Einführung in die Statistik und Datenanalyse (2)

Martin Völkl
martin.voelkl@uni-tuebingen.de

Universität Tübingen
2017-03-21

Aufbau der Vorlesung

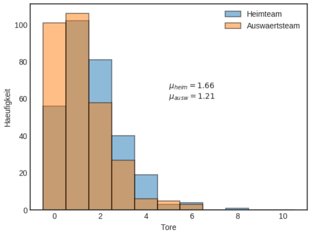
1. Einführung und systematische Unsicherheiten
2. **Werkzeugkiste**
    - Wichtige Wahrscheinlichkeitsverteilungen
    - Fehlerfortpflanzung
3. Wahrscheinlichkeit und Unsicherheit
4. Frequentistische Methoden
5. Bayesianische Methoden

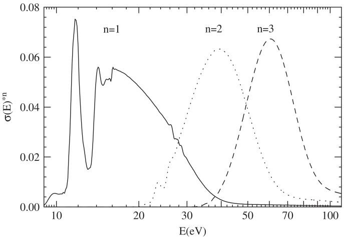
Wahrscheinlichkeitsverteilungen

- "Statistische Unsicherheit" eng verbunden mit "Zufall" und "Wahrscheinlichkeit"
- Verschiedene statistische Prozesse sind mit unterschiedlichen Wahrscheinlichkeitsverteilungen assoziiert
- Wahrscheinlichkeitsverteilung $p(E)$ gibt eine Zahl zwischen 0 und 1 für jedes mögliche Ergebnis zurück
- Summe der Wahrscheinlichkeiten aller (disjunkten) Ergebnisse ist 1

Diskrete und kontinuierliche Verteilungen

- Wahrscheinlichkeit für Ereignisse, z.B. $p(\mathrm{Kopf})$, $p(\mathrm{Zahl}) = 1 - p(\mathrm{Kopf})$
- Auch mehr als 2: Wahrscheinlichkeit für n Tore für Heimmannschaft in Fußballspiel
- Auch kontinuierlicher Fall: Wahrscheinlichkeit für E Energieverlust eines Teilchens bei Kollision $p(E)$
- Exakter: $p(E) \cdot \mathrm{d}E$: Wahrscheinlichkeit, dass E im Intervall $[E, E + \mathrm{d}E]$ liegt
- Diskret: $\sum_i p(h_i) = 1$
- Kontinuierlich: $\int p(x)\mathrm{d}x = 1$



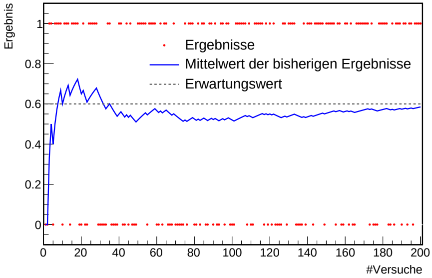


Binomialverteilung

- Basiert auf Prozess mit zwei möglichen Ergebnissen (Kopf/Zahl, 0/1, Erfolg/Misserfolg etc.)
- Für einen Versuch: $p(1) = \theta$, $p(0) = 1 - \theta$
- Gesucht ist die Wahrscheinlichkeit für eine bestimmte Anzahl von Erfolgen bei N Versuchen
- Fragestellung in der Physik typischerweise rückwärts: Gegeben n Erfolge in N Versuchen, was ist der Parameter θ?

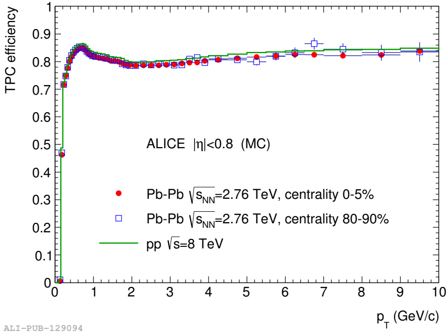
Exkurs: Gesetz der großen Zahlen

- $p(1) = \theta$, $p(0) = 1 - \theta \implies$ Erwartungswert ist θ
- Gesetz der Großen Zahlen: Mit zunehmender Anzahl von Versuchen nähert sich der Mittelwert der Ergebnisse dem Erwartungswert
- Keine Konvergenz im üblichen Sinne
- Annäherung nicht monoton



Binomialverteilung – Beispiele

- Münzwurf
- Zerfall eines Teilchens in einem bestimmten Zerfallskanal
- Detektoreffizienz (Bild)
- Umfragen

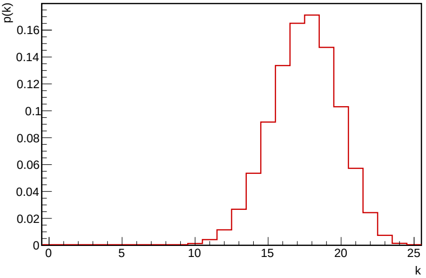


Performance of the ALICE Experiment at the CERN LHC,
ALICE Kollaboration

Binomialverteilung – Eigenschaften

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

- Erwartungswert $\mu = N\theta$
- Varianz: $\sigma^2 = N\theta(1-\theta)$
- Würde die Messung mehrfach wiederholt, fluktuierte die gemessene Anzahl um μ mit einer typischen Abweichung von σ
- Wichtig ist aber die andere Richtung der Schlussfolgerung



Binomialverteilung, $\theta = 0.7$, $N\theta = 17.5$

Binomialverteilung – Unsicherheit

- In welchem Bereich ist das Ergebnis der Parteien zu erwarten?
- Für große N und k: k/N in der Nähe von θ
- Die Schwankungen zwischen Messungen daher abgeschätzt durch

$$\sigma_k^2 = N\theta(1-\theta) \approx N(k/N)(1-k/N)$$

$$\sigma_{(k/N)} \approx \sqrt{\frac{1}{N}}\sqrt{\frac{k}{N}\left(1-\frac{k}{N}\right)}$$

- d.h. die Union könnte mit einem Ergebnis von $(39.5 \pm 1.5)\%$ rechnen
- Die Abhängigkeit $1/\sqrt{N}$ der (relativen) Unsicherheit taucht häufig auf
- Näherung wovon?

Umfrage (Beispiel):

Union: 395
SPD: 238
Linke: 93
Grüne: 82
FDP: 78
AfD: 76
Sonstige: 38

Binomialverteilung – Beispiel: Geburten

Land	Geburten Jungen	Geburten Mädchen	Anteil
Welt	–	–	51.6%
Deutschland	357192	337612	51.4%
China	8900979	7757874	54%
Finland	29627	28017	51.4%
Indien	12849497	11582656	52.6%
Argentinien	383852	370371	50.9%

Daten für 2016, Quelle: Weltbank

Binomialverteilung – Beispiel: Geburten(2)

Land	Geburten Jungen	Geburten Mädchen	Anteil
Welt	–	–	51.6%
Deutschland	357192	337612	$(51.4 \pm 0.06)\%$
China	8900979	7757874	$(54 \pm 0.01)\%$
Finland	29627	28017	$(51.4 \pm 0.2)\%$
Indien	12849497	11582656	$(52.6 \pm 0.01)\%$
Argentinien	383852	370371	$(50.9 \pm 0.05)\%$

Daten für 2016, Quelle: Weltbank

Binomialverteilung – k klein

- Stichprobe in Fabrik: 0 (1) von 100 überprüften Produkten sind fehlerhaft
- In welchem Bereich ist die Fehlerquote der gesamten Produktion zu erwarten?
- k ist nicht groß $\rightarrow$ nicht mit momentanem Werkzeug behandelbar
- Achtung: Unreflektierte Anwendung der Formel ergibt eine Quote von 0 ± 0!

Quelle: Wikipedia

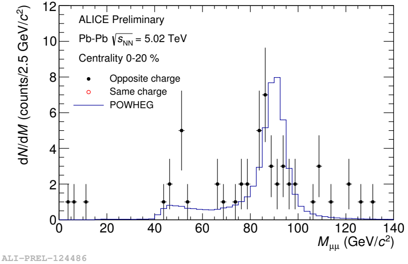
Poissonverteilung

- Typischer Fall: k/N sehr klein (seltene Ereignisse)
- Beispiel: Goldatome in 1l Meerwasser
- Jedes Goldatom hat sehr kleine Wahrscheinlichkeit, in Probe zu gelangen
- Poissonverteilung: Grenzfall Binomialverteilung für $k \ll N$, $\theta \cdot N$ endlich
- Definiere $\lambda = \theta \cdot N$, dann Verteilung ungefähr unabhängig von N
- Vorteil: N muss nicht genau definiert sein (Goldatome in allen Ozeanen oder nur in diesem?
- Für seltene, unabhängige Ereignisse: $p(k|\lambda)$, die Poissonverteilung

Poissonverteilung – Beispiele

Voraussetzung: Anzahl seltener und unabhängiger Ereignisse

- Atome in einem Kubikmeter Weltraum
- Anzahl Zerfälle in radioaktivem Material in Zeitspanne
- Anzahl der Photonen in Kamerapixel in Zeitspanne
- Einträge in bestimmtem Bin in Histogramm
- Von Pferden zu Tode gekickte Soldaten der preußischen Armee (Bortkiewicz 1898)

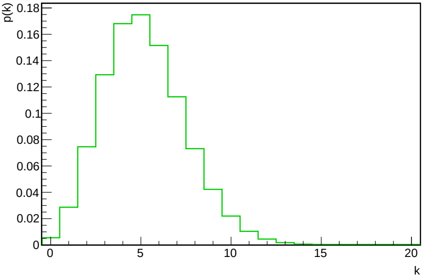


Z Boson Messung, ALICE Kollaboration

Poissonverteilung – Eigenschaften

$$p(k|\lambda) = \frac{\lambda^k \ e^{-\lambda}}{k!}$$

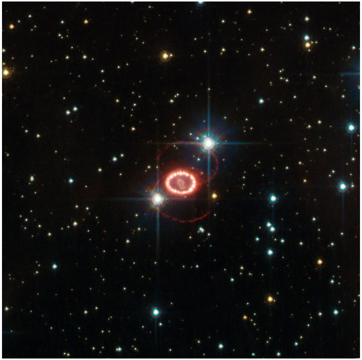
- Erwartungswert $\mu = \lambda$
- Varianz: $\sigma^2 = \lambda$



Poissonverteilung, $\lambda = 5.2$

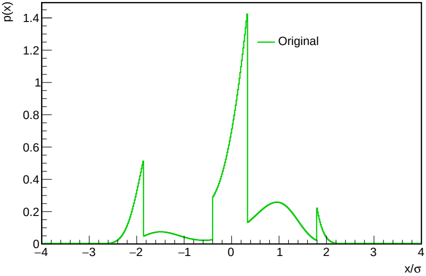
Poissonverteilung – Unsicherheit

- Kamiokande II Messung: 12 Neutrinos $\rightarrow$ Was ist die Unsicherheit auf die Gesamtanzahl?
- Für $k \gg 1$, $k \approx \lambda$
- $\sigma_k^2 = \lambda \approx k$
- Ergebnis also $12 \pm \sqrt{12}$

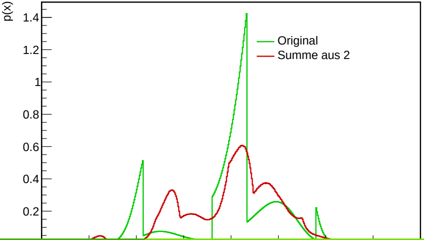


ESA/Hubble & NASA

Normalverteilung – Zentraler Grenzwertsatz

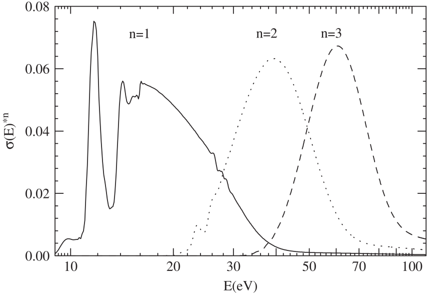


- Addition von (unkorrelierten) Zufallszahlen aus beliebiger Verteilung
- Resultierende Verteilung aus Faltung der Ursprungsverteilung
- Konvergenz zu Gauß-/Normalverteilung



Normalverteilung – Zentraler Grenzwertsatz (2)

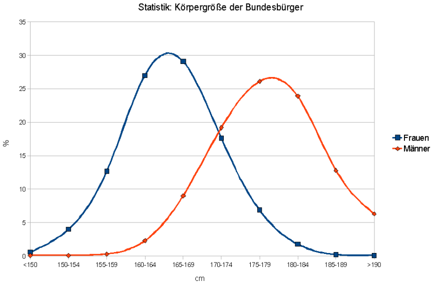
- Konvergenz unter recht allgemeinen Bedingungen
- Notwendige Voraussetzung: Mittelwert und Varianz existieren
- Wichtige Ausnahme: Energieverlust von geladenen Teilchen in Gas: $p(\Delta E) \sim 1/E^2$
- Abweichungen von der Normalverteilung oft an den Enden der Verteilung besonders deutlich



A method to improve tracking and particle identification in TPCs and silicon detectors, Hans Bichsel

Normalverteilung – Beispiel

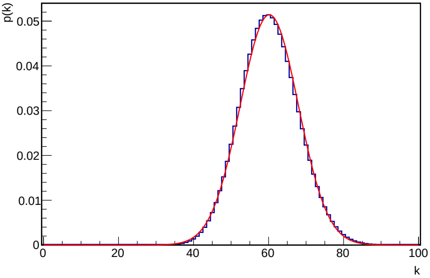
- Größe eines Menschen
- Verschiedene Faktoren (Vererbung, Umwelteinflüsse) haben Einfluss
- Viele recht unabhängig



Quelle: Wikipedia

Normalverteilung – Beispiel (2)

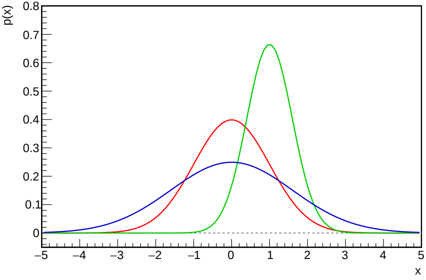
- Binomialverteilung: Summe vieler einzelner Ereignisse
- 0 oder 1 ist auch eine Wahrscheinlichkeitsverteilung
- Binomialverteilung konvergiert gegen Normalverteilung für $n \to \infty$
- Gleiches gilt für Poisson



Poissonverteilung mit $\lambda = 60.2$, Gaußverteilung

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mit Abstand die wichtigste statistische Verteilung in der Physik
- Erwartungswert: μ
- Varianz: σ^2
- Namen: Normalverteilung, Gaußverteilung, Gauß'sche Glockenkurve, Glockenkurve, Laplace-Gauß-Kurve uvm.

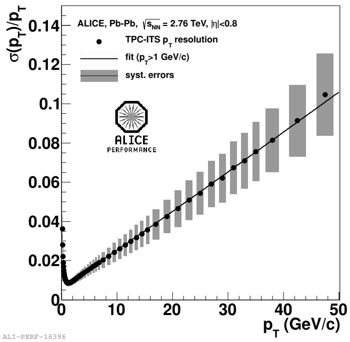


Gaußverteilungen

- Typischer und einfacher Fall: σ aus separater Analyse bekannt
- Angenommen Teilchen gemessen: p_t ist 15 $\mathrm{GeV/c}$
- Mit 3% Unsicherheit:

$$p_\mathrm{t} = 15 \pm 0.45\ \mathrm{GeV/c}$$

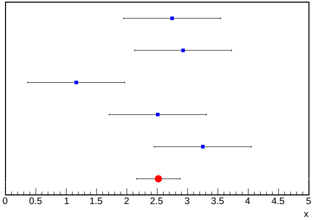
- Für den Moment: Ignoriere Impulsabhängigkeit der Auflösung und Unsicherheit auf die Unsicherheit



ALICE Kollaboration

Normalverteilung – Unsicherheit (2)

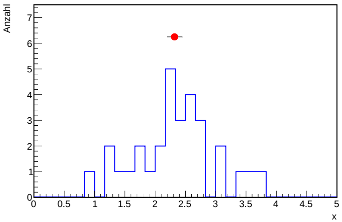
- Ebenfalls typischer Fall: σ aus separater Analyse bekannt, mehrere Messungen
- Arithmetisches Mittel ergibt kleinere Unsicherheiten
- $\mu_{tot} = \frac{1}{N} \sum_i \mu_i$
- $\sigma_{tot} = \frac{\sigma}{\sqrt{N}}$
- Durch zentralen Grenzwertsatz: Auch eingeschränkt anwendbar, wenn Einzelmessung nicht einer Gaußverteilung folgt
- In diesem Fall gibt es aber häufig effizientere Methoden (siehe später)



Mehrere Messungen und Mittelwert, Beispiel (wahrer Wert 2.3)

Normalverteilung – Unsicherheit (3)

- Mittelwert und Varianz unbekannt, mehrere Messungen
- Schätze Standardabweichung aus Varianz der Messungen ab
- $\hat{\sigma}^2 = \frac{\sum_i (x_i - \bar{x})^2}{N-1}$
- Unerwartetes $N-1$: Korrigiert "bias" – mehr dazu später
- $\hat{\mu}_{tot} = \frac{1}{N} \sum_i x_i$
- Und damit: $\sigma_{tot} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N(N-1)}}$



Mehrere Messungen und Mittelwert, Beispiel (wahrer Wert 2.3)

Normalverteilung – Beispiel (4)

- Galtonbrett: Kugeln können in jedem Schritt in zwei Richtungen weiter
- Beispiel von Binomialverteilung ($N = 10$), im Grenzfall normalverteilt
- Modellbeispiel für "random walk" (Zufallsbewegung) $\rightarrow$ Führt zu Normalverteilung der Endzustände
- Z.B. Mehrfachstreuung: Ein geladenes Teilchen fliegt durch Materie und interagiert mit einzelnen Atomen
- (Größtenteils) unabhängig von der genauen Physik ist die Gesamtablenkung normalverteilt (solange der Gesamtenergieverlust gering ist)

Quelle: Wikipedia

Fehlerfortpflanzung

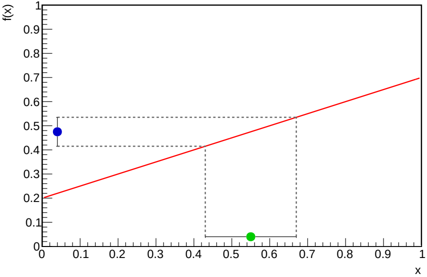
- Mit einer Unsicherheit für x, was ist die Unsicherheit von $f(x)$?
- Z.B. Pendelperiode gemessen als $T \pm \Delta T$: was ist $g \pm \Delta g$?
- Was passiert bei mehreren Quellen von Unsicherheiten?
- Z.B. Pendelperiode und Pendellänge mit endlicher Genauigkeit gemessen
- Zunächst einfache Spezialfälle, dann näherung für allgemeinen Fall
- Wichtig: Alles Folgende unter der Annahme *unkorrelierter* Unsicherheiten

Lineare Funktionen

- $f(x) = ax + b$
- Fortpflanzung intuitiv klar

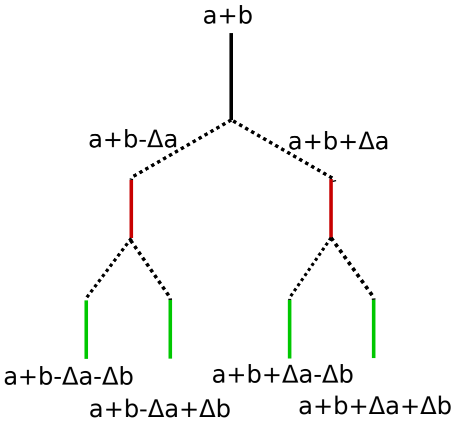
$$\hat{f}(x) = (a\hat{x} + b) \pm (a \cdot \Delta x)$$

- Dies ist in allen Fällen exakt



Summe zweier Messwerte

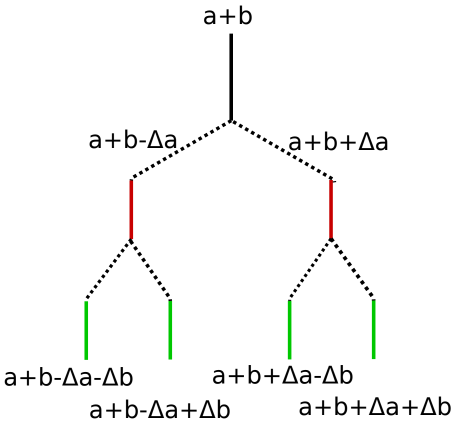
- Heuristische Herleitung (später genauer)
- Stelle die Messung $a \pm \Delta a$ dar durch die beiden Punkte $a + \Delta a$ und $a - \Delta a$
- Diese Verteilung hat $\sigma^2 = (\Delta a)^2$
- Summe aus a und b mit ihren Unsicherheiten ergibt 4 mögliche Kombinationen
- Varianz der Kombinationen ist:



$$\sigma_{tot}^2 = 1/4((\Delta a + \Delta b)^2 + (\Delta a - \Delta b)^2 + (-\Delta a + \Delta b)^2 + (-\Delta a - \Delta b)^2)$$
$$= (\Delta a)^2 + (\Delta b)^2$$

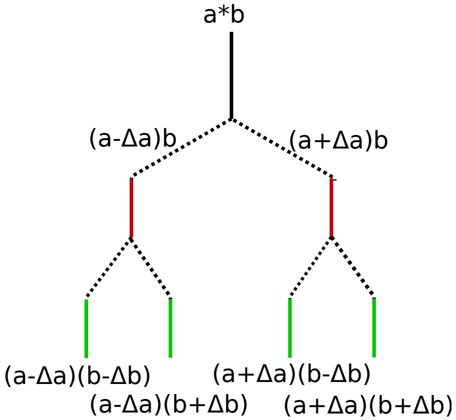
Summer zweier Messwerte (2)

- Bei Addition zweier Werte addieren sich die Varianzen
- D.h. die Unsicherheiten werden quadratisch addiert: $\sigma_{tot} = \sqrt{\sigma_a^2 + \sigma_b^2}$
- Allgemein hängt dies damit zusammen, dass sich die Momente von Funktionen unter Faltungen linear addieren
- Wichtig: Das funktioniert nur, wenn alle Kombinationen gleich wahrscheinlich sind, d.h. die Unsicherheiten unkorreliert sind
- Gilt für statistische und systematisch Unsicherheiten



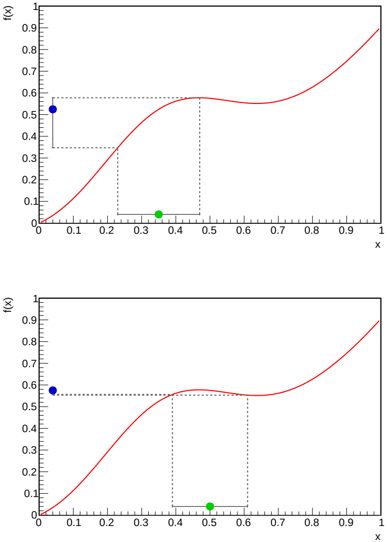
Produkt zweier Messwerte

- Simpler, relative Unsicherheiten zu betrachten: $\sigma_{a,rel} = \sigma_a / a$
- Strenggenommen keine gute Definition, da a unbekannt
- Umformen: $(a \pm \sigma_a)(b \pm \sigma_b) = ab(1 \pm \sigma_{a,rel} \pm \sigma_{b,rel} + \sigma_{a,rel}\sigma_{b,rel}) \approx ab(1 \pm \sigma_{a,rel} \pm \sigma_{b,rel})$
- Gewöhnliche Summe in der Klammer
- Bei Produkten addieren sich also die relativen Unsicherheiten quadratisch
- Näherung für $\sigma_{a,rel}, \sigma_{b,rel} \ll 1$



Allgemeine Funktion eines Parameters

- Für allgemeine Funktion:
  - $f(x)$ nicht mittig zwischen $f(x+\sigma_x)$ und $f(x-\sigma_x)$
  - Am oberen Ende: Viele Eingangswerte ergeben Resultate in kleiner Umgebung
- Pathologischer Fall (rechts)
- Keine intuitive, heuristische Lösung
- Allgemeiner Fall wird in VL 3-5 behandelt
- Hier: Wichtige Näherung

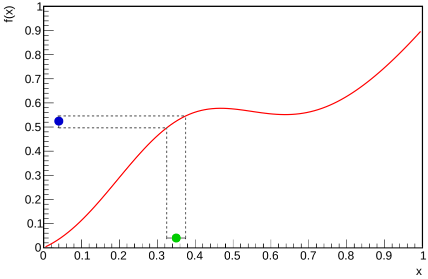


Allgemeine Funktion eines Parameters (2)

- Näherung für kleine Unsicherheiten
- Voraussetzung: Innerhalb ein paar Standardabweichungen ist die Funktion linear
- In dem Fall: Lösung für lineare Funktionen benutzen
- Gemessen $\hat{x} \pm \sigma_x$, dann als Ergebnis:

$$f(\hat{x}) \pm \left| \frac{\partial f}{\partial x} \right| \sigma_x$$

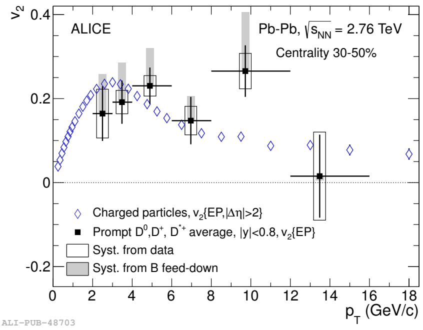
- Bekannt als *Gauß'sches Fehlerfortpflanzungsgesetz*
- Verallgemeinerung auf höhere Ordnungen nichttrivial



Allgemeine Funktion mehrerer Parameter

- Ergebnis bisher:
  - Addition von Unsicherheiten
  - Fortpflanzung durch Funktion
- Verschiedene Messwerte x_i mit Unsicherheiten σ_i, sowie Funktion $f(x_1, \dots, x_n)$
- Unsicherheiten aus einzelnen Quellen werden quadratisch addiert:

$$\sigma_{tot} \approx \sqrt{\sum_i \left(\left| \frac{\partial f}{\partial x_i} \right| \sigma_i \right)^2}$$



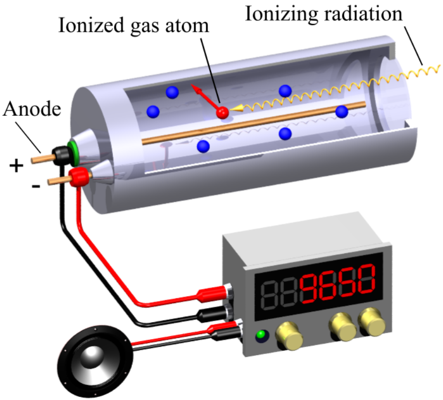
D meson elliptic flow in non-central Pb–Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV, ALICE Kollaboration

$$v_2^{\text{prompt}} = \frac{v_2^{\text{all}}}{f_{\text{prompt}}} - \frac{1 - f_{\text{prompt}}}{f_{\text{prompt}}} v_2^{\text{fd}}$$

$$v_2 = \frac{1}{R_2} \frac{\pi}{4} \frac{N_{\text{in-plane}} - N_{\text{out-of-plane}}}{N_{\text{in-plane}} + N_{\text{out-of-plane}}}$$

Fehlerfortpflanzung – Beispiel

- Ein Messgerät (z.B. Geiger–Müller–Zählrohr) misst Zerfälle in der Probe; gesucht ist die Halbwertszeit des Stoffes:
  - Die Effizienz des Gerätes ist $\epsilon = 5\%$
  - Die Stoffmenge ist $N = 10^{10}$
  - Es werden $n = 8562$ Zerfälle in $t = 5\ \mathrm{min.}$ gemessen
  - Alle Produkte stabil, kein Untergrund
- $n \approx \lambda \epsilon N t = \frac{\ln(2)}{t_{1/2}} \epsilon N t$
- oder $t_{1/2} = \ln(2) \epsilon N t / n$
- Poisson: $n = 8562 \pm \sqrt{8562} = 8562 \pm 93$



Quelle: Wikipedia

Fehlerfortpflanzung – Beispiel

- Ein Messgerät (z.B. Geiger–Müller–Zählrohr) misst Zerfälle in der Probe; gesucht ist die Halbwertszeit des Stoffes:
  - Die Effizienz des Gerätes ist $\epsilon = 5\%$
  - Die Stoffmenge ist $N = 10^{10}$
  - Es werden $n = 8562$ Zerfälle in $t = 5\ \mathrm{min.}$ gemessen
  - Alle Produkte stabil, kein Untergrund
- $n \approx \lambda \epsilon N t = \frac{\ln(2)}{t_{1/2}} \epsilon N t$
- oder $t_{1/2} = \ln(2) \epsilon N t / n$
- Poisson: $\hat{n} = 8562 \pm \sqrt{8562} = 8562 \pm 93$
- Ableitung für Fehlerfortpflanzung:

$$\left| \frac{\partial t_{1/2}}{\partial n} \right| = t_{1/2}(n)/n$$

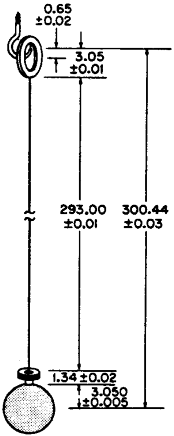
- $t_{1/2} = \ln(2) \epsilon N t / \hat{n} \pm \ln(2) \epsilon N t / \hat{n}^2 \cdot \sigma_n$
- Oder:

$$t_{1/2} = (\ln(2) \epsilon N t / \hat{n})(1 \pm \frac{\sigma_n}{\hat{n}})$$

- D.h. bei Inversion bleibt die relative Unsicherheit gleich
- Somit: $t_{1/2} = 140.5 \pm 1.5 d$

Fehlerfortpflanzung – Beispiel (2)

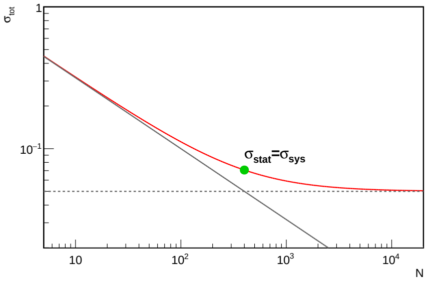
- $g = (2\pi/T_0)^2 \cdot L$
- Länge des Fadens:
  $L = 3.0044 \pm 0.0003\ \mathrm{m}$
- Periode (100 Oszillationen, 10 Messungen): $T = 3.47880 \pm 0.00017\ \mathrm{s}$
- Korrektur zu idealem Pendel
  $T_0 = T - 472\ \mu\mathrm{s} = 3.47833\ \mathrm{s}$
- $|\partial g/\partial T_0|\sigma_T = 2g\frac{\sigma_T}{T_0} = 0.001\ \mathrm{m} \cdot \mathrm{s}^{-2}$
- $|\partial g/\partial L|\sigma_L = g\frac{\sigma_L}{L} = 0.001\ \mathrm{m} \cdot \mathrm{s}^{-2}$
- Somit: $g = 9.803 \pm 0.0014\ \mathrm{m} \cdot \mathrm{s}^{-2}$
- Tatsächlich $g = 9.8026\ \mathrm{m} \cdot \mathrm{s}^{-2}$



The pendulum – Rich physics from a simple system, Robert A. Nelson

Schrumpfende Erträge

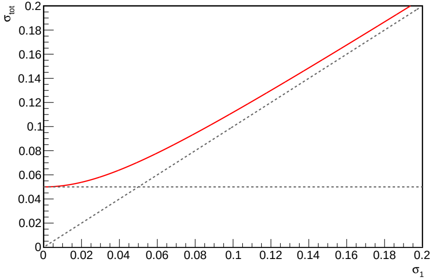
- $\sqrt{\sigma_{stat}^2 + \sigma_{sys}^2}$ dominiert vom größeren der Beiträge
- $\sigma_{stat} \sim 1/\sqrt{n}$
- Falls $\sigma_{stat} \gg \sigma_{sys}$: vierfache Messzeit halbiert Unsicherheit
- Sobald andere Unsicherheiten signifikant: deutlich weniger Fortschritt
- Zur Planung von Experimenten: Ab $\sigma_{stat} \approx \sigma_{sys}$ guter Punkt um aufzuhören
- n typischerweise proportional zu Zeit und laufenden Kosten



Beispiel, $\sigma_{sys} = 0.05$, $\sigma_{stat} = 1/\sqrt{N}$

Schrumpfende Erträge (2)

- $\sqrt{\sigma_1^2 + \sigma_1^2}$ dominiert vom größeren der Beiträge
- $\rightarrow$ es ist viel wichtiger die größeren Beiträge richtig abzuschätzen als die kleinen
- $\rightarrow$ es ist viel wichtiger, keinen großen Beitrag zu vergessen, als alle kleinen Beiträge perfekt abzuschätzen
- Für sehr kleine Unsicherheiten reicht also oft eine grobe Abschätzung



Summe zweier Unsicherheiten, $\sigma_2 = 0.05$

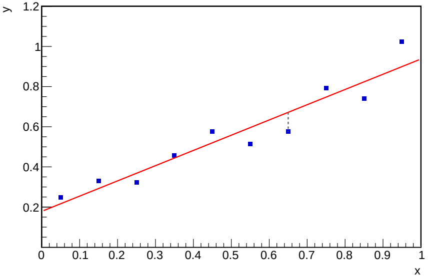
Anwendung: Gewichtete Summe

- Bei arithmetischem Mittel $f = \frac{1}{2}(a_1 + a_2)$ addieren sich die Unsicherheiten quadratisch: $\sigma_{tot} = \frac{1}{2}\sqrt{\sigma_a^2 + \sigma_b^2}$
- Verallgemeinerung: Gewichtetes Mittel $f = \frac{w_1 a_1 + w_2 a_2}{w_1 + w_2}$
- Unsicherheit: $\sigma_{tot} = \frac{1}{w_1 + w_2}\sqrt{(w_1\sigma_1)^2 + (w_2\sigma_2)^2}$

- Gleiches Ergebnis für $w_i \to \alpha w_i$
- σ_{tot} minimal für $\frac{w_1}{w_2} = \frac{\sigma_2^2}{\sigma_1^2}$
- Somit: $w_i = 1/\sigma_i^2$ beste Wahl
- Gewichtung mit inverser Varianz

(Lineare) Regression

- Wollen Funktion $f(x)$ mit freien Parametern durch Messpunkte (x_i, y_i) legen
- Messung von x exakt
- Ideen: Minimiere $S = \sum |y_i - f(x_i)|$ oder $S = \sum (y_i - f(x_i))^2$
- 2. – Methode der kleinsten Fehlerquadrate
- Heuristisch: Differenzierbar in Parametern
- Lineare Funktion $f = a + bx$, dann



$$a = \frac{\bar{y}\left(\sum x_i^2\right) - \bar{x}\left(\sum x_i y_i\right)}{\left(\sum x_i^2\right) - n\bar{x}^2}$$

$$b = \frac{\left(\sum x_i y_i\right) - n\bar{x}\bar{y}}{\left(\sum x_i^2\right) - n\bar{x}^2}$$

(Lineare) Regression (2)

- Falls Unsicherheiten bekannt: Gewichtung
- $S = \sum \frac{(y_i - f(x_i))^2}{\sigma_i^2}$
- Manchmal auch χ^2 genannt
- Genauere Messungen stärker gewichtet
- Begründung später
- Minimierung von S üblicherweise numerisch (üblicherweise in Software enthalten)