Statistical Data Analysis

E. de Wolf

september 2009

Contents

1 Introduction

1.1 Practical Course Information

Teachers

E. de Wolf (tel:5925123, e-mail: e.dewolf@nikhef.nl or e.dewolf@uva.nl) and T. Ketel (tel:5922118, e-mail:tjeerd@nikhef.nl)

Place and time

The course will take place in room H239 at NIKHEF on Monday from 13:30h until 16:30h on the following dates:

August					31
September	7	14	21	28	
October	5	12		26	
November	2	9	16	23	30
December	7				

Date	Subject	Take home exercises	Programming exercises
31/08	ch1-Introduction ch2-Probability	2.1-2.8	ROOT 1: Introduction to ROOT
07/09	ch3-Random variables	3.1-3.8	ROOT 2: propagation of errors
14/09			ROOT 2
21/09	ch4-Distributions	4.1-4.7	ROOT 3: law of large numbers
28/09	ch5-Gaussian distribution		ROOT 4: central limit theorem
05/10			ROOT 3,4
12/10	ch6-Monte Carlo simulation		ROOT 5: simulating a neutrino beam
26/10			ROOT 5
02/11	ch7-Parameter Estimation ch8-Maximum Likelihood method	7.1-7.2	ROOT 6: estimating lifetime of a particle
09/11			ROOT 6
16/11	ch9-Least Square method	9.1-9.2	ROOT 7: fitting a curve ROOT 8: fitting a curve with MINUIT
23/11			ROOT 7,8
30/11			ROOT 9: simulating a gas
07/12	written exam		

Assignments

The course is a hands-on and mainly self-tuition course. Assistance is present on the scheduled course hours. Presence on these hours is mandatory. Hand in assignments according to the following schedule:

Date	Assignment
07/09	Programming exercise ROOT 1 Take home exercises 2.1-2.8
14/09	Take home exercises 3.1-3.8
21/09	Programming exercise ROOT 2
28/09	Take home exercises 4.1-4.7
12/10	Programming exercise ROOT 3 Programming exercise ROOT 4
02/11	Programming exercise ROOT 5
09/11	Take home exercises 7.1-7.2
16/11	Programming exercise ROOT 6
23/11	Take home exercises 9.1-9.2
30/11	Programming exercise ROOT 7 Programming exercise ROOT 8
07/12	Programming exercise ROOT 9 Exam

Assessment

Assessment will be based on the take-home exercises, the programming exercises and a written exam with weights 0.2, 0.3 and 0.5 respectively. Take-home are not marked, but must be handed in according schedule. Programming exercise are marked good, sufficient, fair and must be handed in according schedule.

Access to the written exam will be granted only after the all exercises are handed in.

Required knowledge and experience

Experience with programming C or C++ in a UNIX/LINUX-environment is required.
At *http://www.cplusplus.com/doc/#tutorial* you can find a tutorial for the C++ programming language.

Working environment

Many exercises require graphical presentation of results. Moreover the amount of (programming) work can be strongly reduced by using program-parts (functions) which are available on mathematical libraries. A suitable programming environment is offered by the package called *root*. We will learn you how to work with it; it requires (a little) knowledge of C^{++} .

Reference material

S.Brandt and G.Cowan: "Statistical and Computational Methods for Scientists and Engineers", Springer Verlag, 1998. To follow the course acquisition of the book is not required.

1.2 What is it all about?

Many phenomena in physics are 'of a statistical nature': quantum-physics, statistical physics, ... Generally speaking, measurement of these phenomena does not have one single possible outcome, but there are several possible outcomes, each with its own probability. Measurements are 'disturbed' by statistical fluctuations, i.e. the measured value deviates from the true one; the magnitude of the deviation is described by a probability distribution. How can we deal with this in an accountable way and which tools and methods are available? Hereafter we will present a few examples.

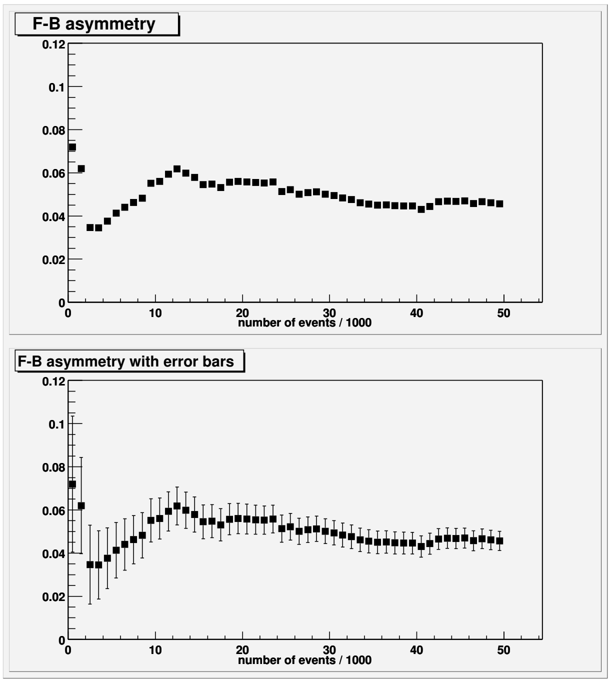


Figure 1: Measured forward-backward asymmetry versus number of observations; without (upper figure) and with errors indicated (lower figure).

An asymmetry-measurement

Suppose, in an experiment beams of electrons and positrons are made to collide. The momentum of the electrons is $p_x = p_y = 0.$, $p_z = 45.0$ Gev/c and the momentum of the positrons is $p_x = p_y = 0.$, $p_z = -45.0$ Gev/c. From all measured collisions we select the final state $\mu^+\mu^-$ and compare the p_z of the μ^+ with the p_z of the e^+: when both p_z's have the same sign, we say that the muon has been produced in the *forward* direction; when they have opposite signs, we say that the muon has been produced in the *backward* direction.

Theory predicts a *forward-backward asymmetry*, and we want to verify what our experiment tells us about it. We plot the measured asymmetry $(N_F - N_B)/N$ as function of the total number of observations N, where N_F is the observed number of forward going muons and N_B the observed number of backward going muons. The experimental result is shown in figure 1. We observe that for a very large number of observations N, the measured asymmetry becomes constant, and there is little doubt about its value. For a small number of observations, the measured value can be different and fluctuates. Can we still draw a conclusion, even when our number of observations is (relatively) small? For that we have to introduce a statistical uncertainty (error) in order to indicate how far our measured value might deviate from the true value. We might ask ourselves other questions: when could one decide that there is indeed an asymmetry? How many observations are needed in order to measure the value of the asymmetry with an accuracy of 1%?

A momentum measurement

We use (again) the final state $\mu^+\mu^-$ in electron-positron collisions, and look at the measured values of the muon momentum. From conservation of energy we know that the value should be equal to 45.0 GeV/c. Since our detector is placed in a magnetic field, the muons follow a circular path. The radius of the circle is a measure for the momentum. In figure 2 we present the results of *two series* of measurements. In the first series (upper plot), the circular path of the muon is followed over a length of 2 meter; in the second series, the circular path is followed over a length of (only) 0.2 meter. Note that, in both series, the measured momentum, although being centred around the expected value, follows a distribution. Apparently the measured value can deviate from the true one by a statistical fluctuation. The width of the distributions is different between the two series. Apparently the statistical fluctuation becomes larger if the path length is smaller. If we use the measured value of the momentum as an estimate of the true value, and that is the only thing we can do, this value must be accompanied by another number, which indicates how far the measured value can deviate from the true one: the statistical uncertainty/error. We can ask the question how large the statistical error is in both series.

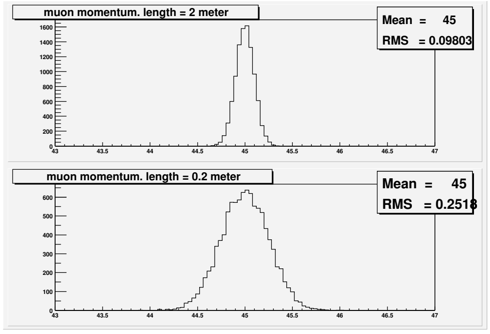


Figure 2: Distributions of measured momentum values

Mass and width of the Z^0

In an other experiment we have measured the total cross section σ of electron-positron collisions, at different values of the collision energy E. The E-values lie around the mass of the Z^0, the 'carrier' of the neutral component of the weak interaction. As a consequence, the total cross section shows a 'resonance' behaviour (see figure 3). We use the experimental result to determine the mass and the width of the Z^0-particle. A (simplified) theoretical prediction for the shape of our result is

$$\sigma(E) = \frac{C}{(s - M_z^2)^2 + s^2 \frac{\Gamma_z^2}{M_z^2}} \tag{1}$$

where $s = E^2$, M_z is the mass of the Z^0 and Γ_z is the decay-width of the Z^0. The questions we may ask ourselves are: What (according to our experiment) are the values of M_z and Γ_z? Since our

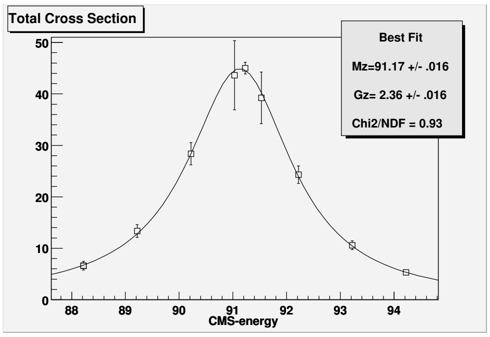


Figure 3: Total cross-section around the Z-mass

measured values of the cross-section have statistical uncertainties, indicated by the vertical errorbars, they differ from point to point. What is the effect of these uncertainties on the values found for M_z and Γ_z? Does the shape of the distribution, predicted by theory, agree with our experimental result?

Simultaneous measurement of two quantities

In yet an other experiment we have measured the total cross-sections of neutrino's and antineutrino's with nucleons (protons and neutrons). From the results we have derived the values of two so-called coupling constants: u_l^2 and d_l^2. We publish our results in an article showing figure 4 with the the following text in the caption:

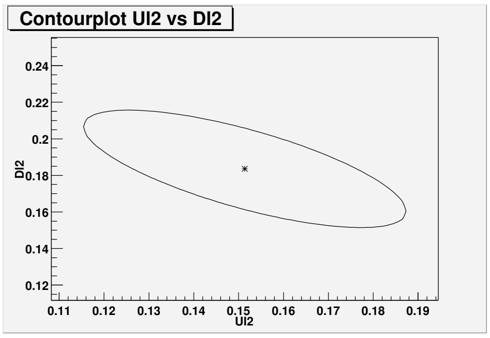


Figure 4: Values of u_l^2 and d_l^2

Two-parameter solution for (u_l^2, d_l^2) The ellipse is the allowed region at 39.3 % C.L. (one standard deviation on each variable separately.)

What is the meaning of this all? When you have finished the course you should be able to answer the questions posed.

1.3 Exercises

Exercise 1.1 - Introduction to *root*

The exercises will frequently require presentation of results in graphical form (histograms, scatter-plots, graphs etc.), the use of random number generators and mathematical operations on vectors and matrices. *root* is a program package/working environment, developed at CERN, which includes many of the desired facilities. So we propose to work on the exercises 'under *root*'.
You can find a (very) extensive description of *root* in: *http://root.cern.ch* This includes: a users guide, a reference guide (with search facility), several tutorials and examples, a reference to C^{++}. The *root*-material on the web site is very extensive. In order to avoid that you have to read it all, we will summarise everything which is needed for the exercises in this section (and sometimes in the exercise itself).

Getting access to *root*

To get access to the *root* package add to the *.cshrc* file in your home directory the following lines:

```
setenv ROOTSYS /public/cern/ROOT/v4.00.06a/rh73_gcc296/root
set path = ( $path $ROOTSYS/bin )
setenv LD_LIBRARY_PATH $ROOTSYS/lib
```

These lines specify where the *root* package is installed at your computer. After this, you can simply start up the *root* package by typing: **root**
You will get a welcome message, and a prompt for your first command which looks like : **root[0]**

Commands to *root*

- Most commands to *root* start with a **.** (a 'dot')
- A (simple) command is the one to stop your *root*-session:
  .quit
- The command to compile a file containing your program code, and make an executable is:
  .L filename
- UNIX/LINUX commands can be given 'under *root*' if preceded by **.!**
  - **.! lpr -Ph238single voorbeeld.ps** to print the postscript file voorbeeld.ps
  - **.! enscript -Ph238single voorbeeld.C** to print program code in file voorbeeld.C

Working with *root*

- Each exercise requires that you write a computer program to solve the problem.
- *root* requires that the name of the file which contains your program ends with **.C**
- *root* expects that your program is written in **C**$^{++}$.
  However : everything you learned about (simple) **C** during the course on "Inleiding Numerieke Natuurkunde" is accepted by the C^{++} compiler. So: you could write simply in C.

- **BUT** : as soon as you want to call functions which are provided by the *root*-package, you are confronted with the fact that these functions have been written in $\mathbf{C}^{++}$: this has consequences for the way these funcions have to be called.
- As you (might) remember, an (independent) C-program should in any case contain a 'main' program (a function called 'main'):

  ```
  int main ()
  {
      return 0 ;
  }
  ```

  Under *root* this is no longer compulsory: your 'main' program can have any name, like in:

  ```
  int voorbeeld ()
  ```

- The name you give becomes relevant if you want to execute your program execution under *root*:
  - Suppose you have a file with *name* **example.C**, which contains :

    ```
    int voorbeeld ()
    {
        return 0 ;
    }
    ```

  - To compile the file with your C++ code (under *root*), you type:
    .L example.C
  - To *execute* the program (under *root*), you type:
    voorbeeld() (note : *without* a **.**) i.e. you type the name of your program including ().

An example program with graphics output (histograms)

Its time now for an example program.

Execution of an example program

To try out execution of an example program:

- First prepare the environment for this exercise:
  - make a subdirectory *exercises1.1* in you home directory by typing the command **mkdir exercise1.1**.
  - go to this subdirectory by typing **cd exercise1.1**. This is now your working directory for this exercise.
- Copy the file */user/uvak/sda/example1.C* to you working directory by typing *cp /user/uvak/sda/example1.C* . (note the dot). Note that the name of the function in this file is **Voorbeeld1**
- Start *root* by typing **root**.

- Compile by typing **.L example1.C**
- Execute by typing **Voorbeeld1()** and see what happens.
- *See* what happens!

Code of the example program

The code of the example program is as follows. Note that there are many useful/relevant inline comments. In addition some parts of it will be explained in the next section.

```
/*                                                                         */
/*  This is the code of /user/uvak/sda/example1.C                      */
/*      The program                                                        */
/*          - generates two sets of random numbers uniformly               */
/*            distributed between 0 and 1                                  */
/*          - makes of each set of random numbers a histogram              */
/*          - makes a scatterplot of the two sets of numbers against       */
/*            each other                                                   */
/*                                                                         */
/*      Give your  'main program' a name                                   */
/*  This name has to be used to execute the  program under ROOT        */
/*                                                                         */
int Voorbeeld1()
{
/*      The following 'lines' are necessary in order to 'reset' some       */
/*      quantities in ROOT to their 'original' values                      */
/*                                                                         */
    gROOT  ->Reset()  ;
    gRandom->SetSeed();

/*                                                                         */
/*      Here starts our program                                            */
/*      type declarations under ROOT :                                     */
/*           integers     are declared with Int_t                          */
/*           real numbers are declared with Float_t                        */
/*           real numbers (double precision) Double_t                      */
/*                                                                         */
#   define nMax 10000
/*                                                                         */
/*      Initialise two histograms (TH1F) and one scatterplot (TH2F)        */
/*                                                                         */
    TH1F *xUniform  = new TH1F(" ","the x distribution",100, 0., 1.);
    xUniform -> GetXaxis() -> SetTitle ("X axis title");
    xUniform -> GetYaxis() -> SetTitle ("number of events/0.01");

    TH1F *yUniform  = new TH1F(" ","the y distribution",100, 0., 1.);
    yUniform -> GetXaxis() -> SetTitle ("X axis title");
    yUniform -> GetYaxis() -> SetTitle ("Number of events/0.01");

    TH2F *xyUniform = new TH2F(" ","x versus y",50, 0., 1., 50, 0., 1.);
    xyUniform -> GetXaxis() -> SetTitle ("xUniform");
    xyUniform -> GetYaxis() -> SetTitle ("yUniform");

/*                                                                          */
/*      Fill two vectors with nMax random numbers from a uniform            */
/*      distribution between 0 and 1 provided by the ROOT-function Rndm()   */
/*      and fill the histograms and the scatterplot using these two vectors */
/*                                                                          */
    Double_t y1[nMax], y2[nMax] ;
    for (Int_t i=0;i<nMax;i++)
    {
       y1[i] = gRandom -> Rndm (1) ;
       y2[i] = gRandom -> Rndm (1) ;
       xUniform ->Fill(y1[i])      ;
       yUniform ->Fill(y2[i])      ;
       xyUniform->Fill(y1[i],y2[i]);
    }

/*                                                                          */
/*      Make histograms and scatterplot visible on computer screen:         */
/*        - open a window (TCanvas)                                         */
/*        - divide the canvas into 2x2 areas (Divide (2,2))                 */
/*        - draw the histograms and the scatterplot in these areas          */
/*          ('go to' an area with 'cd()' and Draw() the histograms)         */
/*                                                                          */
    TCanvas *exam = new TCanvas ("example1", "exercise1", 1) ;
    exam -> Divide (2,2) ;
    exam -> cd(1) ;
    xUniform ->Draw();
    exam -> cd(2) ;
    yUniform ->Draw();
    exam -> cd(3) ;
    xyUniform->Draw();
    exam -> cd(4) ;

/*                                                                          */
/*      Saving the plot results in a file (type .ps or .eps). This          */
/*      can be achieved through a pull-down menu (under 'File')             */
/*      in the Canvas-window                                                */
/*      The .ps file can be printed with lpr -Ph238single 'filename'        */
/*                                                                          */

    return 0 ;
}
```

Initialisation of a histogram

With the command

```
TH1F *xuni  = new TH1F("     ","the xuni distribution",100, 0., 1.);
```

you order *root* to set up provisions for a histogram and you specify its properties in a parameter-list.The meaning of the parameters is :

`"     "`	not relevant for us
`"the xuni distribution"`	gives the title of the histogram
100	specifies the number of intervals/bins
0.	specifies the lower limit
1.	specifies the upper limit

In *root* (C^{++}) a histogram is treated as a *type of a variable* (just like int, float, etc. in ordinary C).

- in the example this histogram(variable) gets the *name* `xuni`
- this name will be used later in the program to *fill* the histogram with quantities, and to *draw* the histogram.

Initialisation of a scatterplot

With the command

```
TH2F *xyuni  = new TH2F("     ","xuni vs yuni",50, 0., 1., 50, 0., 1.);
```

you order *root* to set up provisions for a scatterplot and you specify its properties in a parameter-list. The meaning of the parameters is :

`"     "`	not relevant for us
`"xuni vs yuni"`	gives the title of the scatterplot
50	specifies the number of intervals/bins on the horizontal axis
0.	specifies the lower limit on the horizontal axis
1.	specifies the upper limit on the horizontal axis
50	specifies the number of intervals/bins on the vertical axis
0.	specifies the lower limit on the vertical axis
1.	specifies the upper limit on the vertical axis

In *root* (C^{++}) a scatterplot is treated as a *type of a variable* (just like int, float, etc. in ordinary C). In the example this scatterplot(variable) gets the name `xyuni`. This name will be used later in the program to fill the scatterplot with quantities and to draw it.

Filling histograms and scatter plots

With the command

```
xuni -> Fill (y1) ;
```

you order *root* to accept the quantity whose value is stored in the variable 'y1' as an entry to histogram 'xuni'
With the command

```
xyuni -> Fill (y1, y2) ;
```

you order *root* to accept the quantities whose values are stored in the variables 'y1' and 'y2' as an entry to the scatterplot 'xyuni'

Displaying the graphics results

With the command

```
TCanvas *Vb1 = new TCanvas ("Voorbeeld1", "    ", 1) ;
```

you order *root* to open a window, and you specify its properties through the parameter-list. In this case only the first parameter is relevant `("Voorbeeld1")` ; if you save the content of the window in a file, this will become (part of the) filename.
In *root* (C^{++}) a window is treated as a type of a variable (just like int, float, etc. in ordinary C). In the example the window(variable) gets the name `Vb1` . This name is used when we divide the window in 'blocks', in

```
Vb1 -> Divide (2,2) ;
```

or when we assign a block to a histogram/scatterplot, in

```
Vb1 -> cd (1) ;
```

A histogram or scatterplot is drawn (in the assigned box) through

```
xuni -> Draw () ;
```

Note that `xuni` reperesents the *variable name* we gave to the corresponding histogram.

Saving and printing graphics results

Saving the contents of a window for later viewing or printing can be done through the windows pulldown-menu, under `File`, choosing the option `Save as Canvas.ps`.
The file can be *printed* under *root* with:
.! lpr -Ph238single *filename*

Exercise 1.2 - *root* exercise 1

This exercise should make you acquainted with the *root*-package and some of its (histogram) facilities which will be used frequently. It also introduces a random number generator, the standard-normal distribution and some quantities which characterise a probability distribution, like mean value, variance, skewness, covariance, correlation coefficient and will be introduced in the next chapter in more detail. Moreover, you will learn how to make a printed version of your program code and the results of your program. The example program in */user/uvak/sda/example1.C* can be used as starting point.

First make a subdirectory 'exercise1.2' in your home directory. Then go to this subdirectory and copy the example program to this subdirectory. Start your preferred editor to edit the example program.

To generate random numbers you can use the *root* function *Rannor*. The statements

```
Double_t a, b           ;
gRandom -> Rannor(a,b)  ;
```

return values in the variables a and b. In the user manual of *root* it is claimed that these numbers are independent drawings from the standard-normal distribution. In this exercise we will verify this claim. The standard-normal distribution has the following features:

- its *mean* is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = 0.$$

- its *variance* is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = 1.$$

- its *skewness* is

$$\gamma_1 = \frac{1}{\sigma^3 N}\sum_{i=1}^{N}(x_i - \mu)^3 = 0.$$

To verify whether the *Rannor* results agree with these values, we proceed as follows:

- Make 1000 calls to 'Rannor(a,b)'
- Fill an array xnor(1000) with the values of 'a'.
  Fill an array ynor(1000) with the values of 'b'.
- Calculate for xnor and for ynor the quantities μ, σ^2 and γ_1.
- Print these values, and compare them with the expected ones.

To check if the two sets of drawings are indeed independent we use their covariance and correlation coefficient, which are defined as follows:

- The covariance of two sequences x_i and y_i is defined as:

  $$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

  if x and y are independent, then $cov(x,y) = 0$.

- Their correlation coefficient is defined as :

  $$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

  if x and y are independent, then $\rho(x,y) = 0$.

- Calculate the correlation coefficient of `xnor` and `ynor`. Print the result, and compare with the expected value.

- Make histograms of `xnor` and `ynor`

- Make a scatter plot of `xnor` versus `ynor`

- Are `xnor` and `ynor` indeed independent variables?

- Print your program code in file '*filename.C*' using the Unix-command

  ```
  'enscript -Ph238single filename.C'
  ```

  or alternatively use the root-command

  ```
  '.! enscript -Ph238single filename.C'
  ```

- Print your calculated results as follows:

  - First use the root-command

    ```
    '.x filename.C >> printout'
    ```

    to execute your program and store all print out in the file '*printout*'. This command only works if the name of the file is identical to the name of the function.

  - Then send the file 'printout' to the printer as described above.

HAND IN

For this exercise hand in:

- A print of your program code; make sure that you added your name to the code; make sure that you added sufficient relevant comments to your code for others to understand your program.

- A print of the values found for μ, σ^2, γ_1 and $\rho(x,y)$.

- A print of the histograms and the scatter plot. Make sure that the histograms have relevant (axis-)titles.

2 The concept of probability

Repeated measurement of one or more quantities generally results in a series of different values for each quantity. The differences are caused by the statistical nature of the underlying physical process or by inaccuracies in the measuring device. We can represent the results of the measurements in a table, a histogram or a scatter plot. As an example we take the interaction, mentioned in the previous chapter $e^+ \, e^- \; \rightarrow \; \mu^+ \, \mu^-$ and measure for each interaction the momentum p of the μ^+ and the angle θ between the outgoing μ^+ and the incoming e^+. The results of the measurements are presented in figure 5. We use figure 5 to introduce some definitions and the basic laws of probability.

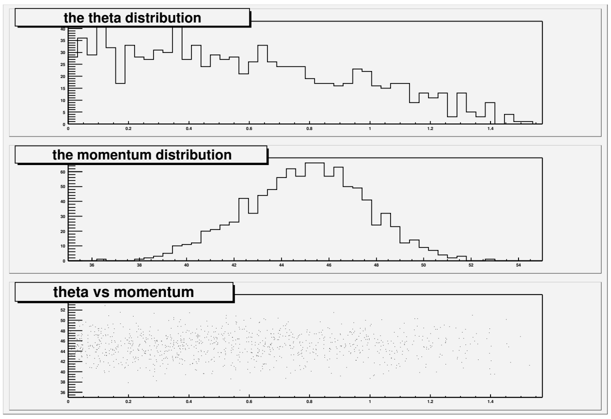


Figure 5: Distributions of measured values of momentum and angle

Define:

event A	the measured value of θ lies between 0.4 and 0.6
event not-A	the measured value of θ *does not* lie between 0.4 and 0.6
event B	the measured value of θ lies between 0.8 and 1.0
event C	the measured value of p lies between 42.5 and 47.5
event (A or C)	the measured value of θ lies between 0.4 and 0.6 **OR**
	the measured value of p lies between 42.5 and 47.5
event (A and C)	the measured value of θ lies between 0.4 and 0.6 **AND**
	the measured value of p lies between 42.5 and 47.5

We then define the observed frequency of event A as follows:

- The total number of measured muons is N
- The number of muons with measured θ between 0.4 and 0.6 is n
- The observed frequency of event A is then defined as:

$$freq(A) = \frac{n}{N} \tag{2}$$

The probability of event A is defined as:

$$P(A) = \lim_{N \to \infty} freq(A) = \lim_{N \to \infty} \frac{n}{N} \tag{3}$$

It is the frequency we would have observed, if the experiment would have been carried on indefinitely. As a consequence of this definition we have :

$$0. \ \leq \ P(A) \ \leq \ 1. \qquad and \qquad P(A) + P(notA) = 1. \tag{4}$$

If event A and event B are mutually exclusive we have :

$$P(A \ \cup \ B) = P(A) + P(B) \tag{5}$$

If event A and event C are not mutually exclusive:

$$P(A \ \cup \ C) = P(A) + P(C) - P(A \ \cap \ C) \tag{6}$$

Suppose that we want to know the probability that the measured value of p lies between 42.5 and 47.5 (event C) if the measured value of θ lies between 0.4 and 0.6 (event A). This conditional probability is defined as:

$$P(C|A) \ = \ \frac{P(A \ \cap \ C)}{P(A)} \tag{7}$$

Bayes theorem connects $P(C|A)$ with $P(A|C)$:

$$P(A|C) \ = \ P(C|A) \ \cdot \ \frac{P(A)}{P(C)} \tag{8}$$

Event A and event C are said to be independent if $P(A|C) = P(A)$. We then have:

$$P(A \ \cap \ C) \ = \ P(A) \ \cdot \ P(C) \tag{9}$$

An application: trigger efficiency

As an example for application of the probability laws, we consider (again) a collision experiment between beams of electrons and positrons. The detector which registers and measures for each collision (which in particle physics is also called an 'event') the collisions products, does not only 'see' the real e^+e^--collision products (the signal), but also the products of (background) processes, like cosmic-rays and interactions with residual gas-molecules in the beam-pipe. The signal events have to be sent to a mass-storage device, for later processing; the background-events have to be rejected. The decision whether a particular event is signal or background, is taken by a trigger-system. A trigger system, however, is rarely 100 percent efficient: there is a small probability that some signal events are lost in a random way. Therefore it is common practice to have several trigger systems working simultaneously to classify a particular event as signal, as soon as it is classified as such by at least one of the trigger systems and to send the data together with the decisions of each of the trigger systems to mass-storage.

Now, suppose that we have two independent trigger systems. We want to calculate the combined efficiency of these, together with the true number of signal events. The procedure is as follows: if C is the number of events registered by both trigger systems, D_1 the number of events registered by system 1 only and D_2 is the number of events registered by system 2 only we have the following conditional probabilities using the independence of the trigger systems:

$$P(1) = P(1|2) = \frac{C}{C + D_2} \quad and \quad P(2) = P(2|1) = \frac{C}{C + D_1} \tag{10}$$

Using the sum law and -again- the independence, we can calculate the combined efficiency:

$$P(1or2) = P(1) + P(2) - P(1and2) = P(1) + P(2) - P(1) \cdot P(2) \tag{11}$$

Substituting our experimental values, we get for the combined efficiency of the two independent trigger systems:

$$P(1or2) = \frac{C \cdot (C + D_1 + D_2)}{(C + D_1) \cdot (C + D_2)} \tag{12}$$

The total number of signal events (N) can now be calculated:

$$P(1) = \frac{C + D_1}{N} \longrightarrow N = \frac{(C + D_1) \cdot (C + D_2)}{C} \tag{13}$$

2.1 Exercises

Exercise 2.1

Given $P(A) = 1/3$, $P(B) = 1/4$ and $P(A \cap B) = 1/6$ find the following probabilities:

a. $P(notA)$
b. $P(notA \cup B)$
c. $P(A \cup notB)$
d. $P(notA \cap notB)$
e. $P(notA \cup notB)$

Exercise 2.2

Given $P(A) = 3/4$ and $P(B) = 3/8$.
a. Show that $P(A \cup B) \geq 3/4$
b. Show that $1/8 \leq P(A \cap B) \leq 3/8$
c. Give inequalities analoguous to a) and b) for $P(A) = 1/3$ and $P(B) = 1/4$.

Exercise 2.3

Five coins are tossed simultaneously. Find the probability of the event A: "at least one head turns up". Assume that the coins are fair.

Exercise 2.4

In producing screws, event A means "screw too slim" and event B means "screw too short". The conditional probability that a slim screw is also too short is $P(B|A) = 0.2$. What is the probability that a screw picked randomly from the produced lot will be both too slim and too short?

Exercise 2.5

A beam of particles consists of a fraction 10^{-4} electrons and the rest fotons. The particles pass through a double-layered detector which gives signals in either zero, one or both layers. The probabilities of these outcomes for electrons (e) and fotons (γ) are:

$$\begin{array}{ll} P(0|e) = 0.001 & P(0|\gamma) = 0.998 \\ P(1|e) = 0.01 & P(1|\gamma) = 0.001 \\ P(2|e) = 0.989 & P(2|\gamma) = 10^{-5} \end{array}$$

a. What is the probability for a particle detected in one layer only to be a photon?

b. What is the probability for a particle detected in both layers to be an electron?

Exercise 2.6

A test for a disease makes a correct diagnosis with probability 95 %. Given that the test for a person is positive, what is the probability that the person really has the disease? Assume that one in every 2000 persons, on average, has the disease.

Exercise 2.7

You are in a game show and you have to choose one of three closed doors, A, B and C. One conceals a car, two conceal goats. You choose A, but A is not opened immediately. Instead, the presenter opens door B to reveal a goat. He offers you the opportunity to change your choice to C. What is the probability to find a car behind the door C?

Exercise 2.8

An imaginary particle can decay in 0, 1 or 2 particles of the same kind with probabilities 1/4, 1/2 and 1/4, respectively. Beginning with one particle, we denote x_i the number of particles in the ith generation. Determine:

a. $P(x_2 > 0)$

b. The probability that $x_1 = 2$, given that $x_2 = 1$.

3 Random variables and their distributions

Repeated measurement of a single quantity which is influenced by random fluctuations, does not produce a single value, but a series of values. The result of a single measurement is a random variable, a random drawing from a probability distribution; the value of such a variable can not be predicted exactly. The set of possible outcomes of the measurement form the sample space, which, by definition, contains only unique values. E.g. possible outcomes for throwing a die are $1, 2, 3, 4, 5, 6$. This is the sample space for throwing a die. The result of the sampling of the parent population is a set of measured values, the sample. The sample may contain identical values: for example, after throwing a die three times, the sample may consist of $3, 6, 3$. The result of the sampling is determined by both the sample space and the probability distribution of the parent population. For throwing a die the probability distribution is $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$. The values in the sample can be represented in a histogram in order to visualise how frequently a particular value falls within a certain range/interval.

The probability distribution of the population is called discrete if the number of possible outcomes is finite; it is called continuous if the outcome can be any number (within a certain range). The distribution is called a probability density distribution or probability density function, abbreviated as *pdf*. In this chapter we will concentrate on continuous random variables and their distributions.

3.1 Probability density distributions of one variable

Consider a random variable x, distributed according to the pdf $f(x)$; $f(x)$ is the probability that the outcome of a measurement is within the range $[x, x + dx]$. The pdf $f(x)$ is normalised such that the total probability is one. $f(x)$ has to satisfy:

$$f(x) \geq 0. \quad \forall x \qquad and \qquad \int_{-\infty}^{\infty} f(x)dx = 1 \tag{14}$$

We then define the *expectation value* of x or the *mean* of x or the *population mean*, as:

$$E(x) =< x >= \int_{-\infty}^{\infty} xf(x)dx \tag{15}$$

$E[x]$ is not a function of x, but depends on the shape of the pdf $f(x)$. If $f(x)$ is concentrated mostly in one region, then $E[x]$ is a measure of where values of x are likely to be observed. However, if $f(x)$ consists of two widely separated peaks $E[x]$ is in the middle between the two peaks where x is seldomly observed.

We define the *variance* of x or the *population variance* as:

$$var(x) = E[(x- < x >)^2] = E[x^2]- < x >^2= \int_{-\infty}^{\infty} \{x- < x >\}^2 f(x)dx \tag{16}$$

The variance is a measure of how widely x is spread around its mean value. The *standard deviation* of x is defined as:

$$\sigma(x) = \sqrt{var(x)} \tag{17}$$

The standard deviation has the same dimension as x. The *skewness* of x is defined as:

$$\gamma = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} \{x- < x >\}^3 f(x)dx \tag{18}$$

For symmetric distributions $\gamma = 0$.

The most probable values of a population, the *mode*, is defined as the value of x for which the pdf is maximum. A pdf can be multimodal. The values of x for which $\int_{-\infty}^{x} f(x)dx = 0.5$ is the *median* of the x.

3.1.1 Cumulative distribution function

The cumulative distribution function (*cdf*) is the probability that the value of a random variable x will be less or equal than a specific value:

$$F(x) = \int_{-\infty}^{x} f(x)dx \tag{19}$$

From this we see that $F(x)$ is monotone and not decreasing, $F(-\infty) = 0$ and $F(+\infty) = 1$ and $0 \leq F(x) \leq 1$. In figure 3.1.1 the pdf $f(x)$ and the cdf $F(x)$ are plotted as function of the random variable x. $F(x)$ represents the shaded area under the pdf to the left of the specified value x.

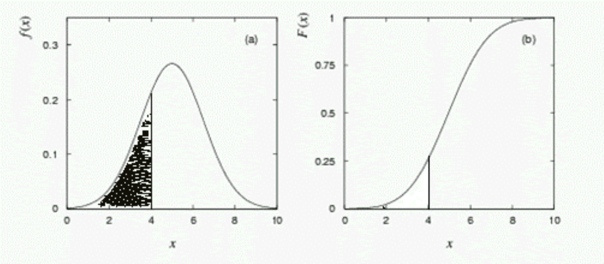


Figure 6: Example of a) a pdf $f(x)$ and b) the corresponding cdf $F(x)$.

3.1.2 The normalised distribution

From a given random variable x with pdf $f(x)$ we often want to obtain a new random variable of the form $y(x) = ax + b$. Then we must know the mean and variance of this new variable $y(x)$. Suppose x has mean $< x >$ and variance $\sigma^2(x)$, then $y(x)$ has mean $< y(x) >= a < x > +b$ and variance $\sigma^2(y(x)) = a^2\sigma^2(x)$. We can proof this as follows: assume the $h(y)$ is the pdf of the random variable y. To a small interval Δx on the x-axis corresponds the probability $f(x)\Delta x$. This probability is equal to the probability $h(y)\Delta yx$, where Δy is the corresponding interval on the y-axis. Thus for the mean of y is:

$$\begin{aligned}
E(y) &= < y >= \int_{-\infty}^{\infty} yh(y)dy \\
&= \int_{-\infty}^{\infty} (ax+b)f(x)dx \\
&= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx \\
&= a\int_{-\infty}^{\infty} xf(x)dx + b \\
&= a < x > +b
\end{aligned}$$

Since

$$y-<y>=(ax+b)-(a<x>+b)=ax-a<x> \tag{20}$$

the variance of y is:

$$\begin{aligned}\sigma^2(y) &= \int_{-\infty}^{\infty}(y-<y>)^2 h(y)dy \\ &= \int_{-\infty}^{\infty}(ax-a<x>)^2 f(x)dx \\ &= a^2\sigma^2(x)\end{aligned}$$

If we transform each measured value of a random variable x with pdf (x) into a new random variable $u = ax+b$ and with $a = \frac{1}{\sigma(x)}$ and $b = -\frac{<x>}{\sigma(x)}$:

$$u(x) = \frac{x-<x>}{\sigma(x)} \tag{21}$$

Then $u(x)$ has the expectation value and variance:

$$\begin{aligned}E[u] &= <u>=\frac{1}{\sigma(x)}(<x>-<x>)=0. \\ \sigma^2(u) &= \frac{1}{\sigma^2(x)}E[(x-<x>)^2]=\frac{\sigma^2(x)}{\sigma^2(x)}=1\end{aligned}$$

$u(x)$ is called a normalised variable. It is also a random variable and has simple properties which makes its use in more involved calculations preferable.

Suppose that we transform each measured value of a variable x with pdf $f(x)$ into another value $u(x)$ as follows:

$$u(x) = \frac{x-<x>}{\sigma(x)} \tag{22}$$

where $<x>$ and $\sigma(x)$ are the mean and the standard deviation of x.

3.1.3 The Chebychev inequality

The variance has a special significance for all distributions because of the Chebyshev's inequality. Suppose that we draw a value for the random variable x from a pdf $f(x)$, with mean $<x>$ and standard deviation σ. Then we consider the 'distance' $d = |x-<x>|$. Intuitively we 'feel' that "small distances will occur more frequently than large distances". This is formalised in Chebychev's inequality as follows:

$$P(|x-<x>|>k\sigma(x))<k^{-2} \tag{23}$$

The Chebychev inequality gives un upper limit on the probability of exceeding any given number of standard deviations, independent of the shape of the function, provided its variance is known. It makes a statement about the probability that a random variable x will be found at a greater distance from the mean than a given number of standard deviations $\sigma(x)$. Another way of phrasing this is that for any set of data at least $100(1-k^2)\%$ of the values are within k standard deviations of the mean. E.g. for any distribution is $P(|x-<x>|>2\sigma(x)) \leq \frac{1}{4}$. This asserts that any random variable x will assume a value further from the mean than two standard deviations with a probability less then $\frac{1}{4}$ or will assume a value closer to the mean then two standard deviations with a probability greater then $\frac{3}{4}$ or at least 75% of the values of x will be within two standards deviations from the mean. On the average it will be more than two standard deviations away from the mean less than $\frac{1}{4}$ of the time. The proof of Chebyschev's inequality goes as follows: consider the variable $t = (x- < x >)^2$ with pdf $g(t)$. The probability P that the distance from the mean d is larger than $k\sigma(x)$ is given by

$$P = P(|x- < x > | > k\sigma(x)) = P((x- < x >)^2 > k^2\sigma^2(x)) = \int_{k^2\sigma^2(x)}^{\infty} g(t)dt \tag{24}$$

We also have :

$$< t >= \sigma^2(x) = \int_0^{\infty} tg(t)dt = \int_0^{k^2\sigma^2(x)} tg(t)dt + \int_{k^2\sigma^2(x)}^{\infty} tg(t)dt \tag{25}$$

Since t and $g(t)$ are positive definite the value of each integral is larger then the one which is obtained by replacing the factor t in the integrand by the lower integration boundary. So :

$$\sigma^2(x) > 0 + k^2\sigma^2(x)\int_{k^2\sigma^2(x)}^{\infty} g(t)dt = k^2\sigma^2(x)P \tag{26}$$

Note that the inequality is valid independent of the 'shape' of $f(x)$. If $f(x)$ is known, a sharper and well defined limit can be calculated.

The Chebychev inequality helps us to interpret the standard deviation in an intuitive way. From the knowledge of the mean and the standard deviation we gain a general impression of where the major portion of a set of data is located and of how much variation there is in the data. A small standard deviation indicates that the values are clustered close to the mean. A large deviation indicates that the values are quite vaired about the mean. Chebychev's inequality shows that of all sets at least 88.9% of the values are within 3 standard deviations of the mean, at least 75.0% are within 2 standard deviations and at least 55.6% are within 1.5 standard deviation.

3.2 Probability density distributions of two variables

Now we consider the case of two random variables, with a joint pdf $f(x, y)$; $f(x, y)$ has to satisfy:

$$f(x, y) \geq 0. \quad \forall x, y \qquad and \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1 \tag{27}$$

We define the marginal distributions of x and y as:

$$g(x) = \int_{-\infty}^{\infty} f(x, y)dy \quad ; \quad h(y) = \int_{-\infty}^{\infty} f(x, y)dx \tag{28}$$

and say that x and y are independent if

$$f(x, y) = g(x)h(y) \tag{29}$$

The marginal distribution $g(x)$ is the probability that occurs no matter what the value of y is. The conditional pdf's are defined as:

$$\begin{aligned} f(y|x) &= \frac{f(x, y)}{g(x)} \\ f(x|y) &= \frac{f(x, y)}{h(y)} \\ g(x) &= \int_{-\infty}^{\infty} f(x|y)dy \\ h(y) &= \int_{-\infty}^{\infty} f(y|x)dx \end{aligned}$$

If x and y are independent then $f(y|x) = \frac{g(x)h(y)}{g(x)} = h(y)$.
We define the expectation values, variances and standard deviations of x and y as:

$$\begin{aligned}
E(x) &= <x>= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dxdy \\
E(y) &= <y>= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dxdy \\
var(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-<x>)^2 f(x,y)dxdy \\
var(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-<y>)^2 f(x,y)dxdy \\
\sigma(x) &= \sqrt{var(x)} \\
\sigma(y) &= \sqrt{var(y)}
\end{aligned}$$

A new quantity is the *covariance* between x and y:

$$\begin{aligned}
cov(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-<x>)(y-<y>)f(x,y)dxdy \\
&= <xy> - <x><y>
\end{aligned}$$

If x and y are independent, then $cov(x,y) = 0$.
The covariance is positive if values of x larger than $<x>$ preferentially appear with values of y larger than $<y>$. The covariance is negative if generally 'x is smaller than $<x>$' implies 'y is smaller than $<y>$'. The covariance vanishes if knowledge of the values of x does not give information about the value of y and v.v.
For both expectation value and covariance there is a sum rule: if a and b are constants, and x and y are random variables, then:

$$\begin{aligned}
E(ax+by) &= <ax+by>= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax+by)f(x,y)dxdy \\
&= a<x> +b<y>= aE(x) + bE(y) \\
\sigma^2(ax+by) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [(ax+by) - (<ax> + <by>)]^2 f(x,y)dxdy \\
&= a^2\sigma^2(x) + b^2\sigma^2(y) + 2ab\cdot cov(x,y)
\end{aligned}$$

With $\sigma^2(x) = var(x), \sigma^2(y) = var(y)$ and $a = b = 1$ we have:

$$var(x+y) = var(x) + var(y) + 2cov(x,y)$$

It is convenient to use the correlation coefficient between x and y which is defined as:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)} \tag{30}$$

If x and y are independent then $\rho(x,y) = 0$.
Consider two normalised variables $u(x)$ and $v(y)$. Then is

$$\begin{aligned}
var(u+v) &= var(u) + var(v) + 2\rho(u,v)\sigma(u)\sigma(v) \\
\sigma^2(u) &= \sigma^2(v) = 1 \\
\sigma^2(u+v) &= 2(1+\rho(u,v)) \\
\sigma^2(u-v) &= 2(1-\rho(u,v)) \\
(\sigma^2 &\geq 0) \rightarrow -1 \leq \rho(u,v) \leq 1 \rightarrow -1 \leq \rho(x,y) \leq 1
\end{aligned}$$

This is called the Schwartz lemma.
Now consider $\rho(u, v) = 1$; then is $\sigma^2(u - v) = 0$, i.e. $u - v$ is constant:

$$u(x) - v(y) = \frac{x- < x >}{\sigma(x)} - \frac{y- < y >}{\sigma(y)} = constant$$

This is always fulfilled if $y = a + bx$, where b is positive. Hence, in a positive linear dependence the correlation coeffient is $+1$, in a negative linear dependence the correlation coefficient is -1.

If the covariance $cov(x, y) = 0$ then x and y are independent and $f(x, y) = g(x)h(y)$:

$$\begin{aligned} cov(x, y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x- < x >)(y- < y >)g(x)h(y)dxdy \\ &= \left(\int_{-\infty}^{\infty} (x- < x >)g(x)dx \right) \left(\int_{-\infty}^{\infty} (y- < y >)h(y)dy \right) \\ &= 0 \end{aligned}$$

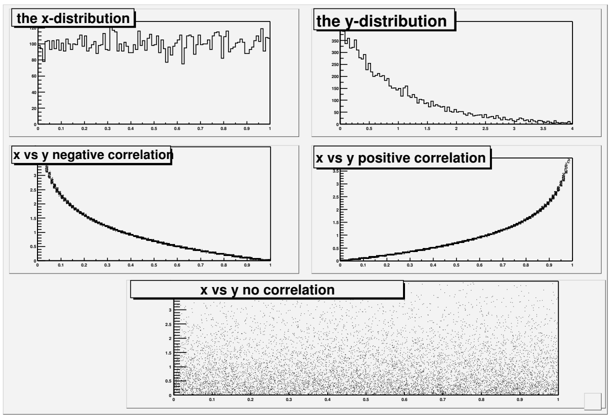


Figure 7: Examples of correlation between variables.

In figure 7 we show three sets of figures with the distributions of two variables x and y; we show the marginal distributions of each of them, and their two-dimensional distribution (x versus y). The correlation coefficients are (in order) -1, +1 and 0.

3.3 Linear functions/transformation of random variables

Consider the following case: we have n (different) random variables $x_1, x_2, ...x_n$, each with its own pdf: $f_1(x), f_2(x), ...f_n(x)$. The expectation values of the pdf's are noted as $< x_1 >, < x_2 >, ... < x_n >$; the variances and covariances as $var(x_1), var(x_2), ...var(x_n)$ and $cov(x_1, x_2), cov(x_1, x_3), ...cov(x_i, x_j)$. We now transform the n variables x_i $(i = 1, 2, ..., n)$ into a set of m different variables $y_1, y_2, ...y_m$, using the following transformation:

$$y_k = a_k + \sum_{i=1}^{n} r_{ki} x_i \qquad (k = 1, 2, ..., m) \tag{31}$$

Or, in matrix notation:

$$Y = A + RX \tag{32}$$

The new variables y_i have their own distributions, expectation values and variances; moreover they may have covariances, since they are derived from the same set x_i. The question is: how are these (new) quantities related to the ones before transformation? The expectation values of the original variables $< x_i >$ form a column vector $< X >$. Their variances $var(x_i)$ and covariances $cov(x_i, x_j)$ form the *covariance matrix*

$$C_x = \begin{pmatrix} cov_{1,1} & cov_{1,2} & ... & cov_{1,n} \\ cov_{2,1} & cov_{2,2} & ... & cov_{2,n} \\ . & . & . & . \\ . & . & . & . \\ cov_{n,1} & cov_{n,2} & ... & cov_{n,n} \end{pmatrix} \tag{33}$$

The covariance matrix is symmetric: $cov_{i,j} = cov_{j,i}$. The diagonal elements are the variances $cov_{i,i} = \sigma^2(x_i) = var_i$. The additive terms a_k in the transformation form a column vector A; the multiplication factors r_{ij} in the transformation form a matrix R:

$$R = \begin{pmatrix} r_{1,1} & r_{1,2} & ... & r_{1,n} \\ r_{2,1} & r_{2,2} & ... & r_{2,n} \\ . & . & . & . \\ . & . & . & . \\ r_{m,1} & r_{m,2} & ... & r_{m,n} \end{pmatrix} \tag{34}$$

The sum rule for expectation values relates the new expectation values Y to the old ones X through:

$$< Y >= A + R < X > \tag{35}$$

The sum rule for covariances relates the new covariance matrix C_y to the old one C_x through:

$$C_y = RC_xR^T \tag{36}$$

where R^T is the transposed matrix of R (swap rows and columns). This equation expresses the law of error propagation. Suppose that the errors (i.e. the standard deviations and variances) and covariances are known, what are the errors of $y(x)$? If the errors are small, the pdf of x will peak in a small region around $< x >$. Then one can perform a Taylor expansion of the function around the expectation value.

As a simple example consider the sum of two random variables: $u = x + y$. Then is:

$$\begin{aligned} R &= \begin{pmatrix} 1 & 1 \end{pmatrix} \\ < u > &= < x > + < y > \\ \sigma_u^2 &= \sigma_x^2 + \sigma_y^2 + 2 \cdot cov(x, y) \end{aligned}$$

If x and y are not correlated the errors are added quadratically: $\sigma_u^2 = \sigma_x^2 + \sigma_y^2$.

3.4 Non-linear functions/transformation of random variables

Consider (again) the case that we have n (different) random variables $x_1, x_2, ...x_n$ with pdf's $f_1(x), f_2(x), ...f_n(x)$, expectation values $< x_1 >, < x_2 >, ... < x_n >$, variances $var(x_1), var(x_2), ...var(x_n)$ and covariances $cov(x_1, x_2), cov(x_1, x_3), ...cov(x_i, x_j)$. We now transform the n variables x_i $(i = 1, 2, ..., n))$ into a set of m different variables $y_1, y_2, ...y_m$, using the following transformation:

$$y_k = f_k(x_1, x_2, ...x_n) \quad (k = 1, 2, ..., m) \tag{37}$$

The new variables y_i have their own distributions, expectation values and variances; moreover they may have covariances, since they are derived from the same set x_i. The question is how these (new) quantities are related to the ones before transformation. The expectation values simply satisfy:

$$< y_k >= f_k(< x_1 >, < x_2 >, ... < x_n >) \tag{38}$$

To calculate the new covariance matrix we proceed as follows: expand $y_k(x_1, x_2, ..., x_n)$ around the expectation values in a Taylor-series:

$$\begin{aligned} y_k(x_1, x_2, ..., x_n) &\simeq y_k(< x_1 >, < x_2 >, ..., < x_n >) + \\ &+ \sum_{j=1}^{n}(x_j - < x_j >)\frac{\partial y_k}{\partial x_j}(\ in\ \ x_j =< x_j >) \end{aligned}$$

If we can neglect the higher-order terms, the problem reduces to the linear case. The coefficients of the transformation matrix R of the previous section have to be written as:

$$r_{kj} = \frac{\partial y_k}{\partial x_j}(\ in\ \ x_j =< x_j >) \tag{39}$$

The sum rule for the new covariance matrix is again:

$$C_y = RC_xR^T \tag{40}$$

with:

$$R = \begin{pmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & ... & \partial y_1/\partial x_n \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 & ... & \partial y_2/\partial x_n \\ . & . & .... & \\ . & . & .... & \\ & & & \\ \partial y_m/\partial x_1 & \partial y_m/\partial x_2 & ... & \partial y_m/\partial x_n \end{pmatrix} (\ in\ \ X =< X >) \tag{41}$$

As a simple example consider the product of two random variables: $u = x \cdot y$.
The transformation matrix R is:

$$\begin{pmatrix} < y > & < x > \end{pmatrix} \tag{42}$$

It follows that:

$$\begin{aligned} \frac{\sigma^2(x \cdot y)}{< x \cdot y >^2} &= \left(\frac{\sigma_x}{< x >}\right)^2 + \left(\frac{\sigma_y}{< y >}\right)^2 + \\ &+ 2 \cdot cov(x, y) \cdot \frac{\sigma_x}{< x >}\frac{\sigma_y}{< y >} \end{aligned}$$

If x and y are not correlated the relative errors add quadratically:

$$(\frac{\sigma_u}{< u >})^2 = (\frac{\sigma_x}{< x >})^2 + (\frac{\sigma_y}{< y >})^2 \tag{43}$$

3.5 Errors and error-propagation

We anticipate a little bit on things which will be treated later on, and give already now an interpretation of the expectation value and the standard deviation. We are dealing with the measurement of a quantity whose outcome is a random variable, drawn from a probability distribution. Repeated measurement of this quantity gives a series of values, which can be represented in a histogram (a frequency distribution). We consider the mean value of the measurements as the best choice for the real value of the quantity. The standard deviation of the measurements is used as a measure for the statistical uncertainty/error in the best value. When we measure n quantities simultaneously, the result is presented as n best values and the corresponding nxn covariance matrix. The diagonal elements of this matrix contain the variances (the square of the standard deviations). The off-diagonal elements contain the covariances. Now, suppose that we have to transform our measured quantities to a set of different quantities (e.g. because the theory of our experiment is expressed in different quantities). The best values of the theoretical quantities are obtained simply by applying the transformation. Their statistical errors and correlation coefficients have to be calculated according to the rules of the previous sections.

As an example, suppose that we have measured the coordinates of a point in the (x, y) plane. The result is: $x = 1$, $y = 1$. The error in x is $\Delta(x) = 0.1$, the error in y is $\Delta(y) = 0.3$. We assume that our measuring device has the property that the measurement errors in x and y are independent. The covariance matrix is then:

$$C_{xy} = \begin{pmatrix} 0.01 & 0. \\ 0. & 0.09 \end{pmatrix} \tag{44}$$

We now change to polar coordinates with the transformation:

$$r = \sqrt{x^2 + y^2} \qquad \phi = \arctan \frac{y}{x} \tag{45}$$

The values of (r, ϕ) for the point $(x = 1, y = 1)$ are: $r = \sqrt{2}$ and $\phi = \pi/4$. The linearised transformation matrix is:

$$R = \begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \phi}{\partial x} & \frac{\partial \phi}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ \frac{-y}{r^2} & \frac{x}{r^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{46}$$

The new covariance matrix becomes:

$$\begin{aligned} C_{r\phi} &= R C_{xy} R^T \\ &= \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0.01 & 0. \\ 0. & 0.09 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} 0.05 & \frac{0.04}{\sqrt{2}} \\ \frac{0.04}{\sqrt{2}} & \frac{0.05}{\sqrt{2}} \end{pmatrix} \end{aligned}$$

Note that the matrix is symmetric around the diagonal, hence, the errors in r and ϕ are correlated. The transformation back from the (r, ϕ)-system to the (x, y)-system ($x = r\cos\phi$, $y = r\sin\phi$) is given by the matrix:

$$R' = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ -\frac{1}{\sqrt{2}} & 1 \end{pmatrix} \tag{47}$$

When we apply this back transformation to the new covariance matrix $C_{r\phi}$, we get the old covariance matrix C_{xy} back. This would not have been the case if we would have ignored the covariances (the off-diagonal elements) .

3.6 Exercises

Exercise 3.1

Verify that the function $f(x) = 1 - |1 - x|$ for $0 < x < 2$ and $f(x) = 0$ elsewhere, is a probability density function.

Exercise 3.2

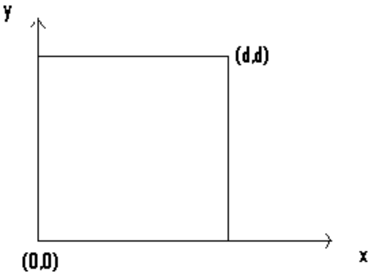
The random variable x has a pdf $f(x) = 0.75(1 - x^2)$ if $-1 \leq x \leq 1$ and zero otherwise.

a. What is the cumulative distribution function $F(x)$?

b. Calculate $P(-\frac{1}{2} \leq x \leq \frac{1}{2})$.

c. Calculate $P(\frac{1}{4} \leq x \leq 2)$.

Exercise 3.3



Suppose particles can fall randomly in a square of side d (see figure). Find $< x >$ and $\sigma(x)$.

Exercise 3.4

Using Chebyshev's inequality show that $P(x = 0) = 1$, if $E(x) = E(x^2) = 0$.

Exercise 3.5

In a study on age and body fat, the percentage of fat in a sample of men aged 40 to 49 produced a mean of 26% and a standard deviation of 6.19%.

a. At least what proportion of the men had percentages of fat between 15% and 37%?

b. If 150 men were included in the sample, at most how many had fat percentages that differ from the mean by more than 15%?

Exercise 3.6

The period of a pendulum is given by $T = 2\pi\sqrt{l/g}$, where l is the length of the pendulum and g is the gravitational acceleration. Compute g and $\Delta g = \sigma_g$ using the measured values $l = 99.8$ cm, $\Delta l = \sigma_l = 0.3$ cm, $T = 2.03$ s and $\Delta T = \sigma_T = 0.05$ s. Assume that the measurements of l and T are not correlated.

Exercise 3.7

m and v are the measured mass and velocity of an object. The measurement errors are $\Delta m = \sigma_m$ and $\Delta v = \sigma_v$. The measurements of m and v are independent; the relative measurements errors are known: $\Delta m/m = a$ and $\Delta v/v = b$. Consider the momentum $p = mv$ and the kinetic energy $E = \frac{1}{2}mv^2$ of the object and compute $\sigma^2(p), \sigma^2(E), cov(p, E)$ and $\rho(p, E)$.

Exercise 3.8

You start with measurements of a and b which are uncorrelated and have variances σ_a^2 and σ_b^2. Calculate the error matrix (the covariance matrix) on the new variables $x = \frac{1}{\sqrt{2}}(a + b)$ and $y = \frac{1}{\sqrt{2}}(a - b)$.

Exercise 3.9 - ***root*** **exercise 2**

This exercise deals with the calculation and propagation of errors. In an experiment collisions have been observed between neutrino's (ν) and antineutrino's ($\overline{\nu}$) with protons (p) and neutrons (n) . The observed collisions were divided over 10 categories. Table 3.6 gives, for each category:

- the physical description of the final state
- the number of observed events, and (between brackets) the statistical uncertainty in that number
- a shorthand notation for the number, which will be used below.

It is very important to note that the categories have been chosen in such a way that the statistical errors are uncorrelated.

ν-beam (pure)	number of events (stat.error)	notation
$\nu + p \to \mu^- + ...$	812 (41)	x_0
$\nu + n \to \mu^- + ...$	1800 (50)	x_1
$\nu + p \to \nu + ...$	400 (33)	x_2
$\nu + n \to \nu + ...$	464 (35)	x_3
$\overline{\nu}$-beam		
(ν-background)		
$\overline{\nu} + p \to \mu^+ + ...$	818 (33)	x_4
$\overline{\nu} + n \to \mu^+ + ...$	356 (22)	x_5
$(\overline{\nu})or(\nu) + p \to (\overline{\nu})or(\nu) + ...$	289 (27)	x_6
$(\overline{\nu})or(\nu) + n \to (\overline{\nu})or(\nu) + ...$	301 (29)	x_7
$\nu + p \to \mu^- + ...$	164 (19)	x_8
$\nu + n \to \mu^- + ...$	393 (24)	x_9

From these ten measured quantities we construct the following four ratios

$$R^{\nu p} = \frac{\nu + p \to \nu + ...}{\nu + p \to \mu^- + ...} = \frac{x_2}{x_0}$$

$$R^{\nu n} = \frac{\nu + n \to \nu + ...}{\nu + n \to \mu^- + ...} = \frac{x_3}{x_1}$$

$$R^{\overline{\nu} p} = \frac{\overline{\nu} + p \to \overline{\nu} + ...}{\overline{\nu} + p \to \mu^+ + ...} = \frac{x_6 - x_8 * x_2/x_0}{x_4}$$

$$R^{\overline{\nu} n} = \frac{\overline{\nu} + n \to \overline{\nu} + ...}{\overline{\nu} + n \to \mu^+ + ...} = \frac{x_7 - x_9 * x_3/x_1}{x_5}$$

Calculate and print the values and statistical errors of $R^{\nu p}, R^{\nu n}, R^{\overline{\nu} p}$ and $R^{\overline{\nu} n}$.
Theory predicts that the four quantities R depend on four 'coupling constants': $u_L^2, d_L^2, u_R^2, d_R^2$. The

relationship is as follows :

$$\begin{pmatrix} u_L^2 \\ d_L^2 \\ u_R^2 \\ d_R^2 \end{pmatrix} = \begin{pmatrix} 0.675 & -0.607 & -0.119 & 0.010 \\ -0.282 & 1.331 & 0.027 & -0.049 \\ -0.133 & 0.060 & 0.477 & -0.078 \\ 0.024 & -0.299 & -0.186 & 0.185 \end{pmatrix} \begin{pmatrix} R^{\nu p} \\ R^{\nu n} \\ R^{\overline{\nu} p} \\ R^{\overline{\nu} n} \end{pmatrix}$$

Calculate and print the values and statistical errors of $u_L^2, d_L^2, u_R^2, d_R^2$ To compare the results of this experiment with other results, the following linear combinations of the coupling constants are needed/relevant

$$u_L^2 + d_L^2, \quad u_L^2 - d_L^2, \quad u_R^2 + d_R^2, \quad u_R^2 - d_R^2$$

Calculate and print the values of these quantities and their statistical errors.

It is a useful practice to represent the results (values found for the coupling constants, their errors and covariances) in a graphical way, like figure 4 in chapter 1. The one standard-deviation contour for two quantities with values a_1 and a_2, standard deviations σ_1 and σ_2, and correlation coefficient ρ is the ellipse defined by

$$\frac{(x_1 - a_1)^2}{\sigma_1^2} + \frac{(x_2 - a_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1 - a_1)}{\sigma_1} \frac{(x_2 - a_2)}{\sigma_2} = 1 - \rho^2$$

Make a contour plot of u_l^2 *versus* d_L^2.

3.6.1 Tools in *root* for matrix-calculus

For this exercise you have to calculate the product of a matrix and a vector, the product of two matrices and a so-called 'triple-product' of matrices: $R \cdot C \cdot R^T$. In *root* you can do that as follows:

- To multiply two matrices, define a matrix a and a matrix b and multiply $c = a * b$:

```
E.g. fill a 3x3 matrix a(nRows, nCols)=a(3,3):
TMatrixD a(3,3) ;
a(0,0) = 3. ;
a(0,1) = 4. ;
a(1,0) = 5. ;
....

And fill a 3x3 matrix b:
TMatrixD b(3,3) ;
b(0,0) = 6. ;
....

and multiply and store the results in the 3x3 matrix c:
TMatrixD c=a*b ;
```

- To multiply a matrix a with a vector (a one-dimensional matrix) b:

```
E.g. define and fill 2x2 matrix a and a 2x1 matrix (a vector) b:
TMatrixD a(2,2) ;
a(0,0)= 8. ;
```

```
...
TMatrixD b(2,1) ;
b(0,0) = 15.;
...
multiply and store the results in the matrix c:
TMatrixD c=a*b  ;
```

- to calculate the triple product $c = a \cdot b \cdot a^T$:

  ```
  E.g. fill the 2x3 matrix a and the 2x1 matrix (vector) b:
  TMatrixD a(2,3) ;
  a(0,0) = 10 ;
  ...
  TMatrixD b(3,3) ;
  b(0,0) = 20 ;
  ...

  define the 3x2 matrix as the transpose of the 2x3 matrix a:
  TMatrixD c(2,3) ;

  and store the results of the triple product in the 2x2 matrix e:
  TMatrixD e = a*b*c.Transpose(a) ;
  ```

Example of a contour plot

An example of a program for the graphical presentation of a contour-line, can be found in: */user/uvak/sda/example2.C*

```
/*                                                                      */
/*           This is the code of /user/uvak/sda/example2.C              */
/*           The program uses the ROOT graphical package to             */
/*           to draw a circle                                           */
/*                                                                      */
int example2 ()
{
    gROOT  ->Reset()  ;
    gRandom->SetSeed();
/*                                                                      */
/*            Here starts our program                                   */
/*                                                                      */
# define nstep 1001
/*                                                                      */
/*         The circle will be drawn as two ’arcs’ :                     */
/*         One with y-values < yCentre ; one with y-values > yCentre    */
/*                                                                      */
    Double_t xCentre  = 1. ;     // x-coordinate of centre of circle
    Double_t yCentre  = 1. ;     // y-coordinate of centre of circle
```

```
    Double_t radius  = 1. ;     // radius of circle
    Double_t xstep = 2. * radius / (nstep-1) ; // stepsize in x-coordinate
    Double_t xVal [nstep] ;   // x-coordinates of circle
    Double_t yVal1[nstep], yVal2[nstep] ; // y-coordinates of circle
/*                                                                      */
/*         Fill the vectors with the x,y coordinates of the circle      */
/*                                                                      */
    Double_t dx;
    for (Int_t i=0; i<nstep; i++)
      {
xVal[i] = xCentre - radius + i * xstep ;
        dx      = xVal[i] - xCentre ;
        yVal1[i] = yCentre + sqrt(radius*radius - dx*dx);
        yVal2[i] = yCentre - sqrt(radius*radius - dx*dx);
      }
/*                                                                      */
/*           Prepare the graphs : gr1 for one circle arc                */
/*                                gr2 for the other circle arc          */
/*           and markerCentre, the marker at the circle center          */
/*                                                                      */
    Int_t marker_id = 3;  // asterix at the circle center
    TGraph *gr1 = new TGraph (nstep, xVal, yVal1) ;
    TGraph *gr2 = new TGraph (nstep, xVal, yVal2) ;
    TMarker *markerCentre =  new TMarker (xCentre, yCentre, marker_id);
/*                                                                      */
/*           open a window for displaying the graphics results
/*                                                                      */
      TCanvas *Coupling = new TCanvas ("example2", "Title", 1) ;
/*                                                                      */
/*            Give the graph a title                                    */
/*                                                                      */
      gr1-> SetTitle ("Contourplot Ul2 vs Dl2") ;
/*                                                                      */
/*            Set min-max values for y-axis                             */
/*                                                                      */
      gr1-> SetMinimum ( yCentre - radius ) ;
      gr1-> SetMaximum ( yCentre + radius ) ;
/*                                                                      */
/*            Draw the two graphs; connect points with a smooth line    */
/*                                                                      */
      gr1->Draw("AC") ;
      gr2->Draw("CP") ;
/*                                                                      */
/*            Give the axes a title                                     */
/*                                                                      */
      gr1-> GetXaxis() ->SetTitle ("Ul2") ;
      gr1-> GetXaxis() ->CenterTitle ( ) ;
      gr1-> GetYaxis() ->SetTitle ("Dl2") ;
      gr1-> GetYaxis() ->CenterTitle ( ) ;
/*                                                                         */
/*           Draw a marker at the circle center                            */
/*                                                                         */
      markerCentre->Draw() ;

      return 0 ;
}
```

3.6.2 Presentation of results

HAND IN :

- A print of your program code
- A print of all the required/calculated quantities
- A print of the contour plot

4 Some important distributions

Table 1 give a number of common pdf's and their means and variances. The notation of the pdf is $f(variable; parameters)$. We will discuss the distributions below except for the normal or Gaussian distribution and the χ^2 distribution which will be discussed later.

distribution	pdf f(variable; parameters)	mean	variance
Uniform	$f(k; a, b) = \frac{1}{(b-a)}$ for $a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	zero otherwise		
Binomial	$f(k; n, p) = \frac{n!}{k!(n-k)!} p^k q^{n-k}$	np	npq
	$k = 0, 1, 2, ..., n$, $0 \le p \le 1$, $q = 1 - p$		
Poisson	$f(n; \lambda) = \frac{1}{n!} \lambda^n e^{-\lambda}, n = 1, 2, ...$	λ	λ
	$\lambda > 0$		
Exponential	$f(x; \lambda) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
	$0 < x < \infty$		
Normal (Gaussian)	$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
χ^2	$f(x; n) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)}$	n	$2n$

Table 1: Common probability density functions.

4.1 The binomial distribution

Consider a random variable/process with two possible outcomes: A and $notA$. For example the toss of a coin, where outcome A means 'head' and outcome $notA$ means 'tail', or the forward-backward asymmetry in $e^+e^- \to \mu^+\mu^-$ interactions, where the outcome A means: the outgoing μ^+ goes 'in the direction of the incoming e$^+$'. Such a process is called a Bernoulli trial. The binomial distribution gives the probability of k successes in n independent Bernouilli trials each having probability p of succes. Set $P(A) = p$ then is $P(notA) = 1 - p = q$. Let us consider the example of the above mentioned forward-backward asymmetry. We observe n collisions in total and ask: what is the probability that in k out of n collisions the μ^+ goes forward (i.e. event A occurs). The answer is:

$$W_k^n = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} = \begin{pmatrix} n \\ k \end{pmatrix} p^k q^{n-k} = f(k; n, p) \tag{48}$$

We assign to each collision a variable x_i: $x_i = 1$ if event A occurs (forward μ^+) and $x_i = 0$ if event $notA$ occurs (backward μ^+). The expectation value and variance of x_i are:

$$\begin{aligned} E(x_i) &= 1 \cdot p + 0 \cdot q = p \\ \sigma^2(x_i) &= (1-p)^2 \cdot p + (0-p)^2 \cdot q = p \cdot q \end{aligned}$$

We now observe n collisions, assign the corresponding value of x_i to each of them, and sum the result in the random variable x:

$$x = \sum_{i=1}^{n} x_i \tag{49}$$

The expectation value, variance and standard deviation of x are:

$$
\begin{aligned}
E(x) &= n \cdot p \\
\sigma^2(x) &= n \cdot p \cdot (1-p) = n \cdot p \cdot q \\
\sigma(x) &= \sqrt{n \cdot p \cdot (1-p)} = \sqrt{n \cdot p \cdot q}
\end{aligned}
$$

Figure 8 shows the probability density distribution $f(x; n, p)$ for $p = 0.3$ and several values of n.

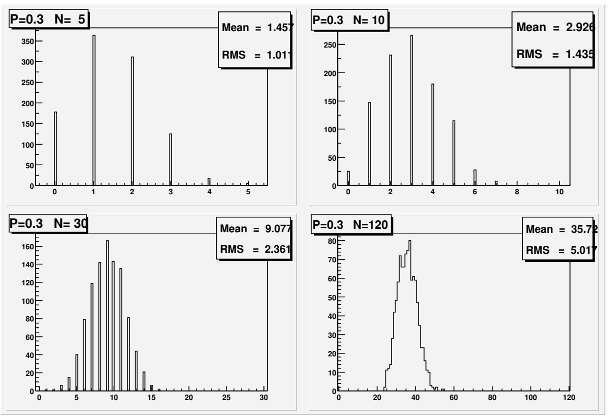


Figure 8: Binomial distributions

They indeed peak around the value np; as n increases the peak, in proportion to the full range of n, becomes progressively narrower. The relative width of the peak also depends on p and (for the same n) peaks with p close to 0 or 1 are narrower than those with p near 0.5. For $p = 0.3$ we expect:

n	$E(x)$	$\sigma(x)$
5	1.5	1.02
10	3.0	1.45
30	9.0	2.51
120	36.0	5.02

(50)

Application to the forward-backward asymmetry in $e^+e^- \rightarrow \mu^+\mu^-$: if F is the observed number of forward muons, B is the observed number of backward muons and $N = F + B$ is the total number of observed collisions, then the asymmetry is defined as:

$$R = \frac{F - B}{N} = \frac{2 \cdot F - N}{N} \tag{51}$$

The standard deviation in F (a measure for the statistical uncertainty in F) is:

$$\sigma(F) = \sqrt{N \cdot p \cdot q} \simeq \sqrt{\frac{F \cdot B}{N}} \tag{52}$$

The standard deviation in R is obtained from $\sigma(F)$ through error-propagation:

$$\sigma(R) = \frac{2}{N}\sqrt{\frac{F \cdot B}{N}} \tag{53}$$

This result is used in the corresponding example in chapter 1 to draw the error bars on the points in the histogram. If the value of r is known (approximately), we can use the formula for $\sigma(r)$ in order to answer the question: how many collisions do we have to observe in order to measure the value of r with a specific precision?

4.2 The multinomial distribution

A logical extension of the binomial distribution deals with experiments where more than two different outcomes are possible. Consider the measurement of a random variable with k possible values $A_1, A_2, ..., A_k$, with probabilities $p_1, p_2, ...p_k$ ($\sum_{i=1}^{k} p_i = 1$). We make N measurements, count how many times each value occurs (N_i), and display the result in a histogram with k intervals/bins. The expectation value for the number of events in interval i is given by:

$$E(N_i) = N \cdot p_i \tag{54}$$

The variance of N_i is:

$$var(N_i) = \sigma^2(N_i) = N \cdot p_i \cdot (1 - p_i) \tag{55}$$

Since $\sum_{i=1}^{k} p_i = 1$ there is also a covariance between the number of events in interval i and j:

$$cov(N_i, N_j) = -N \cdot p_i \cdot p_j \tag{56}$$

The corresponding correlation coefficient is:

$$\rho(N_i, N_j) = -\sqrt{\frac{p_i \cdot p_j}{(1 - p_i) \cdot (1 - p_j)}} \tag{57}$$

When the number of intervals is large, hence the probabilities p_i are small, then:

$$var(n_i) = \sigma^2(N_i) = N \cdot p_i \simeq N_i \tag{58}$$

$$\sigma(N_i) = \sqrt{N \cdot p_i} \simeq \sqrt{N_i} \tag{59}$$

$$\rho(N_i, N_j) \simeq 0. \tag{60}$$

4.3 The law of large numbers

We will illustrate this law using the multinomial distribution. Imagine the measurement of a quantity with k possible values, with probabilities $p_1, p_2, ..., p_k$. Usually the probabilities for different outcomes are not known but have to be obtained from experiment: perform N measurements, and count how frequently each possible outcome has occurred; we note the results as N_i. We choose as best estimate for p_i : N_i/N. This is justified since: $E(N_i) = N \cdot p_i$. A measure for the possible deviation of the best estimate from the true value of p_i is the standard deviation $\sigma(p_i) = \sqrt{p_i(1 - p_i)/N}$. We note that this deviation becomes smaller if the number N of available measurements increases. This is an example of the *law of large numbers*.

4.4 The Poisson distribution

A sample of radio-active nuclei contains a very large number of atoms; a small number of these atoms decay (from time to time). The number of atoms which decay per time interval is characteristic for the atom (its lifetime). The decay itself is a random process, so we can ask the question: what is the probability to observe k decaying atoms per time interval? In fact this decay process is described by the binomial distribution, but counting the number of non-decaying atoms is impossible. Assume that we start with a sample of n atoms and that p is the probability for one atom to decay in a time interval dt and that k atoms decay in this time interval. Then the change in n in time interval dt is $-\frac{dn}{dt} = k$. The expectation value for the change in n per time interval is $E(-\frac{dn}{dt}) = E(k) = -np$. If n is large and p is small we can neglect the change in n. We look at the shape of the binomial distribution if $n \to \infty$, while $E(k) = \lambda = n \cdot p$ remains constant.

$$\begin{aligned}
W_k^n &= \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k} \\
&= \frac{n!}{k!(n-k)!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k} \\
&= \frac{\lambda^k}{k!}\frac{n(n-1)(n-2)...(n-k+1)}{n^k}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k} \\
&= \frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^n(1-\frac{1}{n})(1-\frac{2}{n}).....(1-\frac{k-1}{n})(1-\frac{\lambda}{n})^{-k}
\end{aligned}$$

In the limit $n \to \infty$ the distribution becomes equal to $(\lim_{n\to\infty}(1-\frac{\lambda}{n})^n = e^{-\lambda})$:

$$\lim_{n\to\infty} W_k^n = f(k) = \frac{\lambda^k}{k!}e^{-\lambda} = f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{61}$$

$f(k)$ is the Poisson distribution. It gives the probability of getting k events if the expected number is λ. It is normalised:

$$\sum_{k=0}^{\infty} f(k) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}(1+\lambda+\lambda^2/2!+\lambda^3/3!+.....) = e^{-\lambda}e^{\lambda} = 1$$

Its expectation value is:

$$\begin{aligned}
E(k) &= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda e^{-\lambda}e^{\lambda} = \lambda
\end{aligned}$$

Its variance and standard deviation is determined as follows:

$$\begin{aligned}
E(k^2) &= \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda} k\frac{\lambda\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda}\sum_{j=0}^{\infty}(j+1)\frac{\lambda^j}{j!} = \lambda(\lambda+1)
\end{aligned}$$

$$\sigma^2(k) = E(k^2) - \{E(k)\}^2 = \lambda(\lambda+1) - \lambda^2 = \lambda \tag{62}$$

$$\sigma(k) = \sqrt{\lambda} \tag{63}$$

The Poisson pdf is usually written as $P(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$. As stated before, it is the probability of getting exactly k events in a given interval of x (e.g. in time or space) at an average rate λ per given interval. Figure 9 shows the Poisson distribution for different values of λ .

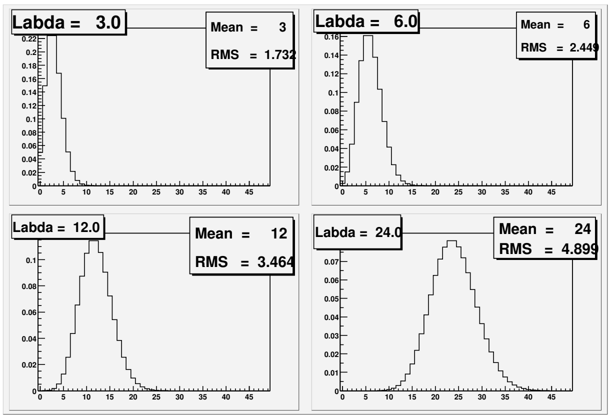


Figure 9: Poisson distributions

We see that for increasing λ the distribution becomes more symmetric. We expect the following values:

λ	$E(x)$	$\sigma(x)$
3	3.	1.73
6	6.	2.45
12	12.	3.46
24	24.	4.90

(64)

As we have obtained the Poisson distribution from the binomial distribution with large n, but constant $\lambda = np$, i.e. small p, we expect it to apply to processes in which a large number of events occur, but of which only very few are of interest to us (i.e. large number of 'trials', few 'successes'). For large λ the Poisson distribution approaches the normal or Gaussian distribution (see chapter 5).

4.5 The exponential distribution

Consider again the radioactive decay of atoms of the previous section. We found that the mean number of atoms that decay in a time interval dt is $E(\frac{dn}{dt}) = -np$. We can rewrite this as $\frac{dE(n)}{dt} = -np$. If we consider $dE(n)$ the actual number of atoms decaying in time interval dt, then $\frac{dn}{dt} = -np$ and thus $n = n_0 e^{-pt}$: the number of undecayed atoms falls exponentially and the pdf for the distribution of individual decays times (lifetimes) is exponential.

If $f(t)$ is the pdf for an individual atom to decay after a time t, then the probability that it decays before time t is the cdf $F(t) = \int_0^t f(t)dt$ and the expected number of decays in time t is $n_0 F(t)$. We find for the probability that an individual atom will decay in time t:

$$f(t;\tau_0) \quad = \quad \frac{1}{\tau_0} e^{-\frac{t}{\tau_0}}$$

Note that this probability distribution is properly normalised. Its expectation value is:

$$E(t) = \int_0^\infty \frac{t}{\tau_0} e^{\left(-\frac{t}{\tau_0}\right)} dt = \tau_0 \tag{65}$$

To calculate its variance we note that:

$$E(t^2) = \int_0^\infty \frac{t^2}{\tau_0} e^{\left(-\frac{t}{\tau_0}\right)} dt = 2\tau_0^2 \tag{66}$$

So the variance and its standard deviation are:

$$\sigma^2(t) = E(t^2) - \{E(t)\}^2 = \tau_0^2 \tag{67}$$

$$\sigma(t) = \tau_0 \tag{68}$$

Figure 10 shows two exponential distributions for different lifetimes.

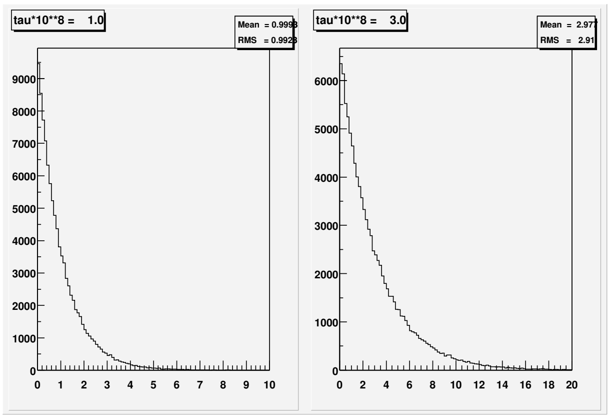


Figure 10: Exponential distributions

4.6 The uniform distribution

The uniform distribution is not of great practical interest but is easy to handle being the simplest distribution of a continuous variable. It is often advantageous to transform a distribution function by means of a transformation of variables into a uniform distribution or, the reverse, to express the given distribution in terms of a uniform distribution. It plays an important role in the simulation of random processes on a computer (the Monte Carlo method). We will deal with this in a later chapter. Here, we only present the definition:

$$f(x) = 1/(b-a) \ \ for \ \ a \leq x < b \qquad f(x) = 0 \ \ for \ \ x < a, x \geq b \tag{69}$$

The distribution is normalised:

$$\int_{-\infty}^{\infty} f(x)dx \quad = \quad 1$$

The expectation value and the variance of x are:

$$\begin{aligned} E(x) &= \frac{a+b}{2} \\ \sigma^2(x) &= \frac{1}{12} \cdot (b-a)^2 \end{aligned}$$

i.e. the standard deviation for a uniform distribution is the width divided by $\sqrt{12}$.

4.7 Exercises

Exercise 4.1

Calculate the probability of obtaining at least two six in rolling a fair die four times.

Exercise 4.2

Calculate the probability of getting a number of heads between 15 and 21 in 36 tosses of a fair coin.

Exercise 4.3

a. Proof the recursion formula:

$$W^n_{k+1} = \frac{n-k}{k+1} \cdot \frac{p}{q} \cdot W^n_k \tag{70}$$

b. Suppose that for a certain production proces a fraction $q = 0.2$ of all pieces produced is defective. That means that in 5 pieces produced the expected number of non-defective pieces is 4. What is the probability that at most 2 pieces of these 5 are free of defects? Use the relationship of a) to simplify the calculation.

Exercise 4.4

A defence system is 99.5% efficient in intercepting ballistic missiles. What is the probability that it will intercept all of 100 missiles launched against it? How many missiles must an agressor launch to have a better than even chance of one or more penetrating the defences? And how many missiles would be needed to ensure a better than even chance of more than two missiles evading the defence? You can only find the answer to this last question by trial.

Exercise 4.5

You are measuring tracks of cosmic ray particles passing through a stack of detectors, of which each with an efficiency of 95% gives the position of the particle that passes the detector. At least three points are need to define a track. How efficient at detecting tracks would a stack of three detectors be? Would using a stack of four or five detectors give a significant improvement?

Exercise 4.6

In a hospital the doctor on duty is called on average three times per night. The number of calls may be considered Poisson distributed. What is the probability for the doctor to have a completely quiet night?

Exercise 4.7

Here are the numbers of neutrino events detected in 10 second intervals by the Irvine-Michigan-Brookhaven experiment on 23 February 1987 - around which time the supernova S1987a was first seen by astronomers:

no. of events	0	1	2	3	4	5	6	7	8	9
no. of intervals	1042	860	307	78	15	3	0	0	0	1
no. of intervals predicted										

a. Ignoring the interval with nine events what is the mean number of neutrino events per interval of 10 seconds?

b. Calculate the predictions for the number of intervals with 0, 1, ..., 9 neutrino events assuming a Poisson distribution for the pdf.

c. Do these predictions agree with the observed number? Are the nine events a statistical fluctuation or did they come from the supernova?

Exercise 4.8 - *root* exercise 3

Problem statement

This exercise deals with the law of large numbers. This law, which is one of the building blocks of the treatment of statistical data, can be summarised as follows:

- Generally speaking, measurements are *random drawings* from a probability distribution.
- When we calculate values of the probability distribution from a finite set of measurements, the calculated values can/will deviate from the true ones. **BUT:**
- the calculated values have a clear relationship to the true ones (they tend to *'cluster'* around the true value), **AND**
- the difference between the calculated values and the true ones *decreases* if the number of available measurements increases.

In this exercise we will *see the law at work* (and thereby verify it). We will do it for the following probability distributions:

- The standard-normal distribution

  $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad (-\infty \leq x \leq \infty)$$

  This distribution is symmetric around $x = 0$.
  Its expectation value is $\int xf(x)dx = 0$.

- The Cauchy distribution

  $$f(x) = \frac{1}{\pi(1+x^2)} \qquad (-\infty \leq x \leq \infty)$$

  This distribution is also symmetric around $x = 0$.
  However, its expectation value $\int xf(x)dx$ is **not defined**.

In order to verify the law we need *recipes* to generate random numbers which follow the standard-normal distribution or the Cauchy distribution. These recipes always start with random numbers from the uniform distribution between 0 and 1. Under *root* these numbers can be obtained with the 'command'

```
x = gRandom -> Rndm (1) ;
```

With two random numbers u and v from the uniform distribution we can get a random number from the standard-normal distribution through :

$$x = \sqrt{-2\ln u} \cdot \cos(2\pi v)$$

With two random numbers u and v from the uniform distribution we can get a random number from the Cauchy distribution through :

$$x = \frac{1-2u}{1-2v}$$

To verify the law, we propose the following procedure (for each distribution):

- generate 50 series of 100 numbers
  calculate the average value for each series
  represent the 50 average values in a histogram
- repeat this procedure for 50 series of 1000 numbers

This gives two histograms for the standard-normal distribution and two for the Cauchy distribution. Verify if the results in the histogram confirm what you expect from the law of large numbers and understand/explain the behaviour of the Cauchy distribution

In the case of the Cauchy distribution we can better work with the median value of the distribution (in stead of the average value).
The median value x_m of a probability distribution $f(x)$ is *defined* as the value for which $P(x < x_m) = 0.5$. Verify this by making histograms of the median values of the series of numbers generated for the Cauchy distribution.

Practical remarks

- An accurate value of π can be obtained with

  ```
  pi = acos (-1.) ;
  ```

- The six histograms can be put in one window/on one page with

  ```
  "window-name" -> Divide (3,2) ;
  ```

Presentation of results

HAND IN :

- A print of your program code
- A print of the histograms

5 The normal or Gaussian distribution

5.1 The characteristic function

Before we introduce the normal or Gaussian distribution and describe its special importance, we need the general concept of the characteristic function. A characteristic function $\Phi(t)$ associated wht the pdf $f(x)$ is the Fourier transform of $f(x)$ or the expectation value of e^{itx}. We can construct a complex random variable from two real ones by $z = x + iy$ and define its expectation value $E(z) = E(x) + iE(y)$. If x is a real random variable with pdf $f(x)$, then its characteristic function is defined as the expectation value of the complex variable e^{itx}:

$$\Phi_x(t) = E\{e^{itx}\} = \int_{-\infty}^{\infty} e^{itx} f(x)dx \tag{71}$$

Note that

$$\Phi_x(0) = 1 \qquad |\Phi_x(t)| \leq 1 \tag{72}$$

In reverse: the pdf $f(x)$ is completely determined by its characteristic function. By inverting the Fourier transform the pdf $f(x)$ can be obtained from $\Phi_x(t)$:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} \Phi_x(t)dt \tag{73}$$

Often it is more convenient to use the characteristic function rather than the original distribution. It is useful in calculating moments, e.g. deriving the expectation value or variance of a random variable: differentiating $\Phi(t)$ with respect to t and evaluating the result at $t = 0$ gives:

$$\frac{d^p\Phi(t)}{dt^p}|_{t=0} \quad = \quad \int_{-\infty}^{\infty} (ix)^p e^0 f(x)dx = i^p E(x^p)$$

From this we see that the first moment ($p = 1$)of the characteristic function is:

$$\Phi_x^{'}(t) = \int_{-\infty}^{\infty} ixe^{itx} f(x)dx \tag{74}$$

$$\Phi_x^{'}(t = 0) = i \int_{-\infty}^{\infty} xf(x)dx = iE(x) = i < x > \tag{75}$$

The second moment ($p = 2$)of the characteristic function is:

$$\Phi_x^{''}(t) = \int_{-\infty}^{\infty} (ix)^2 e^{itx} f(x)dx \tag{76}$$

$$\Phi_x^{''}(t = 0) = i^2 \int_{-\infty}^{\infty} x^2 f(x)dx = -E(x^2) = - < x^2 >= -\{\sigma^2(x) + < x >^2\} \tag{77}$$

In an similar way, the skewness is related to the value of the third derivative in $t = 0$ (the third moment), etc. The Taylor-expansion of the characteristic function can therefore be written as:

$$\Phi_x(t) = 1 \quad + \quad it < x > \quad + \quad \frac{1}{2}(it)^2 \ \{\sigma^2(x) + < x >^2\} \quad + \quad .... = \sum_{r=0}^{\infty} \frac{(it)^r}{r!} E(x^r) \tag{78}$$

The characteristic function can also be used to find the pdf of the sum of independent random variables: if x and y are independent variables, then:

$$\Phi_{x+y} = E\{e^{it(x+y)}\} = E\{e^{itx}e^{ity}\} = E\{e^{itx}\}E\{e^{ity}\} = \Phi_x\Phi_y \tag{79}$$

I.e. the characteristic function of a sum of independent random variables is equal to the product of their repective characteristic functions.
If a and b are constants, then:

$$\Phi_{ax+b}(t) = E\{e^{it(ax+b)}\} = E\{e^{itb}e^{iatx}\} = e^{itb}\Phi_x(at) \tag{80}$$

In special cases:
If $a = 1$ and $b = - < x >$, then:

$$\Phi_{x-<x>}(t) = e^{-it<x>}\Phi_x(t) \tag{81}$$

If $a = 1/\sigma(x)$ and $b = - < x > /\sigma(x)$, then:

$$\Phi_{(x-<x>)/\sigma(x)}(t) = e^{-it<x>/\sigma(x)}\Phi_x(t/\sigma(x)) \tag{82}$$

This describes the relationship between the characteristic functions of a variable and the corresponding normalised variable.

5.2 The Central Limit theorem

The Central Limit theorem states that if independent random variables $x_1, ..., x_n$ are distributed according to any pdf's with finite mean and variance, then the sum $y = \sum_{i=1}^{n} x_i$ will have a pdf that approaches a Gaussian distribution for large n. Mean and variance are given by the sums of corresponding terms from the individual x_i. Therefore, the sum of a large number of fluctuations x_i will be distributed as a Gaussian distribution, even if the x_i themselves are not. We will further discuss this theorem in relation with the statistical measurement error. Imagine that the measurement of a specific quantity is disturbed/deformed by a large number N of independent factors. Each factor causes a small shift in the measured value, away from the true value. This shift is a random variable from a probability distribution $f_i(x)$, with mean value $< x_i >$, and variance $\sigma^2(x_i)$. We now ask ourselves: "what will be the combined effect of the N shifts", in other words what is the effect of adding the N shifts? It is evident that the combined effect will (again) be a random variable y, with a certain probability distribution $f(y)$. We call the mean of this distribution $< y >$, and its variance $\sigma^2(y)$. From the sum rules for expectation values and variances, we know already that:

$$< y >= \sum_{i=1}^{N} < x_i > \qquad \sigma^2(y) = \sum_{i=1}^{N} \sigma^2(x_i) \tag{83}$$

We can say also something about the shape of the distribution $f(y)$, even if the underlying distributions $f_i(x)$ are not known. The shape of $f(y)$ becomes the bell shape of the standard normal distribution when $N \to \infty$. To simplify our 'proof' we assume that all the underlying distributions have expectation values $< x_i >= 0$ and variances $\sigma^2(x_i) = \sigma^2$. It follows that :

$$< y >= 0 \qquad \sigma^2(y) = N\sigma^2 \tag{84}$$

If $\Phi_i(t)$ are the characteristic functions of the underlying distributions, the characteristic function of the combined distribution is given by:

$$\Phi_y(t) = \prod_{i=1}^{N} \Phi_i(t) \tag{85}$$

We now change to the normalised variable $z = y/(\sigma\sqrt{N})$; the characteristic function of z is:

$$\Phi_z(t) = \Phi_y(\frac{t}{\sigma\sqrt{N}}) = \prod_{i=1}^{N} \Phi_i\left(\frac{t}{\sigma\sqrt{N}}\right) \tag{86}$$

Take the logarithm and approximate $\Phi_i(t)$ by its Taylor-expansion

$$ln\Phi_z(t) = \sum_{i=1}^{N} \; ln\Phi_i\left(\frac{t}{\sigma\sqrt{N}}\right) = N \;\; ln\left(1 + \frac{\sigma^2}{2}\frac{(it)^2}{N\sigma^2} + ...\right) \tag{87}$$

Now, use again the Taylor-expansion for

$$ln(1+\epsilon) = \epsilon - \epsilon^2/2 + ... \tag{88}$$

and exponentiate. It follows that

$$ln(\Phi_z) \simeq N\frac{\sigma^2}{2}\frac{(it)^2}{N\sigma^2} = -\frac{t^2}{2} \rightarrow \Phi_z(t) \simeq e^{-t^2/2} \tag{89}$$

This is the characteristic function of the standard normal distribution. Note that, by taking the value of the first and second derivatives of $\Phi_z(t)$ in $t = 0$, we obtain that: $< z >= 0$ and $\sigma(z) = 1$. The corresponding probability distribution is ($\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx = \sqrt{2\pi}$):

$$f(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz}e^{-t^2/2}dt = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} \tag{90}$$

In its general form, the Central Limit Theorem states that the sum of N variables with averages $< x_i >$ and variances σ_i^2 is described by a normal distribution with mean value $\sum_i < x_i >$ and variance $\sum_i < \sigma_i^2 >$.

5.3 The normal or Gaussian distribution

The Central Limit theorem (CLT) shows why the normal or Gaussian pdf is so important. Most of what we measure is the sum of many random variables. If the number of contributions is large the CLT tells us that their total sum is distributed according to the Gaussian distribution. This is often the case. The pdf $f(x)$ of the normal Gaussuan distribution with mean value $< x >= a$ and standard deviation σ is:

$$f(x; a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) \tag{91}$$

Figure 11 shows some examples of Gaussuan distributions. The normal distribution is characteristic for the statistical error in the measurement of quantities. The points of inflection in the bell shape

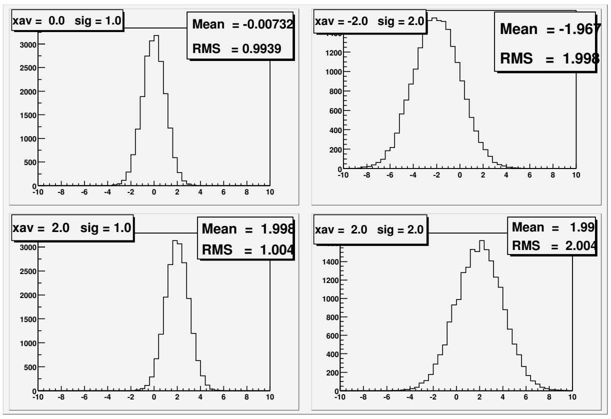


Figure 11: Normal distribution

of the curves are at $x = a \pm \sigma$. The probability that the measured value of a quantity lies within one standard deviation from the true value is:

$$P(|x-a| \leq \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\sigma}^{\sigma} e^{\left(-\frac{(x-a)^2}{2\sigma^2}\right)} dx = 0.6827 \tag{92}$$

This is what we usually call the statistical error in the measured value. Similarly:

$$P(|x-a| \leq 2\sigma) = 0.954 \qquad P(|x-a| \leq 3\sigma) = 0.998 \tag{93}$$

If we want to draw a conclusion with large certainty the quantity 3σ is often used as a measure for the uncertainty in our measurement. In summary: the normal or Gaussian pdf

$$P(x; <x>, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-<x>)^2}{2\sigma}} \tag{94}$$

is a bell shaped curve centred on and symmetric around its mean $<x>$. The width is controlled by its standard deviation σ. It is broad if σ is large, narrow if σ is small. At $x =< x > +\sigma$, $P(x)$ falls to 0.61 of its peak value (at a bit more than half). These are the points of inflection where the second derivative is zero. The Gaussian pdf is normalised to 1:

$$\int_{-\infty}^{\infty} P(x; <x>, \sigma) dx = 1 \tag{95}$$

The mean of the distribution is

$$\int_{-\infty}^{\infty} x P(x; <x>, \sigma) dx = <x> \tag{96}$$

which is also the mode and the median. Its variance is:

$$\int_{-\infty}^{\infty} (x- <x>)^2 P(x; <x>, \sigma) dx = \sigma^2 \tag{97}$$

5.4 The normal distribution for n-dimensional variables

The measured variable becomes a vector $\vec{x} = (x_1, x_2, ..., x_n)$. Also the mean values become a vector $\vec{a}$. The errors (and covariances between them) are contained in the covariance-matrix C the normal distribution is:

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n detC}} e^{\left(-\frac{1}{2}(\vec{x}-\vec{a})^T C^{-1} (\vec{x}-\vec{a})\right)} \tag{98}$$

5.5 The normal distribution for 2-dimensional variables

The covariance matrix C becomes

$$C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{99}$$

Its inverse is

$$C^{-1} = (\frac{1}{\sigma_1^2\sigma_2^2})(\frac{1}{1-\rho^2}) \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \tag{100}$$

Lines (or contours) of constant probability satisfy:

$$(\vec{x}-\vec{a})^T C^{-1} (\vec{x}-\vec{a}) = constant \tag{101}$$

Substitution of $\vec{x}$, $\vec{a}$ and C^{-1} gives (for these contours):

$$\frac{(x_1-a_1)^2}{\sigma_1^2} + \frac{(x_2-a_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-a_1)}{\sigma_1}\frac{(x_2-a_2)}{\sigma_2} = c * (1-\rho^2) \tag{102}$$

The covariance ellipse is defined as the contour with $c = 1$ (see example 4 in chapter 1). In this case the extreme values of the ellips are at $a_1 \pm \sigma_1$ and $a_2 \pm \sigma_2$. The larger the correlation between x_1 and x_2 the thinner the ellipse. The probability that the true values of (x_1, x_2) lie inside the covariance ellipse, is

$$1 - e^{-\frac{1}{2}} = 0.393 \tag{103}$$

The principal axis of the ellipse makes an angle α with the x_1-axis, given by

$$\tan(2\alpha) = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2} \tag{104}$$

If $\rho = 0$, the principal axis of the ellipse falls on top of the x_1-axis (if $\sigma_1 > \sigma_2$) or on top of the x_2-axis (if $\sigma_1 < \sigma_2$). If $\rho > 0$ then: $0 < \alpha < \pi/2$; if $\rho < 0$ then: $\pi/2 < \alpha < \pi$; if $\rho = 1$ or $\rho = -1$ the ellipse becomes a straight line with slope α. A typical example of a covariance ellipse/contour has been presented in chapter 1.

5.6 Exercises

Exercise 5.1 - *root* exercise 4

This exercise deals with the Central Limit Theorem (CLT).

Description

- The 'statement'

  ```
  x = gRandom -> Rndm (1) ;
  ```

  gives a random number from the uniform distribution between 0 and 1 .
  The expectation value of this distribution is $1/2$; its variance is $1/12$.

- We now draw two random numbers from the uniform distribution x_1 and x_2, and calculate their average: $u = (x_1 + x_2)/2$.
  We repeat this 5000 times, and display the distribution of the calculated u's in a histogram.
  From the sum rules of expectation values and variances we expect that :

  - the average value of u will be equal to 0.5 ,
  - the variance of u will be equal to $(1/2) \cdot (1/12)$ and
  - The standard deviation of u will be equal to $\sqrt{(1/2) \cdot (1/12)}$.

- From the CLT we expect that the distribution of the u's will start to look like a normal distribution (although with $N = 2$ we are still far away from the limit $N \rightarrow \infty$).

- We can verify the characteristics of the distribution of the u's with the *root* facilities.

  - Make and fill a histogram with the values of u.
  - After the histogram is drawn with the command

    ```
    "histogram name" -> Draw () ;
    ```

    you can fit and draw a normal distribution with the command

    ```
    "histogram name" -> Fit ("gaus") ;
    ```

  - the average value and standard deviation of the distribution are calculated automatically, and displayed together with the histogram.
  - *Compare these values with what you expect from the sum rules.*

- The CLT predicts that the distribution of the u's will look more and more like a normal distribution if we increase the value of N .
  Therefore repeat the previous procedure with $N = 8, 16$

  - Now $u = (1/N) \sum_{i=1}^{N} x_i$
  - The expectation value of u remains equal to 0.5
  - The standard deviation becomes $1./\sqrt{12 \cdot N}$

- The CLT is valid independent of the shape of the "starting" distribution.
  To verify this we no longer draw x_i from the uniform distribution, but from an exponential distribution $f(x) = e^{-x}$
  - The expectation value of this distribution is $\int_0^\infty xe^{-x}dx = 1$
    Its variance is $\int_0^\infty (x-1)^2 e^{-x}dx = 1$
    So the standard deviation is also equal to 1 .
  - The recipe to get random numbers x distributed according to the exponential distribution $f(x) = e^{-x}$ is as follows:
    - Get y from the uniform distribution between 0 and 1 .
    - Calculate $x = -ln(y)$
- Apply the procedure with $N = 2, 8, 16$ to the x_i drawn from the exponential distribution.

Presentation of results

HAND IN :

- A print of your program code
- A print of the histograms

6 Monte Carlo methods

6.1 Introduction

Monte Carlo methods are used for the numerical simulation of processes which contain an element of chance: an event may or may not occur, or the time/place for an event follows a probability distribution. Monte Carlo methods are used to simulate systems, devices, processes etc. on the computer. There are many applications: the design of nuclear reactors (geometry, shielding); the design of detector systems in particle physics by simulating the collisions one has to measure (resolution, acceptance); the design of networks for telecommunications; the design of systems for traffic regulation. All these applications have in common that one designs complicated systems, which are difficult to change 'afterwards'. Simulation is used to make optimal choices. Simulation in general is broader than Monte Carlo alone (wind tunnel, solving differential equations by iterative methods). Monte Carlo methods are also used to model and study physical systems, like solids (Ising model), gases (molecular dynamics) or collisions between subatomic particles (Lund model, QCD).

6.2 Example: a neutrino beam

6.2.1 Physics background

Imagine that we need a beam of neutrinos (high energy and intensity) in order to study the quark structure of protons and neutrons. The principles of such a beam are as follows: neutrinos and antineutrinos are produced in the decay of π-mesons and K-mesons:

decay mode	decay mode	probability
$\pi^+ \to \mu^+ + \nu_\mu$	$\pi^- \to \mu^- + \overline{\nu_\mu}$	1.00
$K^+ \to \mu^+ + \nu_\mu$	$K^- \to \mu^- + \overline{\nu_\mu}$	0.64
$K^+ \to \pi^+ + \pi^0 \to \mu^+ + \nu_\mu + \pi^0$	$K^- \to \pi^- + \pi^0 \to \mu^- + \nu_\mu + \pi^0$	0.21

The π- and K-mesons are produced by shooting a highly energetic beam of protons (coming from an accelerator) onto a target. They will move in the forward direction, but at arbitrary angles. Once the neutrinos are produced through the decay, their direction of motion can no longer be changed. So, one has to make sure that their 'parents' (π and K) are sharply focused in the forward direction. Conservation of momentum in the decay then guarantees that the neutrinos form a beam too. To obtain neutrinos one needs positively charged 'parents'; for antineutrinos the parents need to have negative charge. So: "wrongly" charged parents have to be 'removed'. The decay not only provides neutrinos, but also muons; these have to be stopped before they can reach the neutrino detector, in which the reaction products are detected.

6.2.2 Objectives of the Monte Carlo program

What do we want to find out with a Monte Carlo program? If all elements in the beam line are defined we want to know how the (anti) neutrino beam profile looks like when it reaches the detector and what its energy spectrum is. Is there any background in the beam (wrong type of neutrino)? Has the background of charged muons been reduced sufficiently? How do these properties depend on the design of the beam line? What is the best (and cheapest) way to build the beam line?

6.2.3 Ingredients of the Monte Carlo program

There are various ingredients of the Monte Carlo program: the proton beam has a certain spread around the beam axis so the location and direction of the proton when it enters the particle production target has to be choosen; the proton has a certain chance to interact inside the target: it has to be choosen whether the interaction occurs and where; when the proton interacts, π- and K-mesons are produced so, it has to be choosen how many mesons are produced and which charge and momentum each of them has. The parent particles move through the focusing part. The change in direction can be calculated directly; there is no element of chance here. In the decay tunnel a fraction of the parents will decay: the decay time must be choosen and its location must be calculated, the decay mode and the momenta of the decay products must be choosen. Follow all decay products to the neutrino detector and decide for each particle whether it is absorbed or not in the shielding.

6.3 The need for random numbers

Every time that a choice has to be made, our computer program has to generate a number for the corresponding quantity. This number has to be random, but it should be generated from a well defined probability distribution, which reflects the physical process. After the pdf that describes the hypothetical population has been choosen a random sample of the hypothetical population must be generated. Then from this random sample the parameter of the hypothetical population must be statistically estimated. Generation of random numbers according to a well defined probability distribution on a computer proceeds in two steps: First, generate a random number from a uniform distribution, between 0 and 1. Then transform this number as if it was generated according to the required (non-uniform) distribution. Literature: D.E. Knuth, The Art of Computer Programming (Volume 2).

6.4 Random numbers - uniformly distributed

6.4.1 The middle square method

To generate random numbers on a computer an algorithm (a recipe) is needed, which can be implemented in a function. An example is the 'middle square method' proposed (in 1946) by John von Neumann; it works as follows: take as first random number an arbitrary number, consisting of 10 digits, for example 5772156649. To get the second random number, square the previous one (this gives 33317792380594909201) and take the 6th to the 15th digit; this gives 7923805949. Subsequent numbers are obtained from the previous one, in the same way: square, and take the 6th to the 15th digit.
Question: each number is completely determined by the previous one; how can the series of numbers be random? Answer: this is true, computer generated series are not random, but pseudo random. By choosing a good algorithm it is possible to obtain a series of numbers which can be used for practical purposes as if they are random. Remark: if one uses numbers of 10 digits, one can never get more than 10^{10} different numbers; sooner or later one gets at a number which has occurred earlier. From that moment onwards the series will start to repeat itself. Comment: it is true that each algorithm has a maximum period, after which it will start to repeat itself. If the maximum period is much longer than the set of numbers which one uses, this will do no harm. The maximum period is determined by the number of digits which is used. Whether a specific algorithm reaches this maximum depends on the quality of the method. Consider as an example the middle square method with two digits. Its maximum period is 100. Start with 99:

$$99; (99^2 = 9801)- > 80; (80^2 = 6400)- > 40; (40^2 = 1600)- > 60; (60^2 = 3600)- > 60; etc.$$

Start with 98:

$$98; (98^2 = 9604) -> 60; (60^2 = 3600) -> 60; etc.$$

Start with 97:

$$97; (97^2 = 9409) -> 40; (40^2 = 1600) -> 60; etc.$$

The longest series is obtained when one starts with 42 or 69; in that case the series 'dies out' after 15 numbers.

6.4.2 The linear congruential sequence

A common algorithm used for pseudo random numbers is the linear congruential sequence

$$x_{n+1} = modulus\{(a * x_n + c), m\}$$

$x_0 > 0$ is an 'arbitrary' starting value (the seed), $a > 0$ is the multiplication factor, $c \geq 0$ is the increment and $(m - 1)$ is the maximum value of a number from the sequence. Remarks: m determines the maximum period of the sequence; its value is commonly chosen as a power of 2 (that makes calculating of the modulus easy); it is also chosen in relation to the word length on the computer (e.g. 32 bits). The choice of a (in connection with c) is extremely important for the quality of the sequence: how random is the result, and how close comes the period to the maximum one. Much theoretical work has been done to formulate a number of conditions on how a and c have to be chosen, if m is given. Apart from this, it is common practice to check the performance of a candidate generator using a number of criteria derived from statistics. Statistics predict a number of properties for series of truly random numbers; the tests verify if the generated numbers agree with them. The algorithm above gives numbers which are uniformly distributed between 0 and $(m - 1)$. Division by m gives a uniform number between 0 and 1. Usually generators for uniform distributions are available as mathematical functions, or on program libraries. The starting value x_0 is 'fixed' in the function itself. This is certainly not random, but it has the advantage that the program is reproducible. The disadvantage is, that rerunning the program does not produce new results. Usually another function is available which can be called to modify the starting value (the seed). The linear congruential method is 'good' for simple applications. For very 'demanding' applications more complicated (and better) algorithms have been developed.

6.5 Random numbers - not uniformly distributed

We have at our disposal a generator which produces random numbers distributed uniformly between 0 and 1. How do we use this to generate random numbers with a different probability distribution? Below we will desribe the algorithm for generating different distributions.

6.5.1 Discrete distributions

A binomial distribution: generate a random sequence of 0 and 1 such that $P(0) = P(1) = 0.5$. Algorithm: take u from the uniform generator. If $u < 0.5$, generate 0, if $u \geq 0.5$, generate 1.

Another binomial distribution: generate a random sequence of 0 and 1, such that $P(0) = 0.6$ and (therefore) $P(1) = 0.4$.

Algorithm: take u from the uniform generator. If $u < 0.6$, generate 0, if $u \geq 0.6$, generate 1.

A multinomial distribution: generate a random sequence of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ such that $P(0) = P(1) = P(2) = ... = P(9) = 0.1$.
Algorithm: take u from the uniform generator. If $u < 0.1$, generate 0, if $0.1 \leq u < 0.2$, generate 1, if $0.2 \leq u < 0.3$, generate 2, and so on.

Another multinomial distribution: generate a random sequence of numbers $x_1, x_2, ..., x_r$ with probabilities $P_1, P_2, ..., P_r$.
Algorithm: take u from the uniform generator. If $u < P_1$, generate x_1, if $P_1 \leq u < P_1 + P_2$, generate x_2, if $P_1 + P_2 \leq u < P_1 + P_2 + P_3$, generate x_3 and so on. In fact we are making use of the cumulative distribution:

$$F_{x_k} = \sum_{i=1}^{k} p_i = P(x < x_k) \quad (k = 1, 2, ..., r)$$

For a discrete probability distribution the cumulative distribution is a monotonous non-decreasing step function with $F(0) = 0$ and $F(r) = 1$ The algorithm is then (see figure 12a): take u from the uniform generator. Search the index k for which: $F(x_k) \leq u < F(x_{k+1})$. Generate: x_k.

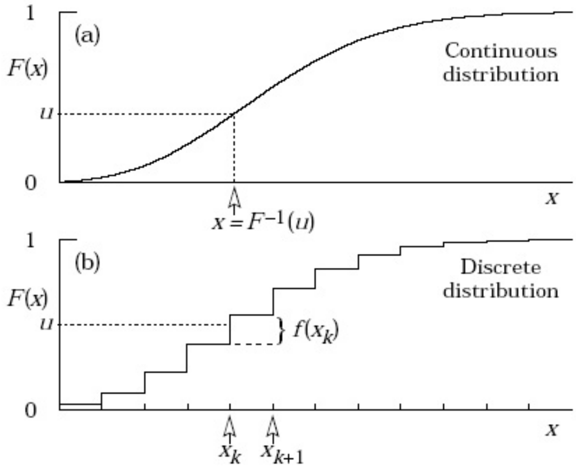


Figure 12: Generation random numbers using the cumulative distribution function

6.5.2 Continuous distributions

It is simple to change the previous algorithm into a procedure to generate random numbers according to a continuous distribution $f(x)$ on the interval (x_{min}, x_{max}) (this is called the Inverse transform method):
Determine the cumulative distribution

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

Take u from the uniform generator and generate x by solving $u = F(x) \rightarrow x = F^{-1}(u)$ (see figure 12b.).

An example: the decay of a particle with lifetime τ_0 is described by the probability distribution

$$f(t) = \frac{1}{\tau_0} e^{-\frac{t}{\tau_0}} \qquad \tau_{min} = 0 \qquad \tau_{max} = \infty \tag{105}$$

The cumulative distribution is:

$$F(t) = \frac{1}{\tau_0} \int_0^t e^{-\frac{x}{\tau_0}} dx = 1 - e^{-\frac{t}{\tau_0}} \tag{106}$$

Algorithm: take x from the uniform generator and calculate t by solving: $x = F(t)$:

$$t = F^{-1}(x) = -\tau_0 ln(1-x) \tag{107}$$

Remark: if x is uniform between 0 and 1, so is $(1-x)$. So we can also use:

$$t = -\tau_0 ln(x) \tag{108}$$

Another example: the 'profile' of a particle beam (the density distribution perpendicular to the beam axis) is described by

$$f(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \qquad r_{min} = 0 \qquad r_{max} = \infty \tag{109}$$

The cumulative distribution is

$$F(r) = \int_0^r \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx = 1 - e^{-\frac{r^2}{2\sigma^2}} \tag{110}$$

Algorithm: take x from the uniform generator and calculate r by solving : $x = F(r)$:

$$r = \sigma \cdot \sqrt{-2 \cdot ln(1-x)} \quad \rightarrow \quad r = \sigma \cdot \sqrt{-2 \cdot ln(x)} \tag{111}$$

Another example (with a problem): in the decay $\mu^- \rightarrow e^- \nu_\mu \overline{\nu}_e$ the energy distribution of the e^- in the rest system is given by

$$f(y) = 6y^2 - 4y^3 \qquad y = \frac{E}{E_{max}} = \frac{E}{52.8MeV} \tag{112}$$

The cumulative distribution is:

$$F(y) = \int_0^y (6x^2 - 4x^3) dx = 2y^3 - y^4 \tag{113}$$

Algorithm: take x from the uniform generator and calculate y by solving : $x = F(y)$; however, this equation can not be solved analytically.

Another example (with a problem): the standard (normal) distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{114}$$

In this case the cumulative distribution has no analytical form.

6.6 Special methods

When the algorithm with the cumulative method does not work, there are alternatives.

6.6.1 Interpolation in a table

In this method the cumulative distribution $F(x)$ (corresponding to $f(x)$) is replaced by a table, and the equation $u = F(x)$ is solved through interpolation between table values. Divide the interval (x_{min},x_{max}) on which the function is defined into N intervals of equal length Δx and calculate

$$F(x_{min} + k\Delta x) \qquad (k = 0, 1, ..., N) \tag{115}$$

In the end points we have

$$F(x_{min}) = 0 \qquad F(x_{max}) = 1 \tag{116}$$

Between the end points we have

$$F(k) = \int_{x_{min}}^{x_{min}+k\Delta x} f(t)dt \tag{117}$$

If $f(x)$ can be integrated analytically this can be calculated directly. If $f(x)$ can not be integrated analytically, $F(k)$ can be approximated numerically:

$$F_k = \Delta x \sum_{i=1}^{k} f\left(x_{min} + \frac{(2i-1)\Delta x}{2}\right) \qquad (k = 1, 2, ..., N) \tag{118}$$

Then take u from the uniform generator and find (in the table) the index j for which:

$$F(x_{j-1}) \leq u < F(_j) \tag{119}$$

Find the corresponding value of x by linear interpolation

$$x = x_{min} + (j-1)\Delta x + \frac{x - F(_{j-1})}{F(x_j) - F(x_{j-1})}\Delta x \tag{120}$$

In this way we find values of x which are distributed according to $f(x)$.

6.6.2 The acceptance-rejection method

A completely different approach (which does not need the knowledge of the cumulative distribution), is the acceptance-rejection method. We illustrate this method with the distribution which was given earlier:

$$f(x) = 6x^2 - 4x^3 \quad 0 \ \leq \ x \ \leq \ 1 \tag{121}$$

Generate pairs of random numbers (x_i, u_i) where x_i is uniformly distributed in the interval available to x $(0 \leq x \leq 1)$ and u_i is uniformly distributed in the range of values assumed by the function $f(x)$ $(0 \leq f(x) \leq 2)$. For each pair (x_i, u_i) we test if $u_i \geq f(x_i)$. If this is true we reject x_i. The set of random numbers x_i that are not rejected then follow a probability density $f(x)$. We can also apply a scale transformation such that the interval (x_{min}, x_{max}) becomes $(0, 1)$ (in our example this is already satisfied) and normalise $f(x)$ such that the maximum value of $f(x)$ on the interval $(0, 1)$ is equal to 1. This would give:

$$f(x) = 3x^2 - 2x^3 \tag{122}$$

Then generate a pair of numbers from the uniform distribution between 0 and 1: (x_1, x_2). Calculate $f(x_1)$. If $f(x_1) \geq x_2$ accept x_1; if $f(x_1) < x_2$ reject x_1. In fact, in this method our function is 'enclosed' in a square. The numbers (x_1, x_2) represent a random point within this square. If the point lies below the function, we accept x_1; if the point lies above the function, we reject x_1. It should be clear from this that acceptance or rejection of x_1 must always be followed by taking a new pair (x_1, x_2).

6.6.3 Methods for the normal distribution

The method of interpolation in a table and the acceptance-rejection method can be 'expensive' in terms of computing time. For some, frequently occurring, probability distributions special methods have been developed. We will mention a special method for the standard normal distribution. The so called direct method works as follows: take x_1 and x_2 from the uniform distribution between 0 and 1. Calculate:

$$y_1 = \sqrt{-2lnx_1} \cdot \cos(2\pi x_2) \qquad y_2 = \sqrt{-2lnx_1} \cdot \sin(2\pi x_2) \tag{123}$$

It can be shown that y_1 and y_2 are independent and follow the standard normal distribution (mean 0, $\sigma = 1$). This method is exact but not very fast; the fast direct method works as follows: take x_1 and x_2 from the uniform distribution between 0 and 1 and calculate d as follows:

$$u_1 = 2x_1 - 1 \qquad u_2 = 2x_2 - 1 \qquad d = u_1^2 + u_2^2 \tag{124}$$

If $d > 1$ reject u_1 and u_2. If $d \leq 1$ calculate

$$y_1 = u_1\sqrt{-2ln(d)/d} \qquad y_2 = u_2\sqrt{-2ln(d)/d} \tag{125}$$

It can be shown (again) that y_1 and y_2 are independent and follow the standard normal distribution. Transformation from the standard normal distribution to a normal distribution with mean value a and width σ goes as follows: generate a number y from the standard normal distribution. Transform $z = a + \sigma \cdot y$.

6.7 Method for the binomial distribution

To generate random numbers from a binomal pdf

$$f(r; n, p) \quad = \quad \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

the following algorithm can be used: Start with $k = 1$ and generate a random number u uniform in [0,1). Compute $P_k = (1-p)^n$ and strore P_k in B. If $u \leq B$ accept $r_k = k$ and stop. If not, compute the $P_{k+1} = P_k \cdot (1-p)^n \frac{p}{1-p} \frac{in-(k+1)}{(k+1)+1}$ and add this to B. If $u \leq B$ accept $r_k = k$ and stop. If not, iterate: calculate $P_{k+2} = P_{k+1} \cdot (1-p)^n \frac{p}{1-p} \frac{in-(k+2)}{(k+2)+1}$ etc.

6.8 Method for the Poisson distribution

To generate random numbers from a Poisson pdf

$$f(n : \lambda) \quad = \quad \frac{\lambda^n e^{-\lambda}}{n!}$$

the following algorithm can be used: Begin with $k = 1$ and $A = 1$. Generate random number u uniform in [0,1). If $uA < e^{-\lambda}$ accept $n_k = k - 1$ and stop. If not, store uA in B and generate a new random number u and calculate uB. If $uB < e^{-\lambda}$ accept $n_{k+1} = (k+1) - 1$ and stop. If not, iterate: replace B with uB, generate a new random number u, calculate uB, if $uB < e^{-\lambda}$ accept $n_{k+2} = (k+2) - 1$ etc.

6.9 Example of Monte Carlo generation

Consider the pdf

$$f(x) = 1 + x^2; \quad -1 < x < 1 \tag{126}$$

This could be an angular distribution with $x = cos\theta$. We will generate random samples of the pdf using different methods.

Weighted events

Generate x_i uniformly in the region $(-1, 1)$ and assign a weight $w_i = 1 + x_i^2$.

Rejection method

Generate x_i uniformly in the region $(-1, 1)$ and r_i in the region $(0, 2)$. Reject the point (x_i, r_i) if $r_i > 1 + x_i^2$.

Inverse transformation method

We have the cumulative distribution

$$F(x) = \int_{-1}^{x} (1 + x^2) dx = x + \frac{x^3}{3} |_{-1}^{x} = x + \frac{x^3}{3} + \frac{4}{3} \tag{127}$$

Hence $F(-1) = 0$ and $F(1) = \frac{8}{3}$. Take u from the uniform generator and calculate $\frac{8}{3}u$, which is then uniformly distributed on $[0, \frac{8}{3}] = [F(-1), F(1)]$. The corresponding value of x is calculated by solving

$$\frac{8}{3}u = F(x) = x + \frac{x^3}{3} + \frac{4}{3} \tag{128}$$

The solution is $x_i = A + B$ with $A = (4u - 2 + s)^{\frac{1}{3}}$, $B = (4u - 2 - s)^{\frac{1}{3}}$, $s = \sqrt{1 + 4(1 - 2u)^2}$

6.10 Exercises

Exercise 6.1 - *root* exercise 5

Problem statement

This exercise deals with the generation of random numbers from several (non-uniform) distributions. We do this using the example of the simulation of a neutrino-beam which has been used at CERN. We use a (very) simplified setup of the beam.

- We assume that at the beginning of the decay tunnel we have a beam of π^+- and K^+-mesons (86 % π^+ and 14 % K^+)
- We assume that the beam is *point like*, and that all particles move along the beam-axis (we call this the x-axis).
- The momentum of the particles follows a normal distribution with an average value of 200 GeV/c and a width of 10 GeV/c.
- Of the possible decay-modes, which produce neutrino's, we only consider
  $\pi^+ \rightarrow \mu^+ + \nu_\mu$ and $K^+ \rightarrow \mu^+ + \nu_\mu$
- The length of the decay tunnel is 300 meter. Particles which do not decay within this distance are absorbed, and can no longer produce neutrinos.
- At a distance of 400 meter behind the decay-tunnel (so 700 meter behind the point where the π^+- and K^+-mesons are created) we have installed our detector, where interactions between the neutrino's and protons/neutrons in the detector material can take place. Our detector can be represented as a circular disc with a radius of 1.50 meter, perpendicular to the beam-axis, and with its centre on the beam-axis.
- Your task is:
  - Calculate the energy-spectrum of the neutrino's which pass through the detector. Do this separately for the neutrino's from the π^+-decay and from the K^+-decay.
  - Check if there is a correlation between the energy of the neutrinos and on the radial position where they hit the detector.

Supplementary (physical) data/facts

- The time-spectrum for the decay of π^+ and K^+ is described by :

  $$f(t) = \frac{1}{\tau_0} e^{-\frac{t}{\tau_0}}$$

  The π^+ lifetime $\tau_0 = 2.603 \cdot 10^{-8}$ s. The K^+ lifetime $\tau_0 = 1.237 \cdot 10^{-8}$ s.
- τ_0 is the lifetime of the particle in its rest system. The distance s travelled in the laboratory system by a particle with momentum p and mass m during a time t is given by

  $$s = \frac{p}{m} ct$$

  where :
  $c = 2.9979 * 10^8$ m/s , the mass of the $\pi^+ = 0.1396$ GeV/c^2, the mass of the $K^+ = 0.4937$ GeV/c^2 and p is the momentum of the decaying particle (in GeV/c). s the distance in m.

- To calculate the momentum components of the neutrino we first go to the **rest system** of the π^+ (or K^+). The conservation laws of momentum and energy give the following equations (we take the π^+ as example)

$$\vec{p}_\mu = -\vec{p}_\nu \qquad m_\pi = \sqrt{p_\mu^2 + m_\mu^2} + \sqrt{p_\nu^2 + m_\nu^2}$$

It follows that (note that we assume $m_\nu = 0.$) :

$$|p_\nu| = \frac{m_\pi^2 - m_\mu^2}{2m_\pi}$$

$|p_\nu|$ can be calculated using $m_\mu = 0.1057 \;\; GeV/c^2$

- The value of $|p_\nu|$ is fixed now , but the direction remains to be calculated. In this case the decay is isotropic (in the rest system) : all directions have equal probability. Moreover there is a symmetry around the beam (x)-axis. To generate this you may proceed as follows :
  - generate $\cos\theta$ from a uniform distribution between -1 and $+1$
  - calculate the transverse momentum component (perpendicular to the beam-axis) p_t, and the longitudinal component p_l

$$p_l = p\cos\theta \qquad p_t = p\sin\theta$$

- We now have for the neutrino

$$p_l \quad p_t \quad |p| \quad E(= |p|)$$

But: these quantities have to be transformed from the rest system to the laboratory system, using a Lorentz transformation. Since we have chosen the beam-axis as x-axis the Lorentz transformation looks as follows

$$\begin{pmatrix} p_l \\ p_t \\ E \end{pmatrix}_{lab-system} = \begin{pmatrix} \gamma & 0 & \beta\gamma \\ 0 & 1 & 0 \\ \beta\gamma & 0 & \gamma \end{pmatrix} \begin{pmatrix} p_l \\ p_t \\ E \end{pmatrix}_{rest-system}$$

β is the velocity of the decaying particle in the laboratory system, expressed in units of the speed on light. For the π-meson :

$$\beta = \frac{|p_\pi|}{\sqrt{p_\pi^2 + m_\pi^2}} \qquad \gamma = \frac{1}{\sqrt{1 - \beta^2}}$$

- We now have the energy of the neutrino in the laboratorium system. We already had the coordinates of the point where the neutrino has been produced. And we know the direction of motion of the neutrino. What remains to be done is to calculate the radial position where the neutrino hits the detector.

Useful intermediate checks

The final result (scatter plots for energy versus radial position in the detector for neutrinos from π^+- and K^+-decay) is reached only after a number of intermediate steps. To verify the correctness of these steps you can perform the following checks

- Plot the momentum of the decaying π- and K-mesons in a histogram. You should get the appropriate normal distribution.
- Plot the calculated decay distances, ignoring the finite length of the decay tunnel. If you take e.g. a maximum decay distance of 10000 meter, you should find an exponential distribution. Also you should find a difference between π- and K-decay.
- Plot the momentum distribution of the neutrinos in the laboratorium system, (ignoring the finite size of the detector). You should find a *flat* distribution.
For neutrinos from π-decay the maximum momentum is about 100 GeV/c.
For neutrinos from K-decay the maximum momentum is about 200 GeV/c.
- When you bring in the finite size of the detector, neutrinos with low momentum will get 'lost'.

Presentation of results

HAND IN :

- A print of your program code
- A print of the histograms and the scatter plots

7 Parameter estimation

7.1 Introduction

So far we calculated the probability of a set of measured values once we knew the pdf that was appropiate to the problem. Now we will address the inverse: if we have a set of measured values which are sampled from a parent pdf, what does this pdf look like? E.g. if we assume a normal distribution for the pdf, can we determine from the measured values its mean and the standard deviation? Or: if we are given that a coin has probability p of landing heads and $(1-p)$ of landing tails, we may ask what is the probability of getting r heads in n trials. The answer is $f(r;n,p) = (nk)p^r(1-p)^{n-r}$ as we have seen before. This is a problem in probability. A problem in statistics would start with the observation of r heads in n trials and ask what is the probability p getting a head on a single row. The answer is not so sharp. A priori we can only state that $0 \leq p \leq 1$. If $r \neq 0$ we can eliminate $p = 0$ and if $r \neq n$ we can also eliminate $p = 1$. Most likely $p = \frac{r}{n}$. Finding the pdf is called parameter estimation. We consider two practical examples.

7.1.1 Example 1: measurement of a single quantity

We measure the value of some quantity X with a measuring device; the result is a number x_1. From experience we know that each measurement has a measurement error. Therefore the measured value x_1 might deviate somewhat from the true value of X. The common way to deal with this is that we repeat our measurement a number of times; this results in a series of numbers x_i $(i = 1, 2, ..., N)$. The numbers will be (slightly) different, since the effect of the measurement error is different in each measurement. When we have finished our measurements, we have to formulate a conclusion about the true value of the quantity X. The common procedure is that we calculate the mean value of all x_i's and present that number as the best estimate of the true value of X. We use the sample mean as the estimate of the true value of X:

$$\tilde{x} =< x >= \frac{1}{N}\sum_{i=1}^{N} x_i \tag{129}$$

We are aware that $\tilde{x}$ can still deviate from X, since we only have done N measurements. To determine the magnitude of this deviation, we compute the standard deviation σ_x of the measured x_i's.

$$\sigma(\tilde{x}) = \sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - < x >)^2} \tag{130}$$

We call $\sigma(\tilde{x})$ the measurement error, and use it as a measure to indicate how far our measured value $\tilde{x}$ could deviate from the true value of X. Where does this come from? How do we justify this?

7.1.2 Example 2: measurement of the lifetime of a particle

The decay of an elementary particle is described by:

$$f(t) = \frac{1}{\tau_0} e^{\left(-\frac{t}{\tau_0}\right)} \tag{131}$$

The lifetime τ_0 is characteristic for the particle. Suppose that we observe the decay of N identical particles, and measure the decay time of each of them. The result is: $t_1, t_2, ..., t_N$. When we

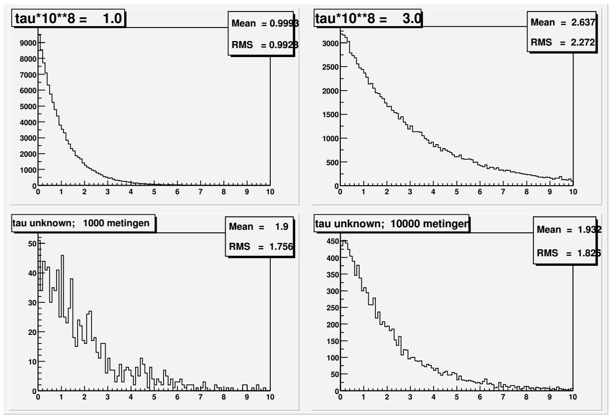


Figure 13: (Exponential) Decay-time distributions

display the result in a histogram (see figure 13) we will see an exponential distribution. Of course this distribution will contain information about the lifetime τ_0 of the decaying particle. The question is: how do we obtain (from the measurements) the best estimate for the lifetime τ_0? Moreover, we are aware that our best estimate might still deviate from the true value. So we need another number to estimate that deviation (the statistical error). How do we do this? How do we justify our method?

7.2 About estimators

In our framework, a series of measurements results in a set of numbers which are random drawings from a (well-known) probability distribution, which is characterised by one or more parameters. An estimator is a method or an algorithm which we apply to the experimental results, in order to obtain an estimate for the true value of the parameters of the corresponding probability distribution (its mean, its variance, the lifetime of the particle, etc). Moreover, we need an algorithm to calculate how far the estimated value could deviate from the true value (the statistical error of our estimated value). In order to define a set of desired properties of an estimator, we imagine that we take several series of measurements of the same quantity. Application of the estimator to the various sets of measurements, will result in a set of estimates. These estimates have (again) a probability distribution, with mean value, variance, etc. A good estimator has to satisfy three criteria, which are related to the probability distribution of the estimated values; we will use the estimation of a particle's lifetime to illustrate them. The individual decay times are drawn from the pdf:

$$f(t) = \frac{1}{\tau_0} e^{\left(-\frac{t}{\tau_0}\right)} \tag{132}$$

where τ_0 is the true lifetime. We observe the decay times of n particles: $(t_1, t_2, ..., t_n)$. We need an estimator $S(t_1, t_2, ..., t_n)$ which, applied to the measurements, gives an estimate $\tilde{\tau_0}$ for the lifetime. We say that our estimator is good if it has the following features:
The first criterium states that the estimator must be unbiased:

$$E\{\tilde{\tau_0}\} = \tau_0 \tag{133}$$

i.e. the most likely value for the estimate is equal to the true value of the estimated parameter. The bias of an estimator is defined as the difference between the expectation value from the estimator (the estimate) and the true value: $b(\tau_0) = E[\tilde{\tau_0}] - \tau_0 = E[\tilde{\tau_o} - \tau_0]$. The estimator is unbiased if the

bias is zero.
The second criterium states that the estimator is consistent. The estimate $\tilde{\tau_0}$ converges to the true value of τ_0 as the number of measurements increases, i.e.

$$\lim_{n\to\infty} \sigma^2(\tilde{\tau_0}) = 0 \tag{134}$$

i.e. the statistical error of the estimated value decreases if the number of measurements increases. Finally, the estimator must be efficient. Efficiency is the reverse of the ratio of the variance $\sigma^2(\tilde{\tau_0})$ to its minimum possible value. $\sigma^2(\tilde{\tau_0})$ is as small as possible. i.e. the estimator uses all the information, which is available in the measurements; we will come back to this later in this chapter.
Application: a good estimator for the lifetime. We know that:

$$\int_0^\infty \frac{t}{\tau_0} e^{\left(-\frac{t}{\tau_0}\right)} dt = \tau_0 \tag{135}$$

So, the lifetime is identical to the mean value of the (exponential) probability distribution and the estimation of the lifetime is reduced to the problem of the estimation of the mean value of a distribution. This will be treated in the next section.

7.3 Confidence Intervals

We may express an estimate $\tilde{\tau_0}$ of a parameter τ_0 as $P(|\tilde{\tau_0} - \tau_0| \geq \delta) \leq a$ or $P(|\tilde{\tau_0} - \tau_0| \leq \delta) \geq 1-a$. In words, the probability is greater than or equal to $(1-a)$ that parameter τ_0 is included in the interval $[\tilde{\tau_0} - \delta, \tilde{\tau_0} + \delta]$. Or, with $100 \cdot (1-a)$ percent confidence, the interval includes τ_0, i.e. a large number of repititions of the measurements that yielded $\tilde{\tau_0}$ will result in the interval including τ_0, $100 \cdot (1-a)$ percent of the time. We call this interval the confidence interval at the level of $100 \cdot (1-a)$ percent confidence. Frequently estimates of parameters are quoted in the form: $\tau_0 = \tilde{\tau_0} \pm (\sigma(\tilde{\tau_0})$. This usually means that the interval $[\tilde{\tau_0} - \sigma(\tilde{\tau_0}), \tilde{\tau_0} + \sigma(\tilde{\tau_0})]$ represents a 65% confidence level if the estimate $\tilde{\tau_0}$ is assumed to be normally distributed.

7.4 Estimating the mean value of a distribution

We have a sample of independent measurements $(x_1, x_2, ..., x_n)$ from an (arbitrary) probability distribution $f(x)$, with mean value $< x >$ and variance $\sigma^2(x)$. We estimate the mean value of the distribution from the sample values with the estimator

$$S_x = \tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{136}$$

$\tilde{x}$ is the sample mean (i.e. the arithmetic mean of the sample values). Verify the bias:

$$\begin{aligned} E\{\tilde{x}\} &= E\left\{\frac{1}{n}\sum_{i=1}^{n} x_i\right\} \\ &= \frac{1}{n} E\left\{\sum_{i=1}^{n} x_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\{x_i\} =< x > \end{aligned}$$

Verify the consistency:

$$\begin{aligned}
\sigma^2(\tilde{x}) &= E\{(\tilde{x} - E\{\tilde{x}\})^2\} \\
&= E\{(\tilde{x} - <x>)^2\} \\
&= E\left\{\left(\frac{x_1 + x_2 ... + x_n}{n} - <x>\right)^2\right\} \\
&= \frac{1}{n^2} E\{(x_1 - <x>)^2 + (x_2 - <x>)^2 ... + (x_n - <x>)^2\} \\
&= \frac{1}{n^2} n\sigma^2(x) = \frac{1}{n}\sigma^2(x)
\end{aligned}$$

Thus $\lim_{n\to\infty} \sigma^2(\tilde{x}) = 0$. In this derivation we used that $(x_1, x_2, ..., x_n)$ are independent.
Verify the efficiency: this will (can only) be done later.

7.5 Estimating the variance of a distribution

We have a sample of independent measurements $(x_1, x_2, ..., x_n)$ from an (arbitrary) probability distribution $f(x)$, with mean $<x>$ and variance $\sigma^2(x)$. We estimate the variance from the sample values with the estimator

$$S_{\sigma^2(x)} = \tilde{\sigma}^2 = \sigma^2(\tilde{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x})^2 \tag{137}$$

where $\tilde{\sigma}^2$ is the sample variance and $\tilde{x}$ is the the sample mean.
Verify the bias of the estimator:

$$\begin{aligned}
E\{\tilde{\sigma}^2\} &= E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x})^2\right\} \\
&= \frac{1}{n}E\left\{\sum_{i=1}^{n}(x_i - <x> + <x> - \tilde{x})^2\right\} \\
&= \frac{1}{n}E\left\{\sum_{i=1}^{n}(x_i - <x>)^2 + \sum_{i=1}^{n}(<x> - \tilde{x})^2 + 2\sum_{i=1}^{n}(x_i - <x>)(<x> - \tilde{x})\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[E\{(x_i - <x>)^2\} - E\{(\tilde{x} - <x>)^2\}\right] \\
&= \frac{1}{n}\left\{n\sigma^2(x) - n\frac{1}{n}\sigma^2(x)\right\} \\
&= \frac{n-1}{n}\sigma^2(x) = \sigma^2(\sigma^2(\tilde{x}))
\end{aligned}$$

So, this estimator is biased, although the bias becomes small for large values of n. Since we can see directly the size of the bias we can change our estimator (our sample variance) to become unbiased (Bessel-correction):

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \tilde{x})^2 \tag{138}$$

Verify the consistency of the estimator: it can be calculated, that the variance of the estimate is given by:

$$V(\tilde{\sigma}^2) = V(\sigma^2(\tilde{x})) = \frac{1}{n}\left[E\{(x - <x>)^4\} - E\{(x - <x>)^2\}\right] \tag{139}$$

So, the estimator is consistent, i.e. $lim_{n\to\infty}V(\tilde{\sigma}^2) = lim_{n\to\infty}V(\sigma^2(\tilde{x})) = 0$.

7.6 The concept of likelihood

In order to introduce the requirement of efficiency of our estimator and to make things more general, we have to introduce the concept of likelihood. We will use again the decay of a particle as an example:

$$f(t) = \frac{1}{\tau_0}e^{\left(-\frac{t}{\tau_0}\right)} \tag{140}$$

We treat the lifetime of the particle as an unknown quantity λ. The probability distribution becomes a function of t and λ:

$$f(t;\lambda) = \frac{1}{\lambda}e^{\left(-\frac{t}{\lambda}\right)} \tag{141}$$

We measure a series of decay times $(t_1, t_2, .., t_n)$ which contains information about the true value of λ. The concept of likelihood is defined as follows: our first measurement yielded a decay time t_1. The probability to get that result is:

$$P(t_1;\lambda) = f(t_1;\lambda) = \frac{1}{\lambda}e^{\left(-\frac{t_1}{\lambda}\right)} \tag{142}$$

Our second measurement yielded a decay time t_2. The probability to get that result is:

$$P(t_2;\lambda) = f(t_2;\lambda) = \frac{1}{\lambda}e^{\left(-\frac{t_2}{\lambda}\right)} \tag{143}$$

The combined probability to find t_1 and t_2 is:

$$P(t_1;\lambda)\cdot P(t_2;\lambda) \tag{144}$$

In general: the combined probability to find, in n measurements, the decay times $t_1, t_2, ..., t_n$ is:

$$L(t_1, t_2, ..., t_n;\lambda) = \prod_{i=1}^{n} P(t_i;\lambda) = \prod_{i=1}^{n} f(t_i;\lambda) = \prod_{i=1}^{n} \frac{1}{\lambda}e^{\left(-\frac{t_i}{\lambda}\right)} \tag{145}$$

This is called the likelihood function. Remark that:

$$\int L(t_1, t_2, ...t_n;\lambda)dt_1dt_2...dt_n = 1 \qquad \forall\lambda \tag{146}$$

7.7 The efficiency of an estimator

One of the requirements for a good estimator is that it should be efficient, i.e. the uncertainty/error of the estimated parameter should be as small as possible. With the concept of likelihood we can prove that there is a lower limit to the error: it is impossible to find an estimator which does a better job. We have a set of measurements $(t_1, t_2, ..., t_n)$. The corresponding likelihood is $L(t_1, t_2, ..., t_n;\lambda)$. We have an estimator for λ: $S_\lambda(t_1, t_2, ..., t_n)$ which gives an estimated value $\tilde{\lambda}$. The expectation value for $\tilde{\lambda}$ is:

$$E(\tilde{\lambda}) = <\tilde{\lambda}> = \int S_\lambda(t_1, t_2, ..., t_n)L(t_1, t_2, ...t_n;\lambda)dt_1dt_2...dt_n = \int S_\lambda \cdot LdT \tag{147}$$

If the estimator is unbiased then:

$$E(\tilde{\lambda}) = \int S_\lambda \cdot L dT = \lambda \tag{148}$$

Differentiate with respect to λ and rewrite:

$$\int S_\lambda \cdot \frac{dL}{d\lambda} dT = 1 \quad \rightarrow \quad \int S_\lambda \cdot L \cdot \frac{d(lnL)}{d\lambda} dT = 1 \tag{149}$$

We already had the normalisation $\int LdT = 1$. Differentiate this with respect to λ and rewrite:

$$\int \frac{dL}{d\lambda} dT = 0 \quad \rightarrow \quad \int L \cdot \frac{d(lnL)}{d\lambda} dT = 0 \quad \rightarrow \quad \int \lambda \cdot L \cdot \frac{d(lnL)}{d\lambda} dT = 0 \tag{150}$$

Combination of equations 149 and 150 gives:

$$\int (S_\lambda - \lambda) \cdot L \cdot \frac{d(lnL)}{d\lambda} dT = 1 \tag{151}$$

We now use the inequality of Cauchy-Schwartz. In its general form, this inequality states:

$$\left\{ \int u^2(x)dx \right\} \cdot \left\{ \int v^2(x)dx \right\} \geq \left\{ \int u(x) \cdot v(x)dx \right\}^2 \tag{152}$$

Applied with:

$$u = (S_\lambda - \lambda) \cdot \sqrt{L} \qquad v = \frac{d(lnL)}{d\lambda} \cdot \sqrt{L} \tag{153}$$

we obtain:

$$\left\{ \int (S_\lambda - \lambda)^2 \cdot LdT \right\} \quad \cdot \quad \left\{ \int \left(\frac{d(lnL)}{d\lambda} \right)^2 \cdot LdT \right\} \geq 1 \tag{154}$$

or

$$< (S_\lambda - \lambda)^2 > \cdot < \left(\frac{d(lnL)}{d\lambda} \right)^2 > \geq 1 \tag{155}$$

or

$$V(S_\lambda) \cdot I(\lambda) \geq 1 \tag{156}$$

where $V(S_\lambda)$ is the variance of the estimator for λ:

$$V(S_\lambda) = \int (S_\lambda - \lambda)^2 \cdot LdT \tag{157}$$

and $I(\lambda)$ is called the information about λ, which is contained in the measurements.

$$I(\lambda) = \int \left(\frac{d(lnL)}{d\lambda} \right)^2 \cdot LdT \tag{158}$$

The consequence of Cauchy-Schwartz is

$$V(S_\lambda) \geq I^{-1}(\lambda) \tag{159}$$

So, the more information about λ is provided by the measurements the smaller the variance of the estimator will be. The quantity $I^{-1}(\lambda)$ is a lower limit to the precision which can be obtained from the data set about the value of λ. It is called the Minimum Variance Bound (MVB). Alternative expressions for the information are (see textbook):

$$I(\lambda) \;=\; \int \left\{\frac{d(lnL)}{d\lambda}\right\}^2 \cdot LdT = E\left\{\left(\frac{d(lnL)}{d\lambda}\right)^2\right\}$$

or, using a slightly different theoretical approach in the derivation (see textbook)

$$I(\lambda) \;=\; -\int \frac{d^2(lnL)}{d\lambda^2} \cdot LdT = -E\left\{\frac{d^2(lnL)}{d\lambda^2}\right\}$$

It is also possible to express $I(\lambda)$ in terms of the underlying probability distribution (see textbook):

$$I(\lambda) = n \cdot E\left\{\left(\frac{d(lnf(x;\lambda))}{d\lambda}\right)^2\right\} = -n \cdot E\left\{\frac{d^2(lnf(x;\lambda))}{d\lambda^2}\right\} \tag{160}$$

Note that the information $I(\lambda)$ increases linearly with the number of measurements n and that $I(\lambda)$ only depends on $f(t;\lambda)$. So, the MVB is independent of the chosen estimator. In our derivation we have never used an explicit form for the estimator S_λ. Therefore, the MVB is valid for whatever estimator one can think of. We call an estimator efficient if it reaches the MVB, i.e. if

$$V(S_\lambda) = I^{-1}(\lambda) \tag{161}$$

In terms of the Cauchy-Schwartz inequality this means that

$$\left\{\int u^2(x)dx\right\} \cdot \left\{\int v^2(x)dx\right\} = \left\{\int u(x) \cdot v(x)dx\right\}^2 \tag{162}$$

This condition is satisfied if $u(x) + a \cdot v(x) = 0$, where a has an arbitrary value (independent of x). We used

$$u = (S_\lambda - \lambda) \cdot \sqrt{L} \qquad v = \frac{d(lnL)}{d\lambda} \cdot \sqrt{L} \tag{163}$$

Substitution of $u(x) = -av(x)$ in 163 gives the following rule for an efficient estimator:

$$\frac{d(lnL)}{d\lambda} = A(\lambda)\{S_\lambda - \lambda\} \tag{164}$$

where $A(\lambda)$ is the arbitrary constant (independent of the measurements t). From 159 we know that an efficient estimator satisfies

$$V(S_\lambda) \cdot I(\lambda) = 1 \tag{165}$$

Combining this with 160 and 164 we obtain for an efficient estimator:

$$V(S_\lambda) = \frac{1}{A(\lambda)} \qquad\qquad \frac{d(lnL)}{d\lambda} = V(S_\lambda)^{-1}\{S_\lambda - \lambda\} \tag{166}$$

7.7.1 An estimator for the binomial distribution

We consider again the forward-backward asymmetry in interactions $e^+e^- \rightarrow \mu^+\mu^-$. Call λ the probability for a forward μ^+. We observe n interactions; in k interactions the μ^+ is produced in the forward direction. What is the best estimate for λ? The likelihood for the binomial distribution is

$$L(k;\lambda) = \left(\begin{array}{c} n \\ k \end{array} \right) \lambda^k (1-\lambda)^{n-k} \tag{167}$$

Take the logarithm and differentiate with respect to λ

$$\begin{aligned} lnL &= kln\lambda + (n-k)ln(1-\lambda) + ln\left(\begin{array}{c} n \\ k \end{array} \right) \\ \frac{d(lnL)}{d\lambda} &= \frac{n}{\lambda(1-\lambda)}\left(\frac{k}{n} - \lambda\right) = V\left(S_\lambda\right)^{-1}\left(S_\lambda - \lambda\right) \end{aligned} \tag{168}$$

The estimator $S_\lambda = \frac{k}{n}$ is unbiased; comparison with 166 shows that the estimator is efficient. Its variance is $V(S_\lambda) = \lambda(1-\lambda)/n$; so it is consistent because $\lim_{n\to\infty}(S_\lambda) = 0$.

7.7.2 An estimator for the exponential distribution

We consider the exponential distribution for the decay of a particle with lifetime λ. We measure the decay times of n particles $(t_1, t_2, ..., t_n)$. What is the best estimate for λ? The likelihood for the exponential distribution is

$$L = \prod_{i=1}^{n} \frac{1}{\lambda} e^{\left(-\frac{t_i}{\lambda}\right)} \tag{169}$$

Take the logarithm and differentiate with respect to λ

$$\frac{d(lnL)}{d\lambda} = \sum_{i=1}^{n}\left(-\frac{1}{\lambda} + \frac{t_i}{\lambda^2}\right) = \frac{n}{\lambda^2}\left(\frac{1}{n}\sum_{i=1}^{n} t_i - \lambda\right) = V\left(S_\lambda\right)^{-1}\left(S_\lambda - \lambda\right) \tag{170}$$

The average value of the observed decay times is an unbiased estimator for the particle's lifetime; comparison with 166 shows that it is an efficient estimator for the particle's lifetime. Its variance is λ^2/n; so it is consistent, because $\lim_{n\to\infty}\left(\frac{\lambda^2}{n}\right) = 0$

7.7.3 An estimator for the normal distribution

We consider a normal distribution with mean value λ and variance σ^2. We have a set of n measurements $(x_1, x_2, ..., x_n)$. What is the best estimate for λ? The likelihood for the normal distribution is

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\left\{-\frac{(x_i-\lambda)^2}{2\sigma^2}\right\}} \tag{171}$$

Take the logarithm and differentiate with respect to λ

$$\frac{d(lnL)}{d\lambda} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \lambda) = \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \lambda\right) = V\left(S_\lambda\right)^{-1}\left(S_\lambda - \lambda\right) \tag{172}$$

The arithmetic mean value of the measurements is an unbiased estimator for the mean value of the normal distribution; comparison with 166 shows that it is an efficient estimator. Its variance is $V\left(S_\lambda\right) = \left(\sigma^2/n\right)$; so it is consistent, because $\lim_{n\to\infty}\left(\frac{\sigma^2}{n}\right) = 0$.

7.7.4 Combining measurements of differing accuracy

Imagine that a quantity has been measured in n different (independent) experiments. Each experiment presents its result with x_i as the best estimated value for the quantity and $\sigma(x_i)$ as statistical error. How can we combine the results, and (what) do we gain by doing so? Each experiment has drawn a value from a normal distribution with mean value λ (the 'true' value of the quantity). Since the statistical errors are different, the underlying normal distributions have different variances σ_i^2. The likelihood to find x_i is

$$\frac{1}{\sigma_i\sqrt{2\pi}} e^{\left\{-\frac{(x_i-\lambda)^2}{2\sigma_i^2}\right\}} \tag{173}$$

The combined likelihood for all n experiments together is

$$L = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{\left\{-\frac{(x_i-\lambda)^2}{2\sigma_i^2}\right\}} \tag{174}$$

Take the logarithm and differentiate with respect to λ

$$\frac{d(lnL)}{d\lambda} = \sum_{i=1}^{n} \frac{x_i-\lambda}{\sigma_i^2} = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} - \sum_{i=1}^{n} \frac{\lambda}{\sigma_i^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left\{ \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} - \lambda \right\} = V\left(S_\lambda\right)^{-1}\left(S_\lambda - \lambda\right) \tag{175}$$

This implies that the weighted mean is an unbiased and efficient estimator for the true value of λ

$$S_\lambda = \left(\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}\right) \cdot \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1} \tag{176}$$

Its variance is

$$V(S_\lambda) = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1} \tag{177}$$

so it is consistent. It is clear that we gain precision by combining the results of the experiments . To show this we consider the case that all experiments have the same statistical error $\sigma_i^2 = \sigma^2$

$$V(S_\lambda) = \left(\frac{n}{\sigma^2}\right)^{-1} = \frac{\sigma^2}{n} \tag{178}$$

7.8 Exercises

Exercise 7.1

Determine the information $I(\lambda)$ of a sample of N measurements

a. obtainded from a normal distribution of known variance σ^2 but unknown mean $\lambda = a$.

b. obtainded from a normal distribution of known mean a but unknown variance $\lambda = \sigma^2$.

7.8.1 Exercise 7.2

In an experiment electrons and positrons collide. The momentum of the electrons is $p_x = p_y = 0.$, $p_z = 45.0GeV/c$ and the momentum of the positrons is $p_x = p_y = 0.$, $p_z = -45.0GeV/c$. 1000 collisions with two muons in the final state $(\mu^+\mu^-)$ were selected. In 550 collisions (out of the 1000) the μ^+ had been produced in the 'forward' direction (i.e. the p_z of the μ^+ had the same sign as the p_z of the positron). The probability that the μ^+ will be produced in forward (backward) direction is p_f (p_b).

a. Determine the best estimate for p_f and its statistical uncertainty.

b. Determine the best estimate of the forward-backward asymmetry $(p_f - p_b)$ and its statistical uncertainty.

c. Give arguments whether this measurement has established the existence of a forward-backward asymmetry clearly or whether the measurement is a statistical fluctuation?

8 The Maximum Likelihood method

8.1 The Maximum Likelihood estimator

We now move on to a general recipe for a good estimator. We have a set of measurements: $(x_1, x_2, ..., x_n)$. They are drawn from a probability distribution $f(x; \lambda)$, where λ is a parameter whose value is unknown. We need an 'algorithm' which can be applied to the measurements in order to get a best estimate for the true value of λ, together with its statistical error. The algorithm should satisfy the three criteria mentioned in the previous chapter. We already defined the likelihood

$$L(\lambda) = \prod_{i=1}^{n} f(x_i; \lambda) \tag{179}$$

The logarithmic likelihood is defined as

$$lnL(\lambda) = \sum_{i=1}^{n} lnf(x_i; \lambda) \tag{180}$$

The Maximum Likelihood (ML) method states that the best estimate for λ is the value which maximises $L(\lambda)$ and (automatically also) $lnL(\lambda)$. This value can be found by solving

$$\frac{dlnL(\lambda)}{d\lambda} = \sum_{i=1}^{n} \frac{dlnf(x_i; \lambda)}{d\lambda} = 0 \tag{181}$$

Note/verify: when we apply this to the binomial, the exponential, and the normal distribution, we find (again) the estimators of the previous chapter.

8.2 Quality of the Maximum Likelihood estimator

We look at the asymptotic shape ($\lim_{n\to\infty}$) of the likelihood-function We note the ML-estimate for λ as $\tilde{\lambda}$, and expand $lnL(\lambda)$ with a Taylor-series around $\lambda = \tilde{\lambda}$

$$\frac{dlnL}{d\lambda}(\lambda) = \frac{dlnL}{d\lambda}\left(\tilde{\lambda}\right) + (\lambda - \tilde{\lambda})\frac{d^2lnL}{d\lambda^2}\left(\tilde{\lambda}\right) + ... = (\lambda - \tilde{\lambda})\frac{d^2lnL}{d\lambda^2}\left(\tilde{\lambda}\right) + ... \tag{182}$$

(Note that the first derivative vanishes in the maximum.) For large values of n, equation 160 from the previous chapter gives the following asymptotic value for the second derivative:

$$\frac{d^2lnL}{d\lambda^2}\left(\tilde{\lambda}\right) = -I(\tilde{\lambda}) \tag{183}$$

so

$$\frac{dlnL}{d\lambda}(\lambda) = -I(\tilde{\lambda})(\lambda - \tilde{\lambda}) \tag{184}$$

Comparison of 184 with 166 shows that for large n the ML-estimate for λ is a good estimator, i.e. unbiased, consistent and efficient. The variance of the estimator can be calculated from the value of the second derivative of the logarithmic likelihood:

$$V(S_\lambda) = V(\tilde{\lambda}) = I^{-1}(\tilde{\lambda}) = \frac{-1}{\sum_{i=1}^{n} \frac{d^2lnf(x_i;\lambda)}{d\lambda^2}|_{\lambda=\tilde{\lambda}}} \tag{185}$$

Note/verify: when we apply this to the binomial, the exponential, and the normal distribution, we find (again) the estimator variances of the previous chapter. Integration of 184 gives

$$lnL(\lambda) = -\frac{I(\tilde{\lambda})}{2}(\lambda - \tilde{\lambda})^2 + const. \tag{186}$$

So, the shape of the logarithmic likelihood function (around its maximum) is a parabola; the width of the parabola is related to the variance of the estimated value. After exponentiation we obtain:

$$L(\lambda) = k \cdot e^{\left\{-\frac{I(\tilde{\lambda})}{2}(\lambda-\tilde{\lambda})^2\right\}} \tag{187}$$

So, the shape of the likelihood function (around its maximum) is a normal distribution; the width of the distribution (σ) is related to the variance of the estimated value.

8.3 The Maximum Likelihood method: numerical procedure

The ML method implies that we have to find the value of a parameter which maximises the value of the likelihood. In some cases this problem can be solved by an analytical method, i.e. by solving the equation $dlnL(\lambda)/d\lambda = 0$. For the estimation of the lifetime of a particle we had

$$\begin{aligned}
f(t;\tau_0) &= \frac{1}{\tau_0}e^{\left(-\frac{t}{\tau_0}\right)} \quad \rightarrow \\
lnL(\tau_0) &= \sum_{i=1}^{n}\left(-ln(\tau_0) - \frac{t_i}{\tau_0}\right) \\
\frac{d(lnL)}{d\tau_0} &= \sum_{i=1}^{n}\left(-\frac{1}{\tau_0} + \frac{t_i}{\tau_0^2}\right) = 0 \quad \rightarrow \quad \tilde{\tau}_0 = \frac{1}{n}\sum_{i=1}^{n} t_i
\end{aligned}$$

Now, suppose that our measurement equipment only registers decay-times which exceed a certain value $t_i > t_{min}$. To cope with this, the probability distribution has to be re-normalised by introducing a weight function

$$\begin{aligned}
g(\tau_0) &= \int_{t_{min}}^{\infty} f(t;\tau_0)dt = \int_{t_{min}}^{\infty} \frac{1}{\tau_0}e^{\left(-\frac{t}{\tau_0}\right)}dt = e^{\left(-\frac{t_{min}}{\tau_0}\right)} \\
&\rightarrow \quad f(t;\tau_0) = \frac{1}{\tau_0}\frac{1}{g(\tau_0)}e^{-\frac{t}{\tau_0}} = \frac{1}{\tau_0}e^{\left(\frac{t_{min}}{\tau_0}\right)}e^{\left(-\frac{t}{\tau_0}\right)}
\end{aligned}$$

The maximum can still be found analytically

$$\begin{aligned}
lnL(\tau_0) &= \sum_{i=1}^{n}\left(-ln(\tau_0) + \frac{t_{min}}{\tau_0} - \frac{t_i}{\tau_0}\right) \\
\frac{d(lnL)}{d\tau_0} &= \sum_{i=1}^{n}\left(-\frac{1}{\tau_0} - \frac{t_{min}}{\tau_0^2} + \frac{t_i}{\tau_0^2}\right) = 0 \quad \rightarrow \quad \tilde{\tau}_0 = \frac{1}{n}\sum_{i=1}^{n}(t_i - t_{min})
\end{aligned}$$

Now, suppose that our measurement equipment only registers decay times which fall between a minimum and a maximum value: $t_{min} < t_i < t_{max}$. To cope with this, the probability distribution has (again) to be re-normalised by introducing a weight function

$$\begin{aligned} g(\tau_0) &= \int_{t_{min}}^{t_{max}} \frac{1}{\tau_0} e^{\left(-\frac{t}{\tau_0}\right)} dt \\ &= e^{\left(-\frac{t_{min}}{\tau_0}\right)} - e^{\left(-\frac{t_{max}}{\tau_0}\right)} \\ &\rightarrow f(t;\tau_0) = \frac{1}{\tau_0}\frac{1}{g(\tau_0)} e^{\left(-\frac{t}{\tau_0}\right)} \end{aligned}$$

The (logarithmic) likelihood and the equation become

$$\begin{aligned} lnL(\tau_0) &= \sum_{i=1}^{n} \left(-ln(\tau_0) - \frac{t_i}{\tau_0} - ln(g(\tau_0)) \right) \\ \frac{d(lnL)}{d\tau_0} &= \sum_{i=1}^{n} \left(-\frac{1}{\tau_0} + \frac{t_i}{\tau_0^2} - \frac{1}{g(\tau_0)} \frac{dg(\tau_0)}{d\tau_0} \right) = 0 \end{aligned}$$

This equation can no longer be solved analytically. When it is impossible to find the maximum of lnL in an analytical way, we have to use numerical methods. In one of the exercises we will describe a general (and simple) method to determine at which value of x the function $f(x)$ has its maximum value. When applied to $lnL(\lambda)$ it will give the best estimate for λ. In order to determine the statistical error on the best value, we note that the shape of $lnL(\lambda)$ in the neighbourhood of the maximum is a parabola:

$$lnL(\lambda) = -\frac{I(\tilde{\lambda})}{2}(\lambda - \tilde{\lambda})^2 + const. \tag{188}$$

The shape of $L(\lambda)$ is a Gaussian distribution $L(\lambda) = k \cdot e^{\left\{-\frac{I(\tilde{\lambda})}{2}(\lambda-\tilde{\lambda})^2\right\}}$. The variance of the distribution, the variance of the estimator is

$$V(S_\lambda) = V(\tilde{\lambda}) = I^{-1}(\tilde{\lambda}) \tag{189}$$

the standard-deviation of the distribution, the standard-deviation of the estimated value is

$$\sigma(S_\lambda) = \sigma(\tilde{\lambda}) = \sqrt{I^{-1}(\tilde{\lambda})} \tag{190}$$

This implies that

$$lnL(\tilde{\lambda} \pm \sigma(\tilde{\lambda})) = lnL(\tilde{\lambda}) - \frac{1}{2} \tag{191}$$

I.e. a change in $\tilde{\lambda}$ of one standard deviation from its maximum likelihood estimate leads to a decrease in the log-likelihood of $\frac{1}{2}$ from its maximum value. We can use this relationship to determine the statistical error of our estimated value in a numerical procedure. The meaning of this error is (as before): the probability that the true value of λ lies within one standard deviation from our best estimate is 0.682

8.4 Estimating several parameters simultaneously

Suppose that we have a set of n measurements $(x_1, x_2, ..., x_n)$ from a normal distribution. We want best estimates for the mean value and for the variance of the normal distribution (simultaneously). The parameter λ becomes a vector $\vec{\lambda}$. The underlying probability distribution becomes $f(x;\vec{\lambda})$. The likelihood and the logarithmic likelihood remain scalar quantities, but they are now a function of more than one variable.

$$L(\vec{\lambda}) = \prod_{i=1}^{n} f(x_i;\vec{\lambda}) \qquad lnL(\vec{\lambda}) = \sum_{i=1}^{n} lnf(x_i;\vec{\lambda}) \tag{192}$$

The ML method still says that the best estimates $\tilde{\vec{\lambda}}$ should maximise $lnL(\tilde{\vec{\lambda}})$. The information I becomes a square matrix. Element (i,j) of this matrix is

$$I_{i,j} = -\frac{\partial^2 lnL}{\partial\lambda_i \partial\lambda_j}\left(\vec{\lambda} = \tilde{\vec{\lambda}}\right) \tag{193}$$

The asymptotic shape of the logarithmic likelihood around the maximum becomes

$$lnL(\vec{\lambda}) = -\frac{1}{2}\left(\vec{\lambda} - \tilde{\vec{\lambda}}\right)^T I\left(\tilde{\vec{\lambda}}\right)\left(\vec{\lambda} - \tilde{\vec{\lambda}}\right) + const. \tag{194}$$

The asymptotic shape of the likelihood around the maximum becomes

$$L(\vec{\lambda}) = k \cdot e^{\left\{-\frac{1}{2}\left(\vec{\lambda}-\tilde{\vec{\lambda}}\right)^T I\left(\tilde{\vec{\lambda}}\right)\left(\vec{\lambda}-\tilde{\vec{\lambda}}\right)\right\}} \tag{195}$$

Application: estimate the mean λ_1 and the width λ_2 of a normal distribution.

$$\begin{aligned} f(x;\lambda_1,\lambda_2) &= \frac{1}{\lambda_2\sqrt{2\pi}} e^{\left\{-\frac{(x-\lambda_1)^2}{2\lambda_2^2}\right\}} \\ lnL(\lambda_1,\lambda_2) &= -\sum_{i=1}^{n}\left\{\frac{(x_i-\lambda_1)^2}{2\lambda_2^2}\right\} - n \cdot ln(\lambda_2) + const. \end{aligned}$$

For the maximum likelihood:

$$\frac{\partial lnL}{\partial\lambda_1} = \sum_{i=1}^{n}\frac{x_i-\lambda_1}{\lambda_2^2} = 0 \qquad \frac{\partial lnL}{\partial\lambda_2} = \sum_{i=1}^{n}\left\{\frac{(x_i-\lambda_1)^2}{\lambda_2^3}\right\} - \frac{n}{\lambda_2} = 0 \tag{196}$$

the solution of these equations is

$$\tilde{\lambda}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \tilde{\lambda}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\tilde{\lambda}_1)^2} \tag{197}$$

the Information-matrix is

$$\frac{\partial^2 lnL}{\partial\lambda_1^2} = -\frac{n}{\lambda_2^2} \qquad \frac{\partial^2 lnL}{\partial\lambda_1\partial\lambda_2} = -\sum_{i=1}^{n}\left\{\frac{2(x_i-\lambda_1)}{\lambda_2^3}\right\} \frac{\partial^2 lnL}{\partial\lambda_2^2} = -\sum_{i=1}^{n}\left\{\frac{3(x_i-\lambda_1)^2}{\lambda_2^4}\right\} + \frac{n}{\lambda_2^2} \tag{198}$$

substitution of $\lambda_1 = \tilde{\lambda}_1$ and $\lambda_2 = \tilde{\lambda}_2$ gives

$$I(\lambda_1,\lambda_2) = \begin{pmatrix} \frac{n}{\tilde{\lambda}_2^2} & 0 \\ 0 & \frac{2n}{\tilde{\lambda}_2^2} \end{pmatrix} \tag{199}$$

the covariance matrix becomes

$$I^{-1}(\lambda_1, \lambda_2) = \begin{pmatrix} \frac{\tilde{\lambda}_2^2}{n} & 0 \\ 0 & \frac{\tilde{\lambda}_2^2}{2n} \end{pmatrix} \tag{200}$$

From this matrix we find

$$\sigma(\tilde{\lambda}_1) = \frac{\tilde{\lambda}_2}{\sqrt{n}} \qquad \sigma(\tilde{\lambda}_2) = \frac{\tilde{\lambda}_2}{\sqrt{2n}} \qquad cov(\tilde{\lambda}_1, \tilde{\lambda}_2) = 0 \tag{201}$$

Note: since $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are calculated from the same set of measurements, we would expect that their errors are correlated. In this specific case, however, the correlation turns out to be $= 0$. In this example we could solve the relevant equations with analytical methods. For problems with more parameters this can become rather tedious. Also one can think of many cases where the analytical approach does not work at all. In those cases we have to rely on numerical methods to find the parameter values which maximise the likelihood, and to calculate the covariance matrix. Our *root* program library contains a 'package' called MINUIT, which can handle this problem. We will use it in one of the exercises.

8.5 Exercises

Exercise 8.1 - *root* exercise 6

Application of the Maximum Likelihood method.

Problem statement

The setup of this exercise is as follows:

- We simulate a series of decay times of a particle with lifetime $\tau_0 = 5$ (the 'unit' in which the lifetime is expressed is not important for this exercise).
- Then we *forget* the value we used for the lifetime, and we are going to *derive* it from the simulated decay times (in a real experiment these would have been our measurements).
- We use the Maximum Likelihood method (ML).
  In this method we assign to the measured decay times $(t_1, t_2, ..., t_n)$ a 'logarithmic Likelihood'

$$lnL(\tau) = \sum_{i=1}^{n} \left\{ -ln\tau - \frac{t_i}{\tau} \right\}$$

- The ML-method states that the best estimate for the lifetime τ_0 is the one which maximises $lnL(\tau)$. Moreover, the statistical uncertainty in the 'best estimate' can be obtained from

$$\frac{d^2 lnL}{d\tau^2}\,(\tau = \tau_0)$$

- The best estimate for the lifetime, and its statistical uncertainty, can be calculated analytically, i.e they can be expressed directly in $(t_1, t_2, ..., t_n)$. The best estimate for the lifetime, and its statistical uncertainty, can be calculated also with numerical methods.

First part of the exercise (analytical method)

The first part of the exercise is :

- Generate 4000 decay times (with $\tau_0 = 5.$) and store them in an array.
- Display the decay times in a histogram and check that they have the proper (exponential) distribution .
- Calculate with the analytical method the best estimate for the lifetime, and its statistical uncertainty. Do this
  - using (only) the first 1000 decay times in the array.
  - using all 4000 decay times in the array.

  Give an interpretation of the result (especially the statistical errors)!

Second and third part of the exercise (numerical method)

A possible **numerical** procedure to solve the problem, goes as follows :

- Define a τ-range which should contain the best value of the lifetime. Take (in this case) $3. < \tau < 7$. Divide this range into k intervals (take $k = 500$).
- The midpoint of each interval corresponds to a specific value of τ; calculate the values $\tau_i (i = 1, 2, ..., k)$, and store them in an array.
- Calculate for each value of τ_i the value of $lnL(\tau_i)$, and store the results in an array.
- Now search in the array for the *maximum value* of lnL ; the corresponding value of τ_i is $\tilde{\tau_0}$, the best estimate for τ_0.

The second part of the exercise is:

- Determine with this numerical method the best estimate for τ_0.
- Do this for the first set of 1000 decay times, and for the full set of 4000 decay times.
- Compare the results with the values from the analytical method.
  NOTE: this comparison makes only sense if you use the **same set of decay times** for the analytical and the numerical method.

A numerical procedure to find the statistical uncertainty in the best estimate can be found using the shape of lnL.

- We have to search (in the array with lnL-values) for the τ-values where

  $$lnL(\tau) = lnL(\tilde{\tau_0}) - 0.5$$

  There are two of these points: one $< \tilde{\tau_0}$, and one $> \tilde{\tau_0}$, so you have to calculate two values for the statistical uncertainty.
- Since our values τ_i are discrete it is very unlikely that we will find (in our arrays) a value of τ_i for which the corresponding value of lnL differs **exactly** 0.5 from the maximum value. But we can look for two values τ_{i-1} and τ_i with :

  $$[lnL(\tau_{i-1}) - lnL(\tilde{\tau_0}) + 0.5] \cdot [lnL(\tau_i) - lnL(\tilde{\tau_0}) + 0.5] < 0$$

  and take their average value for the solution.
  Do this for both points for which $lnL(\tau) = lnL(\tilde{\tau_0}) - 0.5$.

The third part of the exercise is :

- Determine with this numerical method the statistical uncertainty in the best estimate for τ_0.
- Do this for the first set of 1000 decay times, and for the full set of 4000 decay times.
- Compare the results with the values from the analytical method.

Graphical representation of the likelihood (optional)

This (fourth) part of the exercise is optional.
According to theory, the shape of the likelihood function around its maximum approaches a *normal distribution*. We are going to visualise this.

- From the preceding part of the exercise we have two arrays: one with the t_i-values in the mid of each time-interval, one with the corresponding values of lnL .
- First we go from the lnL-values to the L-values , using the exponential function.
  - To avoid numerical problems we subtract the maximum value of lnL from each of the lnL-values, before taking the exponent (from the difference).
  - The *graph* of L versus t can be drawn 'under ROOT' using code similar to $/user/uvak/sda/example2.C$

Presentation of results

HAND IN :

- A print of your program code
- A print of the calculated lifetimes and their statistical errors
- A plot of the likelihood-functions (if made)

9 The method of Least Squares

The method of Least Squares is a way of determining unknown parameters from a set of data. It can be derived from the principle of maximum likelihood or it can be regarded as a sensible estimator in its own right. Least squares should be taken literally: minimise the squared difference between a set of measurements and their predicted values; vary the parameters you want to estimate by adjusting their predicted values so as to be close to the measurements; by squaring the differences greater importance is placed on removing large deviations.

9.1 An example

An experiment has been performed to 'measure' the 'role' of the up-quark inside the proton. The result is expressed in the quark-distribution function $xu(x)$, which represents the role of the up-quark as a function of the Feynman scaling variable x. The function value is measured in several *intervals* of x; the results have statistical uncertainties $\sigma[xu(x)]$. The results are summarized in the following table

x	$xu(x)$	$\sigma[xu(x)]$	x	$xu(x)$	$\sigma[xu(x)]$
0.025	0.230	0.027	0.325	0.432	0.034
0.075	0.371	0.034	0.375	0.392	0.027
0.125	0.459	0.034	0.425	0.322	0.027
0.175	0.486	0.041	0.475	0.257	0.020
0.225	0.527	0.041	0.550	0.196	0.014
0.275	0.500	0.034	0.650	0.088	0.014

(202)

A graphical representation of the distribution is presented in figure 14. There is a theoretical prediction for the shape of this function:

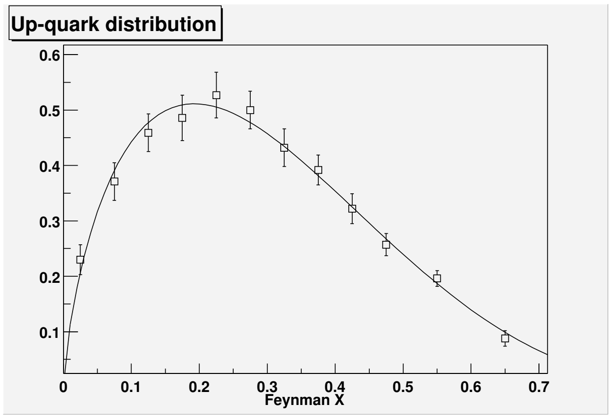


Figure 14: Up-Quark distribution

$$y(x) = Px^{\alpha}(1-x)^{\beta} \tag{203}$$

where P, α and β are parameters with unknown values. It is clear that the experimental result contains information about the values of the unknown parameters: how do we determine the best estimates for these values? Since the data points have statistical uncertainties, also our best estimates have statistical errors: how do we calculate the errors? Since the best values of the three parameters will be calculated from the same set of data points, the results may become correlated: how do we calculate the covariances? It is not a priori evident that the theoretical prediction agrees with the experimental result (regardless of the values of the parameters): (how) can we check this?

9.2 Derivation of the method

We start with a simple problem (from a previous chapter): imagine that we have measured a specific quantity in n different experiments with different instruments. Each experiment presents its result with y_i as the best estimate for the value of the quantity, and σ_i as statistical error $(i = 1, 2, .., n)$. How can we combine the results from the various experiments? Suppose that the errors are normally distributed around zero, so that a measurement corresponds with obtaining a sample from a Gaussian distribution with mean λ and standard deviation σ_i. Since the statistical errors for each experiment are different, the normal distributions have different variances σ_i^2. The likelihood to find the result y_i in the i-th experiment is:

$$\frac{1}{\sigma_i\sqrt{2\pi}}e^{\left\{-\frac{(y_i-\lambda)^2}{2\sigma_i^2}\right\}} \tag{204}$$

The combined likelihood for all n experiments together is

$$L = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}}e^{\left\{-\frac{(y_i-\lambda)^2}{2\sigma_i^2}\right\}} \tag{205}$$

The logarithmic likelihood for all n experiments together is

$$lnL = -\sum_{i=1}^{n}(\ln\sigma_i\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(y_i-\lambda)^2}{2\sigma_i^2} \tag{206}$$

The Maximum Likelihood method says that the best estimate for λ is the value which maximises lnL. It is clear that this is the same value which minimises

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i-\lambda)^2}{\sigma_i^2} \tag{207}$$

This is called the Method of Least Squares: The theoretical value (or the true value) is $y_i^{theory} = \lambda$. We have experimental values y_i^{exp} with statistical errors σ_i. The quantity

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i^{exp} - y_i^{theory})^2}{\sigma_i^2} \tag{208}$$

measures the difference between the theoretical values and the experimental results, where the difference in each point is weighted with the error σ_i. When the values of y_i^{theory} depend on a parameter with unknown value (in this case $y_i^{theory} = \lambda$) we use as best estimate for λ the value which minimises the χ^2. Applied to this case, we get:

$$\frac{d\chi^2}{d\lambda} = -2\sum_{i=1}^{n}\frac{(y_i^{exp}-\lambda)}{\sigma_i^2} = 0 \quad \rightarrow \quad \tilde{\lambda} = \left(\sum_{i=1}^{n}\frac{y_i}{\sigma_i^2}\right)\cdot\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\right)^{-1} \tag{209}$$

the result we found in an earlier chapter. When we want to calculate the statistical error in $\tilde{\lambda}$, we note that $\tilde{\lambda}$ is calculated simply as a linear combination of the measured quantities y_i. Hence, we can apply the rules of error propagation. The covariance matrix of the measured quantities y_i (note that they are independent) is:

$$C_y = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ . & . & .... & \\ . & . & .... & \\ & & & \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix} \tag{210}$$

The transformation matrix is a vector R with components:

$$R_i = \frac{1}{\sigma_i^2}\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\right)^{-1} \tag{211}$$

Error propagation gives:

$$\sigma^2(\tilde{\lambda}) = RC_yR^T = \left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\right)^{-1} \tag{212}$$

(again) a result found in an earlier chapter.

9.3 Fit of a straight line

As an example we consider the measurement of the linear expansion coefficient of some material. Imagine that we have measured the length of a bar at different temperatures. At each temperature the measurement is repeated several times, in order to determine the measurement error (accuracy). Our model (the theory) predicts that there is a linear relationship between length and temperature of the bar. To determine the linear expansion coefficient, we have to fit a straight line through the measured points. We proceed as follows:
Notation:

t_i	the temperatures at which the length is measured.
n	the number of measurements
l_i^{exp}	the (average values) of the measured lengths
$l_i^{theory} = a \cdot t_i + b$	the theoretical predictions
a and b	quantities to be deduced from the experiment.

The χ^2 becomes:

$$\chi^2 = \sum_{i=1}^{n}\left\{\frac{(l_i^{exp} - (at_i + b))^2}{\sigma_i^2}\right\} \tag{213}$$

The best estimates for a and b are those which minimise the χ^2:

$$\frac{\partial\chi^2}{\partial a} = 0 \qquad \frac{\partial\chi^2}{\partial b} = 0 \tag{214}$$

This gives:

$$\sum_{i=1}^{n}\frac{l_it_i}{\sigma_i^2} - a\sum_{i=1}^{n}\frac{t_i^2}{\sigma_i^2} - b\sum_{i=1}^{n}\frac{t_i}{\sigma_i^2} = 0 \tag{215}$$

where $l_i = l_i^{exp}$.

$$\sum_{i=1}^{n} \frac{l_i}{\sigma_i^2} - a\sum_{i=1}^{n} \frac{t_i}{\sigma_i^2} - b\sum_{i=1}^{n} \frac{1}{\sigma_i^2} = 0 \tag{216}$$

From these equations a and b can be solved (we use the abbreviated notation $\sum_{i=1}^{n} = \sum$):

$$\begin{aligned} a &= \frac{\left(\sum \frac{t_i}{\sigma_i^2}\right)\left(\sum \frac{l_i}{\sigma_i^2}\right) - \left(\sum \frac{t_i l_i}{\sigma_i^2}\right)\left(\sum \frac{1}{\sigma_i^2}\right)}{\left(\sum \frac{t_i}{\sigma_i^2}\right)^2 - \left(\sum \frac{t_i^2}{\sigma_i^2}\right)\left(\sum \frac{1}{\sigma_i^2}\right)} \\ b &= -\frac{\left(\sum \frac{t_i^2}{\sigma_i^2}\right)\left(\sum \frac{l_i}{\sigma_i^2}\right) - \left(\sum \frac{t_i l_i}{\sigma_i^2}\right)\left(\sum \frac{t_i}{\sigma_i^2}\right)}{\left(\sum \frac{t_i}{\sigma_i^2}\right)^2 - \left(\sum \frac{t_i^2}{\sigma_i^2}\right)\left(\sum \frac{1}{\sigma_i^2}\right)} \end{aligned}$$

Again we could calculate the statistical errors of a and b through error propagation. This however, becomes rather cumbersome. It is much better to change to the matrix formalism!

9.4 Fit of a polynomial (matrix formalism)

Our measurements form a column vector with n elements $\vec{y}_{exp}$. The measurement errors form a (nxn) covariance matrix. If the measurements are independent, this matrix is:

$$C_y = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ . & . & .... & \\ . & . & .... & \\ & & & \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix} \tag{217}$$

Suppose that theory predicts that our measured points lie on a polynomial of order k ($k < n$):

$$y_i^{theory} = \sum_{j=0}^{k} a_j \cdot (x_i)^j \tag{218}$$

In vector form we can write this as: $\vec{y}_{theory} = A\vec{\Theta}$, where A is a matrix with n rows and $(k+1)$ columns, and $\vec{\Theta}$ is a column vector with $(k+1)$ elements:

$$\vec{y}_{theory} = \begin{pmatrix} 1. & x_1 & x_1^2 & ... & x_1^k \\ 1. & x_2 & x_2^2 & ... & x_2^k \\ . & . & . & ... & \\ . & . & . & ... & \\ 1. & x_n & x_n^2 & ... & x_n^k \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ ... \\ ... \\ a_k \end{pmatrix} \tag{219}$$

The χ^2 can be written as:

$$\chi^2 = (\vec{y}_{exp} - A\vec{\Theta})^T G_y (\vec{y}_{exp} - A\vec{\Theta}) \tag{220}$$

where G_y is the inverse of the covariance matrix:

$$G_y = C_y^{-1} = \left(\begin{array}{cccc} (\sigma_1^2)^{-1} & 0 & ... & 0 \\ 0 & (\sigma_2^2)^{-1} & ... & 0 \\ . & . & .... & \\ . & . & .... & \\ & & & \\ 0 & 0 & ... & (\sigma_n^2)^{-1} \end{array} \right) \tag{221}$$

The best estimate for the unknown quantities $\vec{\Theta}$ is the one which minimises the χ^2. For these values:

$$\frac{d\chi^2}{d\vec{\Theta}} = -2A^T G_y \left(\vec{y}_{exp} - A\vec{\Theta} \right) = 0 \tag{222}$$

It follows that:

$$\vec{\Theta} = \left(A^T G_y A \right)^{-1} \left(A^T G_y \vec{y}_{exp} \right) \tag{223}$$

For the calculation of the statistical errors in the best estimates, we apply the rules of error propagation:

$$C_\Theta = R C_y R^T \tag{224}$$

with:

$$C_y = G_y^{-1} \qquad\qquad R = (A^T G_y A)^{-1} A^T G_y \tag{225}$$

This gives (use $(ABC)^T = C^T B^T A^T$):

$$\begin{aligned} C_\Theta &= \{(A^T G_y A)^{-1} A^T G_y\} G_y^{-1} \{(A^T G_y A)^{-1} A^T G_y\}^T = \\ &= (A^T G_y A)^{-1} A^T (G_y G_y^{-1} G_y^T) A \{(A^T G_y A)^{-1}\}^T \end{aligned}$$

Since C_y is symmetric, G_y, G_y^{-1}, $(A^T G_y A)$ and $(A^T G_y A)^{-1}$ are symmetric too; therefore:

$$C_\Theta = (A^T G_y A)^{-1} A^T G_y A (A^T G_y A)^{-1} = (A^T G_y A)^{-1} \tag{226}$$

In summary, the recipe for the Least Squares Methods is as follows: Store the n measurements in the vector $\vec{y}_{exp}$. Calculate the nxn covariance-matrix C_y. For independent measurements the off-diagonal elements are zero. The diagonal elements contain the variances σ_i^2. Calculate the inverse of the covariance-matrix $G_y = C_y^{-1}$. For independent measurements the off-diagonal elements are (again) zero. The diagonal elements contain the inverses of the variances σ_i^2. Calculate (from the model) the matrix A, with (n) rows and $(k+1)$ columns, containing the partial derivatives of y_i^{theory} with respect to each of the parameters Θ_j. For our polynomial:

$$A = \left(\begin{array}{cccc} \frac{\partial y_1}{\partial a_0} & \frac{\partial y_1}{\partial a_1} & ... & \frac{\partial y_1}{\partial a_k} \\ \frac{\partial y_2}{\partial a_0} & \frac{\partial y_2}{\partial a_1} & ... & \frac{\partial y_2}{\partial a_k} \\ . & . & . & ... \\ . & . & . & ... \\ \frac{\partial y_n}{\partial a_0} & \frac{\partial y_n}{\partial a_1} & ... & \frac{\partial y_n}{\partial a_k} \end{array} \right) = \left(\begin{array}{ccccc} 1. & x_1 & x_1^2 & ... & x_1^k \\ 1. & x_2 & x_2^2 & ... & x_2^k \\ . & . & . & ... & \\ . & . & . & ... & \\ 1. & x_n & x_n^2 & ... & x_n^k \end{array} \right) \tag{227}$$

The best estimate for the parameters $\vec{\Theta}$ is:

$$\vec{\Theta} = \left(A^T G_y A\right)^{-1} \left(A^T G_y \vec{y}_{exp}\right) \tag{228}$$

The covariance matrix of $\vec{\Theta}$ is:

$$C_\Theta = (A^T G_y A)^{-1} \tag{229}$$

The elements on the main diagonal contain the variances of each of the parameters. Their statistical errors are obtained by taking the square-root. The off-diagonal elements contain the covariances between pairs of parameters.

9.5 Linear fit: general case

We can apply the recipe of the previous section, as long as the relationship between $\vec{y}_{theory}$ and the parameters $\vec{\Theta}$ is linear. Consider the example of the quark distribution at the beginning of this chapter. We have a set of measured quantities y_i^{exp} with statistical errors σ_i. Theory gives a prediction for the shape of this distribution:

$$y(x) = Px^\alpha(1-x)^\beta \tag{230}$$

We can make this problem linear by taking the logarithm:

$$ln(y) = ln(P) + \alpha ln(x) + \beta ln(1-x) \tag{231}$$

In doing so, the measurement error in y has to be transformed according to (apply error propagation):

$$\sigma(ln(y)) = \frac{\sigma(y)}{y} \tag{232}$$

We arrive at the following expressions for the measurement vector and the covariance matrix:

$$\vec{y}_{exp} = \begin{pmatrix} lny_1 \\ lny_2 \\ . \\ . \\ lny_n \end{pmatrix} \qquad C_y = \begin{pmatrix} \sigma^2(lny_1) & 0 & 0 & ... & 0 \\ 0 & \sigma^2(lny_2) & 0 & ... & 0 \\ . & . & . & ... & \\ . & . & . & ... & \\ 0 & 0 & 0 & ... & \sigma^2(lny_n) \end{pmatrix} \tag{233}$$

The inverse of the covariance matrix is:

$$G_y = C_y^{-1} = \begin{pmatrix} \sigma^{-2}(lny_1) & 0 & 0 & ... & 0 \\ 0 & \sigma^{-2}(lny_2) & 0 & ... & 0 \\ . & . & . & ... & \\ . & . & . & ... & \\ 0 & 0 & 0 & ... & \sigma^{-2}(lny_n) \end{pmatrix} \tag{234}$$

The transformation matrix and the vector with the unknown parameters are

$$A = \begin{pmatrix} 1. & ln(x_1) & ln(1-x_1) \\ 1. & ln(x_2) & ln(1-x_2) \\ . & . & . \\ . & . & . \\ 1. & ln(x_n) & ln(1-x_n) \end{pmatrix} \qquad \Theta = \begin{pmatrix} lnP \\ \alpha \\ \beta \end{pmatrix} \tag{235}$$

The best estimate for the parameters $\vec{\Theta}$ is:

$$\vec{\Theta} = \left(A^T G_y A\right)^{-1} \left(A^T G_y \vec{y}_{exp}\right) \tag{236}$$

The covariance matrix of $\vec{\Theta}$ is:

$$C_\Theta = (A^T G_y A)^{-1} \tag{237}$$

The elements on the main diagonal contain the variances of each of the parameters. Their statistical errors are obtained by taking the square-root. The off-diagonal elements contain the covariances between pairs of parameters. The value of χ^2 in the minimum is found by substituting the 'fitted' value of $\vec{\Theta}$:

$$\chi^2_{min} = (\vec{y}_{exp} - A\vec{\Theta}_{fitted})^T G_y (\vec{y}_{exp} - A\vec{\Theta}_{fitted}) = \sum_{i=1}^{n} \frac{(y_i^{exp} - y_i^{fitted})^2}{\sigma_i^2} \tag{238}$$

We will see later on that this value χ^2_{min} has a special meaning.

9.6 Linear fit in case of equal errors

Suppose that the measurement errors in all points are equal: σ^2. The covariance matrix C_y becomes equal to $\sigma^2 \cdot \mathbf{1}$, where $\mathbf{1}$ represents the unit-matrix. The inverse of the covariance matrix C_y becomes $G_y = \sigma^{-2} \cdot \mathbf{1}$. The best estimate for the parameters $\vec{\Theta}$ becomes:

$$\vec{\Theta} = \left(A^T A\right)^{-1} \left(A^T \vec{y}_{exp}\right) \tag{239}$$

The covariance matrix of $\vec{\Theta}$ becomes:

$$C_\Theta = \sigma^2 (A^T A)^{-1} \tag{240}$$

The value of χ^2 in the minimum is found (again) by substituting the 'fitted' value of $\vec{\Theta}$:

$$\chi^2_{min} = \sigma^{-2} (\vec{y}_{exp} - A\vec{\Theta}_{fitted})^T (\vec{y}_{exp} - A\vec{\Theta}_{fitted}) = \sum_{i=1}^{n} \frac{(y_i^{exp} - y_i^{fitted})^2}{\sigma^2} \tag{241}$$

The value of χ^2_{min} has (again) a special meaning.

9.7 Linear fit in the absence of errors

We consider (again) the earlier example of a fit to a straight line, in order to get the best estimate for the linear expansion coefficient of a material. The fitting procedure tries to make a straight line which passes as good as possible through the measured points. In doing so, the distance of each measured point to the fitted line is weighted by the error in each point.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i^{exp} - y_i^{fitted})^2}{\sigma_i^2} \tag{242}$$

Even when the measurement errors are unknown, it is still possible to fit a straight line which passes as good as possible through the measured points, although we have to assume that the unknown measurement errors are equal in each point. We then start with putting the covariance matrix $C_y = G_y^{-1} = \mathbf{1}$. The best estimate for the parameters $\vec{\Theta}$ becomes:

$$\vec{\Theta} = \left(A^T A\right)^{-1} \left(A^T \vec{y}_{exp}\right) \tag{243}$$

which is the same as in the case of equal measurement errors. When we want to have a 'measure' for the error in this best estimate, we proceed as follows: substitute the fitted values of the parameters in the χ^2:

$$\chi^2_{min} = (\vec{y}_{exp} - A\vec{\Theta}_{fitted})^T (\vec{y}_{exp} - A\vec{\Theta}_{fitted}) = \sum_{i=1}^{n} (y_i^{exp} - y_i^{fitted})^2 \tag{244}$$

This is the sum of the quadratic distances of the measured points to the fitted line; it is also called the total quadratic residue. From this quantity we can calculate the average quadratic residue per point:

$$\sigma^2_{pres} = \frac{\chi^2_{min}}{n} \tag{245}$$

We use this 'a posteriori' result as a measure for the measurement error and put

$$C_y = \sigma^2_{pres} \cdot \mathbf{1} \qquad \rightarrow G_y = C_y^{-1} = \sigma^{-2}_{pres} \cdot \mathbf{1} \tag{246}$$

The covariance matrix of $\vec{\Theta}$ becomes

$$C_\Theta = \sigma^2_{pres} (A^T A)^{-1} \tag{247}$$

Note, that now

$$\chi^2_{min} = 1 \tag{248}$$

As we will see later on, this has important consequences.

9.8 The χ^2 distribution

In the preceding sections we said that the value of χ^2_{min} has a special meaning. To explain this we have to look at the χ^2 distribution. We start with the standard normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{249}$$

We draw n numbers from this distribution: $(x_1, x_2, ...x_n\ )$, and calculate the quantity

$$\chi^2 = \sum_{i=1}^{n} x_i^2 \tag{250}$$

We repeat this drawing of n numbers several times and calculate each time the quantity χ^2. How will the distribution of the χ^2-s look like? To answer this question, we will have to use the characteristic function, which was introduced earlier. The characteristic function of a quantity x from the standard normal distribution is defined as:

$$\begin{aligned} \Phi_x(t) &= E\{e^{itx}\} = \int_{-\infty}^{\infty} e^{itx} f(x) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\left(itx - \frac{x^2}{2}\right)} dx \end{aligned}$$

Analogously the characteristic function of x^2 is:

$$\Phi_{x^2}(t) = E\{e^{(itx^2)}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\left(itx^2 - \frac{x^2}{2}\right)} dx \tag{251}$$

Using:

$$\int_{-\infty}^{\infty} e^{\left(-\frac{x^2}{2a^2}\right)} dx = a\sqrt{2\pi} \tag{252}$$

we find:

$$\Phi_{x^2}(t) = \frac{1}{\sqrt{1-2it}} \tag{253}$$

To calculate the characteristic function of χ^2, we use the rule that the characteristic function of the sum of a set of variables is equal to the product of their characteristic functions. For the sum of the squares of n drawings, this implies:

$$\Phi_{\chi^2}(t) = (1-2it)^{-(n/2)} \tag{254}$$

The mean value of this χ^2 can be obtained by taking the value of the first derivative of the characteristic function with respect to t, in $t = 0$. Therefore:

$$E(\chi^2) = n \tag{255}$$

When we take the value of the second derivative of the characteristic function with respect to t in $t = 0$, we find

$$E\left\{(\chi^2)^2\right\} = n \cdot (n+2) \tag{256}$$

Therefore, the variance of our χ^2 distribution is

$$\sigma^2(\chi^2) = E\left\{(\chi^2)^2\right\} - \left\{E(\chi^2)\right\}^2 = 2 \cdot n \tag{257}$$

We claim that the probability distribution of our χ^2 is

$$f(\chi^2; n) = \frac{1}{2^\lambda \Gamma(\lambda)} \cdot (\chi^2)^{\lambda-1} \cdot e^{\left(-\frac{\chi^2}{2}\right)} \tag{258}$$

with

$$\lambda = \frac{n}{2} \qquad \Gamma(\lambda) = \int_0^\infty v^{\lambda-1} e^{-v} dv \tag{259}$$

$\Gamma(\lambda)$ is the gamma-function of Euler. We prove this by showing that the characteristic function of (258) is indeed given by (254):

$$\Phi_{\chi^2}(t) = E\{e^{it\chi^2}\} = \frac{1}{2^\lambda \Gamma(\lambda)} \cdot \int_0^\infty (\chi^2)^{\lambda-1} \cdot e^{\left(it\chi^2 - \frac{\chi^2}{2}\right)} d\chi^2 \tag{260}$$

Change the integration variable by substituting

$$v = (1-2it) \cdot \frac{\chi^2}{2} \tag{261}$$

This gives:

$$\begin{aligned} \Phi_{\chi^2}(t) &= \frac{1}{2^\lambda \Gamma(\lambda)} \cdot \int_0^\infty 2^{\lambda-1} \cdot (1-2it)^{-\lambda+1} \cdot v^{\lambda-1} \cdot e^{-v} \cdot 2 \cdot (1-2it)^{-1} dv \\ &= \frac{1}{\Gamma(\lambda)} \cdot (1-2it)^{-\lambda} \cdot \int_0^\infty v^{\lambda-1} \cdot e^{-v} dv = (1-2it)^{-\lambda} \end{aligned}$$

Figure 13 shows examples of χ^2 distributions for several values of n.

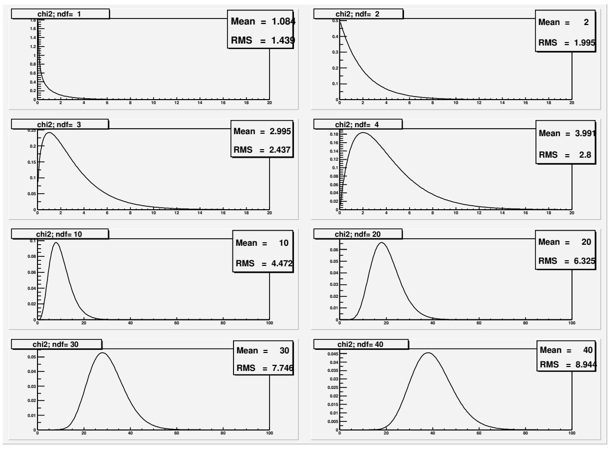


Figure 15: Chi squared distributions

9.9 The χ^2 as hypothesis tester

Consider (again) the example of the linear expansion coefficient of a specific material. We have measured the length of a bar at n different temperatures t_i. The measured lengths (together with their statistical errors) are: $l_i^{exp} \pm \sigma_i$. The model (theory) predicts a linear relationship $l_i^{theory} = a \cdot t_i + b$ We assume that the coefficients a and b are also predicted by the model. So, there are no parameters to be fitted; we simply want to check whether our measurements agree with the model. Of course our measured lengths do not coincide exactly with the lengths predicted by the model, due to the statistical measurement errors. But, if the model is valid, the differences between measurement and model should be compatible with the measurement errors. To make this quantitative we proceed as follows: summarise the total discrepancy between our measurements and the model predictions in

$$\Delta^2 = \sum_{i=1}^{n} \frac{(l_i^{exp} - l_i^{theory})^2}{\sigma_i^2} \tag{262}$$

By taking the square of the differences, we prevent that positive and negative differences cancel in the sum. By dividing each difference by σ_i^2 we express the difference in each point in a 'natural unit'. What do we expect for the value of Δ^2? Each measurement y_i^{exp} is a drawing from a normal distribution with mean value l_i^{theory} (if the model is correct) and width σ_i. Hence, the quantity

$$\frac{(l_i^{exp} - l_i^{theory})}{\sigma_i} \tag{263}$$

is drawn from the standard normal distribution. Therefore, the quantity Δ^2 will follow the χ^2 distribution, which corresponds to the quadratic sum of n drawings from the standard normal distribution. We call n the number of degrees of freedom or the number of constraints. We can use this to test the hypothesis that "the measurements agree with the predictions of the model". A crude argument goes as follows: the expectation value for the χ^2 with n degrees of freedom is n. The width of the χ^2 distribution with n degrees of freedom is $\sqrt{2n}$. If

$$|\Delta^2 - n| >> \sqrt{2n}$$

then our measurements do not agree with the model predictions. In fact we say: the difference between measurement and model is much larger then could be expected from the statistical errors in our measurement. So: the difference is not a coincidence, it is real, the model is wrong. A more precise argument is: the χ^2 distribution for n degrees of freedom is well known, and given by (258). The quantity

$$\int_{\Delta^2}^{\infty} f(\chi^2; n) d\chi^2 = \int_{\Delta^2}^{\infty} \frac{1}{2^{\lambda}\Gamma(\lambda)} \cdot (\chi^2)^{\lambda-1} \cdot e^{-\frac{\chi^2}{2}} d\chi^2 \tag{264}$$

gives the probability to find a weighted quadratic discrepancy $\geq \Delta^2$, if theory and experiment agree. If this probability is very small, we conclude that there is disagreement. A frequently used criterium is the one which corresponds to two standard deviations: if the probability, calculated with formula (264) is smaller than 0.05, then "The hypothesis is rejected at 95% confidence level". It is clear that a good (realistic) determination of the measurement errors σ_i is of crucial importance. If the errors are underestimated the quantity Δ^2 is *blown up*, and one could reject a good hypotheses. If the errors are overestimated the quantity Δ^2 is made smaller and one loses 'decision power'. Note, that a Δ^2 which is much smaller then expected from the χ^2 distribution cannot be interpreted as 'very good agreement' between theory and experiment, but may (also) be a reason to reject the theory. If no measurement errors are available, it is still possible to 'fit' a model and make an 'a posteriori' estimate of the error (using the mean quadratic residue), but it is no longer possible to test an hypothesis. We noted earlier that in this case one always gets $\Delta^2 = 1$.

9.10 The χ^2 test with free parameters

In the previous section we considered the case that the theory (the model) had no free parameters, since we assumed that the coefficients a and b were known quantities. We change to the situation where a and b are unknown quantities, whose values have to be determined from a least squares fit. Still there remains something to test, like: is the model assumption about linearity correct? Or does our experiment support a parabolic relationship: $y = a + b \cdot t + c \cdot t^2$. (How) can we test this? Application of the least squares fit gives a 'best estimate' for a and b. When we substitute these values into the theoretical prediction we can (again) calculate a value for

$$\Delta^2_{min} = \sum_{i=1}^{n} \frac{(l_i^{exp} - l_i^{fitted})^2}{\sigma_i^2} \tag{265}$$

We can use this value of Δ^2_{min} (again) to test our hypothesis by making a comparison with a χ^2 distribution, however, there is a difference: we have taken the values for a and b which make Δ^2 as small as possible. So one should expect a value for Δ^2 which is smaller then the one obtained in the case that there were no free parameters. It can be proven that in the case of n measured points and r free parameters, the Δ^2 should be compared with the χ^2-distribution for $(n - r)$ degrees of freedom. In a sense, one has 'sacrificed' r measured points to estimate the unknown parameters. Only the remaining $(n - r)$ points contribute to the Δ^2.

9.11 Non-linear problems

It may happen that the theoretical prediction is non-linear in the parameters. We take as an example: the total cross section W of e^+e^--collisions as a function of the collision energy E in the energy range around 90 GeV (from chapter 1). The theoretical prediction is (a bit simplified):

$$W(E) = A \cdot \left\{ (E^2 - M_z^2)^2 + E^4 \frac{\Gamma_z^2}{M_z^2} \right\}^{-1} \tag{266}$$

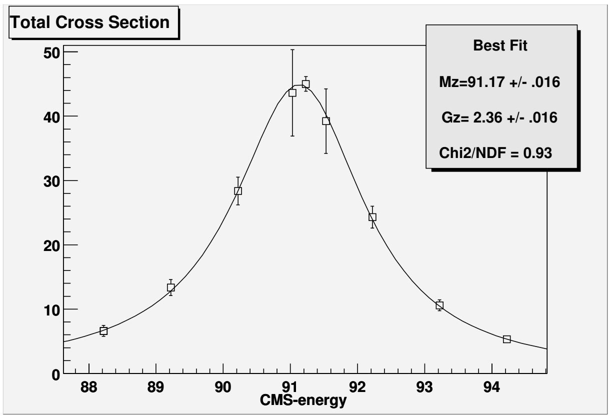


Figure 16: Total cross-section around the Z-mass

The unknown parameters are: A- a normalisation constant, M_z- the mass of the Z-boson and Γ_z- the decay width of the Z-boson. Clearly, the prediction is non-linear in M_z^2. Yet, we can still make up a Δ^2: the total cross section has been measured for a number of values of the collision energy E_i. The results are: $W_i^{exp} \pm \sigma_i$. The theoretical prediction W_i^{theory} for each of the collision energies E_i can be calculated from the formula above. It is a function of (A, M_z, Γ_z). The 'discrepancy' between theory and experiment is

$$\Delta^2(A, M_z, \Gamma_z) = \sum_{i=1}^{n} \frac{(W_i^{exp} - W_i^{theory})^2}{\sigma_i^2} \tag{267}$$

The best estimate for (A, M_z, Γ_z) are the values which minimise Δ^2. In this case it is not possible to solve the problem in an analytical way. So one has to use numerical methods. As mentioned earlier, our program *root* library contains a 'package' (called MINUIT) to deal with this. We not only need a best estimate for the parameters, but also the corresponding covariance-matrix, from which we can calculate the statistical errors of in the best estimates, and their correlation coefficients. This can be done as follows: the quantity Δ^2 can be written (in matrix notation) as:

$$\Delta^2 = (\vec{y}_{exp} - A\vec{\Theta})^T G_y (\vec{y}_{exp} - A\vec{\Theta}) \tag{268}$$

where G_y is the inverse of the covariance matrix of the measurements. It follows that

$$\frac{d\Delta^2}{d\vec{\Theta}} = -2A^T G_y \left(\vec{y}_{exp} - A\vec{\Theta}\right) \quad \rightarrow \quad \frac{d^2\Delta^2}{d\vec{\Theta}^2} = 2A^T G_y A \tag{269}$$

We already saw (via error propagation) that the covariance matrix of $\vec{\Theta}$ is given by

$$C_\Theta = (A^T G_y A)^{-1} \qquad \rightarrow \qquad C_\Theta = \frac{1}{2}\left(\frac{d^2\Delta^2}{d\vec{\Theta}^2}\right)^{-1} \tag{270}$$

The numerical method 'looks' at the function values in the neighbourhood of the minimum and uses them to calculate the elements of the matrix

$$\frac{d^2\Delta^2}{d\vec{\Theta}^2} \tag{271}$$

The covariance matrix can be calculated easily from this. A different way to find the covariance matrix in a numerical procedure can be found by using the fact that there is is a relationship between the value of Δ^2 in the minimum, and at one standard-deviation:

$$\Delta^2_{1std} = \Delta^2_{min} + 1 \tag{272}$$

We verify this for the case that the theoretical prediction has only one parameter. The quantity Δ^2 becomes (θ is a scalar now):

$$\Delta^2(\Theta) = (\vec{y}_{exp} - A\Theta)^T G_y (\vec{y}_{exp} - A\Theta) \tag{273}$$

The value of Θ in the minimum is:

$$\Theta_{min} = \left(A^T G_y A\right)^{-1} \left(A^T G_y \vec{y}_{exp}\right) \tag{274}$$

The variance of Θ is: $(A^T G_y A)^{-1}$. This is a scalar quantity too. It can be shown that

$$\Delta^2(\Theta_{min} + \epsilon) = \Delta^2(\Theta_{min}) + \epsilon^2 (A^T G_y A) \tag{275}$$

At one standard deviation we have:

$$\epsilon = (A^T G_y A)^{-1/2} \qquad \rightarrow \qquad \Delta^2(\Theta_{min} + \epsilon) = \Delta^2(\Theta_{min}) + 1 \tag{276}$$

For the general case of n parameters, the proof is similar, apart from the fact that one standard deviation is then defined as

$$\vec{\epsilon}^T (A^T G_y A) \vec{\epsilon} = 1 \tag{277}$$

9.12 Exercises

Exercise 9.1

At the 1996 Conference on High Energy Physics in Warsaw, 7 different experiments presented their measurement of the quantity A^b_{FB} (the meaning of this quantity is not in the scope of this problem). The presented results were as follows:

Name of the experiment	measured value of A^b_{FB}	measurement error in A^b_{FB}
ALEPH -I	9.65	0.44
DELPHI-I	10.49	0.76
L3	10.30	1.00
OPAL-I	8.92	0.44
ALEPH-II	9.27	0.39
DELPHI-II	9.90	0.72
OPAL-II	9.63	0.67

a. Combination of the seven measurements will provide a better estimate of A^b_{FB} with a new measurement error. Determine this new estimate and the measurement error.

Verify whether it is indeed allowed to combine the results of the seven experiments:

b. Calculate the total χ^2 of the seven measurements with respect to the new estimate of A^b_{FB}.

c. What is the expectation value, the variance and the standard deviation of the χ^2-distribution in that case?

d. What is your conclusion: is it allowed to combine the measurements results or not? And why?

Exercise 9.2 - *root* exercise 7

Application of the Least Square method.

Problem statement

The setup of this exercise is as follows :

- In an experiment we have performed measurements about the 'role' of the up-quark inside the proton. The results are expressed in the *up-quark distribution function* $f(x)$, which has been measured at different values of the so-called Feynman scaling variable x. The measured values of $f(x)$ have statistical errors $\sigma(f(x))$
- The experimental result is given in the following table.

x	$f(x)$	$\sigma(f(x))$	x	$f(x)$	$\sigma(f(x))$
0.025	0.230	0.027	0.325	0.432	0.034
0.075	0.371	0.034	0.375	0.392	0.027
0.125	0.459	0.034	0.425	0.322	0.027
0.175	0.486	0.041	0.475	0.257	0.020
0.225	0.527	0.041	0.550	0.196	0.014
0.275	0.500	0.034	0.650	0.088	0.014

- There is a *theoretical prediction* for the shape of this function :

$$f(x) = Px^{\alpha}(1-x)^{\beta}$$

  where P, α and β are *parameters* whose values are *unknown* .

First part of the exercise

The first part of the exercise is :

- Calculate with the Least Squares method, the 'best values' for the three unknown parameters.
- Calculate the statistical uncertainty in these values.
- Decide whether the shape of the function, as predicted by theory, agrees with the measurement .

Graphical representation of results (optional)

The (optional) second part of the exercise is :

- Give a graphical representation of the measured values of $f(x)$ together with error-bars.
- Include a graphical representation of the fitted curve
- Include in the graphics a print of the fitted values of the unknown parameters, and their statistical errors, and a print of the value of the χ^2

Tools for matrix-calculus

You will have to calculate the inverse of a matrix a. In *root* this can be done as follows:

```
TMatrixD a(3,3) ;
a(0,0)= 5;
a(0,1)- 10 ;
....
a.Invert();
```

Tools for graphics

The *optional* part of this exercise requires that you give a graphical representation of the experimental data and of the results of the fit, using the *root*-facilities.

- an example of how to draw the fitted curve can be found in */user/uvak/sda/example2.C*
- an example of how to draw the experimental data-points (including their error-bars), and of how to include a print of the fitted values, can be found in: */user/uvak/sda/example5.C*

Presentation of results

HAND IN :

- A print of your program code
- A print of the fitted values of the parameters, their statistical errors and the value of the χ^2 of the fit
- A graphical representation of the results (if made)

Exercise 9.3 - *root* exercise 8

Use of MINUIT for a non-linear fit.

Problem statement

In this exercise you have (again) to apply the method of Least Squares, but now the theoretical prediction is a *non-linear* function of the parameters.

- In an experiment we have measured the total cross-section σ of electron-positron collisions, at several values of the collision-energy E.
  The collision energies (between 88 and 95 GeV) lie around the mass of the Z-boson, the carrier of the neutral component of the weak interaction. Therefore, the total cross-section shows a *resonant* behaviour.
- The experimental result is summarised in the following table :

E	σ	error in σ
(GeV)	(nbarn)	(nbarn)
88.22	6.61	0.17
88.28	6.73	0.45
89.22	13.35	0.25
89.28	14.32	0.76
90.22	28.35	0.43
90.28	29.71	1.12
91.03	43.62	1.34
91.23	45.00	0.23
91.28	43.06	1.24
91.53	39.22	1.00
92.22	24.32	0.34
92.28	21.79	1.00
93.22	10.60	0.17
93.28	9.32	0.44
94.22	5.32	0.10
94.28	5.53	0.37
95.03	4.01	0.45

We are going to use these results for the determination of the **mass** and **width** of the Z-boson.

Theoretical prediction

The theoretical prediction for the total cross-section σ as a function of the collision-energy E is :

$$\sigma(E) = A\frac{C_1}{(s-M_z^2)^2 + s^2\frac{\Gamma_z^2}{M_z^2}}\left(\frac{s}{M_z^2} - B\frac{s-M_z^2}{M_z^2}\right)$$

The **known** quantities in this formula are :

- $C_1 = 2.17 * 10^6$
- $s = E^2$

The **unknown** quantities in this formula are :

- M_z :the mass of the Z-boson.
- Γ_z : the width of the Z-boson.
- A and B : two normalisation constants, with values in the order of 1.

First part of the exercise

- Use the Least Squares method to determine the best values for M_z, Γ_z, A and B.
- Determine the statistical errors of the best values.
- Decide whether the shape of the function, as predicted by theory, agrees with the measurement .

Graphical representation of results (optional)

The (optional) second part of the exercise is :

- Give a graphical representation of the measured values of $\sigma(E)$ together with error-bars.
- Include a graphical representation of the fitted curve
- Include in the graphics a print of the fitted values of the unknown parameters, and their statistical errors, and a print of the value of the χ^2

A tool for non-linear fitting : MINUIT

- It is clear that the *analytical* method of the previous exercise can no longer be applied. Instead we have to rely on a *numerical* procedure
  - to find the values of the unknown parameters which minimise the χ^2
  - to use the shape of the χ^2 in the neighbourhood of the minimum for the determination of the statistical errors.
- The *root*-package gives access to a set of functions (collectively called MINUIT) which perform the above tasks.
- You can find an **example** program in
  /user/uvak/sda/example6.C
  In this example MINUIT is used to solve the problem of the (previous) exercise 7 .

- The example program contains a lot of in-line comments to make it clear to you what is going on. In the following sections I will give some additional comments and hints on the most important points.
  In a global way it is good to realize that MINUIT **does not know anything about your problem**. You have to tell **everything**, especially :
  - about the parameters involved in your problem
  - about the function to be minimised

9.12.1 MINUIT: information about your parameters

Information about your parameters is transferred to MINUIT by calling (for each parameter)

```
gMinuit-> mnparm (i, par_name[i], par_start[i], par_range[i],
                                0,           0,    ierflg ) ;
```

The **meaning** of the parameters is as follows :

- **i** : a sequential number of the parameter (type integer)
- **par_name** : a **character string** with the 'name' of the parameter (for printing purposes)
- **par_start** : a **start** value for the minimalisation procedure
- **par_range** : a **start** interval for the minimalisation procedure
- the remaining three parameters are not relevant (for us)

9.12.2 MINUIT: information about your function

MINUIT expects that you provide a **void function fcn** with the following *parameter-list*

```
void fcn(Int_t &dum1, Double_t *dum2, Double_t &Chi2,
                      Double_t *par , Int_t dum3)
```

The **meaning** of the parameters is as follows :

- the parameter **par**.
  this is an **array** in which MINUIT **proposes** a set of **values** for the parameters, in the **order** as specified by your starting information.
- the parameter **Chi2**.
  the (proposed) values in **par** give rise to a specific **theoretical prediction**; combination of this prediction with the **experimental values** and their **errors** results in a χ^2; you have to **calculate** the value of this χ^2 and **return** it to MINUIT through the parameter **Chi2** .
- the remaining three parameters are not relevant (for us)
- Note : to maintain high precision MINUIT always uses **type Double_t** for real numbers; you should do this in your version of fcn too.

9.12.3 MINUIT: its game

We can explain the working of MINUIT using your χ^2 as an example.

- MINUIT copies your starting values into the array **par**, and asks your function fcn for the corresponding value of the χ^2
- Then MINUIT changes the values in **par**, and again asks fcn for the corresponding value of the χ^2. If this value is smaller then the first one, the change in **par** goes apparently in the right direction (we are looking for a minimum).
  If the last value of the χ^2 is larger then the first one, the change in **par** apparently goes in the wrong direction.
- MINUIT has its own strategies to **walk** (in a multi-dimensional space) from a starting point to a point where the function has its minimum value. The coordinates of the points on its trajectory are presented in **par**. Your answer through the χ^2- value tells MINUIT whether the trajectory goes in the right direction or whether the trajectory has to be changed.

9.12.4 MINUIT: hints for the starting values

MINUIT doesn't know anything about the meaning of your parameters, and therefore has no idea about their order of magnitude. That's why you have to give **starting values**. Good starting values for the problem in this exercise can be obtained as follows :

- In the problem statement it was already specified that the values of A and B are in the order of 1. So this is a good starting value.
- In the theoretical prediction for $\sigma(E)$ you can see that $\sigma(E)$ reaches its maximum value for $E = M_z$. The table of the measured values should give you a starting value for M_z .
- The theoretical prediction gives the following expression for the maximum value of σ (calculated with $E = M_z$ and $A = B = 1$)

$$\sigma_{max} = \frac{C_1}{\sigma\Gamma_z^2} = \frac{C_1}{E^2\Gamma_z^2} = \frac{C_1}{M_z^2\Gamma_z^2}$$

  With this you can find a starting value for Γ_z .

Finally you have to give in **par_range** for each parameter a **start interval** . This **must** be a number > 0 ! With this interval you give an **indication** of **how far** the fitted value of each parameter could **deviate** from the starting value. MINUIT is not very sensitive to the value of the starting interval (as long as it is > 0) .

Tools for graphics

The *optional* part of this exercise requires that you give a graphical representation of the experimental data and of the results of the fit, using the *root*-facilities.

- an example of how to draw the fitted curve can be found in */user/uvak/sda/example2.C*
- an example of how to draw the experimental data-points (including their error-bars), and of how to include a print of the fitted values, can be found in: */user/uvak/sda/example5.C*

Presentation of results

HAND IN :

- A print of your program code
- A print of the fitted values of the parameters, their statistical errors and the value of the χ^2 of the fit
- A graphical representation of the results (if made)

10 Exercise Simulation of a gas

(*root* exercise 9)

Problem statement

This exercise is based on the following article:
"A simple computer simulation of molecular collisions leading to the Maxwell distribution"
J.Ftacnik, P.Lichard and J.Pisut
European Journal of Physics, Volume 4(1983), page 68-71.

- It is well known that we can consider a gas (in a closed volume) as a collection of 'free moving' particles (atoms or molecules). The particles move in 'arbitrary' directions, and may collide with each other or with the walls of the volume.
- The velocity of each particle is in principle *random*. However, Statistical Physics tells that the ensemble of the velocities follows the Maxwell distribution.
- In this exercise we will simulate a system of colliding particles
  - We start with a velocity distribution which does not look at all like the one proposed by Maxwell.
  - Then we let the particles collide pairwise, and calculate their velocities after the collision, using the conservation laws of energy and momentum.
  - After a large number of collisions we look at the velocity distribution. We will see that the distribution is the one predicted by Maxwell. Apparently this is the natural result of a large number of collisions.

Model and formulas

For the sake of simplicity we consider a 2-dimensional gas, consisting of point like particles with identical mass, and an interaction mechanism which occurs in collisions between 'hard spheres'. In the article above (and in textbooks about Classical Mechanics) we find :

- When two particles with velocities V_1 and V_2 collide under an angle α (between 0 and 2π), the following holds for the situation after the collision:
  - The angle γ between the particles after the collision is random between (0 and 2π); all angles have the same probability.
  - The velocities after the collision are given by

$$V_{1,2}^2 = \frac{1}{2}\left(V_1^2 + V_2^2\right) \pm \left[\frac{1}{2}\left(V_1^2 - V_2^2\right)\cos\gamma + V_1 V_2 \sin\alpha \sin\gamma\right]$$

- In the article mentioned above (and in textbooks on Statistical Physics) we find the following formula for the Maxwell distribution (in two dimensions !):

$$f(V) \;\sim V \;\; \exp\left(-\frac{V^2}{V_{av}^2}\right)$$

where V_{av}^2 is the average quadratic velocity.

Building the program

We can build the program in the following way :

- Start with a collection of 1000 particles. Give them all the same starting velocity (this is clearly not Maxwell). Store the velocities in a table.
- Simulate a large number of collisions (e.g. 100000), as follows :
  - choose (at random) two particles which are going to collide; this fixes V_1 and V_2 .
  - choose (at random) the collision angle α (uniformly between 0 and 2π)
  - choose (at random) the angle γ after the collision (uniformly between 0 and 2π)
  - calculate the velocities after the collision, and replace (in the table) the old velocities by the new ones.
- Compare the final velocity distribution with the Maxwell prediction. This can be done in the following way :
  - put the experimental distribution in a histogram with e.g. 50 intervals, extending from 0 to V_{max} (the largest velocity in the table).
  - fit the Maxwell distribution to the histogram

Hints and tools

It is advisable to add a few checks in order to verify the working of your program, like :

- make histograms to show how frequently each particle is selected to participate in a collision
- make histograms to show the distribution of the angles α and γ

root provides facilities to fit a user-defined function to a histogram. An example can be found in : */user/uvak/sda/example7.C*

Presentation of results

HAND IN :

- A print of your program code
- A graphical representation of the results