Statistik

– Vorlesung 5 –

Günter Duckeck

J. Elmsheuser

Sommersemester 2008

Inhalt:

- Parameterschätzung
- Maximum Likelihood

Überblick

- Einführung, Beispiele, Wahrscheinlichkeiten, Bayes Theorem
- Beschreibung von Daten und Verteilungen
- Monte Carlo Methoden
- Fehler
- **Parameterschätzung und Maximum-Likelihood**
- χ^2-Methode
- Hypothesentest
- Wahrscheinlichkeiten und Vertrauensintervalle
- Klassifizierung und Neuronale Netze

- Optimierung & Parametrisierung
- Datamining

Zwei grundlegende Fragestellungen in Statistik:

1. Ausgehend von theoretischen Verteilungen und gegebenen Parametern kann man mittels Wahrscheinlichkeitsrechnung Aussagen über die zu erwartenden Daten machen:

$$Theorie \Rightarrow Daten$$

2. Ausgehend von einem Datensatz will man Aussagen über die zugrundelegende Verteilung oder ihr Parameter machen, d.h. statistische Folgerungen *(statistical inference)* aus den Daten ziehen:

$$Daten \Rightarrow Theorie$$

5 Schätzung von Parametern

Im einfachsten Fall ist “statistical inference” die Schätzung von Parametern, d.h. Definition:

Prozedur, die auf den Datensatz (= Stichprobe) angewendet wird, und einen numerischen Wert für einen Parameter oder Eigenschaft der zugrundeliegenden Verteilung liefert

Aus gemessenem Datensatz $x_1, ..., x_n$ soll Parametersatz $a_1, ..., a_m = \vec{a}$ ermittelt werden, z.B. Mittelwert, Standardabweichung, u.a., der diese Daten **am besten** beschreibt.

Schätzung drückt aus, dass man nicht den exakten, wahren Wert bestimmen kann, sondern eine “Zufallsgröße”, die mehr oder weniger stark abweicht. Wiederholen der Messung unter gleichen Bedingungen *(weitere Stichprobe)* liefert i.d.R. anderen Wert.

Beispiele:

- Lebensdauer $\hat{\tau}$ eines instabilen Zustand aus N Messungen der Zerfallszeit t_i.
- Masse eine Teilchens $\hat{M}$ aus N Messungen der invarianten Masse der Zerfallsprodukte $\Rightarrow$ Breit-Wigner Kurve anpassen
- Mittlere Grösse $\hat{g}$ von Studenten

Konvention: Grösse mit "Hut", z.B. $\hat{\tau}$, drückt aus dass $\hat{\tau}$ ein Schätzer des wahren Parameters τ ist.

Mögliche Prozeduren

zur Bestimmung der mittleren Grösse $\hat{g}$ von Studenten:

Suche N repräsentative *(andres Problem)* Studenten aus, dann:

1. Alle aufaddieren und durch N teilen (arithm. Mittel)
2. Nur erste 10 verwenden, Rest ignorieren.
3. Alle multiplizieren und $N - te$ Wurzel nehmen
4. Kleinste und grösste addieren und durch 2 teilen
5. Alle aufaddieren und durch $N - 1$ teilen
6. Mittleren Wert nehmen (Median)
7. obere und untere 25% verwerfen, Rest mitteln
8. Einfach 1.80 m nehmen.
9. ...

Welches die beste Methode ist lässt sich nicht allgemein beantworten sondern hängt von der zugrundeliegenden Verteilung ab.

5.1 Kriterien für Schätzer:

- **Konsistent**

$$\lim_{n\to\infty} \hat{g} = g$$

Schätzwert nähert sich wahrem Wert g an mit zunehmendem N.

(Nicht erfüllt für 2., 8.)

- **Erwartungstreu (unbiased)**

$$\langle \hat{g} \rangle = g$$

Schätzwert ist unverzerrt.

(Nicht erfüllt für 5., 8.)

- **Effizient**
  Varianz $V(\hat{g})$ möglichst klein
- **Robust**
  unempfindlich gegen falsche Daten (Fehlmessungen, Verunreinigungen)

Oft nicht alle Kriterien gleichzeitig erfüllbar, insbesondere **effizient** , **erwartungstreu** und **robust** oft in Konflikt.

5.2 Schätzung von Mittelwert und Varianz

Arithmetischer Mittelwert ist Standard zur Schätzung des Mittelwerts einer Messreihe:

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Wenn zugrundeliegende Verteilung Mittelwert μ Varianz σ_0^2 hat, folgt aus zentralem Grenzwertsatz für die Varianz bzw Standardabweichung von $\bar{x}$:

$$V(\bar{x}) = \langle \bar{x} - \mu \rangle = \frac{\sigma_0^2}{n} \text{ bzw. } \sigma(\bar{x}) = \frac{\sigma_0}{\sqrt{n}}$$

$\bar{x}$ ist

- ist **erwartungstreu** und **konsistent**.
- mit Fehler $\pm\sigma_0/\sqrt{n}$
- Maximal **effizient** nur für bestimmte Verteilungen (Gauss)

Schätzung Varianz

Übliche Definition der **Varianz**

$$V(x) = \frac{\sum (x_i - \bar{x})^2}{N}$$

ist verzerrt:

$$V(x) = \frac{\sum (x_i - \bar{x})^2}{N} = \frac{1}{N} \sum \left[(x_i - \mu) - (\bar{x} - \mu) \right]^2 = \frac{1}{N} \left[\sum (x_i - \mu)^2 - \sum (\bar{x} - \mu)^2 \right]$$

Für den Erwartungswert folgt:

$$\langle V(x) \rangle = \frac{1}{N} \left\langle \sum (x_i - \mu)^2 \right\rangle - \left\langle \sum (\bar{x} - \mu)^2 \right\rangle = \frac{(N-1)}{N} \sigma^2$$

Stattdessen unverzerrter Schätzer für Varianz :

$$\hat{V} = s^2 = \frac{1}{N-1}\sum (x_i - \bar{x})^2, \left\langle s^2 \right\rangle = \sigma_0^2$$

Schätzung s^2 hat natürlich wiederum Unsicherheit, nach einiger Rechnung erhält man:

$$\sigma(s) = \frac{s}{\sqrt{2(n-1)}}$$

Vorsicht mit den Begriffen:

- $\bar{x}$ ist **Schätzer** $\hat{\mu}$ für wahren Mittelwert μ . $\bar{x}$ ist eine Zufallsgrösse, wiederholte Messung von $x_1, ..., x_n$ liefert anderen Wert $\bar{x}$.
- Wegen **CLT** sind die $\bar{x}$ Gauss-verteilt mit Standardabweichung $\pm\sigma_0/\sqrt{n}$, d.h. die Genauigkeit von $\bar{x}$ wächst mit $1/\sqrt{n}$
- s ist **Schätzer** $\hat{\sigma_0}$ für Breite (= Standardabweichung) σ_0 der zugrundeliegenden Verteilung *(unabhängig von n !)*.

5.3 Andere Schätzer für Mittelwert

Beispiel: Gleichverteilung

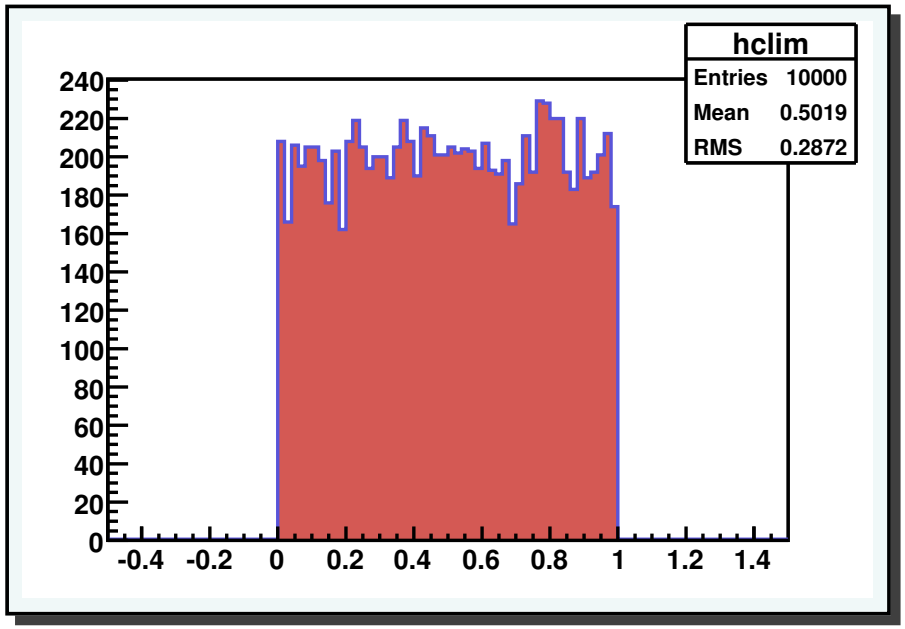
Mittelwert kann über $\bar{x}$ bestimmt werden, ist aber in dem Fall nicht effektiv, d.h. nicht die genaueste Schätzung.

Maximale Effizienz hat

$\hat{\mu} = (x_{max} - x_{min})/2$,

wobei x_{max} und x_{min} grösste und kleinste Werte der Stichprobe sind.

Man kann zeigen, dass die Genauigkeit dieses Schätzers mit $1/n$ wächst.



Beispiel: Effizienz getrimmter Mittelwert

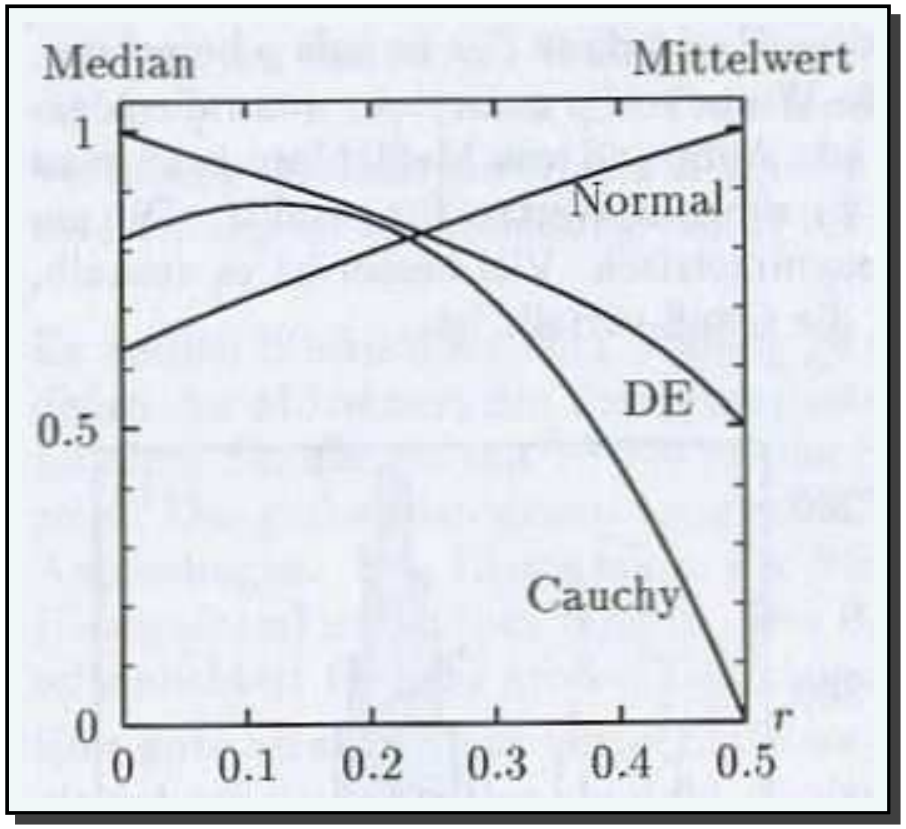
Robuster Schätzer für Mittelwert $\hat{\mu}$ ist der **getrimmte Mittelwert**,

$$\bar{x_r} = \frac{1}{2rn} \sum_{i=a}^{b} x_i$$

wobei a und b so gewählt sind, dass die niedrigsten bzw. höchsten $(1-2r)n/2$ verworfen werden.

Für $r = 0 \rightarrow \bar{x_r} =$ Median und für $r = 0.5 \rightarrow \bar{x_r} = \bar{x}$

Plot zeigt Effizienz für Gauss-, Cauchy- und Doppel-Exponentialverteilung in Abhängigkeit von r .



5.4 Maximum-Likelihood

Allgemeiner mathematischer Ansatz zur Schätzung von Parametern, wenn zugrundeliegende Verteilung $p(x, \vec{a})$ bekannt ist:

Die Wahrscheinlichkeit für einen Datensatz $x_1, ..., x_n$ und zugehörigen Parametersatz $\vec{a}$ lässt sich als Produkt der Einzelwahrscheinlichkeiten darstellen:

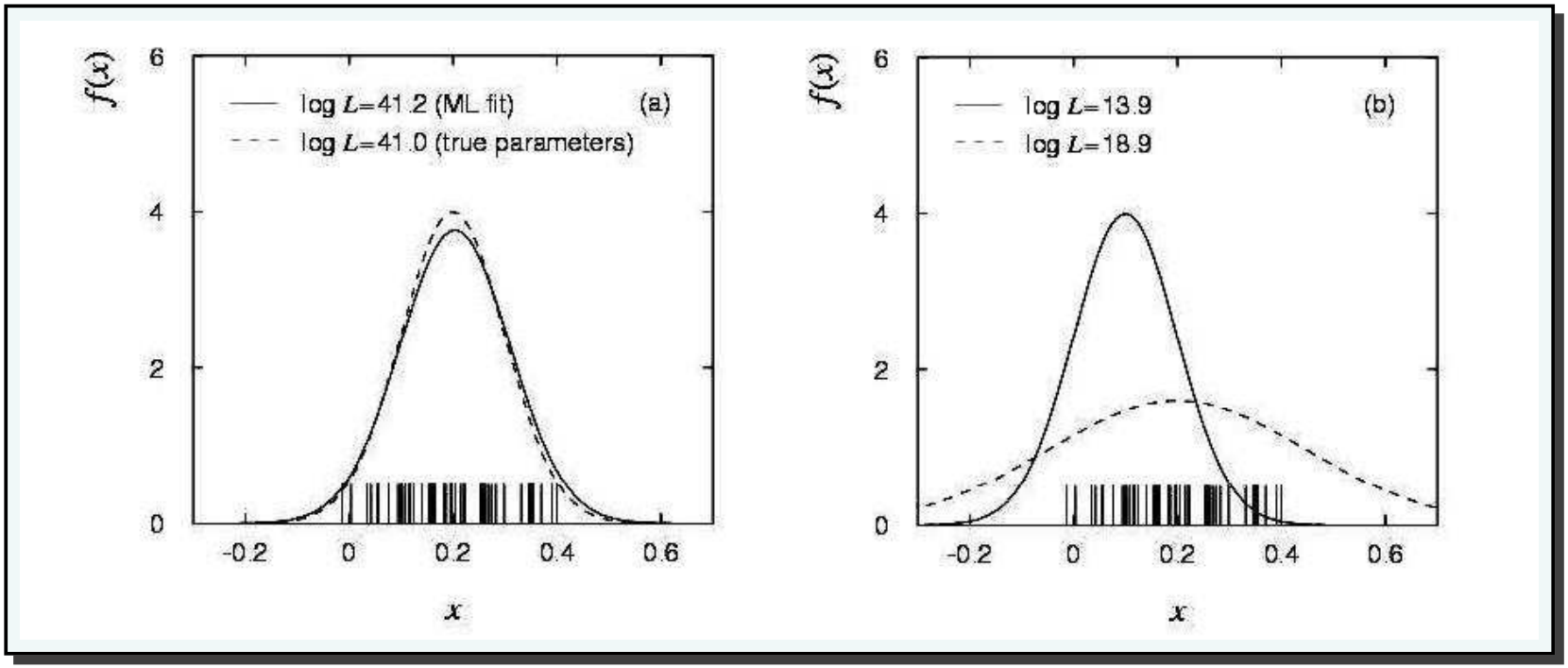
$$\mathcal{L}(x_1, x_2, ..., x_n|\vec{a}) \equiv p(x_1, \vec{a}) \cdot p(x_2, \vec{a}) \cdot ... \cdot p(x_N, \vec{a}) = \Pi\, p(x_i, \vec{a})$$

Die Schätzung $\hat{\vec{a}}$ ist derjenige Parametersatz, der $\mathcal{L} = \Pi\, p(x_i, \vec{a})$ maximiert.

Log-Likelihood: Praktischer ist es statt $\mathcal{L}$ den Logarithmus zu benutzen:

$$\ln \mathcal{L} = \sum_{i=1}^{n} \ln p(x_i, \vec{a})$$

Wenn $\vec{a}$ nahe bei den wahren Werten $\vec{a_0}$ liegt sollte die Wahscheinlichkeit hoch sein gerade den Datensatz $x_1, ..., x_n$ zu finden:



D.h. die Maximum Likelihood Schätzer sind die Parameter $\hat{\vec{a}}$ für die $\mathcal{L}$ maximal ist (bzw. $\ln \mathcal{L}$, Maximum an der gleichen Stelle).

Differenzieren nach $\vec{a}$ liefert Maximum:

$$\frac{\partial \mathcal{L}}{\partial \vec{a}} = 0 \quad \text{bzw.} \quad \frac{\partial \ln \mathcal{L}}{\partial \vec{a}} = 0$$

Nur in wenigen Fällen einfach analytisch lösbar. Im allgemeinen numerische Minimierungs-Verfahren nötig $\Rightarrow$ nächste Vorlesung

Beispiel Lebensdauer Messung:

- Zur Bestimmung der Lebensdauer $\hat{\tau}$ eines radiokativen Kerns hat man N Messungen der Zerfallszeit t_i gemacht.
- Man erwartet eine Exponentialverteilung: $p(t,\tau) = 1/\tau \exp(-t/\tau)$
- Log-Likelihood: $\ln \mathcal{L} = \sum \ln(1/\tau \exp(-t_i/\tau)) = \sum(-t_i/\tau - \ln \tau)$
- Nullstelle Ableitung: $\frac{\partial \mathcal{L}}{\partial \tau} = \sum(\frac{t_i}{\hat{\tau}^2} - \frac{1}{\hat{\tau}}) = 0$
- Simples Resultat:

$$\hat{\tau} = \frac{1}{N}\sum t_i$$

Beispiel Gauss-Verteilung

Likelihood:

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

$$\ln\mathcal{L} = \sum(-\frac{(x_i-\mu)^2}{2\sigma^2} - \ln(2\pi\sigma^2)/2)$$

Mittelwert durch Differenzieren nach μ :

$$\frac{\partial \ln\mathcal{L}}{\partial\mu} = 0 \rightarrow \frac{\sum(x_i-\mu)}{\sigma^2} = 0 \rightarrow \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Varianz durch Differenzieren nach σ^2 :

$$\hat{\sigma^2} = \frac{1}{n}\sum(x_i-\hat{\mu})^2$$

Wichtiges Resultat:

Standard-Mittelwert und -Varianz entsprechen den Maximum Likelihood Schätzern bei Gauss-Verteilung !

Allerdings:

ML Estimate für $\hat{\sigma^2}$ ist verzerrt: $\hat{\sigma^2} = \frac{n-1}{n}\sigma^2$

Unverzerrter Schätzer ist $\hat{s^2} = \frac{1}{n-1}\sum(x_i - \hat{\mu})^2$ (s.o.).

Generelles Feature von ML Schätzern:

- konsistent
- maximal effizient
- aber i.a. Verzerrung möglich

5.5 Varianz von Schätzern

Man kann allgemein zeigen, dass für beliebige Schätzer die 2. Ableitung der Log-Likelihood eine untere Schranke für die Effizienz von Schätzern liefert *(Rao–Cramer–Frechet Ungleichung)*:

$$V(\hat{\theta}) \geq -\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

(gültig für unverzerrte Schätzer $\hat{\theta}$)

$\Rightarrow$ Herleitung s. Literatur (z.B. *R.J. Barlow, Statistics*).

Bsp. Mittelwert bei Gauss-Verteilung:

Bei Gauss-Veteilung

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

lässt sich die Varianz des Mittelwerts relativ einfach analytisch berechnen:

$$V(\hat{\mu}) = -\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \mu^2}\right)^{-1}\Bigg|_{\mu=\hat{\mu}} = -\left(\frac{-N}{\sigma^2}\right)^{-1} = \frac{\sigma^2}{N}$$

Dies ist effizientester Schätzer, d.h. reproduziert Varianz nach **CLT**: $\frac{\sigma^2}{N}$.

Numerische/graphische Verfahren

Gauss– und Exponential-Verteilung sind Ausnahmefälle bei denen sich die Log-Likelihood und Ableitungen leicht analytisch berechnen lassen.

Im allgemeinen Fall sind numerische Verfahren zur Bestimmung der Maxima bzw Ableitungen nötig.

Die Varianz eines ML Schätzers ergibt sich für den allgemeinen Fall aus einer Taylor Reihe:

$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\hat{\theta}) + \left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

$$\rightarrow \ln \mathcal{L}(\theta) \approx \ln \mathcal{L}_{max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}^2}$$

und damit:

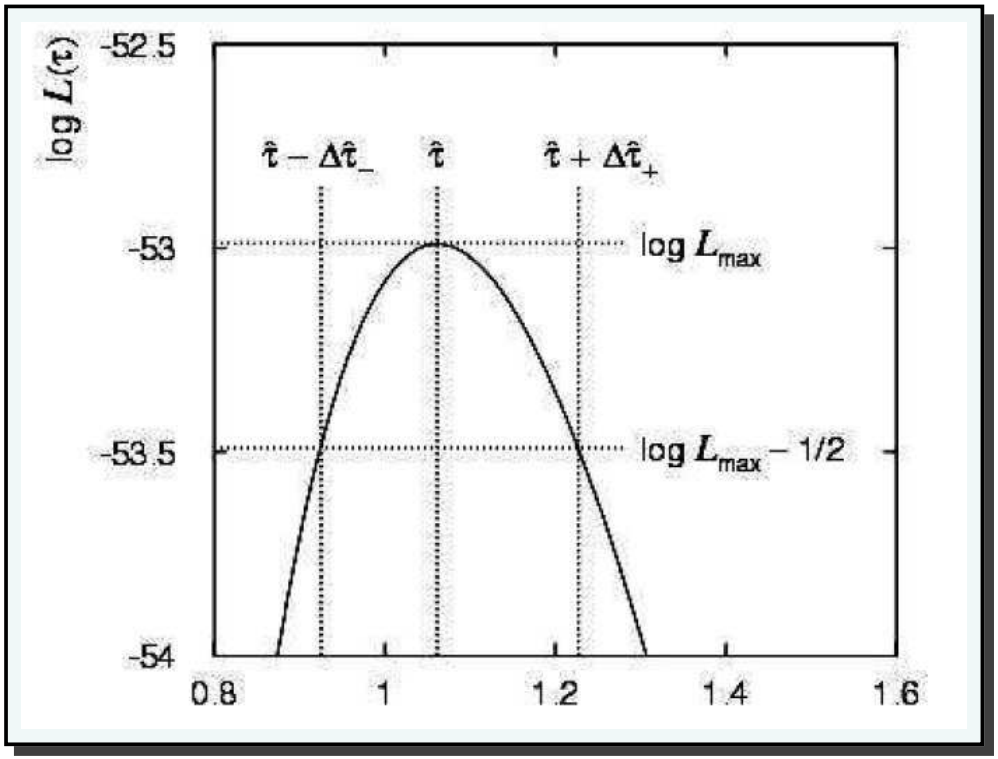
$$\ln \mathcal{L}(\hat{\theta} \pm \sigma) = \ln \mathcal{L}_{max} - \frac{1}{2}$$

Das heisst:

Eine Änderung des Parameters um eine Standardabweichung ändert die Log-Likelihood um 0.5 von ihrem Maximum.

Basis für allgemeine numerische Fehlerdefinition:

Variiere Parameter θ so dass sich Log-Likelihood um 0.5 verringert:

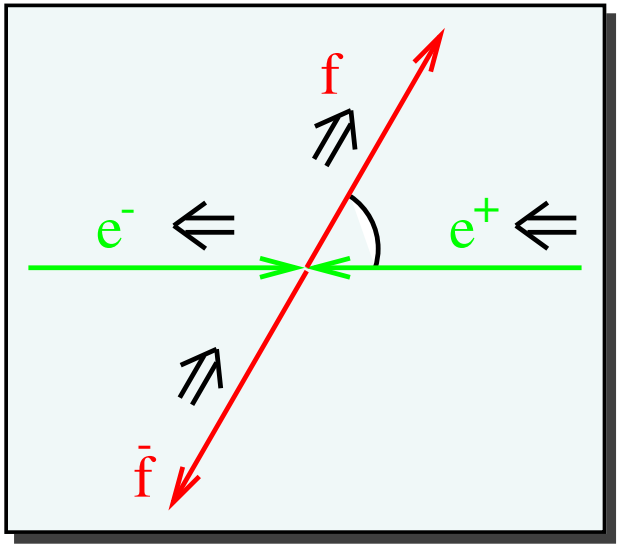


5.6 Beispiel Bestimmung Asymmetrie

Klassische Messung in Teilchenphysik ist Asymmetrie der erzeugten Teilchen in Reaktion.

Parametrisierung:

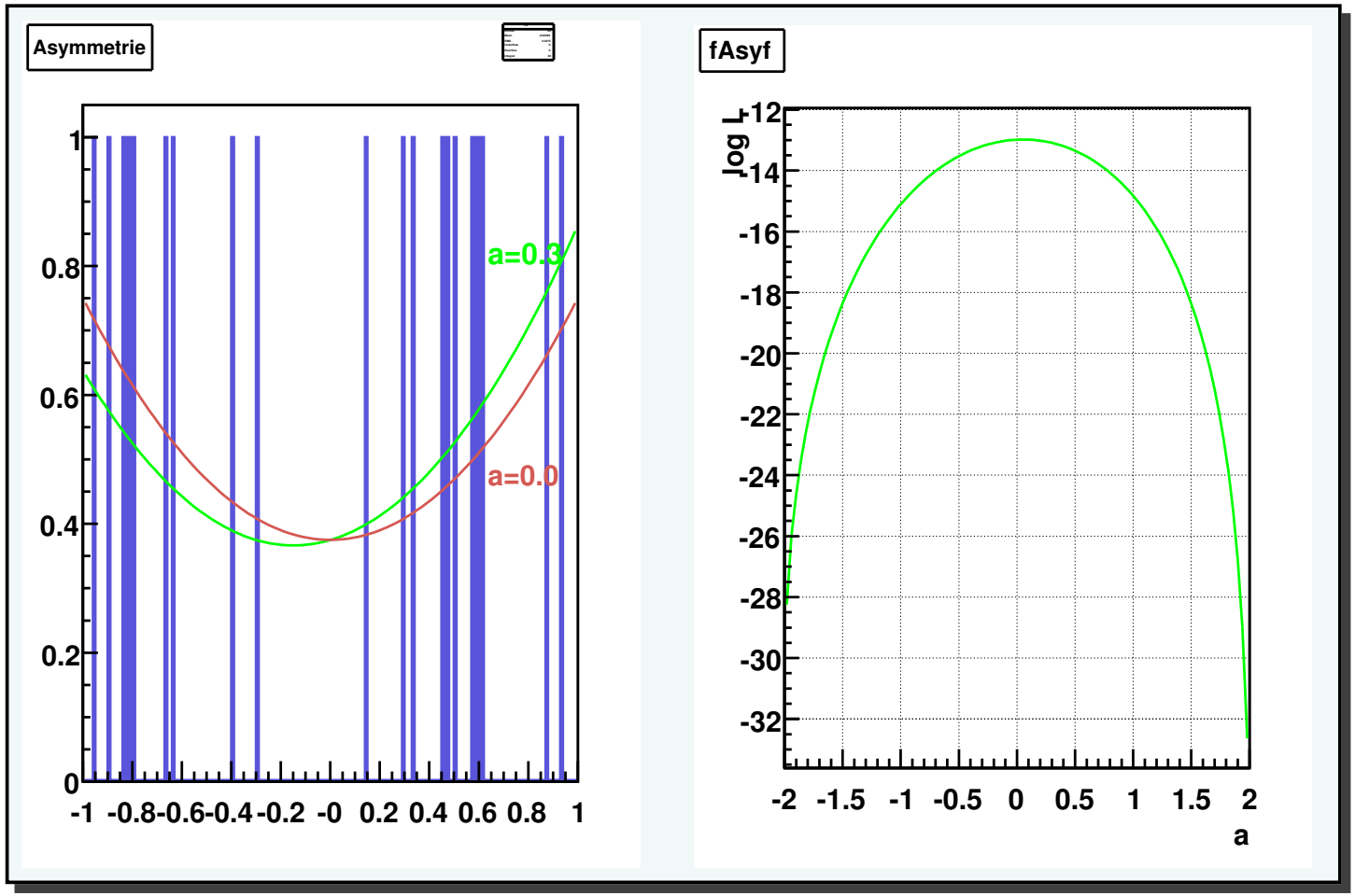
$$P(\cos\theta) = \frac{3}{8}(1 + a \cdot \cos\theta + cos^2\theta)$$



Messung von $x_i = \cos\theta_i$ für N Ereignisse, bestimmen der Log-Likelihood Kurve als Funktion von a :

$$\ln\mathcal{L} = \sum \ln \frac{3}{8}(1 + a \cdot x_i + x_i^2)$$

Asymmetrie: Zufallsdaten und Log-Likelihood



5.7 Zusammenfassung

- Im allgemeinen Fall ist Auswahl geeigneter (konsistent, unverzerrt, effizient und robust) Schätzer nicht-trivial
  - abhängig von Details der zugrundeliegenden Verteilung
- Maximum-Likelihood Methode liefert generelles Verfahren zur Schätzung von Parametern wenn zugrundeliegende Verteilung bekannt ist
  - Im allgemeinen numerische Verfahren zur Bestimmung des Minimums der Log-Likelihood nötig.
  - ML Schätzer sind maximal effizient, d.h. Varianz ist minimal.
- Bei Gauss-Verteilung entsprechen Standard-Verfahren für Mittelwert und Varianz den ML Schätzern.

Ausblick

- Methode der kleinsten Quadrate
- Fitten in ROOT
  - Vordefinierte Funktionen
  - allgemeiner Fit/Minimierung mit Minuit
- Güte eines Fits
- Fitten mehrerer Parameter und Korrelation
- Berücksichtigung korrelierter systematischer Effekte