Universität Zürich[UZH]

Datenanalyse
(PHY231)
Herbstsemester 2017

Olaf Steinkamp

36-J-05 olafs@physik.uzh.ch 044 63 55763

Vorlesungsprogramm

- **Einführung, Messunsicherheiten, Darstellung von Messdaten**
- **Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik**
  - Mittelwert, Standardabweichung, Kovarianz und Korrelation
- **Fehlerfortpflanzungsgesetz**
- **Wahrscheinlichkeitsverteilungen**
  - diskrete Verteilungen, kontinuierliche Verteilungen
  - zentraler Grenzwertsatz
- **Monte-Carlo Methode**
- **Wahrscheinlichkeitsverteilungen II**
  - Faltung zweier Verteilungen
  - Verteilungen zweier Variablen
- **Stichproben und Schätzfunktionen**
  - Maximum-Likelihood Methode
  - Methode der kleinsten Quadrate
- **Interpretation von Messergebnissen**
  - Konfidenzintervalle, Testen von Hypothesen

Beispielprogramme im Verzeichnis

`/disk/puma/da/vorl/ci`

Motivation

Grundsätzlich zwei Gründe ein Experiment durchzuführen

- **will eine bestehende Theorie widerlegen**
  - suche nach Effekten, die mit der Theorie nicht vereinbar sind
  - z.B. Protonenzerfall, neue Elementarteilchen
- **Testen von Hypothesen → nächste Woche**

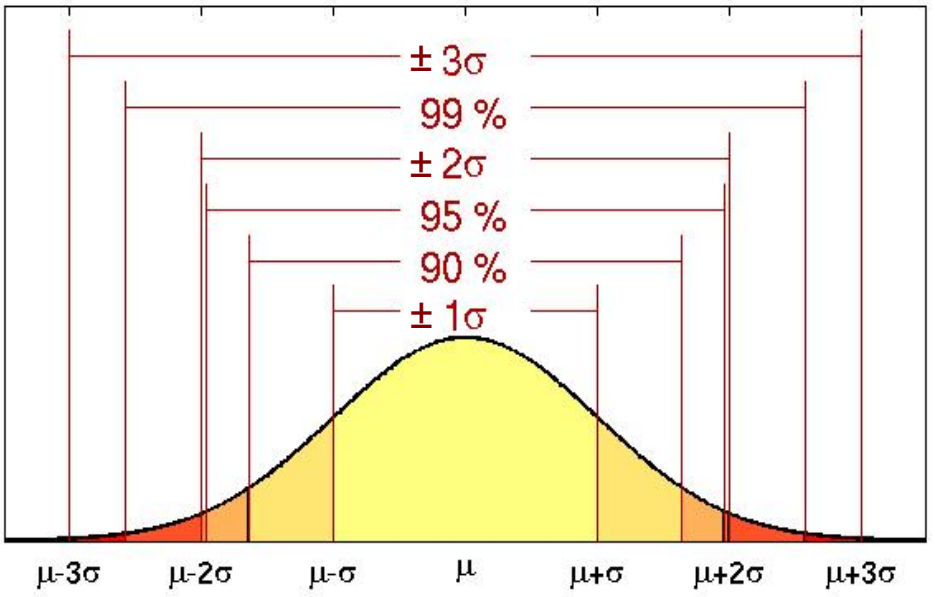
- **will Parameter einer bestehenden Theorie (**genauer**) bestimmen**
  - z.B. mittlere Lebensdauer des Neutrons, Ruhemasse von Neutrinos
- **Konfidenzintervall:**
  - wähle “Konfidenzniveau” (“confidence level”) *CL*, z.B. 68%, 90%, 95%
  - leite aus der Messung ein Intervall ab, das den (wahren) Wert des Parameters mit der Wahrscheinlichkeit *CL* enthält

Konfidenzintervall für Zufallsvariablen

Betrachte Verteilung einer Zufallsvariable *x*

- **Konfidenzintervall $[x_-,x_+]$ zum Konfidenzniveau *CL*:**
  - beliebiger Wert x_i aus der Verteilung liegt mit Wahrscheinlichkeit *CL* in $[x_-,x_+]$
- **Bestimmung von $[x_-,x_+]$:**
  - wähle $[x_-,x_+]$ so, dass es den Bruchteil *CL* aller x_i enthält
- **Beispiel: zentrale Konfidenzintervalle für die Gaußverteilung**

CL	$x_\pm$
90%	μ ± 1.645·σ
95%	μ ± 1.960·σ
99%	μ ± 2.576·σ
68.13%	μ ± 1·σ
95.45%	μ ± 2·σ
99.73%	μ ± 3·σ



Konfidenzintervall für Zufallsvariablen

Zweiseitige Konfidenzintervalle

- **symmetrisches Konfidenzintervall**

$$x_+ - \mu = \mu - x_-$$

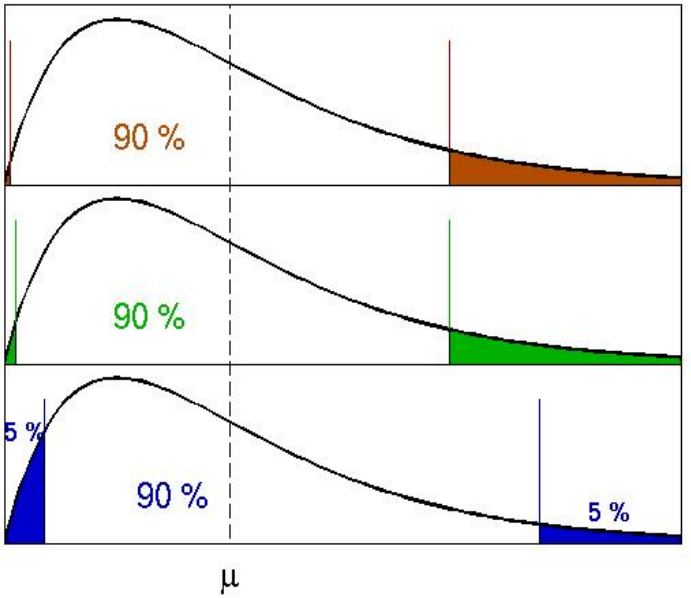
- **kürzestes Konfidenzintervall**

$x_+ - x_-$ **so klein wie moglich**

- **zentrales Konfidenzintervall**

$$\int_{-\infty}^{x_-} p(x)\,dx = \int_{x_+}^{+\infty} p(x)\,dx = (1-CL)/2$$

- **alle identisch für symmetrische Verteilungen**



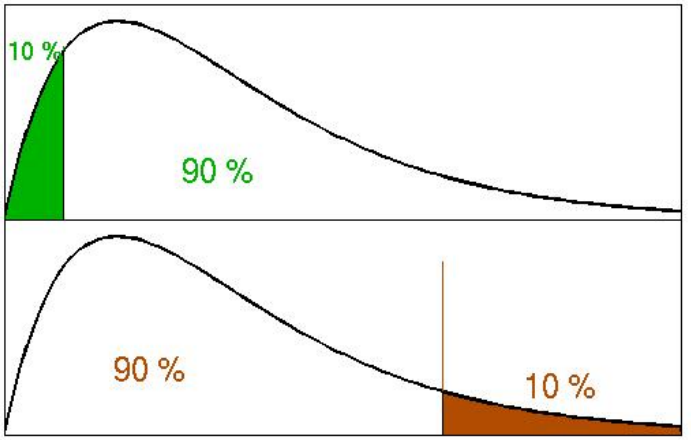
Einseitige Konfidenzintervalle

- **unteres Konfidenzlimit:**

$$\int_{x_-}^{+\infty} p(x)\,dx = CL$$

- **oberes Konfidenzlimit:**

$$\int_{-\infty}^{x_+} p(x)\,dx = CL$$

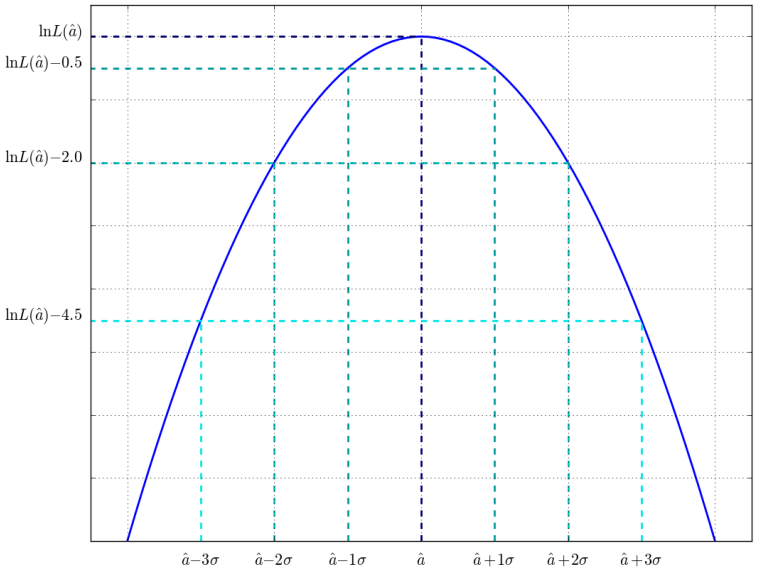


Konfidenzintervall für Schätzwerte

Maximum-likelihood Methode

- großer Stichprobenumfang: gaußförmige Wahrscheinlichkeitsverteilung für Schätzwert

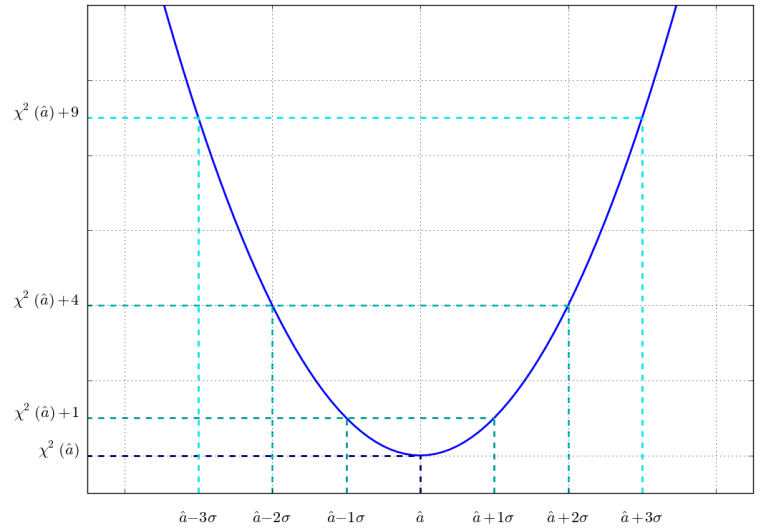
In L(a) – In L($\hat{a}$)	Konfidenzintervall
– 0.5	± 1 σ ⇔ 68.1 %
– 2	± 2 σ ⇔ 95.5 %
– 4.5	± 3 σ ⇔ 99.7 %



Methode kleinster Quadrate

- implizite Annahme gaußverteilter Unsicherheit

$\chi^2(a) - \chi^2(\hat{a})$	Konfidenzintervall
+1.0	± 1 σ ⇔ 68.1 %
+4.0	± 2 σ ⇔ 95.5 %
+9.0	± 3 σ ⇔ 99.7 %

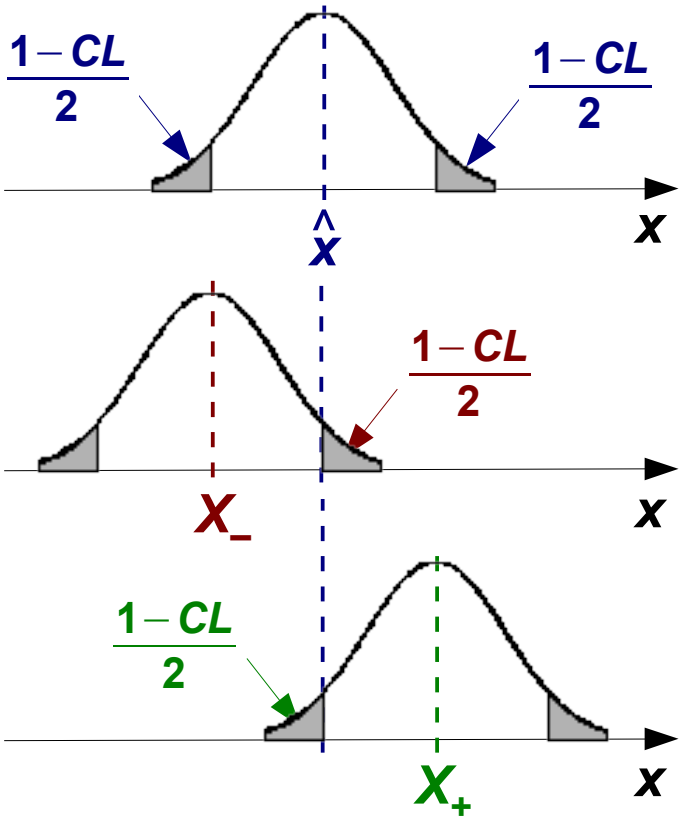


- für *n* Parameter: Konfidenzregionen in *n*-dimensionalen Parameterraum

Konfidenzintervall für den wahren Wert

Messung einer Größe X ergibt einen Schätzwert $\hat{x}$ mit Unsicherheit σ_x

- **will daraus eine Aussage über den wahren Wert X der Größe ableiten**
  - bestimme ein Konfidenzintervall $[X_-, X_+]$, das den wahren Wert X mit Wahrscheinlichkeit CL enthält
- **Ansatz zur Bestimmung von X_- und X_+ :**
  - bei wahrem Wert X_- ist die Wahrscheinlichkeit $(1-CL)/2$, ein Messergebnis $\geq \hat{x}$ zu erhalten
  - bei wahrem Wert X_+ ist die Wahrscheinlichkeit $(1-CL)/2$, ein Messergebnis $x \leq \hat{x}$ zu erhalten
- **für gaußverteilte Messunsicherheiten auf $\hat{x}$: Konfidenzintervall für X = Konfidenzintervall für $\hat{x}$**



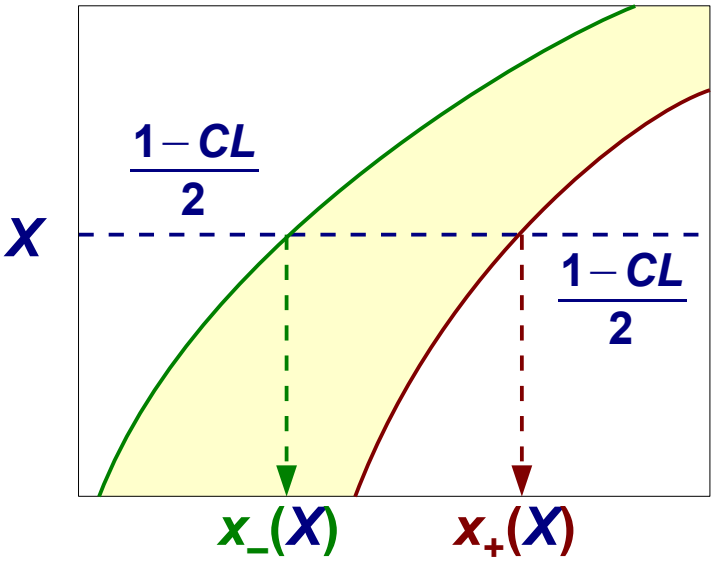
z.B. für X_- :

$$\frac{1-CL}{2} \equiv \frac{1}{\sqrt{2\pi\cdot\sigma^2}} \int_{\hat{x}}^{+\infty} e^{-\frac{(x-X_-)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi\cdot\sigma^2}} \int_{-\infty}^{X_-} e^{-\frac{(x-\hat{x})^2}{2\sigma^2}}\, dx$$

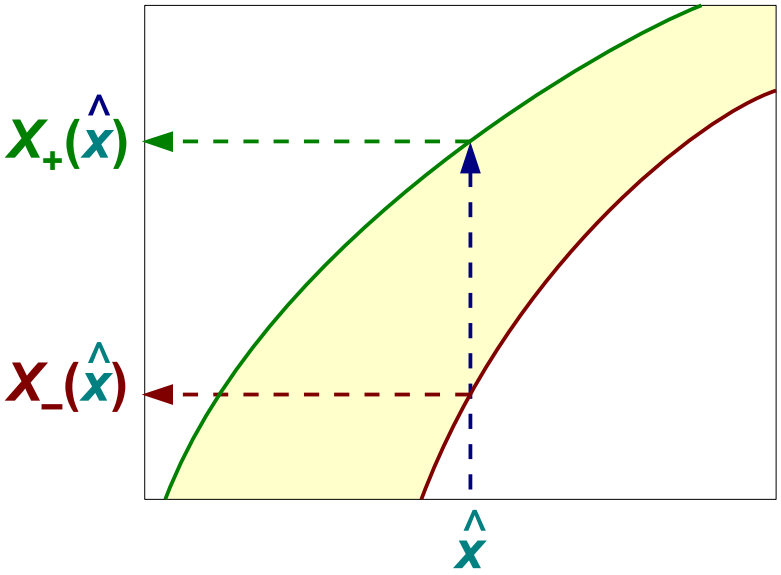
Konfidenzintervall für den wahren Wert

Allgemeiner Fall: Unsicherheiten auf dem Schätzwert nicht gaußverteilt

- **bestimme vor der Messung erwartete Konfidenzintervalle $[x_-,x_+]$ für $\hat{x}$ als Funktion des angenommenen wahren Werts X**
  - benutze zum Beispiel Monte-Carlo Simulation
  - ergibt "Konfidenzband" in der $(\hat{x},X)$-Ebene



- **lese nach der Messung X_- und X_+ beim gemessenen $\hat{x}$ aus dem Konfidenzband ab**



gaußverteilte Unsicherheit auf dem Schätzwert: Konfidenzband begrenzt durch zwei Geraden mit Steigung eins

Konfidenzintervalle an Grenzen des physikalisch erlaubten Bereichs

Beispiel: Experiment zur Messung der Neutrinomasse m_ν

- **messe die Energie E und den Impuls p des Neutrinos, berechne $m^2 = E^2 - p^2$**
- **Messunsicherheit auf m^2 sei gaußverteilt mit $\sigma_{m^2} = 2\ \mathrm{eV}^2$**
- **will 2σ (95.45 %) Konfidenzintervall für den wahren Wert m_ν^2 angeben**
  - Standardrezept: $[\, m_\nu^2{}_- ;\ m_\nu^2{}_+ \,] = [\, m^2 - 2\,\sigma_{m^2} ;\ m^2 + 2\,\sigma_{m^2} \,]$

a) messe $m^2 = E^2 - p^2 = 4\ \mathrm{eV}^2$
→ Standardrezept ergibt Konfidenzintervall [0 eV², 8 eV²] } alles okay

b) messe $m^2 = E^2 - p^2 = -4\ \mathrm{eV}^2$
→ Standardrezept ergibt [-8 eV², 0 eV²] } wahrer Wert des Massenquadrats mit 95.45 % Wahrscheinlichkeit negativ ?

- **divergierende Meinungen darüber, was in Fällen wie b) oder c) zu tun ist**
- **ein möglicher Ansatz: benutze "Bayessche Statistik"**

Bayessches Theorem

Aussage über bedingte Wahrscheinlichkeiten (Thomas Bayes, 1763):

$$p(A \text{ und } B) = p(A) \cdot p(B \mid A) = p(B) \cdot p(A \mid B)$$

$$\Rightarrow \quad p(A \mid B) = \frac{P(B \mid A) \cdot p(A)}{p(B)}$$

Beispiel: Test für eine seltene Krankheit *K*

- im Mittel leiden 7 von 1.000.000 Menschen an der Krankheit: $p(K) = 7 \times 10^{-6}$
- der Test sei zu 99 % effizient: $p(+|K) = 0.99$
- er habe zufällige Ansprechwahrscheinlichkeit von 0.1%: $p(+) = 0.001$
- Wahrscheinlichkeit, dass eine zufällig ausgewählte Person tatsächlich an der Krankheit leidet, wenn der Test positiv angesprochen hat:

$$p(K \mid +) = \frac{p(+ \mid K) \cdot p(K)}{p(+)} = \frac{0.99 \cdot 7 \times 10^{-6}}{0.001} = 6.9 \times 10^{-3} = 0.69\ \%$$

Bayessches Theorem

Angewendet auf die Interpretation von Messergebnissen:

"a-posteriori" Vertrauen in die Theorie (nach der Messung)

Wahrscheinlichkeit, das Ergebnis zu erhalten, wenn die Theorie gilt

"a-priori" Vertrauen in die Theorie (vor der Messung)

$$p(\text{Theorie}\,|\,\text{Ergebnis}) = \frac{p(\text{Ergebnis}\,|\,\text{Theorie})}{p(\text{Ergebnis})} \cdot p(\text{Theorie})$$

Wahrscheinlichkeit, das Ergebnis zu erhalten, egal ob die Theorie gilt oder nicht

- **ein einziges mit der Theorie inkompatibles Messergebnis widerlegt die Theorie**

$$p(\text{Ergebnis}\,|\,\text{Theorie}) = 0 \;\Rightarrow\; p(\text{Theorie}\,|\,\text{Ergebnis}) = 0$$

- **ein von der Theorie vorhergesagtes Ergebnis verstärkt das Vertrauen in die Theorie, wenn dieses Ergebnis ansonsten unerwartet wäre**

$$p(\text{Theorie}\,|\,\text{Ergebnis}) \propto p(\text{Ergebnis}\,|\,\text{Theorie})\,/\,p(\text{Ergebnis})$$

- **Problem: *p* (Theorie) kann nicht objektiv definiert werden**

Bayessches Theorem

Für Messung eines Parameters

- **"Theorie": wahrer Wert *X* des Parameters**
- **"Ergebnis": gemessener Wert $\hat{x}$**

$$\Rightarrow \quad p(X \mid \hat{x}) = \frac{p(\hat{x} \mid X)}{p(\hat{x})} \cdot p(X)$$

- **a-priori Wahrscheinlichkeit *p*(*X*): häufig "vorurteilsfrei"**
  - nehme jeden erlaubten Wert von *X* als gleich wahrscheinlich an
  - an den Grenzen des physikalisch erlaubten Bereichs heisst dies

$$p(X) \equiv \begin{cases} \text{konstant für } X \text{ innerhalb des erlaubten Bereichs} \\ 0 \quad \text{für } X \text{ ausserhalb des erlaubten Bereichs} \end{cases}$$

- **aufgepasst: diese Wahl von *p*(X) ist willkürlich,**
  - beeinflusst aber $p(X \mid x)$ und damit das Konfidenzintervall für *X*
- **Beispiel Messung der Neutrinomasse**

  - wähle $p(m_\nu) = \text{konst für } m_\nu > 0$
  - wähle $p(m_\nu^2) = \text{konst für } m_\nu^2 > 0$

beide Ansätze sind gleich "richtig", resultieren aber in unterschiedlichen Konfidenzintervallen für m_ν

Konfidenzintervalle an Grenzen des physikalisch erlaubten Bereichs

Bayessches Konfidenzintervall für Neutrinomasse, m_ν

- $p(m_\nu)$ "vorurteilsfrei"

$$p(m_\nu) = \begin{cases} 1 & \text{für } m_\nu \geq 0 \\ 0 & \text{für } m_\nu < 0 \end{cases}$$

- $p(m \mid m_\nu)$: gaußverteilte Messunsicherheit

$$p(m \mid m_\nu) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{(m-m_\nu)^2}{2\sigma^2}}$$

- $p(m)$: integriere $p(m \mid m_\nu)$ über alle möglichen Werte von m_ν

$$p(m) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot \int_0^\infty e^{-\frac{(m-m_\nu)^2}{2\sigma^2}} \, dm_\nu$$

- $p(m_\nu \mid m)$: aus Bayes Theorem

$$p(m_\nu \mid m) = e^{-\frac{(m-m_\nu^2)}{2\sigma^2}} \Big/ \int_0^\infty e^{-\frac{(m-m_\nu)^2}{2\sigma^2}} \, dm_\nu$$

- Zahlenbeispiel: 90% oberes Konfidenzlimit für $m = -0.5\,\text{eV}$; $\sigma = 0.2\,\text{eV}$

$$0.9 = \int_0^{(m_\nu)_+} p(m_\nu \mid m) \, dm_\nu = \int_0^{(m_\nu)_+} e^{-\frac{(m-m_\nu)^2}{2\sigma^2}} \, dm_\nu \Big/ \int_0^\infty e^{-\frac{(m-m_\nu)^2}{2\sigma^2}} \, dm_\nu \Rightarrow (m_\nu)_+ = 0.146\,\text{eV}$$

Konfidenzintervalle an Grenzen des physikalisch erlaubten Bereichs

Bayessches Konfidenzintervall für Quadrat der Neutrinomasse, m_ν^2

- $p(m_\nu^2)$ "vorurteilsfrei"

$$p\left(m_\nu^2\right) = \begin{cases} 1 & \text{für } m_\nu^2 \geq 0 \\ 0 & \text{für } m_\nu^2 < 0 \end{cases}$$

- $p(m^2 \mid m_\nu^2)$: gaußverteilte Messunsicherheit

$$p\left(m^2 \mid m_\nu^2\right) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}}$$

- $p(m^2)$: integriere $p(m^2 \mid m_\nu^2)$ über alle möglichen Werte von m_ν^2

$$p\left(m^2\right) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot \int_0^\infty e^{-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}} \, dm_\nu^2$$

- $p(m_\nu^2 \mid m^2)$: aus Bayes Theorem

$$p\left(m_\nu^2 \mid m^2\right) = e^{-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}} \Big/ \int_0^\infty e^{-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}} \, dm_\nu^2$$

$$\sigma\left(m^2\right) = 2|m| \cdot \sigma(m)$$

- Zahlenbeispiel: 90% oberes Konfidenzlimit für $m^2 = -0.25\,\text{eV}^2$; $\sigma = 0.2\,\text{eV}^2$

$$0.9 = \int_0^{(m_\nu^2)_+} p\left(m_\nu^2 \mid m^2\right) dm_\nu^2 = \int_0^{(m_\nu^2)_+} e^{-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}} \, dm_\nu^2 \Big/ \int_0^\infty e^{-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}} \, dm_\nu^2 \Rightarrow \left(m_\nu^2\right)_+ = 0.211\,\text{eV}^2$$

$$\neq (0.146\ \text{eV})^2$$

Zusammenfassung

Konfidenzintervall zu einem Konfidenzniveau *CL*

- **Intervall, das einen Wert mit Wahrscheinlichkeit *CL* enthält**
  - meist: Konfidenzintervall für den wahren Wert eines Parameters
- **gebräuchlichste Konfidenzniveaus: 90 % , 95 % , 99 %**
  - für Gaußverteilung: “1 σ” ($\approx$ 68 %) , “2 σ“ ($\approx$ 95.5 %) , “3 σ“ ($\approx$ 99.7 %)
- **Unsicherheiten gaußverteilt :**
  - Konfidenzintervall für wahren Wert = Konfidenzintervall für Schätzwert
- **Unsicherheiten nicht gaußverteilt :**
  - bestimme vor der Messung Konfidenzband für Schätzwert als Funktion des (angenommenen) wahren Wertes → Monte-Carlo Simulation
  - lese nach der Messung an der Stelle des beobachteten Schätzwerts das Konfizenzintervall für den wahren Wert aus dem Konfidenzband ab
- **Konfidenzintervalle an Grenzen des physikalisch erlaubten Bereichs**
  - verschiedene Ansätze, keiner wirklich ideal → “Religionskriege”