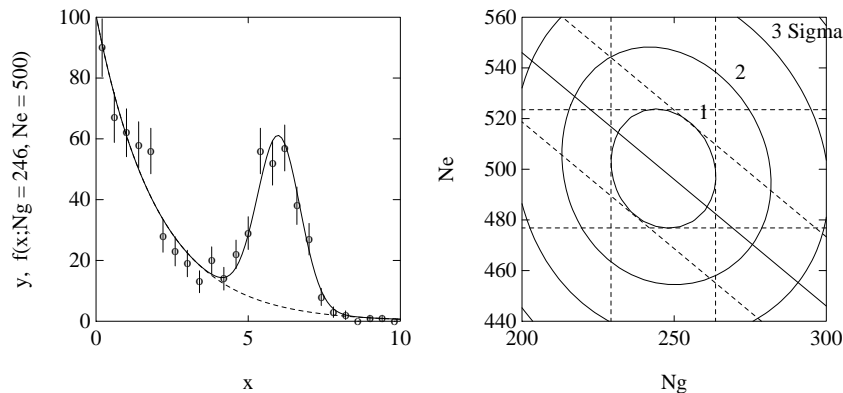


DATENANALYSE

H. Pruis

Physik-Institut der Universität Zürich, SS 2004

Die Datenanalyse befasst sich mit der Auswertung von Messergebnissen. In der Vorlesung werden die theoretischen Grundlagen kurz erläutert und ihre Anwendungen an vielen praktischen Beispielen ausführlich besprochen. Die Beispielen stammen aus VP-Versuchen, Diplom- und Doktorarbeiten. In den Übungen werden gemessene oder simulierte Daten am Computer mit MATLAB gemeinsam analysiert. MATLAB integriert Matrixberechnungen, numerische Analyse, Datenauswertung und Grafik (siehe Figur) in einer benutzerfreundlichen Umgebung.



Likelihood-Analyse eines Gauss-Peaks mit exponentiellem Untergrund

Inhalt:

Einführung: MATLAB, Grafische Darstellung, Histogramm, Verteilungsfunktion, Mittelwert, Varianz, Gauss-Verteilung.

Verteilungen: Binomial-, Poisson-, Exponential-, χ^2 - und Lorentz-Verteilung. Monte-Carlo-Methode, Zweidimensionale Gauss-Verteilung, Korrelation, Faltung.

Methode der kleinsten Quadrate: Anpassung einer Geraden, eines Polynoms und einer allgemeinen linearen Funktion. Anpassung nicht-linearer Funktionen.

Maximum-Likelihood-Methode: Vergleich mit der Methode der kleinsten Quadrate. Anpassung eines Gauss-Peaks mit exponentiellem Untergrund (siehe Figur).

Kapitel 1

Einführung

In diesem Kapitel werden die wichtigsten Begriffe und einfache Anwendungen der Datenanalyse besprochen. Dieser Teil entspricht ungefähr dem Anhang FR (Fehlerrechnung und Statistik) aus der Anleitung zum Anfängerpraktikum (der als bekannt vorausgesetzt wird). Weiter wird eine Einführung in MATLAB gegeben, das wir vor allem in den Übungen ausführlich benutzen werden, um grafische Darstellungen und numerische Auswertungen zu machen.

Die weiteren Kapitel behandeln verschiedene Wahrscheinlichkeitsverteilungen und statistische Methoden.

Literatur:

P. R. Bevington and D. K. Robinson, *Data Reduction and Error Analysis* [Bevi 92]

R. Barlow, *Statistics* [Barl 89]

Beide Bücher geben eine elementare Darstellung der Datenanalyse und sind für den Anfänger sehr geeignet. Inhaltlich überlappen sie am meisten mit der Vorlesung.

S. Brandt, *Datenanalyse* [Bran 99]

Das Buch gibt eine ausführliche Darstellung der Datenanalyse und ist vor allem als theoretische Ergänzung zur Vorlesung zu empfehlen.

W. T. Eadie et al., *Statistical Methods in Experimental Physics* [Eadi 71]

Eine sehr fundamentale, aber nicht einfache Darstellung der Datenanalyse mit vielen Beispielen aus der Teilchenphysik. Dringend empfohlen für künftige Diplomanden und Doktoranden der Teilchenphysik. Es wird empfohlen, mindestens eines dieser Bücher neben der Vorlesung ausführlich zu benutzen.

The Student Edition of MATLAB [MATLAB]

D. M. Etter, *Engineering Problem Solving with MATLAB* [Ette 93]

Beide Bücher geben eine Einführung in MATLAB. Für diejenigen, die eine traditionelle Programmiersprache bevorzugen, ist folgendes Buch sehr zu empfehlen: W. H. Press et al., *Numerical Recipes in C (PASCAL, FORTRAN)* [Press 88] Auch MATLAB Benutzer werden hierin viele nützliche Erläuterungen zu den in den MATLAB-Toolbox vorhandenen Routinen finden.

Das Verzeichnis der weiteren im Vorlesungsskript zitierten Literatur ist am Ende des Skripts zu finden.

1.1 Fehler, Wahrscheinlichkeit und Statistik

Die Datenanalyse befasst sich mit der Auswertung von Messergebnissen. Die exakte Messung einer physikalischen Grösse ist nicht möglich. Jede Messung weist Fehler auf, die systematisch oder zufällig sein können. Das Resultat einer Messung ist damit

$$\boxed{\text{Wert} \pm \text{Fehler}} \quad (1.1)$$

Systematische Fehler treten auf, wenn unbekannte störende Einflüsse das Ergebnis verfälschen. Sind die Störungen bekannt, so kann man sie durch eine Korrektur beseitigen. Der systematische Fehler ist in diesem Fall die abgeschätzte Genauigkeit der Korrektur. Das Wiederholen derselben Messung verkleinert den systematischen Fehler nicht. Beispiele sind ein falsch geeichtes Messinstrument, ein nicht berücksichtigter Untergrund bei einer Aktivitätsmessung und ein Vorurteil des Experimentators (siehe Abb. 1.1). Systematische Fehler sind sehr schwer zu erkennen und werden oft erst durch eine andere Messmethode aufgedeckt. Sie hängen sehr stark von der experimentellen Methode ab und sind deshalb für eine allgemeine Behandlung nicht geeignet. Ihre Minimalisierung gehört eher zur Experimentiertechnik als zur Datenanalyse. Beide Gebiete sind aber nicht zu trennen.

Statistische Fehler (zufällige Fehler, “statistical, random, accidental errors”) treten bei jeder Messung auf. Ihre Ursache sind elementarer statistischer Natur, wie Braunsche Bewegung, thermisches Rauschen und stochastische Folge von diskreten Ereignisse wie radioaktiver Zerfall und Streuung eines Teilchens. Jede Messwert ist damit zufällig (nicht beliebig oder unbestimmt!) und entspricht in der Theorie einer **Zufallsvariablen** (“random variable”). Im Prinzip kann man diese Fehler beliebig klein machen, in dem man die gleiche Messung beliebig oft wiederholt. Die Berechnung der statistischen Fehler ist eine der Hauptaufgaben der Datenanalyse.

Systematische und statistische Fehler werden bei Zwischenresultaten separat behandelt. Sind die Zwischenresultate statistisch voneinander unabhängig, so folgt der totale statistische Fehler aus dem naiven **Fehlerfortpflanzungsgesetz** (siehe Anleitung zum Anfängerpraktikum). Die systematischen Fehler der Zwischenresultate sind fast immer stark miteinander korreliert, was in der Berechnung des totalen systematischen Fehlers berücksichtigt werden muss (Kap. 3.2.3). Im Schlussresultat werden statistischer und systematischer Fehler separat angegeben. Als Gesamtfehler kann man noch die Wurzel aus der Quadratsumme dieser Fehler angeben.

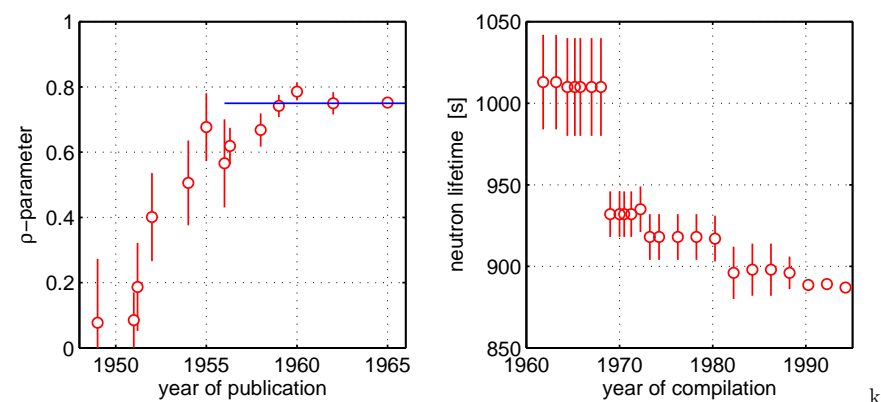


Abbildung 1.1: Links: Experimentelle Bestimmung der ρ -Parameter im μ -Zerfall als Funktion der Zeit [Lee 81]. Die Vorhersage der $V - A$ -Theorie (1956) ist $3/4$. Rechts: Die akzeptierte Lebensdauer des Neutrons als Funktion der Zeit [PDG].

Die Theorie der statistischen Fehler ist die **Wahrscheinlichkeitslehre**. Eine theoretische Wahrscheinlichkeitsfunktion mit bekannten Parametern bestimmt, mit welcher Wahrscheinlichkeit ein Messwert x_k oder ein Bereich von Messwerten ($x_1 \leq x \leq x_2$) auftritt. Die für die Datenanalyse und damit für die Experimentalphysik wichtigste Wahrscheinlichkeitsfunktion (Wahrscheinlichkeitsdichte, Verteilung, “probability density function, distribution”) ist die Gauss-Verteilung (siehe Kap. 1.4, andere wichtige Verteilungen werden in

Kap. 2 besprochen)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(x - \mu)^2/2\sigma^2\right] \quad (1.2)$$

Diese Verteilung beschreibt die Streuungen (statistische Schwankungen) der Messwerte x um den theoretischen Wert μ . Aus ihr folgt, dass die Wahrscheinlichkeit ein Messwert x im Bereich $(\mu - \sigma, \mu + \sigma)$ zu beobachten 68% ist. Das bedeutet aber nicht, dass man bei $N = 100$ Messungen genau 68 Messwerte innerhalb dieses Bereiches beobachtet. Die beobachtete Anzahl schwankt um diesem Wert. Diese Schwankungen werden durch die Binomialverteilung (die fundamentalste Wahrscheinlichkeitsverteilung, siehe Kap. 2.1) beschrieben. Für $N \rightarrow \infty$ wird die Abweichung des beobachteten Wertes vom theoretisch Vorhergesagten (d.h. der statistische Fehler) verschwindend klein. Wie wir sehen werden, ist der

$$\text{statistische Fehler} \propto 1/\sqrt{N} \quad (1.3)$$

Wahrscheinlichkeitsverteilungen spielen auch in der theoretischen Physik eine wichtige Rolle. In der Quantenmechanik ist der Wahrscheinlichkeitsinterpretation fundamental: das Quadrat der Wellenfunktion $|\Psi(x)|^2 dx$ ist die Wahrscheinlichkeit eine Ortskoordinate x im Bereich $(x, x + dx)$ zu beobachten. Auch in der klassischen Physik kommen Wahrscheinlichkeitsverteilungen vor, z.B. die Maxwell'sche Geschwindigkeitsverteilung der Gasmoleküle.

In der **Statistik** (Datenanalyse) ist das Problem umgekehrt: Aus gemessenen Werten soll die zugrunde liegende theoretische Wahrscheinlichkeitsverteilung und ihre Parameter bestimmt werden. Ein bekanntes Beispiel (Anfängerpraktikum Versuch RA) ist die Messung der Aktivität einer radioaktiven Probe als Funktion der Zeit. Die gemessenen Daten entsprechen (sofern keine systematische Fehler vorliegen) einer Exponentialverteilung (siehe Kap. 2.3) mit dem zu bestimmenden Parameter λ (die Zerfallskonstante). Dies ist auch ein Beispiel der Datenreduktion: Die vielen primären Messwerte und ihre Fehler ergeben einen **Schätzwert** (Schätzung, "estimate") für den physikalisch relevanten Parameter und seinen Fehler. Der Schätzwert ist eine Funktion von Zufallsvariablen (die Messwerte) und damit selber auch eine Zufallsvariable. Die wichtigsten statistischen Methoden zur Schätzung von Parametern der theoretischen Verteilungen, sind die Methode der kleinsten Quadraten (Kap. 4) und die Maximum-Likelihood-Methode (Kap. 5).

Die Statistik wurde vor allem im Zusammenhang mit demographischen Untersuchungen entwickelt, was ihrer Fachsprache deutlich anzusehen ist. Statt von Messdaten mit einer zugrunde liegenden theoretischen Verteilung wird von einer **Stichprobe** ("sample") aus einer Population (Grundgesamtheit, "population, parent distribution") gesprochen. Auch der Begriff "Regressionsanalyse" (lineare Anpassung mit der Methode der kleinsten Quadrate) stammt aus der Soziologie.

1.2 Grafik, Histogramm und MATLAB

Der erste Schritt der Datenanalyse ist die grafische Darstellung der Messergebnisse und sollte parallel zur Messung erfolgen: jeder neue Messwert wird laufend in die Grafik eingetragen. In einer Grafik sieht man sofort ob die Daten "vernünftig" sind oder nicht. Als Beispiel betrachten wir die Daten in Abb. 1.1. Die Messwerte des ρ -Parameters im μ -Zerfall sind offensichtlich nicht miteinander verträglich (erstaunlich ist aber, dass je zwei nach einander gemessene Werte konsistent sind!). Das bedeutet, dass systematische Fehler vorliegen und der Mittelwert der Messungen keine Bedeutung hat.

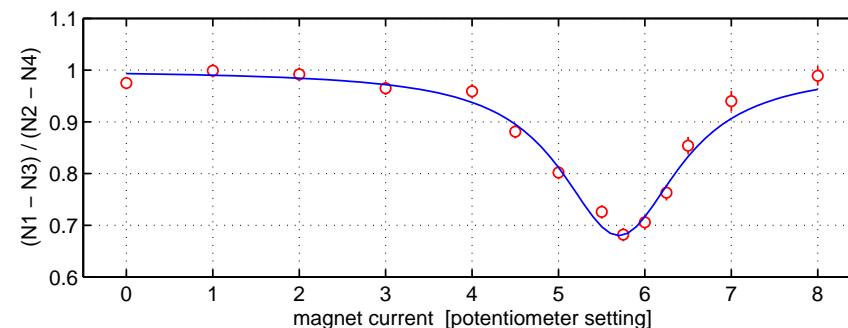


Abbildung 1.2: VP-Versuch "Lambshift": Messdaten und Fitfunktion.

Ist die Messung beendet und sehen die Daten "vernünftig" aus, dann ist der nächste Schritt eine Theoriefunktion "vom Auge" an die Daten anzupassen (qualitative Auswertung). Ein Beispiel aus dem VP-Versuch "Lambshift" ist in Abb. 1.2 gegeben. Die primären Messdaten sind vier Zählraten (N_1 , N_2 , N_3 und N_4), die in Abhängigkeit vom Magnetfeld gemessen werden (siehe VP-Anleitung "Lambshift").

Die relevante Grösse ist das Verhältnis $(N_1 - N_3)/(N_2 - N_4)$, das entsprechend einer umgekehrten Lorentz-Verteilung (Kap. 2.5)

$$f(x) = 1 - \frac{p_3(1 - p_2)}{(x - p_1)^2 + p_3} \quad (1.4)$$

vom Magnetfeld ($\propto x$) abhängen sollte. Der Parameter $p_1 \simeq 5.7$ ist die Lage des Minimums, der Parameter $p_2 \simeq 0.68$ ist der Funktionswert beim Minimum und der Parameter $p_3 \simeq 0.7$ entspricht die Breite $p_3 = \Gamma^2/4$ ($\Gamma = \text{FWHM} = \text{Full Width at Half Maximum}$). Man sieht (Abb. 1.2), dass die Daten mit dieser Verteilung kompatibel sind. Als letzten Schritt (in diesem einfachen Fall) kann man jetzt eine quantitative Anpassung mit der Methode der kleinsten Quadrate machen (Kap. 4) um einen genauen Wert des Lage-Parameters p_1 und seines Fehlers zu erhalten.

MATLAB

Die grafische Darstellung ist mit MATLAB sehr einfach. MATLAB (MATrix LABoratory) integriert Matrixberechnungen, numerische Analyse, Datenauswertung und Grafik in einer benutzerfreundlichen Umgebung. Es steht auf vielen Plattformen (MAC, PC, Linux- und UNIX-Systemen) zur Verfügung und wird in Naturwissenschaften und Technik viel benützt.

Das Grundelement in MATLAB ist eine Matrix; Skalare sind 1×1 Matrizen, ein n -dimensionaler Zeilenvektor ist eine $1 \times n$ Matrix und ein n -dimensionaler Spaltenvektor ist eine $n \times 1$ Matrix. Alle Operatoren sind in Prinzip Matrix-Operatoren, z.B. $A*B$ ist eine Matrix-Multiplikation oder, wenn eine der beiden Matrizen ein Skalar ist, die Multiplikation mit einer Konstante. Auch die elementweise Verknüpfung zweier Matrizen ist möglich, wenn man vor dem Operator einen Punkt schreibt, z.B. $c = A.*B$ ist die elementweise Multiplikation: $c(i,k) = A(i,k)*B(i,k)$.

```
%-----
%   MATLAB program 1.1: Lambshift      (/home/hep/pruys/matlab/daten0102/Kap1)
%   -----
load Pix_ratio
plot_data(Pix_ratio),
grid,                                axis([-0.5 8.5 0.6 1.1])
xlabel('magnet current [potentiometer setting]')
ylabel('(N1 - N3) / (N2 - N4)')
p = [5.7 0.68 0.7];                  %optimize these values!
```

```
x = 0:0.1:8;
fx = 1 - p(3)*(1 - p(2)) ./ ((x - p(1)).^2 + p(3));
plot(x,fx,'b-')
```

```
%-----
%   Pix_ratio: data measured by Pixley in 1994
%-----
0.00  0.975  0.012
1.00  0.999  0.013
2.00  0.992  0.013
3.00  0.965  0.013
4.00  0.959  0.014
4.50  0.881  0.013
5.00  0.802  0.013
5.50  0.726  0.013
5.75  0.682  0.013
6.00  0.706  0.014
6.25  0.763  0.015
6.50  0.854  0.017
7.00  0.940  0.020
8.00  0.989  0.019
%-----
```

Damit wird das Programmieren in MATLAB sehr einfach, z.B. wurde die Abb. 1.2 mit einem 9-Zeilen-Programm erzeugt (siehe Programm 1.1), das sich fast von selbst erklärt. Der `load` Befehl liest ein Datenfile ein (in diesem Fall eine 14×3 Matrix), `axis` spezifiziert die x - und y -Achse und `plot_data(Pix_ratio)` zeichnet die Messpunkte mit den Fehlerbalken in der Grafik. `p = [5.7 0.68 0.7]`; definiert einen 1×3 Parametervektor, `x = 0:0.1:8`; definiert einen Zeilenvektor mit 81 x -Werten,

```
fx = 1 - p(3)*(1 - p(2)) ./ ((x - p(1)).^2 + p(3));
```

erzeugt einen Zeilenvektor mit den entsprechenden 81 Funktionswerten die mit dem Befehl `plot(x,fx,'g-')` gezeichnet werden (man beachte die Punkte vor den Operatoren `/` und `^`).

MATLAB ist ein interaktives System. Man kann ein Programm mit MATLAB-Befehlen (z.B. Programm 1.1) schreiben und anschliessend laufen lassen oder die Befehle einzeln im "Command-Window" ausführen. Es hat ein ausführliches Help-System und viele Beispiele (im Ordner "Demonstrations"). Als Beispiel das "Command-Window" nachdem MATLAB frisch gestartet und Programm 1.1 ausgeführt wurde:

```
>> whos
  Name          Size          Bytes  Class

  Pix_ratio     14x3           336   double array
  fx            1x81           648   double array
  p             1x3            24   double array
  x            1x81           648   double array
```

```
Grand total is 207 elements using 1656 bytes
```

```
>> help plot_data
  plot_data(data,color)
  input:  array data with 2 or 3 columns:
          x(i), y(i), (sigma(i) optional)
          for 2 columns poisson errors are assumed
          color: 'y', 'm', 'c', 'g', 'r', 'b', 'k' (optional)
  output: plot of data y(i) +/- sigma(i)
```

Histogramm

Bei vielen Experimenten ist es notwendig die primären Messdaten für eine sinnvolle grafische Darstellung in einem Histogramm zusammen zu fassen. Die Auswertung kann mit den primären Daten erfolgen (z.B. mit der Maximum-Likelihood-Methode, Kap. 5) oder mit den histogrammierten Daten (z.B. mit der Methode der kleinsten Quadrate, Kap. 4). Typische Beispiele sind die Messungen einer Energie- oder Zeit-Verteilung mit einem Teilchendetektor, z.B. die Messung der Energie der gestreuten Photonen in der Compton-Streuung. Im entsprechenden VP-Versuch stehen die primären Daten (die Energien der einzelne Photonen) jedoch nicht zu Verfügung, sondern nur die mit einem Vielkanalanalysator histogrammierten Daten. Als einfaches Beispiel werden wir deshalb ein Monte-Carlo-Experiment (siehe Kap. 3.1) betrachten: Die Messung der Energie der Elektronen im μ -Zerfall (siehe Vorlesung Teilchenphysik I)

$$\mu^- \rightarrow e^- \nu_\mu \bar{\nu}_e$$

mit einem Detektor mit einer sehr scharfen Energieauflösung. Unter Vernachlässigung der Elektronenmasse ($m_e \ll m_\mu$) variiert die Energie der Elektronen zwischen Null und der maximalen Energie $E_{max} = m_\mu c^2 / 2 \simeq 53$ MeV. Die Tab. 1.1 enthält die simulierten Daten für 180 Ereignisse, wobei die relativen Energien $x_i = E_i / E_{max}$ ($i = 1 \dots 180$) angegeben sind.

Diese Daten sind in einem Histogramm mit 10 Bins¹ zusammengefasst (siehe Abb. 1.3). Bin k ($k = 1 \dots 10$) enthält alle Werte x_i (insgesamt y_k Werte) mit $x_k - \Delta x / 2 \leq x_i < x_k + \Delta x / 2$, wobei $\Delta x = 0.1$ die Binbreite ist. x_k ist die Energie, die der Mitte des Bins entspricht: $x_1 = 0.05$, $x_k = k/10 - 0.05$ und $x_{10} = 0.95$. Als Beispiel sind in der Tabelle alle Werte, die in Bin 4 fallen, fett gedruckt.

Tabelle 1.1: Normierte Energie des Elektrons in 180 μ -Zerfallereignisse (Monte-Carlo-Daten).

	1	2	3	4	5	6	7	8	9
1	0.811	0.794	0.850	0.707	0.919	0.781	0.719	0.719	0.746
2	0.856	0.941	0.544	0.995	0.645	0.787	0.764	0.599	0.757
3	0.748	0.972	0.922	0.949	0.920	0.447	0.828	0.395	0.817
4	0.781	0.842	0.802	0.776	0.706	0.661	0.448	0.983	0.351
5	0.565	0.700	0.787	0.921	0.257	0.423	0.810	0.997	0.831
6	0.763	0.565	0.609	0.339	0.977	0.530	0.965	0.667	0.425
7	0.987	0.401	0.816	0.845	0.814	0.599	0.962	0.274	0.713
8	0.941	0.869	0.975	0.777	0.499	0.351	0.396	0.269	0.825
9	0.456	0.823	0.764	0.726	0.976	0.494	0.693	0.941	0.919
10	0.969	0.979	0.798	0.724	0.979	0.214	0.658	0.657	0.988
11	0.773	0.913	0.582	0.406	0.467	0.512	0.625	0.911	0.591
12	0.316	0.592	0.422	0.972	0.965	0.606	0.691	0.369	0.709
13	0.707	0.705	0.968	0.574	0.288	0.634	0.728	0.319	0.800
14	0.624	0.704	0.488	0.992	0.942	0.304	0.884	0.430	0.886
15	0.780	0.706	0.404	0.864	0.594	0.949	0.508	0.316	0.861
16	0.705	0.882	0.702	0.356	0.989	0.754	0.927	0.270	0.857
17	0.998	0.979	0.457	0.604	0.424	0.522	0.752	0.656	0.605
18	0.989	0.250	0.730	0.643	0.643	0.614	0.452	0.408	0.627
19	0.774	0.971	0.821	0.534	0.974	0.849	0.305	0.881	0.892
20	0.274	0.743	0.766	0.679	0.346	0.720	0.851	0.519	0.929

Ein Histogramm kann auf verschiedene Arten dargestellt werden. Die Abb. 1.3 wurde mit dem MATLAB-Programm 1.2 erzeugt. Das Histogrammieren erfolgt mit dem Befehl `y = hist(data,x)`, wobei `data` die primäre Daten sind (siehe Tab. 1.1) und der `x`-Vektor die Bins definiert (`x = xmin+bin/2:bin:xmax-bin/2`). Die histogrammierten Werte stehen im `y`-Vektor zur Verfügung. Das Histogramm kann

¹Die n Bins werden mit 1 bis n indiziert, da dies der Konvention der meisten Programmiersprachen (MATLAB, PASCAL, FORTRAN) entspricht. Eine wichtige Ausnahme ist C, wo ein Vektor mit n Elementen von 0 bis $n - 1$ indiziert wird. Bedingt durch die digitale Elektronik (die Adressen eines RAMs mit 2^n Bytes laufen von 0 bis $2^n - 1$), laufen die Kanalnummern in einem Vielkanalanalysator von 0 bis $2^n - 1$.

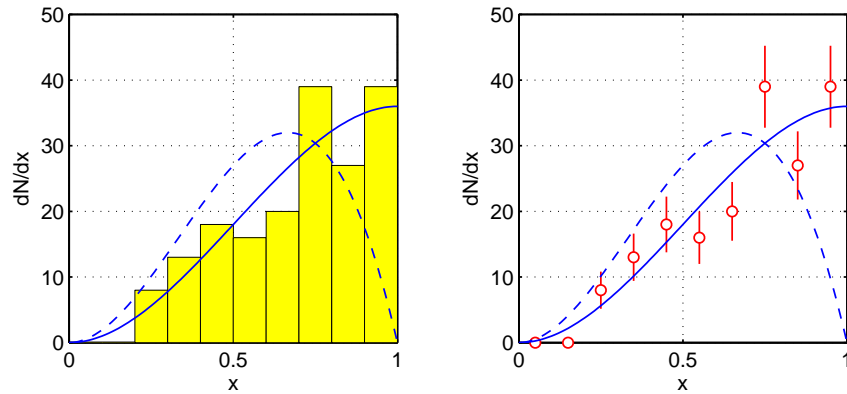


Abbildung 1.3: Energiespektrum des Elektrons in Myon-Zerfall. Die histogrammierte Monte-Carlo-Daten sind auf zwei Arten dargestellt. Die Theoriekurven sind für $\rho = 3/4$ (—) und $\rho = 0$ (- -).

mit dem Befehl `bar(x,y)` oder dem Befehl `plot_data([x' y'])` gezeichnet werden (x und y Zeilenvektoren, $[x' y']$ ist eine Matrix mit zwei Spalten). Die Balken-Darstellung (Abb. 1.3 links) ist eine Flächen-Darstellung und die Gesamtzahl der Ereignisse multipliziert mit der Binbreite entspricht der Gesamtfläche (im Beispiel 180×0.1 , `norm = sum(y)*bin`)

$$\boxed{\sum_{k=1}^n y_k \Delta x = \text{norm}} \quad (1.5)$$

die zur Normierung der Theoriefunktion benutzt werden kann. Die Theoriefunktion ist als MATLAB-Funktion (`function y = Espec(x,rho)`) deklariert, die wie die vordefinierten MATLAB-Funktionen (Befehle) aufgerufen werden kann. Das Integral der Theoriefunktion

$$f(x) = [12(1-x) + 8\rho(4x/3 - 1)]x^2 \quad (1.6)$$

ist

$$\int_{x_{min}}^{x_{max}} f(x) dx = 1$$

Die "Punkt \pm Fehler"-Darstellung (Abb. 1.3 rechts) zeigt die zu erwartenden statistischen Schwankungen in der Anzahl der Ereignisse pro Bin. Diese

Schwankungen entsprechen der Binomialverteilung (2.1) oder der Poisson-Verteilung (Kap. 2.2)

$$\boxed{\text{Histogramm} \Rightarrow \text{Poisson-Statistik} \Rightarrow y_k \pm \sqrt{y_k}} \quad (1.7)$$

Diese Darstellung sollte nicht mit der Darstellung einzelner Messdaten und ihrer Fehler (z.B. Abb. 1.2) verwechselt werden. Aus der Achsenbeschriftung sollte der Unterschied sofort klar sein. Die y -Achse eines Histogramm ist mit "dN/dx" oder "Anzahl Ereignisse pro Bin" mit der Einheit 1/Einheit(x) beschriftet (z.B. x -Achse [MeV] und y -Achse [MeV⁻¹]). Die Darstellung mit den Fehlerbalken ist sehr geeignet für ein Vergleich mit einer Theoriefunktion. Im Beispiel sieht man deutlich, dass der Wert $\rho = 3/4$ mit den Daten verträglich ist (sie wurden mit diesem Wert generiert) und der Wert $\rho = 0$ jedoch nicht.

```

%-----
%   MATLAB program 1.2: Muon Decay
%   -----
load mudata                               % a 180 x 2 matrix
data = mudata(:,2);                       % data in column 2 (1: 1..180)
xmin = 0;                                 xmax = 1;
bin = 0.1;                                x = xmin+bin/2:bin:xmax-bin/2;
y = hist(data,x);                         subplot(121)
b = bar(x,y,1);                           set(b,'FaceColor','y')
axis([0 1 0 50]),                          axis square
xlabel('x'),                               ylabel('dN/dx')
hold on,                                   grid
norm = sum(y)*bin;                         xth = 0:0.01:1;
yth1 = norm*Espec(xth,3/4);                yth2 = norm*Espec(xth,0);
plot(xth,yth1,'b-'),                       plot(xth,yth2,'b--')
hold off,                                  subplot(122)
plot_data([x' y']),                        hold on
axis([0 1 0 50]),                          grid
axis square
plot(xth,yth1,'b-'),                       plot(xth,yth2,'b--')
xlabel('x'),                               ylabel('dN/dx')

%-----
function y = Espec(x,rho)
% calculates energy spectrum of electron in muon decay
% x: normalized electron energy, rho: parameter (= 3/4 for V-A)
y = (12*(1-x) + 8*rho*(4*x/3 - 1)).*x.^2;
%-----

```

1.3 Verteilung, Mittelwert und Varianz

1.3.1 Theoretische Verteilungen

Theoretische **Verteilungen** (“distribution”) können kontinuierlich sein (z.B. die Gauss-Verteilung, Gl. 1.2 oder die Energieverteilung der Elektronen im μ -Zerfall, Gl. 1.6) oder diskret (z.B. die Poisson-Verteilung, Kap. 2.2). Eine kontinuierliche Verteilung $f(x)$ ist eine Wahrscheinlichkeitsdichte (“probability density function”): $f(x)dx$ ist die Wahrscheinlichkeit einen Wert x_i im Intervall $(x, x + dx)$ zu beobachten. Die Wahrscheinlichkeitsfunktion $F(x)$ (“cumulative distribution function”) ist die Wahrscheinlichkeit einen Wert $x_i < x$ zu beobachten

$$F(x) = \int_{x_{min}}^x f(x') dx' \quad F(x_{max}) = \int_{x_{min}}^{x_{max}} f(x) dx = 1 \quad (1.8)$$

Eine diskrete Verteilung P_k ist eine Wahrscheinlichkeitsfunktion: $P_k \equiv P(x_k)$ ist die Wahrscheinlichkeit einen Wert x_k zu beobachten. Die Anzahl möglicher Werte x_k kann endlich oder unendlich sein, die Summe über alle Werte ist Eins:

$$\sum_{k=1}^{k_{max}} P_k = 1 \quad (1.9)$$

Die n möglichen x -Werte werden wir meistens mit $k = 1$ bis $k = k_{max} = n$ numerieren. Nur bei den Verteilungen mit $x_k = 0, 1, 2, \dots, \infty$ ist es sinnvoller mit $k = 0$ anzufangen, damit $x_k = k$ ist.

Zum numerischen Vergleich mit den experimentellen Daten werden wir oft auch eine kontinuierliche Verteilung diskretisieren (histogrammieren mit Binbreite Δx und Werte $x_{k+1} = x_k + \Delta x$)

$$P_k \equiv F(x_k + \Delta x/2) - F(x_k - \Delta x/2) \simeq f(x_k) \Delta x \quad (1.10)$$

Der **Mittelwert** (Erwartungswert von x , “mean”) und die Varianz sind die zwei wichtigsten Parameter einer Verteilung. Der Mittelwert ist

$$\mu \equiv \langle x \rangle \equiv E(x) \equiv \int_{x_{min}}^{x_{max}} x f(x) dx \quad \text{bzw.} \quad \equiv \sum_{k=1}^{k_{max}} x_k P_k \quad (1.11)$$

und die **Varianz**

$$\sigma^2 \equiv \langle (x - \mu)^2 \rangle \equiv E((x - \mu)^2) \equiv \int_{x_{min}}^{x_{max}} (x - \mu)^2 f(x) dx \quad (1.12)$$

$$\text{bzw.} \quad \equiv \sum_{k=1}^{k_{max}} (x_k - \mu)^2 P_k$$

Die Standardabweichung (“standard deviation”) σ (die Quadratwurzel aus der Varianz) ist ein Mass für die Streuung von x um den Mittelwert μ . Der Erwartungswert ist ein linearer Operator

$$E((x - \mu)^2) = E(x^2 - 2x\mu + \mu^2) = E(x^2) - 2\mu E(x) + \mu^2 E(1) = \langle x^2 \rangle - \mu^2 \quad (1.13)$$

Ist der Mittelwert $\mu = 0$, so ist das zweite Moment $\langle x^2 \rangle$ (in diesem Fall = σ^2) der wichtigste Parameter der Verteilung. Ein wichtiges Beispiel ist der mittlere quadratische Radius R_{rms}^2 (rms: “root mean square”) einer kugelsymmetrischen Ladungsverteilung in der Atom- oder Kernphysik

$$R_{rms}^2 \equiv \langle r^2 \rangle = \iiint r^2 |\Psi(\vec{r})|^2 d\vec{r} = 4\pi \int_0^\infty r^2 |\Psi(r)|^2 r^2 dr$$

Die einfachste kontinuierliche Verteilung ist die **Gleichverteilung** (“uniform distribution”)

$$f(x; a, b) = 1/(b - a) \quad \text{mit} \quad a = x_{min} \leq x \leq x_{max} = b \quad (1.14)$$

Der Mittelwert und die Varianz sind

$$\mu = (a + b)/2 \quad \text{und} \quad \sigma^2 = (b - a)^2/12 \quad (1.15)$$

Die Gleichverteilung $f(x; 0, 1)$ ist die Grundverteilung der Monte-Carlo-Methode (Kap. 3.1). Ein physikalisches Beispiel ist die isotrope Winkelverteilung

$$W(\theta) = 1/4\pi \quad \text{mit} \quad \iint W(\theta) d\Omega = 2\pi \int_{-1}^1 W(\theta) d \cos \theta = 1$$

Man beachte, dass nicht θ , sondern $\cos \theta$ gleichverteilt ist!

```

%-----
%   MATLAB program 1.3: mean and standard deviation
%   -----
xmin = 0;                xmax = 1;
bin = 0.01;              x = xmin+bin/2:bin:xmax-bin/2;
rho = 3/4;               y = Espec(x,rho)*bin;
int = sum(y);            out = centroid(x,y);
fprintf('\nint: %5.3f, mean: %5.3f, sigma: %5.3f',int,out(1),out(2))

%-----
function out = centroid(x,y)
%-----
%   input: x: channel number, y: contents
%   output: out(1) = mean: <x>
%           out(2) = variance: sigma = sqrt(<x^2> - <x>^2)
%           out(3) = error of mean: sigma/sqrt(N), N = sum(y)
%
N = sum(y);
xm = sum(x.*y)/N;
sig = sqrt(sum(x.^2 .*y)/N - xm^2);
error = sig/sqrt(N);
out = [xm sig error];

%-----
% output:   int: 1.000, mean: 0.700, sigma: 0.208
%-----

```

Als nächstes Beispiel betrachten wir die Energieverteilung der Elektronen im μ -Zerfall (Gl. 1.6)

$$f(x; \rho) = 12x^2(1-x) + 8\rho x^2(4x/3 - 1) \equiv f_1(x) + \rho f_2(x) \quad (1.16)$$

Der Mittelwert ist

$$\mu = \int_0^1 x f_1(x) dx + \rho \int_0^1 x f_2(x) dx = (9 + 2\rho)/15 \quad (1.17)$$

Das zweite Moment ² ist

$$\langle x^2 \rangle = \int_0^1 x^2 f_1(x) dx + \rho \int_0^1 x^2 f_2(x) dx = (18 + 8\rho)/45$$

²Die Momente einer Verteilung $f(x)$ sind $M_l \equiv \langle x^l \rangle$, wobei $M_0 = 1$ die Normierung ist und $M_1 = \mu$ der Mittelwert.

Die zentralen Momente sind $Z_l \equiv \langle (x - \mu)^l \rangle$, wobei $Z_2 = \sigma^2$ die Varianz ist.

und die Varianz

$$\sigma^2 = \langle x^2 \rangle - \mu^2 = (9 + 4\rho - 4\rho^2)/225 \quad (1.18)$$

Für die beide Werte von ρ gibt dies

$$\rho = 0: \mu = 3/5, \sigma = 1/5 \quad \text{und} \quad \rho = 3/4: \mu = 7/10, \sigma = \sqrt{39}/30$$

Die MATLAB-Funktion `function out = centroid(x,y)` berechnet für diskrete Verteilungen den Mittelwert (Gl. 1.11) und die Varianz (Gl. 1.13). In Programm 1.3 wird diese Funktion aufgerufen, um die Normierung (Gl. 1.8), den Mittelwert und die Standardabweichung der Energieverteilung numerisch zu berechnen. Dazu wurde die kontinuierliche Verteilung in 100 Bins histogrammiert (Gl. 1.10). Die numerischen Werte stimmen mit den analytisch berechneten überein. In dieser numerischen Berechnung wurde das Integral durch eine Summe über die 100 Funktionswerte \times Binbreite ersetzt (`y = Espec(x,rho)*bin; int = sum(y)`). Dies ist die einfachste Art der numerischen Integration, die einer linearen Interpolation der Funktionswerte entspricht (siehe [Pres 88] für bessere Methoden). Man beachte, dass die Normierung, unabhängig von ρ , gleich Eins ist, wie es sein sollte.

1.3.2 Gemessene Verteilungen

Ein normiertes **Histogramm** mit gemessenen Werten y_k/N (Zahl der Ereignisse in Bin k dividiert durch die Gesamtzahl der Ereignisse, $N = \sum y_k$) ist eine Häufigkeitsverteilung, die für $N \rightarrow \infty$ mit der entsprechenden theoretischen Verteilung übereinstimmt (sofern keine systematischen Fehler vorliegen). Der **Mittelwert** (Schwerpunkt, “centroid”) einer Häufigkeitsverteilung ist

$$\bar{x} \equiv \frac{1}{N} \sum_{k=1}^n x_k y_k \quad \text{oder} \quad \bar{k} \equiv \frac{1}{N} \sum_{k=1}^n k y_k \quad (1.19)$$

Für $P_k = y_k/N$ ist der Mittelwert einer gemessenen Verteilung \bar{x} formal identisch mit dem Mittelwert einer theoretischen Verteilung $\langle x \rangle$ (Gl. 1.11), aber im Gegensatz zum $\langle x \rangle$, ist \bar{x} eine Zufallsvariable, die von Messung zu Messung variiert und deshalb einen statistischen Fehler σ_x hat. Bei einer gemessenen Verteilung ist auch die Zuordnung der Binnummer k (Kanalnummer) zur physikalischen Grösse x (z.B. eine Energie) mit (systematischen) Fehlern behaftet. Sie wird mit einer Eichmessung bestimmt. Deshalb ist der relative Fehler in \bar{x} (σ_x/\bar{x}) grösser als der relative Fehler in \bar{k} und ist es besser bei Zwischenresultaten mit k statt x zu rechnen (z.B. $\Delta k = \bar{k}_2 - \bar{k}_1 \Rightarrow \Delta x$ statt $\bar{k}_2 - \bar{k}_1 \Rightarrow \bar{x}_2 - \bar{x}_1 = \Delta x$). In den weiteren Betrachtungen werden wir aber den Eichfehler nicht berücksichtigen und nur noch die Variable x benützen. Die **Varianz** der *gemessenen* Verteilung ist

$$s^2 \equiv \frac{1}{N} \sum_{k=1}^n (x_k - \bar{x})^2 y_k \quad (1.20)$$

und entsprechend für k statt x . Sie ist ein Mass für die Schwankungen von x relativ zu \bar{x} , die einen Faktor $N/(N-1)$ kleiner sind als die Schwankungen von x relativ zum “wahren” Wert $\langle x \rangle$, den wir nicht kennen (sonst würden wir keine Messung machen).

Die besten **Schätzwerte** (Kap. 1.3.5) x_0 und σ_0^2 für den Mittelwert $\langle x \rangle$, bzw. die Varianz σ^2 der zugrundeliegenden *theoretischen* Verteilung sind

$$x_0 = \bar{x} \quad \text{und} \quad \sigma_0^2 = \frac{1}{N-1} \sum_{k=1}^n (x_k - \bar{x})^2 y_k \quad (1.21)$$

Die Zufallsvariable \bar{x} ist eine Funktion von den gemessenen Werten y_k . Der Fehler

von \bar{x} folgt aus dem Fehlerfortpflanzungsgesetz und der Poisson-Statistik

$$\sigma_{\bar{x}}^2 = \sum_{k=1}^n \sigma_{y_k}^2 \left(\frac{\partial \bar{x}}{\partial y_k} \right)^2 = \sum_{k=1}^n y_k \left(\frac{\partial \bar{x}}{\partial y_k} \right)^2 \quad (1.22)$$

Mit der Kettenregel erhält man ($N = \sum y_k$)

$$\frac{\partial \bar{x}}{\partial y_k} = \frac{-1}{N^2} \frac{\partial N}{\partial y_k} \sum_{k=1}^n x_k y_k + \frac{x_k}{N} = \frac{x_k - \bar{x}}{N}$$

Das Quadrat des **Fehlers** des Mittelwerts ist damit

$$\sigma_{\bar{x}}^2 = \frac{1}{N^2} \sum_{k=1}^n (x_k - \bar{x})^2 y_k = \frac{s^2}{N} \quad (1.23)$$

```

%-----
%   MATLAB program 1.4: mean, standard deviation and error of mean
%   -----
load mudata,                               N = length(mudata);
m = mean(mudata(:,2));                       s0 = std(mudata(:,2));
fprintf('\nmean:%6.3f,  sigma:%6.3f,  error:%6.3f',m,s0,s0/sqrt(N))
xmin = 0;                                    xmax = 1;
bin = 0.1;                                   x = xmin+bin/2:bin:xmax-bin/2;
y = hist(mudata(:,2),x);                     out = centroid(x,y);
fprintf('\nmean:%6.3f,  sigma:%6.3f,  error:%6.3f',out(1),out(2),out(3))

%-----
% output:mean: 0.697,  sigma: 0.214,  error: 0.016
%           mean: 0.698,  sigma: 0.209,  error: 0.016
%-----

```

Als Beispiel wurde mit dem Programm 1.4 (siehe auch die Programme 1.2 und 1.3) der Mittelwert mit seinem Fehler und die Varianz der “gemessenen” Energieverteilung der Elektronen im μ -Zerfall (Abb. 1.3) berechnet. Der Mittelwert

$$\bar{x} = 0.698 \pm 0.016 \quad (1.24)$$

stimmt mit dem Mittelwert der theoretischen Verteilung $\mu = 7/10$ für $\rho = 3/4$ (Gl. 1.17) überein und ist im Widerspruch mit dem Wert $\mu = 3/5$ für $\rho = 0$ (eine “6 σ ”-Diskrepanz).

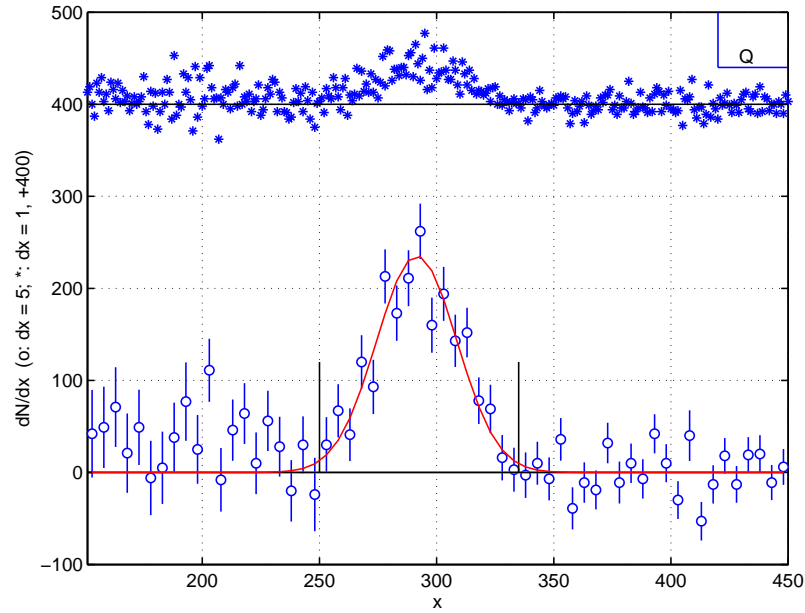


Abbildung 1.4: Differenzspektrale Comptonstreuung (Signal – Untergrund) Oben: ursprüngliches Binning des PHA (Nulllinie bei 400); Unten: Das gleiche Spektrum einen Faktor 5 komprimiert mit Gauß-Fit .

1.3.3 *Differenz zweier gemessenen Verteilungen

Die Formeln werden etwas komplizierter, wenn die experimentelle Verteilung die Differenz zweier gemessenen Verteilungen ist (z.B. Signal – Untergrund)

$$y_k = s_k - u_k, \quad S \equiv \sum_{k=1}^n s_k, \quad U \equiv \sum_{k=1}^n u_k, \quad N = S - U \quad (1.25)$$

Die Formeln für den Mittelwert

$$\bar{x} = \frac{1}{N} \sum_{k=1}^n x_k y_k = \frac{1}{S - U} \sum_{k=1}^n x_k (s_k - u_k) \quad (1.26)$$

und die Varianz

$$s^2 \equiv \frac{1}{N} \sum_{k=1}^n (x_k - \bar{x})^2 y_k = \frac{1}{S - U} \sum_{k=1}^n (x_k - \bar{x})^2 (s_k - u_k) \quad (1.27)$$

sind direkte Verallgemeinerungen der Gln. 1.19 bzw. 1.20. Der Fehler des Mittelwerts muss jedoch mit dem Fehlerfortpflanzungsgesetz neu berechnet werden (vgl. Gl. 1.22)

$$\sigma_{\bar{x}}^2 = \sum_{k=1}^n \left[\sigma_{s_k}^2 \left(\frac{\partial \bar{x}}{\partial s_k} \right)^2 + \sigma_{u_k}^2 \left(\frac{\partial \bar{x}}{\partial u_k} \right)^2 \right] \quad (1.28)$$

Mit der Kettenregel erhält man

$$\frac{\partial \bar{x}}{\partial s_k} = \frac{-\bar{x} + x_k}{S - U}, \quad \frac{\partial \bar{x}}{\partial u_k} = \frac{\bar{x} - x_k}{S - U}$$

Das Quadrat des **Fehlers** des Mittelwerts ist damit

$$\sigma_{\bar{x}}^2 = \frac{1}{(S - U)^2} \sum_{k=1}^n (x_k - \bar{x})^2 (s_k + u_k) \quad (1.29)$$

Dies ist also *nicht* $s^2/(S - U)$, da der Fehler in $y_k = s_k - u_k$ *nicht* $\sqrt{y_k}$ ist, sondern $\sqrt{s_k + u_k}$. Auch der Fehler in der Nettofläche $N = S - U$ ist *nicht* \sqrt{N} , sondern $\sqrt{S + U}$.

Als Beispiel betrachten wir den VP-Versuch “Compton-Streuung” (siehe Anleitung). Das Signal ist das Energiespektrum der im Target gestreuten Photonen und der Untergrund das entsprechende, ohne Target gemessene, Spektrum. Das Differenzspektrum ist im Abb. 1.4 gezeigt. Einmal in der ursprüngliche Auflösung des Vielkanalanalysators (PHA: Puls-Höhen-Analysator), 1 Bin = 1 Kanal und einmal mit 1 Bin = 5 Kanäle. Das Differenzspektrum schwankt rechts vom Peak um die Nulllinie, wie es sein sollte. Links vom Peak haben die meisten Bins einen positiven Inhalt. Dieser Teil des Spektrums entsteht dadurch, dass das auf dem Detektor einfallende Photon nur einen Teil seiner Energie im Detektor verliert. Die Selektion des Peaks ist mit dem Programm 1.5 interaktiv möglich. Zwischen den beiden selektierten Linien wird die Position (Mittelwert, Gl. 1.26) mit Fehler (Gl. 1.29) und die Nettofläche mit Fehler berechnet. Die Resultate hängen von der Peak-Selektion ab (siehe Output von Programm 1.5), die nicht eindeutig ist. Diese Abhängigkeit sollte als systematischer Fehler (im Beispiel klein im Vergleich zum statistischen Fehler) im Gesamtfehler berücksichtigt werden. Zur optischen Unterstützung der Peak-Selektion wird ein Gauß-Peak mit gleichem Mittelwert, gleicher Varianz und gleicher Fläche eingezeichnet.

```

%-----
%   MATLAB program 1.5: evaluates peak in difference spectrum
%   -----
load Sspec,                load Uspec
y = Sspec - Uspec;        sy = Sspec + Uspec;
xa = 151;                  xb = 450;
ya = -100;                 yb = 500;
offset = 400;
x = xa:xb;
y = y(x);                  sy = sy(x);
plot(x,y + offset,'b*'),  hold on
plot([1 max(x)],[offset offset],'k-'), bin = 5;
axis([xa xb ya yb])
x = xa+(bin-1)/2:bin:xb-(bin-1)/2;  nchan = length(x);
y = sum(reshape(y,bin,nchan));      sy = sum(reshape(sy,bin,nchan));
data = [x' y' sqrt(sy)'];          plot_data(data,'b')
plot([1 max(x)],[0 0],'k-'),      xlabel('x')
ylabel('dN/dx (o: dx = 5; *: dx = 1, +400)')
dx = 0.1*(xb - xa);              dy = 0.1*(yb - ya);
lc = 0.2*(yb - ya);              % cursor length
plot([xb-dx xb], [yb-dy yb-dy]),  plot([xb-dx xb-dx], [yb yb-dy])
text(xb-0.7*dx,yb-0.8*dy,'Q'),    grid
fprintf('\n Specify peak boundaries with mouse')
fprintf('\n Stop evaluation by clicking in Q box\n')
fprintf('\n-----')
fprintf('\n xpa xpb position +- error contents +- error')
fprintf('\n-----\n')
xpao = -1;                      xpbo = -1;
k = 1;                           gy = zeros(size(x));
while k < 10
    [xpa, dummy] = ginput(1);
    if dummy > yb-dy
        break;
    end
    plot(x,gy*(x(2) - x(1)), 'w-');  plot([1 max(x)], [0 0], 'k-')
    if k > 1
        plot_data(data(select,1:3), 'b');
    end
    xpa = round(xpa);
    plot([xpao xpao], [0 lc], 'w'),  plot([xpa xpa], [0 lc], 'k')
    [xpb, dummy] = ginput(1);        xpb = round(xpb);
    plot([xpbo xpbo], [0 lc], 'w'),  plot([xpb xpb], [0 lc], 'k')
    select = x >= xpa & x <= xpb;    ypeak = y(select);
end

```

```

xpeak = x(select);           out = centroid(xpeak,ypeak);
m = out(1);                  v = out(2);
gy = sum(ypeak)*gauss(x,m,v); plot(x,gy*(x(2) - x(1)), 'r-')
e = sqrt(sum((xpeak-m).^2 .* sy(select)))/sum(ypeak);
cont = sum(ypeak);          ce = sqrt(sum(sy(select)));
fprintf('%4.0f %4.0f %8.1f', xpa,xpb,m)
fprintf('%8.1f %8.0f %8.0f\n', e,cont,round(ce))
xpao = xpa;                  xpbo = xpb;
k = k+1;
end

```

```

%-----
% output:  xpa xpb position +- error contents +- error
% -----
%          250 335 291.5 1.4 2025 117
%          235 335 291.9 2.0 2011 132
%          235 350 291.8 2.3 2011 138
%          250 350 291.5 1.7 2025 124
%          250 330 291.5 1.3 2022 114
%-----

```

Eine nützliche Funktion im MATLAB ist `reshape(y,bin,nchan)`. Damit wird der n -dimensionale Spaltenvektor y in eine $(bin, nchan)$ -Matrix umgewandelt ($bin * nchan = n$). Mit `sum` werden die Spalten addiert und es entsteht einen $nchan$ -dimensionale Zeilenvektor. Die interaktive Selektion des Peaks ist mit dem Befehl `[xpa, dummy] = ginput(1)` möglich. Damit werden die "mouse"-Koordinaten beim "mouse click" eingelesen. Mit den eingelesenen x -Koordinaten wird der Peak selektiert:

```

select = x >= xpa & x <= xpb;    ypeak = y(select);
xpeak = x(select);

```

Der Mittelwert und Varianz der Peak-Verteilung werden, wie vorher (siehe Programm 1.3) mit dem Befehl `out = centroid(xpeak,ypeak)` berechnet. Der Fehler des Mittelwerts (Gl. 1.29) wird mit dem Befehl

```

e = sqrt(sum((xpeak-m).^2 .* sy(select)))/sum(ypeak);

```

berechnet.

1.3.4 Mittelwert und gewichteter Mittelwert

Bisher haben wir den Mittelwert von theoretischen (Gl. 1.11) und gemessenen Verteilungen (Gl. 1.19) betrachtet. Der Mittelwert einer Serie von Einzelmesswerte x_i ist

$$\bar{x} \equiv \frac{1}{N} \sum_{i=1}^N x_i \quad (1.30)$$

und die Varianz (vgl. die Gln. 1.12 und 1.20)

$$s^2 \equiv \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (1.31)$$

Wie vorher (Gl. 1.22), folgt der Fehler des Mittelwerts aus dem Fehlerfortpflanzungsgesetz,

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^N \sigma_i^2 \left(\frac{\partial \bar{x}}{\partial x_i} \right)^2 \quad (1.32)$$

Der Fehler des Einzelmesswertes ist eine Konstante $\sigma_i = \sigma$.

Mit $\partial \bar{x} / \partial x_i = 1/N$ erhält man (vgl. Gl. 1.23)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Ist σ unbekannt, so kann sie aus der gemessenen Varianz abgeschätzt werden (Gl. 1.21)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \simeq \frac{\sigma_0}{\sqrt{N}} = \frac{s}{\sqrt{N-1}} \quad (1.33)$$

Diese Formeln auf die 180 μ -Zerfallsereignisse (Tab. 1.1) angewandt gibt die gleichen Resultate (siehe Program 1.4) wie für die histogrammierten Daten. Der Mittelwert (Gl. 1.30) wurde mit der MATLAB-Funktion `mean` berechnet. Der Standardabweichung ist in der MATLAB-Funktion `std` mit $N - 1$ statt N (Gl. 1.31) normiert³, was die geringe Differenz ergibt. Mit `std` wird also nicht s , die Abweichung vom Mittelwert \bar{x} , sondern σ_0 , die Schätzung für die Abweichung vom “wahren” Wert μ berechnet (siehe Gl. 1.21).

³The definition of σ is a minefield of alternatives, and to call it the “standard” deviation is something of a sick joke [Barl 89].

Sind die Fehler σ_i unterschiedlich, dann haben die Messungen x_i unterschiedliche Gewichte g_i und der **gewichteter Mittelwert** ist (Kap. 1.3.5)

$$\bar{x} \equiv \frac{\sum_{i=1}^N g_i x_i}{\sum_{i=1}^N g_i}, \quad g_i = \frac{1}{\sigma_i^2} \quad (1.34)$$

Der Fehler des gewichteten Mittelwerts (interner Fehler) folgt wieder aus dem Fehlerfortpflanzungsgesetz (Gl. 1.32 mit $\partial \bar{x} / \partial x_i = g_i / \sum g_i$)

$$\sigma_{int} = \frac{1}{\sqrt{\sum_{i=1}^N g_i}} \quad (1.35)$$

Für ein konstantes Gewicht $g_i = 1/\sigma^2$ erhält man wieder Gl. 1.33.

Die Varianz der mit unterschiedlichen Fehlern gemessenen Werten ist (vgl. die Gln. 1.12, 1.20 und 1.31)

$$s^2 \equiv \frac{\sum_{i=1}^N g_i (x_i - \bar{x})^2}{\sum_{i=1}^N g_i} \quad (1.36)$$

Auch damit lässt sich, wie beim konstanten Messfehler (Gl. 1.33), der Fehler des gewichteten Mittelwerts (externer Fehler) abschätzen

$$\sigma_{ext} = \frac{s}{\sqrt{N-1}} \quad (1.37)$$

Interner und externer Fehler sollte, innerhalb bestimmter Grenzen, gleich sein (siehe Kap. 1.3.5). Als Fehler des Mittelwerts sollte der grösste der Beiden angegeben werden

$$\sigma_{\bar{x}} = \max(\sigma_{int}, \sigma_{ext}) \quad (1.38)$$

```
%-----
%   MATLAB program 1.6: weighted mean value
%   -----
subplot(121)
load n_lifetime,
wmean(n_lifetime)
xlabel('measurement'),
hold off,
load n_lambda,
wmean(n_lambda),
axis('square')
ylabel('lifetime [s]')
subplot(122)
axis('square')
```

```

xlabel('measurement'),          ylabel('\lambda')

%-----
function wmean(data)
%-----
% function wmean(data): calculates (weighted) mean value
% input: array data with 2 or 3 columns: i, y(i), (sig(i) optional)
% output: mean value and errors, chisquare and probability
%
x = data(:,1);                  y = data(:,2);
[n,nc] = size(data);
if nc > 2
    sig = data(:,3);           % third column are errors
else
    var = sum((y - sum(y)/n).^2)/(n-1);
    sig = sqrt(var)*ones(size(x)); % equal errors calculated
end                               % from variance
axis([0 x(n)+1 min(y-1.2*sig) max(y+1.2*sig)])
plot_data([x y sig]),            hold on
g = 1 ./ (sig.*sig);            sm2 = 1/sum(g);
ism = sqrt(sm2);                ym = sum(g.*y)*sm2;
chi2 = sum(g.*(y - ym).^2);      esm = sqrt(chi2/(n-1))*ism;
p = 100*Pchisqr(n-1,chi2);      sm = max(ism,esm);
plot([0 x(n)+1],[ym+sm ym+sm],'g:'), plot([0 x(n)+1],[ym ym],'g-')
plot([0 x(n)+1],[ym-sm ym-sm],'g:')
fprintf('mean = %8.3e, s_int = %8.3e, s_ext = %8.3e\n',ym,ism,esm)
fprintf('chi^2 = %8.3f, dof = %3.0f, prob = %4.1f %%\n',chi2,n-1,p);

%-----
% output:  mean = 8.870e+02, s_int = 1.559e+00, s_ext = 1.935e+00
%          chi^2 = 12.316,   dof = 8,          prob = 13.8 %
%          mean = 1.257e+00, s_int = 2.752e-03, s_ext = 1.762e-03
%          chi^2 = 1.640,   dof = 4,          prob = 80.2 %
%-----

```

Die MATLAB-Funktion `wmean` berechnet den gewichteten Mittelwert, den internen und externen Fehler. Zwei Beispiele wurden mit dem Programm 1.6 berechnet und die Resultate sind in der Abb. 1.5 gezeichnet.

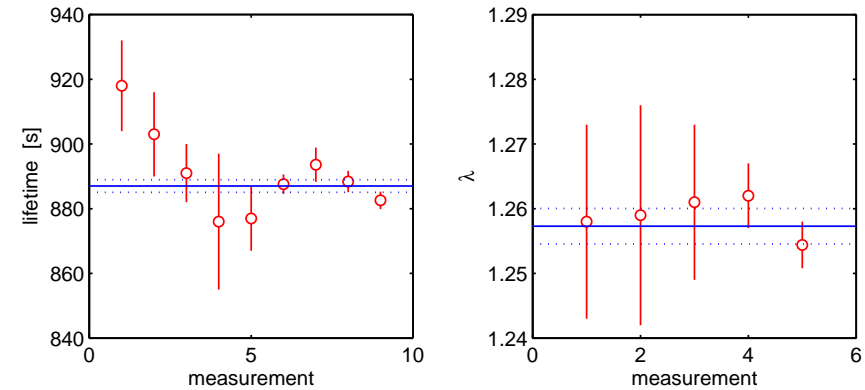


Abbildung 1.5: Zerfallsdaten des Neutrons [PDG 94]: Gewichteter Mittelwert der Lebensdauer und des λ -Parameters ($= |g_A/g_V|$).

1.3.5 Schätzung von Parametern

Das Verhältnis

$$\frac{\sigma_{ext}^2}{\sigma_{int}^2} = \frac{1}{N-1} \sum_{i=1}^N g_i (x_i - \bar{x})^2 \quad (1.39)$$

ist nach einer χ^2 -Verteilung (Kap. 2.4) mit $N-1$ Freiheitsgraden (“dof = degrees of freedom”) verteilt. Die Variable

$$\chi^2 \equiv \sum_{i=1}^N \frac{(x_i - f_i)^2}{\sigma_i^2} \quad (1.40)$$

hat den Erwartungswert und die Varianz

$$\langle \chi^2 \rangle = N - m, \quad \sigma_{\chi^2}^2 = 2(N - m), \quad (1.41)$$

wobei f_i der Erwartungswert von x_i ist und m die Zahl der freien Parameter in der Theoriefunktion $f(x)$. Im einfachsten Fall ist $f_i = f(x_i) = \bar{x}$ eine Konstante, die aus den Daten bestimmt wird ($m = 1$ Parameter). Das ist der Grund für den Faktor $N-1$ statt N in den Formeln. Im ersten Beispiel oben (Programm 1.6, Abb. 1.5) ist der Wert $\chi^2 = 12.3$ mit 8 Freiheitsgraden mehr als 1σ zu gross und die

Wahrscheinlichkeit diesen Wert oder einen noch grösseren zu erhalten entsprechend gering (13.8 %, siehe Kap. 2.4).

Ist $f_i = f_i(a)$ von einem Parameter a abhängig, so ist auch $\chi^2 = \chi^2(a)$ eine Funktion von a . Die optimale **Schätzung** a_0 für den Parameter a erhalten wir aus der Forderung, dass χ^2 minimal sei. Dies ist die **Methode der kleinsten Quadrate** und bedeutet, dass die Ableitung nach a verschwindet

$$\left(\frac{d\chi^2(a)}{da} \right)_{a=a_0} = 0 \quad (1.42)$$

Diese Methode ist sehr wichtig um unbekannte Parameter einer Theorie zu bestimmen (schätzen). Wir werden sie in Kap. 4 ausführlich besprechen und anwenden. Als Beispiele geben wir hier zwei einfache Anwendungen.

Als erstes, die Schätzung für den Mittelwert. In diesem Fall ist die Theoriefunktion gleich dem unbekanntem Parameter $f_i = x$ und die Ableitung ist

$$\frac{d\chi^2}{dx} = -2 \sum_{i=1}^N \frac{x_i - x}{\sigma_i^2} = 0 \quad \text{für } x = x_0 \quad (1.43)$$

Damit erhält man die Formel für den gewichteten Mittelwert (Gl. 1.34)

$$\sum_{i=1}^N \frac{x_i}{\sigma_i^2} = \sum_{i=1}^N \frac{x_0}{\sigma_i^2} \quad \text{oder} \quad x_0 = \frac{\sum_{i=1}^N x_i / \sigma_i^2}{\sum_{i=1}^N 1 / \sigma_i^2} \equiv \bar{x}$$

Neben dieser analytischen Lösung ist auch eine grafische Lösung möglich und oft einfacher. Als Beispiel die optimale Schätzung des ρ -Parameters im μ -Zerfall. Die Daten sind die histogrammierten Ereignisse y_k (Abb. 1.3) mit dem Fehler (Poisson-Statistik) $\sigma_k^2 = y_k$. Der Erwartungswert für die Zahl der Ereignisse im Bin k ist (siehe Gl. 1.10) $f_k = N f(x_k; \rho) \Delta x$, wobei $N = 180$ die Gesamtzahl der Ereignisse ist, $f(x_k; \rho)$ die theoretische Verteilung (Gl. 1.6) und $\Delta x = 0.1$ die Binbreite. Da die zwei ersten Bins keine Ereignisse erhalten⁴, ist die Formel für χ^2 in diesem Fall

$$\chi^2(\rho) = \sum_{k=3}^{10} \frac{(y_k - N f(x_k; \rho) \Delta x)^2}{y_k} \quad (1.44)$$

In Abb. 1.6 ist $\chi^2(\rho)$ dargestellt (Programm 1.7). Der Wert von ρ beim Minimum ist $\rho_0 = 0.74$. Auch der Fehler kann in der Abbildung aus den Schnittpunkten der

⁴ Es ist besser die Gesamtzahl der Ereignisse in den ersten drei Bins zu vergleichen mit der erwarteten Anzahl $N \int_0^{0.3} f(x; \rho) dx$.

Linie $\chi^2(\rho_0) + 1$ mit der χ^2 -Kurve abgelesen werden (siehe Kap. 4.1.1). Der Wert

$$\rho_0 = 0.74 \pm 0.12$$

ist in Übereinstimmung mit dem Wert $\rho = 3/4$ aus der $V - A$ Theorie und im Widerspruch mit dem Wert $\rho = 0$ (eine "6 σ Diskrepanz"). Dieses Resultat hatten wir auch schon mit Hilfe des Mittelwerts μ erhalten (Gl. 1.24). Das ist auch klar, weil es eine eindeutige Beziehung zwischen ρ und μ gibt (Gl. 1.17).

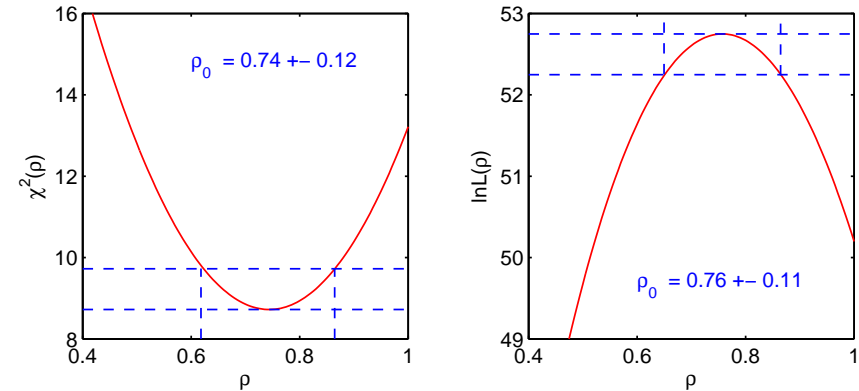


Abbildung 1.6: Bestimmung des ρ -Parameters im Myon-Zerfall mit der Methode der kleinsten Quadrate und der Maximum- Likelihood-Methode.

```

%-----
%   MATLAB program 1.7: least squares and maximum likelihood
%   -----
load mudata;
data = mudata(:,2);
xmin = 0.2;                               xmax = 1;
bin = 0.1;                                 x = xmin+bin/2:bin:xmax-bin/2;
y = hist(data,x);                           N = length(data);
rho = 0.4:0.005:1;                           chi2 = zeros(size(rho));
for k = 1:length(rho)
    chi2(k) = sum((y - N*Especc(x,rho(k))*bin).^2 ./y);
end
subplot(121)
plot(rho,chi2,'-'),                          hold on

```

```

plot([0.4 1],[min(chi2) min(chi2)],'--'), axis([0.4 1 8 16])
plot([0.4 1],[min(chi2)+1 min(chi2)+1], '--')
axis square
xlabel('\rho'), ylabel('\chi^2(\rho)')
disp('specify 1 sigma points (2)'), [x,y] = ginput(2);
plot([x(1) x(1)],[8 y(1)], '--'), plot([x(2) x(2)],[8 y(2)], '--')
txt = sprintf(' = %4.2f +- %4.2f', (x(1)+x(2))/2, (x(2)-x(1))/2);
text(0.6,14.6,['\rho_0' txt])
hold off, lnL = zeros(size(rho));
for k = 1:length(rho)
    lnL(k) = sum(log(Espec(data,rho(k)))));
end
subplot(122)
plot(rho,lnL,'-'), hold on
plot([0.4 1],[max(lnL) max(lnL)], '--'), axis([0.4 1 49 53])
plot([0.4 1],[max(lnL)-0.5 max(lnL)-0.5], '--')
axis square
xlabel('\rho'), ylabel('lnL(\rho)')
disp('specify 1 sigma points (2)'), [x,y] = ginput(2);
plot([x(1) x(1)],[y(1) 53], '--'), plot([x(2) x(2)],[y(2) 53], '--')
txt = sprintf(' = %4.2f +- %4.2f', (x(1)+x(2))/2, (x(2)-x(1))/2);
text(0.6,49.6,['\rho_0' txt])
%-----

```

Eine zweite wichtige Methode zur Schätzung von Parametern ist die **Maximum-Likelihood-Methode**, die hier an einem einfachen Beispiel demonstriert wird (ausführliche Besprechung im Kap. 5). Das Prinzip der Maximum-Likelihood-Methode ist, dass von allen möglichen Werten die beobachteten Werte die grösste Wahrscheinlichkeit haben. Die Wahrscheinlichkeit einen Wert x_i zu messen ist proportional zur Wahrscheinlichkeitsdichte $f(x_i; a)$. Sind die N gemessenen Werte unabhängig voneinander, so ist die Gesamtwahrscheinlichkeit der Messserie

$$L(a) = \prod_{i=1}^N f(x_i; a) \quad (1.45)$$

Dies ist die Likelihoodfunktion mit einem unbekanntem Parameter a . Die beste Schätzwert für a ist der Wert a_0 , der L maximal macht: $L(a_0) \geq L(a)$ für alle a . Statt das Maximum von $L(a)$ ist es gleichwertig das Maximum von

$$\ln L(a) = \sum_{i=1}^N \ln f(x_i; a) \quad (1.46)$$

zu bestimmen. Diese Funktion ist in Abb. 1.6 für den $N = 180$ μ -Zerfälle aus Tab. 1.1 dargestellt (Programm 1.7). Der Schätzwert für den ρ -Parameter und sein Fehler (Schnittpunkte der Linie $\ln L(\rho_0) - 0.5$ mit der $\ln L(\rho)$ -Kurve)

$$\rho_0 = 0.76 \pm 0.11$$

ist in Übereinstimmung mit dem Resultat aus der χ^2 -Auswertung. Unter sehr allgemeinen Bedingungen kann man zeigen (Kap. 5), dass die beiden, auf den ersten Blick so unterschiedlichen Methoden, gleichwertig sind und damit das gleiche Resultat liefern.

1.4 Gauss-Verteilung

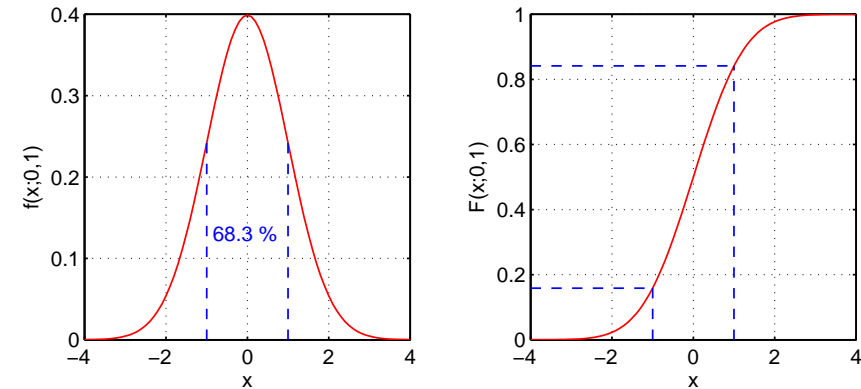


Abbildung 1.7: Wahrscheinlichkeitsdichte und Wahrscheinlichkeitsfunktion der normierten Gauss-Verteilung ($\mu = 0$, $\sigma = 1$).

Die wichtigste Wahrscheinlichkeitsdichte ist, wie wir noch sehen werden (Zentraler Grenzwertsatz, Kap. 3.3.2), die Gauss-Verteilung (Normalverteilung; Gl. 1.2, Abb. 1.7)

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right] \quad (1.47)$$

Der Mittelwert ist μ und die Varianz σ^2 , wie man explizit mit Hilfe der Integrale

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}, \quad \int_0^{\infty} x e^{-ax^2} dx = \frac{1}{2a}, \quad \int_{-\infty}^{\infty} x^2 e^{-ax^2} dx = \frac{1}{2a} \sqrt{\frac{\pi}{a}}$$

und den Definitionen des Mittelwerts (Gl. 1.11) und der Varianz (Gl. 1.12) zeigen kann. Die Wahrscheinlichkeitsfunktion ist

$$F(x; \mu, \sigma) = \int_{-\infty}^x f(x'; \mu, \sigma) dx' \quad (1.48)$$

Numerisch wird sie im MATLAB mit Hilfe der Fehlerfunktion (“error function”)

$$\boxed{\operatorname{erf}(u) \equiv \frac{2}{\sqrt{\pi}} \int_0^u \exp(-t^2) dt} = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{2}u} \exp(-z^2/2) dz \quad (1.49)$$

berechnet (siehe Programm 1.8).

Die Standardform der Gauss-Verteilung ist (Gl. 1.47 mit $\mu = 0$ und $\sigma = 1$)

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp[-z^2/2] \quad (1.50)$$

Jede Gauss-Verteilung lässt sich durch die einfache Variablentransformation

$$z = \frac{x - \mu}{\sigma}$$

auf die Standardform zurückführen.

```

%-----
%   MATLAB program 1.8: Gaussian
%   -----
subplot(121),                x = -4:0.1:4;
plot(x,gauss(x,0,1),'-'),    hold on
plot([-1 -1],[0 gauss(-1,0,1)],'--'), plot([1 1],[0 gauss(1,0,1)],'--')
xlabel('x'),                 ylabel('f(x;0,1)')
grid,                        text(-0.85,0.13,'68.3 %')
axis square,                 subplot(122)
xn = x(x < 0);               Fn = 0.5 - erf(abs(xn)/sqrt(2))/2;
xp = x(x >= 0);              Fp = 0.5 + erf(xp/sqrt(2))/2;
plot(x,[Fn Fp],'--'),       hold on
Fn1 = 0.5 - erf(1/sqrt(2))/2; Fp1 = 1 - Fn1;
plot([-1 -1],[0 Fn1], '--'), plot([1 1],[0 Fp1], '--')
plot([-4 -1],[Fn1 Fn1], '--'), plot([-4 1],[Fp1 Fp1], '--')
xlabel('x'),                 ylabel('F(x;0,1)')
grid,                        axis square

%-----
function y = gauss(x,xm,sig)
x = (x - xm)/sig;
y = exp(-0.5*x.^2)/(sig*sqrt(2*pi));
%-----

```

Sind N Messwerte x_i nach einer Gauss-Verteilung $f(x; \mu, \sigma)$ verteilt, so erwartet man ein Anteil Np mit (Gln. 1.48, 1.49, Abb. 1.7) Sind N Messwerte x_i nach einer Gauss-Verteilung $f(x; \mu, \sigma)$ verteilt, so erwartet man ein Anteil Np mit (Gln. 1.48, 1.49, Abb. 1.7)

$$\begin{aligned} p &= F(\mu + \sigma) - F(\mu - \sigma) = 2[F(\mu + \sigma) - F(0)] \\ &= \operatorname{erf}(1/\sqrt{2}) = 68.3\% \simeq \frac{2}{3} \end{aligned} \quad (1.51)$$

im 1σ -Konfidenzintervall ($\mu - \sigma, \mu + \sigma$). Im 3σ -Intervall sind es schon $\operatorname{erf}(3/\sqrt{2}) = 99.7\%$. Bei einer einzelnen Messung ist also eine 3σ -Abweichung vom Erwartungswert sehr unwahrscheinlich.

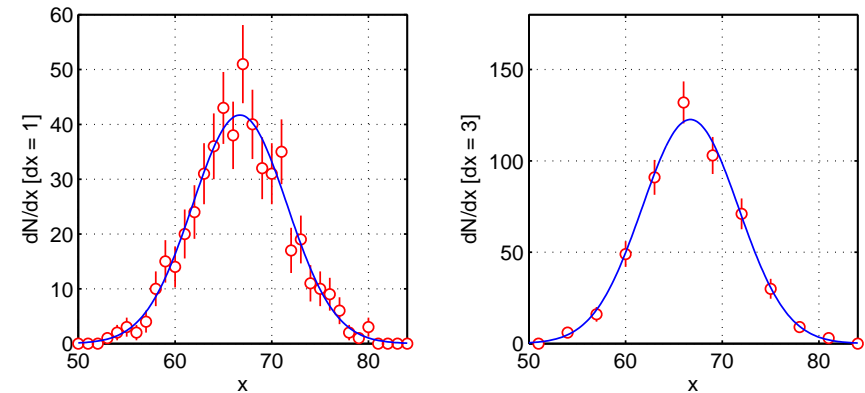


Abbildung 1.8: Gemessene Daten (o; Binbreite 1 und 3), die einer Gauss-Verteilung (—) entsprechen.

In der Datenanalyse ist das Problem umgekehrt. Man hat einen oder mehrere Werte x_i mit einem Instrument mit bekannter Auflösung σ gemessen und möchte eine Aussage über den “wahren” Wert μ machen. Da die Gauss-Verteilung symmetrisch in x und μ ist, gilt für einen Messwert x_0 , dass mit 68.3 % Wahrscheinlichkeit der “wahren” Wert μ im Intervall $(x_0 - \sigma, x_0 + \sigma)$ liegt⁵. Für N Messwerte gilt, dass mit

⁵ Dies ist die Aussage der Bayessche Schule, die über eine Wahrscheinlichkeitsverteilung der “wahren” Wert μ redet. Die “klassische” Schule sagt: es gibt eine festen “wahren” Wert μ und dieser liegt entweder im Intervall (100 %) oder nicht (0 %). Die richtige Aussage nach der “klassische” Schule ist: Mit 68.3 % Wahrscheinlichkeit schliesst das Intervall den “wahren” Wert μ ein. Das Intervall ist variabel, nicht μ . Für eine ausführliche Diskussion siehe [Eadi 71] und [Barl 89]. Physiker denken “klassisch”, reden aber oft wie Bayesianen.

68.3 % Wahrscheinlichkeit der “wahren” Wert μ im Intervall $(\bar{x} - \sigma_x, \bar{x} + \sigma_x)$ liegt, wobei $\bar{x} = \sum x_i/N$ der Mittelwert ist und $\sigma_x = \sigma/\sqrt{N}$ dessen Fehler (siehe Kap. 1.3.4). Werden die N Messwerte mit ihren Fehlern grafisch dargestellt, erwarten wir, dass $2/3 N$ Messpunkte mit ihren Fehlerbalken den “wahren” Wert überlappen (siehe z.B. Abb. 1.5, 5 von 9, bzw, 5 von 5).

In Abb. 1.8 ist sind gemessene Verteilungen mit Gauss-Verteilungen verglichen. Die primären Daten sind 511 Zahlen (Zahl der Ereignisse in einem Detektor, gemessen mit einer Elektronik, die eine Totzeit hat, siehe Kap. 2.3.1), die einmal mit Binbreite 1 und einmal mit Binbreite 3 histogrammiert werden (siehe Programm 1.9. Man beachte, dass ganzzahlige Werte histogrammiert werden). Mittelwert und Varianz beider Verteilungen sind praktisch gleich ($\bar{x}_1 = \bar{x}_3 = 66.7$, $\sigma_1 = 4.9$, $\sigma_3 = 5.0$). Bei gemessenen Verteilungen wird statt σ oft die volle Breite bei halber Höhe angegeben FWHM (“Full Width at Half Maximum”). Für eine Gaussverteilung ist die Beziehung zwischen beiden Grössen

$$\boxed{\text{FWHM} = 2.35\sigma}. \quad (1.52)$$

Die Halbwertsbreite kann man aus der Grafik abschätzen ($\text{FWHM} \simeq 12$) und damit den Fehler des Schwerpunkts ($\sigma_x = \sigma/\sqrt{N} = \text{FWHM}/(2.35\sqrt{511}) \simeq 0.23$; Programm 1.9 gibt 0.22).

```
%-----
%   MATLAB program 1.9: measured Gaussian
%   -----
load gaussdata
bin = 1;                               x1 = 48:bin:84;
subplot(121)
y1 = hist(gaussdata,x1);               plot_data([x1' y1'])
axis([50 84 0 60]),                   ylabel('dN/dx [dx = 1]')
xlabel('x'),                           hold on
out = centroid(x1,y1),                 s = out(2);
xm = out(1);                           plot([50:0.2:84],sum(y1)*bin*gauss([50:0.2:84],xm,s),'b-')
grid,                                   axis square
bin = 3;                               x3 = 48:bin:84;
subplot(122)
y3 = hist(gaussdata,x3);               plot_data([x3' y3'])
axis([50 84 0 180]),                   ylabel('dN/dx [dx = 3]')
xlabel('x'),                           hold on
out = centroid(x3,y3),
```

```
plot([50:0.2:84],sum(y3)*bin*gauss([50:0.2:84],out(1),out(2)),'b-')
grid,                                   axis square
```

```
%-----
%   out =    66.7006    4.8880    0.2162
%   out =    66.6986    4.9856    0.2206
%-----
```