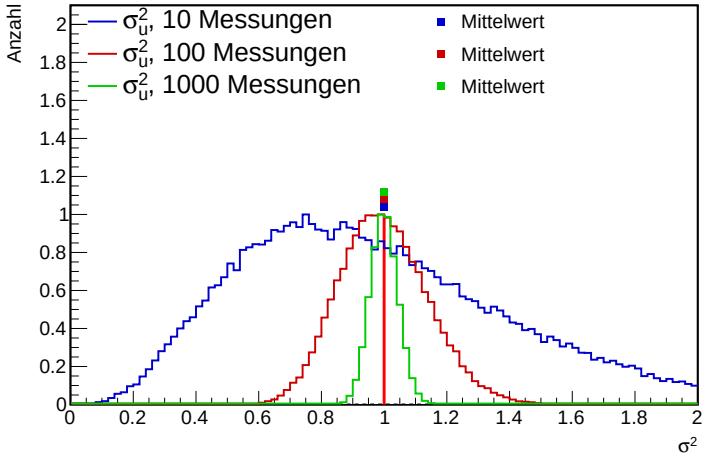




Beispiel – Varianz (2)

- $\hat{\sigma}_b^2 = \frac{\sum(x_i - \bar{x})^2}{N}$ hat Bias
- $\hat{\sigma}_u^2 = \frac{\sum(x_i - \bar{x})^2}{N-1}$ hat keinen Bias
- Beide konsistent
- $\hat{\sigma}_b^2$ hat die niedrigere Varianz bei Wiederholung des Experiments
- Für gegebene Messung erwartungstreue Schätzfunktion mit niedrigster Varianz hat Namen MVUE (minimum variance unbiased estimator)
- Oft nicht leicht zu finden





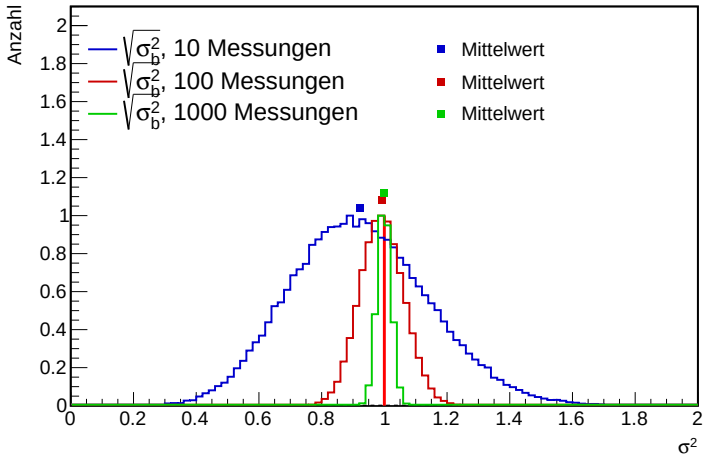
Fehlerfortpflanzung

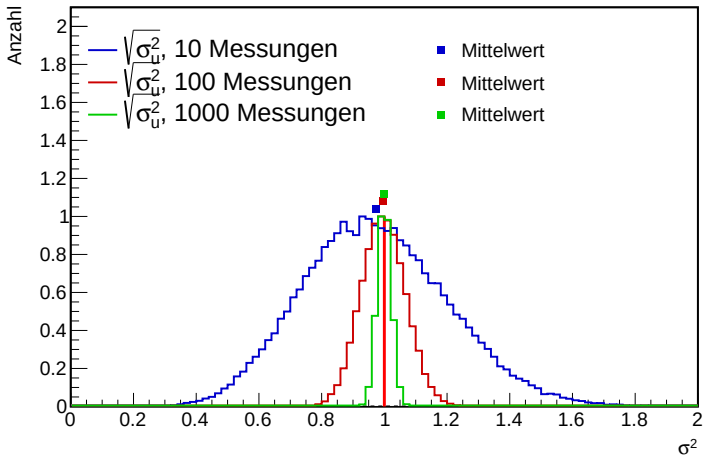
- Schätzfunktion $\hat{x}$ gibt Wissen über x
- Was wissen wir über $f(x)$?
- Beispiel: Was ist σ, wenn wir $\hat{\sigma}^2$ bestimmt haben?
- Idee $f(\hat{x})$
- Grundsätzlich: Suche Schätzfunktion $\hat{f}(x)$
- Untersuche Konsistenz, Varianz und Bias

- Für lineare Funktionen: $\hat{f}(x) = f(\hat{x})$ erhält Erwartungstreue $\rightarrow$ gute Näherung für kleine Unsicherheit des Schätzers
- (Im Prinzip Gauß'sche Fehlerfortpflanzung)
- Wenn $\hat{x}$ Erwartungstreu, kann $f(\hat{x})$ dennoch einen Bias haben

Beispiel – Standardabweichung

- Zur Erinnerung:
  - $\hat{\sigma}_b^2 = \frac{\sum(x_i - \bar{x})^2}{N}$ hat Bias
  - $\hat{\sigma}_u^2 = \frac{\sum(x_i - \bar{x})^2}{N-1}$ hat keinen Bias
- Suche Schätzer für σ
- Versuche: $\hat{\sigma} = \sqrt{(\hat{\sigma^2})}$
- Erwartungstreuer Schätzer für σ:
  $\hat{\sigma} = \frac{\sum(x_i - \bar{x})^2}{N-1} \cdot \sqrt{\frac{N-1}{2}} \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)}$
- Gilt auch nur für Normalverteilung
- Beispiel: Teilchen erzeugt Signal in 4 Detektoren mit unterschiedlichen Funktionsweisen – MVUE für Masse?
- Nicht praktikabel





Erkenntnis

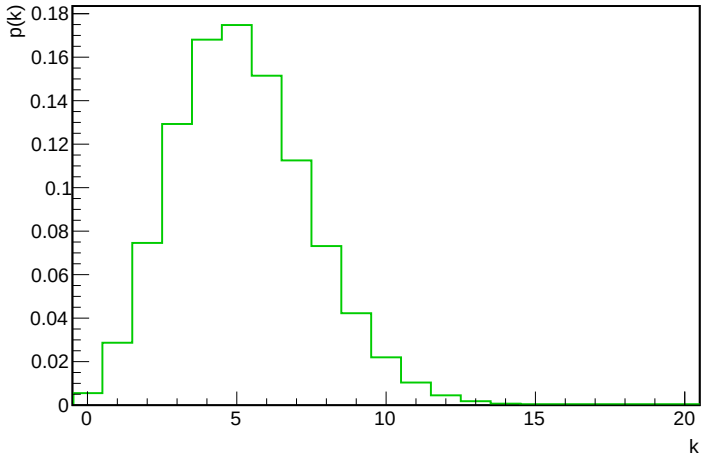
- Mittelwert der Daten ist erwartungstreue Schätzfunktion des Mittelwertes einer beliebigen Verteilung
- Aber nur für Normalverteilung bester Schätzer (minimale Varianz)
- $\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{N-1}$ ist Erwartungstreuer Schätzer für Varianz einer beliebigen Verteilung
- Aber nur für Normalverteilung bester Schätzer (minimale Varianz)
- MVUE für Komplexere Probleme oft zu umständlich

Maximum Likelihood

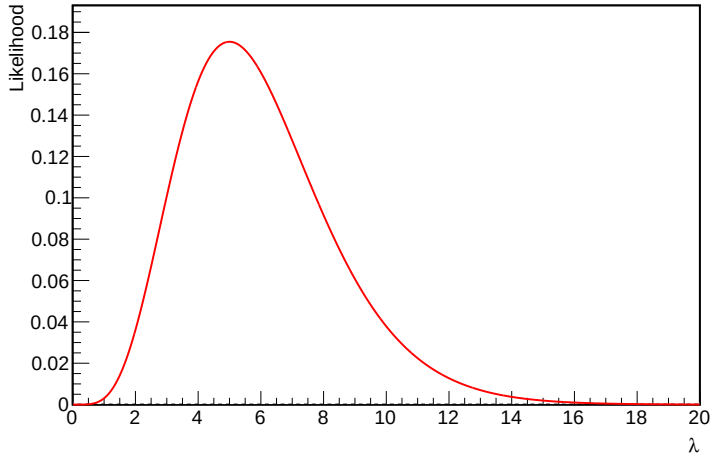
- Daten $\vec{d}$ und Parameter $\vec{\lambda}$
- Kennen $p(\vec{d}|\vec{\lambda})$ (z.B. Poisson)
- "Wahrscheinlichkeit der Daten für ein Parameterset"
- Daten fix, Parameter variabel: "Likelihood"
- $\mathcal{L}(\vec{\lambda}|\vec{d}) \equiv p(\vec{d}|\vec{\lambda})$
- Parameterset, für das die Daten die höchste Wahrscheinlichkeit haben: Maximum Likelihood
- Schätzfunktion mit besonderen Eigenschaften
- Bayes: Verteilung interessant, frequentistisch das Maximum

Maximum Likelihood – Poisson

- Poissonverteilung $p(k|\lambda) = \frac{\lambda^k \ e^{-\lambda}}{k!}$
- Maximum bei $\lambda = k$
- Also $\hat{\lambda}_{\mathrm{ML}} = k$
- Schätzfunktion ist erwartungstreu



Poissonverteilung, $\lambda = 5.2$



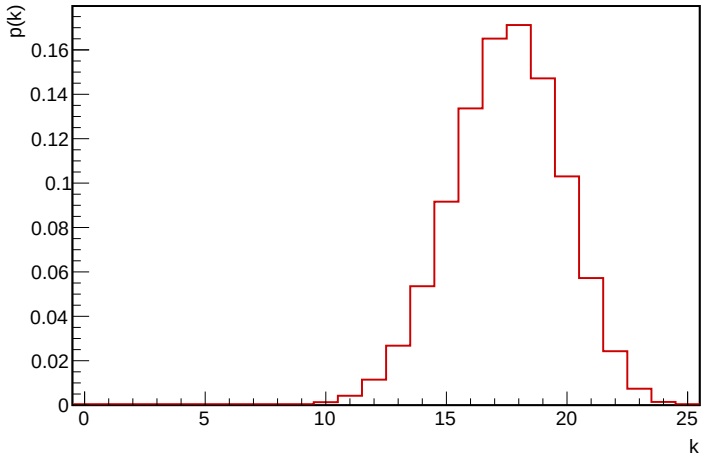
Poissonlikelihood, $k = 5$

Maximum Likelihood – Binomialverteilung

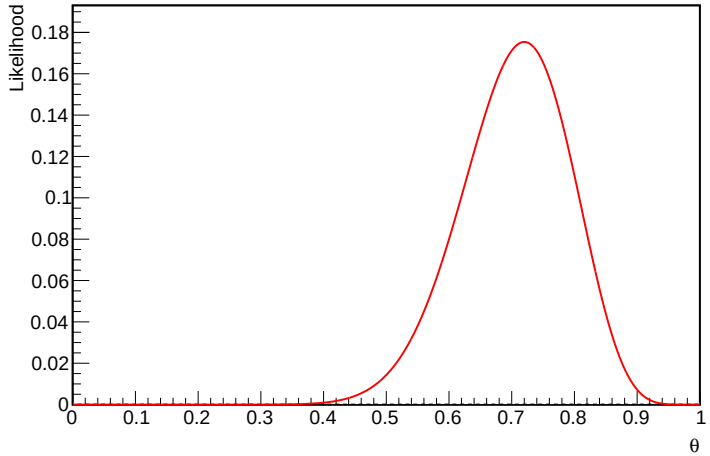
- Binomalverteilung

$$p(k|\theta, N) = \begin{pmatrix} N \\ k \end{pmatrix} \theta^k (1-\theta)^{N-k}$$

- Maximum bei $\lambda = k/n$
- Also $\hat{\theta}_{\mathrm{ML}} = k/N$
- Schätzfunktion ist erwartungstreu



Binomialverteilung, $\theta = 0.7$, $N = 25$



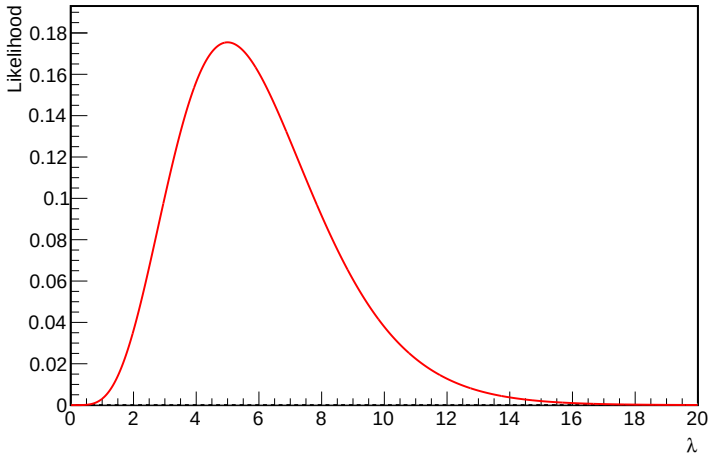
Binomial–likelihood, $k = 18$, $N = 25$

Maximum Likelihood – Wiederholte Messung von Poisson

- Produkt von Poissonverteilungen

$$\mathcal{L}(\lambda) = p(k_1 \dots k_N|\lambda) = \prod_i \frac{\lambda^{k_i}\ e^{-\lambda}}{k_i!}$$

- $\frac{\mathrm{d}\log\mathcal{L}}{\mathrm{d}\lambda} = 0$, dann $\lambda = \frac{\sum k_i}{N}$
- Also $\hat{\lambda}_{\mathrm{ML}} = \frac{\sum k_i}{N}$
- Intuitiv: N gleichlange Messungen enthalten die gleiche Information wie eine N–Mal so lange Messung



Poissonlikelihood, $k = 5$

Maximum Likelihood – Normalverteilungen

- Mehrere Messungen von normaverteilter Zufallsvariable x
- Verteilung mit bekanntem, konstantem σ, unbekanntem μ:

$$\mathcal{L}(\mu) = p(x_1 \dots x_N | \mu, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- Zum Rechnen: Ignoriere konstante Faktoren und nutze Monotonie des Logarithmus
- $\hat{\mu}_{\mathrm{ML}} = \frac{\sum x_i}{N}$
- ML Schätzer ist Mittelwert
- Für variierendes σ_i ergibt sich:

$$\hat{\mu}_{\mathrm{ML}} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$$

  – das Ergebnis der gewichteten Summe aus mehreren Messungen

Maximum Likelihood – Normalverteilungen (2)

- Mehrere Messungen von normaverteilter Zufallsvariable x
- Unbekanntes σ und μ:

$$\mathcal{L}(\mu) = p(x_1 \dots x_N | \mu, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

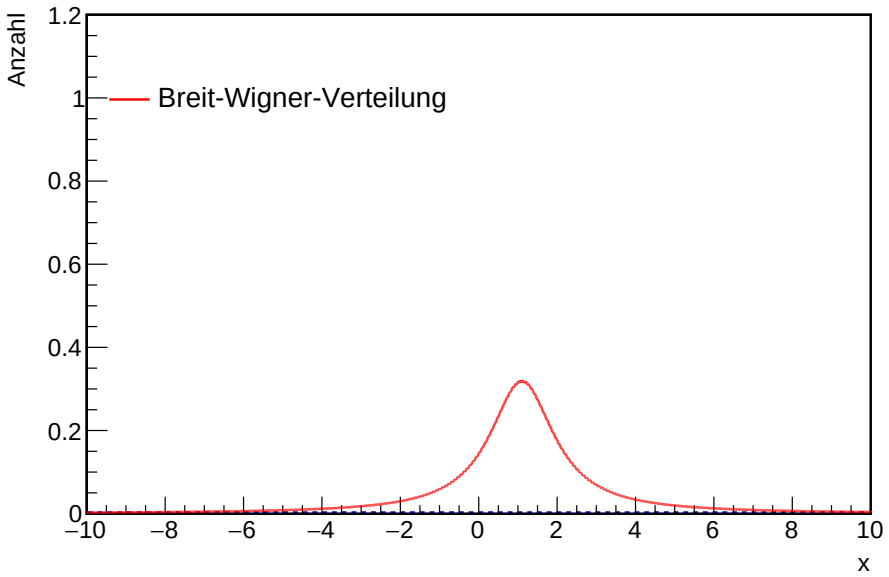
- Selbes Ergebnis für $\hat{\mu}_{\mathrm{ML}} = \frac{\sum x_i}{N}$
- $$\hat{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{N}$$
- ML hier nicht erwartungstreu

Warum Maximum Likelihood?

- Maximum Likelihood ist oft der intuitiv benutzte Schätzer
- Nicht allgemein erwartungstreu
- Warum also ML?

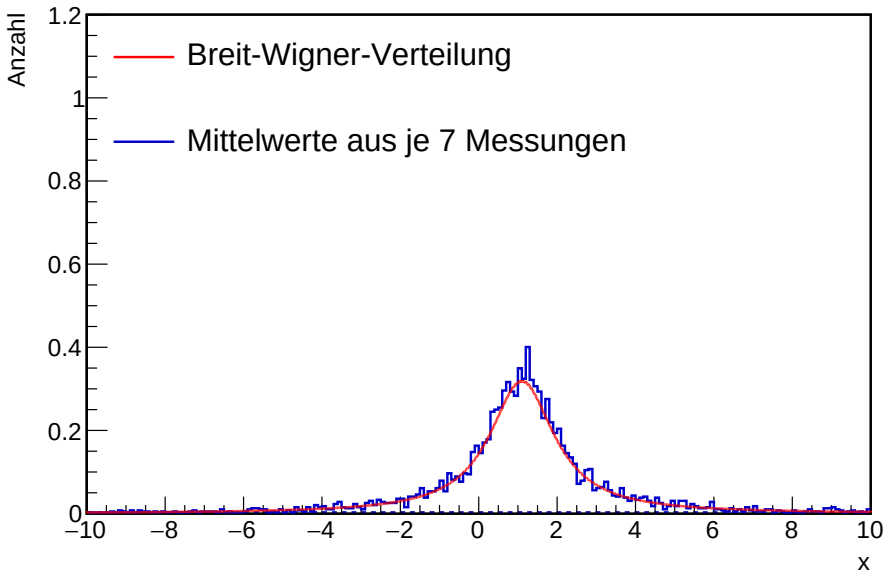
Breit–Wigner-Verteilung

- Messwerte verteilt wie $p(x|\mu) = \frac{\pi^{-1}}{1+(x-\mu)^2}$
- Vorkommen z.B. Verteilung der invarianten Masse von Resonanzen
- Mittelwert und Varianz divergieren
- Intuitive Schätzfunktion: Mittelwert der Messungen



Breit–Wigner-Verteilung (2)

- Mittelwert aus 7 Messungen
- 7 Messungen virtuell mehrfach wiederholt
- Verteilung des Mittelwertes gleich der Verteilung eines Wertes
- Mittelwert verbessert das Ergebnis überhaupt nicht
- Mittelwert–Schätzfunktion hat $\sigma = \infty$
- Schätzfunktion ist "unendlich schlecht"

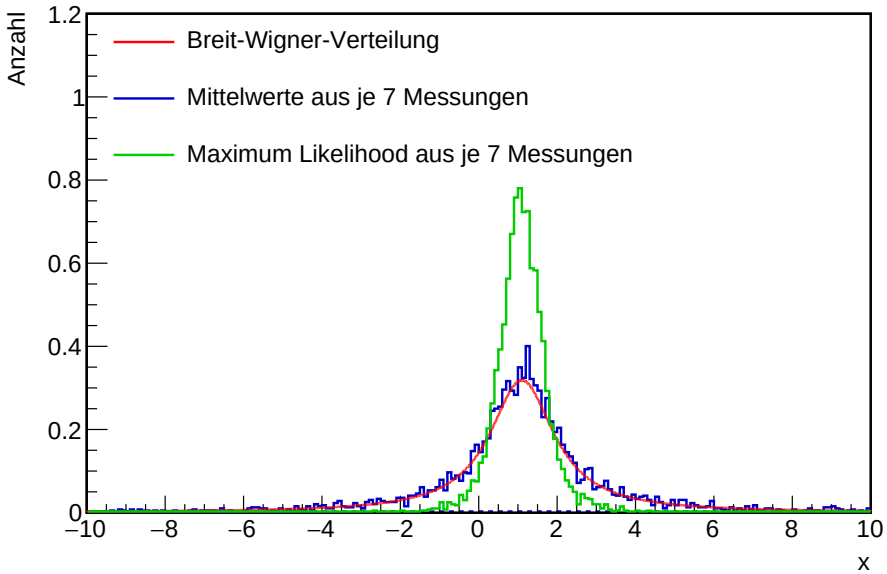


Breit–Wigner-Verteilung (3)

- Likelihood:

$$\mathcal{L}(\mu) = \prod_i \frac{\pi^{-1}}{1 + (x_i - \mu)^2}$$

- Maximierung über $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\mu} = 0$ schwierig (Polynom von Ordnung $2n - 1$)
- Numerische Maximierung hilft
- Maximum–Likelihood Schätzer weit überlegen



Die Cramér–Rao Ungleichung

- Unbekannter Parameter θ und Likelihood $\mathcal{L}(\theta)$
- Beliebige erwartungstreue Schätzfunktion $\hat{\theta}$
- Dann gilt für die Varianz des Schätzers die Ungleichung

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{-\mathrm{E}\left[\frac{\mathrm{d}^2 \log \mathcal{L}(\theta)}{\mathrm{d}\theta^2}\right]}$$

- $E[\,]$ ist der Erwartungswert über alle möglichen Daten
- Schwer auszurechenen. Für Normalverteilung ergibt sich σ^2
- Definiert "bestmögliche Schätzfunktion"

Warum ist die max. Likelihood Methode so gut?

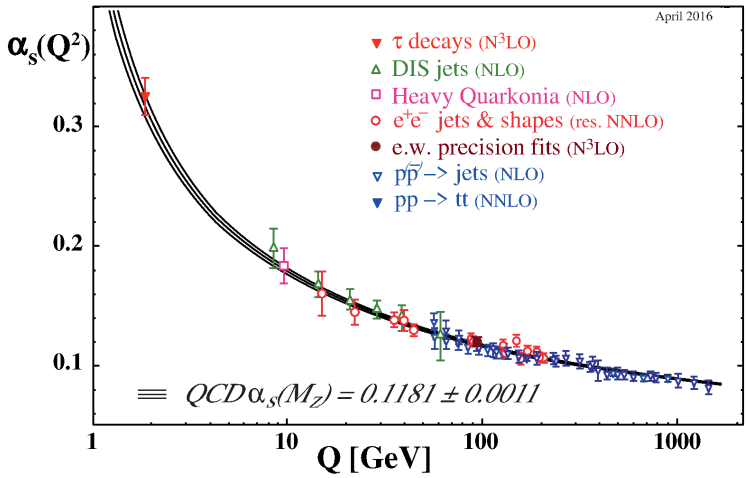
- Im Grenzfall vieler Messungen ist die Maximum Likelihood Schätzfunktion erwartungstreu
- Im Grenzfall vieler Messungen wird die Likelihood eine Normalverteilung
- In Grenzfall vieler Messungen erfüllt die ML Gleichheit bei der Cramér–Rao Ungleichung, ist also asymptotisch der "bestmögliche Schätzer"
- Die Unsicherheit lässt sich dann aus der zweiten Ableitung bestimmen $\sigma^2_{\mathrm{ML}} \approx \left(\frac{\mathrm{d}^2 \log \mathcal{L}(\theta)}{\mathrm{d}\theta^2} \right)^{-1}$

$\rightarrow$ Für eine große Menge von Anwendungen ist der ML Schätzer einfach zu bestimmen und sehr gut

Maximum likelihood fits

- Beispiel: Parameter einer Funktion (z.B. $\alpha_s(M_Z)$) soll aus mehreren Messungen bestimmt werden
- Nichts Neues hier: $p(n|\lambda)$ wird zu $p(n|\lambda(\alpha))$, mit parameter α
- Bei mehreren Messungen: Produkt der Wahrschienlichkeiten – daher auch Produkt der Likelihoods
- Beispiel: Erwarteter Wert bei x ist $f(x|\alpha)$, Messungen bei $x_1, x_2 \ldots$ ergeben Werte $\phi_1, \phi_2 \ldots$ mit Unsicherheiten $\sigma_1, \sigma_2 \ldots$ mit Gauß'schen Fluktuationen.
- Likelihood ist:

$$L(\alpha) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(\phi_i - f(x_i|\alpha))^2}{2\sigma_i^2}}$$



Quelle: Particle data group

Maximum likelihood fits (2)

- Maximiere:

$$L(\alpha) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(\phi_i - f(x_i|\alpha))^2}{2\sigma_i^2}}$$

- Für Maximierung können die Wurzeln ignoriert werden, da sie nicht von α abhängen:

$$L'(\alpha) = \prod_i e^{-\frac{(\phi_i - f(x_i|\alpha))^2}{2\sigma_i^2}}$$

- Maximierung der Likelihood ist Maximierung des logs:

$$-2 \cdot \log L(\alpha) = \sum_i \frac{(\phi_i - f(x_i|\alpha))^2}{\sigma_i^2}$$

- Dies ist die χ^2 Funktion aus Vorlesung 2, ML ist äquivalent zu Minimierung von χ^2

Maximum likelihood fits (3)

- Was, wenn die Fluktuationen nicht aus einer Gaußverteilung stammen?
- Beispiel: Poisson

$$L(\alpha) = \prod_i \frac{f(x_i|\alpha)^{\phi_i} e^{-f(x_i|\alpha)}}{\phi_i!}$$

- Log, ignoriere konstante Terme

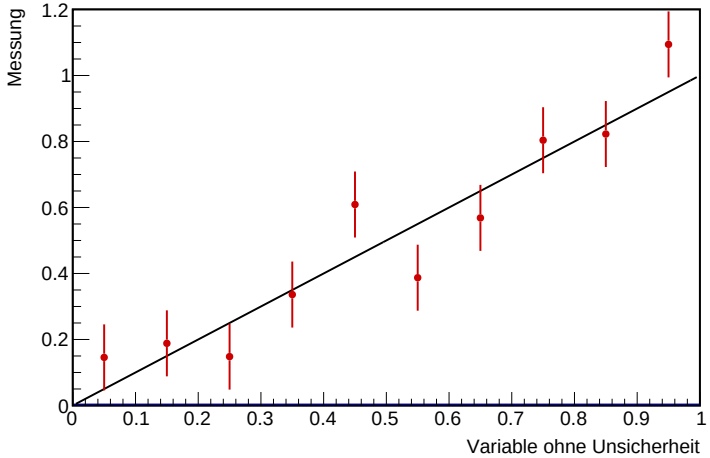
$$\log L(\alpha) = \sum_i \phi_i \cdot f(x_i|\alpha) - f(x_i|\alpha)$$

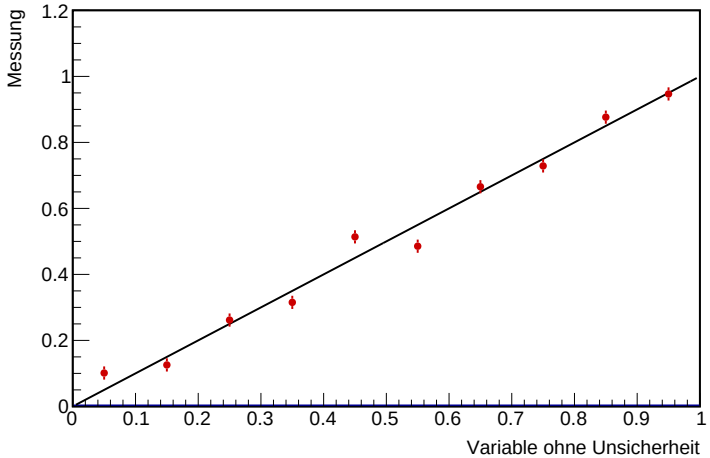
- Für mehrere Variablen, ersetze $f(x_i|\alpha) \to f(x_i|\alpha, \beta \ldots)$
- Maximierung üblicherweise numerisch

Anpassungsgüte (goodness of fit)

- Schlussfolgerung bezüglich Daten und Modell?
- Wie mit kleineren Unsicherheiten?
- Objektives Kriterium?
- "Wie weit weg vom Modell im Verhältnis zur Unsicherheit?"
- χ^2 so eine Größe:

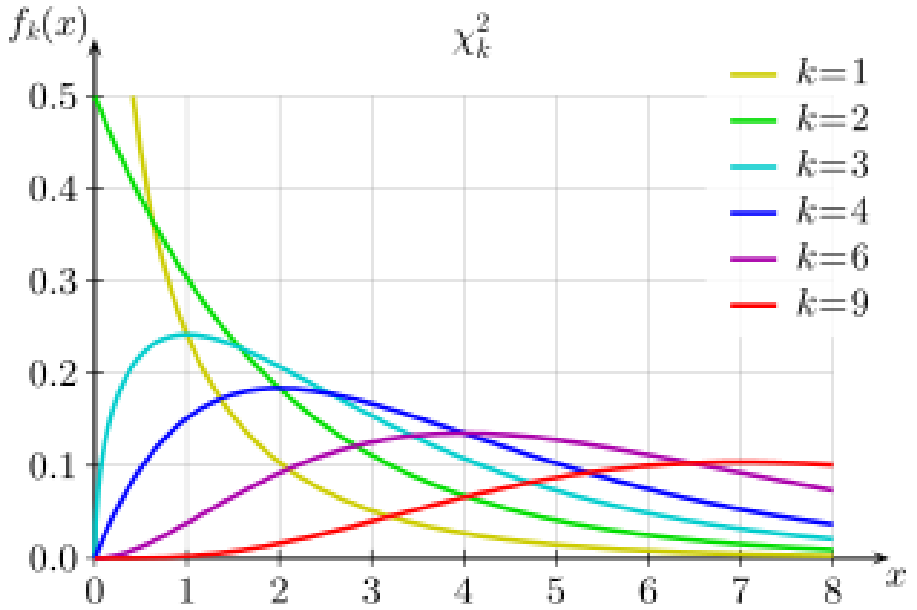
$$\chi^2 = \sum_i \left(\frac{(\phi_i - f(x_i)}{\sigma_i} \right)^2$$





Die χ^2 Funktion

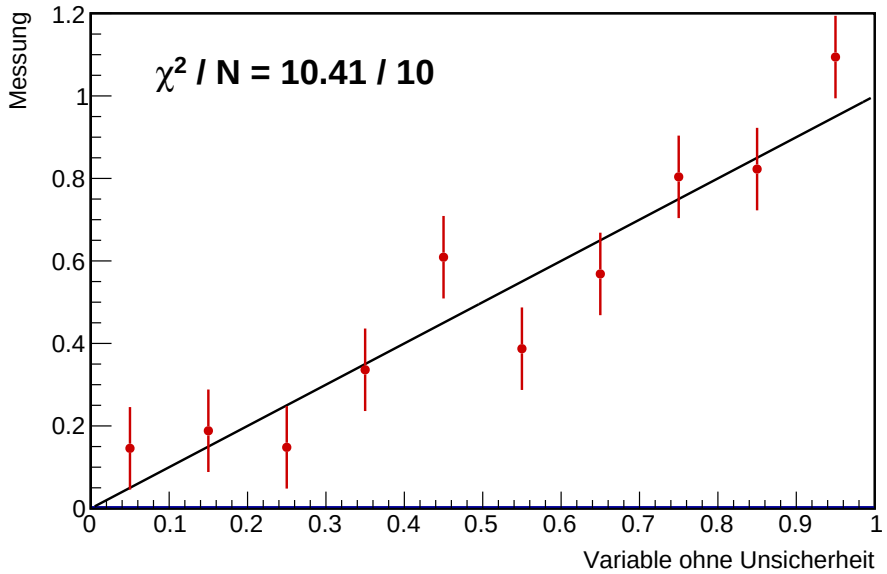
- χ^2-Funktion: Verteilung der Summe der Quadrate mehrerer (k) normalverteilter Zufallsvariablen
- "Das, was sich für χ^2 ergibt, falls das Modell richtig ist"
- Mittelwert ist k
- Bei k Messungen, Vergleiche χ^2 des Modells mit k

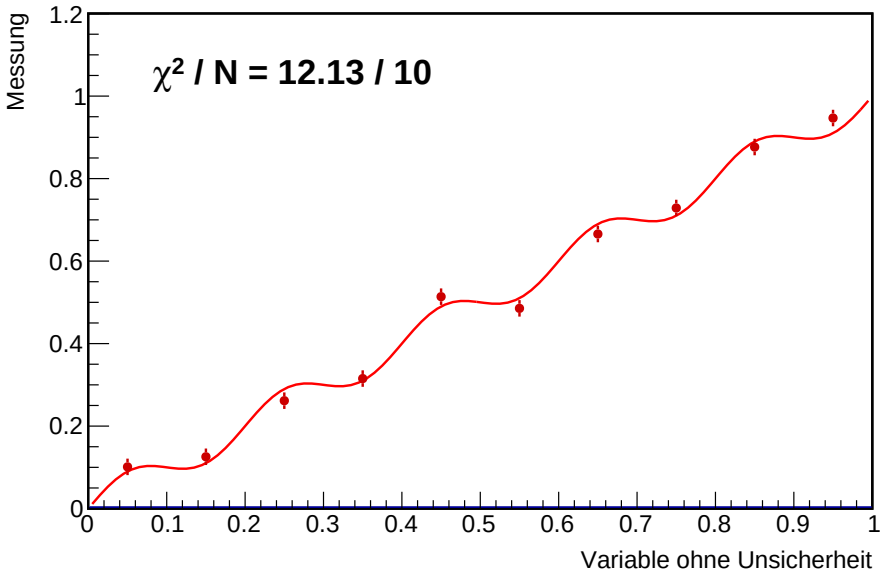


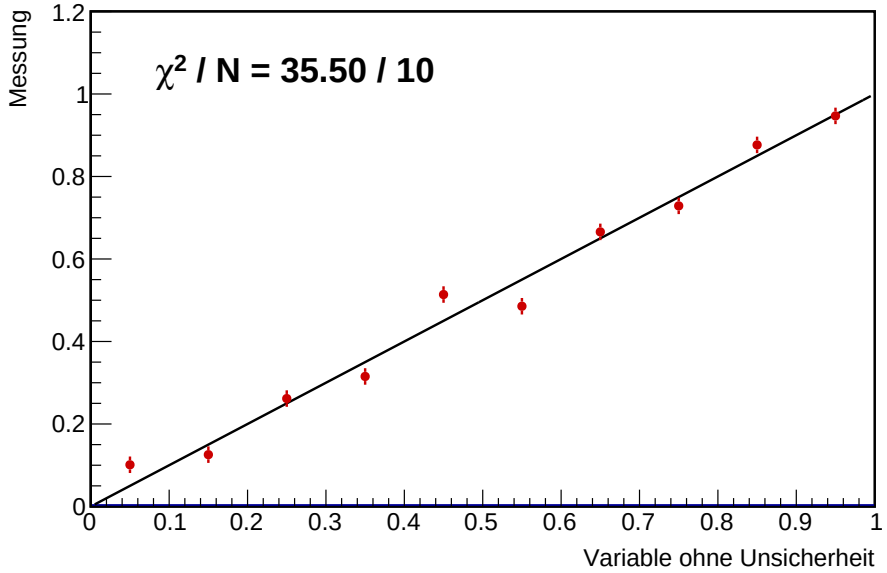
Quelle: Wikipedia

Anpassungsgüte (2)

- Mit großen Unsicherheiten χ^2 nahe 1
- Eine bessere Messung zeigt aber Diskrepanzen zum Modell
- χ^2 quantifiziert diese Abweichung

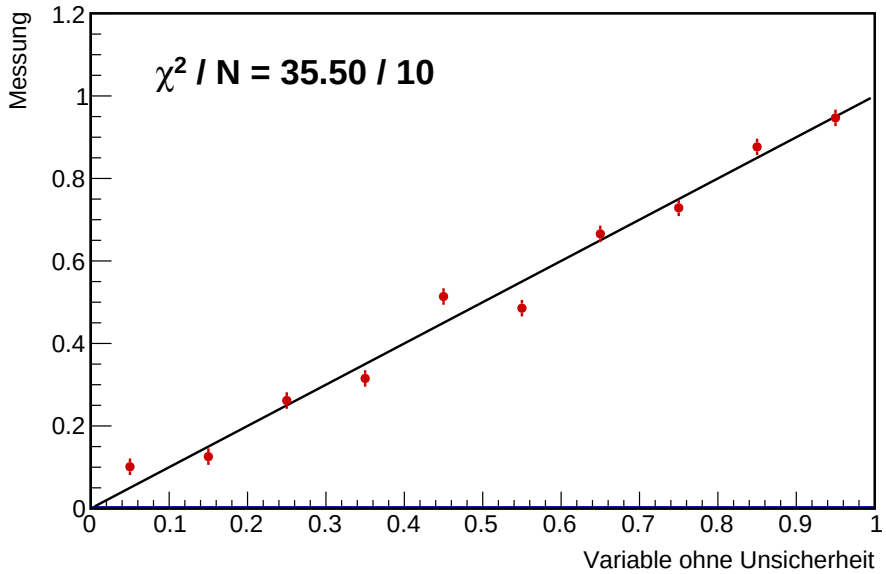






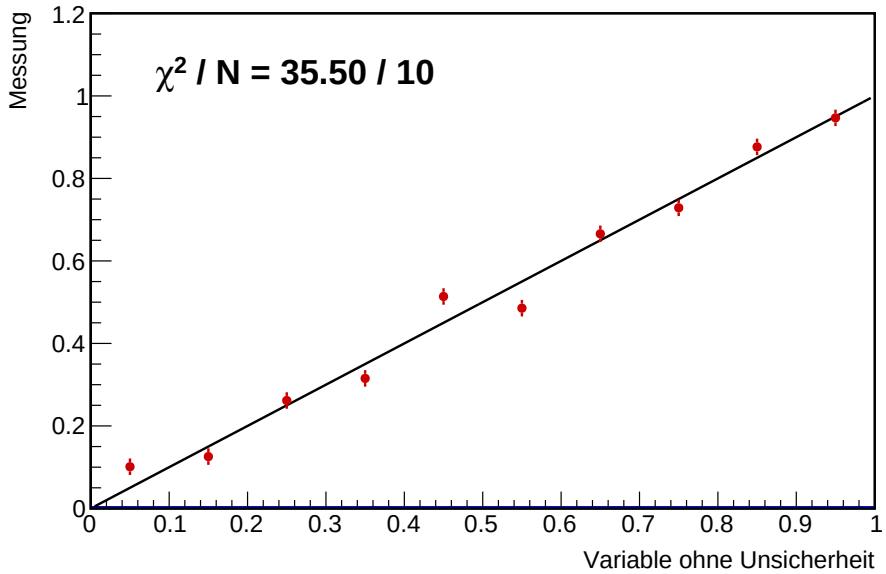
Hypothesentests und p–Wert

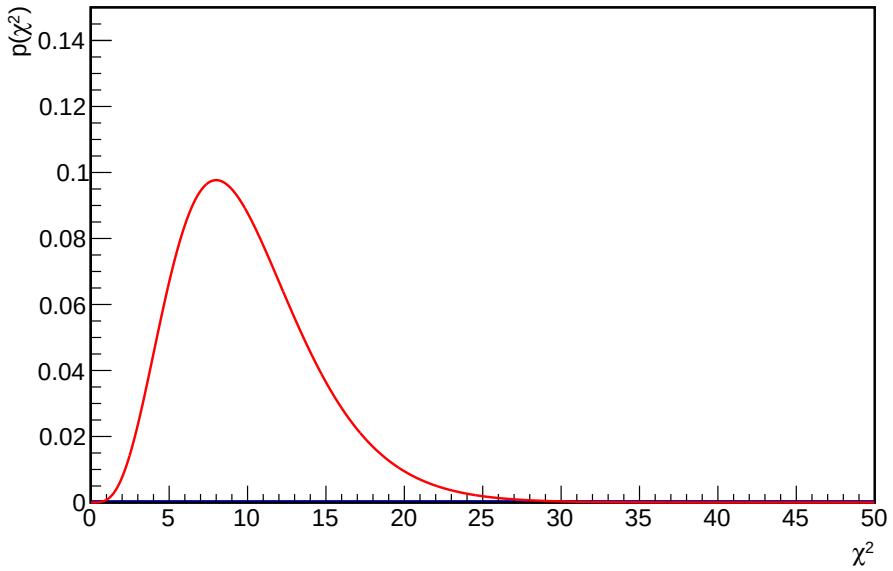
- Hier Frage nicht: "Wie gut ist der Fit?", sondern "Ist die Fitfunktion richtig?"
- Frequentistisch: Es git keine Wahrscheinlichkeit für ein Modell!
- Also kann die Frage "Wie sicher sind wir uns bezüglich des Modells?" nicht beantwortet werden
- Stattdessen Frage umkehren: Wie außergewöhnlich wären die Daten, wenn das Modell wahr wäre?



Hypothesentests und p–Wert (2)

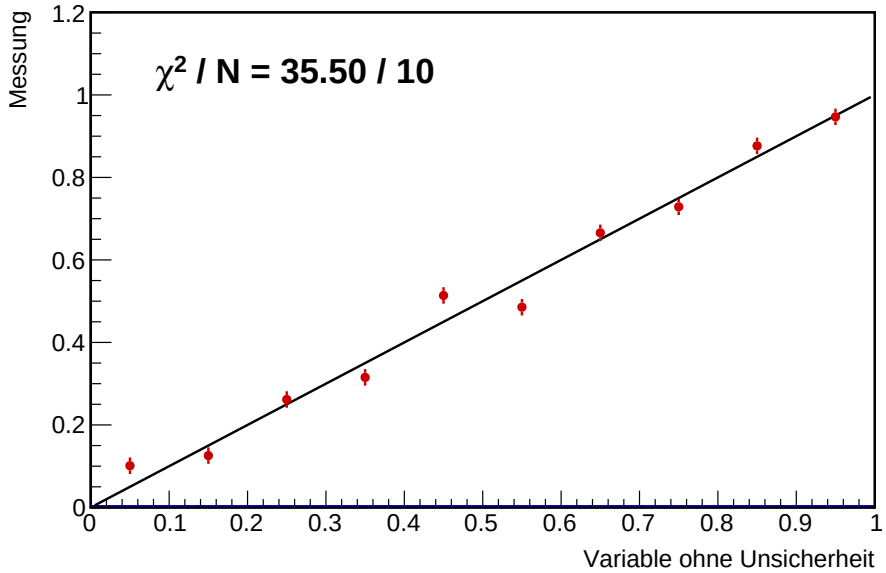
- Vorgehen
  - Wähle ein Maß für die Abweichung von einer Hypothese (H_0) hin zu einer anderen
  - Zum Beispiel: Mittelwert, χ^2, ...
  - Berechne die Wahrscheinlichkeit, ein "mindestens so extremes" Ergebnis zu Erhalten, falls die Hypothese wahr ist
- Die resultierende Wahrscheinlichkeit heißt p–Wert
- p–Wert für lineares Modell: 0.000102

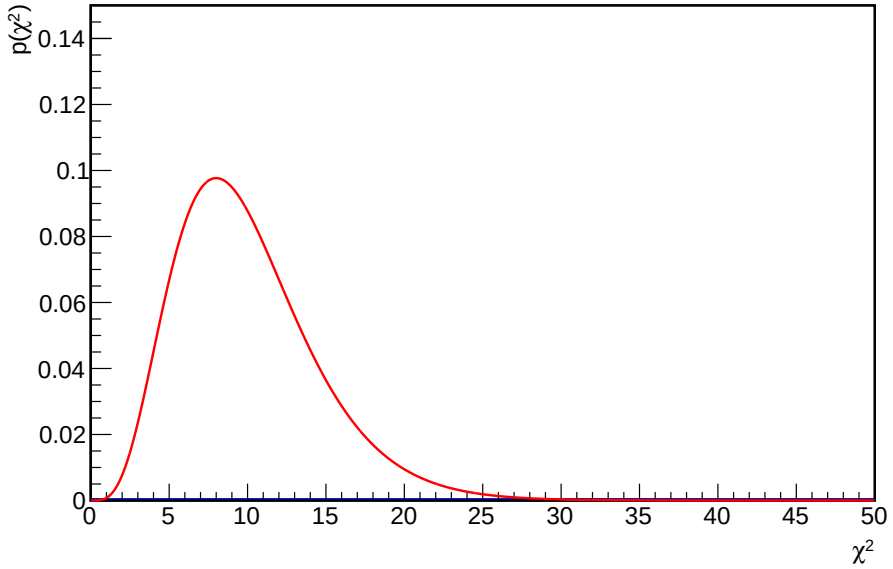




Hypothesentests und p–Wert (3)

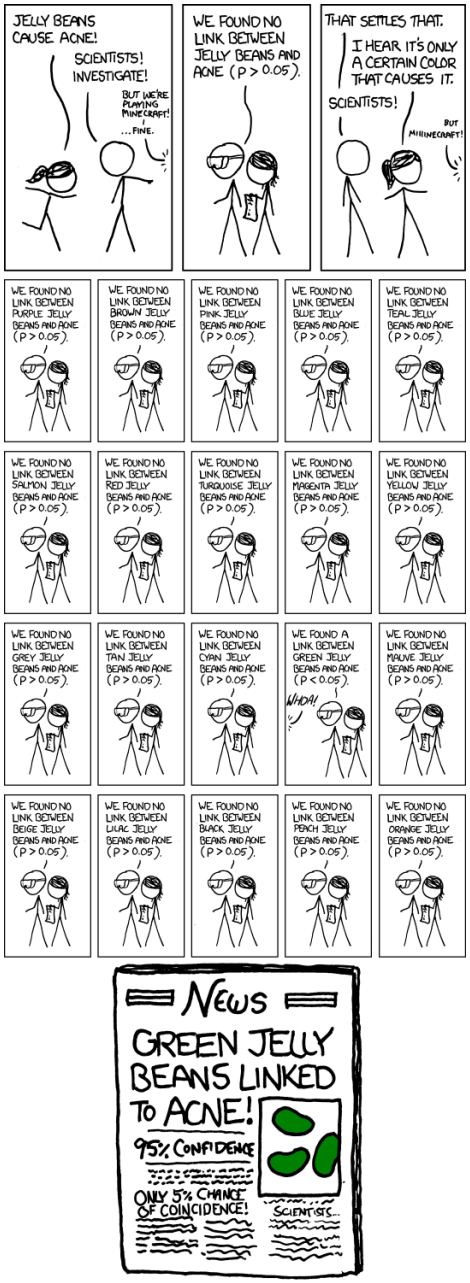
- p–Wert differenziert 2 Modelle (H_0,H_1), aber nur eines (H_0) taucht auf
- In Biologie, Sozialwissenschaften, Grenze oft bei 0.05
- $p < 0.05 \rightarrow$ verwirf Nullhypothese
- p–Wert < 0.05 heißt nicht:
  - ... dass H_0 mit 95% Wahrscheinlichkeit falsch ist
  - ... dass H_1 mit 95% Wahrscheinlichkeit wahr ist
  - ... dass die Wahrscheinlichkeit der Messung, gegeben H_0, 0.05 ist





Probleme mit dem p–Wert

- $p < 0.05$ taucht in einem von 20 Fällen zufällig auf, wenn H_0 wahr ist
- Wird das Experiment oft genug durchgeführt, ergibt sich das irgendwann



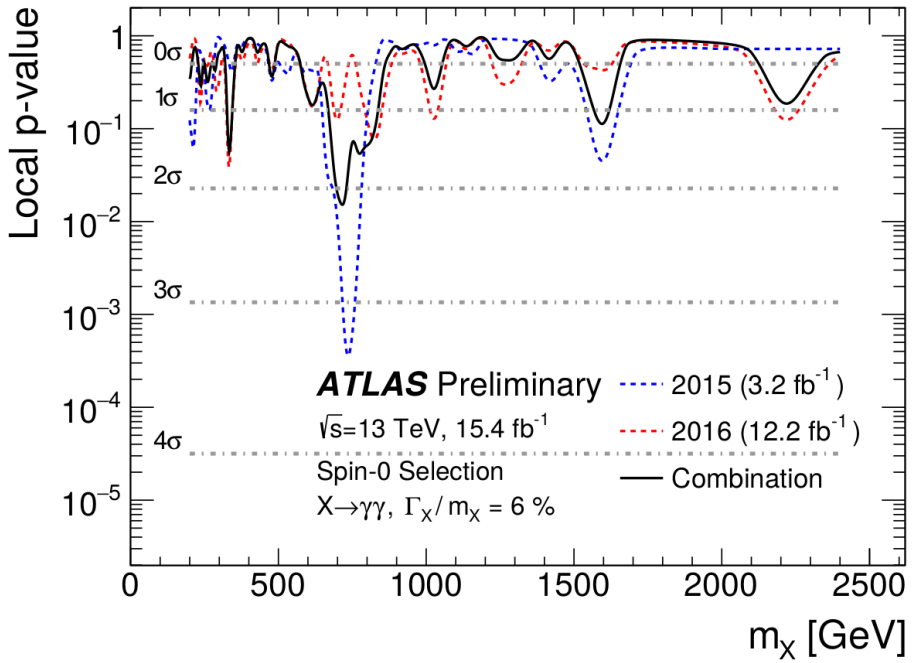
Quelle: xkcd

Beispiel: Münzwurf

- Münzwurf (H/T): HTHTHTHTHT
- Ist die Münze fair?
- Möglichkeit: vergleiche Differenz von H und T
- $\rightarrow$ p–Wert is 1 (genau gleichviele)
- Möglichkeit 2: Untersuche, wie oft das Ergebnis zum Vorherigen wechselt
- $\rightarrow$ 9 Wechsel bei 4.5 erwarteten: p–Wert ist 0.0039
- Ergebnis hängt von der Teststatistik ab!

ATLAS Diphoton excess

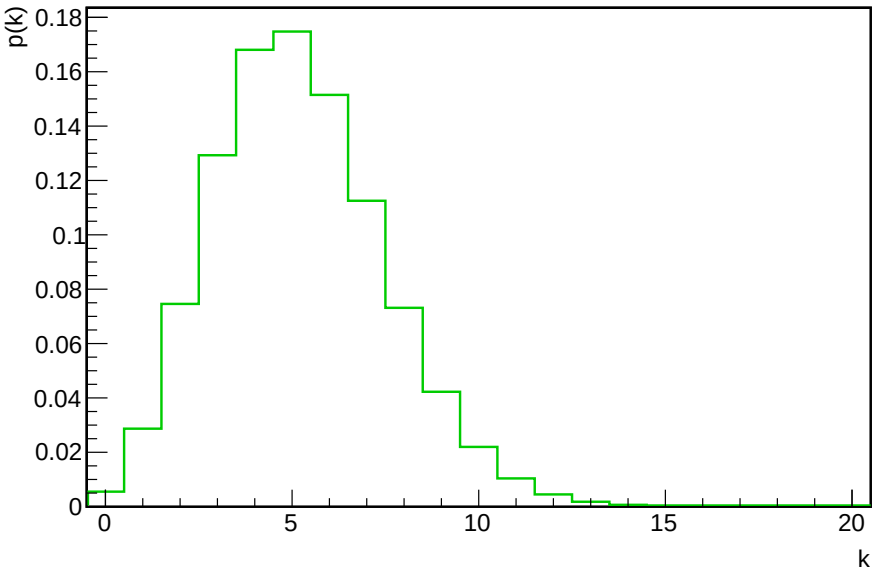
- In ATLAS Messung wurde ein "Teilchen" mit $p < 0.001$ gefunden
- Mit mehr Daten verschwand der Effekt
- p–Werte mit Vorsicht genießen!



Quelle: ATLAS Kollaboration

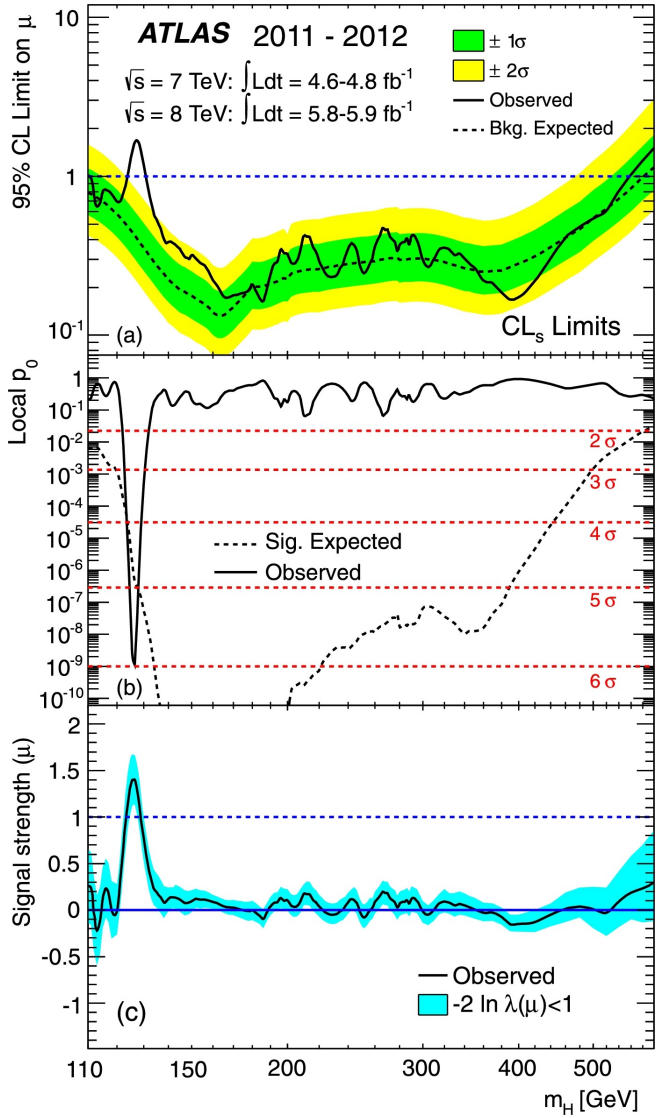
Konfidenzintervalle

- Erinnerung: Poisson, Messung von n $\rightarrow \mu = n \pm \sqrt{n}$
- Klappt nur für große n
- Was bei $n = 0$ order $n = 1$?
- $n = 0$ sollte immer noch $\mu = 1000$ ausschließen $\rightarrow$ kann man das quantifizieren?
- Antwort: Nein! Denn es gibt keine Wahrscheinlichkeit für $\mu = 1000$ (frequentistisch)
- Also wieder "tricksen" wie bei den p–Werten



Konfidenzintervalle (2)

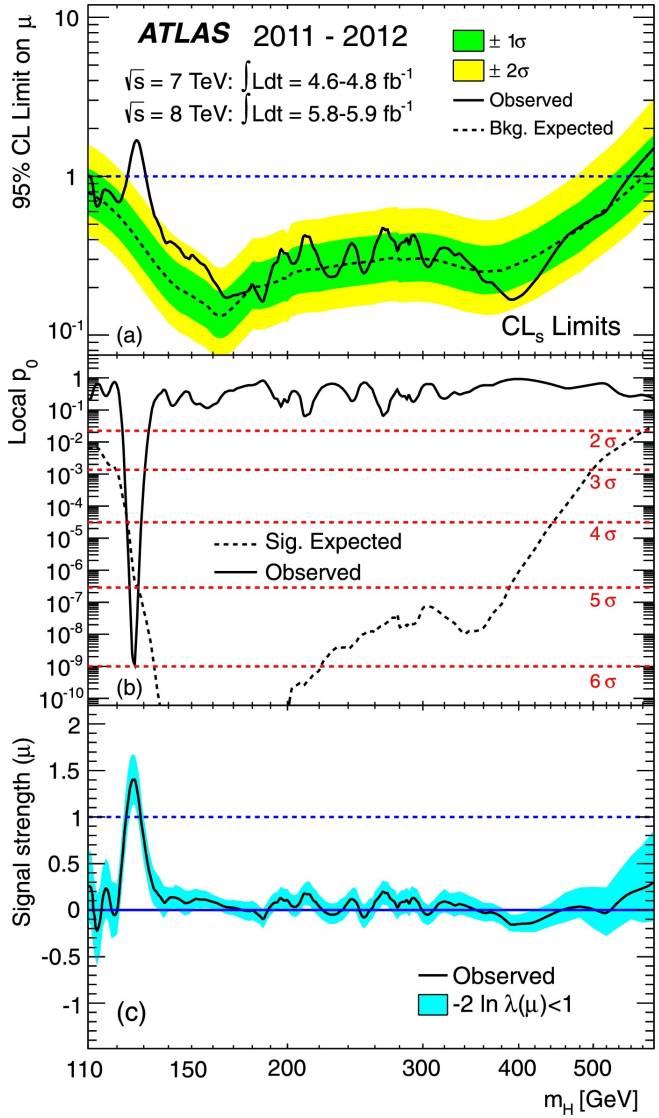
- Parameterbestimmung: Parameter μ, Daten $\vec{d}$
- Vorgehen:
  - Mache eine Regel, wie man gemessenen Daten ein Intervall zuordnet
  - Regel sollte so funktionieren, dass für jedes beliebige wahre μ, μ in 95% der Messungen im Intervall liegt
- Resultat genannt: "95% Konfidenzintervall"
- 95% enspricht etwa 2σ bei Normalverteilung



Quelle: ATLAS Kollaboration

Konfidenzintervalle (3)

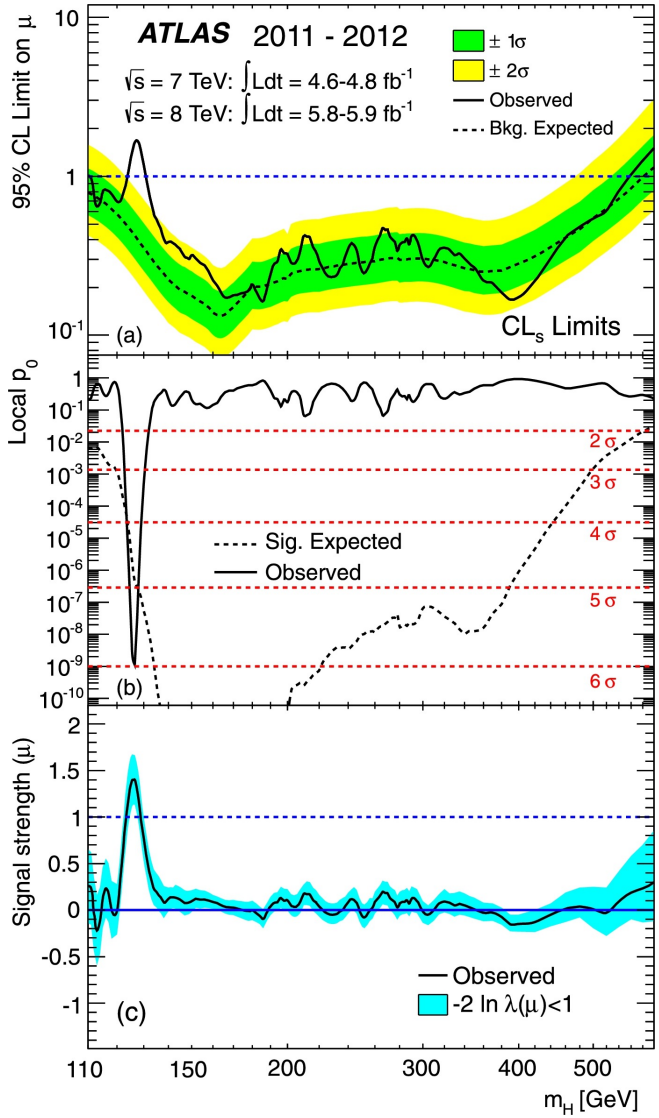
- 95% Konfidenzintervall heißt nicht ...
  - ..., dass das gemessene Intervall den Wert mit 95% Wahrscheinlichkeit enthält
  - ..., dass bei Wiederholung der Messung 95% der Messwerte im Intervall liegen
- Für Normalverteilung mit festem σ einfach zu Berechnen
- Für viele andere Verteilungen existieren Tabellen



Quelle: ATLAS Kollaboration

Konfidenzintervalle (3)

- 95% Konfidenzintervall heißt ...
  - ..., dass das gemessene Intervall den Wert mit 95% Wahrscheinlichkeit enthält
  - ..., dass bei Wiederholung der Messung 95% der Messwerte im Intervall liegen
- Für Normalverteilung mit festem σ einfach zu Berechnen
- Für viele andere Verteilungen existieren Tabellen



Quelle: ATLAS Kollaboration

Konfidenzintervalle (4)

- Poisson Verteilung für $n = 0$
- Tabelle zeigt 95% Konfidenzintervall von $0 - 3$

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	μ_{lo}	μ_{up}	μ_{lo}	μ_{up}
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51
6	3.15	10.53	2.61	11.84
7	3.89	11.77	3.29	13.15
8	4.66	12.99	3.98	14.43
9	5.43	14.21	4.70	15.71
10	6.22	15.41	5.43	16.96

Quelle: Particle Data Group

Zusammenfassung

- **Schätzfunktionen** erzeugen Schätzwert aus Daten
  $\rightarrow$ Sie müssen bezüglich ihrer Eigenschaften analysiert werden
- **Maximum Likelihood** ist bestmöglicher Schätzer im Limit großer Datenmengen
- χ^2/N gibt die Abweichung der Daten zu einem Modell bezogen auf ihre Unsicherheit an
- **p–Wert** für den Test von Hypothesen $\rightarrow$ Bezeichnet Wahrscheinlichkeit, bei Wiederholung des Experiments ein bezüglich der gewählten Teststatistik mindestens so extremes Ergebnis zu erhalten, falls die Hypothese wahr ist
- **Konfidenzintervalle** sind eine Vorschrift, aus Daten so Intervalle zu konstruieren, dass bei Wiederholung des Experimentes das Intervall mit einer gegebenen Wahrscheinlichkeit den wahren Wert umschließt