

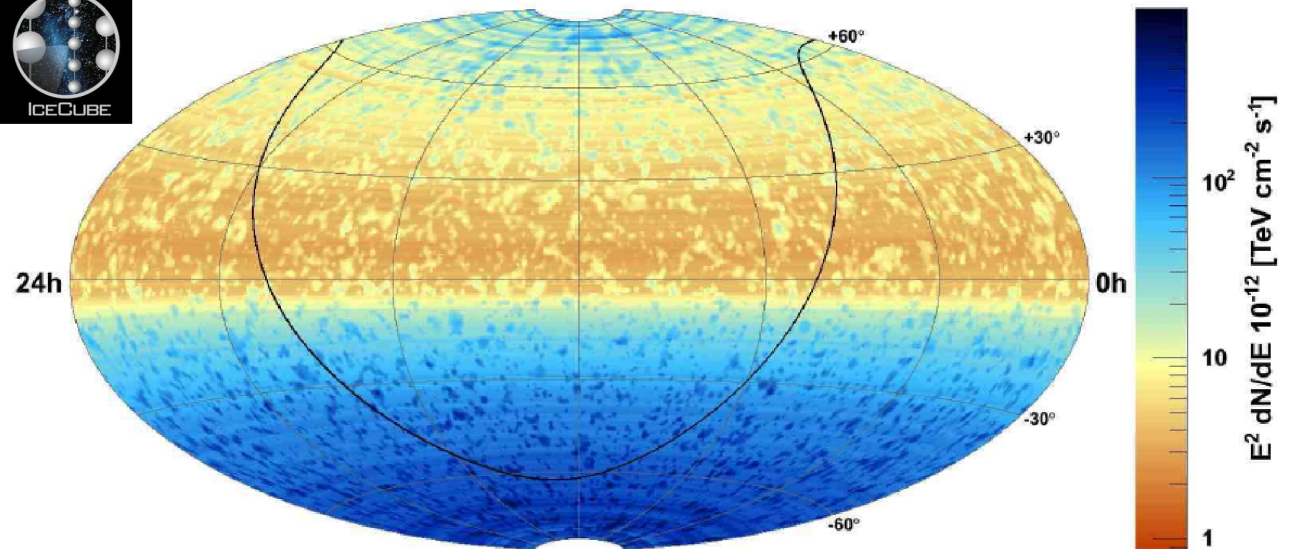
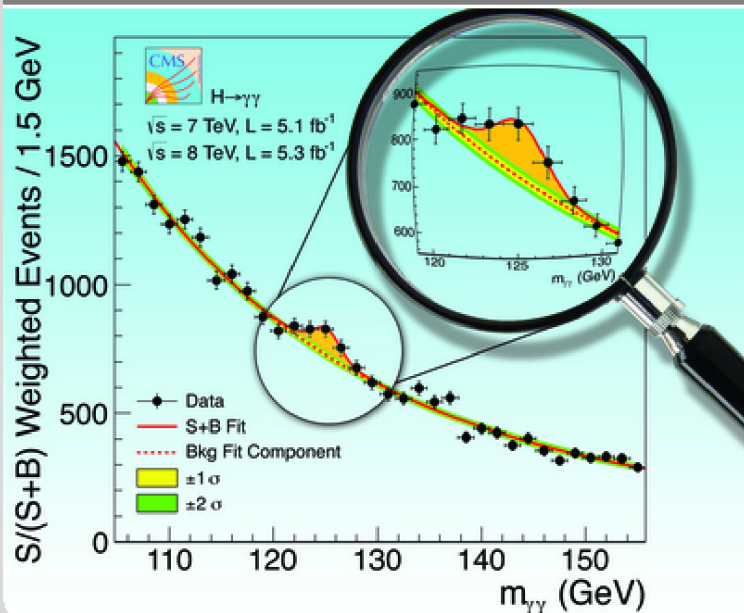
Moderne Methoden der Datenanalyse

30. Mai 2017

Ralf Ulrich, Andreas Meyer

Institut für Kernphysik, Institut für Experimentelle Kernphysik

Master-Kurs SS 2017

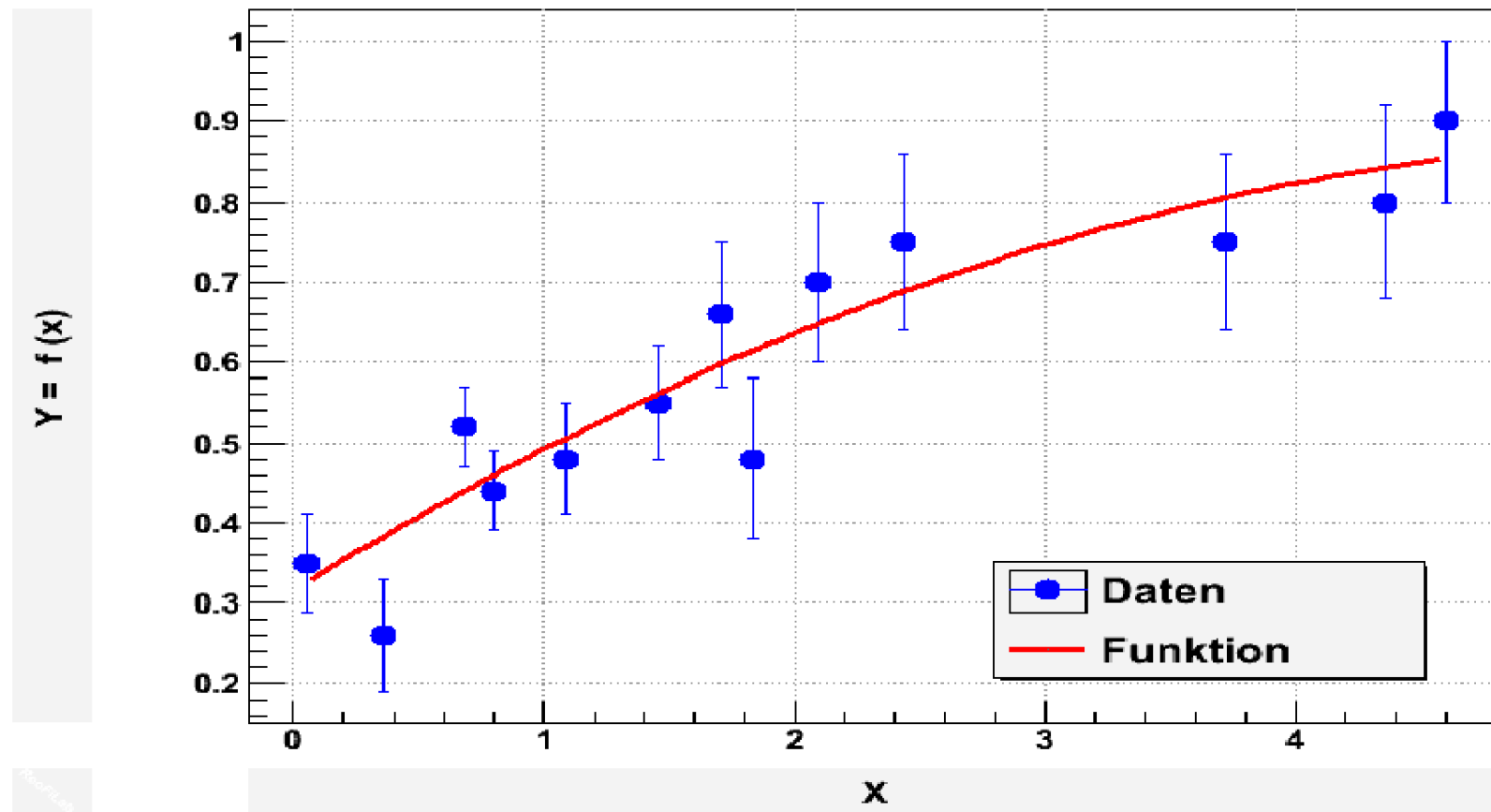


Themen, Volesungsprogramm

- Einführung: Überblick und grundlegende Konzepte (1)
- Einführung: Überblick und grundlegende Konzepte (2)
- Zufallszahlen und Monte-Carlo Methoden
- **Hypothesentests (1)**
- **Parameterschätzung**
- **Parameterschätzung (Goodness-of-fit)** 
- **Optimierungs- und Parametrisierungsmethoden**
- Konfidenzintervalle
- Hypothesentests (2)
- Klassifikation
- Klassifikation
- Klassifikation
- Messen und Entfalten
- Systematische Unsicherheiten

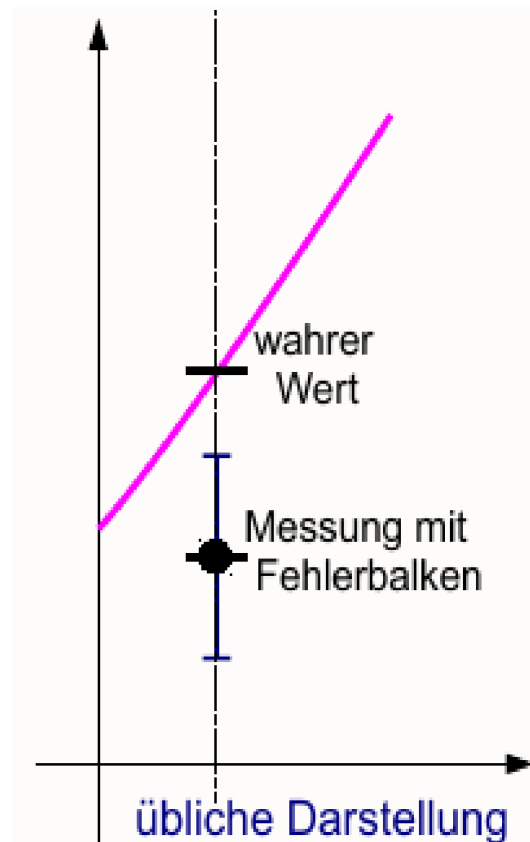
Was bedeuten Fehlerbalken?

- Und was hat dies für Konsequenzen für die Parameteranpassung?



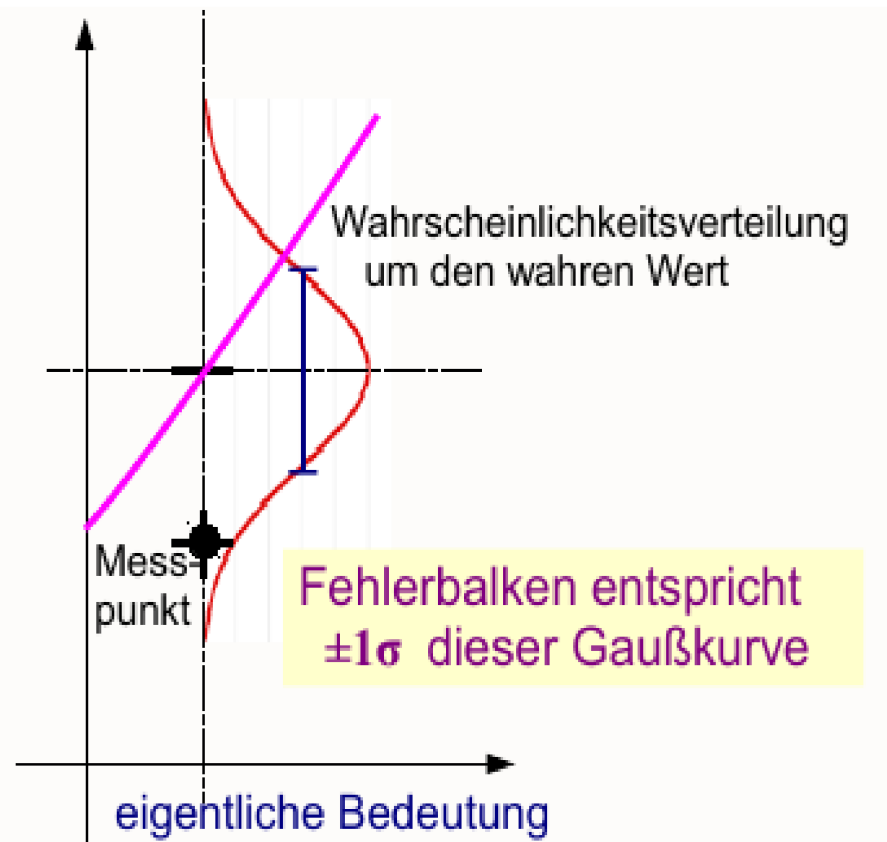
Interpretation von Fehlerbalken

■ Parameteranpassung



=

■ Vergleich Messung-Model



Messung mit Unsicherheiten: Beobachtung eines Messergebnisses, das einer Zufallszahl aus einer Verteilung um den wahren Wert entspricht.

Kleinste quadratische Abweichungen

„Least Squares oder auch χ^2 “

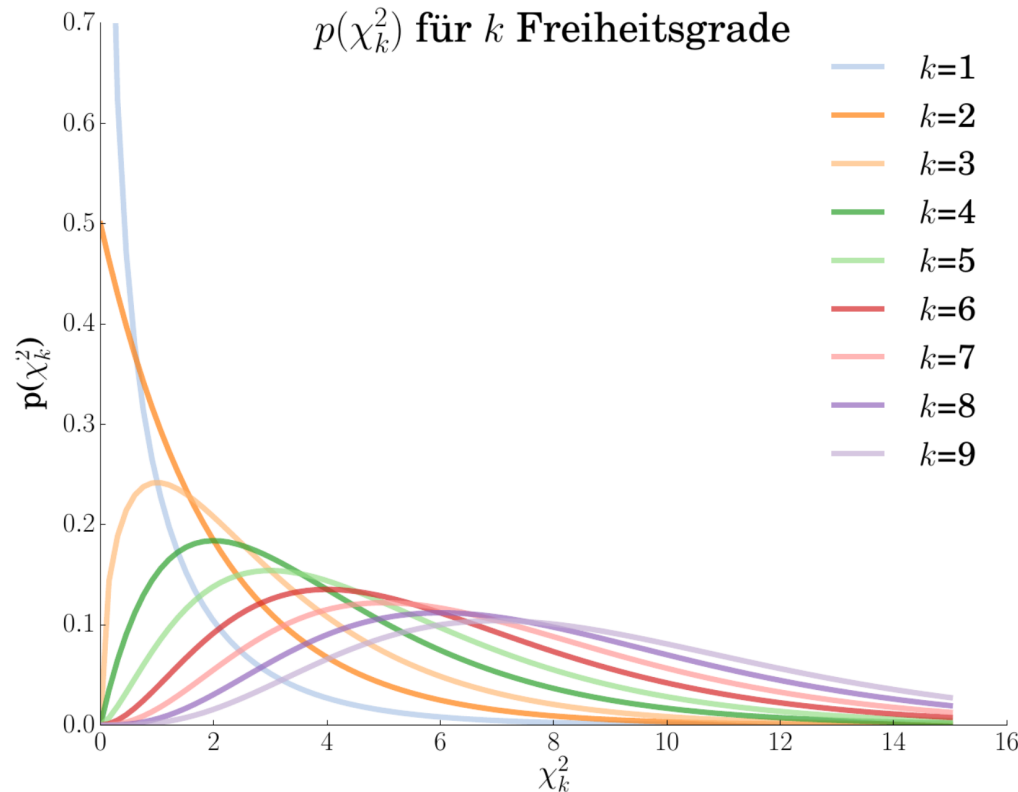
Methode der kleinsten quadratischen Abweichungen, χ^2 (~Wiederholung)

- Die quadratische Abweichung einer beliebigen Funktion $f(x|\vec{a})$ von einem gemessenen Satz von $i=1, \dots, N$ Datenpunkten x_i, y_i mit Gaußförmigen Unsicherheiten σ_i

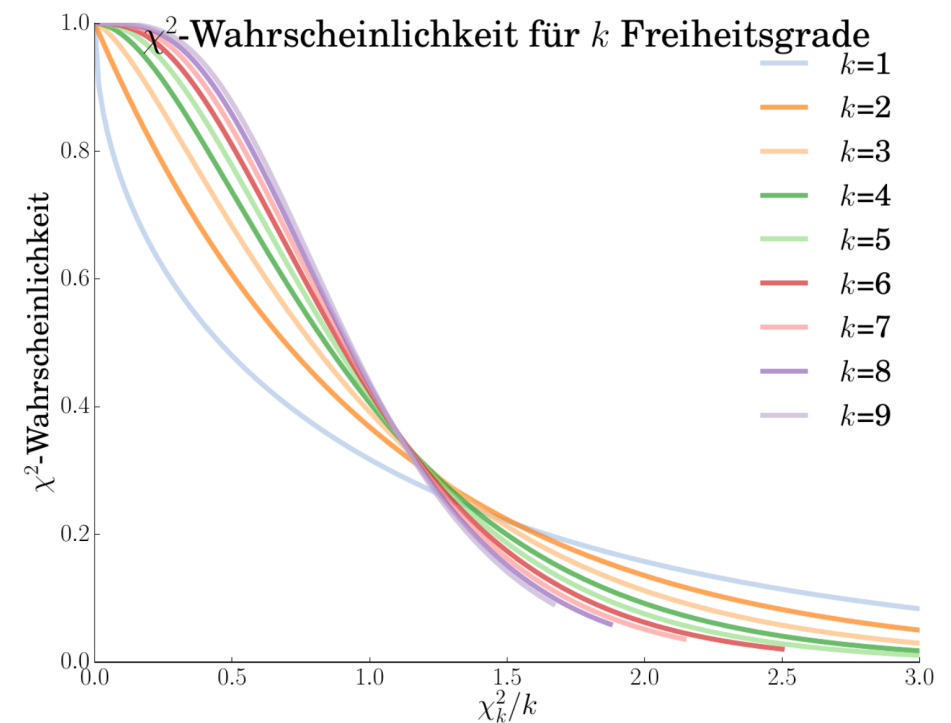
$$\chi^2 = \sum_i \frac{(f(x_i|\vec{a}) - y_i)^2}{\sigma_i^2}$$

- Jeder Punkt trägt im Mittel den Wert 1 zum χ^2 bei
- Aber: Nach Parameteranpassung wird χ^2 kleiner ($\rightarrow$ Minimierung!)
- Für jeden angepassten Parameter sinkt effektiv die Anzahl der Freiheitsgrade um 1

Zusammenhang mit χ^2 und χ^2 - Wahrscheinlichkeit



$$P = \int_{\chi_k^2}^{\infty} p(\chi_k'^2) d\chi_k'^2$$



■ Erwartungswert

$$E[\chi_k^2] = k$$

■ Varianz

$$V[\chi_k^2] = 2k$$

■ Limit für große k : Gaußverteilung mit $\mu = k$ und $\sigma = \sqrt{2k}$

Wiederholung: Lineares χ^2 Problem

- Für ein lineares Problem $y_i = \sum_j f_j(x_i) a_j = A_{ij} a_j$ oder $\vec{y} = A\vec{a}$

mit Kovarianzmatrix $\mathbf{V}$ der y_i

- Finde Minimum von:
bzgl. $\vec{a}$, also

$$\chi^2(\vec{a}) = (\vec{y} - A\vec{a})^T W (\vec{y} - A\vec{a})$$

$$\frac{\partial \chi^2}{\partial a_i} = 0$$

*Definition des
Schätzers*

- Mit der Gewichtsmatrix:

$$W = V^{-1}$$

- Optimale Parameter:

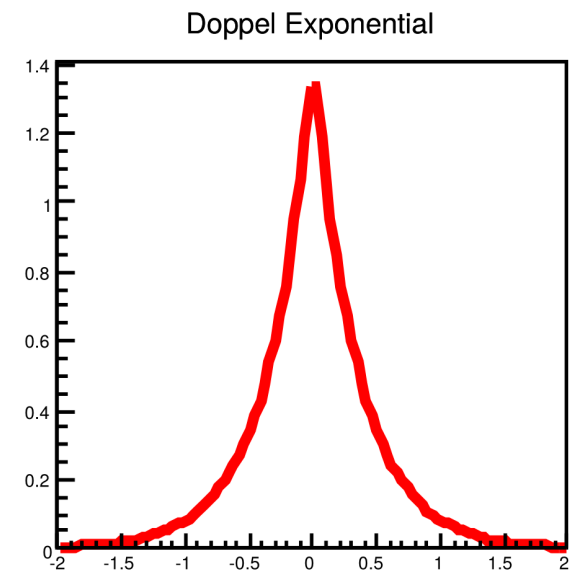
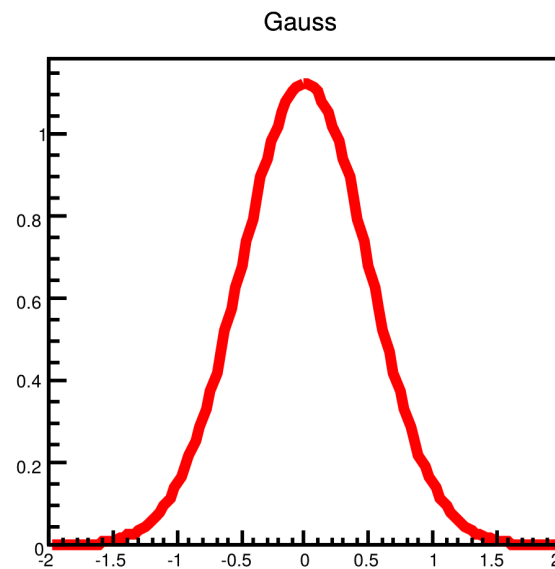
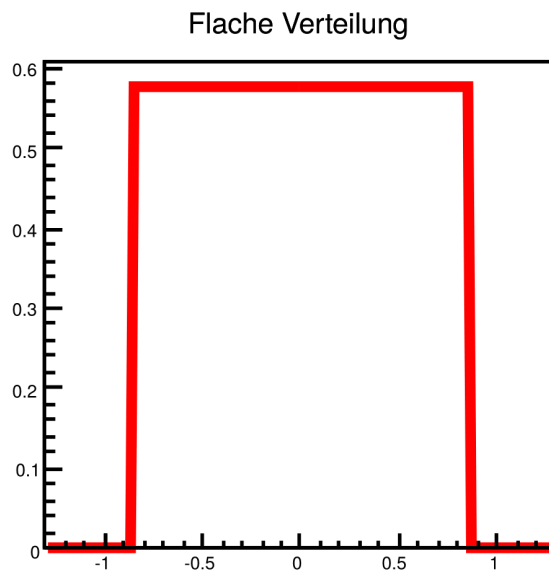
$$\hat{\vec{a}} = (A^T W A)^{-1} A^T W \vec{y}$$

- Unsicherheiten

$$V(\hat{a}) = (A^T W A)^{-1}$$

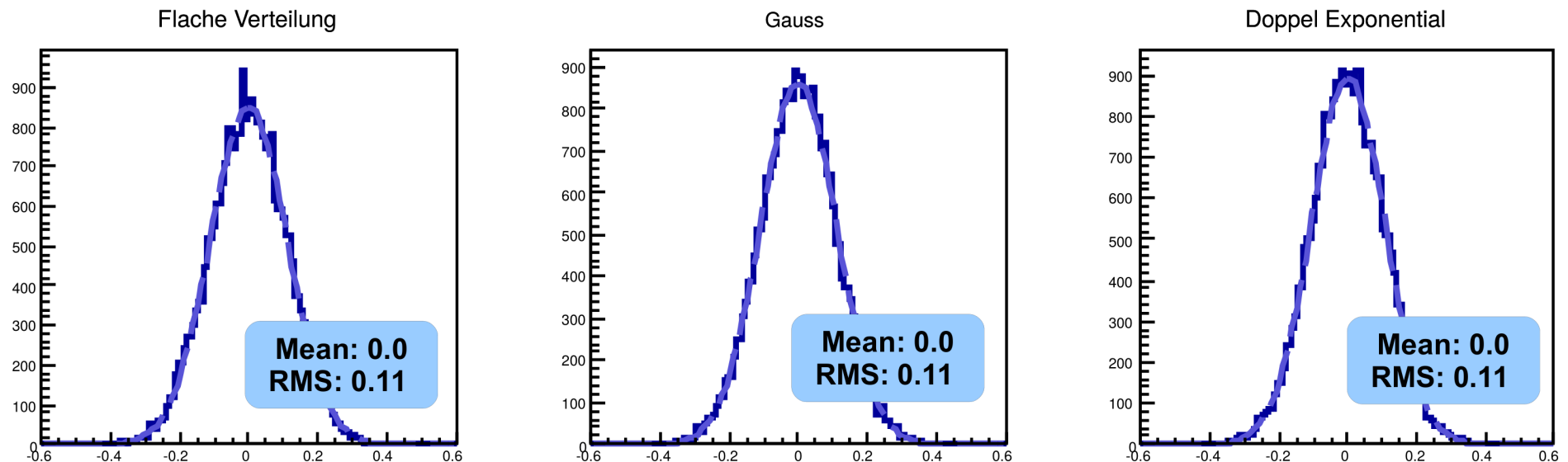
Abhängigkeit von der Messfehler-Verteilung

- Anpassen einer Geraden $y = a + bx$ an 20 Datenpunkte
- n.d.f = $20 - 2 = 18$
- Betrachte drei verschiedene Verteilungen für die Unsicherheiten der Einzelmessungen. Alle mit Mittelwert 0 und Standardabweichung 0.5.



Einfluss auf die Parameterschätzung

- 10000 Monte Carlo Realisationen
- Alle Parameter Verteilungen sind gaußisch, die Breite kompatibel zur Erwartung aus Fehlerfortpflanzung (für beide Parameter)
- Hier die Verteilungen (für Parameter a) : $\hat{a} - a_0$



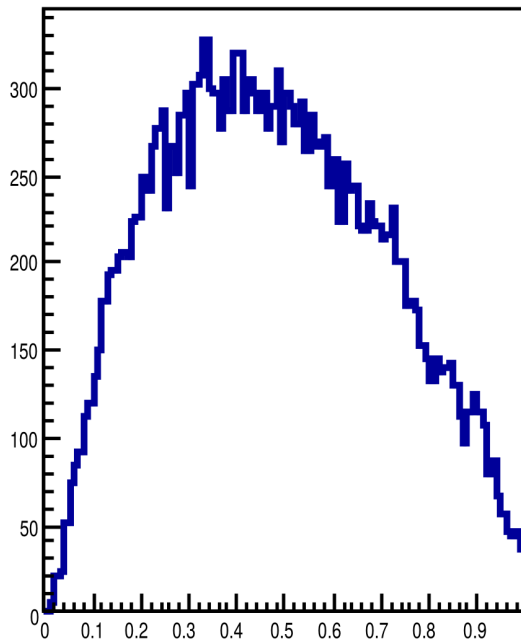
Methode der kleinsten Quadrate funktioniert wenn Mittelwerte und Standardabweichungen der Daten bekannt, sowie die Fehler symmetrisch sind

Effekt auf die χ^2 -Wahrscheinlichkeit

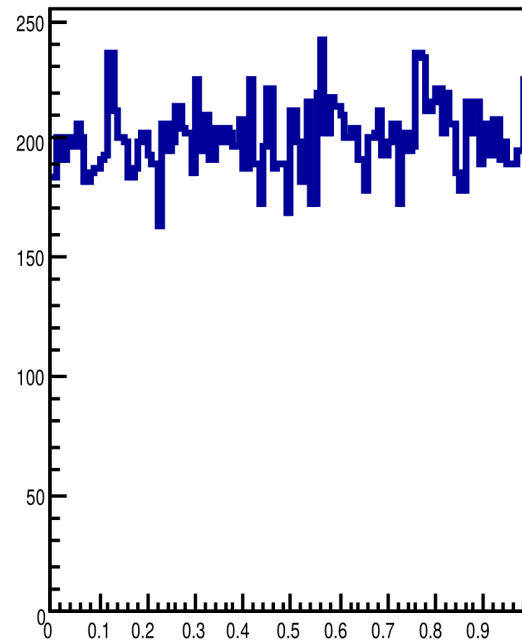
- Berechne jeweils die χ^2 -Wahrscheinlichkeit

$$Prob(\chi^2) = \int_{\chi^2}^{\infty} P_{\chi^2}(s) ds$$

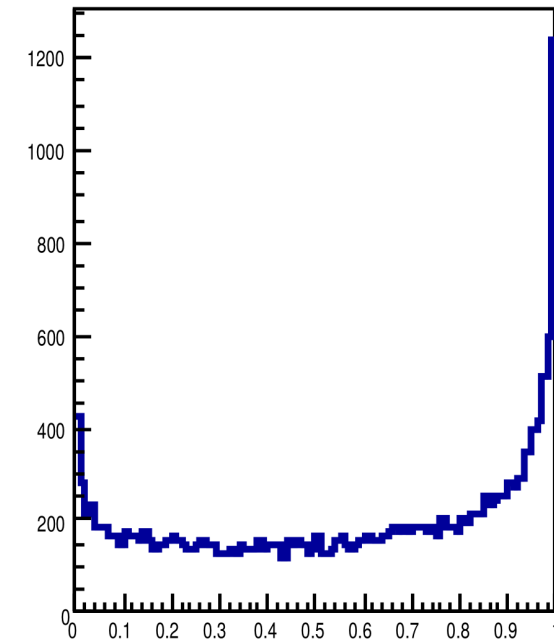
Flache Verteilung



Gauss



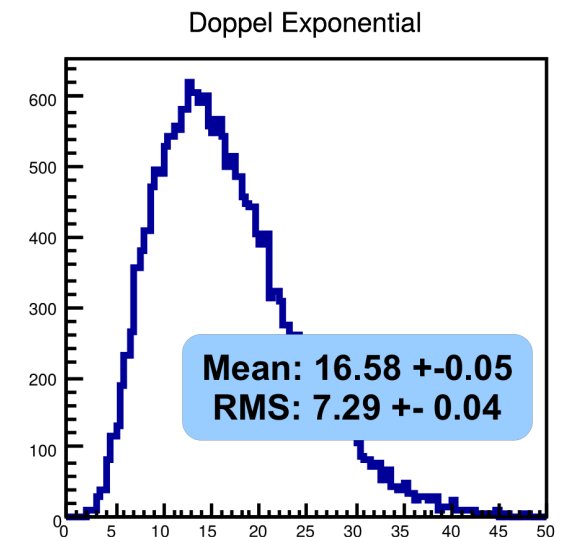
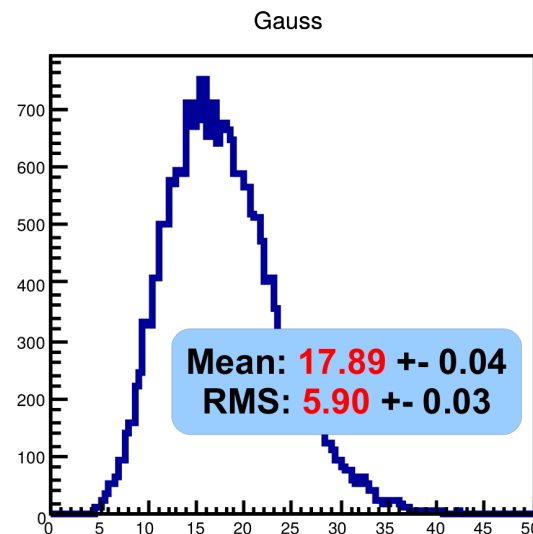
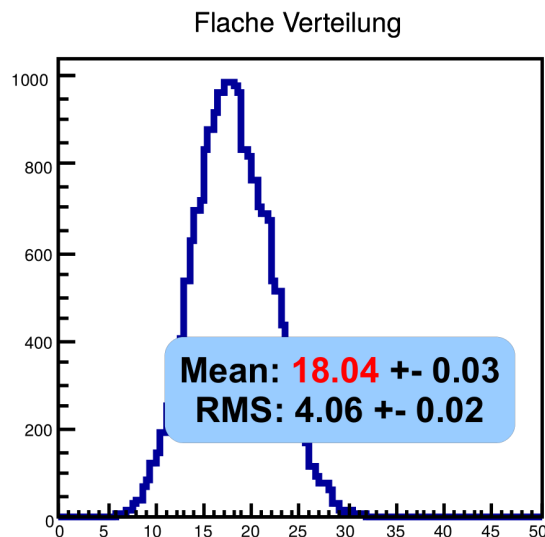
Doppel Exponential



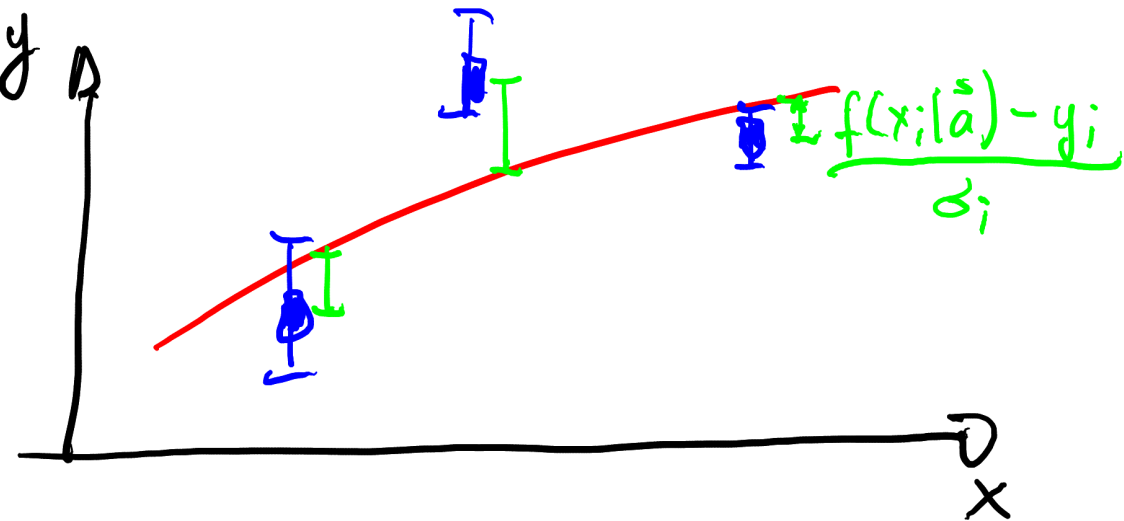
Bei gaußförmigen Fehlern gilt $Prob(\chi^2) = \text{const}$, folgt also einer χ^2 -Verteilung.
Ansonsten ist dies nicht der Fall !

Effekt auf die χ^2 -Verteilung

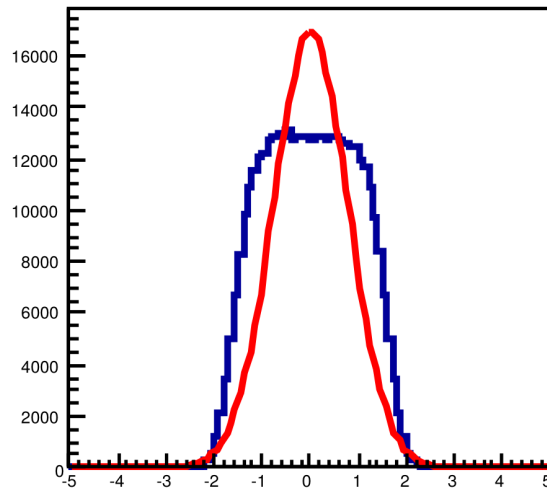
- $E[\chi^2] = \text{Ndf} = 20-2 = 18$
- $V[\chi^2] = 2 \text{ Ndf} = 36$, d.h. RMS = 6



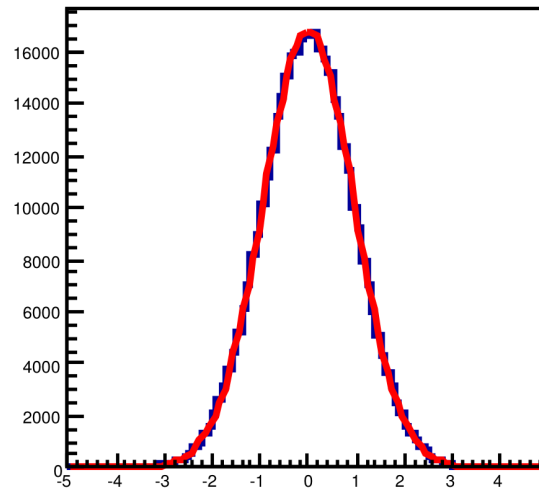
„Pull“ Verteilung



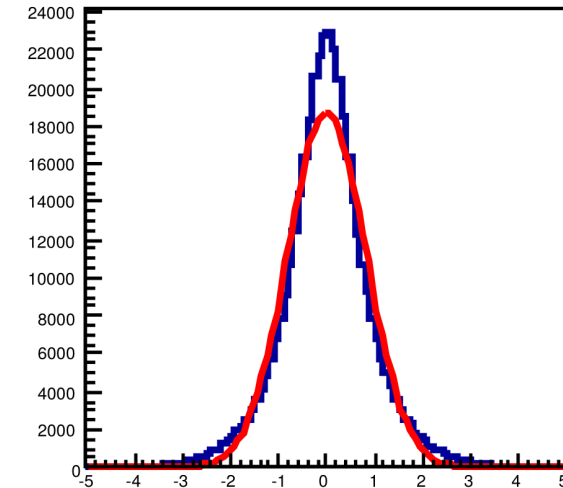
Flache Verteilung



Gauss



Doppel Exponential



- Erwarte Gaußverteilung mit $\mu=0$ und $\sigma=1$
- Jeder Datenpunkt trägt im Mittel den Wert 1 zum X^2 bei

χ^2 Statistik

■ $\chi^2 \ll 1 \cdot k$

Abweichung von der Statistik.

- Fehler zu groß ?
- Zu viele Parameter ?

■ $\chi^2 \sim 1 \cdot k$

Dies ist die Erwartung aus der Statistik.

■ $\chi^2 \gg 1 \cdot k$

Abweichung von der Statistik.

- Fehler zu klein ?
- Zu wenige Parameter ?

χ^2 Verteilung kann verwendet werden, um χ^2 Werte zu interpretieren und Konfidenzbereiche zu quantifizieren

Vergleich χ^2 und Likelihood

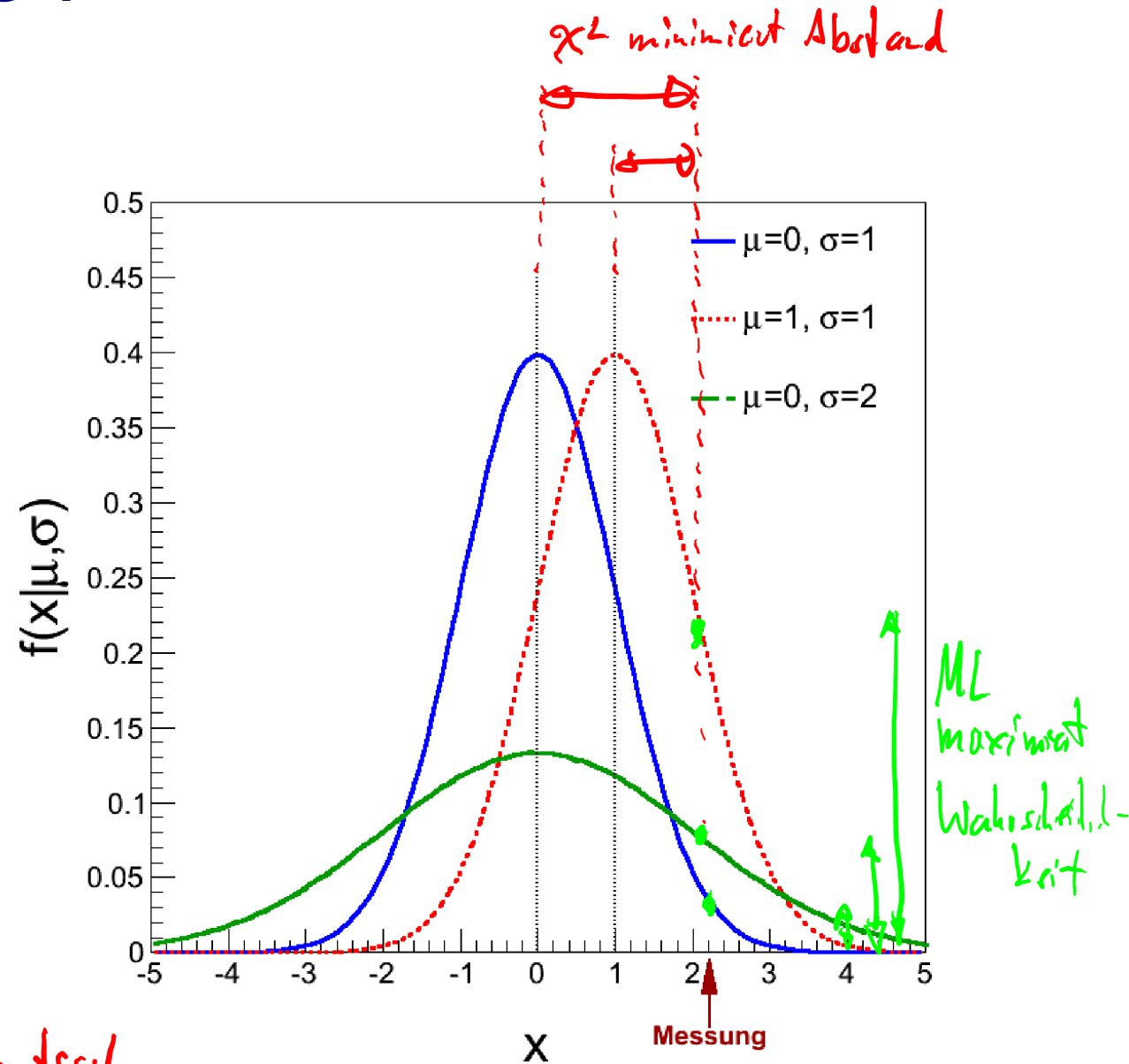
Welche Verteilung passt am besten zur Messung

In diesem Beispiel bevorzugen beide Methoden die gleiche Verteilung

Das muss nicht so sein!

Die kleinsten Quadrate unterscheiden nicht zwischen blauen und grünen Verteilung

⇒ Die Methoden nicht identisch



Zusammenhang $-\log L$ und χ^2

Für Gauß-förmig verteilte Messungen y_i ist die

χ^2 -Methode äquivalent zur $-\log L$ -Methode

$$P(y_i | \vec{a}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[- \frac{(y_i - f(x_i | \vec{a}))^2}{2\sigma_i^2} \right]$$

$$-\log L(y_i | \vec{a}) = \underbrace{\frac{1}{2} \left(\frac{y_i - f(x_i | \vec{a})}{\sigma_i} \right)^2}_{\chi_i^2} + \underbrace{\sum_i \log(\sqrt{2\pi}\sigma_i)}_{\text{const bzgl. } \vec{a}}$$

Minimieren von $-\log L$

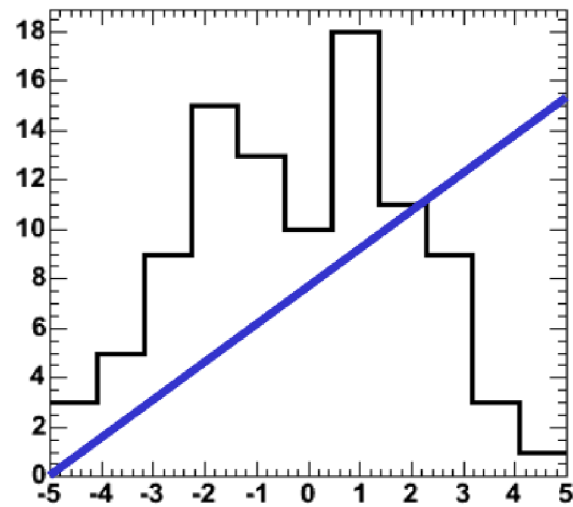
$\Leftrightarrow$ Minimieren von χ^2

$$\Delta(-\log L) = \frac{1}{2} (\Delta \chi^2)$$

$$\frac{\partial^2(-\log L)}{\partial a_i \partial a_j} = \frac{1}{2} \frac{\partial^2(\chi^2)}{\partial a_i \partial a_j}$$

- Bei nicht Gauß-förmigen Fehlern $\chi^2 \neq ML$
- Bei unbekannten Fehlern ist χ^2 optimal
- χ^2 linearisieren $\Rightarrow$ analytische Lösung

„Goodness of fit“



‘Not a good fit’

Pearson Test, z.B. χ^2 nach ML

■ Pearson-Test Statistik:

$$\chi_M^2 = \sum_{i=1}^N \frac{(f(x_i) - y_i)^2}{f(x_i)}$$

(hier typisches Beispiel für Poisson-Verteilung mit Varianz $V[x] = f(x_i)$)

- Bin-Größe eventuell dynamisch wählen. y_i sollte >5 , besser >10 sein
- Große Bins führen zu Verlust von Information
- Allgemein anwendbar, in vielen Situationen mit guten Ergebnissen

Likelihood Ratio

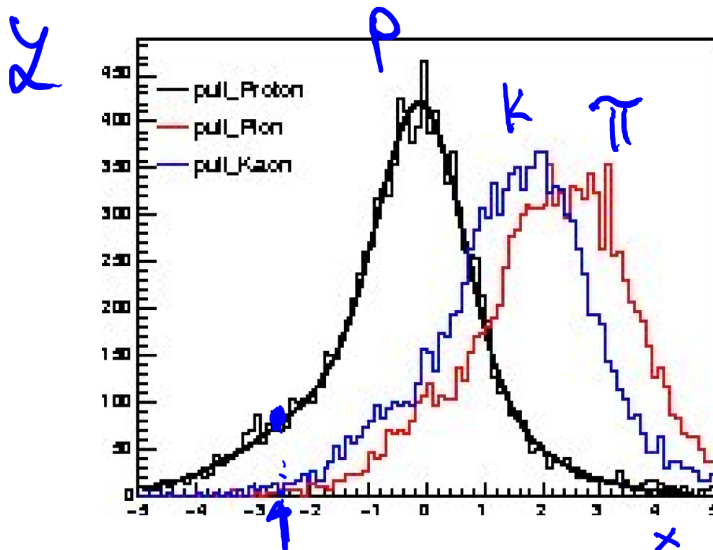
- Der absolute Wert einer Likelihood hat keine Bedeutung
- Aber Likelihood-Ratio (für gleiche N) ist die beste Möglichkeit verschiedene Modelle miteinander zu vergleichen

$$\Lambda = \frac{\mathcal{L}_1(a|\vec{x})}{\mathcal{L}_2(a|\vec{x})}$$

Neyman and Pearson

On the Problem of the Most Efficient Tests of Statistical Hypotheses

Phil. Trans. R. Soc. Lond. A 1 January 1933 vol. 231 no. 694-706 289-337



- Beobachtetes Teilchen muss Kaon, Pion oder Proton sein → welche Hypothese ist die wahrscheinlichste
- Likelihood-Ratio ist gegeben durch $L(\text{data}|\text{proton})/L(\text{data}|\text{proton oder pion oder kaon})$
- Vorsicht: die Prior-Wahrscheinlichkeiten sind sehr wichtig
- Auch die „Tails“ sind sehr wichtig!

Details der Anpassungs-Güte

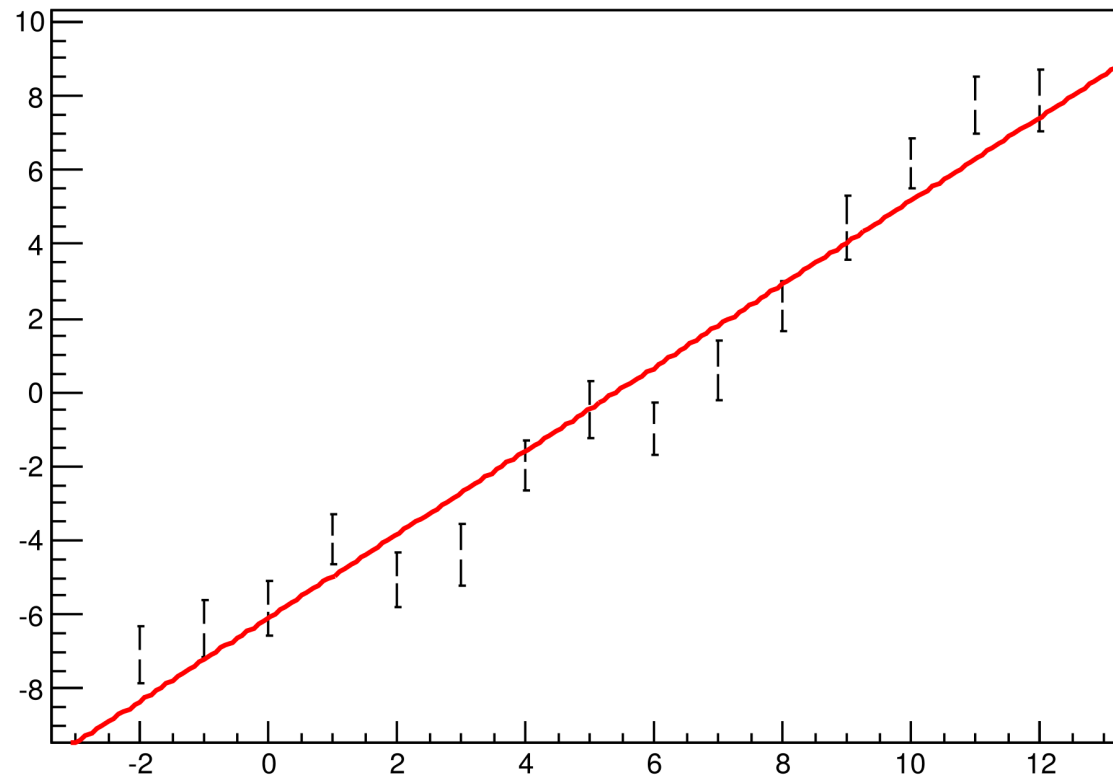
- Likelihood und χ^2 Methoden quantifizieren nur globale Übereinstimmung oder nicht-Übereinstimmung

aber

- Auch weitere Details sind wichtig: Vorzeichen, Sequenzen, Krümmung, etc.

Ist dieses Modell „gut“ ?

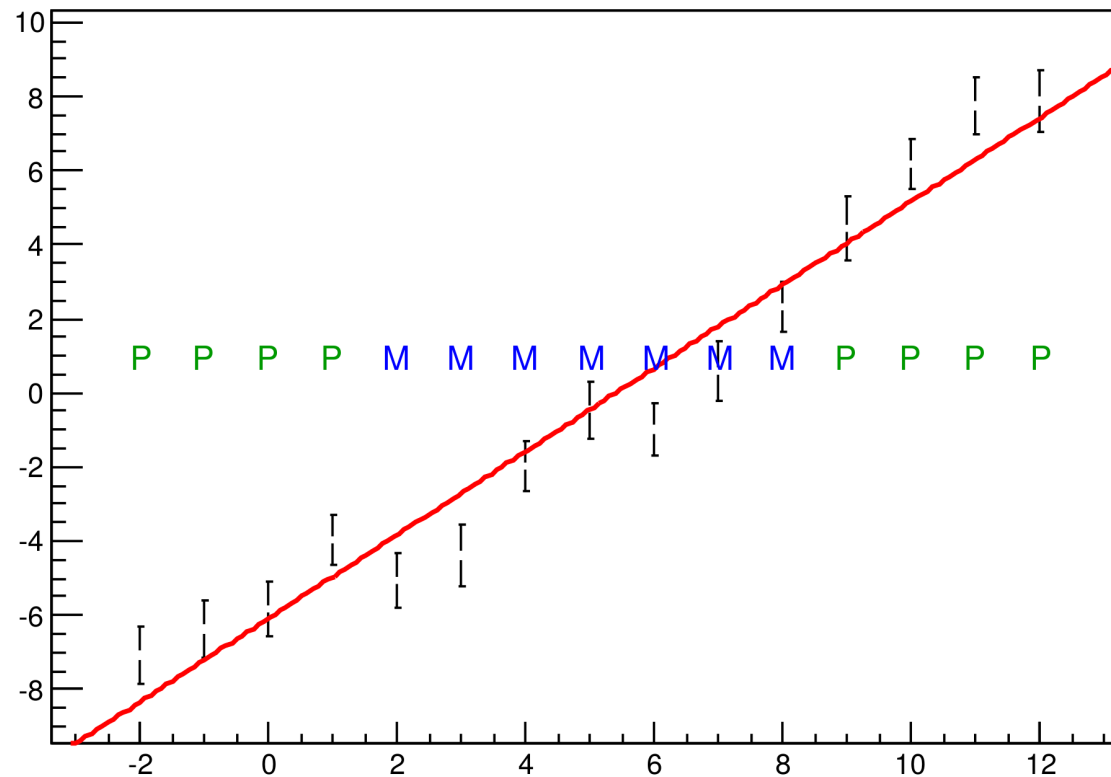
$$\chi^2/\text{ndf} = 27.7154/13 = 2.13196$$



Run Test

Untersucht die Verteilung von „runs“, in welchen die Messungen oberhalb bzw. unterhalb des Modells liegen

$$\chi^2/\text{ndf} = 27.7154/13 = 2.13196$$



N_A , N_B : Anzahl von Bins ober- bzw unterhalb des Modells, und $N_A + N_B = N$

Run Test Statistik

siehe z.B. Cowan 8.3.2

- Die erwartete Häufigkeit von Runs **r** kann aus Kombinatorik errechnet werden

- Es kann gezeigt werden, dass

- $$E[r] = \langle r \rangle = 1 + \frac{2N_A N_B}{N}$$

- $$V[r] = \frac{2N_A N_B (2N_A N_B - N)}{N^2 (N - 1)}$$

- Für $N > 10$ bis 15, kann $E[r]$ und $V[r]$ zur Approximation einer Gauß-verteilung verwendet werden, $\mu = \langle r \rangle$ $\sigma = \sqrt{V[r]}$

- Ansonsten: Nachschlagen in Tabellen

Anwendung auf Beispiel

■ $N_A = 8, N_B = 7, N = 15 \rightarrow$ Gauß Näherung mit $\mu=8.5$ und $\sigma=1.8$

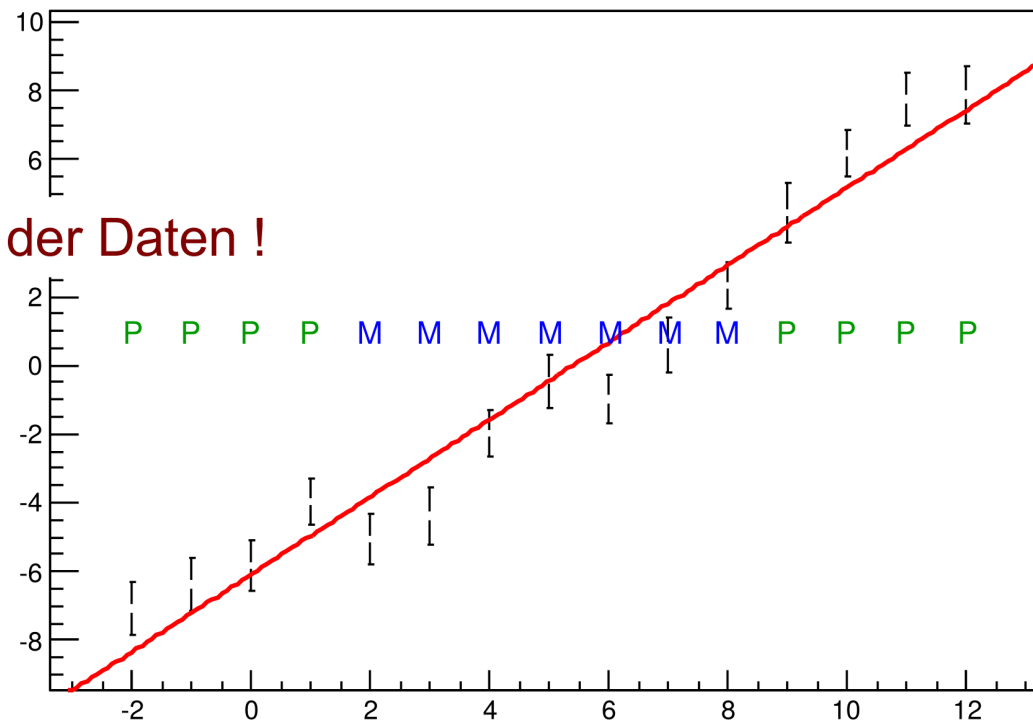
■ $r = 3$, d.h. $(r-\mu)/\sigma = -3.05$

■ zweiseitiger Test:
$$p = 2 \int_{-\infty}^{(r-\mu)/\sigma} \text{Gauss}(x) dx = 0.0022$$

■ $p < 1\%$ für $r=3$
oder schlechter

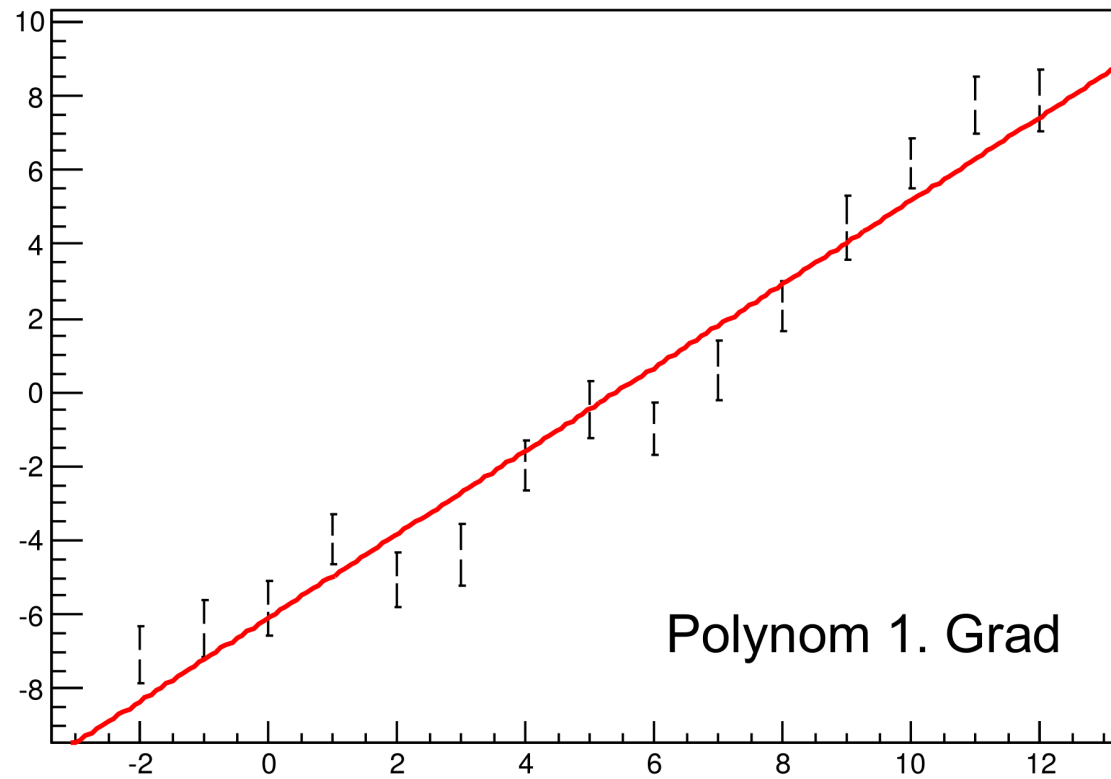
$$\chi^2/\text{ndf} = 27.7154/13 = 2.13196$$

→ schlechte Beschreibung der Daten !



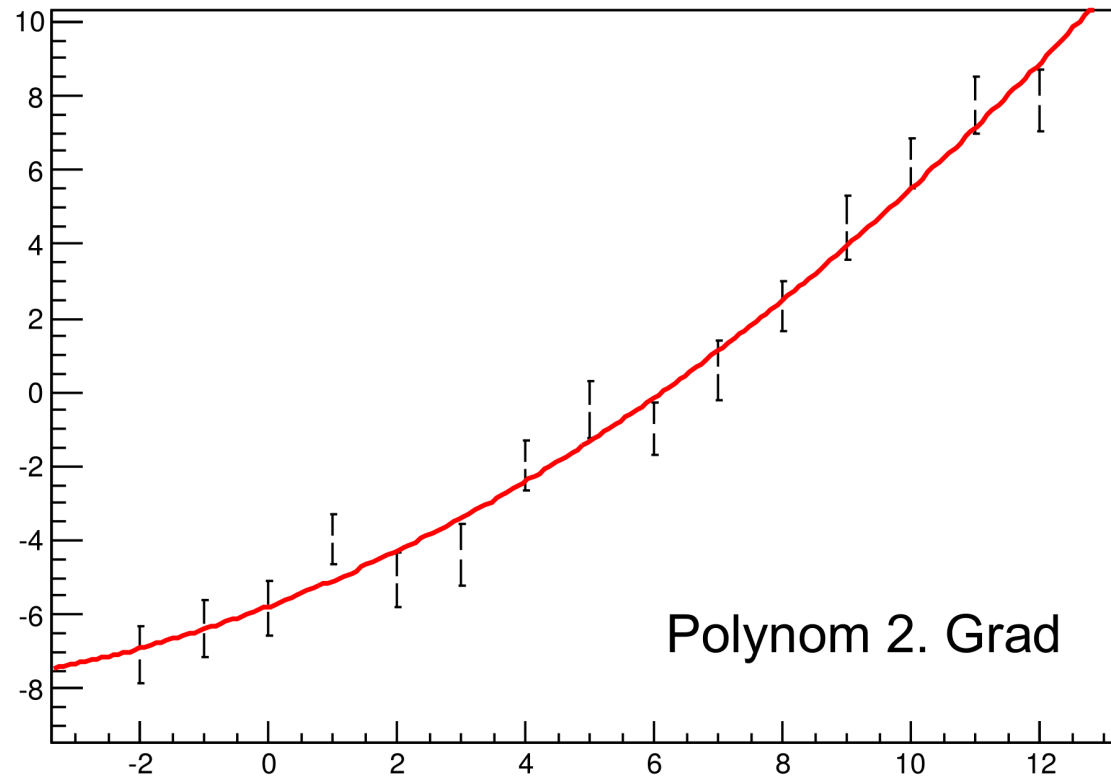
Overfitting

$$\chi^2/\text{ndf} = 27.7154/13 = 2.13196$$



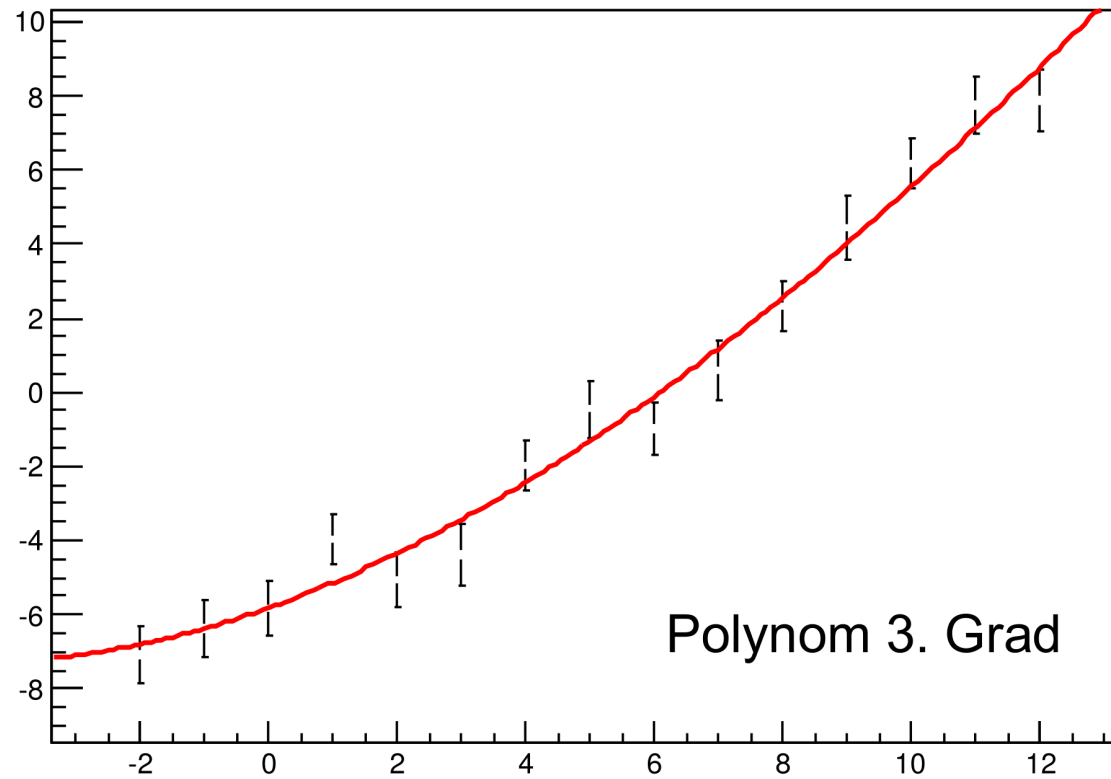
Overfitting

$$\chi^2/\text{ndf} = 12.0749/12 = 1.00624$$

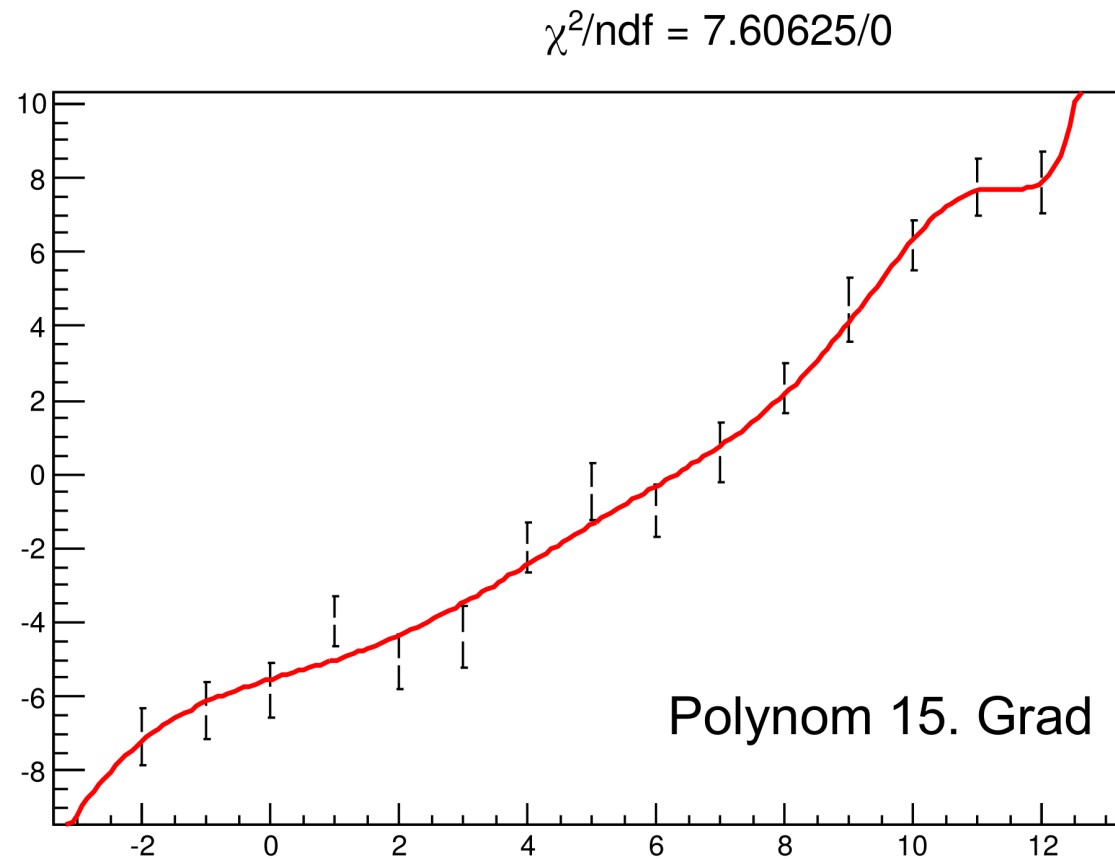


Overfitting

$$\chi^2/\text{ndf} = 11.9862/11 = 1.08966$$



Overfitting



Indiz: χ^2 geht gegen Null, Freiheitsgrade gehen auch gegen Null....

Kolmogorov-Smirnov (KS) Test (→ Form)

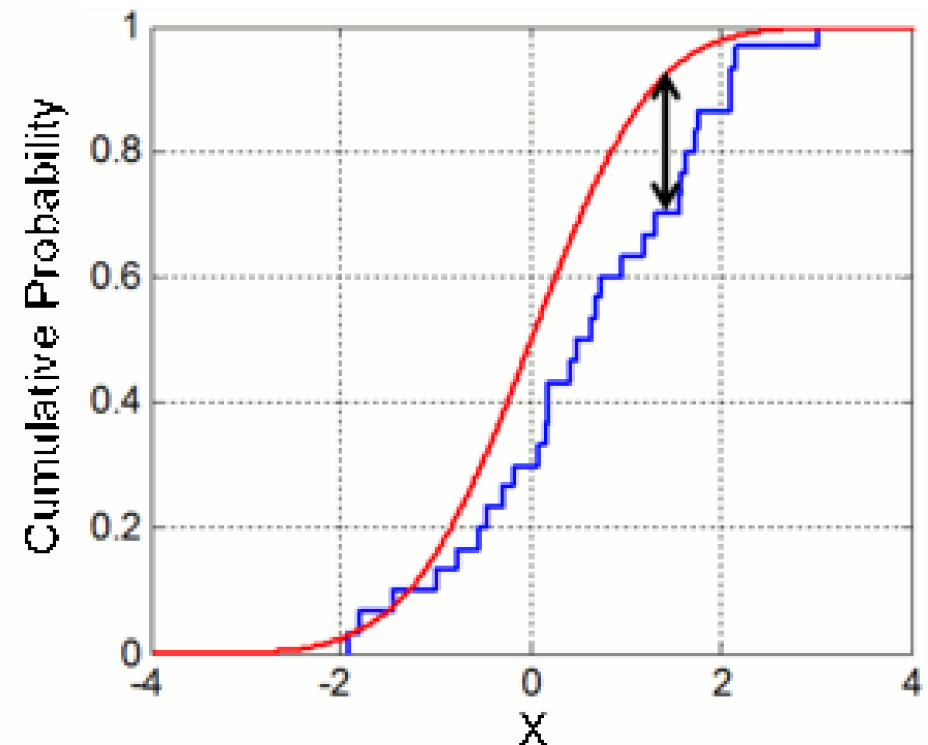
- Vergleiche die kumulative Verteilungsfunktionen zweier Verteilungen oder zwischen Verteilung und Modell

- Kein Binning !

- Messen der größten vertikalen Abweichung

$$D = \max |\text{cum}(x) - \text{cum}(P)|$$

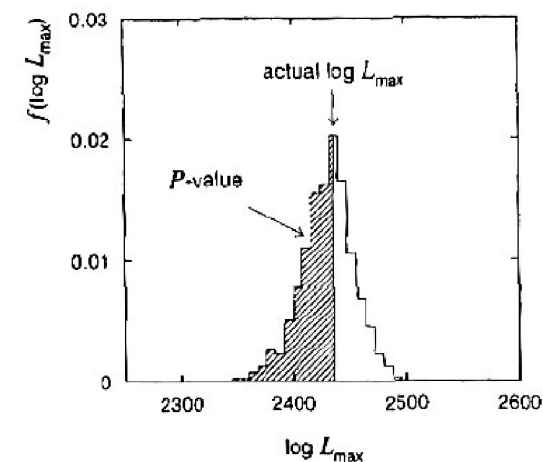
- Testgröße $d = D\sqrt{N}$



- Vergleiche mit tabellarisierten kritischen Werten von d !

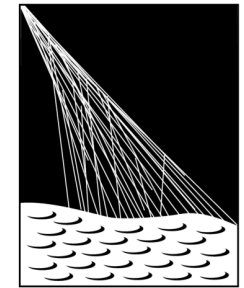
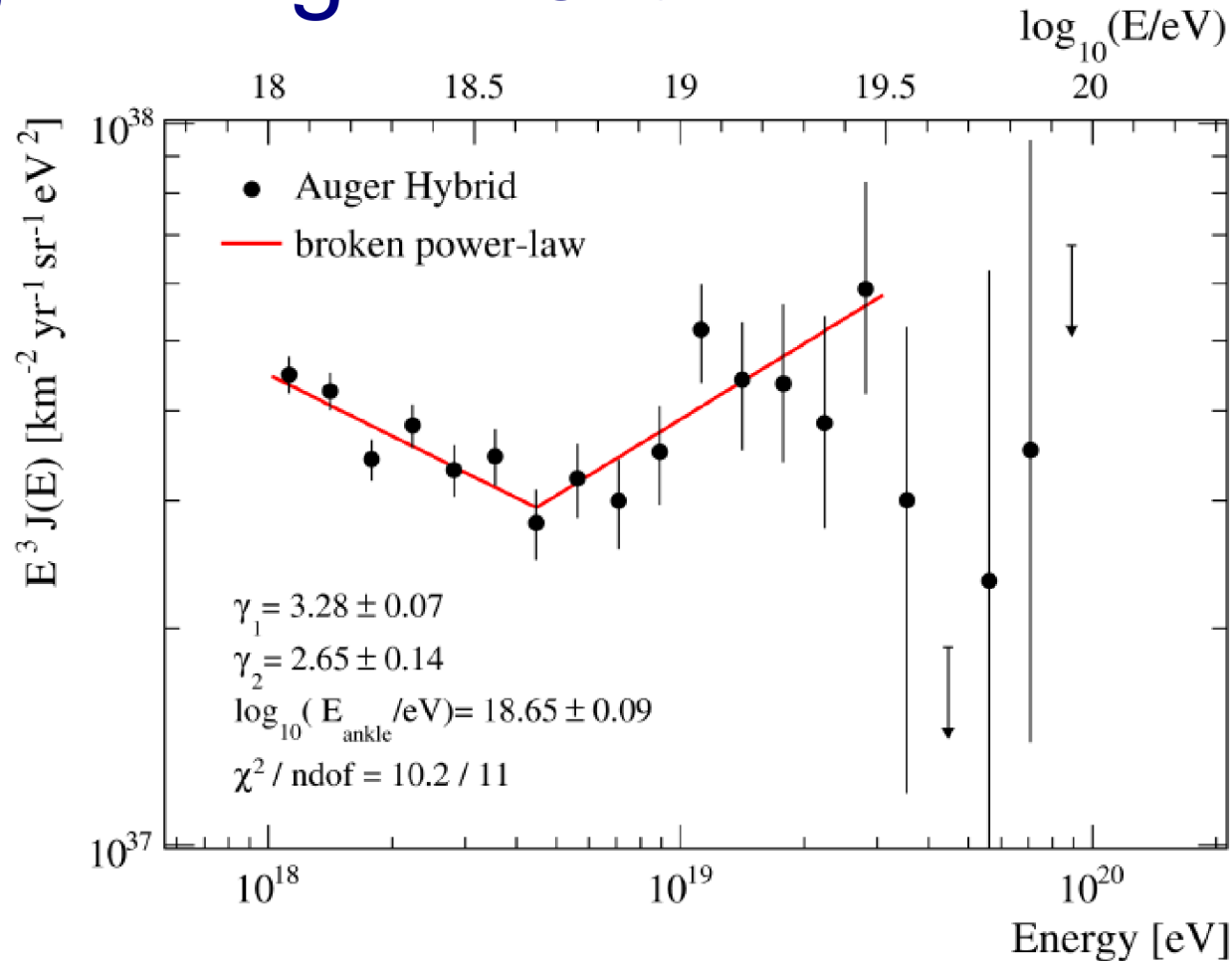
Anmerkungen, allgemeiner Ansatz

- Gerade in mehr-dimensionalen Situationen ist eine „Goodness of fit“ sehr komplex. Es gibt keine universelle Methode/Lösung. Hier wird auch immer noch geforscht.
- Heutzutage ist Rechenleistung oft ausreichend vorhanden, um Toy-MC Studien zu machen, und so p-Werte, Bias, Effizienz etc. abzuschätzen:
 - 1. Verwende „best-fit“ Parameter und deren Unsicherheiten
 - 2. Erzeuge Monte Carlo Daten basierend darauf
 - 3. Berechne Likelihood der Toy Daten
 - 4. Wiederhole Punkte 2+3 sooft es technisch möglich ist
 - 5. Erzeuge Histogramm der so erzeugten Toy-Likelihood und vergleiche die echte Messung mit dem Histogramm
 - 6. Berechne p-Wert der echten Likelihood



Der Einfluss des Histogrammierens auf die Parameterschätzung

Der „Binning“ Effekt

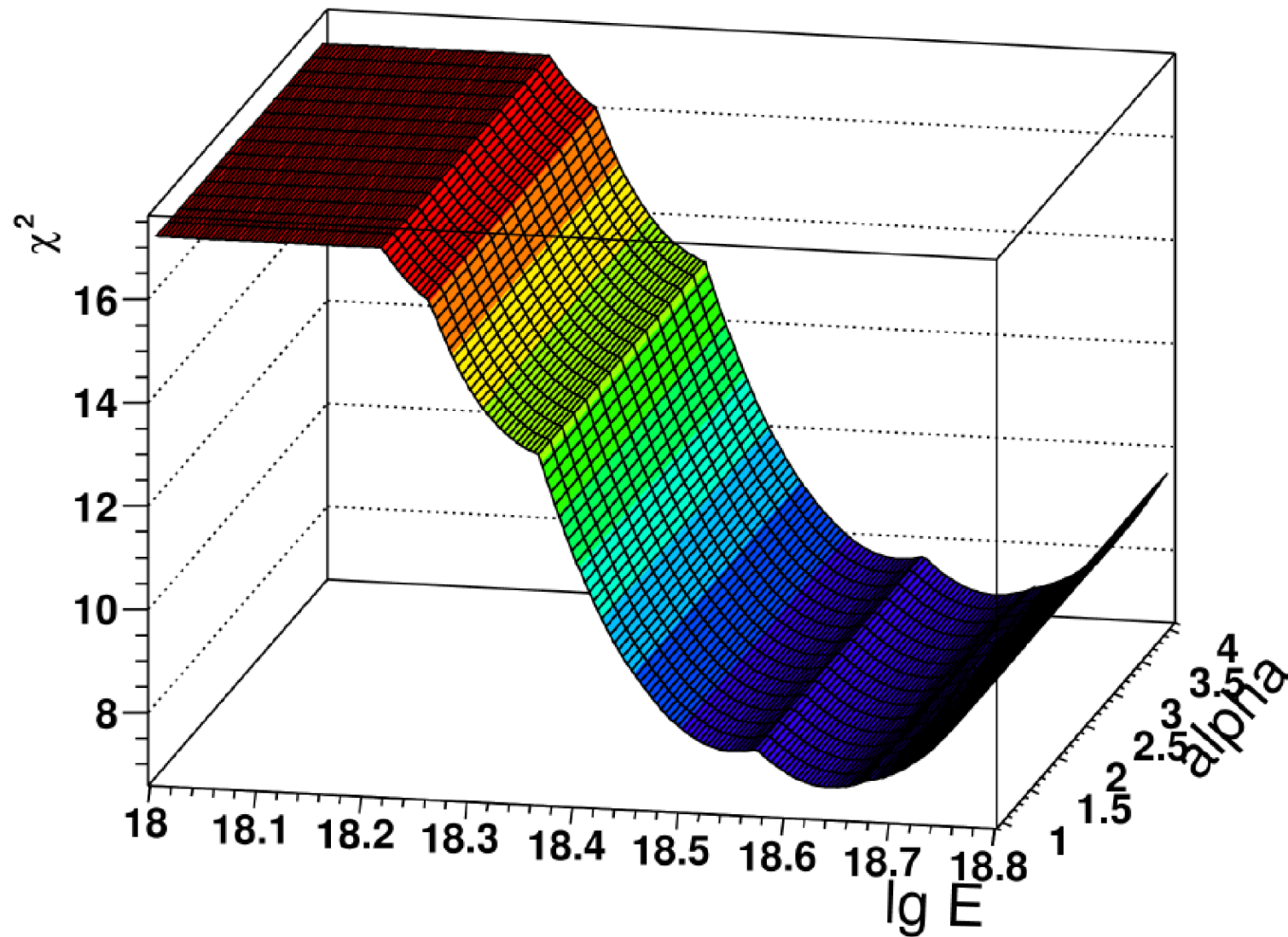


PIERRE
 AUGER
 OBSERVATORY

- Energieverteilung der kosmischen Strahlung und der „Knöchel“ bei $10^{18.6}$ eV (*Phys.Lett. B685 (2010) 239-246*)

- Generisches Problem: **Anpassen von Modell Parameter an Daten**³³

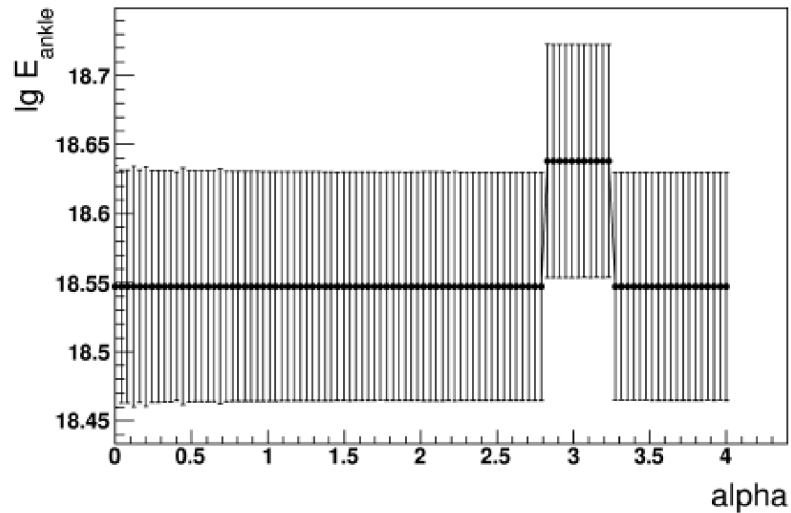
Die χ^2 Landschaft



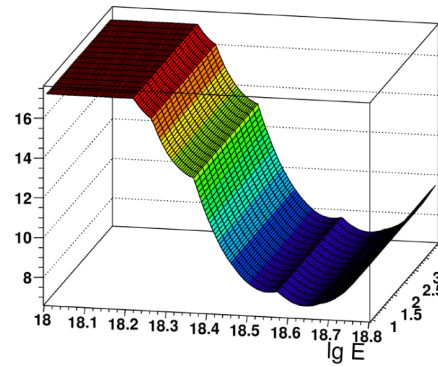
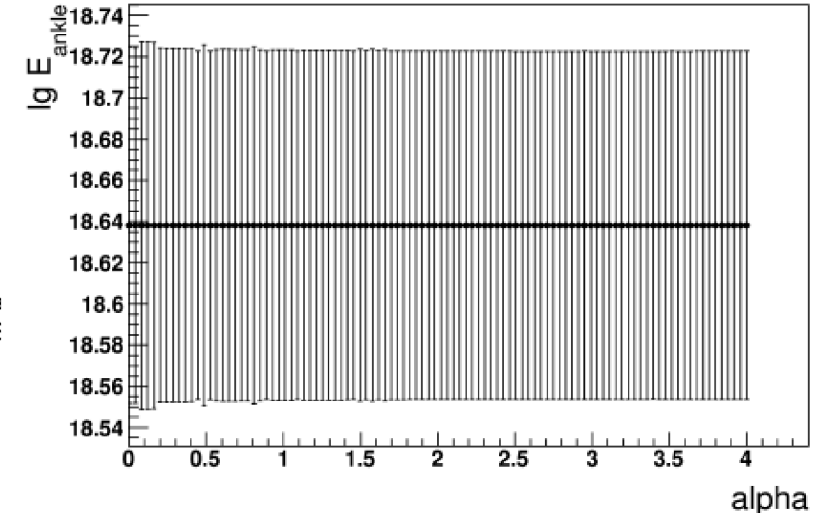
- Aus dem Scan von zwei Modellparametern berechnet

Form und Richtung sind Wichtig

Startwert „Links vom Minimum“



Startwert „Rechts vom Minimum“

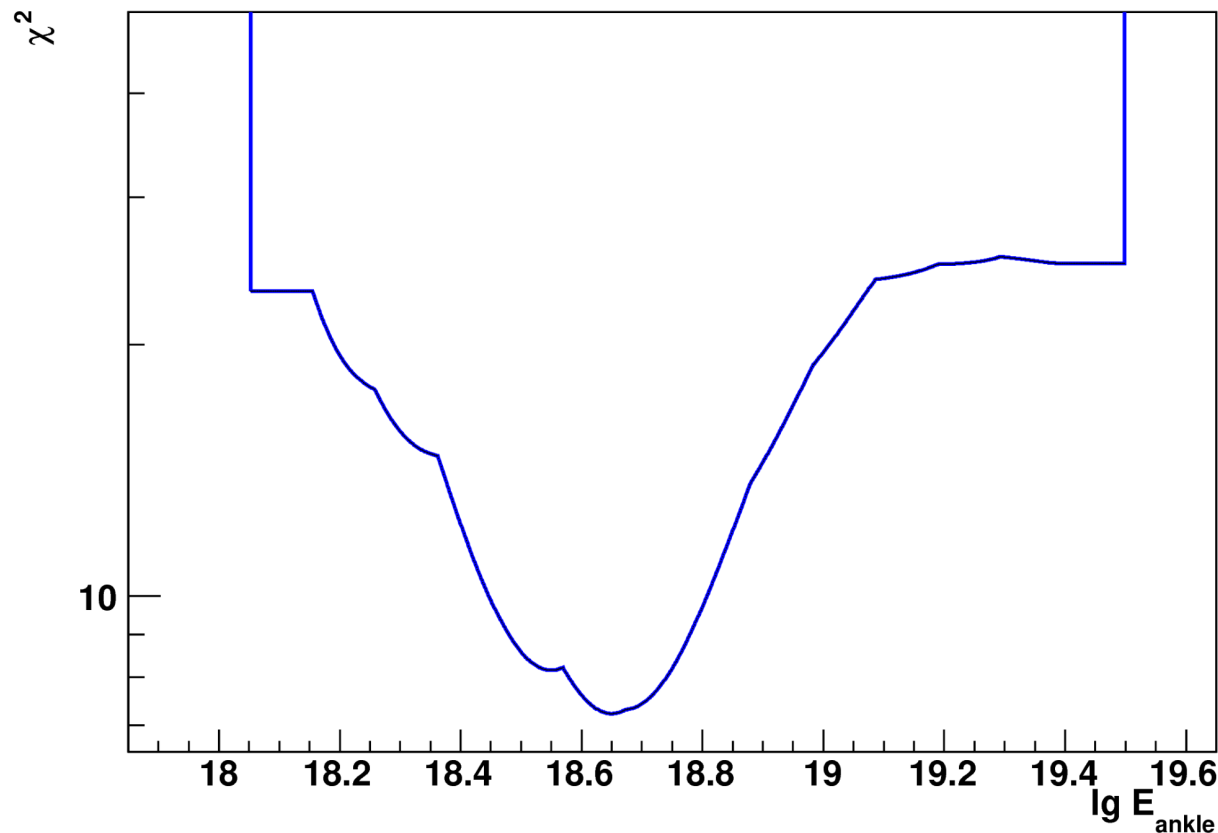


Analytischer χ^2 Scan

linearisiertes Problem:

$$\begin{pmatrix} \lg I_1 \\ \lg I_2 \\ \vdots \\ \lg I_m \\ \lg I_{m+1} \\ \vdots \\ \lg I_n \end{pmatrix} = \begin{pmatrix} 1 & \lg E_1 - \lg E_{\text{ankle}} & 0 \\ 1 & \vdots & \vdots \\ 1 & \lg E_m - \lg E_{\text{ankle}} & 0 \\ 1 & 0 & \lg E_{m+1} - \lg E_{\text{ankle}} \\ 1 & \vdots & \vdots \\ 1 & 0 & \lg E_n - \lg E_{\text{ankle}} \end{pmatrix} \cdot \begin{pmatrix} \lg I_{\text{norm}} \\ s_1 \\ s_2 \end{pmatrix}$$

analytical χ^2 scan



Likelihood und Bayes Theorem

Likelihood und Bayes Theorem

- Likelihood ist keine Wahrscheinlichkeitsdichte für a sondern für x
- Mit Bayeschen Theorem kann man daraus eine Wahrscheinlichkeitsdichte für a konstruieren

$$f(a|\vec{x}) = \frac{\mathcal{L}(\vec{x}|a)f_{\text{prior}}(a)}{\int \mathcal{L}(\vec{x}|a')f_{\text{prior}}(a')da'}$$

- Für gleich-verteilten Prior $f_{\text{prior}} = \text{const}$

$$f(a|\vec{x}) = \frac{\mathcal{L}(a)}{\text{Normierung}}$$

- Einfluss auf Schätzer

$$\hat{a}_{\text{Bayes}} = \hat{a} + \sigma_{\hat{a}}^2 \frac{1}{f_{\text{prior}}} \frac{df_{\text{prior}}}{da}$$

Maximum Likelihood Binomial Fit mit verschiedenen a priori Annahmen

Erinnerung:

Wahrscheinlichkeit für die Beobachtung eines bestimmten Ereignisses ist p .
Wahrscheinlichkeit, bei n Versuchen k solcher Ereignisse zu beobachten
ist gegeben durch die **Binomialverteilung** [$P(k) \sim p^k (1-p)^{n-k}$]

Eine Münze wird N mal geworfen.

Was kann man über die Wahrscheinlichkeit p ,
dass „Kopf“ geworfen wird, sagen?

0 Versuche: Wir wissen nichts, p flach in $[0,1]$.

1. Versuch: Kopf $L_1(p) = p$, $p(\text{Zahl})=0$

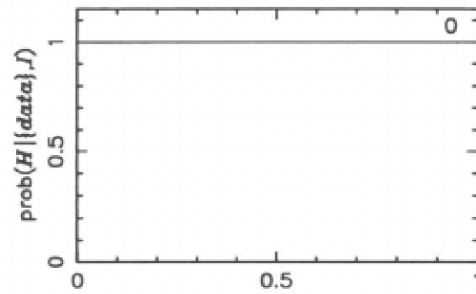
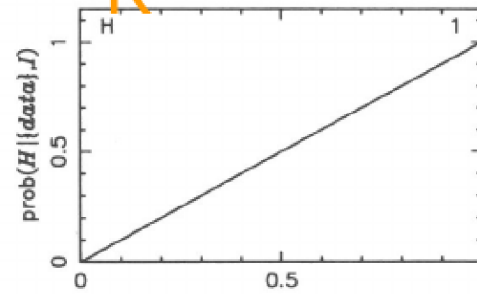
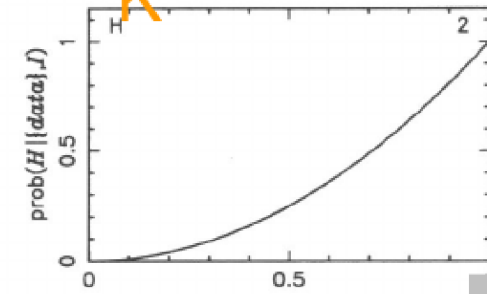
2. Versuch: Kopf $L_2(p) = L_1 p = p^2$

3. Versuch: Zahl $L_3(p) = L_2 (1-p) = p^2 (1-p) = p^3 - p^2$

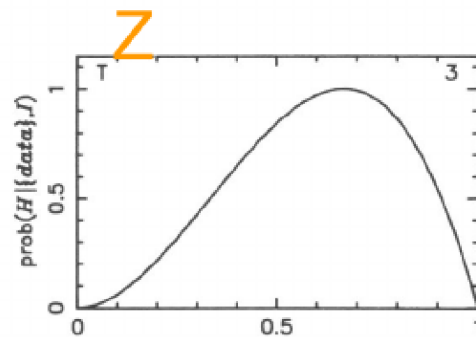
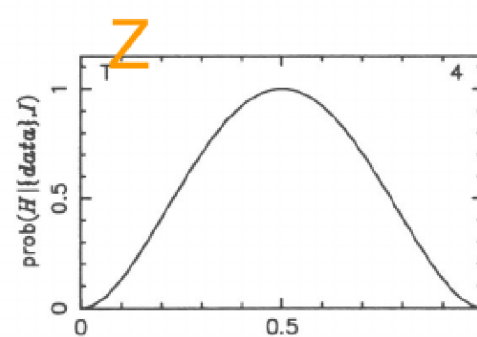
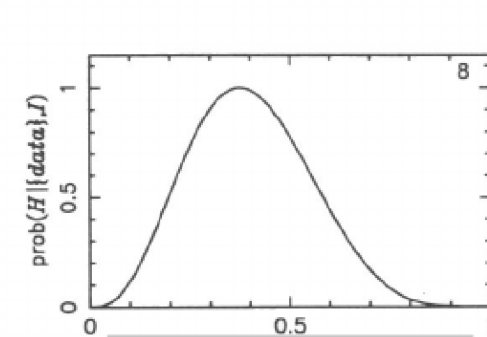
3. Versuch: Zahl $L_4(p) = L_3 (1-p) = p^2 (1-p)^2 = p^4 - 2p^3 + p^2$

...

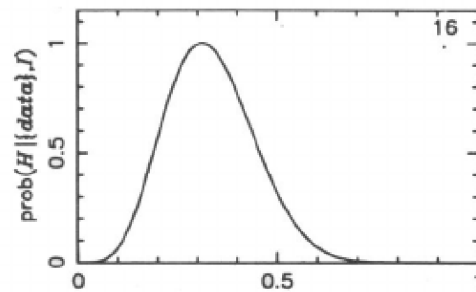
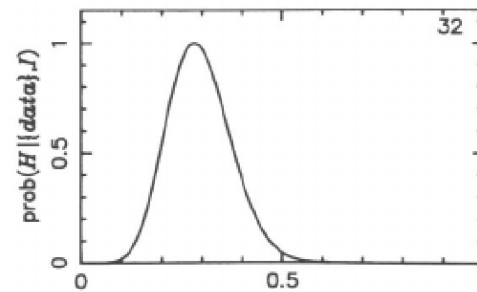
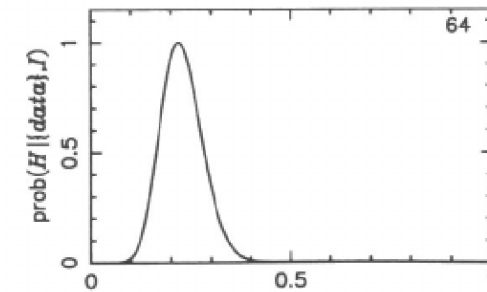
n . Versuch: p wird immer genauer, im Grenzfall $n \rightarrow \infty$ Gauß-förmig verteilt.


 $L_0(p_{\text{Kopf}})$

 $L_1(p_{\text{Kopf}})$

 $L_2(p_{\text{Kopf}})$

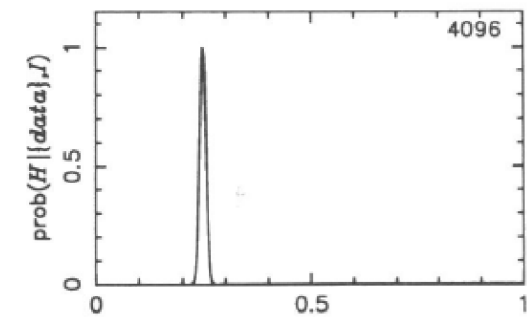
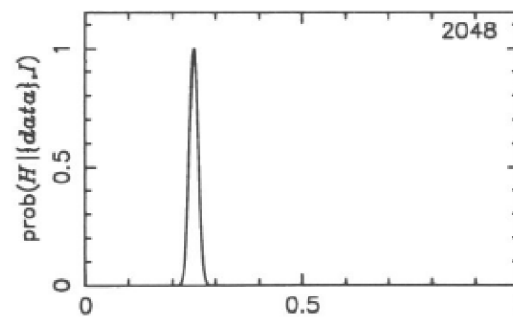
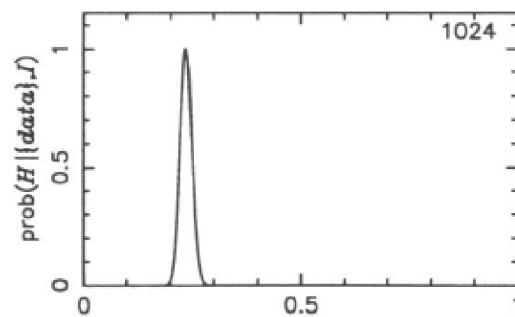
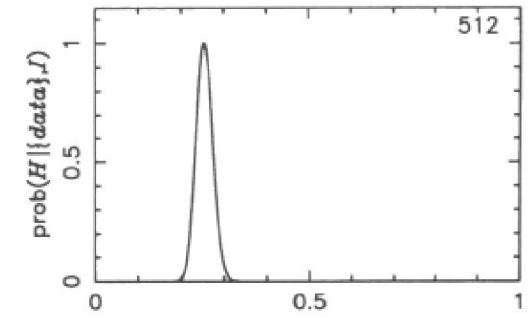
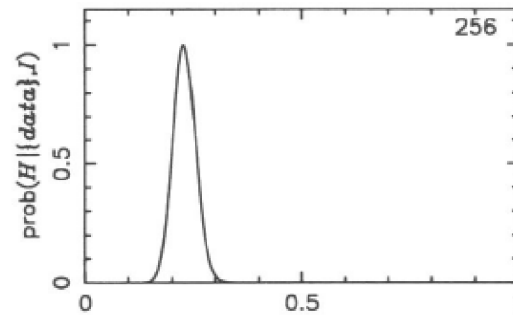
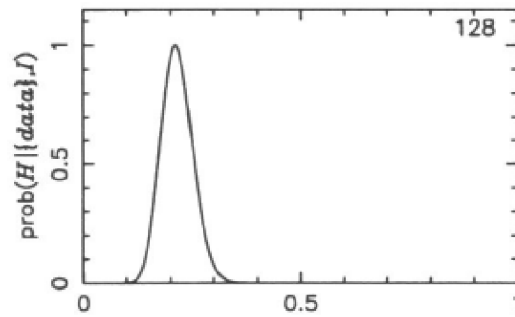
→ p


 $L_3(p_{\text{Kopf}})$

 $L_4(p_{\text{Kopf}})$

 $L_8(p_{\text{Kopf}})$

→ p


 $L_{16}(p_{\text{Kopf}})$

 $L_{32}(p_{\text{Kopf}})$

 $L_{64}(p_{\text{Kopf}})$

→ p

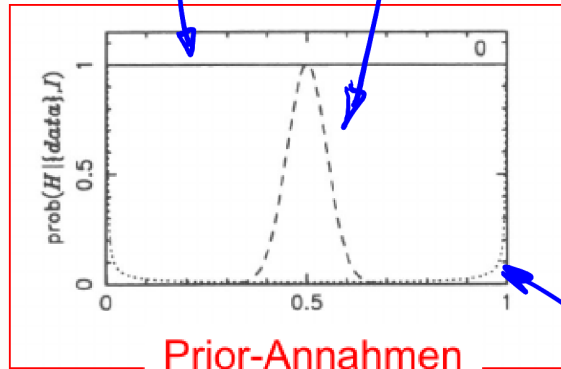


Fehler wird mit $1/\sqrt{n}$ kleiner

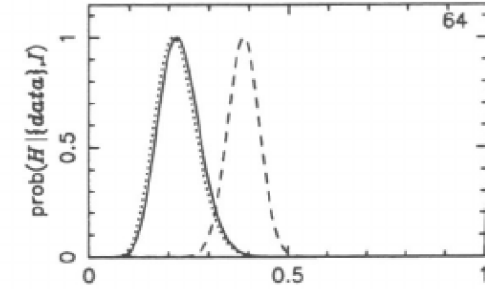
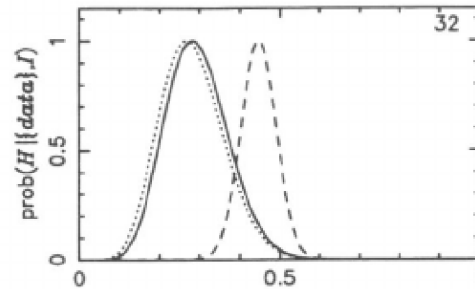
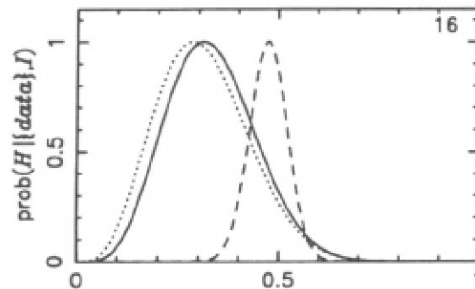
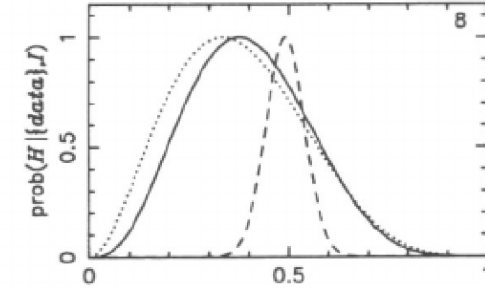
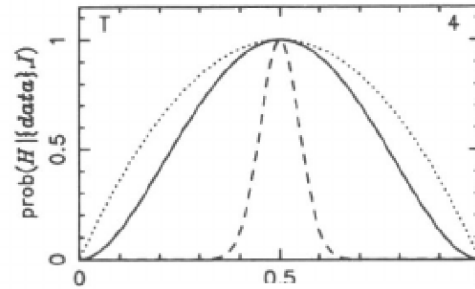
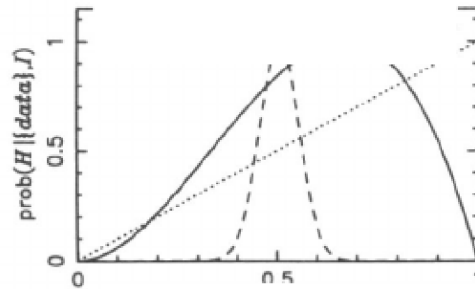
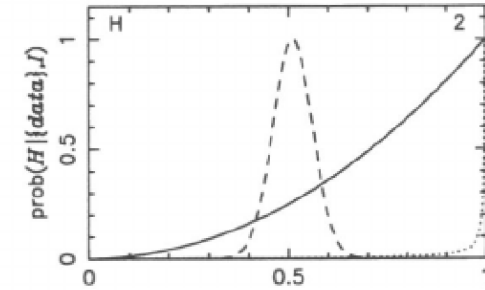
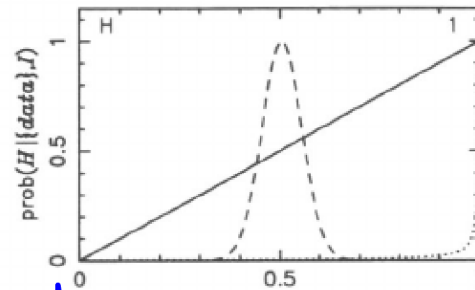
Münze ist falsch !

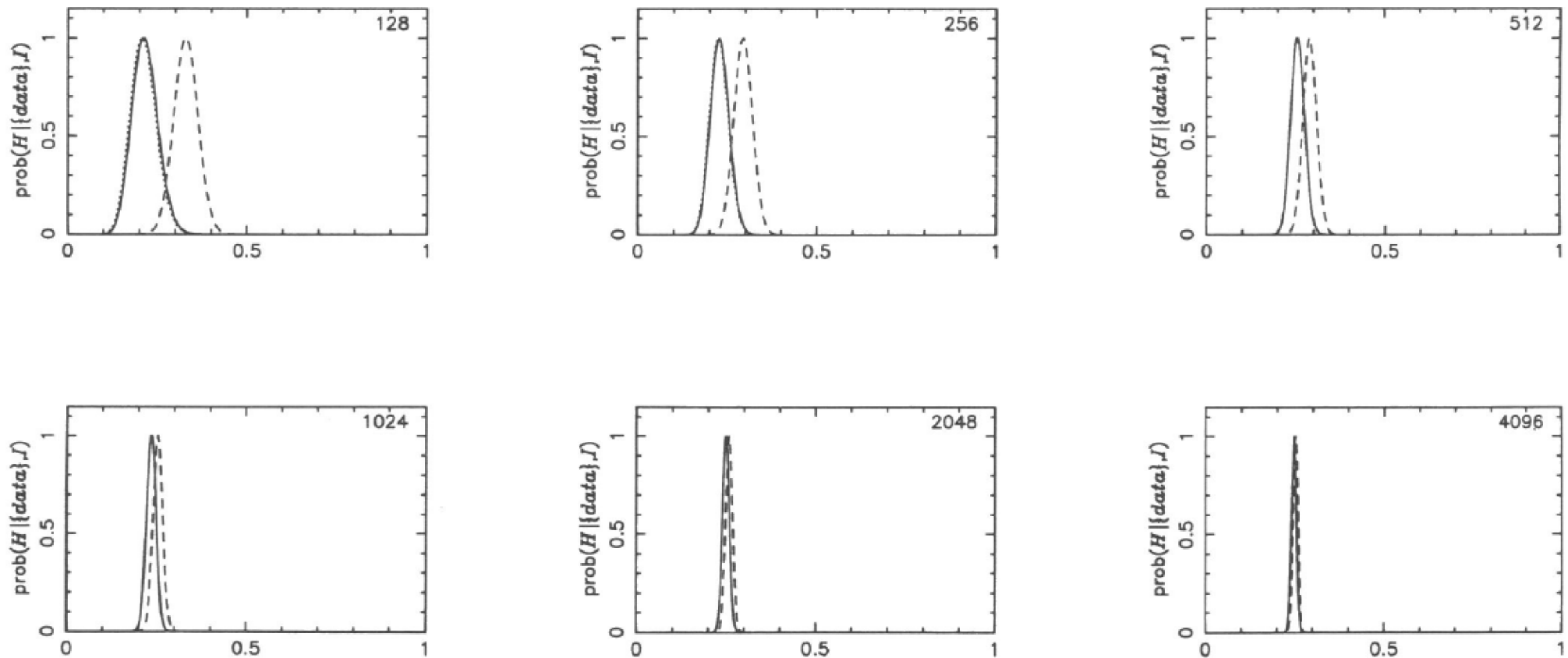
flach, keine Information
zufällig

Maximiere $L(x_i | p) * f_{\text{prior}}(p)$



baised





Fehler wird mit $1/\sqrt{n}$ kleiner

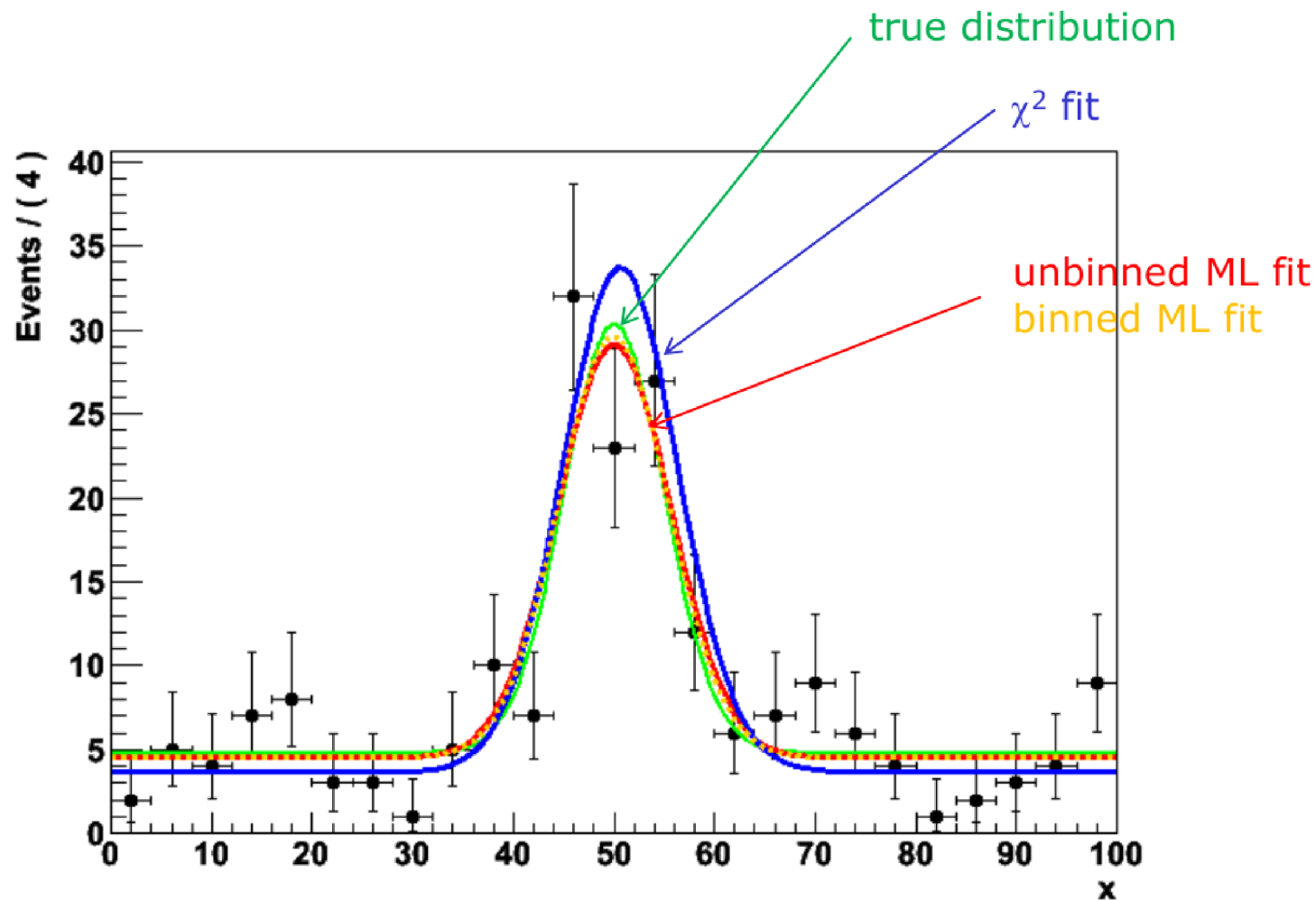
Münze ist (auch jetzt noch) falsch !

Einfluss des Priors bei hoher Statistik vernachlässigbar

Vergleich:

	Maximum - Likelihood	Kleinste Quadrate
Voraussetzung	PDF exakt bekannt	Mittelwert und Varianz bekannt
Methode	Höhe der PDF	Abweichung vom Mittelwert der PDF
Effizienz	maximal	maximal bei linearen Problemen
Komplexität	aufwändig, meist nicht-linear	oft linear und exakt lösbar
Robustheit	nein - PDF muss exakt bekannt sein	nein („Ausreißer“)
korrelierte Datenfehler	u.U. kompliziert	einfach über Kovarianzmatrix
Güte der Anpassung	nein	ja: χ^2 -Wahrscheinlichkeit
Bias (dt. Verzerrung)	i.A. ja	nein bei linearen Problemen (wenn Fehler der Daten unverzerrt sind)
Spezialfall	identisch bei Gauß-förmigen Fehlern	

Beispiel/Vergleich bei kleinen Stichproben



Wouter Verkerke, NIKHEF

ML vs. χ^2 – Was soll ich jetzt tun?

- χ^2 fit is fastest, easiest
 - Works fine at high statistics
 - Gives absolute goodness-of-fit indication
 - Make (incorrect) Gaussian error assumption on low statistics bins
 - Has bias proportional to $1/N$
 - Misses information with feature size $<$ bin size
- Full Maximum Likelihood estimators most robust
 - No Gaussian assumption made at low statistics
 - No information lost due to binning
 - Gives best error of all methods (especially at low statistics)
 - No intrinsic goodness-of-fit measure, i.e. no way to tell if 'best' is actually 'pretty bad'
 - Has bias proportional to $1/N$
 - Can be computationally expensive for large N
- Binned Maximum Likelihood in between
 - Much faster than full Maximum Likelihood
 - Correct Poisson treatment of low statistics bins
 - Misses information with feature size $<$ bin size
 - Has bias proportional to $1/N$

$$-\ln L(p)_{\text{binned}} = \sum_{\text{bins}} n_{\text{bin}} \ln F(\bar{x}_{\text{bin-center}}; \vec{p})$$

Fehler in x- und y-Richtung

1. Ableitung nutzen, um x-Fehler in y-Fehler umzurechnen und Quadratisch zu den y-Fehlern addieren

Iteratives Verfahren:

1. Anpassung ohne x-Fehler
2. $f'(x_i)$ bilden und neue Fehler bestimmen:

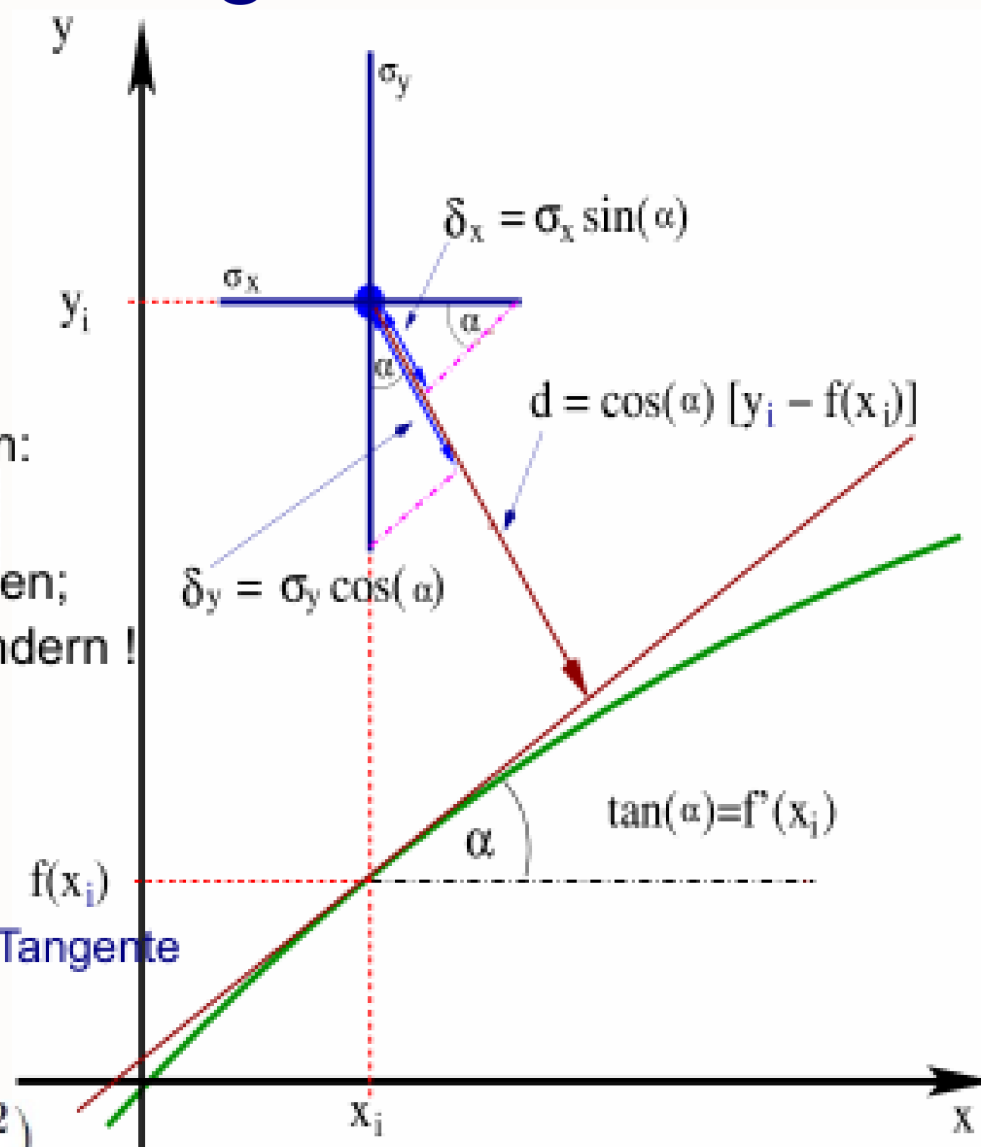
$$\sigma_i^2 = \sigma_{y_i}^2 + f'(x_i) \cdot \sigma_{x_i}^2$$
3. Schritt analog 2 zur Kontrolle wiederholen;
 χ^2 am Minimum darf sich nicht stark ändern !

Geometrische Interpretation:

Minimierung des auf projizierte Messfehler normierten Abstands d der Punkte von der Tangente an die Funktion:

$$\chi^2 = d^2 / (\cos(\alpha) \sigma_y^2 + \sin(\alpha) \sigma_x^2)$$

Implementierung: Methode Fit() der Root-Klasse TGraphErrors



Zusammenfassung

- Jeder Schätzer kann konsistent, treu, effizient und invariant sein.
 - ML-Schätzer Theorem: ML treu, wenn es einen treuen Schätzer gibt!
 - Im Allgemeinen: jeder Schätzer sollte genau charakterisiert werden!
 - X^2 -Schätzer entsprechen in sehr vielen praktischen Fällen dem MLE, aber Vorsicht bei: Poisson, Multinomial, Binomial, und anderen asymmetrischen Verteilungen.
- Bayesian Analyse mit flachem Prior entspricht Likelihood Analyse
 - Likelihood Analyse kann mit Prior Verteilung gekoppelt werden
 - Für große N wird der Einfluss des Priors gering
- Goodness-of-Fit: Im 1-dimensionalen sehr gute Methoden: X^2 , Pearson, L-Ratio, KS, Run-Test, etc.
- Im Allgemeinen sehr komplex und keine eindeutige Lösung: Toy-MC Studien und p-Werte