Statistical Methods in Experimental Physics

For internal use only.
Furthermore, please be aware that these notes have only been checked by me. It would therefore not be too surprising if some typos have survived to this stage. Therefore: *use the notes only for your conceptual understanding and overview of the subject.* If a formula does not look quite right to you it is probably because of a typo. If a formula disagrees with the textbook, it is almost certainly because of a typo in these lecture notes. If you notice any such typos, please send an email to: alessandro.pastore@york.ac.uk.

Chapter 1

Basic measures of data

1.1 Data types

We can identify two main categories of data

Discrete data *i.e* the outcome of tossing a coin, the number of physics graduates ...

Continuous data *i.e* the heigth of students

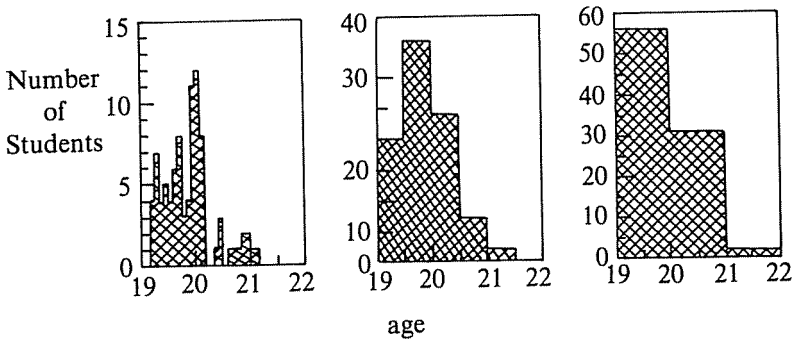

Fig. 2.2 The ages (in years) of a group of second year students, showing the effects of choosing different bin sizes for the same data.

Figure 1.1: Age distribution, Barlow p. 5.

A critical first step in data analysis is very often the decisions relating to displaying the data. Often, as you see here, the data are displayed in a histogram. In figure 1.1 this is shown for a set of single-variable data, but it could have been multi-variable. The effect of different choices in binning is clearly seen.

Requirements for representation of data: effective, easy to grasp, honest. There are many ways to lie (to yourself) with statistics: non-causal correlations (examples in figure), binning of data (peaks), displaying data (scale), comparing incompatible data (salaries).

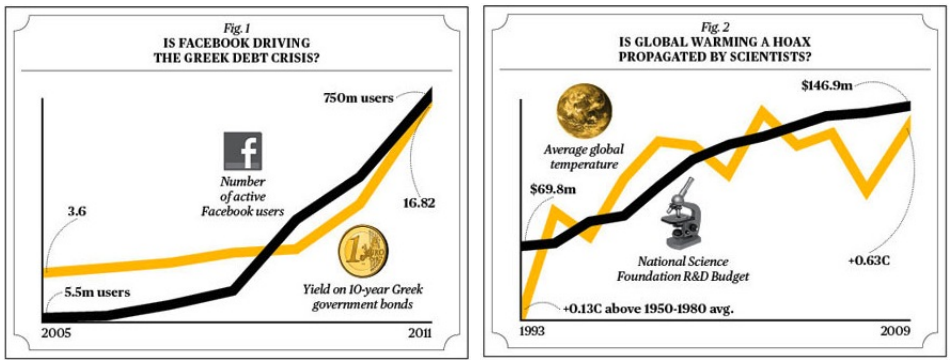

Figure 1.2: Non-causal correlations [V. Chandrasekaran, *Correlation or Causation*, Bloomberg Businessweek, 1 Dec 2011].

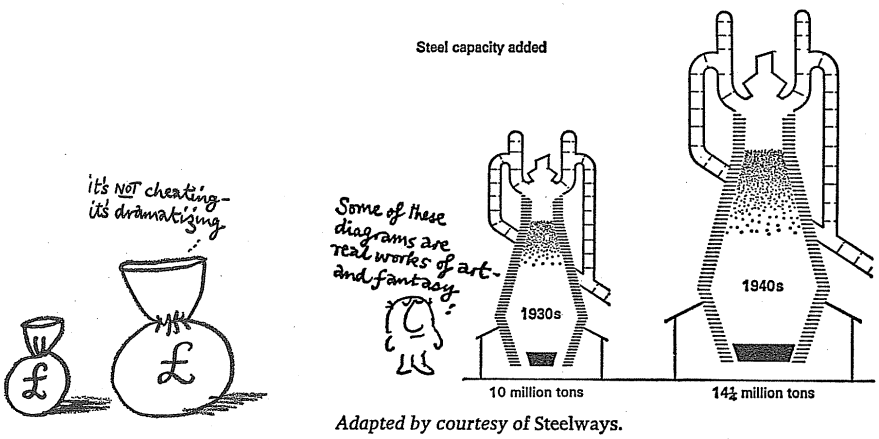

Figure 1.3: Graphical manipulation and scale illusions (D. Huff, *How to lie with statistics*).

1.2 Describing the data

To describe a set of data, we can summerize the main information by using some indicators, as for example the *mean*.

There are different possible definitions for the *mean* of a set of data

When Dewey was elected Governor in 1942, the minimum teacher's salary in some districts was as low as $900 a year. Today the school teachers in New York State enjoy the highest salaries in the world. Upon Governor Dewey's recommendation, based on the findings of a Committee he appointed, the Legislature in 1947 appropriated $32,000,000 out of a state surplus to provide an immediate increase in the salaries of school teachers. As a result the minimum salaries of teachers in New York City range from $2,500 to $5,325.

Figure 1.4: Comparing incompatible data [D. Huff, *How to lie with statistics*, Penguin, 1991]. This is particularly critical in physics, where unless we are careful, the conditions under which two measurements are taken may affect the results in ways we hadn't predicted and didn't notice, making a comparison between the measurements meaningless.

1.2.1 Arithmetic mean

Given a set of data $\{x_1, x_2, \ldots, x_N\}$ of N points, we can calculate the mean value as

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \overline{f(x)} &= \frac{1}{N}\sum_{i=1}^{N} f(x_i) \end{aligned}$$

In the case of binned data, we have

$$\bar{x} = \frac{1}{N}\sum_{j=1} n_j x_j$$

where now the index j runs over the bins and n_j is the quantity of data in each bin.

Example: Scott took 7 math test with results: 89,73,84,91,87,77,94. The arithmetic mean is

$$\bar{x} = \frac{89+73+84+91+87+77+94}{7} = 85 \tag{1.1}$$

1.2.2 Alternative definitions

There are 3 alternative definitions:

- Geometric mean: $\mu_g = \sqrt[N]{x_1 x_2 \dots x_n} \Rightarrow \ln(\mu_g) = \frac{1}{N}\sum_i \ln x_i$

  Geometric mean is used in finance to calculate average growth rates. *Example*: Consider a stock that grows by 10% in year 1, then decreases of 20% in year 2 and then up again of 30% in year 3.

$$\mu_g = \sqrt[3]{(1+0.1)(1-0.2)(1+0.3} = 1.046 \tag{1.2}$$

$$\mu_a = \frac{1.1+0.8+1.03}{3} = 1.066 \tag{1.3}$$

- Harmonic mean : $\mu_h = \frac{N}{1/x_1+1/x_2+\dots+1/x_n}$

  Example Joe drives a car at 20mph for 1h and 30 mph in the 2nd hour of the journey. We use the here the simple arithmetic average

$$\mu_a = \frac{20+30}{2} = 25mph \tag{1.4}$$

  Now Joes continues driving and for 1/2 of the jurney he drives at 30mph and for the other 1/2 at 20mph. What's the average speed per segment?

  In this case is more appropriate to use the harmonic mean

$$\mu_h = \frac{2}{\frac{1}{20}+\frac{1}{30}} = 24mph \tag{1.5}$$

It is important that they satisfy the relation

$$\mu_a \geq \mu_g \geq \mu_h \tag{1.6}$$

it is also useful to define the peak value (Mode)and the central data point (Median).

1.2.3 Dispersion of data

To characterize the data, the mean value is not sufficien since it does not give any information concerning the distribution of data aroun the mean value $\bar{x}$. It is thus useful to define the *variance* $V(x)$ as

$$V(x) = \overline{(x-\bar{x})^2} = \frac{1}{N}\sum_i (x_i-\bar{x})^2 = \overline{x^2} - \bar{x}^2, \tag{1.7}$$

The last equality can be easily proved[1]

[1]To demonstrate the equality we start from the definition of $V(x)$

The standard deviation of the set is thus $\sigma = \sqrt{V(x)} = \sqrt{\bar{x^2} - \bar{x}^2}$. The standard deviation help us describing the data *Example* Given the 2 data set

	Set1	Set2
x_1	3	1
x_2	4	2
x_3	4	4
x_4	5	5
x_5	6	7
x_6	8	11

We notice that

$$\mu_{Set1} = \mu_{Set2} = 5 \tag{1.8}$$

The deviation is quite different...

$$\sigma_{Set1} = 1.63\sigma_{Set2} = 3.33 \tag{1.9}$$

In the second data set the data are more scattered around the mean value μ.

1.2.4 Correlations

Having multi-variable sets of data as in the case

$$\{(x_1, y_1), (x_2, y_2), \dots\}$$

we can use the individual means $\bar{x}, \bar{y}$ and variances $V(x), V(y)$ to characterize the data, but we can also check if there are correlations between them. For example if x, y, represent a simultaneous mesurments, we verify if and

---

$$\begin{aligned} V(x) &= \frac{1}{N}\sum_i (x_i - \bar{x})^2 \\ &= \frac{1}{N}\sum_i x_i^2 - 2x_i\bar{x} + \bar{x}^2 \\ &= \frac{1}{N}\sum_i x_i^2 - 2\bar{x}\frac{1}{N}\sum_i x_i + \bar{x}^2 \\ &= \bar{x^2} - 2\bar{x}^2 + \bar{x} = \bar{x^2} - \bar{x}^2 \end{aligned}$$

how the measurment of x impacts the measurment of y. To this purpose we define the covariance

$$\mathrm{cov}(x,y) = \frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\cdot\bar{y}$$

Notice that $cov(x,x) = V(x)$. The covariance is used to define the correlation coefficient

$$\rho_{xy} = \mathrm{cov}(x,y)/\sigma_x\sigma_y \tag{1.10}$$

the main advantage of using ρ is that it is a pure number and it is included in the interval $[-1,1]$. When $\rho = 0$ the data are not correlated, when $\rho = \pm 1$ the data are correlated/anti-correlated. In figure 1.5), we show some set of data and their correlation coefficient ρ

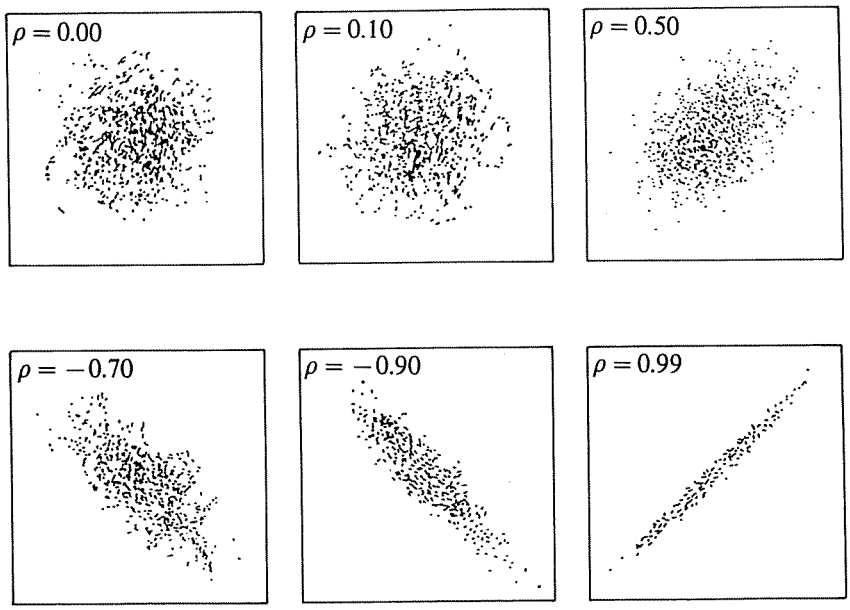

Fig. 2.4 Scatter plots showing examples of correlation. The scales and origin of the axes are irrelevant (see text) and are therefore not shown.

Figure 1.5: Correlations in a two-parameter data set. Absolute scale on parameters has no impact on ρ.

The concept of correlations can be generalized to multi-variable sets of data, by defining 2 by 2 correlations .

Example We want to test if there is a correlation between the number of vouchers offered by a shop and the number of sales per day.

x (voucher)	y (sales)
1	8
3	6
2	9
5	4
8	3
7	3
12	2
2	7
4	7

We can calculate the average value of x,y and the variances

$$\bar{x} = 4.89 \; \sigma_x = 3.34 \tag{1.11}$$

$$\bar{y} = 5.44 \; \sigma_y = 2.36 \tag{1.12}$$

To check if there is a correlation in the data we calculate

$$cov = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = -7.172 \tag{1.13}$$

and

$$\rho = -0.908 \tag{1.14}$$

There is anti-correlation!!!!

Chapter 2

Probability

2.1 General properties

Probability is a deductive discipline relying on pre-existing informations. Statistics deals with observations of the real world.

Example The dice is thrown and lands one side up. What is the probability we get a "3"?

2.2 Definition and meaning

The mathematical definition of probability states that: for a set of events $E_1, E_2, \ldots$, the probability of each $P(E)$ is a function ascribed to each event which complies with the following:

1. $P(E) \geq 0$ for any event E.
2. $P(E_1||E_2) = P(E_1) + P(E_2)$ if E_1 and E_2 are mutually exclusive.
3. $\sum_i P(E_i) = 1$ over all mutually exclusive events.

Note, that as a direct consequence of 1 and 3, we know that the probability of the negated of $\bar{E}$, that is the probability that E does not happen is:

$$P(\bar{E}) = \sum_i P(E_i) - P(E) = 1 - P(E)$$

2.2.1 Meaning of "probability"

How to define the concept of *probability*?

The frequentist view on probability

Repeat an experiment N times and let N_x be the number of times X happens. The probability of X is then defined as:

$$\frac{N_X}{N} \to P(X) \text{ for } N \to \infty$$

This means we cannot define P for a single experiment, but is dependent on the existence of an ensemble of (identical) experiments. As a consequence we must require:

Unique ensemble: If the choice of the collective ensemble changes $P(X)$, $P(X)$ is not well-defined. Examples of different ensembles are that of the risk of a driver damaging a car. You will get different values of the probability depending on whether you restrict the ensemble to: students, elderly, men, women in the 40–50 year age-bracket, foreigners, etc. If I belong to three such groups can you get the risk that I crash my car in the next year from any of these sample studies? The answer, strictly speaking, is no (and this was in fact put into UK law recently, resulting in a significant rise in insurance cost for female drivers).

Repeatable experiments: We cannot (in the frequentist view) meaningfully define: P(snow next christmas); P(rain tomorrow); P(England winning the 2013 Six Nations); P(Big-bang giving rise to the laws of nature we observe today).

Objective

Alternatively, we may see probability as an intrinsic property of a coin, a die, or an experiment. The problem is that this view on probability only carries predictive power if the "object" in question carries an intrinsic symmetry giving rise to the probability. In particular this is a problem for most continuous stochastic variable because the assumed probability distribution must be different for different functions of the variable ($\ln(x)$, $\sin(x)$, $\sqrt{x}$, x^2, ...). So in practice the "objective" view on probability is rarely used, though to be honest, this is probably(!) how most of us think of daily-life probabilities, i.e. as an intrinsic property.

Subjective

This view on probability is known as the Bayesian view. It is not, however, called subjective because the probability is just about who you are, what you think about the matter in question. Rather, it is called "subjective" because the probability depends on the information available to the "subject" evaluating the probability. We will return to this approach to probability next week.

Example

We have a jar with 3 yellow balls, 2 red, 2 green and 1 blue. $P(yellow) = \frac{3}{8}$. P(yellow in 3 rounds (no replacement))=?

$$P = \frac{3}{8} \times \frac{2}{7} \times \frac{1}{6} \tag{2.1}$$

P(not yellow)=P($\bar{yellow}$)?

$$P(\bar{yellow}) = 1 - P(yellow) = 1 - \frac{3}{8} = \frac{5}{8} \tag{2.2}$$

Example Suppose we have 2 dices and the sum of the faces is 7. What is the probability of 1 face having a "5"?

We need to calculate P("5"having $sum = 7$) =?

There are 36 possible combinations (6x6), but only 6 giving a sum of 7.

$$(1-6)\ (4-3)\ (2-5)\ (5-2)\ (3-4)\ (6-1) \tag{2.3}$$

So $P(sum = 7) = 6/36 = \frac{1}{6}$ We have

$$P(\text{"5" and } sum = 7) = \frac{2}{36} = \frac{1}{18} \tag{2.4}$$

So we have

$$P(\text{"5"having } sum = 7) = \frac{P(\text{"5" and } sum = 7)}{P(sum = 7)} = \frac{1}{3} \tag{2.5}$$

2.3 Bayesian analysis

We get the important result about conditional probability.

The probability of event A happens given B $P(A|B)$. If the two event are independent then $P(A|B) = P(A)$

We get the Bayes theorem

$$P(A|B)P(B) = P(A\&B) = P(B|A)P(A) \tag{2.6}$$

or in a more conveninet way

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.7}$$

The Bayesian view on probability is known as the subjective view. Not because the probability is just about who you are, what you think about the matter in question. Rather, it is called "subjective" because the probability depends on the information available to the "subject" evaluating the probability:

$$\begin{aligned} P(a|b)P(b) &= P(a\&b) = P(b|a)P(a) \\ P(\text{theory}|\text{result}) &= \frac{P(\text{result}|\text{theory})}{P(\text{result})}P(\text{theory}) \end{aligned}$$

This is a general statement about probabilities, and is derived solely from the mathematical definition, so the statement is independent on our interpretation of the meaning of probabilities. It can therefore be used to combine (prior) knowledge about the reliability of a theory $P(\text{theory})$ with knowledge about the intrinsic nature of the experiment $P(\text{result})$ (and knowledge about how the theory restricts the result and thereby changes the probabilities for individual measurements $P(\text{result}|\text{theory})$. By using this as above, we can evaluate how the probability of the theory has changed now that we can add this new measurement $P(\text{theory}|\text{result})$ which is exactly what we are interested in. However, it is critical for this to work, that we have a detailed (and reliable) evaluation of what the probability (distribution) for the model was before the experiment. A typical choice, in cases where we know nothing beforehand, is for example a uniform distribution for the parameter of interest x. However, uniform in which parameter and over which range ($\ln(x)$, $\sin(x)$, $\sqrt{x}$, x^2, ...)?

2.3.1 Analysis of the Monty-Hall problem

Optional item: Though Bayesian analysis is also known in some areas of physics, most usage of the method has been in other fields. As a simple example of the methodology, let us consider the Monty Hall problem, in which the participant of a show is offered the chance of winning either a car or a goat. These are hid behind three doors (say a read, a green, and a blue), two hiding a goat and one a car. The show then goes as follows: the participant chooses a door, after which the host (knowing where the goats are) opens one of the other two (naturally not one that hides the car). Following this, the participant is again offered whether to stick to the original choice or change to the other unopened door.

The Monty Hall problem is in essence a problem of evaluating probabilities as they change with the information available, i.e. conditional probabilities as they develop throughout the show. For a Bayesian analysis of the problem, let us by A_r, A_g, and A_b denote the events where the red, green, and blue door hides the car (as opposed to a goat). This clearly means

$$P(A_r) = P(A_g) = P(A_b) = \frac{1}{3},$$

because of symmetry of the system. This is therefore an example of an "objective" probability.

Let then similarly C_r, C_g, and C_b denote the choice of the participant of one of the three doors. As the choice in itself does not yield any additional information about the system, the conditional probabilities, given for example the choice C_r, are still:

$$P(A_r|C_r) = P(A_g|C_r) = P(A_b|C_r) = \frac{1}{3}.$$

However, when the host subsequently opens one of the other two doors (say the green door), should the participant change his/her choice?

We could, naively, think that the probabilities now must be:

$$P(A_r|H_g) = P(A_b|H_g) = \frac{1}{2},$$

and that it therefore is irrelevant whether the participant changes his/her choice, since the probabilities are anyway 50:50. However, the probability we should be calculating is:

$$P(A_r|H_g, C_r),$$

the probability of the event A_r given the fact that both C_r and H_g has happened. Bayes' theorem stated that:

$$P(A|B)P(B) = P(B|A)P(A).$$

Furthermore, it doesn't change anything in the Bayesian analysis that all of these might be calculated under a common condition, i.e. we also have:

$$P(A|B, C)P(B|C) = P(B|A, C)P(A|C).$$

For the present example, this means:

$$P(A_r|H_g, C_r) \cdot P(H_g|C_r) = P(H_g|A_r, C_r) \cdot P(A_r|C_r),$$

such that:

$$P(A_r|H_g, C_r) = \frac{P(H_g|A_r, C_r) \cdot P(A_r|C_r)}{P(H_g|C_r)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

Similarly, the probability for the car to be behind the blue door is:

$$P(A_b|H_g, C_r) = \frac{P(H_g|A_b, C_r) \cdot P(A_b|C_r)}{P(H_g|C_r)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3},$$

since the host will certainly would open the green door if the car was to be behind the blue, that is: $P(H_g|A_b, C_r) = 1$. Likewise, we can evaluate the

probability of the car being behind the green door even without looking, since the host would not open the green door if the car was behind that door:

$$P(A_g|H_g, C_r) = \frac{P(H_g|A_g, C_r) \cdot P(A_g|C_r)}{P(H_g|C_r)} = \frac{0 \cdot \frac{1}{3}}{\frac{1}{2}} = 0.$$

We can therefore conclude that the participant can double his/her chances for winning the car by taking the other closed door, from a probability of $\frac{1}{3}$ to a probability of $\frac{2}{3}$.

2.3.2 Example:drug test

Suppose we have a drug test with 99% reliability. This means 99 positive tests out 100 on actual drug users and 99 negative tests (out of 100) on non drug users. Let assume the probability of randomly selecting one person to be a drug user is P(User)=0.5.

What is the probability of having a drug user in case of positive test? $P(User|+)$

$$\begin{aligned} P(User|+) &= \frac{P(+|User)P(User)}{P(+)} \\ &= \frac{P(+|User)P(User)}{P(+|User)P(User) + P(+|Non-User)P(Non-User)} \\ &= \frac{0.99 \times 0.05}{0.99 \times 0.05 + 0.01 \times 0.995} \end{aligned}$$

2.3.3 Example: lamp test

A desk lamp was found to be defective (D). Knowing that there are 3 factories:A,B,C that produce that lamp with following data

	% oflamps produced	% defects
A	0.35	0.015
B	0.35	0.01
C	0.3	0.02

If a randomly selcted lamp has a defect, which is the probability of being produced in C?

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \tag{2.8}$$

We need to calculate the probability of having a defect

$$P(D) = P(A|D)P(A)+P(B|D)P(B)+P(C|D)P(C) = 0.015*0.35+0.01*0.35+0.02*0.3 \tag{2.9}$$

so we have

$$P(C|D) = \frac{0.02 * 0.3}{0.01475} = 0.407 \tag{2.10}$$

and similarly we can calculate

$$P(B|D) = 0.237 \tag{2.11}$$

$$P(A|D) = 0.356 \tag{2.12}$$

Chapter 3

Probability Distributions

Let's consider now an experiment that has more than 2 possible outcomes A_j with $j = 1, 2, \ldots$ and $P(A_j)$. We shoudl respect the laws of probability

- The probability is always a positive number

$$P(A_j) \geq 0 \ \ \forall \ \ A_j \tag{3.1}$$

- The probability is normalised

$$\sum_j P(A_j) = 1 \tag{3.2}$$

- If 2 event are intependent

$$P(A_1 \& A_2) = P(A_1)P(A_2) \tag{3.3}$$

We now consider the outcome of an experiment a continuum. Let x be any real number, we define a probability distribution $f(x)$ a functiona having the following properties

$$f(x) \ \geq \ 0 \tag{3.4}$$

$$\int dx f(x) \ = \ 1 \tag{3.5}$$

$$P(a \leq x \leq b) \ = \ \int_a^b f(x)dx \tag{3.6}$$

Since we deal with continuos functions most of the time we need to define the probability in a range. $f(x)$ is called probability density distribution.

Example Consider the function

$$f(x) = \begin{cases} 0 & x < 0 \\ a\exp(-ax) & x \geq 0 \end{cases} \tag{3.7}$$

where $a > 0$. We notice that

- It is single valued
- it is positive everywhere
- it is normalised

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{0}^{\infty} a\exp(-ax)dx & (3.8) \\ &= -exp(-ax)|_0^{\infty} = 1 & (3.9) \end{aligned}$$

and

$$\begin{aligned} \int_{x_a}^{x_b} f(x)dx &= \int_{x_a}^{x_b} a\exp(-ax)dx & (3.10) \\ &= \exp(-ax_a) - \exp(-ax_b) & (3.11) \end{aligned}$$

We define a cumulative probability distribution (CDF) as

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{3.12}$$

To characterise the distribution, we can define the *moments* defined as

$$M_m = \int_{-\infty}^{\infty} x^m f(x)dx = \langle x^m \rangle \tag{3.13}$$

The M_0 moment is just the normalisation

$$M_0 = \int_{-\infty}^{\infty} x^0 f(x)dx = 1 \tag{3.14}$$

The M_1 is the mean

$$M_1 = \int_{-\infty}^{\infty} x^1 f(x)dx = \tag{3.15}$$

Example

$$\begin{aligned}\langle x\rangle &= \int_{-\infty}^{\infty} xa\exp(-ax)dx &(3.16)\\ &= -\frac{1}{a}(1+ax)\exp(-ax)|_0^{\infty} &(3.17)\\ &= \frac{1}{a} &(3.18)\end{aligned}$$

The CDF is

$$F(x) = \int_{\infty}^{y} a\exp(-ax)dx = 1-\exp^{-ax} \tag{3.19}$$

The variance of a distribution can be calculated as

$$\begin{aligned}\sigma^2 &= \langle (x-\mu)^2\rangle &(3.20)\\ &= \langle x^2 - 2x\mu + \mu^2\rangle &(3.21)\\ &= \langle x^2\rangle - 2\mu\langle x\rangle + \mu^2 &(3.22)\\ &= M_2 - 2\mu^2 + \mu^2 &(3.23)\\ &= M_2 - M_1^2 &(3.24)\end{aligned}$$

Example
We can calculate the vairance of the exponentail distribution as

$$M_2 = \int_0^{\infty} x^2 a\exp(-ax) = \frac{2}{a^2}M_1 = \int_0^{\infty} xa\exp(-ax) = \frac{1}{a} \tag{3.25}$$

so we have

$$\sigma^2 = M_2 - M_1^2 = \frac{1}{a^2} \tag{3.26}$$

3.1 Famous/Common distributions

We have different probability distributions, but hte most know/used are

- Binomial distribution
- Poisson distribution
- Gaussian distribution
- Student's t distribution

other distributions exist, but they are less interesting for nuclear physics.

3.1.1 Binomial distribution

It is used to describe process with a given number of identical trials with two possible outcomes that we define as

- p: success probability
- $1-p$ failure probability

this is the distribution to describe the outcome of tossing a coin for example

The probability distribution of having a number of success events r out of n trials is given by

$$P_B(r;p,n) = p^r(1-p)^{n-r}\frac{n!}{r!(n-r)!}. \tag{3.27}$$

Here, the first two terms are the probability for the drawing of r "successes" followed by $n-r$ "failures". The last term is the number of possible reorderings of the events. T The main properties of this distribution are

$$\langle r\rangle = pn \tag{3.28}$$
$$V(r) = np(1-p) \tag{3.29}$$

3.1.2 Poisson distribution

The Poisson distribution deals with random events and is defined only by it's fixed average expected number of events (it is used for example to describe radioactive decays) and is in practice the infinite limit of the binomial distribution for an (infinitely) large pot with (infinitely) small individual probabilities. Let us define the expectation value for the number of successful draws (events) as $\lambda = \langle r\rangle = pn$ for a total (very large) number of draws as n at probability p. The probability for each draw to be successful must then be $p = \lambda/n$, and the probability for obtaining r events is given by the binomial distribution $P(r;p,n) = P(r;\lambda/n,n)$.

$$P(r,\lambda/n,n) = \left(\frac{\lambda}{n}\right)^r\left(1-\frac{\lambda}{n}\right)^{n-r}\frac{n!}{r!(n-r)!}$$

In the limit where n approaches infinity while λ is kept fixed (for a fixed value of r): $(1-\lambda/n)^{n-r} \to (1-\lambda/n)^n \to e^{-\lambda}$ (where that latter is the Taylor expansion of the exponential). $n!/(n-r)! \to n^r$ as the former is the product of the r numbers from $(n-r+1)$ up to n. The derivation of this is detailed in the following.

$$\frac{n!}{r!(n-r)!} = \frac{1\cdot 2\cdot\cdots\cdot n}{1\cdot 2\cdot\cdots\cdot n-r} = n\cdot(n-1)\cdot\cdots\cdot(n-r+1)$$

$$
\begin{aligned}
&= n^r \left(1 \cdot (1 - \frac{1}{n}) \cdot (1 - \frac{2}{n}) \cdot \cdots \cdot (1 - \frac{r-1}{n})\right) \\
&\quad \text{which to first order in } 1/n \text{ yields :} \\
&\to n^r \left(1 - \frac{\sum_{i=1}^{r-1} i}{n}\right) = n^r \left(1 - \frac{(r-1)r}{2n}\right) \\
&\to n^r \text{as } n \to \infty \text{ for fixed } r
\end{aligned}
$$

and similarly :

$$
\begin{aligned}
\left(1 - \frac{\lambda}{n}\right)^{n-r} &= \left(1 - \frac{\lambda}{n}\right)^{n} \cdot \left(1 - \frac{\lambda}{n}\right)^{-r} \\
&\to e^{-\lambda} \cdot 1
\end{aligned}
$$

yielding :

$$
P(r, \lambda/n, n) \to \lambda^r \frac{1}{n^r} e^{-\lambda} n^r \frac{1}{r!} = \frac{e^{-\lambda}\lambda^r}{r!}
$$

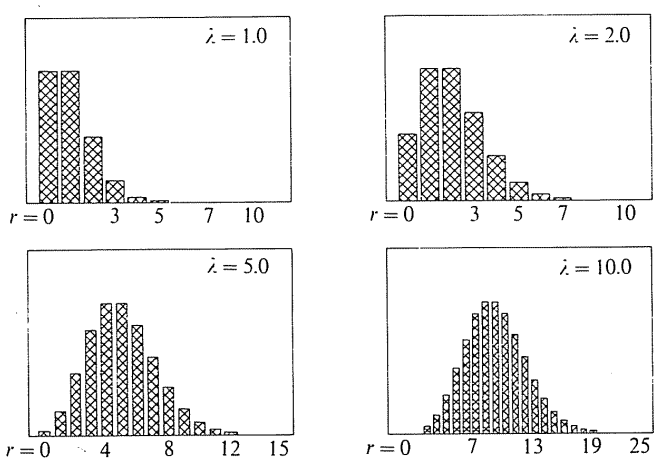

Figure 3.1: Poisson distributions for varying mean values λ are shown. The vertical scale is arbitrary.

The properties of this distribution are

$$\langle r \rangle = \lambda \tag{3.30}$$

$$V(r) = \lambda \tag{3.31}$$

3.1.3 Gaussian Distribution

Another common distribution is the Gaussian one. It reads

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The properties of the distribution are

$$\langle r \rangle = \mu \tag{3.32}$$
$$V(x) = \sigma^2 \tag{3.33}$$

In Fig.3.2, we show the Gaussian distribution for $\mu = 0$.

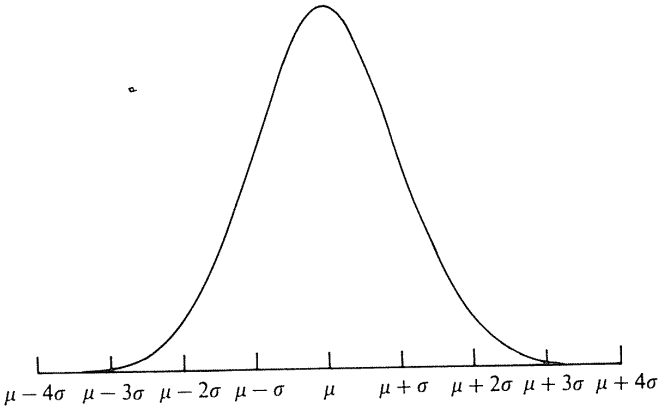

Fig. 3.3. The Gaussian distribution.

Figure 3.2: Gaussian distribution, for $\mu = 0$, $\sigma = 1$ called the nomal distribution, Barlow p. 34.

This distribution is arguably the most well-known continuous probability distributions. This is in part because it is a limit for many other probability distributions, such as for example the Poisson distribution:

$$P_{Poiss}(r;\lambda) \to P_{Gauss}(r;\lambda,\sqrt{\lambda}) \text{ as } \lambda \to \infty$$

In figure 3.1 the approach towards a gaussian distribution os clearly visible. Also, the Gaussian distribution is extremely important in error handling, as almost all errors are Gaussian.

3.1.4 Student's t distribution

This distribution takes the name after N.S. Gossert who used *Student* pseudonym to publish his results. The distribution looks like

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{3.34}$$

where Γ is a function of the form

$$\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)dx \tag{3.35}$$

We will use this distribution later for hypothesis testing. The main features of the distribution are

- The mean $\langle x\rangle = 0$ for $\nu > 1$
- The variance is $V = \frac{\nu}{\nu-2}$ for $\nu > 2$ and ∞ for $1 \leq \nu \leq 2$

Chapter 4

Errors

4.1 The Central Limit Theorem

Measurments acquire errors from many different sources. A very important result is that (almost) all statistical fluctuations or errors are Gaussian[1]. This can be explained by advokating the Central Limit Theorem (CLT).

Specifically, the central limit theorem states:

Central Limit Theorem:
If $\{x_1, x_2, \ldots, x_N\}$ are independent stochastic variables each with mean μ_i and variances V_i, and we define $X = \sum_i x_i$, then:

(a) $\langle X \rangle = \sum_i \mu_i$ (always)

(b) $V(X) = \sum_i V_i$ (independent)

(c) X becomes Gaussian for $N \to \infty$ ("well behaved")

For (c), x_i must be 'nicely behaved' which particularly means that $\langle x^r \rangle$ must be well-defined for all r (and not increase too rapidly with r). Furthermore, is should be noted that (c) is best (*i.e.* the convergence is fastest) for the central part of the distribution, slower for the tails of the distribution.

In the following we will thus show the proofs of the theorem

4.1.1 Proofs of (a) and (b)

Let's consider the first staement (a). We have used $X = \sum_i x_i$ Since integrals and sums commute (i.e. the order in which they are taken does not matter),

[1] For example the heights of British men, the measurement errors when measuring the voltage across a resistor, etc.

we have:

$$\langle X \rangle \;=\; \langle \sum_i x_i \rangle = \sum_i \langle x_i \rangle = \sum_i \mu_i$$

for the second proof (b), we start from the definition

$$\begin{aligned}
V(X) &= \langle (X - \langle X \rangle)^2 \rangle \\
&= \langle (\sum_i x_i - \sum_i \mu_i)^2 \rangle \\
&= \langle (\sum_i (x_i - \mu_i))^2 \rangle \\
&= \langle \sum_i (x_i - \mu_i)^2 + \sum_{i \neq j} (x_i - \mu_i)(x_j - \mu_j) \rangle \\
&= \langle \sum_i (x_i - \mu_i)^2 \rangle + \langle \sum_{i \neq j} (x_i - \mu_i)(x_j - \mu_j) \rangle \\
&= \sum_i \langle (x_i - \mu_i)^2 \rangle + \sum_{i \neq j} \langle (x_i - \mu_i)(x_j - \mu_j) \rangle \\
&= \sum_i V_i + \sum_{i \neq j} \mathrm{cov}_{ij} = \sum_i V_i
\end{aligned}$$

as in the latter term, the sum of all covariance terms, each of which are required to be zero in the assumption that all $\{x_i\}$ are independent. For (c) see example in figure 4.1 and proof in appendix 2.

Example: random numbers

Let's consider the distributions shown in Fig.4.1.

- **A** Histogram of 5000 random numbers taken from a uniform distribution with $\mu = \frac{1}{2}$ and $V(x) = \frac{1}{12}$
- **B** Histogram of 5000 numbers each is the sum of pair of random numbers of the type **A** $X = x_1 + x_2$. The mean value is $\langle x \rangle = 1$ and $V(x) = \frac{1}{6}$.
- **C** Histogram of 5000 numbers each is the sum of three randome numbers
- **D** Histogram of 5000 numbers each is the sum of twelve randome numbers. The mean value is $\langle x \rangle = 6$ and $V(x) = 1$.

We observe that the distribution **D** approaches a Gaussian. The convergence in the central part is quite fast, while the tails are not still well converged.

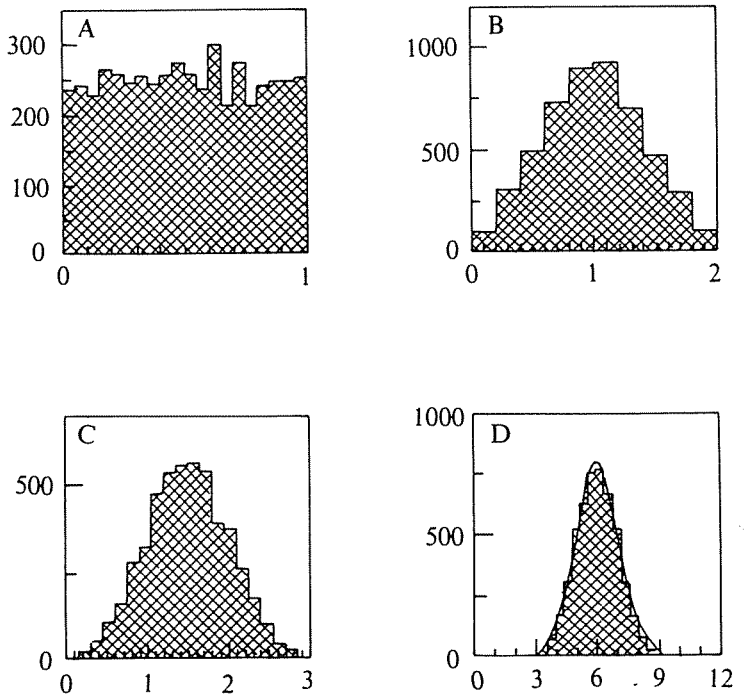

Fig. 4.1. The CLT at work.

Figure 4.1: The central limit theorem in action for uniform distributions.

4.2 Working with errors

The expressions for the expectation value of a sum of random variables and the resulting variance as based on (a) and (b) in the central limit theorem is in fact the foundation for much of our practical work with errors. For example, let's say that we plan to make N measurements x_i of a single physical observable, they therefore all have the same expectation values $\mu_i = \mu$ and if they are measured in the same way, also the same variance $V = \sigma^2$. These measurements are all independent stochastic variables as required in the CLT.

$$\begin{aligned} X &\equiv \sum_i x_i \\ \langle X \rangle &= \sum \mu = N\mu \end{aligned}$$

we can thus define the *a*verage of the average $\langle \bar{x} \rangle$ as

$$\begin{aligned}\langle\bar{x}\rangle &= \langle\frac{1}{N}X\rangle = \frac{1}{N}\langle\sum_i x_i\rangle \\ &= \frac{1}{N}\sum_i\langle x_i\rangle = \frac{1}{N}N\mu = \mu\end{aligned}$$

Similarly, the variance of the mean is:

$$\begin{aligned}V(\bar{x}) &= V\left(\frac{1}{N}\sum_i x_i\right) \\ &= \left\langle\left(\left(\frac{1}{N}\sum_i x_i\right)-\mu\right)^2\right\rangle \\ &= \frac{1}{N^2}\left\langle\left(\left(\sum_i x_i\right)-N\mu\right)^2\right\rangle \\ &= \frac{1}{N^2}\left\langle(X-\langle X\rangle)^2\right\rangle = \frac{1}{N^2}V(X) \\ &= \frac{1}{N^2}\sum_i V_i = \frac{1}{N^2}\sum_i\sigma^2 \quad \text{(as } x_i, x_j \text{ are independent for } i\neq j\text{)} \\ &= \frac{\sigma^2}{N}\end{aligned}$$

So we have $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$. By increasing the number of measurements we can thus reduce the errors (not the systematic one!)

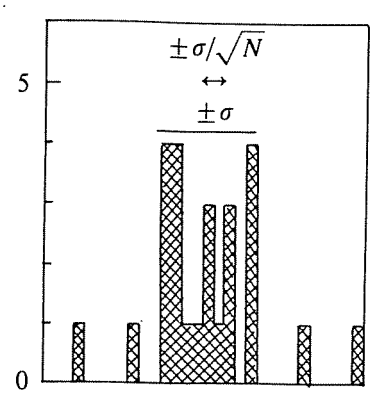

Fig. 4.2. Standard errors.

Figure 4.2: Distribution of measurements according to the standard error. The scaling of the uncertainty in the mean, $\sigma(\bar{x}) = \sigma/\sqrt{N}$, is indicated.

Example: Eggs

The weigth of eggs produced by a farmer have a standard deviation of $\sigma = 10$g. He feeds a group of hens at extra vitamins whcih will pay back if there is a weigth increase of 2 g per egg. He measures 25 eggs from vitamin fed abd he founds an average increase of 3 g. Does it prove anything useful?

$$\sigma_{\bar{x}} = \frac{10}{\sqrt{25}} = 2 \tag{4.1}$$

The increase is only $1.5\sigma_{\bar{x}}$. Not really significant.

4.3 Error propagation

To understand the concept of error propagation, let's consider a function $f(x)$ which is calculate starting from the stochastic variable x

$$f(x) = ax + b \tag{4.2}$$

a and b are constants. The variance of $V(f)$ reads

$$\begin{aligned} V(f) &= \langle f^2 \rangle - \langle f \rangle^2 \\ &= \langle a^2x^2 + 2abx + b^2 \rangle - (a\langle x \rangle + b)^2 \\ &= a^2\langle x^2 \rangle + 2ab\langle x \rangle + b^2 - a^2\langle x \rangle^2 - 2ab\langle x \rangle - b^2 \\ &= a^2\left(\langle x^2 \rangle - \langle x \rangle^2\right) \\ &= a^2 V(x) \end{aligned}$$

In the general case, the uncertainty σ_f depends not only on the uncertainty in x, σ_x, but also on the value of x.

Assuming that errors are small, we can Taylor expand the function f(x) around the observed value x_0

$$f(x) \approx f(x_0) + (x - x_0)\left(\frac{df}{dx}\right)\Big|_{x=x_0} + \dots$$

from which we can deduce

$$V(f) \approx \left(\frac{df}{dx}\right)^2 V(x)\sigma_f \approx \left|\frac{df}{dx}\right| \sigma_x \tag{4.3}$$

Example

θ is an anlge with error $\sigma_\theta = \pm 0.01$ rad. Which is the error on $\sin\theta$?

$$\sigma_{\sin\theta} = \left|\frac{d\sin\theta}{d\theta}\right| \sigma_\theta = 0.01|\cos\theta| \tag{4.4}$$

4.3.1 Functions of two variables

For functions of two variables, we have

$$\begin{aligned} f(x) &= ax + by + c \\ V(f) &= a^2 V(x) + b^2 V(x) + 2ab\ \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) &= \langle xy \rangle - \langle x \rangle \langle y \rangle \end{aligned}$$

in the case of *independent* variables (*i.e* cox(x,y)=0) we have

$$\sigma_f^2 = \frac{\partial f}{\partial x}^2 \sigma_x^2 + \frac{\partial f}{\partial y}^2 \sigma_y^2 + \dots$$

Example
Calculate error propagation of

$$f(\theta) = A \sin\theta + B \cos\theta \tag{4.5}$$

$$\sigma_f^2 = \sin^2\theta \sigma_A^2 + \cos^2\theta \sigma_B^2 + (A\cos\theta - B\sin\theta)\sigma_\theta^2 \tag{4.6}$$

We observe that the errors added in quadrature. In the general case the errors will be correlated so it useful to introduce a different notation

Now suppose there are several m different functions $f_1, f_2, \dots f_m$ of n different variables $x_1 \dots x_n$ The variables x_i have an error associated with them as a result the functions f_j will be correlated to one another (even if the x are not) because different f_j share the same x_i.

The variance of each function is

$$V(f_i) = \langle f_i^2 \rangle - \langle f_i \rangle^2 \tag{4.7}$$

by making Taylor expansion around the mean we have

$$f_i \approx f(\mu_1, ...\mu_N) + \left(\frac{\partial f_i}{\partial x_1}\right)(x_1 - \mu_1) + \dots \tag{4.8}$$

We insert this expansion into the expression for the variance and we get

$$V(f_i) = \left(\frac{\partial f_i}{\partial x_1}\right)^2 \langle (x_1 - \mu_1)^2 \rangle + ... + 2\frac{\partial f_i}{\partial x_2}\frac{\partial f_i}{\partial x_1}\langle (x_1 - \mu_1)(x_2 - \mu_2)\rangle \tag{4.9}$$

$$= \sum_j \left(\frac{\partial f_i}{\partial x_1}\right) V(x_i) + \sum_j \sum_{k \neq j} \frac{\partial f_i}{\partial x_j}\frac{\partial f_i}{\partial x_k} cov(x_i, x_j) \tag{4.10}$$

The matrix V_x is also called *error matrix*

$$V_{ij} = cov(x_{(i)}, x_{(j)}) \tag{4.11}$$

For several functions, the results for the functions will in many cases be correlated, and in the general case of several functions (f_k) of several variables (x_i), the uncertainties on the function values and correlations between variables is most practically computed using a matrix formalism. The covariance matrix (or error matrix) for the variables (V_x) is then defined as the matrix where entry (i, j) is the covariance $\mathrm{cov}(x_i, x_j)$, such that the matrix is symmetric with the individual-variable uncertainties along the diagonal. The corresponding covariance matrix for the functions (V_f) may then be found as the matrix product:

$$V_f = G V_x \tilde{G},$$

where the matrix G is defined with entries $G_{ki} = \frac{\partial f_k}{\partial x_i}$, and $\widetilde{G}$ is the transposed of G, in Matlab and Octave defined as G'.

Example

Let's consider a traking chamber. The position of a particle is measured in cylindrical polar coordinates (r, ϕ, z). Assuming that the error over r is zero, while we have $\sigma_\phi^2, \sigma_z^2$ for the other 2 coordinates (no correlation!). What are the errors on the cartesian coordinates (x, y, z)?

Having the transformation

$$\begin{aligned} x &= r cos\phi \\ y &= r sin\phi \\ z &= z \end{aligned}$$

we can calculate the G matrix as

$$G = \begin{pmatrix} cos\phi & -r sin\phi & 0 \\ sin\phi & r cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \; ; \; V_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_\phi^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \tag{4.12}$$

so that we have

$$V_{xyz} = \begin{pmatrix} \sigma_\phi^2 y^2 & -\sigma_\phi^2 xy & 0 \\ -\sigma_\phi^2 xy & \sigma_\phi^2 x^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \tag{4.13}$$

4.4 Systematic errors

4.4.1 Estimating systematic errors

An error that affects several measurements in the same way are denoted systematic. As an example we can consider a voltmeter measurements which have a built-in systematic error (percentage of the measured value). Background on measurements introduce a systematic zero-point error. Systematic errors always introduce correlations between measurements, and are therefore cases where the previous formalism is particularly useful. Naturally, if you explicitly know the exact value of a systematic shift, it is no longer an error, but you can simply correct for it thereby eliminating the error. Very often, however, we cannot correct for all errors, and will have to include the systematic errors in the calculation of the uncertainty of the result in question. These systematic errors come in three categories:

1. Explicitly known errors (such as that of the voltmeter).
2. Known but must be evaluated (this is the usual case, and an example is given in the following).
3. Unexpected and unidentified (means your result is wrong, but you won't even know that).

As an example of the second case where we can evaluate the potential error, let us assume that we are using a ^{210}Pb source of an (original) activity of 370 kBq. However, with the half life being only $T_{1/2} = 21$ y, the reduction at the time of use must be taken into account if we want to compare to a measured rate in a detector. We can estimate the error in the following way:

- Let's say we know it to be very unlikely for the source to be more than 5 years old.
- This means the minimum possible activity is 314 kBq, so the actual activity must be somewhere in this range.
- If we assume a uniform distribution over the possible activity range, we get $\sigma = (\max - \min)/\sqrt{12} = 16$ kBq.

Assuming an actual activity of 342 ± 16 kBq is therefore a good (and probably conservative) estimate, yielding a systematic uncertainty of 5 % to be taken into account the comparison to measurements.

4.4.2 Handling systematic errors

Let's assume that two stochastical variables x and y have (independent) random errors σ_x,σ_y and a common systematic error s. To describe the errors in x, y data, we must describe both the errors on x and y as well as

the x, y covariance. For this, it is practical to describe the values of x and y as arising from the sum of two components, one carrying only the random error and one carrying only the systematic error:

$$\begin{aligned} V(x) &= \sigma_1^2 + s^2 \\ V(y) &= \sigma_2^2 + s^2 \\ cov(x,y) &= s^2 \end{aligned}$$

So that we can write the error matrix as

$$V_{xy} = \begin{pmatrix} \sigma_x^2 + s^2 & s^2 \\ s^2 & \sigma_y^2 + s^2 \end{pmatrix}$$

A similar calculation can be carried out if the systematic error is a relative error, or in cases where more than two variables are in play.

4.4.3 G-matrix example, one linear function

As an example in the use of the G-matrix notation, let's say we have the two measurements x, y with the above errors. Now if the physical parameter we are after is instead: $f(x,y) = ax + by$, then:

$$\begin{aligned} G &= \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = (a,b) \\ \widetilde{G} &= \begin{pmatrix} a \\ b \end{pmatrix}. \end{aligned}$$

This yields for the (1×1) error matrix for f (i.e. the variance V_f):

$$\begin{aligned} V_f &= GV_{xy}\widetilde{G} = (a,b) \begin{pmatrix} \sigma_x^2 + s^2 & s^2 \\ s^2 & \sigma_y^2 + s^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \\ &= a^2\sigma_x^2 + b^2\sigma_y^2 + (a+b)^2\, s^2. \end{aligned}$$

The last expression emphasises that the effect of the systematic error depends critically on the values of a and b, particularly whether they are of the same or opposite sign. If the systematic uncertainty is an absolute uncertainty (rather than relative) and if a and b are of the same absolute value but of opposite sign, the systematic uncertainty cancels out because it affects x and y in the same way, and therefore cancels out in the difference. A similar argument can be made for ratios x/y when the error is a relative error. The previous expression can also be re-written as

$$V_f = a^2\left(\sigma_x^2 + s^2\right) + b^2\left(\sigma_y^2 + s^2\right) + 2ab \cdot s^2$$

which is exactly the same as was derived for a general function of two variables. This is seen as the total errors on x and y in the present case are $\sqrt{\sigma_x^2 + s^2}$ and $\sqrt{\sigma_y^2 + s^2}$ with the covariance between x and y being s^2.

Chapter 5

Estimators

In statistics 'estimation' is a technical term at difference from the meaning of everyday life. It has nothing to do with approximation

An *estimator* is a procedure applied to the data sample which gives a numerical value for a property of the parent population or a property/parameter of the parent distribution. Suppose that the quantity we want to measure is called a. $\hat{a}$ is an estimator.

Let's consider an estimator $\hat{a}$ to find the height a of all students of the University on the basis of a sample N. Let's consider the different estimators

1. Add up all the heigths and divide by N
2. Add up the first 10 heights and divide by 10. Ignore the rest
3. Add up all the heigths and divide by N-1
4. Trow away the data and give 1.8 as answer
5. Add up the second, fourth, sixth,... heights and divide by N/2 for N even and (N-1)/2 for N odd

How to chose the estimator? It has to respond to some criteria

Consistent: $\lim_{N\to\infty} \hat{a} = a$, that is you can get as close to the true value as you want, as long as you have a large enough data set. In the previous example 1 is consistent:

$$\hat{\mu} = \frac{x_1 ... x_N}{N} = \bar{x} \tag{5.1}$$

for N going to infinity $\bar{x} \to \mu$: law of big numbers. 3 nad 5 are also consistent since N-1 or N/2 make little difference when $N \to \infty$. On the contrary 2 and 4 are not consistent.

Unbiased: $\langle \hat{a} \rangle = a \, \forall \, N$, that is however large or small your data set may be, you should on average expect to get the right answer. The expectation value of the estimator is equal to the true value. For 1 we have

$$\langle \hat{\mu} \rangle = \langle \frac{x_1 \dots x_N}{N} \rangle = \frac{1}{N} \left(\langle x_1 \rangle + \langle x_N \rangle \right) = \frac{N \langle x \rangle}{N} = \mu \tag{5.2}$$

For 3 we have

$$\langle \hat{\mu} \rangle = \langle \frac{x_1 \dots x_N}{N-1} \rangle = \frac{1}{N-1} \left(\langle x_1 \rangle + \langle x_N \rangle \right) = \frac{N \langle x \rangle}{N-1} \tag{5.3}$$

so 3 is biased. While for 5

$$\langle \hat{\mu} \rangle = \langle \frac{x_2 \dots x_N}{N/2} \rangle = \frac{1}{N/2} \left(\langle x_2 \rangle + \langle x_N \rangle \right) = \frac{N/2 \langle x \rangle}{N/2} = \mu \tag{5.4}$$

Efficient: $V(\hat{a})$ is small, that is the fluctuations around the true value is for a given size of the data set smaller than for less efficient estimators. In general if the variance is smaller we prefer the estimator, so the main difference between 1 and 5 is that 5 uses only half of the data set thus its variance is $\sqrt{2}$ larger.

This leads on to why among the eight estimators, not all are equally good. Among the different estimators presented here, the estimator 1 is *consistent*, *unbiased* and is the most *efficient*.

5.1 Specific estimators

We consider here some specific estimators

5.1.1 Estimating the mean

Two estimators are commonly used, given a data set $\{x_i, \sigma_i\}$. Depending on the situation we have:

$$\text{unweigthed mean} \quad \begin{cases} \bar{x} &= \frac{1}{N} \sum_i x_i \\ V(\bar{x}) &= \frac{1}{N^2} \sum_i \sigma_i^2 \end{cases}$$

$$\text{weigthed mean} \quad \begin{cases} \bar{x} &= \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \\ V(\bar{x}) &= \frac{1}{\sum_i 1/\sigma_i^2} \end{cases}$$

Both estimators are *consistent* and *unbiased*. We observe that the second type should be preferable if the data points are not equally well determined, i.e. if σ_i are different.

5.1.2 Estimating the variance

We consider the ideal case where the true mean is known μ. The estimator of the variance is thus

$$\widehat{V(x)} = \frac{1}{N}\sum_i (x_i - \mu)^2$$

We can show that it is consistent and unbiased

$$\begin{aligned}\left\langle \widehat{V(x)} \right\rangle &= \frac{1}{N} N \left\langle (x-\mu)^2 \right\rangle \\ &= \left\langle (x-\mu)^2 \right\rangle = V(x)\end{aligned}$$

Now let's consider the case where μ is not know. An obvious remedy is to use $\bar{x}$ so that

$$\widehat{V(x)} = \frac{1}{N}\sum_i (x_i - \bar{x})^2$$

We can prove that such an estimator is biased

$$\begin{aligned}\widehat{V(x)} &= \frac{1}{N}\sum_i (x_i - \bar{x})^2 \\ &= \frac{1}{N}\sum_i \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \frac{1}{N}\left(\sum_i x_i^2 - 2\bar{x}\sum_i x_i + \sum_i \bar{x}^2\right) \\ &= \frac{1}{N}\left(\sum_i x_i^2 - 2N\bar{x}\bar{x} + \sum_i \bar{x}^2\right) \\ &= \frac{1}{N}\left(\sum_i x_i^2 - 2\sum_i \bar{x}^2 + \sum_i \bar{x}^2\right) \\ &= \frac{1}{N}\sum_i \left(x_i^2 - \bar{x}^2\right)\end{aligned}$$

We now take the expectation value of the estimator

$$\begin{aligned}\left\langle \widehat{V(x)} \right\rangle &= \left\langle x^2 - \bar{x}^2 \right\rangle \\ &= \left\langle x^2 \right\rangle - \left\langle \bar{x}^2 \right\rangle \\ &= \left\langle x^2 \right\rangle - \langle x \rangle^2 + \langle \bar{x} \rangle^2 - \left\langle \bar{x}^2 \right\rangle \qquad \text{[CLT]}\end{aligned}$$

The latter identity $\langle x \rangle = \langle \bar{x} \rangle$ comes from the central limit theorem (CLT). So we have

$$\begin{aligned} \left\langle \widehat{V(x)} \right\rangle &= V(x) - V(\bar{x}) \\ &= V(x) - \frac{1}{N} V(x) \quad \text{[CLT]} \\ &= \frac{N-1}{N} V(x) \end{aligned}$$

The bias fall as $1/N$ so for large data set this can be neglected. A way to correct the bias is to defined

$$s^2 \equiv \widehat{V(x)} \equiv \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

where the multiplication factor $\frac{N}{N-1}$ is known as Bessel's correction. So $\widehat{V(x)} \equiv \frac{1}{N} \sum_i (x_i - \bar{x})^2$ is biased, but the unbiased estimator $\widehat{V(x)} \equiv \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$ is less efficient. Either way, defining $\widehat{\sigma} \equiv \sqrt{\widehat{V(x)}}$ gives:

$$\begin{aligned} V\left(\widehat{V(x)}\right) = V(\widehat{\sigma}^2) &= \left.\frac{\mathrm{d}\sigma^2}{\mathrm{d}\sigma}\right|^2_{\sigma=\hat{\sigma}} V(\widehat{\sigma}) = (2\hat{\sigma})^2 V(\widehat{\sigma}) \\ &= 4\hat{\sigma}^2 V(\widehat{\sigma}) \\ \Rightarrow & \\ \sigma(\widehat{\sigma}) &= \sqrt{V(\widehat{\sigma})} = \sqrt{V(\hat{\sigma}^2)\frac{1}{4\hat{\sigma}^2}} = \sqrt{\frac{2\hat{\sigma}^4}{N}\frac{1}{4\hat{\sigma}^2}} \\ &= \frac{\widehat{\sigma}}{\sqrt{2N}} \text{ or } \frac{\widehat{\sigma}}{\sqrt{2(N-1)}}, \end{aligned}$$

The first is for the biased, but efficient, maximum-likelihood estimator of σ^2 (see next chapter), and does not include the Bessel correction. The second is the unbiased estimator of σ^2, $\widehat{\sigma^2} = s^2$, which includes the Bessel correction, but is less efficient.[1]

[1]Note, that there is no guarantee that because s^2 is an unbiased estimator of the variance (σ^2), s is an unbiased estimator of σ. However, because s^2 is unbiased we will nevertheless refer to s as the unbiased estimator.

Chapter 6

The Maximum Likelihood estimator

6.1 Definition and use

Given a data sample $X = \{x_1, x_2, \ldots, x_N\}$ one applies an estimator $\hat{a}$ for the quantity a. The data values x_i are drawn from some probability density function $P(x, a)$ which depends on a. The form of P is given and a specified. The probability of a data set is the product of the individual probabilities.

$$\begin{aligned} L(x_1, x_2, \ldots, x_N; a) &= P(x_1; a)P(x_2; a)\ldots P(x_N; a) \\ &= \Pi_i P(x_i; a) \end{aligned}$$

This product is called *likelihood* in Fig.6.1 we show an example. It follows that the expectation value of any function of the sample is found by integrating over all possible values of all the x_i weighted by the total probability

$$\begin{aligned} \langle f(x_1, \ldots, x_N)\rangle &= \int \ldots \int f(x_1, \ldots, x_N)L(x_1, \ldots, x_N)dx_1 \ldots dx_N \\ &= \int fLdX \end{aligned}$$

In particular for some estimators we have

$$\begin{aligned} \langle \hat{a}\rangle &= \int \hat{a}LdX \\ \langle \hat{a}^2\rangle &= \int \hat{a}^2LdX \end{aligned}$$

where dX is shorthand notation for the variables of the sample.

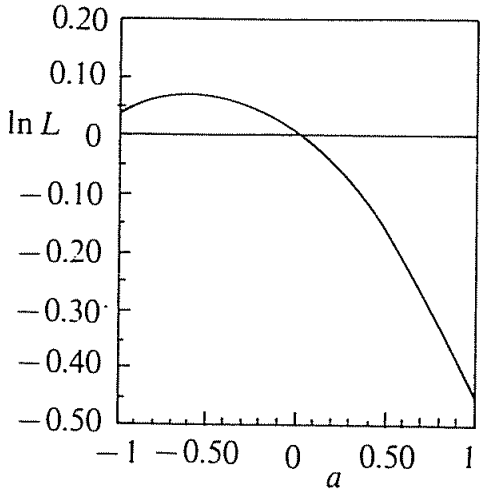

Fig. 5.1. A likelihood function.

Figure 6.1: Likelihood function for one parameter.

Maximum Likelihood Estimator:

The ML estimator for a parameter a is then the procedure which evaluates the parameter value $\hat{a}$ which makes the actual observations X as likely as possible, that is the set of parameters $\hat{a}$ which maximises $L(X;a)$. In practice, the logarithm of L (the log-likelihood function) is more practical to work with computationally, as sums are computationally much easier to handle than products:

$$\begin{aligned} \ln L(X;a) &= \ln\left(\prod_i P(x_i;a)\right) \\ &= \sum_i \ln P(x_i;a), \end{aligned}$$

The ML estimator $\hat{a}$ is then the value of a which maximises $\ln L(X;a)$ (or minimises $-\ln L(X;a)$). This can be found (in some cases analytically) by:

$$\left.\frac{\mathrm{d}\ln L}{\mathrm{d}\,a}\right|_{a=\hat{a}} = 0$$

6.1.1 Example: lifetime

For decays with a lifetime τ, the (normalised) probability distribution as a function of time t is: $P(t;\tau) = \frac{1}{\tau}\exp(-t/\tau)$. We calculate the function $\ln L$

for this distribution

$$\begin{aligned} \ln L &= \sum_i \ln\left(\frac{1}{\tau}\exp(-t/\tau)\right) \\ &= \sum_i \left(-\ln\tau - t_i/\tau\right) \end{aligned}$$

differentiating with respect to τ and putting to zero we obtain the estimator $\hat{\tau}$

$$\left.\frac{\mathrm{d}\,\ln L}{\mathrm{d}\,\tau}\right|_{\tau=\hat{\tau}} = 0 \Leftrightarrow \sum_i (t_i - \hat{\tau}) = 0 \Leftrightarrow \hat{\tau} = \frac{1}{N}\sum_i t_i$$

This can also be done of the time measurement is restricted, i.e. $P(t;\tau)$ doesn't extend to $t = \infty$, which is always the case experimentally (at least in principle).

6.1.2 Example: lifetime with acceptance

Assuming that the time acceptance is T, *i.e.* the instrument does not register any decay bigger than T. The new probability function reads

$$P(t;\tau) = \frac{1}{\tau}\exp(-t/\tau)\frac{1}{1-e^{-T/\tau}} \tag{6.1}$$

We can calculate

$$\ln L = \sum_i \left[-\frac{t}{\tau} - \ln(\tau - e^{-T/\tau}\tau)\right] \tag{6.2}$$

and differentiating respect to τ and then equating to zero.

$$\left.\frac{d\ln L}{d\tau}\right|_{\tau=\hat{\tau}} = 0 = \sum_i \left(\frac{t_i}{\tau^2} + \frac{T}{\tau^2(-1+e^{T/\tau})} - \frac{1}{\tau}\right) \tag{6.3}$$

so we have

$$\hat{\tau} = \frac{1}{N}\sum_i t_i + \frac{Te^{-T/\hat{\tau}}}{1-e^{-T/\hat{\tau}}} \tag{6.4}$$

which has to be solved numerically

6.2 Expectation values of functions

The likelihood function $L(X;a)$ has further use, since it is the multidimensional probability density for observing the data set X. As with the continuous probability distributions we looked at previously we therefore have, for any function of such data $f(X)$:

$$\begin{aligned}
\langle f(X)\rangle &= \int_X f(X)L(X;a)\mathrm{d}X \\
\langle \hat{a}\rangle &= \int_X \hat{a}(X)L(X;a)\mathrm{d}X \\
\left\langle \hat{a}^2\right\rangle &= \int_X \hat{a}(X)^2L(X;a)\mathrm{d}X \\
\left\langle \left(\frac{\mathrm{d}\,\ln L(X;a)}{\mathrm{d}a}\right)^2\right\rangle &= \int_X \left(\frac{\mathrm{d}\,\ln L(X;a)}{\mathrm{d}a}\right)^2 L(X;a)\mathrm{d}X.
\end{aligned}$$

The latter is a particularly useful function, since it can be used to give a quantitative, absolute, measure of the "efficiency" we discussed previously. This is called the Minimum Variance Bound (MLB), and states that:

$$V(\hat{a}) \geq \frac{1}{\left\langle \left(\frac{\mathrm{d}\,\ln L}{\mathrm{d}a}\right)^2\right\rangle},$$

for all unbiased estimators $\hat{a}$. So the best we can hope for in terms of efficiency given a set of data X is obtaining a variance $V(\hat{a})$ as low as that defined by the MVB.

6.3 Specific ML estimators: Weighed mean

The gaussian weighed mean for set of data $\{x_i\}$ with a common mean μ and different (non-zero) uncertainties σ_i can be found using the ML formalism:

$$\begin{aligned}
P(x_i;\mu,\sigma_i) &= \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \\
\ln L &= -\sum_i \ln(\sigma_i\sqrt{2\pi}) - \sum_i \frac{(x_i-\mu)^2}{2\sigma_i^2},
\end{aligned}$$

such that:

$$\begin{aligned}
0 &= \left.\frac{\mathrm{d}\,\ln L}{\mathrm{d}\,\mu}\right|_{\mu=\hat{\mu}} \\
\Leftrightarrow & \\
0 &= \sum_i \frac{2(x_i-\mu)}{2\sigma_i^2}
\end{aligned}$$

$$\begin{aligned} & \Leftrightarrow \\ \sum_i \frac{x_i}{\sigma_i^2} &= \sum_i \frac{1}{\sigma_i}\hat{\mu} \\ & \Leftrightarrow \\ \hat{\mu} &= \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i} \end{aligned}$$

which shows that the weighed mean estimator as we know it is in fact the ML estimator of the mean, assuming Gaussian errors on the data points.

6.4 Properties of the ML estimator

So how good are ML estimators?

Consistency: Usually, ML estimators are consistent. That is for increasing data sets, the estimator approaches the true value of the parameter.

Efficiency: A consistent ML estimator is unbiased in the $N \to \infty$ limit, and:

$$V(\hat{a}) = \text{MVB} = \frac{1}{\left\langle \left(\frac{\mathrm{d}\,\ln L}{\mathrm{d}a}\right)^2 \right\rangle} = \frac{-1}{\left\langle \left(\frac{\mathrm{d}^2\,\ln L}{\mathrm{d}a^2}\right) \right\rangle}$$

Biasedness: However, the ML estimator is often biased. The ML estimator for the spread of a gaussian data set is:

$$\hat{\sigma} = \frac{1}{N}\sum_i (x_i - \bar{x})^2,$$

which as we know is biased (does not include Bessel's correction). This can be seen by calculating the likelihood function for a data set $\{x_i\}$ with a common mean μ and uncertainty σ:

$$P(x_i; \mu, \sigma_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

and find the maximum, where the partial derivatives are zero:

$$\frac{\partial\,\ln L}{\partial\,\mu} = 0\ , \frac{\partial\,\ln L}{\partial\,\sigma} = 0$$

6.5 Errors on ML estimators

The error on a ML estimator can be found as discussed in relation to the efficiency:

$$\widehat{\sigma}_{\hat{a}}^2 = V(\hat{a}) = \text{MVB} = \frac{1}{\left\langle \left(\frac{\mathrm{d}\,\ln L}{\mathrm{d}a}\right)^2 \right\rangle} = \frac{-1}{\left\langle \left(\frac{\mathrm{d}^2\,\ln L}{\mathrm{d}a^2}\right) \right\rangle}$$

for unbiased, efficient, ML estimators. In particular for any consistent ML estimator at $N \to \infty$.

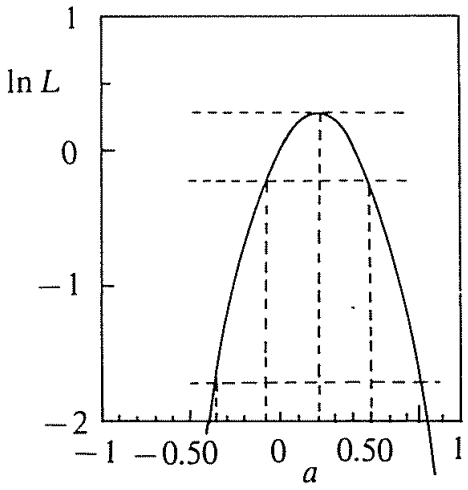

Fig. 5.2. A log likelihood function showing the 1σ and 2σ limits.

Figure 6.2: Errors on Maximum-Likelihood estimators in one dimension.

If the likelihood function is well described by a parabola around it's maximum, this is equivalent to finding the distances from the estimated value, where $\ln L$ has decreased by $1/2$ ($1\sigma(\hat{a})$), $2^2 \cdot 1/2$ ($2\sigma(\hat{a})$), or $3^2 \cdot 1/2$ ($3\sigma(\hat{a})$). Or equivalently, you may minimise $-2\ln L$ and find the distances at which it increases by 1, 2^2, 3^2, etc. This function is particularly useful for two reasons. The first is practical, the second theoretical:

- Minimisation routines are more commonly implemented compared to maximisation routines. Minimising $-2\ln L$ (or at least $-\ln L$) would therefore often be the preferred option, practically.
- $-2\ln L$ behaves like a χ^2 function for large data sets, which is seen in part by the increase by 1, 2^2, 3^2, etc, for 1σ, 2σ, 3σ etc.

In fact, where the local region around the maximum/minimum is non-parabolic, using the increase is a better measure of the error than the curvature.

For several variables (N), the vector of estimated parameters $\widehat{\vec{a}}$ is found by minimising the N-dimensional function $-2\ln L$, as shown for two parameters in figure 6.3. The inverse of the covariance matrix is then given

by:

$$\mathrm{cov}^{-1}(a_i, a_j) = - \left. \frac{\partial^2 \ln L}{\partial a_i \partial a_j} \right|_{\vec{a}=\widehat{\vec{a}}},$$

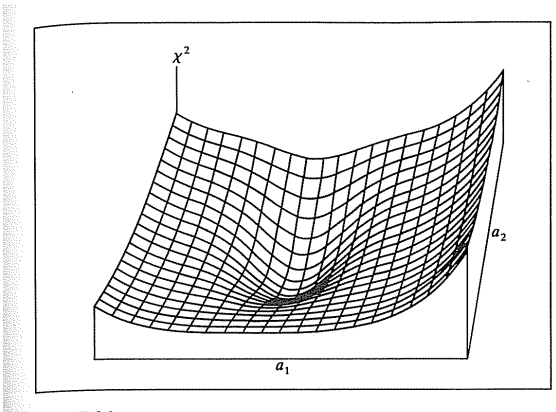

FIGURE 8.2
Chi-square hypersurface as a function of two parameters.

Figure 6.3: $-2 \ln L$ in two dimensions.

6.6 Summary for the ML estimator

- ML estimator $\hat{a}$ does not give the "most likely" value for a, rather it takes the value of a which makes the data set as likely as possible.
- In the large-N limit, the ML estimator gives: $\hat{a}$ unbiased; $V(\hat{a}) =$ MVB, which means it is an "optimal" estimator.
- No need to bin the data set.
- However: for small N, ML is usually (in the general case) biased, and $V(\hat{a}) >$ MVB (typically).
- The ML estimator does not give goodness-of-fit.
- Maximisation of the likelihood can be numerically heavy. Minimise $-2 \ln L$ instead, using well-tested minimisation routines.

Chapter 7

Fitting of data

7.1 Least squares fitting

For a function $f(x_i; a)$ and a data set $\{x_i, y_i, \sigma_i\}$, assuming for each data point y_i is drawn from a Gaussian distribution of mean $f(x_i; a)$ and spread σ_i, we know that the likelihood function must obey:

$$\begin{aligned} -2\ln L &= -2\sum_i \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_i - f(x_i;a))^2}{2\sigma_i^2}\right)\right) \\ &= 2\sum_i \ln\left(\sigma\sqrt{2\pi}\right) + \sum_i \frac{(y_i - f(x_i;a))^2}{\sigma_i^2} \end{aligned}$$

Instead of minimising $-2\ln L$ we may therefore equivalently define and minimise:

$$\chi^2 \equiv \sum_i \frac{(y_i - f(x_i;a))^2}{\sigma_i^2}$$

I.e. chi-squared minimisation, as you know it, is in fact the maximum-likelihood estimator of the function parameters a. This means, as you may recall, that it comes with the nice properties of the ML estimator:

- It obeys invariance, such that the optimum function is independent of the specific parameterisation of the function.
- It is consistent (at least typically).
- The bias is most often small.
- It is efficient asymptotically ($N \to \infty$), and $\pm 1\sigma$ can be found by identifying where χ^2 changes by 1 from it's minimum:

$$V(\hat{a}) = -\frac{1}{\frac{\partial^2 \ln L}{\partial a^2}} = \frac{2}{\frac{\partial^2 \chi^2}{\partial a^2}}$$

7.1.1 χ^2

The χ^2 is a statistical distribution which is built as a sum of squares of k-gaussian (random) variables

$$\chi^2 = \sum_i^k z_i^2 \tag{7.1}$$

The corresponding PDF for the case of a continuos variable can be expressed as

$$\chi^2 = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} e^{-x/2} x^{k/2-1} \tag{7.2}$$

We can calculate the mean μ and the variance V and we get

$$\mu = k \tag{7.3}$$
$$V = 2k \tag{7.4}$$

7.2 Linear least squares

7.2.1 Fitting a straight line

For a straight line fit to a data set $\{x_i, y_i\}$ with common uncertainty σ we have $f(x_i; m, c) = m \cdot x_i + c$, and:

$$\chi^2 = \sum_i \frac{(y_i - f(x_i; a))^2}{\sigma^2} = \sum_i \frac{(y_i - mx_i - c)^2}{\sigma^2}$$

We now differentiate and equate to zero as

$$\begin{aligned} \left.\frac{\partial \chi^2}{\partial c}\right|_{m=\widehat{m},c=\widehat{c}} &= \frac{1}{\sigma^2} \sum_i -2\left(y_i - \widehat{m}x_i - \widehat{c}\right) = 0 \\ &\Rightarrow \\ 0 &= \overline{y} - \widehat{m}\overline{x} - \widehat{c} \end{aligned}$$

and

$$\begin{aligned} \left.\frac{\partial \chi^2}{\partial m}\right|_{m=\widehat{m},c=\widehat{c}} &= \frac{1}{\sigma^2} \sum_i -2\left(y_i - \widehat{m}x_i - \widehat{c}\right) x_i = 0 \\ &\Rightarrow \\ 0 &= \overline{xy} - \widehat{m}\overline{x^2} - \widehat{c}\overline{x} \end{aligned}$$

$$\begin{aligned} &= \overline{xy} - \widehat{m}\overline{x^2} - (\overline{y} - \widehat{m}\overline{x})\,\overline{x} \\ &= \overline{xy} - \widehat{m}\overline{x^2} - \overline{y}\,\overline{x} + \widehat{m}\overline{x}^2 \end{aligned}$$

so we have

$$\hat{m} = \frac{\bar{x}\bar{y} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{cov(x,y)}{V(x)} \tag{7.5}$$

$$\hat{c} = \bar{y} - \hat{m}\bar{x} \tag{7.6}$$

We thereby have established that $\widehat{m}$ and $\widehat{c}$ are linear in y_i

7.2.2 Uncertainties and covariances for straight-line fit

We can now calculate the errors on previous quantities by acting with derivatives over the ML.

$$cov^{-1}(a_i, a_j) = \frac{1}{2}\frac{\partial^2\chi}{\partial a_i \partial a_j}\bigg|_{\vec{a}=\hat{a}} \tag{7.7}$$

$$\begin{aligned} \frac{1}{2}\frac{\partial^2\chi^2}{\partial c^2}\bigg|_{m=\widehat{m},c=\widehat{c}} &= \frac{1}{\sigma^2}\sum_i 1 = \frac{N}{\sigma^2} \\ \frac{1}{2}\frac{\partial^2\chi^2}{\partial m\partial c}\bigg|_{m=\widehat{m},c=\widehat{c}} &= \frac{1}{\sigma^2}\sum_i x_i = \frac{N}{\sigma^2}\overline{x} \\ \frac{1}{2}\frac{\partial^2\chi^2}{\partial m^2}\bigg|_{m=\widehat{m},c=\widehat{c}} &= \frac{1}{\sigma^2}\sum_i x_i^2 = \frac{N}{\sigma^2}\overline{x^2}. \end{aligned}$$

We therefore get the inverse of the covariance matrix for (m, c) to be:

$$(V_{cm})^{-1} = \frac{N}{\sigma^2}\begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}.$$

The easiest way of inverting the matrix (if you don't want to go through it by hand) is to use Matlab's analytical tools:

Matlab command line input and output for symbolic matrix inversion, where `x` and `x2` denotes $\overline{x}$ and $\overline{x^2}$ respectively:

```
>> syms x2 x
>> Vm1 = [x2,x;x,1]
Vm1 =
[ x2, x]
[ x, 1]
>> Vxx = inv(Vm1)
Vxx =
[ 1/(- x^2 + x2), -x/(- x^2 + x2)]
[ -x/(- x^2 + x2), x2/(- x^2 + x2)]
```

That is, we get:

$$V_{cm} = \frac{\sigma^2}{N \cdot \left(\overline{x^2} - \overline{x}^2\right)} \begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{pmatrix} = \frac{\sigma^2}{N \cdot V(x)} \begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{pmatrix},$$

where σ is the (common) uncertainty for the $\{y_i\}$ measurements and $V(x)$ is the variance in the $\{x_i\}$ data. This means the variances (and covariance) in the fitted parameters c and m scales with the square of the uncertainty in $\{y_i\}$, as it should, and inversely with both N and $V(x)$, such that a data set with many measurements and an extended measurement range has a reduced uncertainty (and variance) on the fitted parameters.

7.2.3 General function linear in all parameters

The methods from section 7.2.1 can be generalised to cover functions f which are linear in all parameters $\mathbf{a}$ (the vector of parameters) :

$$f(x; a) = \sum_r c_r(x) a_r \tag{7.8}$$

however complicated the functions $c_r(x)$ may be. For example given the two functions

$$f(x, \mathbf{a}) = a_1 + a_2 x^4 + a_3 \cos(x) + a_4 \exp(-x^4) \tag{7.9}$$

$$g(x, \mathbf{a}) = a_1 + \cos(a_2 + x) + a_3 x + a_4 x^3 \tag{7.10}$$

the function $g(x, \mathbf{a})$ is not linear in its parameters! We can express Eq.7.8 in matrix form. Let's consider the case of a function with 3 parameters

a_1, a_2, a_3. We have

$$C \cdot a = \begin{pmatrix} c_1(x_1) & c_2(x_1) & c_3(x_1) \\ c_1(x_2) & c_2(x_2) & c_3(x_2) \\ \vdots & \vdots & \vdots \\ c_1(x_N) & c_2(x_N) & c_3(x_N) \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{pmatrix}.$$

If V_y is the (known) covariance matrix for the measured values of y (assuming no error on x), we can therefore express χ^2 on matrix form:

$$\chi^2 = \left(\widetilde{y} - \widetilde{a}\widetilde{C}\right) V_y^{-1} (y - Ca)$$

which gives the following χ^2 estimator (also ML estimator) of the parameter vector a

$$\widehat{a} = \left(\widetilde{C}V(y)^{-1}C\right)^{-1} \widetilde{C}V_y^{-1} \cdot y = M \cdot y$$

with the notation that $\widetilde{C}$ is the transposed of C, in Matlab denoted C'. For the case above with 3 parameters and N data points we have:

$$M \cdot y = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1N} \\ m_{21} & m_{22} & \dots & m_{2N} \\ m_{31} & m_{32} & \dots & m_{3N} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \end{pmatrix} = \widehat{a}$$

This is a linear transformation of $y \to \widehat{a}$, and it is straight forward to find the G-matrix that converts the covariance matrix for y into a covariance matrix for $\widehat{a}$. This was calculated for any (locally) linear function of observed parameters (in this case: data points), and we therefore get:

$$G_{ki} = \frac{\partial a_k}{\partial y_i} = \frac{\partial}{\partial y_i} \sum_{j=1}^{N} m_{kj} y_j = m_{ki}$$

The G-matrix for the transformation is therefore the M-matrix described above, and the covariance matrix for the estimator becomes:

$$V(\widehat{a}) = M V_y \widetilde{M}$$

This expression can be simplified by using that for matrices A, B, we have: $\widetilde{AB} = \widetilde{B}\,\widetilde{A}$ and for any invertible square matrix A we have $I = \widetilde{I} = \widetilde{AA^{-1}1} = \widetilde{A^{-1}}\,\widetilde{A}$ which means that $\widetilde{A}^{-1} = \widetilde{A^{-1}}$. Specifically for a symmetric matrix we have: $\widetilde{V_y^{-1}} = \widetilde{V}^{-1} = V_y^{-1}$.

$$V(\widehat{a}) = \left(\widetilde{C}V_y^{-1}C\right)^{-1} \widetilde{C}\cancel{V_y^{-1}} \cdot \cancel{V_y} \cdot \widetilde{\left(\widetilde{C}V_y^{-1}C\right)^{-1}\widetilde{C}V_y^{-1}}\Big)$$

$$
\begin{aligned}
&= \left(\widetilde{C}V_y^{-1}C\right)^{-1}\widetilde{C}\widetilde{V_y^{-1}}C\,\widetilde{\left(\widetilde{C}V_y^{-1}C\right)^{-1}} \\
&= \cancel{\left(\widetilde{C}V_y^{-1}C\right)^{-1}}\cancel{\left(\widetilde{C}V_y^{-1}C\right)}\left(\widetilde{C}\widetilde{V_y^{-1}}\widetilde{C}\right)^{-1} \\
&= \left(\widetilde{C}V_y^{-1}C\right)^{-1}
\end{aligned}
$$

7.2.4 Non-linear least squared

When the function is non-linear in the parameters we have to minimise χ^2 numerically and estimate the errors (asymptotically) from:

$$\Delta\chi^2 = 1 \quad \text{or} \quad V(\widehat{a})^{-1} = \frac{1}{2}\frac{\partial^2\chi^2}{\partial a^2}.$$

For more than one parameter, the inverse of the covariance matrix is found as:

$$V_y^{-1}(a_i, a_j) = \frac{1}{2}\frac{\partial^2\chi^2}{\partial a_i \partial a_j}\bigg|_{a=\widehat{a}}$$

7.3 Binned counting data

7.3.1 χ^2 fitting of binned data

For binned counting data, each measurement (y_i) is Poisson distributed around the true (physical) function value $f_{true}(x_i)$. For a Poisson distribution we have that the mean and the variance are equal and we call them λ . In the limit of $\lambda \to \infty$ the distribution becomes Gaussian, it makes sense to approximate the likelihood function with that of a Gaussian distribution, and we therefore get:

$$\chi^2 \equiv \sum_i \frac{(y_i - f(x_i; a))^2}{f(x_i; a_{true})}$$

However, we do not know a_{true} and we must therefore choose one further approximation, which gives rise to Pearsson's and (modified) Neymann's χ^2 minimisation, where $f(x; a_{true})$ is replaced with the estimated function and the measured values respectively:

$$
\begin{aligned}
\chi^2_P &\equiv \sum_i \frac{(y_i - f(x_i; a))^2}{f(x_i; a)} \\
\chi^2_N &\equiv \sum_i \frac{(y_i - f(x_i; a))^2}{\max(y_i, 1)}
\end{aligned}
$$

Both work reasonably for large number of counts per bin ($y_i > 10$), but performs poorly when a significant part of the bins have lower count numbers

than that. That I refer to the latter as "modified" Neymann only refers to the fact that cases where $y_i = 0$ are explicitly taken into account, which is unimportant for high-statistics data where it's use is best justified. However, note that

- With Neymann χ^2 you artificially assign too low a weight to data points y_i that are higher than the true value for that bin, $f(x; a_{true})$, and too high a weight to data points that fall below the true value. This means Neymann fitting will *systematically* underestimate the function f.
- On the other hand, since with Pearson's χ^2 we weigh each point with the *estimated* function, we can decrease the value of χ^2 slightly, by artificially increasing the function slightly relative to what would have been the minimum had we been able to use the physical (true) function instead.

If you look at this effect as a function of the average count number per bin, you see the following [Hauschild]. In fact, for $f > 10$ the picture simplifies slightly, with the biases converging to $1/2$ for Pearsson's and -1 for Neymann's. For $f < 10$ the behaviour is complicated, as you see that the bias is dependent on the number of counts in the different regimes, and for Pearsson's can even change sign. This means the bias in two regions of the fit may be different. This is something you should consider very carefully when fitting counting spectra using χ^2.

7.3.2 Likelihood fitting for binned data

The correct maximum-likelihood estimator for fitting of a set of counting data instead uses the Poisson distribution rather than the Gaussian approximation:

$$\begin{aligned} L &= \prod_i P_{Poiss}(y_i; f_a(x_i)) \\ &= \prod_i \frac{(f_a(x_i))^{y_i}}{y_i!} \exp\left(-f_a(x_i)\right) \\ &\Rightarrow \\ -2\ln L &= 2\sum_i \left(f_a(x_i) - y_i \ln f_a(x_i) + \ln y_i!\right), \end{aligned}$$

and since the last term is a constant depending only on the data, it may be discarded during minimisation. This type of minimisation is (at least for a constant function) unbiased for counting data, and will therefore typically (for a general function) be less biased than χ^2_P and χ^2_N. There are cases, however, where also this estimator is significantly biased.

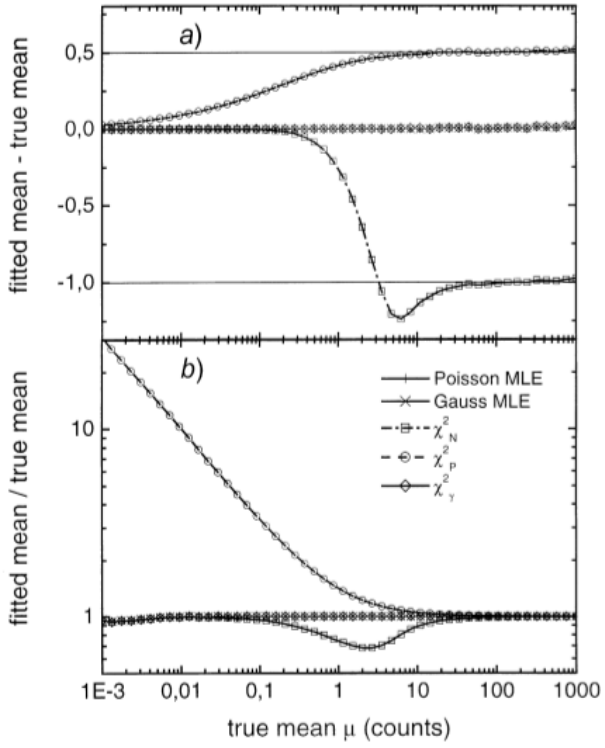

Fig. 1. Deviation of fitted mean from true mean value when using the five statistics under study. (a) For large μ Pearson's χ^2_P leads to an overestimation of 0.5 whereas Neyman's χ^2_N underestimates the true mean by -1. (b) Looking at the ratio of fitted and true mean, one clearly sees that using Pearson's χ^2_P, the fitted mean deviates extremely from the true one for very small μ.

Figure 7.1: Bias for χ^2 fitting [T. Hauschild and M. Jentschel, *Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments*, Nucl. Instrum. Methods A, 457:384, 2001]. See also course EARL resource list.

As discussed before, this function has similar properties to the χ^2 distribution, particularly that (asymptotically) $\pm\sigma$ can be found from:

$$\Delta(-2\ln L) = 1$$

because asymptotically, $-2\ln L$ behaves like a χ^2 distribution, that is Δ-values of 1, 2^2, and 3^2 give us the 1-σ, 2-σ, and 3-σ levels respectively:

```
octave-3.2.3:1> chi2cdf(1,1)
ans = 0.68269
octave-3.2.3:2> chi2cdf(4,1)
ans = 0.95450
octave-3.2.3:3> chi2cdf(9,1)
ans = 0.99730
```

This can be seen by evaluating the $\chi^2(x;n)$ distribution for $n = 1$. Equivalently, the curvature of $-2\ln L$ around the minimum can be used as discussed for the general ML-estimator (when correlations are manageable).

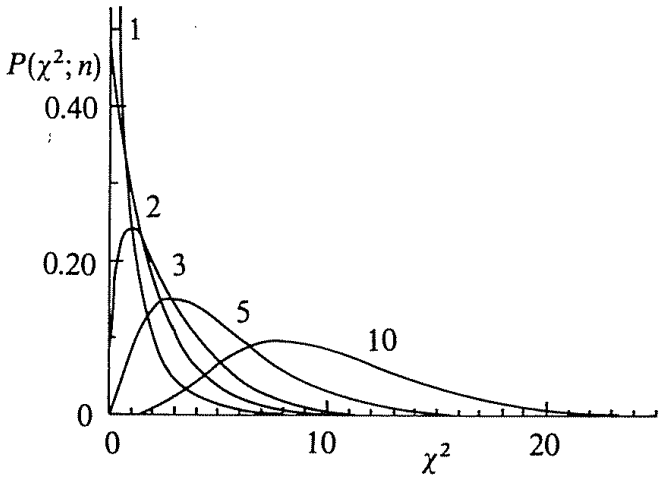

Figure 7.2: Some χ^2 distributions, for $n_{\rm deg} = 1$–10, [RB:fig6.4:p107].

7.4 Evaluating goodness-of-fit with χ^2

On the other hand, the χ^2 function is useful for other purposes, since it follows a well-known distribution depending only on the so-called *number-of-degrees-of-freedom* in the fit. This is given by: $n_{\rm deg} = N_{\rm data} - n_{\rm par}$ (n in the following). The χ^2 distribution is shown in for some $n_{\rm deg}$ between 1 and 10. Furthermore, the distribution converges to a Gaussian (yes also χ^2) as $n \to \infty$. The known Gaussian distribution can therefore be used to evaluate how good the fit is. However, the convergence of $\sqrt{2\chi^2}$, is much faster and should therefore be used instead. It converges to a Gaussian of $\mu = \sqrt{2n-1}$ and $V = 1$.

Example: for fitting of a binned data set of 45 bins to a parabolic function, the fit function has three parameters, that is: $n = 45 - 3 = 42$. If the best fit gave $\chi^2 = 73$, we therefore have: $\sqrt{2\chi^2} = 12.1$ and $\sqrt{2n-1} = 9.1$, and χ^2 is therefore 3σ away from the expected centroid of the function. The

probability that this happens by chance (unlikely data set) is only:

$$1 - CDF_{Gauss}(x = 3, \mu = 0, \sigma = 1) = 1 - 0.9987 = 0.0013.$$

That is, the probability that this would happen by mere chance is only 0.13%, and we must conclude that the data is not well described by the fitted function. If on the other hand, the fit yielded $\chi^2 = 51$ we would have $\sqrt{2\chi^2} = 10.1$ which is only one σ above the centroid of the Gaussian. This will happen by chance in 16% of data sets, and the fit should therefore not be excluded.

As opposed to χ^2 fitting, log-likelihood fitting for binned data does not directly give a goodness-of-fit. To do this, we should instead of $-2\ln L$ calculate what is known as the likelihood-χ^2:

$$\chi^2_\lambda = -2\ln L(\{y_i; f_a(x_i)\}) + 2\ln L(\{y_i; f_{optimal}(x_i)\}),$$

where $f_{optimal}$ is a function that is allowed to vary freely, without constraints on the functional form. The optimal values of $f(x_i)$ are exactly the y_i values, and we therefore get:

$$\begin{aligned}
\chi^2_\lambda &= -2\ln L(\{y_i; f_a(x_i)\}) + 2\ln L(\{y_i; y_i\}) \\
&= 2\sum_i \left(f_a(x_i) - y_i \ln f_a(x_i) - y_i + y_i \ln y_i\right) \\
&= 2\sum_i \left(f_a(x_i) - y_i \left(1 + \ln \frac{f_a(x_i)}{\max(1, y_i)}\right)\right)
\end{aligned}$$

In the limit of large data sets, χ^2_λ also follows a χ^2 distribution for the number of degrees of freedom: $n_{\rm deg} = N_{\rm data} - n_{\rm par}$. However, the correspondence is not exact, and you have to use alternative techniques if you want a precise goodness-of-fit evaluation.

Chapter 8

Probability and Confidence

Confidence levels appear as a part of descriptive statistics, as a way to describe the spread of a distribution

8.1 Probability and measurement

For a specific measurement of the electron mass m_e (the physical, true value) with a Gaussian error $\sigma = 10\,\text{keV}$ we write the result as:

$$m = 520 \pm 10\,\text{keV},$$

which means that the probability distribution when we performed the measurement was:

$$P(m; m_e) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(m - m_e)^2/2\sigma^2)$$

However, what in fact we wanted to say something about was actually: $P(m_e; m)$. For which we for example wanted to say that there is 68% probability that m_e lies in the range $m \pm \sigma$:

$$P(m_e \in [m - \sigma, m + \sigma]) = 68\%$$

Strictly speaking (in the objective as well as the frequentist sense of the word) saying this is nonsense, since m_e is a physical parameter and therefore m_e has one and only one value, it either falls within the range or not. The problem is that m is the stochastic variable (and therefore follows a random distribution) m_e is not. Instead what we really mean when we say this is in fact:

$$P(m, \sigma \text{ are such that } m_e \in [m - \sigma, m + \sigma]) = 68\%$$

That is, what we should aim to define limits $m \pm \sigma$ such that in 68% of repeated experiments the interval obtained from the experimental evaluation will contain the true value m_e. We therefore have to be very careful about

whether we talk about the probability distribution for the measurement, or the probability that the limits derived from the experiment include the true (physical) parameter value.

8.1.1 Probability distribution for the measurement

In the following, it will be important to distinguish between measured values and physical values, so in in the following, lower case will be used for measurements (x) and upper case for the physical values (X). So, still looking at the probability distribution for the measurement x (or m as above), we can define confidence regions for the measurement for any confidence level C such that:

$$P(x_- \leq x \leq x_+) = \int_{x_-}^{x_+} P(x)dx = C$$

as shown in fig. 8.1. For any given value of C, this definition allows us to

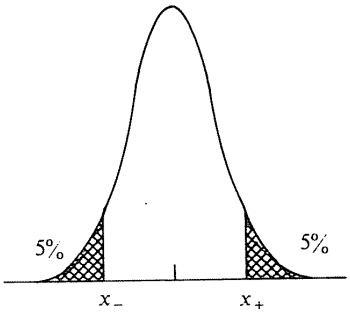

Fig. 7.1. The 90% central confidence interval for a Gaussian distribution.

Figure 8.1:

move the region around within the peak, as we wish, as long as the area within the region is kept constant. There are five such (commonly used) conventions for the confidence regions:

1. *Symmetric:* $x_+ - \mu = \mu - x_-$ (μ is the expectation value for x).
2. *Shortest:* given C, $[x_-, x_+]$ is chosen such that $x_+ - x_-$ is minimal.
3. *Central:* identical area tails, $\int_{-\infty}^{x_-} P(x)dx = \int_{x_+}^{\infty} P(x)dx = (1 - C)/2$.
4. *One-tailed upper limit:* $C = P(x \leq x_+) = \int_{-\infty}^{x_+} P(x)dx$.
5. *One-tailed lower limit:* $C = P(x_- \leq x) = \int_{x_-}^{\infty} P(x)dx$.

For the Gaussian distribution, 1–3 are all the same region. Commonly used confidence levels are (with the two-tailed central confidence region):

- 68.3%: 1-σ
- 90.0%: 1.64-σ
- 95.0%: 1.96-σ
- 95.4%: 2-σ
- 99.0%: 2.58-σ

and with the one-tailed confidence regions:

- 90.0%: one tail from 1.28-σ outwards (for Gaussian).
- 95.0%: one tail of the central two-tailed 90% confidence interval.
- $2.9 \cdot 10^{-7}$: one tail, 5-σ (required to claim discovery of a new particle)

8.1.2 Confidence intervals in estimation

Suppose we want know the value of a parameter X, and have estimated it from data, giving a result x. We know the variance $V(x)$ of our measurment. How to use this information to assess a confidence interval for X?

The naive aswer would be X lies within $x - \sigma$ and $x + \sigma$ at 68%.

Counter-Example

The weight of an empty dish is 25.3±0.14 g. A sample of powder is placed on the dish and the combined weigth measured as 25.50±0.14 g. By subtraction and combining the errors the weight of the powder is 0.2±0.2 g.

If we look at probabilities, using the naive interpretation, we have that 32% of probability of being outside the limits and 16% of probability of having a negative weight!!!

In reality for each value of x , there is a probability distribution $P(x; X_{true})$

So, how can we turn the probability distribution for the measurement:

$$P_{\text{meas}}(x; X_{\text{true}})$$

around, to get:

$$[x_-, x_+] \text{ such that } P(x_- \leq X_{\text{true}} \leq x_+) \geq C$$

Independent of the actual measurement of x, we can for any (assumed) true value of the

This is reported for example in Fig .8.2, for every value of parameter θ we build the dristribution $P(t, \Theta)$ and we build the two lines whown in figure corresponding (for example) to 90% cofidence level. The corresponding area is called confidence belt. In Fig.8.3 we have explicityl drawn to distributions (a Gaussian, Poisson,...) for two values of the true parameter X. This is done before the measurment is done.

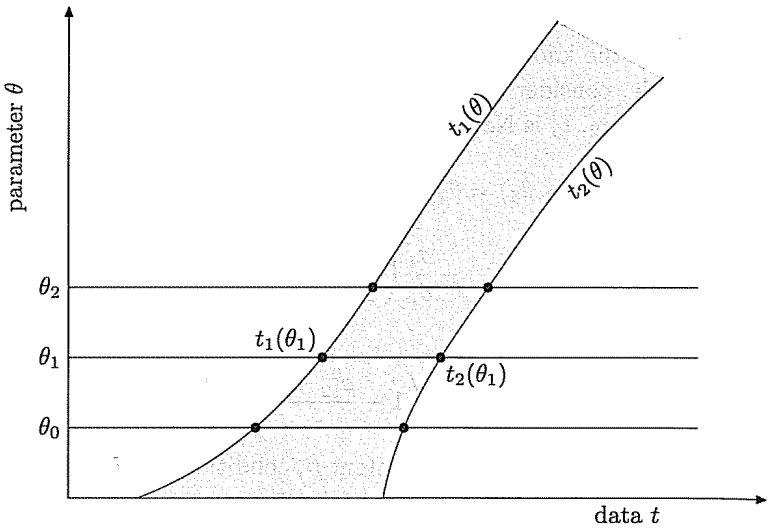

Figure 8.2:

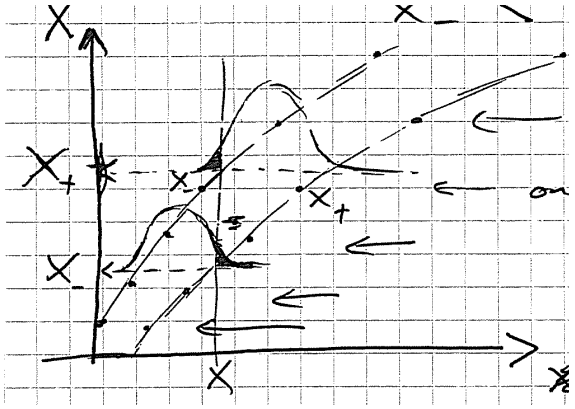

Figure 8.3: Confidence

Once the measurment is done we report is on the horizontal axis and we can draw a vertical line to extract X_- and X_+ intervals for X.

The confidence region for the measurement x can then be found from the areas under the tails of the distribution, giving us two points in the diagram. When this is done for all values of X, we obtain the two lines shown in the figure.

When we then measure a specific value x_m, we can then obtain the lower and upper limits for X (X_-, X_+) by reading off the diagram vertically as follows:

X_- is the value of X for which x_m is the upper limit of x.

X_+ is the value of X for which x_m is the lower limit of x.

As an example, let's say we want the $C = 90\%$ central confidence region for X. We then have:

X_- is the lowest possible value of X for which the chance of measuring x as high as (or higher than) x_m is at most 5%.

X_+ is the highest possible value of X for which the chance of measuring x as low as (or lower than) x_m is at most 5%.

That is, if the physical value X_{true}, is larger than X_+ then the probability of getting a measurement as low as the measured value is (at most) 5%. This means in at most 5% of such experiments will we find an upper limit X_+ such that the true value X_{true} is higher than X_+. Or equivalently: only in at most 5% of the experiments will we find an upper limit X_+ that is lower than the true, physical, value of the parameter. Similarly for X_-, only in (at most) 5% of experiments will we find a value X_- that is higher than the true value X_{true}. In conclusion, at least 90% of repeated experiments will the deduced values X_- and X_+, be such that $X_- \leq X_{\text{true}} \leq X_+$, and we therefore have:

$$P(X_- \leq X_{\text{true}} \leq X_+) \geq 90\%.$$

8.2 Examples of confidence regions

8.2.1 Gaussian confidence limits

For a measurement x_m with Gaussian error σ, we will find the central confidence region $[X_-, X_+]$ for a confidence level of C. X_- is the (true) value of X for which the probability for measuring a value as high (or higher than) x_m by chance is (at most) $(1 - C)/2$:

$$\int_{x_m}^{\infty} P(x'; X_-)dx' \leq (1 - C)/2$$

and X_+ is the (true) value of X, for which the probability of measuring a value as low (or lower than) x_m is (at most) $(1 - C)/2$.

$$\int_{-\infty}^{x_m} P(x'; X_+)dx' \leq (1 - C)/2$$

which for the Gaussian distribution is (for the lower limit X_-):

$$(1 - C)/2 \quad = \quad \int_{x_m}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x' - X_-)^2/2\sigma^2)dx'$$

$$= \int_{-\infty}^{X_-} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x' - x_m)^2/2\sigma^2)dx'$$

because the Gaussian probability distribution has two properties that are very convenient in this context: it is symmetric and the shape is independent of the centroid value, as can be seen in fig. 8.4. This means that, for

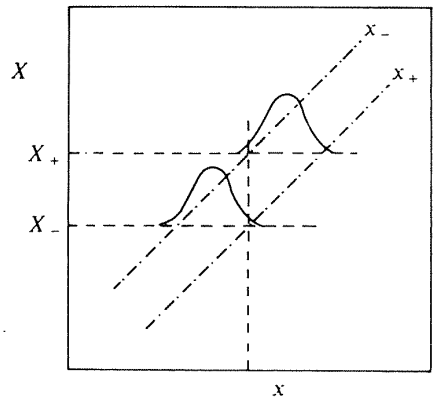

Fig. 7.3. The confidence diagram for a Gaussian.

Figure 8.4: Gaussian confidence band (for constant uncertainty) for measured and physical values [Barlow, fig 7.3].

example, for $C = 90\%$ the integral should be 0.05, which is the case when $X_- = x_m - 1.64\sigma$, exactly the same as the corresponding region for the measurement. However, this is only because of the two properties: *symmetry* and *independence of parameter value.*

8.2.2 Poisson confidence limits

The confidence limits for the Poisson distribution, on the other hand, are more involved. This is because the Poisson distribution lacks exactly those two properties that proved so convenient for the Gaussian distribution: the probability distribution for Poisson-distributed data is *asymmetric* and the shape of the distribution *depends on the value of the centroid.* Additionally, the Poisson distribution is a discrete distribution, which as you will see also complicates matters slightly.

For a measurement n_m drawn from a Poisson distribution, the limits on the true (physical) value N can then be found as follows. The central confidence region C_c is found by evaluating the upper and lower limits at a one-tailed confidence level $C_{o-t} = 1 - (1 - C_c)/2 = 1/2 + C_c/2$. The upper limit N_+ is then the value of N for which the probability of measuring r as

low as (or lower than) the measured n_m is (at most) $1 - C_{o.t.}$. That is:

$$P(n \leq n_m; N_+) = \sum_{r=0}^{n_m} P(r; N_+) \leq 1 - C_{o.t.} = (1 - C_c)/2.$$

You should note that the sum *includes* the measurement n_m, since we are calculating the probability for measuring a value lower than or equal to the measured value. Similarly, the lower limit N_- is found as the value of N for which the probability of measuring r as high as (or higher than) n_m is (at most) $1 - C_{o.t.}$. That is:

$$P(n \geq n_m; N_-) = \sum_{r=n_m}^{\infty} P(r; N_-) \leq 1 - C_{o.t.} = (1 - C_c)/2,$$

or equivalently:

$$\sum_{r=0}^{n_m - 1} P(r; N_-) \geq C_{o.t.} = 1 - (1 - C_c)/2,$$

since in practice it is always most convenient to calculate the finite sum starting from zero. Note, however, the difference in the summations here (up to $n_m - 1$ instead of n_m). This is because when finding N_-, we are evaluating the probability of achieving a count number higher than or equal to the measured n_m. For the finite sum from zero n_m is therefore not included in this case.

Specific values of N_+ and N_- for measured count numbers n_m from 0 to 10, given for one-tailed 90%, 95%, and 99% limits. For $n_m = 5$, we for example have: $N_+ = 10.51$, and $N_- = 1.97$ (one-tailed), so that with 90% confidence we can state that the true value N and the limits N_-, N_+ obey: $N_- \leq N \leq N_+$, as illustrated in figure 8.6. Note that because a 91% confidence region also obeys the requirements for a 90% confidence region, this means the slightly larger region $1.9 \leq N \leq 10.6$ is also a valid 90% confidence region, whereas $2.0 \leq N \leq 10.5$ is technically not a 90% confidence region. When we round off the values for N_- and N_+ to an appropriate number of significant figures we should therefore round outwards, enlarging the region rather than rounding off to the nearest value.

Signal with background

For cases where the measured signal n_m results from the sum $N^{\text{tot}} = N^{\text{sig}} + N^{\text{bgr}}$ of a physical signal and a background, the arguments above still hold for the total, as the sum of two Poisson-distributed stochastic variables is itself a Poisson distributed stochastic variable with an expectation value equal to the sum of the two expectation values. This is shown [in RB:3.3.3]

TABLE 7.1.
SOME POISSON LIMITS

	Upper			Lower		
	90%	95%	99%	90%	95%	99%
$n=0$	2.30	3.00	4.61	—	—	—
1	3.89	4.74	6.64	0.11	0.05	0.01
2	5.32	6.30	8.41	0.53	0.36	0.15
3	6.68	7.75	10.05	1.10	0.82	0.44
4	7.99	9.15	11.60	1.74	1.37	0.82
5	9.27	10.51	13.11	2.43	1.97	1.28
6	10.53	11.84	14.57	3.15	2.61	1.79
7	11.77	13.15	16.00	3.89	3.29	2.33
8	12.99	14.43	17.40	4.66	3.98	2.91
9	14.21	15.71	18.78	5.43	4.70	3.51
10	15.41	16.96	20.14	6.22	5.43	4.13

Figure 8.5: One-tailed confidence limits (90%, 95%, and 99%) for the Poisson distribution for a measured number of counts n.

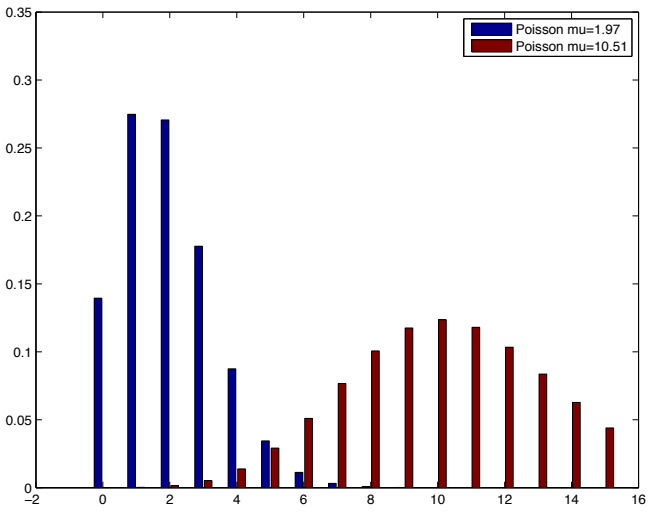

Figure 8.6: Poisson probability distributions for the measured number of counts n for two Poisson distributions with mean values $N = 1.97$ and $N = 10.51$ respectively. The first has an upper-tail integral (sum) from $n = 5$ and up of 5% and the second has a lower-tail integral (up to and including $n = 5$) of 5%. The 95% one-tailed confidence regions or 90% two tailed region for N with $n_m = 5$ measured counts is therefore: $N_- = 1.97$ and $N_+ = 10.51$.

by calculating for two event types (say a and b):

$$\begin{aligned} P(r)_{\lambda_a,\lambda_b} &= \sum_{r_a}^{r} P(r_a;\lambda_a)P(r-r_a;\lambda_b) \\ &= \cdots = \frac{1}{r!}e^{-(\lambda_a+\lambda_b)}(\lambda_a+\lambda_b)^r \\ &= P(r;\lambda_a+\lambda_b). \end{aligned}$$

If we assume the (average) background level to be known exactly, we can then using the methods above find the limits on N^{tot}: N_+^{tot}, N_-^{tot}, and we therefore obtain limits on N^{sig} as:

$$\begin{aligned} N_+^{\mathrm{sig}} &= N_+^{\mathrm{tot}} - N^{\mathrm{bgr}} \\ N_-^{\mathrm{sig}} &= N_-^{\mathrm{tot}} - N^{\mathrm{bgr}} \end{aligned}$$

So if $N^{bgr} = 0.45$ in the experiment above, we get:

$$\begin{array}{rcrcrcr} N_+^{\mathrm{sig}} & = & 10.51 - 0.45 & = & 10.06 & = & 10.1 \\ N_-^{\mathrm{sig}} & = & 1.97 - 0.45 & = & 1.52 & = & 1.5 \end{array}$$

where again we have rounded off to the 0.1 level by rounding outwards.

8.2.3 Two-dimensional confidence regions

Very often, when we wish to display confidence regions for parameters, we want to know the appropriate regions for each parameter independently, as done in the preceding. That is, that we can state with a confidence of C that a given region overlaps with true value the physical parameter. This can be seen in fig. 8.7 is the $\pm\sigma$ confidence band for parameter 2 given independently of the value of parameter 1 (the total horizontal band). Similarly if you imagine the vertical lines continued to $\pm\infty$, this would correspond to the confidence band for parameter 1 given independently of parameter 2.

For two-dimensional confidence regions (for parameters θ_1, θ_2), the question stated is slightly different, namely: give a region in the two-dimensional parameter space for which you can state with a confidence C that the region contains the true physical parameter pair: $\theta_{1,0}, \theta_{2,0}$. The ellipsis seen in the figure signifies the points in 2D parameter space at which the χ^2 function (or $-2\ln L$ if you wish) has increased from it's minimum by one (less inside, more outside). As you can see, this shows that for parameter 1, at $\pm\sigma$ there is a value of parameter 2 at which the change in χ^2 is as low as one. However it is also clear from the drawing that the confidence region defined by the ellipsis must then be less than the 68% corresponding to the full horizontal (or vertical) band. Instead, the confidence region defined by this region is found by evaluating the χ^2 distribution for two degrees of freedom, and we therefore find that this is the 39.3% confidence region:

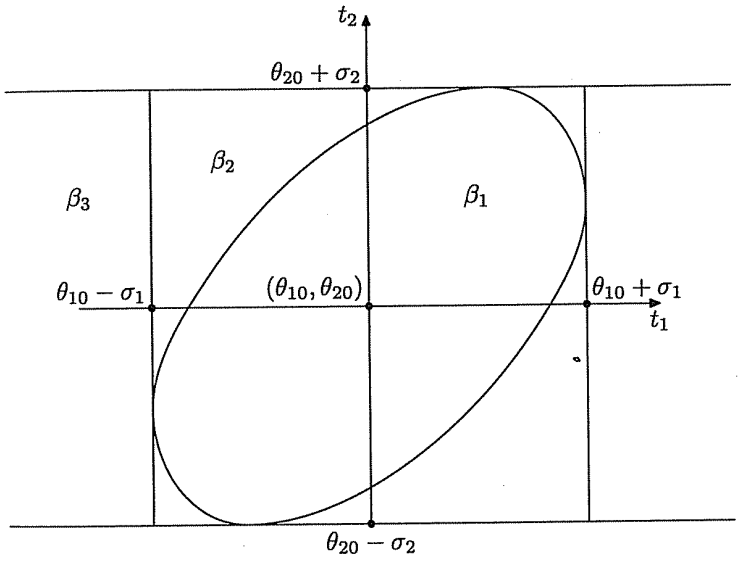

Figure 8.7: For two-dimensional confidence regions (parameters θ_1, θ_2). One-dimensional $\pm\sigma$ band as well as 2D 39% region are shown.

```
octave-3.2.3:1> chi2cdf(1,2)
ans = 0.39347
octave-3.2.3:2> chi2cdf(1,1)
ans = 0.68269
octave-3.2.3:3> chi2cdf(2.30,2)
ans = 0.68336
octave-3.2.3:4> chi2cdf(4.61,2)
ans = 0.90024
```

On the other hand, the 68.3% or 90% confidence regions, are found from an increase in χ^2 of 2.30 and 4.61 respectively.

As with one-dimensional regions, we have the additional complication of the difference between confidence regions for the measurement and that of the true parameter. This is indicated in fig. 8.8. However, as before, when the errors are Gaussian, there is no practical difference between the two evaluations. For most symmetric errors, the translation is therefore trivial.

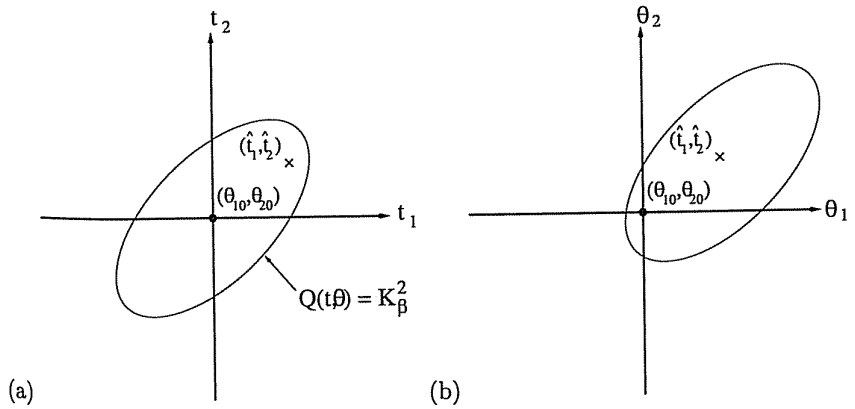

Figure 8.8: Translation between measurement space (a) and true parameter space (b) for 2D confidence regions when errors are Gaussian.

Chapter 8

Hypothesis testing

Sometimes the information you want from the data is not a number, but the yes-or-no answer to a factual question. For the answer to be simple, the question posed must be precisely specified. This is done by expressing it as an assertion that some *hypothesis* is true. You then construct a *test* and apply it to the data and the hypothesis is accepted/rejected depending on the result of the test.

Consider as examples therefore the hypothesis:

- The data are drawn from a Poisson distribution of mean larger than 3.4.

It is then (most often) important to consider the hypothesis in comparison to the alternative hypothesis. However, what is the alternative?

- The data are drawn from a Poisson distribution of mean $\lambda < 3.4$.
- The data are not drawn from a Poisson distribution.

You therefore have to consider what the alternative hypothesis is from (careful) investigation of the physics case in question, and make sure you clarify this by formulating the hypothesis very precisely. For example, it would be very difficult to distinguish between a Poisson distribution of $\lambda = 16$ and a Gaussian distribution with $\mu = 16$ and $\sigma = 4$. Deciding (and describing explicitly) the alternative hypothesis is therefore critical.

8.0.1 Errors in tests

When the answer to a hypothesis ($\mathcal{H}$) is yes/no, there are two ways of making mistakes, denoted type-I and type-II errors:

Type-I: Rejecting a hypothesis that in reality was true, simply because you were unlucky in getting an atypical dataset because of statistical fluctuations.

Type-II: Accepting a wrong hypothesis, that is, accepting $\mathcal{H}$ when in fact the alternative $\mathcal{A}$ was true.

Which of the two types of errors is most problematic naturally depends on the nature of the decision, whether the decision is to accept the existence of a new particle, to approve the use of a medical drug, or something entirely different.

For a test based on accepting the hypothesis if the value of a parameter x is lower than a cutoff value x_{cut}, as shown in figure 8.1a, the probability of a Type-I error is quantified by the *significance* of the test:

$$\alpha = \int_{x_{cut}}^{\infty} P(x|\mathcal{H})dx,$$

where $\alpha \leq 5\%$ means the test is significant at the 5% level. Naturally, this can be generalised to tests where the decision is based on a set of parameters, by calculating the corresponding multidimensional integral outside the acceptance region.

Similarly, the probability of Type-II errors can be quantified by the β-value for the test, defined by the integral over the probability distribution for the parameter x assuming the alternative hypothesis $\mathcal{A}$ is true:

$$\beta = \int_{-\infty}^{x_{cut}} P(x|\mathcal{A})dx,$$

as shown in figure 8.1b. In practice, however, we instead quantify this by the *power*, $1-\beta$, of the test. That is, the power describes how good the test is at not allowing false acceptances to sneak in. This means a good test has a low value for the significance and a large power. However, it is typically such that the smaller you make α (significant at the 10%, 5%, 1% level), the larger you make β and therefore the more likely you are to accept a false alternative hypothesis. This is the trade-off we have to decide on.

Barlow gives a very important note of caution here: the significance α (which describes our ability to accept $\mathcal{H}$ when it is correct) is often very well described mathematically because $\mathcal{H}$ is well defined. $\mathcal{A}$ on the other hand may well be more vaguely defined such that the power $(1-\beta)$ is less well-defined. However, the power is very often more important, in that it is the hardest of the two to optimise. As an example of this, look at the following test, with a significance level of, for example, $\alpha = 0.05$:

> When given your data to test $\mathcal{H}$ against $\mathcal{A}$: first generate a random number x from a uniform distribution [0,1]; then reject $\mathcal{H}$ if $x < \alpha$.

This test is significant at the 5% level. However, the power $(1-\beta)$ is also only 5%, and the test is therefore not useful. It is the combination of a low significance level and a high power that together makes a good test.

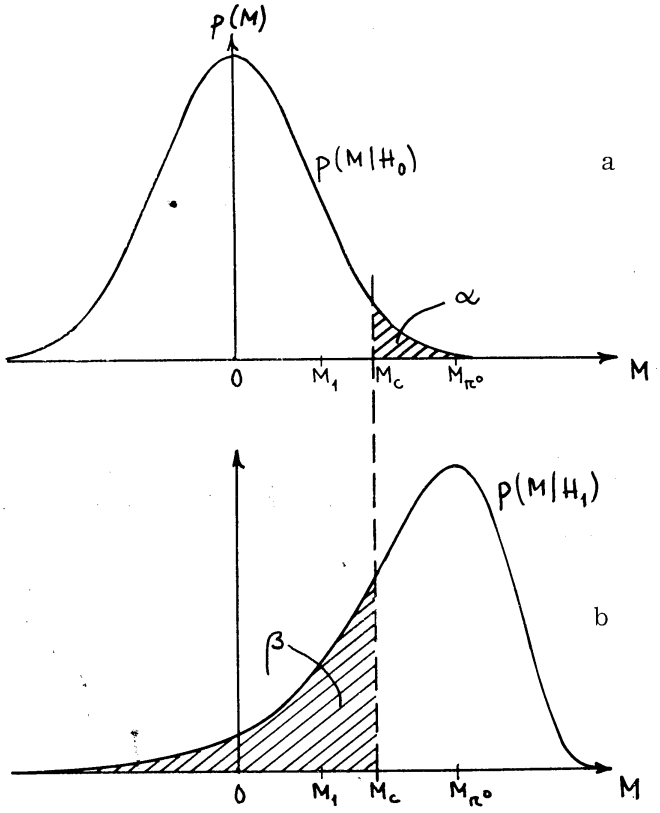

Figure 8.1: Significance (α) and power ($1-\beta$) of tests. $\mathcal{H}_0$ is the hypothesis under question, with $\mathcal{H}_1$ the alternative hypothesis.

8.0.2 Two well-defined hypothesis

Optional item: In a case where we know $P(x|\mathcal{H})$ and $P(x|\mathcal{A})$ as a function of a parameter x, and given one event (measurement of x) wish to determine whether to place it in group $\mathcal{H}$ or $\mathcal{A}$. Let us require that the test should have a given (small) significance of α:

$$\alpha = \int_{\text{accepted}} P(x|\mathcal{H})dx.$$

We should then optimise the test such that the power is as large as possible while still complying with the requirement set by α. That is to minimise β:

$$\beta = \int_{\text{accepted}} P(x|\mathcal{A})dx.$$

This can be done by choosing a value c and rejecting the hypothesis $\mathcal{H}$ for

all regions where

$$\frac{P(x|\mathcal{A})}{P(x|\mathcal{H})} > c,$$

and match c to give us the required significance level α. In practice, we would adjust c such that we get a α, $1-\beta$ tradeoff we can live with. As an example of this, Barlow describes how to use a measured density for samples from a mine to determine whether the sample is a valuable opal or an invaluable (or less valuable) quartz crystal.

8.1 The null-hypothesis, hypothesis testing in practice

Often, however, in experimental applications of hypothesis testing, the experiment aims at discovering an effect that is different from what is generally accepted, such as:

- Assymmetry of an angular distribution.
- Time-dependence of the fine-structure constant.
- Deviations from the expected signal as function of a spatial or energy variation.

In such cases, it doesn't prove anything that we can show that our measurement is *consistent* with, say, a time-dependent fine-structure constant. Only if we can prove that the measurement is *inconsistent* with a time-independent fine-structure constant have we proven that it is time-dependent.

Therefore, what we set out to do is to disprove the alternative hypothesis, the so-called "null-hypothesis" ($\mathcal{H}$) claiming that the effect we are looking for is not present. If we can disprove this, we will have proved that some effect other than the accepted is present. Whether it is the effect we were suggesting or some other effect is another matter.

8.1.1 A simple example: the biased coin

Optional item: Let us say you are suspecting your friends coin to be biased in favour of heads and want to test whether it is indeed biased by throwing the coin 15 times. We then assume the null-hypothesis:

$\mathcal{H}_0$: The coin has probability $p \leq \frac{1}{2}$ for heads.

This can technically be seen as a whole range of hypothesis with p anywhere from zero to $p = \frac{1}{2}$. The latter, however, is the hardest of these cases to discriminate against, and it is therefore sufficient to reject this hypothesis. (If, with a large number of observed heads, $p = \frac{1}{2}$ is rejected at a significance

level of α then so is all values below $p = \frac{1}{2}$.) With that hypothesis, the probability for (randomly) getting a number of heads h in the range of 10 to 15 is:

$$\begin{array}{lll} \mathrm{P}(h = 15) = 0.003\% & , & \mathrm{P}(h = 12) = 1.4\% \\ \mathrm{P}(h = 14) = 0.05\% & , & \mathrm{P}(h = 11) = 4.2\% \\ \mathrm{P}(h = 13) = 0.3\% & , & \mathrm{P}(h = 10) = 9.2\% \end{array}$$

such that the summed probabilities are:

$$\begin{array}{l} \mathrm{P}(h \geq 12) = 1.7\% \\ \mathrm{P}(h \geq 11) = 5.9\% \\ \mathrm{P}(h \geq 10) = 15.1\% \end{array}$$

So to our rejection level should be $h \geq 12$ if we wish to reject the null-hypothesis at a (one-tailed) significance level of 5%. In other words, the critical value for a 5% significance level with 15 observations in total is 12. Similarly, the critical value is 11 if we accept to work at a significance level of (only) 10%.

8.1.2 Example: testing batteries

A manifacturer claims its batteries last longer than 1000 hours. A random sample made of 900 batteries has $\mu = 980$h and $\sigma = 100$h. Is the statement of the manifacturer true correct to 5%?

- $\mathcal{H}_0$: the batteries last $\mu = 1000$ or $\mu \leq 1000$
- $\mathcal{H}_a$: the batteries last $\mu \geq 1000$

Since we want to test the hypothesis for the mean, having σ known and a large population (n¿30). We can assume that the mean is normally distributed. We use for this testing the $z$distribution. This is a Gaussian distribution with mean 0 and $\sigma = 1$.

For our test, we have to make the change of variable

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{8.1}$$

where $\bar{x}$ and σ are the mean and the variance of the sample. So in our case we have

$$z = \frac{980 - 1000}{\frac{100}{\sqrt{900}}} = -6 \tag{8.2}$$

Using tables, we find that to reject the null hypothesis $\mathcal{H}_0$ we should have $z > 1.95$ i.e. the area of the normal distribution from $-\infty$ to $z = 1.64$ is 95%. So in our case $z = -6$ is well below the rejection value, so we have to accept the $\mathcal{H}_0$ hypothesis.

8.1.3 Exmple: 2 tail distribution

Let's take a sample of students. I say *I think the average age of the student population is 30y.*

We thus have two hypothesis

- $\mathcal{H}_0$: the average age is 30 y
- $\mathcal{H}_a$: the average age is NOT 30 y.

We assume that the population is large and using (CLT) we assume it is thus distributed with a Gaussian with a $\sigma = \sqrt{20}$

We extract a small sample from thge population, say 10 and we find $\bar{x} = 27$ We use again the z distribution since we now that variance of the parent distribution. To have 5% of accuracy we need to find the limits of z containing 95% of the area of the distribution.

Remember that this case we have to deal with a 2 tail distribution, so since it symmetric we have to find a lower value z_l and upper value z_u taht defines the 95% of the area. From the bales we find $z_u = 1.96$ and $z_l = -1.96$.

We calculate the actual value of z with the numbers

$$z = \frac{27 - 30}{\sqrt{20}/\sqrt{10}} = -2.12 \tag{8.3}$$

$\rightarrow$ we have thus to reject the $\mathcal{H}_0$ hypothesis.

Now what happens if σ of the parent distribution is not known? So we derive standard deviation directly from the sample and let assume is $s = 3$. We can not use in this case the z distribution, we use the t distribution. The distribution is defined by the number of degrees of freedon of the sample which is given

$$n_{dof} = n_{sample} - 1 = 10 - 1 = 9 \tag{8.4}$$

If the sample size goes to infinity the t-distribution tends to the z distribution.

We calculate the t value

$$t = \frac{27 - 30}{s/\sqrt{10}} = -3.16 \tag{8.5}$$

From the tables we define again the two limit value (two tail distribution) t_u, t_l. From the tables we extract $t_u = 2.262$ and $t_l = -2.262$ that contain 95% of the distributions. So we reject $\mathcal{H}_0$ hypothesis even in this case.

8.1.4 Example: customer in a shop

I want to buy a shop and the owner claims that the distribution of the customer is given in Tab.8.1.4

Day of the week	Expected	Observed
M	10%	30
T	10%	14
W	15%	34
T	20%	45
F	30%	57
S	15%	20

We have 2 hypothesis

- $\mathcal{H}_0$ the owner distribution is correct to 5% confidence
- $\mathcal{H}_a$ is not correct

We use a χ^2 distribution to test the owner distribution. We have thus observed 200 customers. We transform the percentage in numbers in Tab.8.1.4

Day of the week	Expected
M	20
T	20
W	30
T	40
F	60
S	30

The χ^2 is defined using Pearson approximation

$$\chi^2 = \sum_i \frac{(\Theta_i - E_i)}{E_i} = 11.44 \tag{8.6}$$

here Θ_i is the observed and E_i the expected value. The number of degrees of freedom is $n_{dof} = 6 - 1 = 5$.

For our case to exclude the distribution at 5% we need to find the critical value (from tables) and we have $\chi = 11.07$ is the limit to have 95% confidence. Given the value we have we need to reject the hypothesis.

8.1.5 Example: testing a Poisson distribution

We want to test our guess about a Poisson distribution. We take the case of a call center monitored for 16h. We have recorded the number of call per h. See Tab.8.1.5.

number of calls per h	Frequency
0	53
1	62
2	33
3	12
≥ 3	0

We test the hypothesis

- $\mathcal{H}_0$ the distribution follow the Poisson distribution with mean 1 call per h.
- $\mathcal{H}_a$ it does not follow Poisson with mean 1.

Since we have a large number of counts per bin (n¿5) we can make the test using χ^2

$$\chi^2 = \sum_i \frac{(\Theta_i - E_i)}{E_i} \tag{8.7}$$

If the fit is good we expect χ^2 to be small. The expected values are calculated from the Poisson distribution $E_i = p_i N_{events}$ So we have

$$\begin{aligned} E_1 &= P(x=0)160 \\ E_2 &= P(x=1)160 \\ E_3 &= P(x=2)160 \\ E_4 &= P(x=3)160 \\ E_5 &= P(x>3)160 = (1 - P(x=0) - P(x=1) - P(x=2) - P(x=3))160 \end{aligned}$$

We get $\chi^2 = 4.71$ We compare the limit value of χ^2 with 5 d.o.f. at 95% accuracy. From tables we get $\chi_c = 9.49$ thus we accept our $\mathcal{H}_0$ hypothesis.

8.1.6 Goodness-of-fit using χ^2

As discussed previously, the χ^2 value can be used to check how good a fit is. For a fit $f(x)$ to a data set of n measurements $\{x_i, y_i, \sigma_i\}$, we have:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

where σ_i may either be known directly from the measurement: $\sigma_i = \sigma$ (constant), $\sigma_i = \alpha \cdot y_i$ (fixed ratio errors), $\sigma_i = \sqrt{y_i}$ (binned data, Neymann), or may be derived from the fit function: $\sigma_i = \sqrt{f(x_i)}$ (binned data, Pearsson).

CRITICAL χ^2 VALUES

	$P = 10\%$	$= 5\%$	$= 2\%$	$= 1\%$
$n = 1$	2.71	3.84	5.41	6.63
2	4.61	5.99	7.82	9.21
3	6.25	7.82	9.84	11.34
4	7.78	9.49	11.67	13.28
5	9.24	11.07	13.39	15.09
6	10.64	12.59	15.03	16.81
7	12.02	14.07	16.62	18.47
8	13.36	15.51	18.17	20.09
9	14.68	16.92	19.68	21.67
10	15.99	18.31	21.16	23.21
11	17.27	19.68	22.62	24.72
12	18.55	21.03	24.05	26.22
13	19.81	22.36	25.47	27.69
14	21.06	23.68	26.87	29.14
15	22.31	25.00	28.26	30.58
16	23.54	26.30	29.63	32.00
17	24.77	27.59	31.00	33.41
18	25.99	28.87	32.35	34.81
19	27.20	30.14	33.69	36.19
20	28.41	31.41	35.02	37.57
21	29.62	32.67	36.34	38.93
22	30.81	33.92	37.66	40.29
23	32.01	35.17	38.97	41.64
24	33.20	36.42	40.27	42.98
25	34.38	37.65	41.57	44.31
26	35.56	38.89	42.86	45.64
27	36.74	40.11	44.14	46.96
28	37.92	41.34	45.42	48.28
29	39.09	42.56	46.69	49.59
30	40.26	43.77	47.96	50.89

Figure 8.2: Critical values for the χ^2 distributions with $n = 1 \ldots 30$ at significance levels of 1–10%. R.J. Barlow, table 8.1, p. 151.

For $n = 5$ degrees of freedom ($n = n_{bin} - n_{par}$), if we get $\chi^2 = 14.2$, we see from Barlow's table on χ^2 significance values on p. 151 (figure 8.2). That is:

$$\int_{13.39}^{\infty} P(\chi^2; n = 5) d\chi^2 = 0.02$$
$$\int_{15.09}^{\infty} P(\chi^2; n = 5) d\chi^2 = 0.01$$

and we therefore know that the hypothesis:

$\mathcal{H}_0$: $f(x)$ describes the data

is rejected at the 2% significance level. At higher values of the number of

degrees of freedom: $n \geq 30$, the Gaussian approximation:

$$P_{\sqrt{2\chi^2}}(x, n) \approx P_{Gauss}(x, \sqrt{2n-1}, 1)$$

works well.

8.1.7 Goodness-of-fit using the log-likelihood ratio

For low-statistics counting-data $\{x_i, n_i\}$, as we have seen, neither χ^2_P nor χ^2_N worked well for the estimation, as both were biased even for a constant fit function. We therefore instead minimised the log-likelihood function:

$$-2\ln L = 2\sum_i f(x_i) - n_i \ln f(x_i) + \ln n_i!.$$

This, however, did not give any form of goodness-of-fit. However, we may instead define the maximum-likelihood ratio:

$$\lambda = \frac{\max_f L\left(data|f(x)\right)}{L\left(data|optimal\ model\right)}.$$

The optimal model to describe the data m is in this case obviously the model defined such that: $m(x_i) = n_i$. This means we have:

$$\begin{aligned} -2\ln\lambda &= 2\sum_i \left(f(x_i) - n_i \ln f(x_i) + \cancel{\ln n_i!}\right) - 2\sum_i \left(n_i - n_i \ln n_i + \cancel{\ln n_i!}\right) \\ &= 2\sum_i \left(\left(f(x_i) - n_i\right) - n_i \ln \frac{f(x_i)}{n_i} \right). \end{aligned}$$

This function is referred to as the likelihood-χ^2, since asymptotically (for large data sets) it converges to a chi^2 distribution, that is the significance levels are identical (asymptotically) to that of the χ^2 distribution, and we denote it:

$$\chi^2_\lambda = -2\ln\lambda$$

Furthermore, since χ^2_λ is as good as $-2\ln L$ for parameter estimation (differing only by a constant), we may as well use this version, the maximum-likelihood ratio, for parameter estimation also. We thereby have a combined method for parameter estimation and goodness-of-fit for binned data.

There is, however, one note of caution to make: χ^2_λ is biased for low count numbers ($n_i < 10$) as seen in figure 8.3. The best procedure for low count numbers would therefore be to use χ^2_λ for parameter estimation and χ^2_P for goodness-of-fit. Alternatively, the bias in χ^2_λ must be corrected for on a channel-by channel basis.

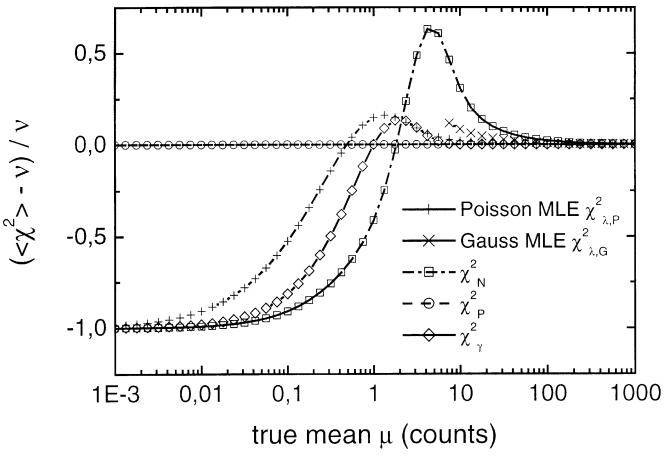

Figure 8.3: χ^2_λ per degree of freedom for low statistics, Hauschild-2001

Number of degrees of freedom, Matlab v.s. Barlow

Note, when you use a goodness of fit test, that Matlab and Barlow use different convensions for the number of fit parameters, and therefore use a different expression for the number of degrees of freedom (though the exact expression is well-hidden in the case of Matlab). Specifically, Barlow uses: $ndeg = nbin - npar$, whereas Matlab uses $ndeg = nbin - npar - 1$, as seen for example in the `chi2gof()` routine. Which of the two is correct hinges on what you count as a parameter, as you will see in the following:

Barlow always counts all parameters (for an exponential function $f(t) = a\exp(-t/\tau)$, for example, we have two parameters: a and τ). We then have $ndeg = nbin - npar = 10 - 2 = 8$ for a data set with 8 data points (bins). This is the basis for Barlow's calculation of the degrees of freedom. In Matlab on the other hand, as exemplified in the `chi2gof()` routine, the absolute scaling is not counted as a free parameter (that is, it is not included in *nparams* as given below). This, however, does not change the fact that the absolute scale also reduces the number of degrees of freedom in the same way as other parameters do. If we therefore only count tau as a parameter in the experimental distribution (or lambda in a Poisson distribution as below), we must instead use the convention: $ndeg = nbin - npar - 1$. Below is calculated the χ^2 g.o.f. for a set of counts over 6 bins. The data, $c(i)$, for each of the bins is the number of intervals with that count number, $b(i)$, such that for example 16 measurements have obtained exactly 1 count. The claim, as in exercise 8.2 [R.J. Barlow], is that the data follows a Poisson distribution, and therefore that the number of events as a function of bin number follows a Poisson probability distribution: $P(b) = P_{Poiss}(b, \hat{\lambda})$, scaled by the total number of events. Additionally, the Matlab

`chi2gof()` routine approximates the error on each bin by the square-root of the expected number of counts, $e(i)$, and therefore uses $\chi^2 = \sum((c(i) - e(i))^2/e(i))$. The fitted parameters below are effectively the mean count number per interval ($\hat{\lambda}$) giving the shape of the Poisson distribution, and the scale (n) defining the total area under the function.

```
b = 0:5
c = [6 16 10 12 4 2]
n = sum(c)
lambdaHat = sum(b.*c)/n
e = n*poisspdf(b,lambdaHat)
[h,p,st] = chi2gof(b,'ctrs',b,'frequency',c,...
                   'expected',e,'nparams',1)
% result:  %
% st =
% chi2stat:  2.5550
% df:  3
% edges:  [-0.5000 0.5000 1.5000 2.5000 3.5000 5.5000]
% O: [6 16 10 12 6]
% E: [7.0429 13.8041 13.5280 8.8383 6.0284]
```

In the result, we first note that the last two bins have been combined. Matlab's `chi2gof()` does this automatically to ensure that the count numbers are at least above 5. Second, we note that from the 5 bins remaining after the automatic re-binning, we have $ndeg = 3$ degrees of freedom (here written as `df`, accessed as `st.df`). This in spite of having quoted only one free parameter for the function ('nparams',1). We therefore see that the scaling parameter has been taken into account automatically in the calculation of the number of degrees of freedom (the -1 in ndeg = nbin-npar-1). Therefore, in summary: though both can be used correctly, the two conventions must be kept separate to avoid inconsistencies. Furthermore, as you may already have noted, I would recommend that you use Barlow's ndeg = nbin-npar but I should emphasise that because of the potential misunderstandings, you will always have to be explicit about your definition of npar and ndeg.

8.1.8 The Run Test

Optional item: As an alternative to the χ^2 test we can instead (or in addition) use the run-test. This test is only sensitive to the sign of $y_i - f(x_i)$, which is ignored entirely in the χ^2 test. An example of the run-test for a straight-line fit is shown in figure 8.4. The number of runs is then found by assigning one of two values to each data point: above (A) or below (B), depending on whether the data point is above or below the fit function. For the straight-line fit shown in the figure, the data set is described as: AAABBBBBBAAA. This means we have tree runs in the present case. The

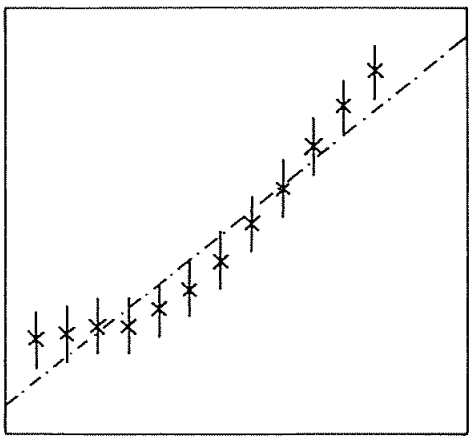

Figure 8.4: A straight-line fit through 12 data points, Barlow figure 8.3. The run-test evaluates how (un)likely it is to only have three runs (as is the case here) in a 12-point data set.

expectation value and variance of the number of runs, $\langle r \rangle$ and $V(r)$ are given in the book (p. 154). For the present example, the probability of getting at most 3 runs assuming 6 points above the fitting function and 6 points below, we get:

$$\begin{aligned} P(r \leq 3 | N_A = N_B = 6) &= P(r = 3 | N_A = N_B = 6) \\ &\quad + P(r = 2 | N_A = N_B = 6) \\ &= \frac{10}{924} + \frac{2}{924} = \frac{1}{77} \\ &= 1.299\%. \end{aligned}$$

This means even though the χ^2 value in the present example is in fact reasonable (you can see that about a third of the data points fall outside 1-σ), $\mathcal{H}_0$ can still be rejected at the 1.3% significance level (or 2% significance if you want integer percentages). For comparison, the Gaussian approximation as carried out by Barlow gives $P(r \leq 3 | N_A = N_B = 6) = 0.82\%$, indicating a significance level of 1%.

In general the run-test is not as strong as the χ^2 test, but as it provides *independent* information, it can still be useful:

$$P_{total} = P_{\chi^2} \times P_{run}.$$

Particularly, in cases where the χ^2 test accepts the hypothesis and the run-test rejects it this tells us something very specific about the data (ask): *the uncertainties are probably either overestimated or correlated*, which is why the χ^2 test does not reject the hypothesis.

8.2 Kolmogorov-Smirnov test(s)

The Kolmogorov test (or Kolmogorow-Smirnov test, as it is also known) is particularly good for low-statistics data. It is best used for un-binned data, and can therefore be used without the information-loss introduced by binning of a data set. The fundamental idea behind this test is to compare the cumulative distribution function (CDF) of the data set to the expected function. On the other hand, the tests we have looked at so far have compared the distribution of counts to what is in essence a probability density function (PDF). The CDF is the integral over the PDF, as indicated in figure 8.5. Where the PDF ($f(x)$) is large, the CDF is steep ($F(x)$). $F(x)$ furthermore

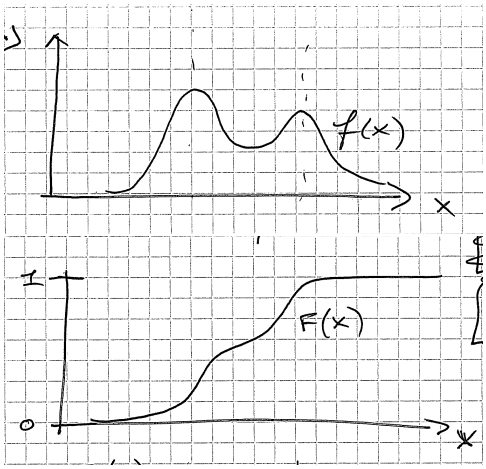

Figure 8.5: Probability density function (upper) and cumulative distribution function (lower).

is by definition normalised, so even if your function $f(x)$ has an area corresponding to the number of counts you have observed in your experiment, $F(x)$ must still be normalised such that it approaches 1 for $x \to \infty$.

8.2.1 One sample and a theoretical function

To perform the test for a data set with N x-values $\{x_i\}$, we first calculate the CDF ($F(x)$) for the theoretical function as shown in figure 8.6:

$$F(x) = \int_{-\infty}^{x} P(x')dx'.$$

This is compared to the empirical cumulative distribution function:

$$cum(x) = \sum_{x' \leq x} \frac{1}{N},$$

which starts at 0 for $x = -\infty$ and increase by $\frac{1}{N}$ for each data point as we proceed along the horizontal axis.

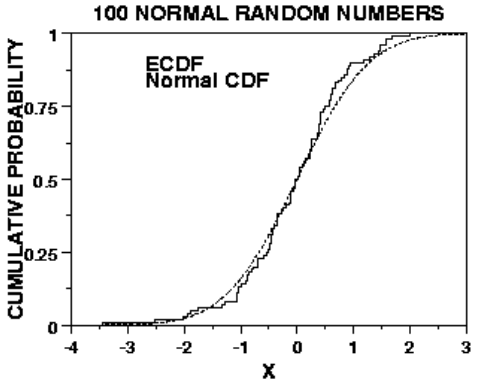

Figure 8.6: Kolmogorow-Smirnov test for Gaussian-distributed data, 100 data points. The continuous dashed line describes the theoretical Gaussian function of $\mu = 0$ and $\sigma = 1$. The discrete function (ECDF) describes the empirical CDF.

The test is then performed by evaluating the maximum difference between the two functions:

$$D = \max |cum(x) - F(x)|,$$

which because $cum(x)$ is discontinuous must be evaluated as:

$$D = \max_{1 \le i \le N} \left(F(x_i) - \frac{i-1}{N}, \frac{i}{N} - F(x) \right).$$

Naturally, the critical values for D will depend strongly on N for a given significance level (small relative variations for large number of data points). In fact, it scales exactly as $\frac{1}{\sqrt{N}}$, because the random variation in the number of counts below a given point, say $x = 0$ for example, scales as:

$$\sigma_D(x) = \frac{1}{\sqrt{N \int_{-\infty}^{x} P(x')dx'}}$$

and we therefore use the parameter:

$$d = D\sqrt{N}$$

for the evaluation. The critical values for this parameter (d) for significance levels of 1-20% are given in table 8.1. Note, however, that for these critical values to hold the function can not have been fitted to the distribution. The only "free parameter" in the comparison is therefore the total number

Table 8.1: Critical values of d for significance levels of the Kolmogorov-Smirnov test in the 1-20% range.

Critical value of d	Significance level
1.63	1%
1.36	5%
1.22	10%
1.07	20%

of events in the data set, N. If any parameter related to the shape of the distribution has been adjusted in the comparison of the data to the theoretical function, the critical values do not hold. This effect is exactly the same as the dependency of the critical values for the χ^2 distribution on the number of degrees of freedom, and because of that dependent on the number of parameters in the fit, as discussed in the previous section.

8.2.2 Two-sample comparison

Optional item: Similarly, for two data sets, $\{x_i\}$ and $\{y_j\}$, with N_x and N_y data points each. If these are proposed to arise from the same distribution, we can calculate the differences:

$$D = \max |cum(x) - F(x)|,$$

and

$$d = D\sqrt{\frac{N_x N_y}{N_x + N_y}}.$$

The critical values for d are then the same as those shown in table 8.1.

8.3 Other comparative tests

Optional item: More specific methods for the comparison of two data samples also exist, when only specific information about the distributions is required, and some information about the distributions is known. Barlow describes examples of such tests when we wish to know whether the two samples have the same:

Mean: for Gaussian distributions with known σ; for Gaussian distributions with unknown σ, as well as for the specific example of correlated samples.

Spread/Variance: using the F-distribution to evaluate whether the variances of two samples are significantly different.

Chapter 9

Bootstrap

Let suppose we have a sample extracted from a given population. Bootstrap is a method very efficient and simple to extract information and estimate errors directly from the sample. You want to ask a question of a population but you can't. So you take a sample and ask the question of it instead. Now, how confident you should be that the sample answer is close to the population answer obviously depends on the structure of population. One way you might learn about this is to take samples from the population again and again, ask them the question, and see how variable the sample answers tended to be. Since this isn't possible you can either make some assumptions about the shape of the population, or you can use the information in the sample you actually have to learn about it.

Imagine you decide to make assumptions, e.g. that it is Normal, or Bernoulli or some other convenient fiction. Following the previous strategy you could again learn about how much the answer to your question when asked of a sample might vary depending on which particular sample you happened to get by repeatedly generating samples of the same size as the one you have and asking them the same question. That would be straightforward to the extent that you chose computationally convenient assumptions. (Indeed particularly convenient assumptions plus non-trivial math may allow you to bypass the sampling part altogether, but we will deliberately ignore that here.)

This seems like a good idea provided you are happy to make the assumptions. Imagine you are not. An alternative is to take the sample you have and sample from it instead. You can do this because the sample you have is also a population, just a very small discrete one; it looks like the histogram of your data. Sampling 'with replacement' is just a convenient way to treat the sample like it's a population and to sample from it in a way that reflects its shape.

Let suppose that we have a statistical sample $\vec{x} = (x_1, x_2, \ldots, x_N)$. These numbers follow a distribution that is not known. Instead of guessing the

underlying distribution, *i.e.*

- Gaussain → continuos data
- Poisson → binned data

We can use the bootstrap method to obtain extra informations from our data. We assume that the sample is now the population and we resample it to get new distribution

$$\vec{x} \rightarrow \text{re-sampling} \rightarrow \vec{x}^* \tag{9.1}$$

We stress that $\vec{x}^*$ is not obtained by taking any extra new measurment but simply re-sampling (using random numbers) the previous set. For example we get

$$\vec{x}^* = (x_3, x_1, x_N, ...x_N, ...x_1) \tag{9.2}$$

Notice the new vector has the same length of the previous one, i.e. N. Repetitions of data points are allowed. I can apply an estimator $\hat{a}$ on the original sample $\vec{x}$, but also on the new one. Let suppose we repeat this operaton B times.

So from our original data sample $\vec{x}$ I draw B new samples $\vec{x}^B$ each of them having its own estimator $\hat{a}^B$. B should be large...$\approx 10^{3-4}$ to follow nicely a Gaussian. I can now plot the distribution of the estimators $\hat{a}$ and extract informations about their distribution, i.e. the variance. The distribution of the estimator is a Gaussian (CLT).

We define

$$\sigma_B^2 = \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{a}(\vec{x}) - \hat{a}_i(\vec{x}^*)\right)^2 \tag{9.3}$$

to get the error on the estimator used on our sample. For small values of re-sampling a Poisson distribution should be used instead, but is is slighlty more complicated.

It is worth noticing that the error on our estimate depend on

$$V(\sigma_B) \approx \frac{1}{N^2} + \frac{1}{NB} \tag{9.4}$$

We can only operate on B, so we can make it large enough so that the error induced by our bootstrapping is negligible compared to the one done by the original sampling size. An example of script

```
n = 700;
%lambda=5;
%x=normrnd(lambda,lambda,1,n)
x = randn(1,n);
%x = chi2rnd(10,1,n);
% Mean of x
mu = mean(x)
ss = sum((x-mu).^2)/(n-1)
B = 100000;
% Initialize muStar
muStar = zeros(B,1);
ssStar = zeros(B,1);
% Loop over B bootstraps
for b=1:B
% Uniform random numbers over 1...n
u = ceil(n*rand(n,1));
% x-star sample simulation
xStar = x(u);
% Mean of x-star
muStar(b) = mean(xStar);
ssStar(b)= sum((xStar-muStar(b)).^2)/(n-1);
end
figure(1)
xx = min(muStar):.01:max(muStar);
hist(muStar,xx)

figure(2)
yy = min(ssStar):.01:max(ssStar);
hist(ssStar,yy)

s2 = 1/(n-1)*sum((x-mu).^2)
stdErr = s2/n
bootstrapStdErr = sum((muStar-mu).^2)/(B-1)
bootstrapStdErr = sum((ssStar-ss).^2)/(B-1)
```

Chapter 10

Additional notes

10.1 Random number generators

All random number generators available to you will be pseudo-random, i.e. they will be deterministically determined but with good randomised properties and (typically) a long repetition chain, such that many independent random numbers can be generated. Which set of the chain of pseudo-random numbers is used is determined by the "seed" of the random number generator. The seed is decided either by direct/default input or by the time of your running of the simulation. Make sure you are aware of the properties of the generators available in your statistical package.

10.1.1 Direct (pseudo-)random generators

- Uniform distribution [0,1] is generally always available. However, many different qualities are available.
- Many distributions are often available. In general it is safer to use a well-tested implementation (Matlab, Root, or other) than to write your own.
- Sometimes, however, it may be necessary to write your own generator based on the supplied generators, primarily based on the uniform generators. This could include:
  - Random direction in 4-π: with x_1, x_2 as uniform [0,1] random numbers, calculate: $\phi = 2\pi \cdot x_1$ and $\theta = \text{acos}(2 \cdot x_2 - 1)$.
  - If you can calculate the cumulative distribution function, $F(x)$, of the distribution $f(x)$ you wish to simulate: because $F(x)$ is monotonic with $(-\infty, \infty) \to [0, 1]$, the inverse F^{-1} exists with $[0, 1] \to (-\infty, \infty)$. Random numbers x can then be generated based on the generation of one uniform variable [0,1], x_1, by calculating $x = F^{-1}(x_1)$.

10.1.2 Acceptance-reject (von Neumann sampling)

If you only have $f(x)$ available, but know that the function is effectively confined to a specific interval in x, you can use the Acceptance-reject method (also known as von Neumann sampling). This sampling method, as shown

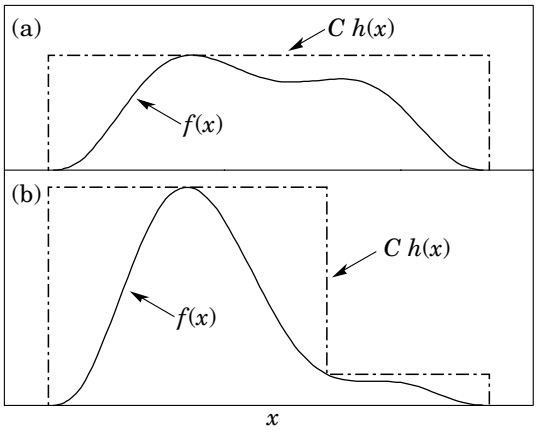

Figure 10.1: Von-Neumann sampling [Particle Data Group, Eidelman et al., 2004, see VLE-EARL].

in figure 10.1, is based on the observation that when we sample a positive function $f(x)$ confined to the region $[x_1, x_2]$, we are essentially sampling the area under the curve (for each point along the horizontal axis, the sampling rate scales with the value of the function. The sampling procedure is therefore:

1. Evaluate the confining region $[x_{min}, x_{max}]$ and the maximum value f_0 for the function $f(x)$.
2. While insufficient events have been generated do the following:
   (a) Generate two uniform random numbers $x, y \in (0, 1)$.
   (b) Scale $y' = y * f_0$, $x' = x_{min} + x * (x_{max} - x_{min})$.
   (c) If $y' \leq f(x')$ accept the event.
   (d) If $f(x') < y'$ reject the event.

If the maximum value f_0 for the function is unknown a priori, add after 2b the test: if $f(x') > f_0$, increase f_0 and restart the simulation from 2. The increase can for example be done by scaling $f_0^{\text{new}} = 2 \cdot f_0^{\text{old}}$. Following this procedure, you will at most be generating four times as many random numbers as you would have generated had you known the exact maximum of the function.

Variations of this exist for which the envelope (in the figure denoted $C \cdot h(x)$) is more complicated than a simple uniform distribution. These include for example cases where $h(x)$ is peaked or region-scaled Neumann sampling, where the simulation is effectively performed separately for different regions that are normalised appropriately to each other. This is very useful when the function $f(x)$ has large variations (for example strong peaks with long, shallow tails), since the efficiency of the generation (the number of accepted events relative to the number of uniform random numbers generated) scales as the ratio of the area under the function and the area of the enveloping function $C \cdot h(x)$. For the case shown in figure 10.1b, the gain by introducing the separate regions is only about 30%, but in other cases the gain can orders of magnitude.

The von Neumann sampling method can similarly be used for multi-dimensional distributions $f(x_1, x_2, \ldots, x_n)$ for which you generate the relevant $x_1, x_2, \ldots, x_n, y$ random numbers and accept the event if:

$$f(x_1, x_2, \ldots, x_n) \leq f_0 \cdot y.$$

In this case it is very often necessary to use a more complex enveloping function to obtain an effective sampling. Furthermore, you should note that to obtain an effective sampling of N in each of the n dimensions, you will need N^n accepted samples, which grows rapidly with the dimensionality n of the problem.

10.2 Computational errors

Optional item: The precision and stability of numerical calculations is not a central part of the present course. However, some notes of caution are worth emphasising:

10.2.1 Subtraction

Do not subtract numbers of the same magnitude:

$$1000024 - 1000003 = 1.0000 \cdot 10^{-6} - 1.0000 \cdot 10^{6} = 0$$

for five decimal-place numerical precision. This can for example give problems when calculating the variance of a data set:

$$V(x) = \overline{x^2} - \overline{x}^2$$

can be calculated in a single loop over the data, whereas the equivalent:

$$V(x) = \overline{(x - \overline{x})^2}$$

requires two loops. However, the first is prone to numerical errors when the data set has a combination of large mean $\overline{x}$ and large number of elements N, resulting in very large values of both $\sum x^2$ and $\sum x$ compared to the result, $V(x)$.

10.2.2 Addition

Do not add large and small numbers:

$$1000000 - 1.000000 = 1.0000 \cdot 10^6 - 1 = 1.0000 \cdot 10^6,$$

for six decimal-place numerical precision. Successive additions and subtractions will further enhance this problem. An example of this is in the evaluation of the zero-points of a quadratic equation ($y = ax^2 + bx + c$) through evaluation of the discriminant, this type of error will be dominant for cases where $4ac << b^2$:

$$\begin{aligned} D &= b^2 - 4ac \\ x &= \frac{-b \pm \sqrt{D}}{2a} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \end{aligned}$$

Alternatively, the solutions can be evaluated through:

$$q = \frac{1}{2}\left(b + \text{sign}(b)\sqrt{b^2 - 4ac}\right)$$

$$x_1 = q/a, \quad x_2 = c/q.$$

In general it is therefore worth considering carefully, how you implement a routine. Though it may be the same physics, the numerical errors of two different implementations may well prove significant.

10.2.3 Matrix inversion

This involves many successive additions and subtractions, some of which may well be of the type discussed above. There are therefore two rules you should use in this context:

1. Don't invert the matrix if you don't have to. For linear equations, for example, there are well-tested routines that are less prone to numerical errors

2. If you need the matrix inversion, test against numerical errors by multiplying the matrix and it's inverse after the calculation to test: $A \cdot A^{-1} = I$ and $A^{-1} \cdot A = I$.

10.2.4 Polynomial curve fitting

The standard implementation of an n^{th} order polynomial, $\sum a_n x^n$, is often not the best choice, since the resulting fit parameters may end up being highly correlated. Exactly which expression is best for the polynomials is case dependent. However, the Chebychev polynomials for example are orthogonal over $x \in [-1, 1]$, and therefore works well in this regime. The standard implementations of splines (local polynomial fitting) are similarly optimised for independence of parameters in local fits.

10.3 Key texts, available through VLE-EARL

Barlow, Roger (1989). Statistics, A guide to the use of statistical methods in the physical sciences. Wiley. This is the key textbook for the couse.

Introduction to Kolmogorov-Smirnov tests by the National Institute of Standards and Technology, NIST/SEMATECH e-Handbook of Statistical Methods [http://www.itl.nist.gov/div898/handbook/].

S. Baker and R. D. Cousins, 1984, Clarification of the use of chi-square and likelihood functions in fits to histograms, Nucl. Instrum. Methods, vol. 221, pp. 437-442. This is a good and detailed overview of the use of chi-square and likelihood functions in fits to histograms.

T. Hauschild and M. Jentschel, 2001, Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments, Nuclear Instruments and Methods in Physics Research A, vol. 457, pp. 384-401. This is a comparison of fitting methods for counting experiments (Maximum Likelihood, and three different Chi2 methods).

James, F (2006). Statistical methods in experimental physics. Hackensack, NJ ; World Scientific. This is particularly relevant in relation to confidence regions, hypothesis testing, and goodness-of-fit tests, in which this book is more detailed than the key-text by R.J. Barlow. First edition, published in 1971, is often referred to as Eadie's (Eadie, Drijard, James, Roos, and Sadoulet).

Particle Data Group, 2004, Mathematical tools or statistics, Monte Carlo, Group Theory, Physics Letters B, vol. 592, pp. 275-297. This is a 20-page review of relevant mathematical tools, by the Particle Data Group.

Y. Jading and K. Riisager, 1996, Systematic errors in Chi2-fitting of Poisson distributions, Nuclear Instruments and Methods in Physics Research Section A, vol. 372, pp. 289-292. This is a comparison of fitting methods for counting experiments (errors in Chi2 methods).

U.C. Bergmann and K. Riisager, 2002, A systematic error in maximum likelihood fitting, Nuclear Instruments and Methods in Physics Research Section A, vol. 489, pp. 444-447. This is about highly-correlated errors and the effect on the fitting of low-statistics data.