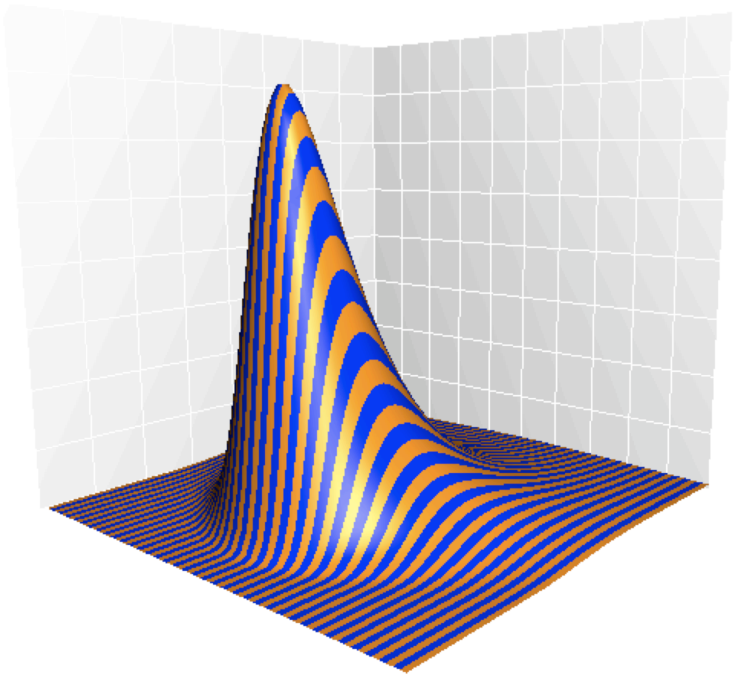
NIKHEF-2009-027

Introduction to Bayesian Inference

M. Botje

Nikhef, PO Box 41882, 1009DB Amsterdam, the Netherlands

October 20, 2009
(Last updated on December 15, 2013)



Abstract

In these lectures we cover—from a Bayesian perspective—the definition of probability, elementary probability calculus and assignment, selection of least informative probabilities by the maximum entropy principle, parameter estimation, systematic error propagation and model selection.

Lectures given at the BND School, Rathen, Germany, September 18–19, 2009, at the FANTOM International Research School, Groningen, the Netherlands, November 7–10, 2011, and at the Nikhef Topical Lectures series, December 11–13, 2013.

Contents

The cover plot shows the joint posterior distribution of the mean and width of a Gaussian distribution, estimated from four draws of this distribution assuming a uniform prior for the mean and a Jeffreys prior for the width, as is described in Section 6.1 of this write-up.

1 Introduction

The Frequentist and Bayesian approaches to statistics differ in the definition of probability. For a Frequentist, the probability of an event is the relative *frequency* of the occurrence of that event in an infinitely large set of repeated observations under identical conditions. Roughly speaking, probability is, in this view, taken to be a property of the world around us. Bayesian probability, on the other hand, is not defined as a frequency of occurrence but as the *plausibility* that a proposition is true, given the available information. Bayesian probability is thus not *per se* a property of the world around us, but more reflects our state of knowledge about that world. These different views have, as we will see, far-reaching consequences when it comes to data analysis since Bayesians can assign probabilities to propositions, or hypotheses, while Frequentists cannot.

In these lectures we present the basic principles and techniques underlying Bayesian statistics or, rather, Bayesian *inference*. Such inference is the process of determining the plausibility of a conclusion, or a set of conclusions, which we draw from the available data and prior information.

Since we derive in this write-up (almost) everything from scratch, little reference is made to the literature. So let us start by giving some useful references below:

A good introduction to Bayesian methods is given in the book by Sivia '*Data Analysis—a Bayesian Tutorial*' [Sivia06]. More extensive, with many worked-out examples in Mathematica, is the book by P. Gregory '*Bayesian Logical Data Analysis for the Physical Sciences*' [Greg05]. We also mention the monumental work by Jaynes, '*Probability Theory—The Logic of Science*' [Jay03] but this book is certainly not for the fainthearted. Unfortunately Jaynes died before the book was finished so that it is incomplete. It is available in print (Cambridge University Press) but a free (preliminary) copy can still be found on the website given in [Jay98]. For those who want to refresh their memory on Frequentist methods we recommend '*Statistical Data Analysis*' by G. Cowan [Cowan98] and '*Statistical Methods in Experimental Physics*' by F. James [James06].

Many references in these notes are from an interesting collection of papers that can be found on `http://www.astro.cornell.edu/staff/loredo/bayes`, and links therein. From these we mention the nice review (from an astronomers perspective) of T. Loredo '*From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics*' [Lor90]. To get a grasp of the basic ideas and their historical development, we recommend '*Bayesian Methods: General Background*' by E.T. Jaynes [Jay85]. A good summary of Bayesian methods from a particle physicist view can be found in the article '*Bayesian Inference in Processing Experimental Data*' by G. D'Agostini [Agost03]. Illuminating case studies are presented in '*An Introduction to Parameter Estimation using Bayesian Probability Theory*' [Bret90] and '*An Introduction to Model Selection using Probability Theory as Logic*' [Bret96] by G.L. Bretthorst.

Finally, there are of course these lecture notes which can be found, together with the lectures themselves, on `http://www.nikhef.nl/user/h24/bayes`.

> **Exercise 1.1**: Several of the works referred to above are written by astronomers. Can you give reasons why Bayesian methods tend to be more popular among astronomers than among particle physicists?

2 Bayesian Probability

Bayesian probability is a measure of the plausibility of a proposition. It can be viewed as a quantity that interpolates between 'true' and 'false' in case we do not have sufficient information to draw firm conclusions. In this section we will establish the link between logic and probability and develop probability calculus from a Bayesian viewpoint.

2.1 Plausible inference

In Aristotelian logic a **proposition** can be either true or false. In the following we will denote a proposition by a capital letter like A and represent 'true' or 'false' by the Boolean values 1 and 0, respectively. The operation of **negation** (denoted by $\overline{A}$ or, equivalently, by $\sim A$) turns a true proposition into a false one and *vice versa*.

Two propositions can be linked together to form a **compound proposition**. The state of such a compound proposition depends on the states of the two input propositions and on the way these are linked together. It is not difficult to see that there are exactly 16 different ways in which two propositions can be combined.[1] All these have specific names and symbols in formal logic but here we will be concerned with only a few of these, namely, the **tautology** ($\top$), the **contradiction** ($\bot$), the **and** ($\wedge$),[2] the **or** ($\vee$) and the **implication** 'if A then B' ($\Rightarrow$). The truth tables of these binary relations are

A	B	$A \top B$	$A \bot B$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$
0	0	1	0	0	0	1
0	1	1	0	0	1	1
1	0	1	0	0	1	0
1	1	1	0	1	1	1

(2.1)

Note that the tautology is always true and the contradiction always false, independent of the value of the input propositions.

> **Exercise 2.1**: Show that $A \wedge \overline{A}$ is a contradiction and $A \vee \overline{A}$ a tautology.

Two important relations between the logical operations 'and' and 'or' are given by the **de Morgan's laws**

$$\overline{A \wedge B} = \overline{A} \vee \overline{B} \qquad \text{and} \qquad \overline{A \vee B} = \overline{A} \wedge \overline{B}. \tag{2.2}$$

We note here a remarkable duality possessed by logical equations in that they can be transformed into other valid equations by interchanging the operations $\wedge$ and $\vee$.

> **Exercise 2.2**: Prove (2.2). Hint: this is easiest done by verifying that the truth tables of the left and right-hand sides of the equations are the same. Once it is shown that the first equation in (2.2) is valid, then duality guarantees that the second equation is also valid.

[1] Each input proposition can be true or false so that the two propositions define four possible input states. The compound proposition can thus be encoded in 4 bits by specifying the output bit (true or false) of each input state. Five out of the 16 possible output words are listed in the truth table (2.1).

[2] The conjunction $A \wedge B$ will often be written as the juxtaposition AB since it looks neat in long expressions or as (A, B) since we are accustomed to that in mathematical notation.

One may ask the question how many logical functions are necessary to generate all others. The answer (as every electronics engineer knows) is that only one function is sufficient, namely the 'nand' defined by

$$A \uparrow B \equiv \overline{A \wedge B} = \overline{A} \vee \overline{B}. \tag{2.3}$$

Exercise 2.3: Express $\overline{A}$, $A \wedge B$ and $A \vee B$ in terms of the 'nand' operator.

The reasoning process by which conclusions are drawn from a set of input propositions is called **inference**. If there is enough input information we apply **deductive inference** which allows us to draw firm conclusions, that is, the conclusion can be shown to be either true or false. Mathematical proofs, for instance, are based on deductive inferences. If there is not enough input information we apply **inductive inference** which does not allow us to draw a firm conclusion. The difference between deductive and inductive reasoning can be illustrated by the following simple example:

P1: Roses are red P2: This flower is a rose	$\rightarrow$	This flower is red (deduction)
P1: Roses are red P2: This flower is red	$\rightarrow$	This flower is perhaps a rose (induction)

Induction thus leaves us in a state of uncertainty about our conclusion. However, the statement that the flower is red increases the *probability* that we are dealing with a rose as can easily be seen from the fact that—provided all roses are red—the fraction of roses in the population of red flowers must be larger than that in the population of all flowers.

It was already known in the ancient world that deductive reasoning can be broken down into a chain of **strong syllogisms**,[3] the two types of which are

Major premise:	If A is true then B is true	If A is true then B is true
Minor premise:	A is true	B is false
Conclusion:	B is true	A is false

Inductive reasoning, on the other hand, contains one or more **weak syllogisms**

Major premise:	If A is true then B is true	If A is true then B is true
Minor premise:	A is false	B is true
Conclusion:	B is less probable	A is more probable

The first proposition in all four syllogisms above can be recognised as the implication $A \Rightarrow B$ for which the truth table is given in (2.1). It is straight forward to check from this truth table the validity of the conclusions given above.

Exercise 2.4: (i) Show that it follows from $A \Rightarrow B$ that $\overline{B} \Rightarrow \overline{A}$; (ii) Check that the conclusions of the above syllogisms are consistent with the truth table of implication as given in (2.1).

[3] A syllogism is a triplet of related propositions consisting of a *major premise*, a *minor premise* and a *conclusion*.

In inductive reasoning then, we are in a state of uncertainty about the validity (true or false) of the conclusion we wish to draw. This is nothing special because we all the time conduct, often intuitively, inductive reasoning to cope with questions such as: Shall I cross this road? Should I bring my raincoat? Can I trust this bank?

The steps taken in answering such questions are very succinctly phrased by Jaynes as follows [Jay85]: (i) Try to foresee all the possibilities that might arise; (ii) judge how likely each is, based on everything you can see and all your past experience; (iii) in the light of this, judge what the probable consequences of various actions would be; (iv) now make your decision. The last two steps belong to the field of **decision theory** which is not covered in these lectures. The first two steps belong to the field of **plausible inference**, that is, the art of reasoning in the presence of uncertainty.

The first step in formalising the inductive reasoning process is to define a measure $P(A|I)$ of the plausibility (or degree of belief) that a proposition A is true, given the information I. It may seem quite an arbitrary business to attempt to quantify something like a 'degree of belief' but this is not so.

Cox (1946) has, in a seminal paper [Cox46], formulated the rules of plausible inference and plausibility calculus by basing them on several *desiderata*. These desiderata are not axioms (they don't postulate true propositions) but a list of properties that a sensible measure of plausibility measure should possess. Here is how Jaynes formulates them [Jay03] (my wording):

(I) The measure of plausibility is a real number. This is because real numbers are continuous and transitive which means that if $a < b$ and $b < c$ then $a < c$. We need this because otherwise we cannot order propositions by degree of plausibility;

(II) Agreement with common sense. By this desideratum is meant that plausibility should increase continuously and monotonically when more supporting evidence for the truth of a proposition is supplied. At the same time, the plausibility of the negated proposition must decrease. It is also required that plausible inference must become Aristotelean logic (Boolean algebra) in the limit that we deal with propositions which are certain to be true or false on the evidence;

(III) Consistency.

- (a) The plausibility of a conclusion depends on the relevant information and not on the path of reasoning by which the conclusion is reached;
- (b) All available relevant information (and not some selection) should be taken into account while all irrelevant information should be ignored;
- (c) Equivalent states of knowledge must be represented by equal plausibility assignments.

It turns out that these desiderata are so restrictive that they *completely* determine the algebra of plausibility. To the surprise of many, this algebra appeared to be identical to that of classical probability as defined by the axioms of Kolmogorov (see below for these axioms). Plausibility is thus, for a Bayesian at least, identical to probability.[4] It

[4]The intimate connection between probability and logic is reflected in the title of Jaynes' book: '*Probability Theory—The Logic of Science*'.

is worth noting that the desiderata—which are the foundation of Bayesian probability theory—do not make any mention of random variables, frequencies, ensembles, or hypothetical repetitions of an experiment. The derivation of the Kolmogorov axioms from the desiderata is beyond the scope of these lectures; we refer to Loredo [Lor90] for a very clear derivation of the product rule. Full derivations (covering many pages) can be found in Cox [Cox61], Jaynes [Jay03] and Gregory [Greg05].[5]

Finally, we remark that the desiderata are built, from the start, into the theory so that they cannot be violated in properly conducted Bayesian inference. Later on in this write-up we will encounter examples where they are violated in Frequentist inference.

2.2 Probability calculus

We will now derive several useful formula, starting from the fundamental axioms of probability calculus and taking the viewpoint of a *Homo Bayesiensis*, when appropriate. As already mentioned above, $P(A|I)$ is a real number which, partially from the desiderata and partially by convention, is bounded to $P(A|I) = 1$ (0) when we are certain that the proposition A is true (false). The two **Kolmogorov axioms** that define probability calculus are the **sum rule**

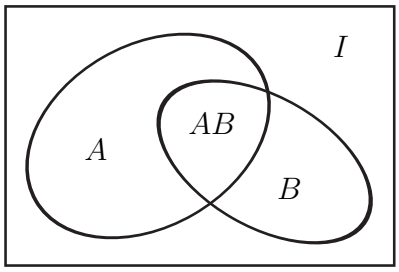
$$P(A \vee B|I) = P(A|I) + P(B|I) - P(AB|I) \tag{2.4}$$

and the **product rule**

$$P(AB|I) = P(A|BI)P(B|I). \tag{2.5}$$

Let us, at this point, spell-out the difference between AB ('A and B') and $A|B$ ('A given B'): In AB, B can be true or false while in $A|B$, B is assumed to be true and cannot be false. The following terminology is often used for the probabilities occurring in the product rule (2.5): $P(AB|I)$ is called the **joint probability**, $P(A|BI)$ the **conditional probability** and $P(B|I)$ the **marginal probability**.

Probabilities can be represented in a Venn diagram by the (normalised) areas of the sub-sets A and B of a given set I. The sum rule is then trivially understood from the following diagram.



The product rule (2.5) normalises the conjunction AB to the set B, instead of to I: $P(AB|I)$ corresponds to the area AB normalised to I, $P(A|BI)$ corresponds to AB normalised to B and $P(B|I)$ corresponds to B normalised to I.

[5] Cox employed a somewhat unorthodox mathematics where he did not start from *axioms*, but from *desiderata* which were cast into *functional equations*. The Kolmogorov axioms follow from solving these equations. Two years later, Shannon used similar methods in his foundation of communication theory.

Because $A \vee \overline{A}$ is a tautology (always true) and $A\overline{A}$ a contradiction (always false) we find from (2.4)

$$P(A|I) + P(\overline{A}|I) = 1 \tag{2.6}$$

which often is taken as an axiom instead of (2.4).[6]

Exercise 2.5: Derive the sum rule (2.4) from the axioms (2.5) and (2.6).

If A and B are **mutually exclusive propositions** (they cannot both be true) then, because AB is a contradiction, $P(AB|I) = 0$ and (2.4) becomes

$$P(A \vee B|I) = P(A|I) + P(B|I) \qquad (A \text{ and } B \text{ exclusive}). \tag{2.7}$$

If A and B are **independent** (the knowledge of B does not give us information on A and *vice versa*),[7] then $P(A|BI) = P(A|I)$ and (2.5) becomes

$$P(AB|I) = P(A|I)P(B|I) \qquad (A \text{ and } B \text{ independent}). \tag{2.8}$$

Because $AB = BA$ we see from (2.5) that $P(A|BI)P(B|I) = P(B|AI)P(A|I)$. From this we obtain the rule for **conditional probability inversion**, also known as **Bayes' theorem**:

$$P(H|DI) = \frac{P(D|HI)P(H|I)}{P(D|I)}. \tag{2.9}$$

In (2.9) we replaced A and B by D and H to indicate that in the following these propositions will refer to 'data' and 'hypothesis', respectively. From this we see that Bayes' theorem models a learning process in the sense that it specifies how to update the knowledge on H when new information D becomes available. In words, this updating process reads as follows: The probability $P(H|DI)$ of a hypothesis, given the data, is equal to the probability $P(H|I)$ of the hypothesis, given the background information alone (that is, without considering the data) multiplied by the probability $P(D|HI)$ that the hypothesis, when true, just yields that data. In Bayesian parlance $P(H|DI)$ is called the **posterior probability**, $P(D|HI)$ the **likelihood**, $P(H|I)$ the **prior probability** and $P(D|I)$ the **evidence**. Note that (2.9) only makes sense when the evidence $P(D|I)$ is non-zero.

Exercise 2.6: Investigate probability inversion

$$P(B|AI) = \frac{P(A|BI)P(B|I)}{P(A|I)}$$

in case the propositions A and B are (i) mutually exclusive, (ii) logically independent, (iii) both. You can take the marginal probabilities $P(A|I)$ and $P(B|I)$ to be both non-zero.

[6] Eq. (2.6) is then called the sum rule and (2.4) the 'extended sum rule'.

[7] We are talking here about a *logical* dependence which could be defined as follows: A and B are logically dependent when learning about A implies that we also will learn something about B. Note that logical dependence does not necessarily imply *causal* dependence. Causal dependence does, on the other hand, always imply logical dependence.

We remark that Bayes' theorem is valid in both the Bayesian and Frequentist worlds because it follows directly from axiom (2.5) of probability calculus. What differs is the *interpretation* of probability: for a Bayesian, probability is a measure of plausibility so that it makes perfect sense to convert $P(D|HI)$ into $P(H|DI)$ for data D and hypothesis H. For a Frequentist, on the other hand, probabilities are properties of *random variables* and, although it makes sense to talk about $P(D|HI)$, it does not make sense to talk about $P(H|DI)$ because a hypothesis H is a proposition and not a random variable. More on Bayesian versus Frequentist in Section 2.5.

Not being aware of the consequences of probability inversion easily leads to flawed reasoning. To see this, consider the case of Mr. White who goes to a doctor for an AIDS test. This test is known to be 100% efficient (the test never fails to detect AIDS). A few days later poor Mr. White learns that he is positive. Does this mean that he has AIDS? Most people (including, perhaps, Mr. White himself and his doctor) would say 'yes' because they fail to realise that, in general,

$$P(\text{positive}|\text{AIDS}) \neq P(\text{AIDS}|\text{positive}).$$

Here are two more examples that should make the point clear:

$$P(\text{rain}|\text{clouds}) \neq P(\text{clouds}|\text{rain}), \qquad P(\text{woman}|\text{pregnant}) \neq P(\text{pregnant}|\text{woman}).$$

Right?

In the next section we will learn how to deal with Mr. White's test (and with the opinion of his doctor).

2.3 Exhaustive and exclusive sets of hypotheses

Let us now consider the important case that H can be expanded into an exhaustive set of mutually exclusive hypotheses $\{H_i\}$, that is, into a set of which one and only one hypothesis is true.[8] Note that this implies, by definition, that H itself is a tautology. Trivial properties of such a complete set of hypotheses are[9]

$$P(H_i, H_j|I) = P(H_i|I)\,\delta_{ij} \tag{2.10}$$

and

$$\sum_i P(H_i|I) = P(\bigvee_i H_i|I) = 1 \qquad \text{(normalisation)} \tag{2.11}$$

where we used the sum rule (2.7) in the first equality and the fact that the logical sum of the H_i is a tautology in the second equality. Eq. (2.11) is the extension of the sum-rule axiom (2.6) and is called the **normalisation condition**.

Similarly it is straight forward to show that

$$\sum_i P(D, H_i|I) = P(D, \bigvee_i H_i|I) = P(D|I). \tag{2.12}$$

[8] A trivial example is the complete set $H_1 : x < a$ and $H_2 : x \geq a$ with x and a real numbers.

[9] Here and in the following we write the conjunction AB as A, B.

This operation is called **marginalisation**[10] and plays a very important role in Bayesian analysis since it allows us to eliminate sets of hypotheses which are necessary in the formulation of a problem but are otherwise of no interest ('nuisance parameters').

The inverse of marginalisation is the **expansion** of a probability: Using the product rule we can re-write (2.12) in reverse order as

$$P(D|I) = \sum_i P(D, H_i|I) = \sum_i P(D|H_i, I)P(H_i|I) \tag{2.13}$$

which states that the probability of D can be written as the weighted sum of the probabilities of a complete set of hypotheses $\{H_i\}$. The weights are just given by the probability that H_i, when true, gives D. In this way we have expanded $P(D|I)$ on a basis of probabilities $P(H_i|I)$.[11] Expansion is often used in **probability assignment** because it allows us to express a compound probability in terms of known elementary probabilities.

Using (2.13), Bayes' theorem (2.9) can, for a complete set of hypotheses, be written as

$$P(H_i|D, I) = \frac{P(D|H_i, I)P(H_i|I)}{\sum_i P(D|H_i, I)P(H_i|I)}, \tag{2.14}$$

from which it is seen that the denominator is just a normalisation constant.

If we calculate with (2.14) the posteriors for all the hypotheses H_i in the set, we obtain a spectrum of probabilities which, in the continuum limit, goes over to a probability density distribution (see Section 2.4). Note that in computing this spectrum the term $P(D|H_i, I)$ is taken to be a function of the hypotheses for fixed data. It is then called a **likelihood function**; note that this is *not* a probability. On the other hand, if $P(D|H_i, I)$ is regarded as a function of the data for fixed hypothesis it is not called a likelihood but, instead, a **sampling probability**.

> **Exercise 2.7**: Mr. White is positive on an AIDS test. The probability of a positive test is 98% for a person who has AIDS (efficiency) and 3% for a person who has not (false-positive). Given that a fraction $\mu = 1\%$ of the population is infected, what is the probability that Mr. White has AIDS? What would be this probability for full efficiency and for zero false-positives? Note that Bayesian probabilities are by no means fixed since they can change when new information becomes available. For instance, suppose that two months after the test a more thorough investigation of the population reveals that $\mu = 0.1\%$, instead of 1%. What is now the probability that Mr. White has AIDS?

> **Exercise 2.8**: What would be the probability that Mr. White has AIDS given the prior information $\mu = 0$ (nobody has AIDS) or $\mu = 1$ (everybody has AIDS)? Note that both these statements on μ encode prior *certainties*. Convince yourself that, according to Bayes' theorem, no amount of data can ever change a prior certainty.

Up to now we have explicitly kept the probabilities conditional to 'I' in all expressions as a reminder that Bayesian probabilities are always defined in relation to some background

[10] A projection of a two-dimensional distribution $f(x, y)$ on the x or y axis is called a *marginal* distribution. Because (2.12) is projecting out $P(D|I)$ it is called marginalisation.

[11] Note that (2.13) is similar to the closure relation in quantum mechanics $\langle D|I\rangle = \sum_i \langle D|H_i\rangle\langle H_i|I\rangle$.

information. This does not encompass 'all that is known' but, instead, all background information known to us that is relevant for our inference. In the following we will be a bit liberal and sometimes omit 'I' when it clutters the notation.

It is very important to realise that this background information must be the *same* for all probabilities in a given expression; if this is not the case, calculations may lead to paradoxical results.

> **Exercise 2.9**: Suppose we use Bayes' theorem $P(H|DI) \propto P(D|HI)P(H|I)$ to update the prior probability $P(H|I)$ with our data D. Now consider the following: take $P(H|DI)$ as a better estimate of the prior and use Bayes' theorem again to improve on posterior!
>
> Right?

2.4 Continuous variables

The formalism presented above describes the probability calculus of propositions or, equivalently, of discrete variables (which can be thought of as an index labelling a set of propositions). To extend this discrete algebra to continuous variables, consider the propositions

$$A : r < a, \qquad B : r < b, \qquad C : a \leq r < b$$

for a real variable r and two fixed real numbers a and b with $a < b$. Because we have the Boolean relation $B = A \vee C$ and because A and C are mutually exclusive we find from the sum rule (2.7)

$$P(a \leq r < b|I) = P(r < b|I) - P(r < a|I) \equiv G(b) - G(a). \tag{2.15}$$

In (2.15) we have introduced the **cumulative distribution** $G(x) \equiv P(r < x|I)$ which obviously is a monotonically increasing function of x. The **probability density** p is defined by

$$p(x|I) = \lim_{\delta \to 0} \frac{P(x \leq r < x + \delta|I)}{\delta} = \frac{\mathrm{d}G(x)}{\mathrm{d}x} \tag{2.16}$$

(note that p is positive definite) so that (2.15) can also be written as

$$P(a \leq r < b|I) = \int_a^b p(r|I)\,\mathrm{d}r. \tag{2.17}$$

In terms of probability densities, the product rule (2.5) can now be written as

$$p(x, y|I) = p(x|y, I)\,p(y|I). \tag{2.18}$$

Likewise, the normalisation condition (2.11) can be written as

$$\int p(x|I)\,\mathrm{d}x = 1, \tag{2.19}$$

the marginalisation/expansion (2.13) as

$$p(x|I) = \int p(x, y|I)\,\mathrm{d}y = \int p(x|y, I)\,p(y|I)\,\mathrm{d}y \tag{2.20}$$

and Bayes' theorem (2.14) as

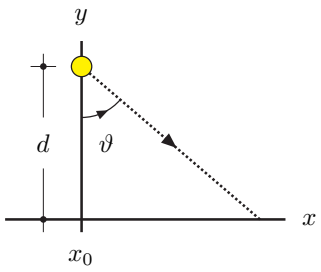
$$p(y|x,I) = \frac{p(x|y,I)\,p(y|I)}{\int p(x|y,I)\,p(y|I)\,\mathrm{d}y}. \tag{2.21}$$

Exercise 2.10: A counter produces a yes/no signal S when it is traversed by a pion. Given are the efficiency $P(S|\pi,I) = \varepsilon$ and miss-identification probability $P(S|\sim\pi,I) = \delta$. The fraction of pions in the beam is $P(\pi|I) = \mu$. What is the probability $P(\pi|S,I)$ that a particle which generates a signal is a pion in case (i) μ is known and (ii) μ is unknown? In the latter case assume a uniform prior distribution for μ in the range $0 \le \mu \le 1$.

We make four remarks: (1)—Probabilities are dimensionless numbers so that the dimension of a density is the reciprocal of the dimension of the variable. This implies that $p(x)$ transforms when we make a change of variable $x \to f(x)$. The size of the infinitesimal element $\mathrm{d}x$ corresponding to $\mathrm{d}f$ is given by $\mathrm{d}x = |\mathrm{d}x/\mathrm{d}f|\ \mathrm{d}f$; because the probability content of this element must be invariant we have

$$p(f|I)\,\mathrm{d}f = p(x|I)\,\mathrm{d}x = p(x|I)\left|\frac{\mathrm{d}x}{\mathrm{d}f}\right|\mathrm{d}f \qquad \text{and thus} \qquad p(f|I) = p(x|I)\left|\frac{\mathrm{d}x}{\mathrm{d}f}\right|. \tag{2.22}$$

Exercise 2.11: A lighthouse at sea is positioned a distance d from the coast.



This lighthouse emits collimated light pulses at random times in random directions, that is, the distribution of pulses is uniform in ϑ. Derive an expression for the probability to observe a light pulse as a function of the position x along the coast. (From Sivia [Sivia06].)

(2)—Without prior information it is tempting to chose a uniform distribution for the prior density $p(y|I)$ in Bayes' theorem (2.21). However, the distribution of a transformed variable $z = f(y)$ will then, in general, not be uniform. Uniformity thus seems to be not a very good criterion to characterise un-informative densities: we will come back to this in Section 5 where we introduce entropy as a measure of information content. Note, however, that in an iterative learning process the choice of initial prior becomes increasingly less important because the posterior obtained at one step can be taken as the prior in the next step.[12]

Exercise 2.12: We put two pion counters in the beam both with efficiency $P(S|\pi,I) = \varepsilon$ and miss-identification probability $P(S|\sim\pi,I) = \delta$. The fraction of pions in the beam is $P(\pi|I) = \mu$. A particle traverses and both counters give a positive signal. What is

[12]The rate of convergence can be much affected by the initial choice of prior, see Section 5.1 for a nice example taken from the book by Sivia.

the probability that the particle is a pion? Calculate this probability by first taking the posterior of the measurement in counter (1) as the prior for the measurement in counter (2) and second by considering the two responses $(S1, S2)$ as one measurement and using Bayes' theorem directly. In order to get the same result in both calculations an assumption has to be made on the measurements in counters (1) and (2). What is this assumption?

(3)—The equations (2.20) and (2.21) constitute, together with their discrete equivalents, the core of Bayesian inference. Indeed, apart from the approximations and transformations described in Section 3 and the maximum entropy principle described in Section 5, the remainder of these lectures will be not much more than repeated applications of *expansion*, *probability inversion* and *marginalisation*.

(4)—Plausible inference is, strictly speaking, always conducted in terms of probabilities instead of probability densities. A density $p(x|I)$ is turned into a probability by multiplying it with the infinitesimal element $\mathrm{d}x$. For conditional probabilities $\mathrm{d}x$ should refer to the random variable (in front of the vertical bar) and not to the condition (behind the vertical bar); thus $p(x|y, I)\mathrm{d}x$ is a probability but $p(x|y, I)\mathrm{d}y$ is not although it may make perfect sense mathematically. It is a good habit to keep track of these infinitesimal elements in a calculation even when most of them will cancel in the end.

2.5 Bayesian versus Frequentist inference (I)

The Bayesian interpretation of probability as a 'degree of belief' is not new since this was just the concept used by the early founders of probability theory like Bernoulli (1713), Bayes (1763), Laplace (1812) and others. Laplace, in particular, was very successful in applying Bayesian methods to celestial mechanics[13] and other fields of investigation. However, many objections were raised to Bayesian methods which caused them to fall into discredit at the beginning of the 20^{th} century, in favour of Frequentist approaches, much advocated by Fisher.

One reason for this was that the rules of probability calculus were known to apply to probability defined in terms of frequencies, but that no compelling reason could be given why they would apply to probability defined as a degree of belief. We have seen how the work of Cox has solved this problem in 1946.

Furthermore, probability as a degree of belief was considered to be *subjective*, and thus unfit for use in a scienctific argument. Bayesian probabilities are, of course, subjective in the sense that they depend on the amount of available information which may differ from one person to another. But that does not mean that these probabilities are *arbitrary*. Indeed, desideratum III of Section 2.1 requires that two people who are in equivalent states of knowledge about a proposition must assign equal probabilities to that proposition.

Another important feature of Bayesian inference is that it requires a prior probability density as input. Here it is not possible to fall-back on an interpretation of probability in terms of outcomes of a repeated experiment since the prior should reflect our knowledge—or lack of knowledge—*before* we do any experiment. Because of this, there

[13] Laplace determined, to very good precision, the mass of Saturn (and its uncertainty) from the limited set of astronomical data that were available to him.

exists at present no generally accepted method to assign these priors, which is seen by some as an insurmountable problem, and by others as just a technical difficulty. We will see in Section 5 of these notes that guidance is provided by the principle of insufficient reason, symmetry arguments, and maximum entropy. Note, however, that prior assignment is presently a large and active field of research which is beyond the scope of these lectures; for a review of recent developments you may consult [Kass96].

To avoid the problems mentioned above, the Frequentist defines probability as the frequency of the occurrence of an event in an infinitely large number of repetitions of the experiment or, equivalently, in an infinitely large *ensemble* of identical systems. Such a definition of probability is consistent with the Kolmogorov axioms and is also objective since its definition does not depend on an observer. But it also implies that Frequentist theory denies probability assignment to a hypothesis, since the hypothesis must be either true or false in all the repetitions of the experiment.[14] The inversion of $P(D|HI)$ to $P(H|DI)$ with Bayes' theorem is thus invalidated. This removes, in Frequentist inference, the need to specify that disturbing prior probability $P(H|I)$.

Because Bayes' theorem cannot be used, an hypothesis (*e.g.* the value of a parameter) is, in the Frequentist approach, accessed via a so-called *statistic* which is a function of the data and thus a *random variable* with a distribution that can be derived from the sampling distribution of the data. Such a statistic is called an **estimator** when it is used to estimate the value of a parameter (*e.g.* sample mean and sample variance to estimate the mean and variance of an underlying sampling distributition) or a **test statistic** when it is used access the validity of an hypotheses, or to discriminate between hypotheses (*e.g.* χ^2, z-statistic, t-statistic, F-statistic). There is no fundamental rule to construct a statistic and one has to base the choice on properties like *consistency*, *bias*, *efficiency* and *robustness* for estimators, and on *power* or *error-1* and *error-2* probabilities for test statistics. Bayesian inference does not make use of a statistic, parameter values being accessed via the posterior (Section 6) and hypotheses via model selection (Section 8).

One important feature of Bayesian inference is that the posterior distribution is always conditional on the data, that is, conclusions are always based on the data and prior information. The Frequentist approach, on the other hand, allows in so-called 'goodness of fit' tests that the validity of a hypothesis is based on hypothetical repetitions of the experiment which never took place.[15] The possible implications of this are nicely illustrated by the following example, taken from the book by Berger and Wolpert [Berg88].

Suppose we have a parameter θ and a random variable X which yields either $\theta - \sigma$ or $\theta + \sigma$, each with 50% probability. Our experiment consists of two observations x_1 and x_2 from which we want to calculate an estimate m of θ:

$$m = \begin{cases} \frac{1}{2}(x_1 + x_2) & \text{when } x_1 \neq x_2 \\ x_1 - \sigma & \text{when } x_1 = x_2 \end{cases}$$

The question now is what 'confidence' we can assign to this result, that is, what is

[14] A model parameter thus has for a Frequentist a fixed, but unknown, value which is also the view of a Bayesian since the probability distribution that he assigns to the parameter does not describe how it *fluctuates* but how *uncertain* we are of its value.

[15] Some people compare this to a judge which convicts a suspect on grounds of evidence that *could* have been produced, but never was. In a famous Dutch court case, Lucia de B. fell victim to this and served 7 years in prison.

the probability that the statement '$\theta = m$' is true? We can answer this question by constructing the **sample space** which consists of 4 events, each equally probable:

Event	x_1	x_2	m	$\theta = m$
E1	$\theta - \sigma$	$\theta - \sigma$	$\theta - 2\sigma$	F
E2	$\theta + \sigma$	$\theta + \sigma$	θ	T
E3	$\theta - \sigma$	$\theta + \sigma$	θ	T
E4	$\theta + \sigma$	$\theta - \sigma$	θ	T

Thus, if the experiment is repeated many times there will be a 75% chance of obtaining the right answer, and this is the confidence that an orthodox Frequentist would assign to the measurement. However, when we observe $x_1 = x_2$ we are 50% sure that the answer is right, and when we observe $x_1 \neq x_2$ we are 100% sure. This is because after the measurement the sample space has collapsed to {E1,E2} in the first case, and to {E3,E4} in the second. This collapse occurs when we 'condition on the data' which in Bayesian inference is a trivial consequence of Bayes' theorem. Indeed, Bayesian analysis leads to a posterior probability which is *always* conditional on the data. In Frequentist inference, on the other hand, one has to adhere to the **likelihood principle** which states that all experimental evidence about an unknown quantity θ is contained in the likelihood function (or a multiple thereof) of θ for given data. Berger and Wolpert mention in the introduction of their book that Bayesian analysis seems to be the most realistic implementation of the likelihood principle, and they state that 'Many Bayesians became Bayesians only because the likelihood principle left them little choice' [Berg88, p.2]. But note that the principle is not universally accepted by Frequentists.

Somewhat related to this is the observation that likelihoods may not only depend on the relevant information carried by the data but also on *how* the data were actually obtained. A well known example is the so-called *optional stopping problem* which we will discuss in Section 4.4. It turns out that Bayesian inference automatically discards information on the stopping strategy as being irrelevant—in accordance with desideratum IIIb in Section 2.1—while orthodox Frequentist inference does not, although a way-out is provided by the likelihood principle.

With these remarks we leave the Bayesian-Frequentist comparison for what it is and refer to the abundant literature on the subject, see *e.g.* [James00] for recent discussions.

3 Posterior Representation

The full result of Bayesian inference is the posterior distribution. However, instead of publishing this distribution in the form of a parametrisation, table, plot or computer program it is often more convenient to summarise the posterior—or any other probability distribution—in terms of a few parameters.

3.1 Measures of location and spread

The **expectation value** of a function $f(x)$ is defined by[16]

$$< f > = \int f(x)\, p(x|I)\, \mathrm{d}x. \tag{3.1}$$

Here the integration domain is understood to be the definition range of the distribution $p(x|I)$. The ***k*-th moment** of a distribution is the expectation value $< x^k >$. From (2.19) it immediately follows that the zeroth moment $< x^0 > = 1$. The first moment is called the **mean** of the distribution and is a location measure

$$\mu = \bar{x} = < x > = \int x\, p(x|I)\, \mathrm{d}x. \tag{3.2}$$

The **variance** σ^2 is the second moment about the mean

$$\sigma^2 = < \Delta x^2 > = < (x-\mu)^2 > = \int (x-\mu)^2\, p(x|I)\, \mathrm{d}x. \tag{3.3}$$

The square root of the variance is called the **standard deviation** and is a measure of the width of the distribution.

> **Exercise 3.1**: Show that the variance is related to the first and second moments by $< \Delta x^2 > = < x^2 > - < x >^2$.

The width of a **multivariate** distribution is characterised by the **covariance matrix**:

$$V_{ij} = < \Delta x_i \Delta x_j > = \int \cdots \int (x_i - \mu_i)(x_j - \mu_j)\, p(x_1, \ldots, x_n|I)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, \tag{3.4}$$

where μ_i is given by

$$\mu_i = \bar{x}_i = < x_i > = \int \cdots \int x_i\, p(x_1, \ldots, x_n|I)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

The covariance matrix is obviously symmetric.

> **Exercise 3.2**: Show that the off-diagonal elements of V_{ij} vanish when $x_1, \ldots, x_n$ are independent variables.

A correlation between the variables is better judged from the matrix of **correlation coefficients** which is defined by

$$\rho_{ij} = \frac{V_{ij}}{\sqrt{V_{ii} V_{jj}}} = \frac{V_{ij}}{\sigma_i \sigma_j}. \tag{3.5}$$

It can be shown that $-1 \leq \rho_{ij} \leq +1$.

[16] We discuss here only continuous variables; the expressions for discrete variables are obtained by replacing the integrals with sums.

The position of the maximum of a probability density function is called the **mode**[17] which often is taken as a location parameter (provided the distribution has a single maximum). For the general case of an n-dimensional distribution one finds the mode by minimising the function $L(\boldsymbol{x}) = -\ln p(\boldsymbol{x}|I)$. Expanding L around some point $\hat{\boldsymbol{x}}$ we can write

$$L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}}) + \sum_{i=1}^{n} \frac{\partial L(\hat{\boldsymbol{x}})}{\partial x_i} \Delta x_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 L(\hat{\boldsymbol{x}})}{\partial x_i \partial x_j} \Delta x_i \Delta x_j + \cdots \tag{3.6}$$

with $\Delta x_i \equiv x_i - \hat{x}_i$. We now take $\hat{\boldsymbol{x}}$ to be the mode, that is, the point where p is maximum and L is minimum. With this choice $\hat{\boldsymbol{x}}$ is a solution of the set of equations

$$\frac{\partial L(\hat{\boldsymbol{x}})}{\partial x_i} = 0 \tag{3.7}$$

so that the second term in (3.6) vanishes. Up to second order, the expansion can now be written in matrix notation as

$$L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}}) + \tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}})\boldsymbol{H}(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \cdots , \tag{3.8}$$

where the **Hessian matrix** of second derivatives is defined by

$$H_{ij} \equiv \frac{\partial^2 L(\hat{\boldsymbol{x}})}{\partial x_i \partial x_j}. \tag{3.9}$$

Taking the exponent of (3.8) gives for our approximation of the probability density in the neighbourhood of the mode:

$$p(\boldsymbol{x}|I) \approx C \exp[-\tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}})\boldsymbol{H}(\boldsymbol{x} - \hat{\boldsymbol{x}})] \tag{3.10}$$

where C is a constant which can be adjusted to $C = p(\hat{\boldsymbol{x}}|I)$ or to a value that normalises the right-hand side of (3.10). In the latter case the posterior is approximated by a normalised **multivariate Gaussian** in $\boldsymbol{x}$-space

$$p(\boldsymbol{x}|I) \approx \frac{1}{\sqrt{(2\pi)^n \, |\boldsymbol{V}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}})\boldsymbol{V}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{x}})\right], \tag{3.11}$$

where the *inverse* of the Hessian is identified with the covariance matrix $\boldsymbol{V}$ of the Gaussian[18] and where $|\boldsymbol{V}|$ in the normalisation term denotes the determinant of $\boldsymbol{V}$.[19] One should always bear in mind that the approximation (3.10) or (3.11) will, by construction, have the same *mode* as the posterior, but not necessarily the same *shape*, unless the posterior happens to be Gaussian of course.

Sometimes the distribution $p(\boldsymbol{x}|I)$ is such that the mode and Hessian can be calculated analytically. In most cases, however, minimisation programs like MINUIT are used to

[17] We denote the mode by $\hat{x}$ to distinguish it from the mean $\bar{x}$. For symmetric distributions this distinction is irrelevant since then $\hat{x} = \bar{x}$.

[18] This is why we have expanded in (3.6) the *logarithm* instead of the distribution itself: only then the inverse of the second derivative matrix is equal to the covariance matrix of a multivariate Gaussian.

[19] There is no problem with $\sqrt{|\boldsymbol{V}|}$ since $|\boldsymbol{V}|$ is positive definite as will be shown in Section 3.3.

determine numerically $\hat{\boldsymbol{x}}$ and $\boldsymbol{V} = \boldsymbol{H}^{-1}$ from $L(\boldsymbol{x})$ (the function L is then calculated in the subroutine `fcn` provided by the user).

The approximation (3.11) is easily marginalised. It can be shown (see Appendix A) that integrating a multivariate Gaussian over one variable x_i is equivalent to deleting the corresponding row and column i in the covariance matrix $\boldsymbol{V}$. This defines a new covariance matrix $\boldsymbol{V}'$ and, by inversion, a new Hessian $\boldsymbol{H}'$. Replacing $\boldsymbol{V}$ by $\boldsymbol{V}'$ and n by $(n-1)$ in (3.11) then obtains the integrated Gaussian. It is now easy to see that integration over all but one x_i gives

$$p(x_i|I) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{x}_i}{\sigma_i}\right)^2\right] \tag{3.12}$$

where σ_i^2 is the diagonal element V_{ii} of the covariance matrix $\boldsymbol{V}$.[20]

A continuous density $p(x|I)$ defined on an interval $[a, b]$ can also be described by the cumulative distribution

$$G(x) = \int_a^x p(y|I)\mathrm{d}y \qquad a \le x \le b,$$

with, obviously, $G(a) = 0$ and $G(b) = 1$. The value x_α for which $G(x_\alpha) = \alpha$ is called the **α-quantile** of the distribution. The 50% quantile, which divides the probability content in equal parts, is called the **median**.[21] The complement $1 - G(x)$ is, in Frequentist hypothesis testing, called the **p-value** of an observation x (small p-values being unlikely with x residing in the right-hand tail of the distribution).

Let us close this section by making the remark that one may very well encounter distributions for which the mean and variance do not exist because the integrals (3.2) or (3.3) are divergent. An example of this is the **Cauchy distribution**

$$p(x|I) = \frac{1}{\pi}\frac{1}{1+x^2}, \tag{3.13}$$

which we plot in Fig. 1.

Exercise 3.3: The Cauchy distribution is often called the **Breit-Wigner** distribution which usually is parametrised as

$$p(x|x_0, \Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(\Gamma/2)^2 + (x - x_0)^2}.$$

(i) Show that Γ is the FWHM (full width at half maximum). (ii) For simplicity set $x_0 = 0$ and $\Gamma = 2$ and calculate the Gaussian approximation (3.11) of the Breit-Wigner. Use a plotting program to check if this approximation is reasonable.

[20] The error on a 'fitted' parameter given by MINUIT is the diagonal element of the covariance matrix and is thus the width of the *marginal* distribution of this parameter.

[21] The median is often preferred as a measure of central tendency, because it is insensitive to outliers. In statistical language the median is called a *robust* estimator; the mean is clearly *not* robust.

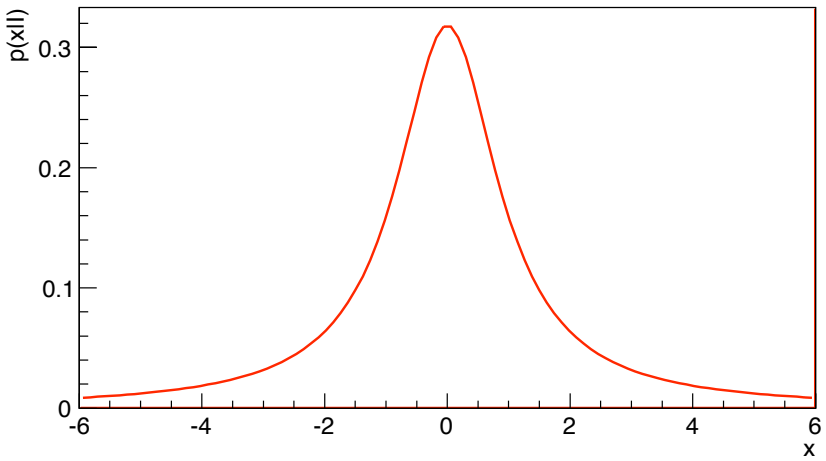


Figure 1: The Cauchy distribution.

3.2 Transformations

In this section we briefly describe how to construct probability densities of functions of a (multi-dimensional) random variable. We start by calculating the probability density $p(z|I)$ of a single function $z = f(\boldsymbol{x})$ from a given distribution $p(\boldsymbol{x}|I)$ of n variables $\boldsymbol{x}$. Expansion in the variable $\boldsymbol{x}$ gives

$$\begin{aligned} p(z|I) &= \int p(z, \boldsymbol{x}|I)\, \mathrm{d}\boldsymbol{x} = \int p(z|\boldsymbol{x}, I)\, p(\boldsymbol{x}|I)\, \mathrm{d}\boldsymbol{x} = \\ &= \int \delta[z - f(\boldsymbol{x})]\, p(\boldsymbol{x}|I)\, \mathrm{d}\boldsymbol{x} \end{aligned} \tag{3.14}$$

where we have made the trivial assignment $p(z|\boldsymbol{x}, I) = \delta[z - f(\boldsymbol{x})]$. This assignment guarantees that the integral only receives contributions from the hyperplane $f(\boldsymbol{x}) = z$.

As an example consider two *independent* variables x and y distributed according to $p(x, y|I) = f(x)g(y)$. Using (3.14) we find that the distribution of the sum $z = x + y$ is given by the **Fourier convolution** of f and g

$$p(z|I) = \int f(x)g(z - x)\, \mathrm{d}x = \int f(z - y)g(y)\, \mathrm{d}y. \tag{3.15}$$

Likewise we find that the product $z = xy$ is distributed according to the **Mellin convolution** of f and g

$$p(z|I) = \int f(x)g(z/x)\frac{\mathrm{d}x}{|x|} = \int f(z/y)g(y)\,\frac{\mathrm{d}y}{|y|}, \tag{3.16}$$

provided that the definition ranges do not include $x = 0$ and $y = 0$.

Exercise 3.4: Use (3.14) to derive Eqs. (3.15) and (3.16).

In case of a **coordinate transformation** it may be convenient to just use the Jacobian as we have done in (2.22) in Section 2.4. By a coordinate transformation we mean a

mapping $\mathbb{R}^n \to \mathbb{R}^n$ by a set of n functions

$$\boldsymbol{z}(\boldsymbol{x}) = \{z_1(\boldsymbol{x}), \ldots, z_n(\boldsymbol{x})\},$$

for which there exists an inverse transformation

$$\boldsymbol{x}(\boldsymbol{z}) = \{x_1(\boldsymbol{z}), \ldots, x_n(\boldsymbol{z})\}.$$

The probability density of the transformed variables is then given by

$$q(\boldsymbol{z}|I) = p[\boldsymbol{x}(\boldsymbol{z})|I] \ |\boldsymbol{J}| \tag{3.17}$$

where $|\boldsymbol{J}|$ is the absolute value of the determinant of the **Jacobian matrix**

$$J_{ik} = \frac{\partial x_i}{\partial z_k}. \tag{3.18}$$

Exercise 3.5: Let x and y be two independent variables distributed according to $p(x, y|I) = f(x)g(y)$. Let $u = x + y$ and $v = x - y$. Use (3.17) to obtain an expression for $p(u, v|I)$ in terms of f and g and show, by integrating over v, that the marginal distribution of u is given by (3.15). Likewise, define $u = xy$ and $v = x/y$ and show that the marginal distribution of u is given by (3.16).

The above, although it formally settles the issue of how to deal with functions of random variables, often gives rise to tedious algebra as can be seen from the following exercise:[22]

Exercise 3.6: Two variables x_1 and x_2 are independently Gaussian distributed:

$$p(x_i|I) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right] \qquad i = 1, 2.$$

Show, by carrying out the integral in (3.15), that the variable $z = x_1 + x_2$ is Gaussian distributed with mean $\mu = \mu_1 + \mu_2$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

A simple way to obtain numerical results is to generate $p(\boldsymbol{x}|I)$ by Monte Carlo, calculate $F(\boldsymbol{x})$ at each generation and then histogram the result.

However, if we are content with summarising the distributions by mean and covariance, and if $F(\boldsymbol{x})$ is not strongly varying, then we may use a very simple transformation rule, known as **linear error propagation**. Let $F_\lambda(\boldsymbol{x})$ be one of a set of m functions of $\boldsymbol{x}$. Linear approximation gives

$$\Delta F_\lambda \equiv F_\lambda(\boldsymbol{x}) - F_\lambda(\bar{\boldsymbol{x}}) = \sum_{i=1}^{n} \frac{\partial F_\lambda(\bar{\boldsymbol{x}})}{\partial x_i} \Delta x_i \tag{3.19}$$

with $\Delta x_i = x_i - \bar{x}_i$. Now multiplying (3.19) by the expression for ΔF_μ and averaging obtains

$$< \Delta F_\lambda \Delta F_\mu > = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial F_\lambda}{\partial x_i} \frac{\partial F_\mu}{\partial x_j} < \Delta x_i \Delta x_j > . \tag{3.20}$$

[22] Later on we will use Fourier transforms (characteristic functions) to make life much easier.

Eq. (3.20) can be written in compact matrix notation as

$$\boldsymbol{V}_F = \boldsymbol{D}\boldsymbol{V}_x\boldsymbol{D}^{\mathrm{T}} \tag{3.21}$$

where $\boldsymbol{D}$ denotes the $m \times n$ derivative matrix $D_{\lambda i} = \partial F_\lambda / \partial x_i$ and $\boldsymbol{D}^{\mathrm{T}}$ its transpose. Well known applications are the quadratic addition of errors for a sum of independent variables

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2 \quad \text{for} \quad z = x_1 + x_2 + \cdots + x_n \tag{3.22}$$

and the quadratic addition of *relative* errors for a product of independent variables

$$\left(\frac{\sigma}{z}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2 + \cdots + \left(\frac{\sigma_n}{x_n}\right)^2 \quad \text{for} \quad z = x_1 x_2 \cdots x_n. \tag{3.23}$$

Exercise 3.7: Use (3.20) to derive the two propagation rules (3.22) and (3.23).

Exercise 3.8: A counter is traversed by N particles and fires n times. Since $n \subseteq N$ these counts are not independent but n and $m = N - n$ are. Assume Poisson errors (Section 4.6) $\sigma_n = \sqrt{n}$ and $\sigma_m = \sqrt{m}$ and use (3.20) or (3.21) to show that the error on the efficiency $\varepsilon = n/N$ is given by

$$\sigma_\varepsilon = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}}$$

This is known as the binomial error, see Section 4.2.

Exercise 3.9: Let x be a random variable distributed according to $p(x|I)$. Show that the cumulative distribution of x is uniform.

3.3 The covariance matrix revisited

In this section we investigate in some more detail the properties of the covariance matrix which, together with the mean, fully characterises the multivariate Gaussian (3.11).

In Section 3.1 we have already remarked that $\boldsymbol{V}$ is symmetric but not every symmetric matrix can serve as a covariance matrix. To see this, consider a function $f(\boldsymbol{x})$ of a set of Gaussian random variables $\boldsymbol{x}$. For the variance of f we have according to (3.21)

$$\sigma^2 = <\Delta f^2> = \boldsymbol{d}\,\boldsymbol{V}\boldsymbol{d},$$

where $\boldsymbol{d}$ is the vector of derivatives $\partial f / \partial x_i$. But since σ^2 is positive for any function f it follows that the following inequality must hold:

$$\boldsymbol{d}\,\boldsymbol{V}\boldsymbol{d} > 0 \quad \text{for any vector} \quad \boldsymbol{d}. \tag{3.24}$$

A matrix that possesses this property is called **positive definite**.

A covariance matrix can be diagonalised by a unitary transformation. This can be seen from Fig. 2 where we show the one standard deviation contour of two correlated Gaussian variables (x_1, x_2) and two uncorrelated variables (y_1, y_2). It is clear from these plots that the two error ellipses are related by a simple rotation. A pure rotation is not

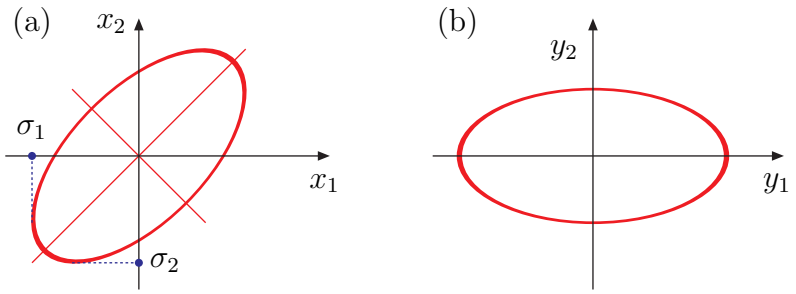


Figure 2: The one standard deviation contour of a two dimensional Gaussian for (a) correlated variables x_1 and x_2 and (b) uncorrelated variables y_1 and y_2. The marginal distributions of x_1 and x_2 have a standard deviation of σ_1 and σ_2, respectively.

the only way to diagonalise the covariance matrix since the rotation can be combined with a scale transformation along y_1 or y_2.

The rotation $\boldsymbol{U}$ which diagonalises $\boldsymbol{V}$ must, according to the transformation rule (3.20), satisfy the relation

$$\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{L} \quad \Rightarrow \quad \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{L} \tag{3.25}$$

where $\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes a diagonal matrix. In (3.25) we have used the property $\boldsymbol{U}^{-1} = \boldsymbol{U}^{\mathrm{T}}$ of an orthogonal transformation. Let the columns i of $\boldsymbol{U}^{\mathrm{T}}$ be denoted by the set of vectors $\boldsymbol{u}^{(i)}$, that is,

$$u_j^{(i)} = U_{ji}^{\mathrm{T}} = U_{ij}. \tag{3.26}$$

It is then easy to see that (3.25) corresponds to the set of **eigenvalue equations**

$$\boldsymbol{V}\boldsymbol{u}^{(i)} = \lambda_i \boldsymbol{u}^{(i)}. \tag{3.27}$$

Thus, the rotation matrix $\boldsymbol{U}$ and the vector of diagonal elements λ_i is determined by the complete set of eigenvectors and eigenvalues of the covariance matrix $\boldsymbol{V}$. From the eigenvalue spectrum is it easy to calculate the **condition number** $\kappa = \lambda_{\mathrm{max}}/\lambda_{\mathrm{min}}$ of the covariance matrix. A large condition number means that the matrix is ill-conditioned, that is, susceptible to round-off errors.

Exercise 3.10: Show that (3.25) is equivalent to (3.27).

Exercise 3.11: (i) Show that for a symmetric matrix $\boldsymbol{V}$ and two arbitrary vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ the following relation holds $\boldsymbol{y}\boldsymbol{V}\boldsymbol{x} = \boldsymbol{x}\boldsymbol{V}\boldsymbol{y}$; (ii) Show that the eigenvectors $\boldsymbol{u}^i$ and $\boldsymbol{u}^j$ of a symmetric matrix $\boldsymbol{V}$ are orthogonal, that is, $\boldsymbol{u}^i\boldsymbol{u}^j = 0$ for $i \neq j$; (iii) Show that the eigenvalues of a positive definite symmetric matrix $\boldsymbol{V}$ are all positive.

The normalisation factor of the multivariate Gaussian (3.11) is proportional to the square root of the determinant $|\boldsymbol{V}|$, which only makes sense if this determinant is positive definite. Indeed,

$$|\boldsymbol{V}| = |\boldsymbol{U}|\,|\boldsymbol{V}|\,\left|\boldsymbol{U}^{\mathrm{T}}\right| = \left|\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right| = |\boldsymbol{L}| = \prod_i \lambda_i > 0,$$

where we have used the fact that $|\boldsymbol{U}| = 1$ and that all the eigenvalues of $\boldsymbol{V}$ are positive.

4 Basic Probability Assignment

In Section 2.3 we have introduced the operation of *expansion* which allows us to assign a compound probability by writing it as a sum of known elementary probabilities.

Bernoulli (1713) was among the first to formulate a rule for an elementary probability assignment which was, later on, called the **principle of insufficient reason**, also known as the **principle of indifference**:

> If for a set of N exclusive and exhaustive propositions there is no evidence to prefer any proposition over the others then each proposition must be assigned an equal probability $1/N$.

As a very basic application of expansion in combination with Bernoulli's principle we will discuss, in the next section, drawing balls from an urn. In passing we will make some observations about probabilities which are *very* Bayesian and which you may find quite surprising if you have never encountered them before. This is because probability is for you as a Frequentist a property of the urn and its contents, while for you as a Bayesian it is a measure of what you *know* about the urn and its contents.

We then proceed by deriving the Binomial distribution in subsection 4.2 using nothing else but the sum and product rules of probability calculus. Multinomial and Poisson distributions are introduced in the subsections 4.5 and 4.6. In the last subsection we will introduce a new tool, the characteristic function, and derive the Gauss distribution as the limit of a sum of arbitrarily distributed random variables.

4.1 Bernoulli's urn

Consider an experiment where balls are drawn from an urn. Let the urn contain N balls and let the balls be labelled $i = 1, \ldots, N$. We can now define the exhaustive and exclusive set of hypotheses

$$H_i = \text{'this ball has label } i\text{'}, \qquad i = 1, \ldots, N.$$

Since we have no information on which ball we will draw we use the principle of insufficient reason to assign the probability to get ball 'i' at the first draw:

$$P(H_i|N, I) = \frac{1}{N}. \tag{4.1}$$

Next, we consider the case that R balls are coloured red and $W = N - R$ are coloured white. We define the exhaustive and exclusive set of hypotheses

$$\begin{aligned} H_R &= \text{'this ball is red'} \\ H_W &= \text{'this ball is white'}. \end{aligned}$$

We now want to assign the probability that the first ball we draw will be red. To solve this problem we expand this probability into the hypothesis space $\{H_i\}$ which gives

$$\begin{aligned} P(H_R|I) &= \sum_{i=1}^{N} P(H_R, H_i|I) = \sum_{i=1}^{N} P(H_R|H_i, I)P(H_i|I) \\ &= \frac{1}{N}\sum_{i=1}^{N} P(H_R|H_i, I) = \frac{R}{N} \end{aligned} \tag{4.2}$$

where, in the last step, we have made the trivial probability assignment

$$P(H_R|H_i, I) = \begin{cases} 1 & \text{if ball '}i\text{' is red} \\ 0 & \text{otherwise.} \end{cases}$$

Next, we assign the probability that the second ball will be red. This probability depends on how we draw the balls:

1. We draw the first ball, put it back in the urn and shake the urn. The latter action may be called 'randomisation' but from a Bayesian point of view the purpose of shaking the urn is, in fact, to destroy all *information* we might have on the whereabouts of this ball after it was put back in the urn (it would most likely end-up in the top layer of balls). Since this **drawing with replacement** does not change the contents of the urn and since the shaking destroys all previously accumulated information, the probability of drawing a red ball a second time is equal to that of drawing a red ball the first time:

   $$P(R_2|I) = \frac{R}{N},$$

   where R_2 stands for the hypothesis 'the second ball is red'.

2. We record the colour of the first ball, lay it aside, and then draw the second ball. Obviously the content of the urn has changed after the first draw. Depending on the colour of the first ball we assign:

   $$\begin{aligned} P(R_2|R_1, I) &= \frac{R-1}{N-1} \\ P(R_2|W_1, I) &= \frac{R}{N-1} \end{aligned}$$

   Exercise 4.1: Draw the first ball blindly and put it aside (*i.e.* without knowing its colour). Show that the probability for the second draw to be red is the same as that for the first draw to be red: $P(R_2|I) = R/N$. This result is hard to swallow if you insist that probability is a property of the urns content, which has changed after the first draw!

In the above we have seen that the probability of the second draw may depend on the outcome of the first draw. We will now show that the probability of the first draw may depend on the outcome of the second draw! Consider the following situation: The first draw is blind and the ball is put aside (without knowing its colour). The probability

that this first ball is red is R/N, as we have shown in (4.2). A second ball is drawn and it turns out to be red. What is now the probability that the first ball was red? Bayes' theorem immediately shows that it is *not* R/N:

$$P(R_1|R_2,I) = \frac{P(R_2|R_1,I)P(R_1|I)}{P(R_2|R_1,I)P(R_1|I)+P(R_2|W_1,I)P(W_1|I)} = \frac{R-1}{N-1}.$$

If this argument fails to convince you, take the extreme case of an urn containing one red and one white ball. The probability of a red ball at the first draw is 1/2. Lay the ball aside (without knowing its colour) and take the second ball. If it is red, then the probability that the first ball was red is zero and not 1/2. The fact that the second draw influences the probability of the first draw has of course nothing to do with a *causal* relation but, instead, with a *logical* relation.

Exercise 4.2: Draw a first ball blindly and put it back in the urn. The colour of a second draw is red. What is the probability that the first draw was red?

4.2 Binomial distribution

We now make N draws from the urn, putting the ball back after each draw and shaking the urn. In this way the probability that a draw is red is the same for all draws: $h = R/N$. (We call the probability 'h' for 'heads' in coin flipping which will be our archetypical random process later on in these notes.) What is the probability that we find n red balls in our sample of N draws? Again, we seek to expand this probability into a combination of elementary ones which are easy to assign. Let us start with the hypothesis

$$S_j = \text{"the } N \text{ balls are drawn in the sequence labelled '}j\text{' "}$$

where $j = 1,\ldots,2^N$ is the index in a list of all possible sequences (of length N) of white and red draws. The set of hypotheses $\{S_j\}$ is obviously exclusive and exhaustive. The draws are independent, that is, the probability of the k^{th} draw does not depend on the outcome of the other draws (remember that this is only true for drawing with replacement). Thus we find from the product rule

$$P(S_j|I) = P(C_1,\ldots,C_N|I) = \prod_{k=1}^{N} P(C_k|I) = h^{n_j}(1-h)^{N-n_j}, \tag{4.3}$$

where C_k stands for red or white at the k^{th} draw and where n_j is the number of red draws in the sequence j. Having assigned the probability of each element in the set $\{S_j\}$, we now expand our probability of n red balls into this set:

$$\begin{aligned} P(n|I) &= \sum_{j=1}^{2^N} P(n,S_j|I) = \sum_{j=1}^{2^N} P(n|S_j,I)P(S_j|I) \\ &= \left[\sum_{j=1}^{2^N} \delta(n-n_j)\right] h^n(1-h)^{N-n} \end{aligned} \tag{4.4}$$

where we have assigned the trivial probability

$$P(n,|S_j,I) = \delta(n-n_j) = \begin{cases} 1 & \text{when the sequence } S_j \text{ contains } n \text{ red draws} \\ 0 & \text{otherwise.} \end{cases}$$

The sum inside the square brackets in (4.4) counts the number of sequences in the set $\{S_j\}$ which have just n red draws. It is an exercise in combinatorics to show that this number is given by the binomial coefficient. Thus we obtain

$$P(n|h,N) = \frac{N!}{n!(N-n)!}\, h^n (1-h)^{N-n} = \binom{N}{n} h^n (1-h)^{N-n}. \tag{4.5}$$

This is called the **binomial distribution** which applies to all processes where the outcome is binary (red or white, head or tail, yes or no, absent or present *etc.*), provided that the probability h of the outcome of a single draw is the same for all draws. In Fig. 3 we show the distribution of red draws for $N = (10, 20, 40)$ trials for an urn with $h = 0.25$.

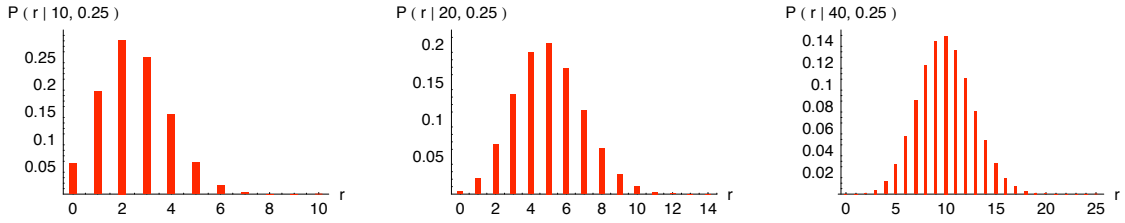


Figure 3: The binomial distribution to observe r red balls in $N = (10, 20, 40)$ draws from an urn containing a fraction $h = 0.25$ of red balls.

The binomial probabilities are just the terms of the binomial expansion

$$(a+b)^N = \sum_{n=0}^{N} \binom{N}{n} a^n\, b^{N-n} \tag{4.6}$$

with $a = h$ and $b = 1 - h$. From this it follows immediately that

$$\sum_{n=0}^{N} P(n|h,N) = 1.$$

The condition of independence of the trials is important and may not be fulfilled: for instance, suppose we scoop a handful of balls out of the urn and count the number n of red balls in this sample. Does n follow the binomial distribution? The answer is 'no' since we did not perform draws with replacement, as required. This can also be seen from the extreme situation where we take *all* balls out of the urn. Then n would not be distributed at all: it would just be R.

The first and second moments and the variance of the binomial distribution are

$$\begin{aligned} <n> &= \sum_{n=0}^{N} nP(n|h,N) = Nh \\ <n^2> &= \sum_{n=0}^{N} n^2 P(n|h,N) = Nh(1-h) + N^2h^2 \\ <\Delta n^2> &= <n^2> - <n>^2 = Nh(1-h) \end{aligned} \tag{4.7}$$

If we now define the ratio $\mu = n/N$ then it follows immediately from (4.7) that

$$<\mu> = h \qquad <\Delta\mu^2> = \frac{h(1-h)}{N} \tag{4.8}$$

The square root of this variance is called the **binomial error** which we have already encountered in Exercise 3.8. It is seen that the variance vanishes in the limit of large N and thus that μ converges to h in that limit. This fundamental relation between a probability and a limiting relative frequency was first discovered by Bernoulli and is called the **law of large numbers**. This law is, of course, the basis for the Frequentist definition of probability.

For a uniform prior, the Binomial posterior is given by $p(h|n,N)\mathrm{d}h = CP(n|h,N)\mathrm{d}h = Ch^n(1-h)^{(N-n)}\mathrm{d}h$. Integrating this distribution to calculate the normalisation constant, we find for the posterior, and its *mode*,

$$p(h|n,N)\,\mathrm{d}h = \frac{(N+1)!}{n!(N-n)!}\,h^n(1-h)^{(N-n)}\,\mathrm{d}h \qquad \hat{h} = \frac{n}{N} \pm \sqrt{\frac{\hat{h}(1-\hat{h})}{N}}. \tag{4.9}$$

Here we have characterised the width of the posterior by the inverse of the Hessian.

Exercise 4.3: A counter is traversed by N particles and fires N times. Calculating the efficiency and error from (4.8) gives $\varepsilon = 1 \pm 0$ which is an unacceptable estimate for the error. Derive an expression for the lower limit ε_α corresponding to the α confidence interval defined by the equation

$$\int_{\varepsilon_\alpha}^{1} p(\varepsilon|N,N)\,\mathrm{d}\varepsilon = \alpha.$$

Show that for $N = 4$ and $\alpha = 0.65$ the result on the efficiency can be reported as

$$\varepsilon = 1\ ^{+0}_{-0.19} \qquad (65\%\ \mathrm{CL})$$

4.3 The negative binomial

Instead of drawing N balls from the urn (with replacement), we may decide to draw balls as many times as is necessary to observe n red balls. Because the last draw must by definition be red, and because the probability of this draw does not depend on the previous draws, the probability of N draws is given by

$$\begin{aligned} P(N|n,h) &= P(n-1 \text{ red balls in } N-1 \text{ draws}) \times P(\text{one red ball in one draw}) \\ &= \frac{(N-1)!}{(n-1)!(N-n)!}\,h^N\,(1-h)^{N-n} \qquad n \geq 1,\ N \geq n. \end{aligned} \tag{4.10}$$

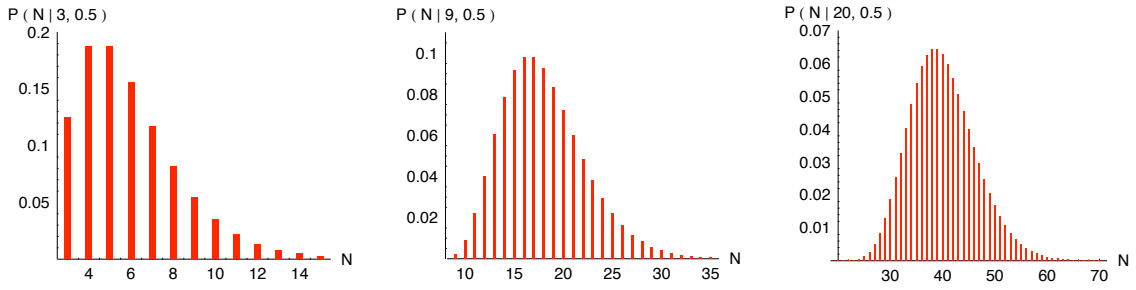


Figure 4: The negative binomial distribution $P(N|n,h)$ of the number of trials N needed to observe $n = (3, 9, 20)$ red balls in draws from an urn with 50% red balls ($h = 0.5$).

This distribution is known as the **negative binomial**. In Fig. 4 we show this distribution for $h = 0.5$ and $n = (3, 9, 20)$ red balls.

It can be shown that $P(N|n,h)$ is properly normalised

$$\sum_{N=n}^{\infty} P(N|n,h) = 1.$$

The first and second moments and the variance of this distribution are

$$\begin{aligned} <N> &= \sum_{N=n}^{\infty} NP(N|n,h) = \frac{n}{h} \\ <N^2> &= \sum_{N=n}^{\infty} N^2 P(N|n,h) = \frac{n(1-h)}{h^2} + \frac{n^2}{h^2} \\ <\Delta N^2> &= \frac{n(1-h)}{h^2} \end{aligned} \tag{4.11}$$

If we define the ratio $Q = N/n$ as our statistic for $z \equiv 1/h$ it follows directly from (4.11) that the average and variance of Q are given by

$$<Q> = z \qquad <\Delta Q^2> = \frac{z(z-1)}{N} \tag{4.12}$$

4.4 The stopping problem

There is a curious problem in the analysis of counting experiments, related to the fact that the sampling distribution (or likelihood) depends on the strategy we have adopted to halt the experiment. In a simple coin flipping experiment, for instance, the sampling distribution is different when we chose to stop after a fixed amount of N throws or chose to stop after the observation of a fixed amount of n heads. In the first case the number of heads is the random variable while in the second case it is the number of throws. In the following we will have a closer look at these two stopping strategies. As we will see, Frequentist inference is sensitive to the stopping rules but Bayesian inference is *not*.

We denote the probability of heads in coin flipping by h. If we stop the experiment after N throws, the likelihood of observing n heads is given by the binomial distribution

$$P(n|N,h,I) = \frac{N!}{n!(N-n)!}\, h^n (1-h)^{N-n}. \tag{4.13}$$

The expectation value of the *statistic* $R = n/N$ is, from (4.8),

$$< R > = < \frac{n}{N} > = \frac{< n >}{N} = h. \tag{4.14}$$

If we stop after observing n heads, the likelihood is given by the negative binomial

$$P(N|n,h,I) = \frac{(N-1)!}{(n-1)!(N-n)!}\; h^n\; (1-h)^{N-n}. \tag{4.15}$$

But the expectation value of R over the negative binomial is *not* equal to h. This can easily be seen, without any explicit calculation, from the fact that N and not n is the random variable and that the reciprocal of an expectation value is not the expectation value of the reciprocal. Indeed, using (4.11) for the expectation value $< N >$ of the negative binomial we find

$$\frac{n}{< N >} = h \quad \text{but} \quad < R > = < \frac{n}{N} > = n < \frac{1}{N} > \neq \frac{n}{< N >}.$$

The expectation value $< R >$ of our estimator R thus depends not only on the data (n heads in N throws) but also on the stopping strategy! It follows that a Frequentist cannot analyse these data unless the stopping strategy is known, since he needs this knowledge to construct a meaningful statistic. For a Bayesian the situation is different.

From the binomial likelihood (4.13) and Bayes' theorem, we obtain for the posterior of h

$$p(h|n,N,I) = C\, P(n|N,h,I)\, p(h|I) = C\, h^n\; (1-h)^{N-n}\; p(h|I), \tag{4.16}$$

where $p(h|I)$ is the prior for h and C is a normalisation constant. Taking, instead, the negative binomial likelihood (4.15), we get

$$p'(h|n,N,I) = C'\, P(N|n,h,I)\, p(h|I) = C'\, h^n\; (1-h)^{N-n}\; p(h|I). \tag{4.17}$$

Normalisation gives $C' = C$ and thus $p' = p$.

Here we have encountered a very nice property of Bayesian inference, namely its ability to discard information which is irrelevant. This is in accordance with Cox' desideratum of consistency which states that conclusions should depend on relevant information only. Frequentist inference does *not* possess this property since the stopping rule must be specified in order to construct the likelihood and a meaningful statistic.[23]

It can be shown (Exercise 4.4 below) that if we stop *at random*—for instance when a certain amount of time has elapsed—the number of heads is Poisson distributed. Also in this case, Bayesian inference is not sensitive to the stopping rule.

[23]Note that the dependence on the stopping strategy disappears in the posterior normalisation step. This is equivalent to saying that normalisation factors in the likelihood are irrelevant, and this is—not surprisingly—also stated in the 'likelihood principle' of Frequentist statistics.

Exercise 4.4: We flip a coin at an average rate of R flips per second and decide to stop when a time Δt has elapsed. This gives a Poisson distribution for N

$$P(N|\mu) = \frac{\mu^N}{N!} e^{-\mu},$$

where $\mu = R\Delta t$. Derive an expression for the sampling distribution $P(n|h, \mu)$ for n heads. Show that the posterior distribution of h is, again, the same as (4.16).

4.5 Multinomial distribution

A generalisation of the binomial distribution is the **multinomial distribution** which applies to N independent trials where the outcome of each trial is among a set of k alternatives with probability p_i. Examples are drawing from an urn containing balls with k different colours, the throwing of a dice ($k = 6$) or distributing N independent events over the bins of a histogram.

The multinomial distribution can be written as

$$P(\boldsymbol{n}|\boldsymbol{p}, N) = \frac{N!}{n_1! \cdots n_k!} \, p_1^{n_1} \cdots p_k^{n_k} \tag{4.18}$$

where $\boldsymbol{n} = (n_1, \ldots, n_k)$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$ are vectors subject to the constraints

$$\sum_{i=1}^{k} n_i = N \qquad \text{and} \qquad \sum_{i=1}^{k} p_i = 1. \tag{4.19}$$

The multinomial probabilities are just the terms of the expansion

$$(p_1 + \cdots + p_k)^N$$

from which the normalisation of $P(\boldsymbol{n}|\boldsymbol{p}, N)$ immediately follows. The average, variance and covariance are given by

$$\begin{aligned} <n_i> &= Np_i \\ <\Delta n_i^2> &= Np_i(1 - p_i) \\ <\Delta n_i \Delta n_j> &= -Np_ip_j \quad \text{for} \quad i \neq j. \end{aligned} \tag{4.20}$$

Marginalisation is achieved by adding in (4.18) two or more variables n_i and their corresponding probabilities p_i.

Exercise 4.5: Use the addition rule above to show that the marginal distribution of each n_i in (4.18) is given by the binomial distribution $P(n_i|p_i, N)$ as defined in (4.5).

The conditional distribution on, say, the count n_k is given by

$$P(\boldsymbol{m}|n_k, \boldsymbol{q}, M) = \frac{M!}{n_1! \cdots n_{k-1}!} \, q_1^{n_1} \cdots q_{k-1}^{n_{k-1}}$$

where

$$\boldsymbol{m} = (n_1, \ldots, n_{k-1}), \quad \boldsymbol{q} = \frac{1}{s}(p_i, \ldots, p_{k-1}), \quad s = \sum_{i=1}^{k-1} p_i \quad \text{and} \quad M = N - n_k.$$

Exercise 4.6: Derive the expression for the conditional probability by dividing the joint probability (4.18) by the marginal (binomial) probability $P(n_k|p_k, N)$.

Exercise 4.7: We fill the k bins of a histogram according to the probabilities

$$(p_1, \ldots, p_{k-1}, p_k = 1 - S) \quad \text{where} \quad S = \sum_{i=1}^{k-1} p_i.$$

The histogram is filled at a rate of R entries per second and we stop the accumulation after a time Δt has elapsed. The likelihood for a total content N of the histogram is then

$$P(N|\mu) = \frac{\mu^N}{N!}\, e^{-\mu} \quad \text{with} \quad \mu = R\Delta t.$$

Assume uniform priors for $(p_1, \ldots, p_{k-1})$ and μ and show that the posterior of the quantities $c_i = \mu p_i$ is given by a product of independent Poisson distributions

$$p(c_1, \ldots, c_k|n_1, \ldots, n_k) = \prod_{i=1}^{k} \frac{c_i^{n_i}}{n_i!} e^{-c_i}.$$

Show also that the posterior for $(p_1, \ldots, p_{k-1})$ is given by

$$p(p_1, \ldots, p_{k-1}|n_1, \ldots, n_k) \propto \left(\prod_{i=1}^{k-1} p_i^{n_i} \right) (1 - S)^{n_k}.$$

Note that this is a generalisation of (4.16).

4.6 Poisson distribution

Here we consider 'events' or 'counts' which occur randomly in time (or space). The counting rate R is supposed to be given, that is, we know the average number of counts $\mu = R\Delta t$ in a given time interval Δt. There are several ways to derive an expression for the probability $P(n|\mu)$ to observe n events in a time interval with contains, on average, μ events. Our derivation is based on the fact that this probability distribution is a limiting case of the binomial distribution.

For this, we divide the interval Δt in N sub-intervals δt. The probability to observe an event in such a sub-interval is then $p = \mu/N$, see (4.2). Now we can always make N so large and δt so small that the number of events in each sub-interval is either one or zero. The probability to find n events in N sub-intervals is then equal to the (binomial) probability to find n successes in N trials:

$$P(n|N) = \frac{N!}{n!(N-n)!} \left(\frac{\mu}{N}\right)^n \left(1 - \frac{\mu}{N}\right)^{N-n}.$$

Re-arranging some terms and taking the limit $N \to \infty$ then yields the desired result

$$P(n|\mu) = \lim_{N\to\infty} \frac{N!}{(N-n)!(N-\mu)^n} \frac{\mu^n}{n!} \left(1 - \frac{\mu}{N}\right)^N = \frac{\mu^n}{n!}\, e^{-\mu}. \tag{4.21}$$

This distribution is known as the **Poisson distribution**. The normalisation, average and variance are given by, respectively,

$$\sum_{n=0}^{\infty} P(n|\mu) = 1, \quad <n> = \mu \quad \text{and} \quad <\Delta n^2> = \mu. \tag{4.22}$$

Exercise 4.8: A counter is traversed by beam particles at an average rate of R particles per second. (i) If we observe n counts in a time interval Δt, derive an expression for the posterior distribution of $\mu = R\Delta t$, given that the prior for μ is uniform. Calculate mean, variance, mode and width (inverse of the Hessian) of this posterior. (ii) Give an expression for the probability $p(\tau|R, I)\,\mathrm{d}\tau$ that the time interval between the passage of two particles is between τ and $\tau + \mathrm{d}\tau$.

4.7 Gauss distribution

The sum of many small random fluctuations follows the **Gauss distribution** (also called the **normal distribution**), irrespective of the distribution of each of the terms contributing to the sum (see below for some restrictions). This fact, known as the **central limit theorem**, is often held responsible for the dominant presence of the Gauss distribution in statistical data analysis. But there are also other reasons for the ubiquitous Gauss distribution, as we will see in Section 5.

To prove the central limit theorem we first have to introduce the **characteristic function**, which is nothing else than the Fourier transform of a probability density. The Fourier transform, and its inverse, of a distribution $p(x)$ is defined by

$$\phi(k) = \int_{-\infty}^{\infty} e^{ikx} p(x)\,\mathrm{d}x \qquad p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ikx}\,\phi(k)\,\mathrm{d}k \tag{4.23}$$

This transformation plays an important role in proving many theorems related to sums of random variables and moments of probability distributions. This is because a Fourier transform turns a Fourier convolution in x-space, see (3.15), into a product in k-space.[24] To see this, consider a joint distribution of n *independent* variables

$$p(\boldsymbol{x}|I) = f_1(x_1)\cdots f_n(x_n).$$

Using (3.14) we write for the transform of the distribution of the sum $z = \sum x_i$

$$\begin{aligned}
\phi(k) &= \int_{-\infty}^{\infty}\mathrm{d}z\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathrm{d}x_1\cdots\mathrm{d}x_n\,\exp(ikz)\,f_1(x_1)\cdots f_n(x_n)\,\delta(z-\sum_{i=1}^{n}x_i)\\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathrm{d}x_1\cdots\mathrm{d}x_n\,\exp\left(ik\sum_{i=1}^{n}x_i\right)\,f_1(x_1)\cdots f_n(x_n)\\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathrm{d}x_1\cdots\mathrm{d}x_n\,\exp(ikx_1)f_1(x_1)\cdots\exp(ikx_n)f_n(x_n)\\
&= \phi_1(k)\cdots\phi_n(k).
\end{aligned} \tag{4.24}$$

The transform of a sum of independent random variables is thus the product of the transforms of each variable.

The moments of a distribution are related to the derivatives of the transform at $k = 0$:

$$\frac{\mathrm{d}^n\phi(k)}{\mathrm{d}k^n} = \int_{-\infty}^{\infty}(ix)^n\,e^{ikx}p(x)\,\mathrm{d}x \quad\Rightarrow\quad \frac{\mathrm{d}^n\phi(0)}{\mathrm{d}k^n} = i^n\,<x^n>\,. \tag{4.25}$$

[24] A Mellin transform turns a Mellin convolution (3.16) in x-space into a product in k-space. We will not discuss Mellin transforms in these notes.

The characteristic functions (Fourier transforms) of many distributions can be found in, for instance, the particle data book [Eid04]. Of importance to us is the Gauss distribution and its transform

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \phi(k) = \exp\left(i\mu k - \frac{1}{2}\sigma^2 k^2\right) \tag{4.26}$$

To prove the central limit theorem we consider the sum of a large set of n random variables $s = \sum x_j$. Each x_j is distributed *independently* according to $f_j(x_j)$ with mean μ_j and a standard deviation σ which we take, for the moment, to be the same for all f_j. To simplify the algebra, we do not consider the sum itself but rather

$$z = \sum_{j=1}^{n} y_j = \sum_{j=1}^{n} \frac{x_j - \mu_j}{\sqrt{n}} = \frac{s-\mu}{\sqrt{n}} \tag{4.27}$$

where we have set $\mu = \sum \mu_i$. Now take the Fourier transform $\phi_j(k)$ of the distribution of y_j and make a Taylor expansion around $k = 0$. Using (4.25) we find

$$\begin{aligned} \phi_j(k) &= \sum_{m=0}^{\infty} \frac{k^m}{m!} \frac{\mathrm{d}^m \phi_j(0)}{\mathrm{d}k^m} = \sum_{m=0}^{\infty} \frac{(ik)^m < y_j^m >}{m!} \\ &= 1 + \sum_{m=2}^{\infty} \frac{(ik)^m < (x_j - \mu_j)^m >}{m!\, n^{m/2}} = 1 - \frac{k^2\sigma^2}{2n} + \mathrm{O}(n^{-3/2}) \end{aligned} \tag{4.28}$$

Taking only the first two terms of this expansion we find from (4.24) for the characteristic function of z

$$\phi(k) = \left(1 - \frac{k^2\sigma^2}{2n}\right)^n \;\rightarrow\; \exp\left(-\frac{1}{2}\sigma^2 k^2\right) \quad \text{for } n \rightarrow \infty. \tag{4.29}$$

But this is just the characteristic function of a Gaussian with mean zero and width σ. Transforming back to the sum s we find

$$p(s) = \frac{1}{\sigma\sqrt{2\pi n}} \exp\left[\frac{(s-\mu)^2}{n\sigma^2}\right] \tag{4.30}$$

It can be shown that the central limit theorem also applies when the widths of the individual distributions are different in which case the variance of the Gauss is $\sigma^2 = \sum \sigma_i^2$ instead of $n\sigma^2$ as in (4.30). However, the theorem breaks down when one or more individual widths are much larger than the others, allowing for one or more variables x_i to occasionally dominate the sum. It is also required that all μ_i and σ_i exist so that the theorem does not apply to, for instance, a sum of Cauchy distributed variables.

Exercise 4.9: Apply (4.25) to the characteristic function (4.26) to show that the mean and variance of a Gauss distribution are μ and σ^2, respectively.

Exercise 4.10: In Exercise 3.6 we have derived the distribution of the sum of two Gaussian distributed variables by explicitly calculating the convolution integral (3.15). Derive the same result by using the characteristic function (4.26). Convince yourself that the 'central limit theorem' *always* applies to sums of Gaussian distributed variables even for a finite number of terms or large differences in width.

5 Least Informative Probabilities

5.1 Impact of prior knowledge

The impact of the prior on the outcome of plausible inference is nicely illustrated by a very instructive example, taken from Sivia [Sivia06], were the bias of a coin is determined from the observation of the number of heads in N throws.

Let us first recapitulate what constitutes a **well posed** problem so that we can apply Bayesian inference.

- First, we need to define a complete set of hypotheses. For our coin flipping experiment this will be the value of the probability h to obtain a head in a single throw. The definition range is $0 \leq h \leq 1$.
- Second, we need a model which relates the set of hypotheses to the data. In other words, we need to construct the likelihood $P(D|H, I)$ for all the hypotheses in the set. In our case this is the binomial probability to observe n heads in N throws of a coin with bias h

$$P(n|h, N, I) = \frac{N!}{n!(N-n)!}\, h^n (1-h)^{N-n}. \tag{5.1}$$

- Finally, we need to specify the prior probability $p(h|I)\,\mathrm{d}h$.

If we observe n heads in N throws, the posterior distribution of h is given by Bayes' theorem

$$p(h|n, N, I)\,\mathrm{d}h = C\, h^n (1-h)^{N-n}\, p(h|I)\,\mathrm{d}h, \tag{5.2}$$

where C is a normalisation constant which presently is of no interest to us. In Fig. 5 these posteriors are shown for 10, 100 and 1000 throws of a coin with bias $h = 0.25$ for a flat prior (top row of plots), a strong prior preference for $h = 0.5$ (middle row) and a prior which excludes the possibility that $h < 0.5$ (bottom row). It is seen that the flat prior converges nicely to the correct answer $h = 0.25$ when the number of throws increases. The second prior does this too, but more slowly. This is not surprising because we have encoded, in this case, quite a strong prior belief that the coin is unbiased and it takes a lot of evidence from the data to change that belief. In the last choice of prior we see that the posterior cannot go below $h = 0.5$, because we have excluded this region by setting the prior to zero. This is an illustration of the fact that no amount of data can change certainties encoded by the prior as we have already remarked in Exercise 2.8. This can of course be turned into an advantage since it allows us to exclude physically forbidden regions from the posterior, like a negative mass for instance.

Two important lessons can be learnt from this exercise in coin flipping. First, we learn that the conclusion derived from the data depends on prior knowledge. This is not a weak point of Bayesian inference but a strong one, because prior information *must* play a role, as the following simple argument will show. If we observe 255 heads in 1000 throws of the coin then, without prior knowledge about the coin (or the throwing process), we would conclude that it is biased. Now suppose that we have convinced

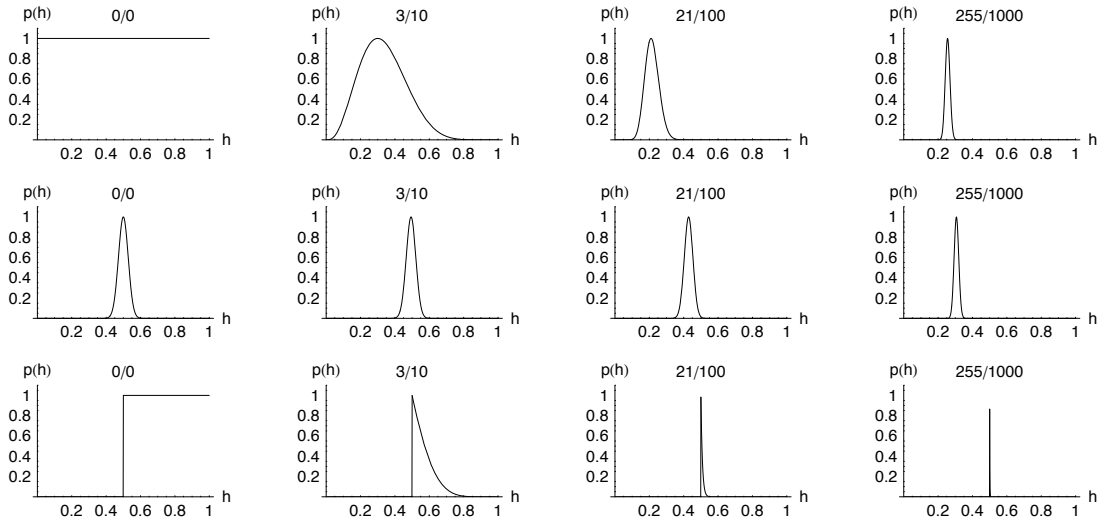


Figure 5: The posterior density $p(h|n, N)$ for n heads in N flips of a coin with bias $h = 0.25$. In the top row of plots the prior is uniform, in the middle row it is strongly peaked around h = 0.5 while in the bottom row the region $h < 0.5$ has been excluded. The posterior densities are scaled to unit maximum for ease of comparison.

ourselves beforehand, by careful measurement of its properties, that the coin is fair. The observation of 255 heads in 1000 throws would then lead to the conclusion that we just have witnessed a very rare event, or that the coin has been exchanged, or that something went wrong with the counting, or that some mechanism controls the throws, or whatever other explanation we may come up with, but *not* that our coin is biased!

Second, we learn that unsupported information should not enter into the prior because it may need a lot of data to converge to the correct result in case this information turns out to be wrong. The maximum entropy principle provides a means to construct priors which have the property that they are consistent with given boundary conditions but are otherwise maximally un-informative.

Let us close this section by making a few general remarks on priors. First, it is clear that when the likelihood is narrow compared to the prior, it does not matter very much what prior distribution we chose. On the other hand, when the likelihood is so wide that it competes with any reasonable prior then this simply means that the experiment does not carry much information on the subject we want to investigate. In such a case it should not come as a surprise that answers become dominated by prior knowledge or assumptions. Of course there is nothing wrong with that, as long as these prior assumptions are clearly stated (alternatively one could try to look for better data!). The prior also plays an important role when the likelihood peaks near a physical boundary or, as very well may happen, resides in an unphysical region (likelihoods related to neutrino mass measurements are a famous example; for another example see Section 7.1 in this write-up). In such cases, the information on the boundary is mostly (or exclusively) contained in the prior and not in the data.

Before we proceed with the maximum entropy principle let us first, in the next section, make a few remarks on symmetry considerations in the assignment of probabilities.

5.2 Symmetry considerations

In Section 4 we have introduced Bernoulli's principle of insufficient reason which states that, in absence of relevant information, equal probability should be assigned to members of an enumerable exclusive set of hypotheses. Below we give a very simple invariance argument supporting this principle.

Suppose we would plot the probabilities assigned to each hypothesis in a bar chart. If we are in a state of complete ignorance about these hypotheses then it obviously should not matter how they would be ordered in such a chart. Stated differently, in absence of additional information the set of hypotheses is invariant under permutations. But our bar chart of probabilities can *only* be invariant under permutations if all the probabilities are the same, hence Bernoulli's principle.

Similarly, translation invariance implies that the least informative probability distribution of a so-called **location parameter** is uniform. Indeed, a translation invariant probability should obey the relation

$$p(x|I)\,\mathrm{d}x = p(x+a|I)\,\mathrm{d}(x+a) = p(x+a|I)\,\mathrm{d}x, \tag{5.3}$$

which can be satisfied only when $p(x|I)$ is a constant.

A somewhat less intuitive assignment is related to positive definite **scale parameters**. Scale invariance implies that for $r > 0$ and $\alpha > 0$

$$p(r|I)\,\mathrm{d}r = p(\alpha r|I)\,\mathrm{d}(\alpha r) = \alpha\, p(\alpha r|I)\,\mathrm{d}r. \tag{5.4}$$

But this is only possible when $p(r|I) \propto 1/r$. This probability assignment is called a **Jeffreys prior**. Note that a Jeffreys prior is *uniform* in $\ln(r)$, which means that it assigns equal probability per decade instead of per unit interval as does a uniform prior.

Both the uniform and the Jeffreys prior cannot be normalised when the variable ranges are $x \in [-\infty, \infty]$ or $r \in [0, \infty]$. Such un-normalisable distributions are called **improper**. The way to deal with improper distributions is to normalise them on a finite interval and take the limits to infinity (or zero) at the *end* of the calculation.[25] The posterior should, of course, always remain finite. If not, you may have to carefully reformulate your problem or it may be that your data simply do not carry enough information. In the exercise below we illustrate this by an estimate of the decay rate R from an observation of n counts in a time interval Δt. Assuming a Poisson likelihood and a Jeffreys prior it turns out that the posterior is only finite when at least one count is observed; for small R we have to be patient and wait for that one count!.

> **Exercise 5.1**: We make an inference on a counting rate R by observing the number of counts n in a time interval Δt. Assume that the likelihood $P(n|R\Delta t)$ is Poisson distributed as defined by (4.21). Assume further a Jeffreys prior for R, defined on the *positive* interval $R \in [a, b]$. Show that (i) for $\Delta t = 0$ the posterior is equal to the prior and that we cannot take the limits $a \to 0$ or $b \to \infty$; (ii) when $\Delta t > 0$ but still $n = 0$ we can take the limit $b \to \infty$ but not $a \to 0$; (iii) that both limits can be taken once $n > 0$. (From Gull [Gull88].)

[25] Limits should always be taken at the end of a mathematical calculation, if one does not want to run into all kind of paradoxes. Often it is also important to exactly specify *how* the limits are taken.

Finally, let us remark that we have only touched here upon very simple cases so that the above may seem quite trivial. However, in Jaynes [Jay03] you can find several examples which are far from trivial. In his book, Jaynes also includes a beautiful argument, due to the astronomer J. Herschel (1850), where the Gaussian distribution is derived from symmetry considerations alone. Here is how it goes:

Herschel was looking for the probability distribution which quantifies the uncertainty in a measurement of the position of a star in the sky. He defined an orthogonal coordinate system (x, y), centred on the true star position with x running horizontal (azimuth) and y vertical (altitude). The corresponding polar coordinate system is denoted by (r, φ). Next, Herschel postulated

1. The position in x does not yield information on that in y and *vice versa*.
2. The uncertainty distribution does not depend on φ.

According to the first postulate the unknown distribution should factorise in x and y. Together with the second postulate we obtain the following functional equation

$$p(x, y)\mathrm{d}x\mathrm{d}y = f(x)g(y)\mathrm{d}x\mathrm{d}y = h(r)r\mathrm{d}r\mathrm{d}\varphi.$$

The second postulate also enforces $f() = g()$ so that we have

$$f(x)f(y) = h(\sqrt{x^2 + y^2}). \tag{5.5}$$

Setting x (or y) to zero obtains, for real argument z

$$f(0)f(z) = h(|z|).$$

Substituting this result in the right-hand side of (5.5) gives

$$f(x)f(y) = f(0)f(\sqrt{x^2 + y^2}).$$

Defining $u(z) = \ln[f(z)/f(0)]$ the above equation can be written as

$$u(x) + u(y) = u(\sqrt{x^2 + y^2}),$$

which only can be satisfied if $u(z) = \alpha z^2$. For $\alpha > 0$ we then have

$$f(z) = f(0)\exp(\pm\alpha z^2),$$

where only the negative exponent is acceptable since the distribution should be normalisable. Thus,

$$f(z) = f(0)\exp(-\alpha z^2) \qquad (\alpha > 0).$$

Normalising the distribution and denoting the variance by σ^2 we find

$$\begin{aligned} \iint_{-\infty}^{\infty} f(0)\exp[-\alpha(x^2 + y^2)]\mathrm{d}x\mathrm{d}y &= \frac{\pi f(0)}{\alpha} = 1, \\ \iint_{-\infty}^{\infty} f(0)\, x^2 \exp[-\alpha(x^2 + y^2)]\mathrm{d}x\mathrm{d}y &= \frac{\pi f(0)}{2\alpha^2} = \sigma^2. \end{aligned}$$

Solving for $f(0)$ and α then gives for the uncertainty distribution

$$p(x, y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right),$$

which is just the expression for an uncorrelated two-dimensional Gaussian distribution with equal width in x and y, see also (3.11) in Section 3.1.

5.3 Maximum entropy principle

In the assignments we have made up to now (mainly in Section 4), there was always enough information to unambiguously determine the probability distribution. For instance if we apply the principle of insufficient reason to a fair dice then this leads to a probability assignment of 1/6 for each face $i = 1, \ldots, 6$. Note that this corresponds to an expectation value of $< i >= 3.5$. But what probability should we assign to each of the six faces when no information is given about the dice except that, say, $< i >= 4.5$? There are obviously an infinite number of probability distributions which satisfy this constraint so that we have to look elsewhere for a criterion to select one of these.

Jaynes (1957) has proposed to take the distribution which is the least informative by maximising the **entropy** (MAXENT). The concept of entropy as a measure of information was first introduced by Shannon (1948) in his pioneering paper on information theory [Shan48]. The entropy carried by a probability distribution is defined by[26]

$$\begin{aligned} S[p_1, \ldots, p_n] &= -\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right) \qquad \text{(discrete case)} \\ S[p] &= -\int p(x) \ln\left[\frac{p(x)}{m(x)}\right] \mathrm{d}x \quad \text{(continuous case)} \end{aligned} \tag{5.6}$$

where m_i or $m(x)$ is the so-called **Lebesgue measure** which satisfies

$$\sum_{i=1}^{n} m_i = 1 \quad \text{or} \qquad \int m(x)\, \mathrm{d}x = 1. \tag{5.7}$$

Roughly speaking, the Lebesgue measure associates to each subspace of the sample space a positive real number that measures the 'size' of that subspace. Note that the measure $m(x)$ makes the entropy invariant under coordinate transformations since both p and m transform in the same way.

Some formal insight can be gained by maximising (5.6), imposing only the normalisation constraint and nothing else. Restricting ourselves to the discrete case we find, using the method of Lagrange multipliers, that the following equation has to be satisfied

$$\delta\left[\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right) + \lambda\left(\sum_{i=1}^{n} p_i - 1\right)\right] = 0. \tag{5.8}$$

Differentiation of (5.8) to p_i leads to the equation

$$\ln\left(\frac{p_i}{m_i}\right) + 1 + \lambda = 0,$$

so that

$$p_i = m_i e^{-(\lambda+1)}. \tag{5.9}$$

[26]The definition (5.6) is such that larger entropies correspond to smaller information content of the probability distribution.

Imposing the normalisation constraint $\sum p_i = 1$ we find, using (5.7),

$$\sum_{i=1}^{n} p_i = \left(\sum_{i=1}^{n} m_i \right) e^{-(\lambda+1)} = e^{-(\lambda+1)} = 1.$$

Substituting this result into (5.9), it follows that

$$p_i = m_i \quad \text{(discrete case)}, \qquad p(x)\,\mathrm{d}x = m(x)\,\mathrm{d}x \quad \text{(continuous case)}. \tag{5.10}$$

It follows that the Lebesgue measure m_i or $m(x)$ is just the least informative probability distribution in complete absence of information. This '*Ur*-prior' thus describes the structure of our sample space and to determine it we have, again, to look elsewhere and use symmetry arguments (see Section 5.2) or we just make an *ansatz* which can always be revised later, if necessary.

Let us now assume that we have chosen some Lebesgue measure m, say uniform, and proceed by imposing further constraints on the probability distribution, like specifying moments, expectation values, *etc.* Such constraints are called **testable information** because one can always verify afterward that the MAXENT distribution indeed satisfies the constraint. We can write, in the discrete case, the constraints as a set of k independent weighted sums of the probabilities p_i

$$\sum_{i=1}^{n} w_{ji}\, p_i = \beta_j \qquad j = 1, \ldots, k. \tag{5.11}$$

Using Lagrange multipliers we maximise the entropy by solving the equation

$$\delta \left[\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right) + \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) + \sum_{j=1}^{k} \lambda_j \left(\sum_{i=1}^{n} w_{ji} p_i - \beta_j \right) \right] = 0. \tag{5.12}$$

Differentiating to p_i gives the equation

$$\ln\left(\frac{p_i}{m_i}\right) + 1 + \lambda_0 + \sum_{j=1}^{k} \lambda_j w_{ji} = 0 \quad \Rightarrow \quad p_i = m_i \exp(-1-\lambda_0) \exp\left(-\sum_{j=1}^{k} \lambda_j w_{ji} \right).$$

Imposing the normalisation condition $\sum p_i = 1$ to solve for λ_0, we find

$$p_i = \frac{1}{Z}\, m_i \exp\left(-\sum_{j=1}^{k} \lambda_j w_{ji} \right) \tag{5.13}$$

where we have introduced the **partition function** (normalisation sum)

$$Z(\lambda_1, \ldots, \lambda_k) = \sum_{i=1}^{n} m_i \exp\left(-\sum_{j=1}^{k} \lambda_j w_{ji} \right). \tag{5.14}$$

Such partition functions play a very important role because our constraints (5.11) are encoded in Z through (see Exercise 5.2)

$$-\frac{\partial \ln Z}{\partial \lambda_j} = \beta_j. \tag{5.15}$$

The formal solution (5.13) guarantees that the normalisation condition is obeyed but we still have to solve for the unknown Lagrange multipliers $\lambda_1, \ldots, \lambda_k$ from the equations (5.15) or, equivalently, by substituting (5.13) into (5.11). This often has to be done numerically.

For a continuous distribution $p(x|I)$, the above reads as follows. Let

$$\int f_j(x)\, p(x|I)\, \mathrm{d}x = \beta_j \qquad j = 1, \ldots, k \tag{5.16}$$

be a set of k testable constraints. The distribution that maximises the entropy is then given by

$$p(x|I) = \frac{1}{Z}\, m(x)\, \exp\left[-\sum_{j=1}^{k} \lambda_j f_j(x)\right]. \tag{5.17}$$

Here the partition function Z (normalisation integral) is defined by

$$Z(\lambda_1, \ldots, \lambda_k) = \int m(x)\, \exp\left[-\sum_{j=1}^{k} \lambda_j f_j(x)\right]\, \mathrm{d}x. \tag{5.18}$$

The values of the Lagrange multipliers λ_i are either found by solving (5.15), or by substituting (5.17) back into (5.16).

> **Exercise 5.2**: Prove (5.15) by differentiating the logarithm of the partition function (5.14) or (5.18).

5.4 MAXENT distributions

In this section we will derive from the maximum entropy principle a few well known distributions: the uniform, exponential, Gauss, and Poisson distributions. The fact that they can be derived from MAXENT sheds some new light on the origin of these distributions, namely that they are not necessarily related to some underlying random process, as is assumed in Frequentist theory (and also in Section 4) but that they can also be viewed as least informative distributions. With the MAXENT assignment, we indeed have moved far away from random variables, repeated observations, and the like.

If there are no constraints, $f(x) = 0$ in (5.16) so that it immediately follows from (5.17), (5.18) and (5.7) that

$$p(x|I) = m(x).$$

For a sample space without structure this gives a uniform distribution, in accordance with the continuum limit of Bernoulli's principle of insufficient reason.

Let us now consider a continuous distribution defined on $[0, \infty]$ and impose a constraint on the mean

$$< x > = \int_0^\infty x p(x|I)\, \mathrm{d}x = \mu \tag{5.19}$$

so that $f(x) = x$ in (5.16). From (5.17) and (5.18) we have, assuming a uniform Lebesgue measure,

$$p(x|I) = e^{-\lambda x} \left[\int_0^\infty e^{-\lambda x}\, \mathrm{d}x\right]^{-1} = \lambda e^{-\lambda x}.$$

Substituting this into (5.19) leads to

$$\int_0^\infty x\,\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda} = \mu$$

from which we find that x follows an **exponential distribution**

$$p(x|\mu, I) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right). \tag{5.20}$$

The moments of this distribution are given by

$$<x^n> = n!\,\mu^n \text{ so that } \quad <x> = \mu, \quad <x^2> = 2\mu^2 \text{ and } \quad <\Delta x^2> = \mu^2.$$

Another interesting case is a continuous distribution defined on $[-\infty, \infty]$ with a constraint on the variance

$$<\Delta x^2> = \int_{-\infty}^{\infty}(x-\mu)^2\,p(x|I)\,\mathrm{d}x = \sigma^2$$

so that $f(x) = (x-\mu)^2$ in (5.16). We find from (5.17), after normalisation,

$$p(x|I) = \sqrt{\frac{\lambda}{\pi}}\exp\left[-\lambda(x-\mu)^2\right].$$

The constraint on the variance allows us to solve for λ:

$$<\Delta x^2> = \sqrt{\frac{\lambda}{\pi}}\int_{-\infty}^{\infty}(x-\mu)^2\exp\left[-\lambda(x-\mu)^2\right]\,\mathrm{d}x = \frac{1}{2\lambda} = \sigma^2,$$

so that x turns out to be Gaussian distributed

$$p(x|\mu, \sigma, I) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \tag{5.21}$$

This is a very important result because it means that we do not necessarily have to invoke the central limit theorem to justify a Gaussian. It can be applied to all cases where we want to describe noise of which nothing is known, except its level (characterised by the value of the variance).

This is then the fourth time we encounter the Gaussian: in Section 3.1 as a convenient approximation of the posterior in the neighbourhood of the mode, in Section 4.7 as the limiting distribution of a sum of random variables, in Section 5.2 as a consequence of symmetry constraints and in this section as the least informative distribution consistent with a constraint on the variance.

As an example of a non-uniform Lebesgue measure we will now derive the Poisson distribution from the maximum entropy principle. We want to know the distribution $P(n|I)$ of n counts in a time interval Δt when there is nothing given but the average

$$\sum_{n=0}^{\infty} nP(n|I) = \mu. \tag{5.22}$$

To find the Lebesgue measure of the time interval Δt we divide it into a very large number (M) of intervals δt. A particular distribution of counts over these infinitesimal boxes is called a micro-state. If the micro-states are independent and equally probable then it follows from the sum rule that the probability of observing n counts is proportional to the number of micro-states which have n boxes occupied, which is given by the binomial coefficient. For large $M \gg n$ this becomes $M^n/n!$, as is easy to show by using the Stirling approximation for $M!$. Upon normalisation we then have for the Lebesgue measure

$$m(n) = \frac{M^n}{n!} e^{-M}. \tag{5.23}$$

Inserting this result in (5.13) we get

$$P(n|I) = \frac{C\, e^{-M} \left(M e^{-\lambda}\right)^n}{n!}$$

with

$$\frac{1}{C} = e^{-M} \sum_{n=0}^{\infty} \frac{\left(M e^{-\lambda}\right)^n}{n!} = e^{-M} \exp\left(M e^{-\lambda}\right),$$

so that

$$P(n|I) = \frac{\left(M e^{-\lambda}\right)^n}{\exp\left(M e^{-\lambda}\right) n!}. \tag{5.24}$$

To calculate the average we observe that

$$\sum_{n=0}^{\infty} \frac{n \left(M e^{-\lambda}\right)^n}{n!} = -\frac{\partial}{\partial \lambda} \sum_{n=0}^{\infty} \frac{\left(M e^{-\lambda}\right)^n}{n!} = -\frac{\partial}{\partial \lambda} \exp\left(M e^{-\lambda}\right) = M e^{-\lambda} \exp\left(M e^{-\lambda}\right)$$

Combining this with (5.24) we find from the constraint (5.22)

$$M e^{-\lambda} = \mu \quad \Rightarrow \quad P(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \tag{5.25}$$

which is the same result as derived in Section 4.6.

6 Parameter Estimation

In data analysis the measurements are often described by a parametrised model. In hypothesis testing such a model is called a **composite hypothesis** (*i.e.* one with parameters) in contrast to a **simple hypothesis** (without parameters). Given a composite hypothesis, the problem is how to extract information on the parameters from the data. This is called **parameter estimation**. It is important to realise that the composite hypothesis is assumed here to be true; investigating the plausibility of the hypothesis itself, by comparing it to a set of alternatives, is called 'model selection'. This will be the subject of Section 8.

The relation between the model and the data is encoded in the likelihood function

$$p(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{s}, I)$$

where $\boldsymbol{d}$ denotes a vector of data points and $\boldsymbol{\theta}$ and $\boldsymbol{s}$ are the model parameters which we have sub-divided in two classes:

1. The class $\boldsymbol{\theta}$ of parameters of interest;

2. The class $\boldsymbol{s}$ of so-called **nuisance parameters** which are necessary to model the data but are otherwise of no interest. These parameters often describe the systematic uncertainties due to detector calibration, acceptance corrections and so on. Input parameters also belong to this class like, for instance, an input value of the strong coupling constant $\alpha_\mathrm{s} \pm \Delta\alpha_\mathrm{s}$, taken from the literature.

There may also be parameters in the model which have known values. These are, if not explicitly listed, included in the background information 'I'.

Given a prior distribution for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{s}$, Bayes' theorem gives for the joint posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{s}|\boldsymbol{d}, I)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{s} = \frac{p(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{s}, I)p(\boldsymbol{\theta}, \boldsymbol{s}|I)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{s}}{\int p(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{s}, I)p(\boldsymbol{\theta}, \boldsymbol{s}|I)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{s}}. \tag{6.1}$$

The posterior of the parameters $\boldsymbol{\theta}$ is then obtained by marginalisation of the nuisance parameters $\boldsymbol{s}$:

$$p(\boldsymbol{\theta}|\boldsymbol{d}, I) = \int p(\boldsymbol{\theta}, \boldsymbol{s}|\boldsymbol{d}, I)\,\mathrm{d}\boldsymbol{s}. \tag{6.2}$$

As we have discussed in Section 5, choosing appropriate priors for the parameters $\boldsymbol{\theta}$ may, or may not be a delicate issue (often it is not). However, a very nice feature of priors is that unphysical regions can be excluded from the posterior by simply setting it to zero. In this way it is—to give an example—impossible to obtain a negative value for the neutrino mass even when that would be preferred by the likelihood. The priors for $\boldsymbol{s}$ are assumed to be known from detector studies (Monte Carlo simulations) or, in case of external parameters, from the literature. Note that the marginalisation (6.2) provides a very elegant way to propagate the uncertainties in the parameters $\boldsymbol{s}$ to the posterior distribution of $\boldsymbol{\theta}$ ('systematic error propagation', see Section 6.4).

Bayesian parameter estimation is thus fully described by the equations (6.1) and (6.2). But the evaluation of these two innocent looking formulae may need a lot of sophistication to properly assign the probabilities and to compute the integrals. These may be far from trivial tasks, in particular when we deal with complicated detectors and/or when our parameter space has a large number of dimensions. Considerable simplifications occur when two or more variables are independent (the probability distributions then factorise), when the distributions are Gaussian or when the model is linear in the parameters.

In the following subsections we will discuss a few simple cases which are frequently encountered in data analysis.

6.1 Gaussian sampling

One of the most simple parameter estimation problems is to find the mean and/or variance of a Gaussian distribution from which a sample of n measurements is drawn.

Suppose that we know the width σ of the Gaussian (resolution of our measuring device) and want to find the best estimate for the mean (or mode) μ from a set of n independent

observations d_i. Since the measurements are independent, we can use the product rule (2.8) and write for the likelihood

$$p(\boldsymbol{d}|\mu,\sigma) = \prod_{i=1}^{n} p(d_i|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{d_i-\mu}{\sigma}\right)^2\right].$$

The quantity $(d_i - \mu)/\sigma$ is called a **residual**. Assuming a uniform prior for μ the posterior becomes

$$p(\mu|\boldsymbol{d},\sigma) = C \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{d_i-\mu}{\sigma}\right)^2\right]. \tag{6.3}$$

To calculate the normalisation constant it is convenient to write

$$\sum_{i=1}^{n}(d_i-\mu)^2 = V + n(\bar{d}-\mu)^2$$

where

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad V = \sum_{i=1}^{n}(d_i-\bar{d})^2.$$

The quantities $\bar{d}$ and $S^2 \equiv V/(n-1)$ are called the **sample mean** and **sample variance**, respectively (we will see later why V is divided by $n-1$ and not by n). The constant C is now obtained from

$$\frac{1}{C} = \exp\left(-\frac{V}{2\sigma^2}\right)\int_{-\infty}^{\infty}\exp\left[-\frac{n(\bar{d}-\mu)^2}{2\sigma^2}\right]\mathrm{d}\mu = \sqrt{\frac{2\pi\sigma^2}{n}}\exp\left(-\frac{V}{2\sigma^2}\right).$$

Inserting this result in (6.3) we find

$$p(\mu|\bar{d},\sigma,n) = \sqrt{\frac{n}{2\pi\sigma^2}}\exp\left[-\frac{n}{2}\left(\frac{\bar{d}-\mu}{\sigma}\right)^2\right]. \tag{6.4}$$

But this is just a Gaussian with mean $\bar{d}$ and width $\sigma/\sqrt{n}$. Thus we have the well known result

$$\hat{\mu} = \bar{\mu} = \bar{d} \pm \frac{\sigma}{\sqrt{n}}. \tag{6.5}$$

Exercise 6.1: Derive (6.5) directly from (6.3) by expanding $L = -\ln p$ using equations (3.6), (3.7) and (3.9) in Section 3.1. Calculate the width as the inverse of the Hessian.

Now suppose that not only μ but also σ is unknown. Assuming a Jeffreys prior

$$p(\sigma|I) = \begin{cases} 0 & \text{for } \sigma \le 0 \\ 1/\sigma & \text{for } \sigma > 0 \end{cases}$$

we find for the posterior

$$p(\mu,\sigma|\bar{d},V,n) \propto \frac{1}{\sigma^{n+1}}\exp\left[-\frac{V+n(\bar{d}-\mu)^2}{2\sigma^2}\right]. \tag{6.6}$$

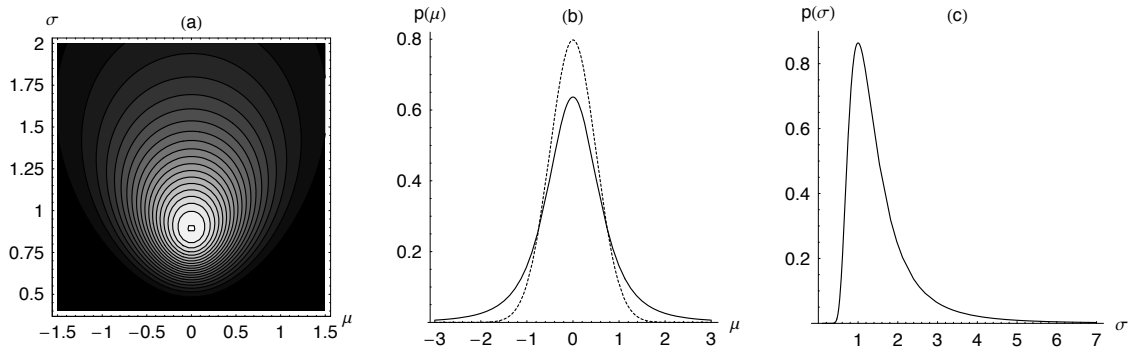


Figure 6: (a) The joint distribution $p(\mu, \sigma|\bar{d}, V, n)$ for $\bar{d} = 0$, $V = 4$ and $n = 4$; (b) The marginal distribution $p(\mu|\bar{d}, V, n)$ (full curve), compared to a Gaussian with $\sigma = 1/2$ (dashed curve); (c) The marginal distribution $p(\sigma|V, n)$.

In Fig. 6a we show this joint distribution for four samples ($n = 4$) drawn from a Gauss distribution of zero mean and unit width. For this plot, the random variables in (6.6) were set to $\bar{d} = 0$ and $V = 4$.

The posterior for μ is found by integrating over σ.

$$p(\mu|\bar{d}, V, n) = \int_0^\infty p(\mu, \sigma|\bar{d}, V, n)\, \mathrm{d}\sigma = C \left[\frac{1}{V + n(\bar{d} - \mu)^2}\right]^{n/2}. \tag{6.7}$$

When $n = 1$, the distribution (6.7) is improper (cannot be normalised on $[-\infty, \infty]$). Calculating the normalisation constant C for $n \geq 2$ we find the **Student-t distribution**:

$$p(\mu|\bar{d}, V, n) = \frac{\Gamma[n/2]}{\Gamma[(n-1)/2]} \sqrt{\frac{n}{\pi V}} \left[\frac{V}{V + n(\bar{d} - \mu)^2}\right]^{n/2}. \tag{6.8}$$

To obtain a more universal form for this distribution, define the scaled variable $t = (\bar{d} - \mu)/\Delta$, where Δ^2 is the variance of the sample mean $\bar{d}$. This variance is taken to be the sample variance S^2, divided by n: $\Delta^2 = S^2/n = V/n(n-1)$. In terms of t, the distribution reads

$$p(t|\nu) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \tag{6.9}$$

which has the number of degrees of freedom $\nu \equiv n - 1$ as the only parameter. This is the expression for the Student-t distribution with ν degrees of freedom, usually found in the literature.

Exercise 6.2: Derive (6.9) from (6.8).

The Student-t distribution for $\nu = 1$ degree of freedom is equal to the Cauchy distribution (3.13) which has no well defined moments (except normalisation). For $\nu \geq 2$ the first moment is given by $< t >= 0$ and for $\nu \geq 3$ the second moment is given by $< t^2 >= \nu/(\nu - 2)$. From this it is easy to see that the mean and standard deviation of (6.8) are given by

$$< \mu >= \bar{d} \qquad (n \geq 3) \qquad \text{and} \qquad < \Delta\mu^2 >= \frac{V}{n(n-3)} \qquad (n \geq 4). \tag{6.10}$$

Exercise 6.3: Derive (6.10).

In Fig. 6b we show the marginal distribution (6.8) for $\bar{d} = 0$, $V = 4$ and $n = 4$ (full curve), compared to a Gaussian with zero mean and width $\sigma/\sqrt{n} = 1/2$. It is seen that the Student-t distribution is similar to a Gaussian but has much longer tails.

Likewise, we can integrate (6.6) over μ to obtain the posterior for σ:

$$p(\sigma|\bar{d}, V, n) = \int_{-\infty}^{\infty} p(\mu, \sigma|\bar{d}, V, n)\, \mathrm{d}\mu = \frac{C}{\sigma^n} \exp\left(-\frac{V}{2\sigma^2}\right). \tag{6.11}$$

Integrating this equation over σ to calculate the normalisation constant C we find the **chi-squared distribution** for $\nu = n - 1$ degrees of freedom

$$p(\sigma|V, n) = 2\left(\frac{V}{2}\right)^{(n-1)/2} \frac{1}{\Gamma[(n-1)/2]} \frac{1}{\sigma^n} \exp\left(-\frac{V}{2\sigma^2}\right). \tag{6.12}$$

This distribution is shown for $V = 4$ and $n = 4$ in Fig. 6c.

Exercise 6.4: Transform $\chi^2 = V/\sigma^2$ and show that (6.12) can be written in the more familiar form with $\nu = n - 1$ as the only parameter

$$p(\sigma|V, n)\, \mathrm{d}\sigma \to p(\chi^2|\nu)\, \mathrm{d}\chi^2 = \frac{\left(\chi^2\right)^{\alpha-1} \exp(-\frac{1}{2}\chi^2)}{2^\alpha \Gamma(\alpha)}\, \mathrm{d}\chi^2,$$

where $\alpha = \nu/2$. Use the definition of the Gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, \mathrm{d}t$$

and the property $\Gamma(z + 1) = z\Gamma(z)$ to show that the mean and variance of the χ^2 distribution are given by ν and 2ν, respectively.

Exercise 6.5: Show by expanding the negative logarithm of (6.11) that the maximum of the posterior is located at $\hat{\sigma} = \sqrt{V/n}$.

I can be shown that the sampling distribution of the random variable V is also χ^2 distributed with $\nu = n - 1$ degrees of freedom; the proof is somewhat lengthy and we refer for this to standard textbooks. Using the result $<\chi^2>= \nu$, we find

$$<\frac{V}{\sigma^2}>= \frac{<V>}{\sigma^2} = n - 1 \qquad \Rightarrow \qquad \sigma^2 = \frac{<V>}{n-1}. \tag{6.13}$$

This justifies the definition of the sample variance $S^2 \equiv V/(n - 1)$, mentioned above, because its distribution has a mean value of σ^2.

6.2 Bayesian versus Frequentist inference (II)

In the previous section we have seen how to conduct Bayesian inference on the unknown mean (and/or width) of a Gaussian sampling distribution. At this point it is interesting

to compare this to inference done in the Frequentist way. For this comparison we will reduce the inference problem to one which is as simple as possible.

Suppose we have done *one* measurement $x = 3$ drawn from a Gaussian sampling distribution $p(x|\mu)$ with an unknown mean μ and a known width $\sigma = 5$. Given the observation $x = 3$, we now ask what is the best value of μ and its uncertainty. Assuming a uniform prior, Bayes theorem gives for the posterior

$$p(\mu|x)\mathrm{d}\mu = C\; p(x|\mu)\; p(\mu)\mathrm{d}\mu = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{d}\mu, \tag{6.14}$$

which is shown in Fig. 7a for $x = 3$. This plot is then the full result of the Bayesian

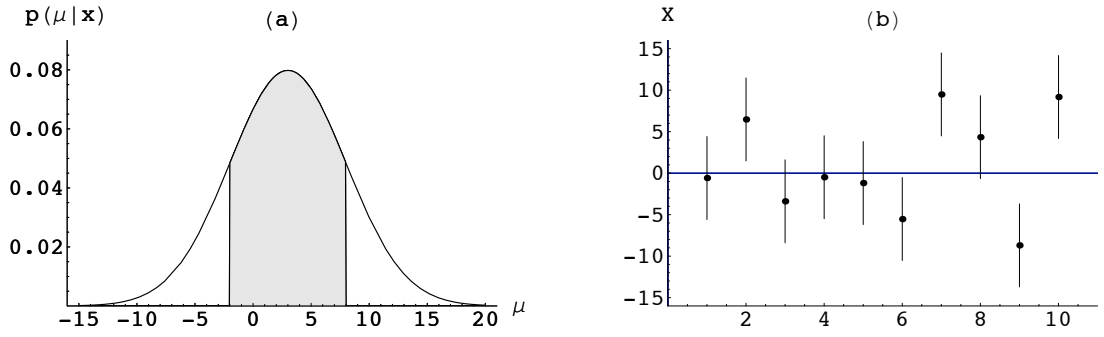


Figure 7: (a) Bayesian inference: The posterior distribution $p(\mu|x)$ for one measurement $x = 3$ drawn from a Gaussian distribution with unknown mean μ and known width $\sigma = 5$. The shaded area indicates the $\pm 1\sigma$ credible interval containing 68.3% probability. (b) Frequentist inference: confidence intervals $x \pm \sigma$ (vertical bars) of the first ten measurements drawn from a Gaussian distribution with $\mu = 0$ (unknown) and $\sigma = 5$ (known). Five out of ten confidence intervals contain the unknown value of μ. The limit for an infinite number of measurements is that 68.3% of the intervals contain μ.

inference which we may summarise, if we wish, by giving the $\pm 1\sigma$ **credible interval**[27] as

$$\mu = 3 \pm 5 \;\; (68.3\%\ \mathrm{CL}) \qquad \text{or as} \qquad P(-2 < \mu < 8) = 68.3\%. \tag{6.15}$$

In Frequentist statistics the outcome of the measurement is described by a **random variable** which we denote here by the capital letter X. This random variable takes on different values $(x_1, x_2, \ldots)$ at every repetition of the experiment according to some known sampling distribution, which, in this example, is our Gaussian with unknown mean μ and known $\sigma = 5$. It is important to note that X is a random variable but that the realisations $(x_1, x_2, \ldots)$ of X are *not* random variables.

Now to make an inference on μ from the observations we first have to construct a **statistic** which is a function of the data and which is in one way or another related to μ. This statistic is thus also a random variable with a known distribution which can be derived from the sampling distribution of the data. A convenient statistic for the inference on μ is the sample mean $\bar{X}$ which reduces in our case (only one measurement) to X itself.

[27] Intervals defined on Bayesian posteriors are called *credible* intervals to distinguish them from *confidence* intervals which are defined on the likelihood or sampling distribution and play a very specific role in Frequentist statistics.

Once an appropriate statistic is defined, various probability statements can be made like, for our statistic X,

$$P(\mu - \sigma < X < \mu + \sigma) = 68.3\%.$$

Subtracting μ and X in the inequality above and then reversing the signs we can re-write this as

$$P(X - \sigma < \mu < X + \sigma) = 68.3\%. \tag{6.16}$$

Having observed $x = 3$ it is tempting substitute this in (6.16) to obtain

$$P(x - \sigma < \mu < x + \sigma) = P(-2 < \mu < 8) = 68.3\% \tag{6.17}$$

which is, by the way, the same statement as given in (6.15).

But in the Frequentist world (6.17) does not make sense!

This follows immediately from the fact that Frequentist probabilistic statements can only be made about random variables but the argument of P in (6.17) does not contain any random variable. A little thought will reveal that accepting (6.17) would allow for probability inversion *without* using Bayes' theorem or even specifying a prior. This clearly violates the elementary rules of probability calculus and thus the desiderata of plausible inference as given in Section 2.1.

Thus, as a Frequentist we cannot go beyond (6.16) which says, in fact, that the **confidence interval** $[x - \sigma, x + \sigma]$ associated with each possible measurement x will contain the unknown μ in 68.3% of the cases in the limit of an infinite amount of repetitions of the experiment. This fact is illustrated in Fig. 7b for the first 10 measurements. The property that the unknown μ is contained in a given fraction of the confidence intervals is called **coverage**. This pre-defined fraction (68.3% in our case) is called a **confidence level** (CL).[28] For more on Frequentist interval estimation we refer to [Cowan98, Ch. 9], [James06, Ch. 9], and discussions in [James00] or [Zech02].

Before we enter again the Bayesian world, let us wrap-up this section with the remark that a statement like

$$\mu = 3 \pm 5 \qquad (68.3\%\ \text{CL})$$

has quite a different meaning in Bayesian and Frequentist inference, and that

$$P(-2 < \mu < 8) = 68.3\%$$

is a statement which does not exist in the Frequentist world.

6.3 Maximum likelihood and least squares

In this section we consider the case that the data can be described by a function $f(x; \boldsymbol{\theta})$ of a variable x depending on a set of parameters $\boldsymbol{\theta}$. For simplicity we will consider only functions of one variable x; the extension to more dimensions is trivial. Suppose that

[28] For a given distribution, there is an infinity of confidence intervals corresponding to a given confidence level. The interval can then be fixed by additional criteria like requiring equal probability $(1 - \text{CL})/2$ in the left- and right-hand tails, or taking the shortest possible interval, *etc.*

we have made a series of measurements $\{d_i\}$ at the sample points $\{x_i\}$ and that each measurement is distributed according to some sampling distribution $p_i(d_i|\mu_i, \sigma_i)$. Here μ_i and σ_i characterise the position and the width of the sampling distribution p_i of data point d_i. We parametrise the positions μ_i by the function f:

$$\mu_i(\boldsymbol{\theta}) = f(x_i; \boldsymbol{\theta}).$$

If the measurements are *independent* we can write for the likelihood

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) = \prod_{i=1}^{n} p_i[d_i|\mu_i(\boldsymbol{\theta}), \sigma_i].$$

Introducing the somewhat more compact notation $p_i(d_i|\boldsymbol{\theta}, I)$ for the sampling distributions, we write for the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{d}, I) = C \left(\prod_{i=1}^{n} p_i(d_i|\boldsymbol{\theta}, I) \right) p(\boldsymbol{\theta}|I) \tag{6.18}$$

where C is a normalisation constant and $p(\boldsymbol{\theta}|I)$ is the joint prior distribution of the parameters $\boldsymbol{\theta}$. The position and width of the posterior can be found by minimising

$$L(\boldsymbol{\theta}) = -\ln[p(\boldsymbol{\theta}|\boldsymbol{d}, I)] = -\ln(C) - \ln[p(\boldsymbol{\theta}|I)] - \sum_{i=1}^{n} \ln[p_i(d_i|\boldsymbol{\theta}, I)] \tag{6.19}$$

as described in Section 3.1, equations (3.6)–(3.9). In practice this is often done numerically by presenting $L(\boldsymbol{\theta})$ to a minimisation program like MINUIT. Note that this procedure can be carried out for any sampling distribution p_i be it Binomial, Poisson, Cauchy, Gauss or whatever. In case the prior $p(\boldsymbol{\theta}|I)$ is chosen to be uniform, the second term in (6.19) is constant so that the maximum of the posterior coincides with the maximum of the likelihood. The procedure is then called a **maximum likelihood fit**.

The most common case encountered in data analysis is when the sampling distributions are Gaussian. For a uniform prior, (6.19) reduces to

$$L(\boldsymbol{\theta}) = \text{constant} + \tfrac{1}{2}\chi^2 = \text{constant} + \tfrac{1}{2} \sum_{i=1}^{n} \left[\frac{d_i - \mu_i(\boldsymbol{\theta})}{\sigma_i} \right]^2. \tag{6.20}$$

We then speak of **$\boldsymbol{\chi^2}$ minimisation** or **least squares minimisation**. When the function $f(x; \boldsymbol{\theta})$ is *linear* in the parameters, the minimisation can be reduced to a single matrix inversion, as we will now show.

A function which is linear in the parameters can generically be expressed by

$$f(x; \boldsymbol{\theta}) = \sum_{\lambda=1}^{m} \theta_\lambda f_\lambda(x) \tag{6.21}$$

where the f_λ are a set of functions of x and the θ_λ are the coefficients to be determined from the data. We denote by $w_i \equiv 1/\sigma_i^2$ the weight of each data point and write for the log posterior

$$L(\boldsymbol{\theta}) = \tfrac{1}{2} \sum_{i=1}^{n} w_i \left[d_i - f(x_i; \boldsymbol{\theta}) \right]^2 = \tfrac{1}{2} \sum_{i=1}^{n} w_i \left[d_i - \sum_{\lambda=1}^{m} \theta_\lambda f_\lambda(x_i) \right]^2. \tag{6.22}$$

The mode $\hat{\boldsymbol{\theta}}$ is found by setting the derivative of L to all parameters to zero

$$\frac{\partial L(\hat{\boldsymbol{\theta}})}{\partial \theta_\mu} = -\sum_{i=1}^{n} w_i \left[d_i - \sum_{\lambda=1}^{m} \hat{\theta}_\lambda f_\lambda(x_i) \right] f_\mu(x_i) = 0. \tag{6.23}$$

We can write this equation in vector notation as

$$\boldsymbol{b} = \boldsymbol{W}\hat{\boldsymbol{\theta}} \qquad \text{so that} \qquad \hat{\boldsymbol{\theta}} = \boldsymbol{W}^{-1}\boldsymbol{b} \tag{6.24}$$

where the (symmetric) matrix $\boldsymbol{W}$ and the vector $\boldsymbol{b}$ are given by

$$W_{\lambda\mu} = \sum_{i=1}^{n} w_i \, f_\lambda(x_i) f_\mu(x_i) \qquad \text{and} \qquad b_\mu = \sum_{i=1}^{n} w_i \, d_i \, f_\mu(x_i). \tag{6.25}$$

Differentiating (6.22) twice to θ_λ yields an expression for the Hessian

$$H_{\lambda\mu} = \frac{\partial^2 L(\hat{\boldsymbol{\theta}})}{\partial \theta_\lambda \partial \theta_\mu} = W_{\lambda\mu}. \tag{6.26}$$

Higher derivatives vanish so that the quadratic expansion (3.8) in Section 3.1 is exact.

To summarise, when the function to be fitted is linear in the parameters we can build a vector $\boldsymbol{b}$ and a matrix $\boldsymbol{W}$ as defined by (6.25). The posterior (assuming uniform priors) is then a multivariate Gaussian with mean (or mode) $\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} = \boldsymbol{W}^{-1}\boldsymbol{b}$ and covariance matrix $\boldsymbol{V} \equiv \boldsymbol{H}^{-1} = \boldsymbol{W}^{-1}$. In this way, a fit to the data is reduced to one matrix inversion and does not need starting values for the parameters, nor iterations, nor convergence criteria.

Exercise 6.6: Calculate the matrix $\boldsymbol{W}$ and the vector $\boldsymbol{b}$ of (6.25) for a polynomial parametrisation of the data

$$f(x; \boldsymbol{a}) = a_1 + a_2 x + a_3 x^2 + a_4 x^3 + \cdots .$$

Exercise 6.7: Show that a fit to a constant results in the **weighted average** of the data

$$\hat{a}_1 = \frac{\sum_i w_i d_i}{\sum_i w_i} \pm \frac{1}{\sqrt{\sum_i w_i}}.$$

6.4 Correlated data errors

Data are often subject to sources of uncertainty which cause a simultaneous fluctuation of more than one data point. We will call these correlated uncertainties **systematic**, in contrast to point to point un-correlated errors, which we will call **statistical**.

To propagate the systematic uncertainties to the parameters $\boldsymbol{\theta}$ of interest one often offsets the data by each systematic error in turn, redo the analysis, and then add the deviations from the optimal values $\hat{\boldsymbol{\theta}}$ in quadrature. Such an intuitive *ad hoc* procedure (**offset method**) has no sound theoretical foundation and may even spoil your result by assigning errors which are far too large, see [Alekh00] for an illustrative example and

also Exercise 6.10 below. The method of systematic error propagation described below is taken from the CTEQ group [Stump02] who used it to propagate the experimental systematic errors in their global QCD analyses of deep inelastic and collider data.

To take systematic errors into account we include them in the data model. How to do this properly, depends of course on the experiment; here we will restrict ourselves to a linear parametrisation with Gaussian distributed parameters which has the advantage that it is easily incorporated in a least squares minimisation procedure. This model, as it stands, does not handle asymmetric errors. However, in case we deal with several systematic sources these asymmetries tend to vanish by virtue of the *central limit theorem*.

In Fig. 8 we show a systematic distortion of a set of data points

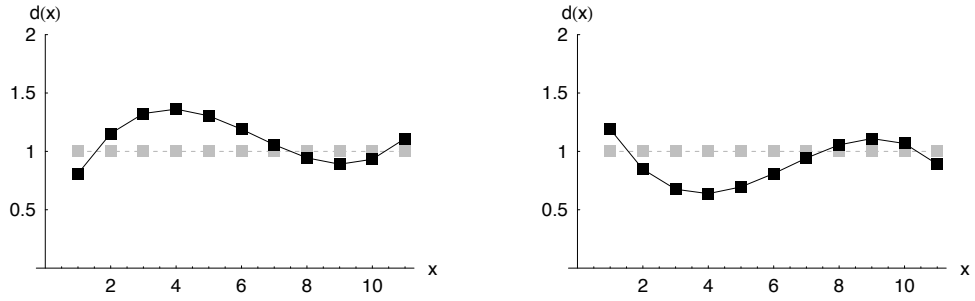


Figure 8: Systematic distortion (black symbols) of a set of data points (grey symbols) for two values of the interpolation parameter $s = +1$ (left) and $s = -1$ (right).

$$d_i \to d_i + s\Delta_i.$$

Here Δ_i is a list of systematic deviations and s is an interpolation parameter which controls the amount of systematic shift applied to the data. Usually there are several sources of systematic error stemming from uncertainties in the detector calibration, acceptance, efficiency and so on. For m such sources, the data model can be written as

$$d_i = t_i(\boldsymbol{\theta}) + r_i + \sum_{\lambda=1}^{m} s_\lambda \Delta_{i\lambda}, \tag{6.27}$$

where $t_i(\boldsymbol{\theta}) = f(x_i;\boldsymbol{\theta})$ is the theory prediction containing the parameters $\boldsymbol{\theta}$ of interest and $\Delta_{i\lambda}$ is the correlated error on point i stemming from source λ. In (6.27), the uncorrelated statistical fluctuations of the data are described by the independent Gaussian random variables r_i of zero mean and variance σ_i^2. The s_λ are independent Gaussian random variables of zero mean and unit variance which account for the systematic fluctuations. The joint distribution of $\boldsymbol{r}$ and $\boldsymbol{s}$ is thus given by

$$p(\boldsymbol{r},\boldsymbol{s}|I) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{r_i^2}{2\sigma_i^2}\right) \prod_{\lambda=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s_\lambda^2}{2}\right).$$

The covariance matrix of this joint distribution is trivial:

$$< r_i r_j > = \sigma_i^2\, \delta_{ij} \qquad < s_\lambda s_\mu > = \delta_{\lambda\mu} \qquad < r_i s_\lambda > = 0. \tag{6.28}$$

Because the data are a linear combination of the Gaussian random variables $\boldsymbol{r}$ and $\boldsymbol{s}$, it follows that $\boldsymbol{d}$ is also Gaussian distributed

$$p(\boldsymbol{d}|I) = \frac{1}{\sqrt{(2\pi)^n\,|\boldsymbol{V}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{d}-\bar{\boldsymbol{d}})\boldsymbol{V}^{-1}(\boldsymbol{d}-\bar{\boldsymbol{d}})\right]. \tag{6.29}$$

The mean $\bar{\boldsymbol{d}}$ is found by taking the average of (6.27)

$$\bar{d}_i = <d_i> = t_i(\boldsymbol{\theta}) + <r_i> + \sum_{\lambda=1}^{m} <s_\lambda>\,\Delta_{i\lambda} = t_i(\boldsymbol{\theta}). \tag{6.30}$$

A transformation of (6.28) by linear error propagation (see Section 3.2) gives for the covariance matrix

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & \sigma_2^2 + S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \cdots & \sigma_n^2 + S_{nn} \end{pmatrix} \quad \text{with} \quad S_{ij} = \sum_{\lambda=1}^{m} \Delta_{i\lambda}\Delta_{j\lambda}. \tag{6.31}$$

Exercise 6.8: Use the propagation formula (3.20) in Section 3.2 to derive (6.31) from (6.27) and (6.28).

It is more easy to calculate this covariance matrix by directly averaging the product $\Delta d_i \Delta d_j$. Because all the cross terms vanish by virtue of (6.28) we immediately obtain

$$\begin{aligned} V_{ij} = <\Delta d_i \Delta d_j> &= <(r_i + \sum_\lambda s_\lambda \Delta_{i\lambda})(r_j + \sum_\lambda s_\lambda \Delta_{j\lambda})> \\ &= <r_i r_j> + \sum_\lambda \sum_\mu \Delta_{i\lambda}\Delta_{j\mu} <s_\lambda s_\mu> + \cdots \\ &= \sigma_i^2 \delta_{ij} + \sum_\lambda \Delta_{i\lambda}\Delta_{j\lambda} \end{aligned}$$

which is the same as given in (6.31).

Inserting (6.30) in (6.29) and assuming a uniform prior, the log posterior of the parameters $\boldsymbol{\theta}$ can be written as

$$L(\boldsymbol{\theta}) = -\ln[p(\boldsymbol{\theta}|\boldsymbol{d})] = \text{Constant} + \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[d_i - t_i(\boldsymbol{\theta})]V_{ij}^{-1}[d_j - t_j(\boldsymbol{\theta})]. \tag{6.32}$$

Minimising L defined by (6.32) is impractical because we need the inverse of the covariance matrix (6.31) which can become very large. Furthermore, when the systematic errors dominate, the matrix might—numerically—be uncomfortably close to a matrix with the simple structure $V_{ij} = \Delta_i \Delta_j$, which is singular.

Fortunately, (6.32) can be cast into an alternative form which avoids the inversion of large matrices [Stump02]. Our derivation of this result is based on the standard steps taken in a Bayesian inference: (i) use the data model to write an expression for the likelihood; (ii) define prior probabilities; (iii) calculate posterior probabilities with Bayes' theorem and (iv) integrate over the nuisance parameters.

The likelihood $p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{s})$ is calculated from the expansion in the variables $\boldsymbol{r}$

$$p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{s}) = \int \mathrm{d}\boldsymbol{r}\, p(\boldsymbol{d},\boldsymbol{r}|\boldsymbol{\theta},\boldsymbol{s}) = \int \mathrm{d}\boldsymbol{r}\, p(\boldsymbol{d}|\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{s})\, p(\boldsymbol{r}|\boldsymbol{\theta},\boldsymbol{s}) \tag{6.33}$$

The data model (6.27) is incorporated through the trivial assignment

$$p(\boldsymbol{d}|\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{s}) = \prod_{i=1}^{n} \delta[r_i + t_i(\boldsymbol{\theta}) + \sum_{\lambda=1}^{m} s_\lambda \Delta_{i\lambda} - d_i]. \tag{6.34}$$

As already discussed above, the distribution of $\boldsymbol{r}$ is taken to be

$$p(\boldsymbol{r}|\boldsymbol{\theta},\boldsymbol{s}) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{r_i^2}{2\sigma_i^2}\right). \tag{6.35}$$

Inserting (6.34) and (6.35) in (6.33) yields, after integration over $\boldsymbol{r}$,

$$p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{s}) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(\frac{d_i - t_i(\boldsymbol{\theta}) - \sum_\lambda s_\lambda \Delta_{i\lambda}}{\sigma_i}\right)^2\right]. \tag{6.36}$$

Assuming a Gaussian prior for $\boldsymbol{s}$

$$p(\boldsymbol{s}|I) = (2\pi)^{-m/2} \prod_{\lambda=1}^{m} \exp(-\tfrac{1}{2}s_\lambda^2)$$

and a uniform prior for $\boldsymbol{\theta}$, the joint posterior distribution can be written as

$$p(\boldsymbol{\theta},\boldsymbol{s}|\boldsymbol{d}) = C \exp\left[-\tfrac{1}{2}\sum_{i=1}^{n} w_i \left(d_i - t_i(\boldsymbol{\theta}) - \sum_{\lambda=1}^{m} s_\lambda \Delta_{i\lambda}\right)^2 - \tfrac{1}{2}\sum_{\lambda=1}^{m} s_\lambda^2\right] \tag{6.37}$$

where $w_i = 1/\sigma_i^2$. The log posterior $L = -\ln p$ can now numerically be minimised (for instance by MINUIT) with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{s}$. Marginalisation of the nuisance parameters $\boldsymbol{s}$, as described in Section 3.1, then yields the desired result. Clearly we now got rid of our large covariance matrix (6.31) at the expense of extending the parameter space from $\boldsymbol{\theta}$ to $\boldsymbol{\theta} \times \boldsymbol{s}$. In a global data analysis where many experiments are combined, the number of systematic sources $\boldsymbol{s}$ can become quite large so that minimising L of (6.37) may still not be very attractive.

However, the fact that L is *linear* in $\boldsymbol{s}$ allows us to analytically carry out the minimisation and marginalisation with respect to $\boldsymbol{s}$. For this, we expand L like in (3.6) but only to $\boldsymbol{s}$ and not to $\boldsymbol{\theta}$ (it is easy to show that this expansion is exact *i.e.* that higher derivatives in s_λ vanish):

$$L(\boldsymbol{\theta},\boldsymbol{s}) = L(\boldsymbol{\theta},\hat{\boldsymbol{s}}) + \sum_\lambda \frac{\partial L(\boldsymbol{\theta},\hat{\boldsymbol{s}})}{\partial s_\lambda}\Delta s_\lambda + \frac{1}{2}\sum_\lambda\sum_\mu \frac{\partial^2 L(\boldsymbol{\theta},\hat{\boldsymbol{s}})}{\partial s_\lambda \partial s_\mu}\Delta s_\lambda \Delta s_\mu.$$

Solving the equations $\partial L/\partial s_\mu = 0$ we find

$$L(\boldsymbol{\theta},\boldsymbol{s}) = L(\boldsymbol{\theta},\hat{\boldsymbol{s}}) + \tfrac{1}{2}(\boldsymbol{s}-\hat{\boldsymbol{s}})\boldsymbol{S}(\boldsymbol{s}-\hat{\boldsymbol{s}}) \tag{6.38}$$

with

$$\begin{aligned} \hat{\boldsymbol{s}} &= \boldsymbol{S}^{-1}\boldsymbol{b} \\ S_{\lambda\mu} &= \delta_{\lambda\mu} + \sum_{i=1}^{n} w_i \Delta_{i\lambda}\Delta_{i\mu} \\ b_\lambda(\boldsymbol{\theta}) &= \sum_{i=1}^{n} w_i [d_i - t_i(\boldsymbol{\theta})]\Delta_{i\lambda}. \end{aligned} \tag{6.39}$$

and

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{s}}) = \tfrac{1}{2}\sum_{i=1}^{n} w_i \left(d_i - t_i(\boldsymbol{\theta}) - \sum_{\lambda=1}^{m} \hat{s}_\lambda \Delta_{i\lambda} \right)^2 - \tfrac{1}{2}\sum_{\lambda=1}^{m} \hat{s}_\lambda^2. \tag{6.40}$$

Exercise 6.9: Derive the equations (6.38)–(6.40).

Taking the exponent of (6.38) gives for the posterior

$$p(\boldsymbol{\theta}, \boldsymbol{s}|\boldsymbol{d}) = C \exp\left[-L(\boldsymbol{\theta}, \hat{\boldsymbol{s}})\right] \exp\left[-\tfrac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{s}})\boldsymbol{S}(\boldsymbol{s} - \hat{\boldsymbol{s}})\right]$$

which yields upon integrating over the nuisance parameters $\boldsymbol{s}$

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = C' \exp\left[-L(\boldsymbol{\theta}, \hat{\boldsymbol{s}})\right] \tag{6.41}$$

The log posterior (6.40) can now numerically be minimised with respect to the parameters $\boldsymbol{\theta}$. Instead of the $n \times n$ covariance matrix $\boldsymbol{V}$ of (6.31) only an $m \times m$ matrix $\boldsymbol{S}$ has to be inverted with m the number of systematic sources.

The solution (6.39) for $\hat{\boldsymbol{s}}$ can be substituted back into (6.40). This leads after straight forward algebra to the following very compact and elegant representation of the posterior [Stump02]

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = C' \exp\left\{ -\tfrac{1}{2}\left[\sum_{i=1}^{n} w_i (d_i - t_i)^2 - \boldsymbol{b}\, \boldsymbol{S}^{-1} \boldsymbol{b} \right] \right\}. \tag{6.42}$$

The first term in the exponent is the usual χ^2 in absence of correlated errors while the second term takes into account the systematic correlations. Note that $\boldsymbol{S}$ does not depend on $\boldsymbol{\theta}$ so that $\boldsymbol{S}^{-1}$ can be computed once and for all. The vector $\boldsymbol{b}$, on the other hand, does depend on $\boldsymbol{\theta}$ so that it has to be recalculated at each minimisation step.

Although the posteriors defined by (6.32) and (6.42) look very different, it can be shown (tedious algebra) that they are mathematically identical [Stump09]. In other words, minimising (6.32) leads to the same result for $\hat{\boldsymbol{\theta}}$ as minimising the negative logarithm of (6.42).

It is clear that the uncertainty on parameters derived from a given data sample should decrease when new data are added to the sample. This is because additional data will always *increase* the available information even when these data are very uncertain. From this it follows immediately that the error obtained from an analysis of the total sample can never be larger than the error obtained from an analysis of any sub-sample of the data. It turns out that error estimates based on equations (6.39)–(6.42) do meet this requirement but that the offset method—mentioned in the beginning of this section—does *not*. This issue is investigated further in the following exercise.

Exercise 6.10: We make n measurements $d_i \pm \sigma_i$ of the temperature in a room. The measurements have a common systematic offset error Δ. Calculate the best estimate $\hat{\mu}$ of the temperature and the total error (statistical $\oplus$ systematic) by: (i) using the offset method mentioned at the beginning of this section and (ii) using (6.42). To simplify the algebra assume that all data and errors have the same value: $d_i = d$, $\sigma_i = \sigma$.

A second set of n measurements is added using another thermometer which has the same resolution σ but *no* offset uncertainty. Calculate the best estimate $\hat{\mu}$ and the total error from both data sets using either the offset method or (6.42). Again, assume that $d_i = d$ and $\sigma_i = \sigma$ to simplify the algebra.

Now let $n \to \infty$. Which error estimate makes sense in this limit and which does not?

The systematic errors described in this section were treated as *offsets*. Another important class are scale or normalisation errors which should be treated somewhat differently because they affect not only the position but also the width of the sampling distributions. We will come back to this in the next section.

7 A Few Examples

In this section we work-out a few simple cases of parameter estimation which are often encountered in data analysis but cannot be solved by out-of-the-box least square minimisation. These examples are the fitting of a signal drowned in background, the fitting of a sparsely filled histogram, the treatment of normalisation uncertainties and the accounting for poorly known experimental errors. In Section 7.5 we describe how to treat data with errors on both the x and y co-ordinate. That section is particularly interesting since it provides an example of somewhat more advanced Bayesian reasoning.

7.1 Signal drowned in background

In this section we describe a typical case where the likelihood prefers a negative value for a positive definite quantity. Defining a confidence interval in such a case is not so obvious in Frequentist statistics. But in our Bayesian approach the solution is trivial as is illustrated by the following example where a negative counting rate is found after background subtraction.

A search was made by the NA49 experiment at the CERN SPS for D^0 production in a large sample of 4×10^6 Pb-Pb collisions at a beam energy of 158 GeV per nucleon [Alt06]. Since NA49 does not have secondary vertex reconstruction capabilities, all pairs of positive and negative tracks in the event were accumulated in invariant mass spectra assuming that the tracks originate from the decays $D^0 \to K^-\pi^+$ or $\bar{D}^0 \to K^+\pi^-$. In the left-hand side plot of Fig. 9 we show the invariant mass spectrum of the D^0 candidates. The vertical lines indicate a ± 90 MeV window around the nominal D^0 mass. The large combinatorial background is due to a multiplicity of approximately 1400 charged tracks per event giving, for each event, about 5×10^5 entries in the histogram. In the right-hand side plot we show the invariant mass spectrum after background subtraction. Clearly no signal is observed.

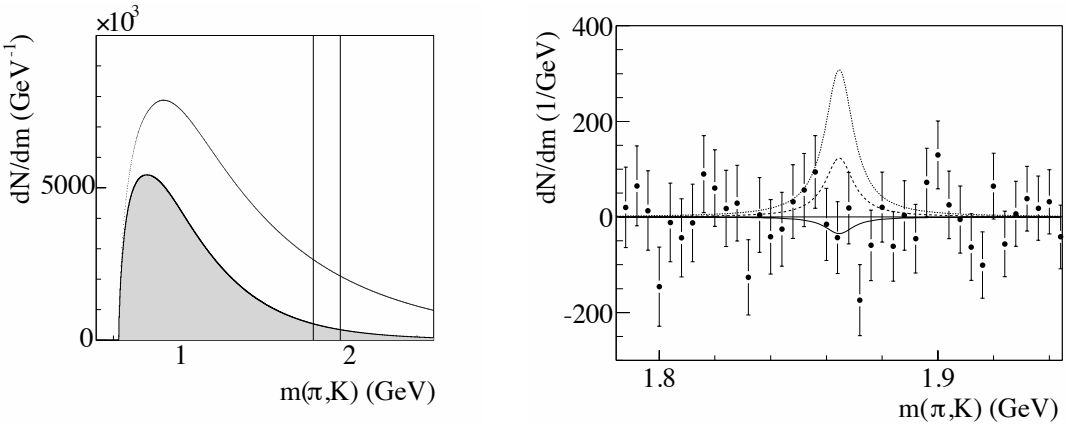


Figure 9: Left: The invariant mass distribution of D^0 candidates in 158A GeV Pb-Pb collisions at the CERN SPS. The open (shaded) histograms are before (after) applying cuts to improve the significance. The vertical lines indicate a ±90 MeV window around the nominal D^0 mass. Right: The $D^0 + \bar{D}^0$ invariant mass spectrum after background subtraction. The full curve is a fit to the data assuming a fixed signal shape. The other curves correspond to model predictions of the D^0 yield.

A least squares fit to the data of a Cauchy line shape (with fixed position and FWHM but free normalisation n) on top of a polynomial background yielded a *negative* value for the yield $\hat{n}(D^0 + \bar{D}^0) = -0.36 \pm 0.74$ per event as is shown by the full curve in the right-hand side plot of Fig. 9. As already mentioned above, this is a typical example of a case where the likelihood favours an outcome which is unphysical.

To calculate an upper limit on the D^0 yield, Bayesian inference is used as follows. In the fit to the invariant mass spectrum, the parameters $\boldsymbol{\theta} = (n, \boldsymbol{a})$ are introduced where n is the D^0 yield of interest, and $\boldsymbol{a}$ are the background shape (nuisance) parameters. The likelihood of the data $\boldsymbol{d}$ is now written as a multivariate Gaussian in parameter space:

$$p(\boldsymbol{d}|\boldsymbol{\theta}, I) = \frac{1}{\sqrt{(2\pi)^p\,|\boldsymbol{V}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\boldsymbol{V}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \tag{7.1}$$

where p is the number of parameters and $\hat{\boldsymbol{\theta}}$ and V are the best values and covariance matrix as obtained from MINUIT, see also Section 3.1. Taking flat priors for the background parameters, but not for the yield n

$$p(\boldsymbol{\theta}|I) = p(n, \boldsymbol{a}|I) \propto p(n|I),$$

leads to the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{d}, I) = \frac{C}{\sqrt{(2\pi)^p\,|\boldsymbol{V}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\boldsymbol{V}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] p(n|I) \tag{7.2}$$

where C is a normalisation constant and $p(n|I)$ is the prior for the D^0 yield n. The posterior for n is now obtained by integrating (7.2) over the background parameters $\boldsymbol{a}$. As explained in Section 3.1 this yields a one-dimensional Gauss with a variance given by the diagonal element $\sigma^2 = V_{nn}$ of the covariance matrix. Thus we have

$$p(n|\boldsymbol{d}, I) = \frac{C}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{n - \hat{n}}{\sigma}\right)^2\right] p(n|I) = C\,\mathcal{N}(n; \hat{n}, \sigma)\,p(n|I) \tag{7.3}$$

where we have introduced the short-hand notation $\mathcal{N}()$ for a one-dimensional Gaussian distribution. In (7.3), $\hat{n} = -0.36$ and $\sigma = 0.74$ as obtained from the fit to the data.

As a last step we encode in the prior $p(n|I)$ our knowledge that n is positive definite

$$P(n|I) \propto \theta(n) \qquad \text{with} \qquad \theta(n) = \begin{cases} 0 & \text{for } n < 0 \\ 1 & \text{for } n \geq 0. \end{cases} \tag{7.4}$$

Inserting (7.4) in (7.3) and integrating over n to calculate the constant C we find

$$p(n|\boldsymbol{d}, I) = \mathcal{N}(n; \hat{n}, \sigma)\, \theta(n) \left[\int_0^\infty \mathcal{N}(n; \hat{n}, \sigma)\, \mathrm{d}n \right]^{-1}. \tag{7.5}$$

The posterior distribution is thus a truncated Gaussian with mean and variance as obtained from the fit. This Gaussian is set to zero for $n < 0$ and re-normalised to unity for $n \geq 0$. The upper limit ($n_{\max}$), corresponding to a given confidence level (CL) is then calculated by (numerically) solving the equation

$$\mathrm{CL} = \int_0^{n_{\max}} p(n|\boldsymbol{d}, I)\, \mathrm{d}n = \int_0^{n_{\max}} \mathcal{N}(n; \hat{n}, \sigma)\, \mathrm{d}n \left[\int_0^\infty \mathcal{N}(n; \hat{n}, \sigma)\, \mathrm{d}n \right]^{-1}. \tag{7.6}$$

Using the numbers quoted above this gives $n(\mathrm{D}^0 + \bar{\mathrm{D}}^0) < 1.5$ per event at 98% CL.

7.2 Sparsely populated histogram

As stated in Section 6.3, a χ^2 minimisation is only valid when the data are Gaussian distributed. Another important requirement—which is easily forgotten—is that the fit model should not affect the data errors.[29] The results of a χ^2 analysis can be seriously biased when these requirements are not met, as is the case in the following simple example.

In Fig. 10 we show a histogram filled with a uniform background. There are 10 counts observed in 50 bins so that the average background rate is $R = 10/50 = 0.2$ counts per bin. However, a least squares fit (with ROOT) to a zero-degree polynomial (that is, to a constant), yields $\hat{R} = 1.06$, as is shown by the full line in the plot. The reason for this biased result is twofold. First, the counts in a bin are Poisson distributed, and not Gaussian. Second, assuming Poisson errors, the program minimises[30]

$$\chi^2 = \sum_{n_i > 0} \frac{(n_i - R)^2}{n_i}.$$

This χ^2 discards empty bins which still carry a lot of information.

The proper way to analyse these data is to start from the correct posterior. Using a flat prior for $R \geq 0$ we find

$$p(R|\boldsymbol{n}, I) \propto \prod_i \frac{R^{n_i}}{n_i!}\, e^{-R} \quad \text{and} \quad L = -\ln(p) = \sum_i \left[R - n_i \ln(R) \right] + \text{constant}. \tag{7.7}$$

[29] See (7.22) below, where the errors *are* affected.

[30] By default, ROOT minimises a χ^2 but there is an option to use the log likelihood (7.7).

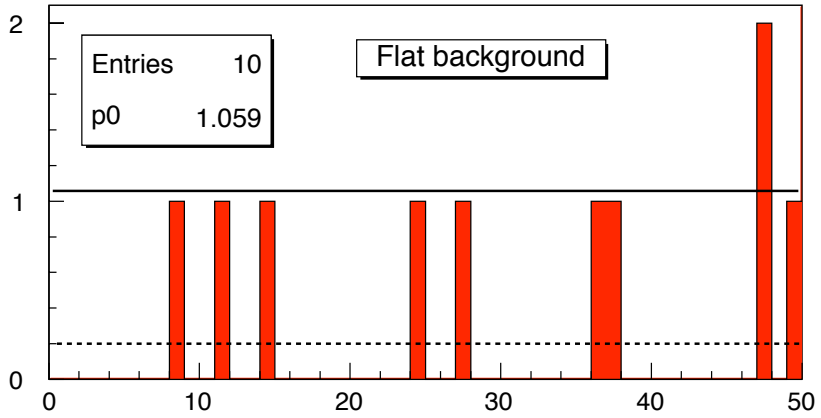


Figure 10: Sparse histogram filled with a uniform distribution. The full line shows the result of a χ^2 fit to the data. The dashed line corresponds to the likelihood fit described in the text.

The mode $\hat{R}$ is found by setting the derivative to R of the log posterior to zero:

$$\frac{\mathrm{d}L}{\mathrm{d}R} = \sum_i \left(1 - \frac{n_i}{R}\right) = 0 \qquad \text{whence} \qquad \hat{R} = \frac{\sum n_i}{n_{\text{bins}}} = \frac{10}{50} = 0.2,$$

which is the correct answer. In Fig. 11 is shown the posterior distribution (7.7), together

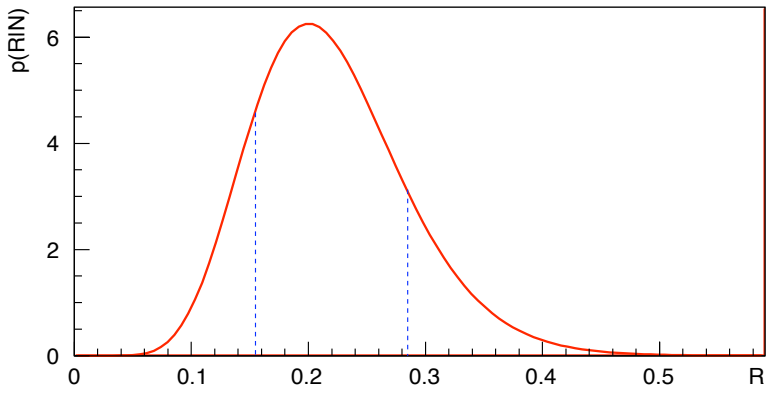


Figure 11: The posterior distribution of the background rate R (counts per bin) for the data shown in Fig. 10. The region between the vertical lines (16% and 84% quantiles) contains 68% probability.

with the 16% and 84% quantiles which happen to be located at $R = 0.155$ and 0.285, respectively. Thus we may summarise our result as

$$R = 0.20^{+0.08}_{-0.04} \qquad (68\%\ \text{CL}) \qquad \text{or as} \qquad P(0.16 < R < 0.28) = 0.68.$$

The lesson to be learnt here is that one should never start from some *ad hoc* χ^2 but always from the likelihood which, when multiplied by the prior, gives the posterior. The log posterior can then be analysed using programs like MINUIT, for instance. However, one should be aware of a little catch here: the parameter errors (inverse of the Hessian) calculated by MINUIT correspond to

$$\chi^2(\hat{\theta} + \Delta\theta) = \chi^2(\hat{\theta}) + 1 \qquad \text{but for } L \text{ we have} \qquad L(\hat{\theta} + \Delta\theta) = L(\hat{\theta}) + \tfrac{1}{2}.$$

Exercise 7.1: Where does the $\chi^2 + 1$ rule come from, and why is the rule different when we consider the log posterior (or log likelihood) instead of χ^2?

7.3 Normalisation uncertainties

When parameters are estimated from a combination of data sets, each with their own normalisation uncertainty Δ, it may be beneficial to allow the data to float within this uncertainty. In case of a single data set (to keep the notation simple) a multiplicative normalisation parameter N is introduced and the χ^2 is often defined by

$$\chi^2(N, \boldsymbol{\theta}) = \sum_i \left[\frac{Nd_i - f_i(\boldsymbol{\theta})}{\sigma_i}\right]^2 + \left(\frac{N-1}{\Delta}\right)^2 . \tag{7.8}$$

Here the last term is a so-called 'penalty chi-squared', introduced to constrain N to within the quoted error Δ, thus avoiding a possible collapse $\chi^2 = 0$ for $N = 0$.[31] In Fig. 12 we show data fitted to a straight line with a 5% floating normalisation, using

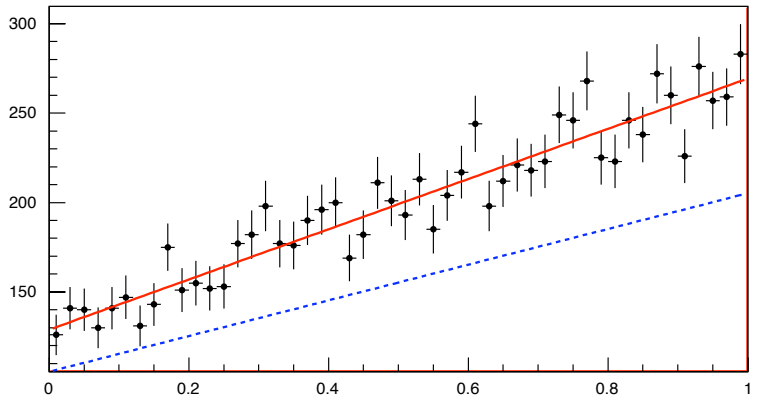


Figure 12: Data fitted to a linear dependence with a 5% floating normalisation uncertainty. The dashed line shows the result when minimising (7.8) and the full line when minimising (7.11).

the χ^2 definition (7.8). The result, plotted as the dashed line in the figure, completely misses the data! This bias is explained in [Agost94, Take96] but its origin is clear: normalisation affects both d and σ so that the χ^2 definition (7.8) is just wrong.

So let us first write down the correct likelihood[32]

$$p(\boldsymbol{d}|N, \boldsymbol{\theta}, I) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{Nd_i - f_i(\boldsymbol{\theta})}{N\sigma_i}\right]^2\right\} . \tag{7.9}$$

We assume a Gaussian prior for N

$$p(N|I) \propto \exp\left[-\frac{1}{2}\left(\frac{N-1}{\Delta}\right)^2\right] \tag{7.10}$$

[31] Such a collapse occurs when there exists some $\hat{\boldsymbol{\theta}}$ for which $f_i(\hat{\boldsymbol{\theta}}) = 0$ for all data points i.

[32] One may wonder why σ is scaled by N in the exponent but not in the normalisation constant. We leave it as an exercise to show that the normalisation factors of (7.9) are, indeed, correct.

so that (minus) the log posterior becomes

$$\chi^2(N,\boldsymbol{\theta}) = \sum_i \left[\frac{Nd_i - f_i(\boldsymbol{\theta})}{N\sigma_i}\right]^2 + \left(\frac{N-1}{\Delta}\right)^2. \tag{7.11}$$

It is now also understood that the last term in (7.8) is not some magic penalty χ^2, but just a Bayesian prior. Note, however, that N is positive definite so that a Gaussian prior is inadequate when Δ is large. It would then be better to choose a lognormal distribution,[33] for instance. But the Gaussian prior is fine for our 5% normalisation error, and minimising (7.11) clearly does a good job, as is shown by the full line in Fig. 12.

7.4 Uncertain experimental errors

In a χ^2 analysis of a large body of experimental data, it often happens that a value of $\chi^2/\nu > 1$ is observed. For example, a global parton distribution fit by the CTEQ group [Pump02], yields $\chi^2 = 1954/1811 = 1.08$. This corresponds to a p-value of 1%. Instead of rejecting QCD as the theory of the strong interaction on the grounds of this p-value, it is more reasonable to suspect that, among other possible causes, the published experimental errors might be somewhat underestimated, or that there are unknown correlations in the data.

To obtain a rough estimate of the effect, one can artificially increase the experimental errors by some common factor α.

> **Exercise 7.2**: Show that when the experimental errors are scaled by α, the χ^2 is scaled by $1/\alpha^2$ and the (co)variances of the estimated parameters by α^2.

A reasonable choice is $\alpha = \sqrt{\chi^2/\nu}$ since that makes the χ^2 value equal to the number of degrees of freedom, as would be the case for a perfect fit. The covariance matrix of the fitted parameters must then be multiplied by a factor α^2. Note that such scaling makes, by construction, the data globally compatible with the fit model; we try, in a sense, to *determine* the experimental error by observing the scatter of the data around the optimal value. For the CTEQ fit above, the scale factor would give only a modest increase of 4% in the parameter errors.

While globally underestimated (or overestimated) errors are a nuisance—they spoil the comparison between the data and the model—it is far more dangerous to have outliers in the data. This is because Gauss distributions carry little probability in the tails so that a data point which is many standard deviations off can exert an enormous pull on the fit. Outlier sensitivity is considerably reduced if we allow the experimental errors to be distributed towards larger values, instead of fixing them to the published ones. The Gaussian probability distributions of the data then acquire long tails, somewhat similar to the Student-t distribution described in Section 6.1. Below we will use a very simple *ansatz* for the distribution of errors, taken from the book by Sivia [Sivia06].

The proposed distribution of the error on a data point is

$$p(\sigma|\sigma_0) = \frac{\sigma_0}{\sigma^2} \qquad \text{for} \qquad \sigma \geq \sigma_0, \tag{7.12}$$

[33] If $x \in (0,\infty]$ and $\ln(x)$ is normally distributed, then x is lognormally distributed.

and zero otherwise. Assuming that $p(d|\mu,\sigma)$ is Gaussian distributed, we obtain for the likelihood of a data point d

$$p(d|\mu,\sigma_0) = \int_0^\infty p(d,|\mu,\sigma)p(\sigma|\sigma_0)\,\mathrm{d}\sigma = \frac{1}{\sqrt{2\pi}}\int_{\sigma_0}^\infty \frac{\sigma_0}{\sigma^3}\exp\left[-\frac{(d-\mu)^2}{2\sigma^2}\right]\,\mathrm{d}\sigma. \qquad (7.13)$$

Transforming $z = 1/\sigma$, this integral is not difficult to evaluate; the result is

$$p(d,|\mu,\sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}}\left[\frac{1-\exp(-\Delta^2/2)}{\Delta^2}\right] \qquad \text{with} \qquad \Delta = \frac{d-\mu}{\sigma_0}. \qquad (7.14)$$

The likelihood is shown for $\mu = 0$ and $\sigma_0 = 1$ in Fig. 13. It clearly has much longer

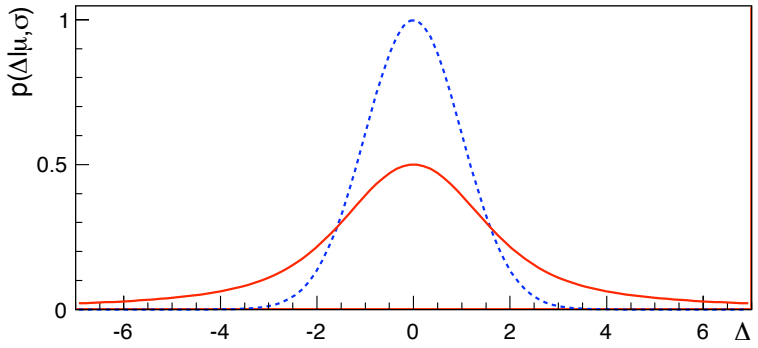


Figure 13: The likelihood (7.14) of a measurement with uncertain errors (full curve), compared to a Gaussian of unit width (dashed curve). The distributions are scaled to unit maximum for the Gauss.

tails than the corresponding Gauss distribution (dashed curve in the figure). In Fig. 14 we plot data that follow a straight line, but has two outliers. These outliers spoil a

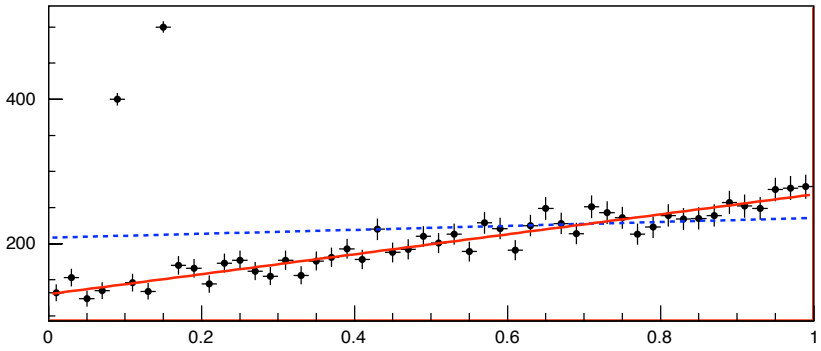


Figure 14: Data along a straight line with two outliers that spoil the least squares fit, shown by the dashed curve. A minimisation of the log likelihood (7.15) is insensitive to the outliers (full curve).

least squares fit as is shown by the dashed curve. However, the full line in the figure shows that there is hardly any outlier sensitivity when we minimise the log likelihood (or, rather, the log posterior)

$$L = -\sum_i \log\left[\frac{1-\exp(-\Delta_i^2/2)}{\Delta_i^2}\right]. \qquad (7.15)$$

In the above, we have used a very simple-minded $1/\sigma^2$ distribution of the experimental errors; see [Agost99] for a more sophisticated assignment.

7.5 Errors on both x and y

Sometimes we want to estimate the parameters $\boldsymbol{\theta}$ of a functional relationship

$$y = f(x; \boldsymbol{\theta}) \tag{7.16}$$

from a data set where both x and y have errors, be it correlated or uncorrelated. How to deal with this is very nicely explained by D'Agostini in [Agost05], and we will closely follow here his line of argument.

The variables in the problem are the measured points x_i and y_i, their means μ_{x_i} and μ_{y_i}, and the parameters $\boldsymbol{\theta}$. We are interested in the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, I)$. The first step is to relate this conditional probability to the joint density $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I)$.

Using the product rule we may write

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I) = p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, I) p(\boldsymbol{x}, \boldsymbol{y}|I). \tag{7.17}$$

On the other hand, marginalisation gives

$$p(\boldsymbol{x}, \boldsymbol{y}|I) = \iiint p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y \mathrm{d}\boldsymbol{\theta} \tag{7.18}$$

so that we obtain, combining (7.17) and (7.18)

$$p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, I) = \frac{p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I)}{\iiint p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y \mathrm{d}\boldsymbol{\theta}}$$

Marginalisation over $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ then yields the desired result

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}, I) = \frac{\iint p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I) \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y}{\iiint p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y \mathrm{d}\boldsymbol{\theta}}. \tag{7.19}$$

Our task is now to determine the joint probability density $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I)$ from the background information I at our disposal. By successive application of the product rule, we can write this joint density as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I) &= p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) \\ &\times\ p(\boldsymbol{y}|\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) \\ &\times\ p(\boldsymbol{\mu}_y|\boldsymbol{\mu}_x, \boldsymbol{\theta}, I) \\ &\times\ p(\boldsymbol{\mu}_x|\boldsymbol{\theta}, I) \\ &\times\ p(\boldsymbol{\theta}|I) \end{aligned} \tag{7.20}$$

There are of course many possible orderings of the conditional probabilities in this chain (also called a **Bayesian network**) and one should try to chose the sequence which best accommodates the model relations between the variables (we have done that here, as

we will see below). The last density $p(\boldsymbol{\theta}|I)$ in (7.20) is a prior: priors, as we have emphasised several times in these lectures, do *always* enter into an inference problem.

Now we proceed by actually incorporating the relations between the variables in the conditional probabilities of (7.20). These relations are of course not universal, but depend on the problem we want to solve; the ones given below are just toy assumptions.

1. We assume that $\boldsymbol{x}$ depends only on $\boldsymbol{\mu}_x$, so that $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) = p(\boldsymbol{x}|\boldsymbol{\mu}_x, I)$;
2. Likewise, we assume that there are no correlations between $\boldsymbol{x}$ and $\boldsymbol{y}$ so that $\boldsymbol{y}$ depends only on $\boldsymbol{\mu}_y$. Therefore $p(\boldsymbol{y}|\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, I) = p(\boldsymbol{y}|\boldsymbol{\mu}_y, I)$;
3. The relation between $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ is given by (7.16) so that
$$p(\boldsymbol{\mu}_y|\boldsymbol{\mu}_x, \boldsymbol{\theta}, I) = \delta[\boldsymbol{\mu}_y - f(\boldsymbol{\mu}_x; \boldsymbol{\theta})];$$
4. Finally, we regard $\boldsymbol{\mu}_x$ as an independent variable with a prior density $p(\boldsymbol{\mu}_x|I)$.

In Fig. 15 we show a diagram that represents (7.20). The full (dashed) arrows indicate

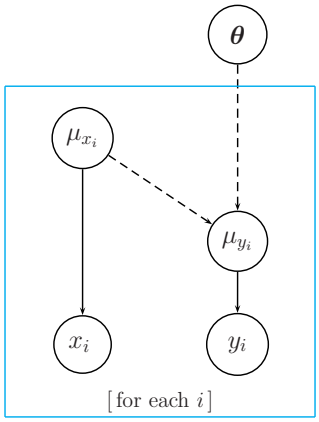


Figure 15: Bayesian network diagram showing the relations between the variables of the inference problem described in the text. Full (dashed) arrows represent probabilistic (functional) relations. Figure taken from [Agost05].

the probabilistic (functional) relations between the variables. Note that the variables $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_x$ have no arrows pointing to them so that these are priors.

We take a uniform prior for $\boldsymbol{\mu}_x$ so that we have,

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|I) \propto p(\boldsymbol{x}|\boldsymbol{\mu}_x, I)\; p(\boldsymbol{y}|\boldsymbol{\mu}_y, I)\; \delta[\boldsymbol{\mu}_y - f(\boldsymbol{\mu}_x; \boldsymbol{\theta})]\; p(\boldsymbol{\theta}|I). \tag{7.21}$$

If we assume Gaussian distributions for $\boldsymbol{x}$ and $\boldsymbol{y}$ and a linear relation

$$\boldsymbol{\mu}_y = a\,\boldsymbol{\mu}_x + b,$$

then it is an exercise in Gaussian integration to obtain from (7.21) and (7.19)

$$p(a, b|\boldsymbol{x}, \boldsymbol{y}) \propto \prod_i \frac{1}{\sqrt{\sigma_{y_i}^2 + a^2\sigma_{x_i}^2}} \exp\left[-\frac{(y_i - ax_i - b)^2}{2(\sigma_{y_i}^2 + a^2\sigma_{x_i}^2)}\right] p(a, b|I). \tag{7.22}$$

Exercise 7.3: Carry out the integrals leading to (7.22).

Note that the errors on x_i and y_i are added in quadrature, but with σ_{x_i} weighted by the slope parameter a. Note also that the straight line fit has become non-linear in the parameters, so that the negative logarithm of (7.22) has to be minimised numerically, by MINUIT, for instance.[34] In Fig. 16 we show a linear fit to data which have errors on both x and y.

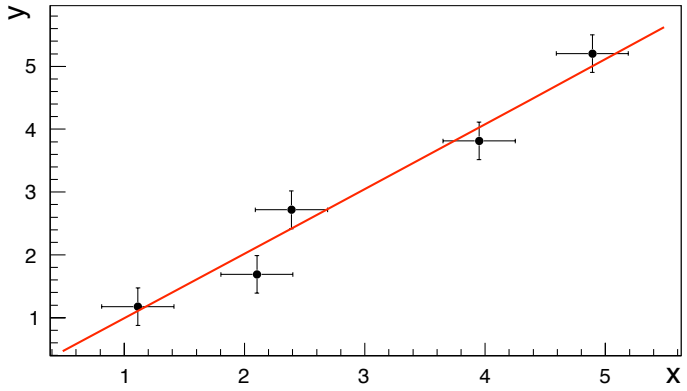


Figure 16: Straight line fit using (7.22) to data with (uncorrelated) errors on both x and y.

Exercise 7.4: Repeat the analysis above, but assume that the errors on x and y are correlated. How does the Bayesian network diagram now look?

8 Bayesian Hypothesis Testing

In the previous sections we have derived posterior distributions of model parameters under the assumption that the model hypothesis is true. In other words, we have only investigated the model parameters, and not the model itself.

But how do we handle the case when the data do, in fact, indicate that the model hypothesis might be wrong? The Bayesian answer is to populate an enlarged hypothesis space with several competing models and then ask the question which model is preferred by the data. A requirement is that the extended set of model hypotheses must form a reasonably complete collection of exclusive alternatives.[35] The procedure to pick the best alternative is called **model selection**. As we will see, this model selection is not only based on the quality of the data description ('goodness of fit') but also on a factor which penalises models which have a larger number parameters. Bayesian model selection thus automatically applies *Occam's razor* in preferring, to a certain degree, the more simple model.

[34]The method `TGraph::Fit` in ROOT can handle errors on both the x and y coordinate, using a generalisation of (7.22) where the slope parameter a is replaced by $f'(x)$.

[35]In particle ID, for instance, the hypothesis space is usually taken to contain e, μ, π, K, p and d.

8.1 Model selection

As already mentioned several times before, Bayesian inference can only asses the plausibility of an hypothesis when this hypothesis is a member of an exclusive and exhaustive set. One can of course always complement a given hypothesis H by its negation $\overline{H}$ but this does not bring us very far since $\overline{H}$ is usually too vague a condition for a meaningful probability assignment.[36] Thus, in general, we have to include our model H into a finite set of mutually exclusive and exhaustive alternatives $\{H_k\}$. This obviously restricts the outcome of our selection procedure to one of these alternatives but has the virtue that we can use Bayes' theorem to assign posterior probabilities to each of the H_k

$$P(H_k|D,I) = \frac{P(D|H_k,I)\,P(H_k|I)}{\sum_i P(D|H_i,I)\,P(H_i|I)}. \tag{8.1}$$

To avoid calculating the denominator, one often works with the so-called **odds ratio** (*i.e.* a ratio of probabilities)

$$O_{kj} = \underbrace{\frac{P(H_k|D,I)}{P(H_j|D,I)}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_k|I)}{P(H_j|I)}}_{\text{Prior odds}} \times \underbrace{\frac{P(D|H_k,I)}{P(D|H_j,I)}}_{\text{Bayes' factor}}. \tag{8.2}$$

The first term on the right-hand side is called the **prior odds** and the second term the **Bayes' factor**.

The selection problem can thus be solved by calculating the odds with (8.2) and accept hypothesis k if O_{kj} is much larger than one, declare the data to be inconclusive if the ratio is about unity and reject k in favour of one of the alternatives if O_{kj} turns out to be much smaller than one. The prior odds are usually set to unity unless there is strong prior evidence in favour of one of the hypotheses. However, when the hypotheses in (8.2) are *composite*, then not only the prior odds depend on prior information but also the Bayes' factor.

To see this, we follow Sivia [Sivia06] in working out an illustrative example where the choice is between a parameter-free hypothesis H_0 and an alternative H_1 with one free parameter λ. Let us denote a set of data points by $\boldsymbol{d}$ and expand the probability density $p(\boldsymbol{d}|H_1)$ into the parameter λ

$$p(\boldsymbol{d}|H_1) = \int p(\boldsymbol{d},\lambda|H_1)\,\mathrm{d}\lambda = \int p(\boldsymbol{d}|\lambda,H_1)\,p(\lambda|H_1)\,\mathrm{d}\lambda. \tag{8.3}$$

To evaluate (8.3) we assume a uniform prior for λ in a finite range $\Delta\lambda$ and write

$$p(\lambda|H_1) = \frac{1}{\Delta\lambda}. \tag{8.4}$$

Gaussian approximation of the likelihood (which is a function of λ) gives

$$p(\boldsymbol{d}|\lambda,H_1) \approx p(\boldsymbol{d}|\hat{\lambda},H_1)\exp\left[-\frac{1}{2}\left(\frac{\lambda-\hat{\lambda}}{\sigma}\right)^2\right]. \tag{8.5}$$

[36] We could, for instance, describe a detector response to pions by a probability $P(d|\pi)$. However, it would be very hard to assign something like a 'not-pion probability' $P(d|\sim\pi)$ without specifying the detector response to members of an alternative set of particles like electrons, kaons, protons *etc.*

Here $\hat{\lambda}$ is the mode of the likelihood and σ is the inverse of the Hessian, as is described in Section 3.1. Inserting (8.4) and (8.5) in (8.3) gives, upon integration over λ,

$$p(\boldsymbol{d}|H_1) \approx p(\boldsymbol{d}|\hat{\lambda}, H_1)\, \frac{\sigma\sqrt{2\pi}}{\Delta\lambda}. \tag{8.6}$$

Exercise 8.1: Derive an expression for the posterior of λ by inserting Eqs. (8.4), (8.5) and (8.6) in Bayes theorem

$$p(\lambda|\boldsymbol{d}, H_1) = \frac{p(\boldsymbol{d}|\lambda, H_1)\, p(\lambda|H_1)}{p(\boldsymbol{d}|H_1)}.$$

The result should look familiar.

Thus we find for the odds ratio (8.2)

$$\underbrace{\frac{P(H_0|\boldsymbol{d}, I)}{P(H_1|\boldsymbol{d}, I)}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_0|I)}{P(H_1|I)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\boldsymbol{d}|H_0)}{p(\boldsymbol{d}|H_1, \hat{\lambda})}}_{\text{Likelihood ratio}} \times \underbrace{\frac{\Delta\lambda}{\sigma\sqrt{2\pi}}}_{\text{Occam factor}} \tag{8.7}$$

As already mentioned above, the prior odds can be set to unity unless there is information which gives us prior preference for one model over another. The likelihood ratio will, in general, be smaller than unity and therefore favour the model H_1. This is because the additional flexibility of an adjustable parameter usually yields a better description of the data. This preference for models with more parameters leads to the well known phenomenon that one can 'fit an elephant' with enough free parameters.[37] This clearly illustrates the inadequacy of using the fit quality as the only criterion in model selection. Indeed, such a criterion alone could never favour a simpler model.

Intuitively we would prefer a model that gives a good description of the data in a wide range of parameter values over one with many fine-tuned parameters, unless the latter would provide a significantly better fit. Such an application of Occam's razor is encoded by the so-called **Occam factor** in (8.7). This factor tends to favour H_0 since it penalises H_1 for reducing a wide parameter range $\Delta\lambda$ to a smaller range σ allowed by the fit. Here we immediately face the problem that H_0 would always be favoured in case $\Delta\lambda$ is set to infinity. Prior assignment is thus a more sensitive issue in model selection problems than it is in parameter estimation problems.

In case H_i and H_j both have one free parameter (μ and λ) the odds ratio becomes

$$\frac{P(H_i|\boldsymbol{d}, I)}{P(H_j|\boldsymbol{d}, I)} = \frac{P(H_i|I)}{P(H_j|I)} \times \frac{p(\boldsymbol{d}|H_i, \hat{\mu})}{p(\boldsymbol{d}|H_j, \hat{\lambda})} \times \frac{\Delta\lambda}{\sigma_\lambda}\, \frac{\sigma_\mu}{\Delta\mu}. \tag{8.8}$$

For similar prior ranges $\Delta\lambda$ and $\Delta\mu$ the likelihood ratio has to overcome the penalty factor $\sigma_\mu/\sigma_\lambda$. This factor favours the model for which the likelihood has the largest width. It may seem a bit strange that the less discriminating model is favoured but inspection of (8.6) shows that the **evidence** $P(D|H)$ carried by the data tends to be larger for models with a larger ratio $\sigma/\Delta\lambda$, that is, for models which cause a smaller collapse of the hypothesis space when confronted with the data. Note that the prior ranges enter as a ratio in (8.8) and that they vanish for parameters that are common between the hypotheses. Note also that the Occam factor penalises each free parameter, so that models with many parameters may get very strongly disfavoured by this factor.

[37] Including as many parameters as data points will cause any model to perfectly describe the data.

Exercise 8.2: Generalise (8.8) to the case where H_i has n free parameters $\boldsymbol{\lambda}$ and H_j has m free parameters $\boldsymbol{\mu}$ (with $n \neq m$).

From the above it should be clear that model selection is less straight-forward than parameter estimation because the answer depends on the hypothesis space explored, and on the prior ranges of the parameters. But these issues may be less of a problem than one might initially think, as the example in the next section shows.

8.2 Example: is there a signal or not?

To illustrate model selection, we will search for a signal (s), superimposed on a uniform background (b) in a histogram. This is similar to the D^0 search described in Section 7.1, except that in the D^0 search there was no doubt on the validity of the fit model: we know that D^0 *are* produced in heavy-ion collisions, and we also know where the signal should show up in the invariant mass spectrum. In the analysis presented below we are not sure that a signal exists (like in a Higgs search) so that we have to consider two alternative hypotheses, namely, H_0 := 'no signal exists' and H_1 := 'a signal exists'.

The data are accumulated in a histogram that covers a range Δx in some variable x (invariant mass, for instance). The signal—if it exists at all—has an unknown position which should lie somewhere in the measured range Δx. Furthermore, the anticipated signal has a very narrow width so that it will contribute to the yield of a single bin only. The background distribution is known to be uniform. As an example, we show in Fig. 17 data accumulated in a histogram with 20 bins. A signal with a significance of $\mathcal{S} = s/\sqrt{b} = 6$ is clearly visible above a uniform background.

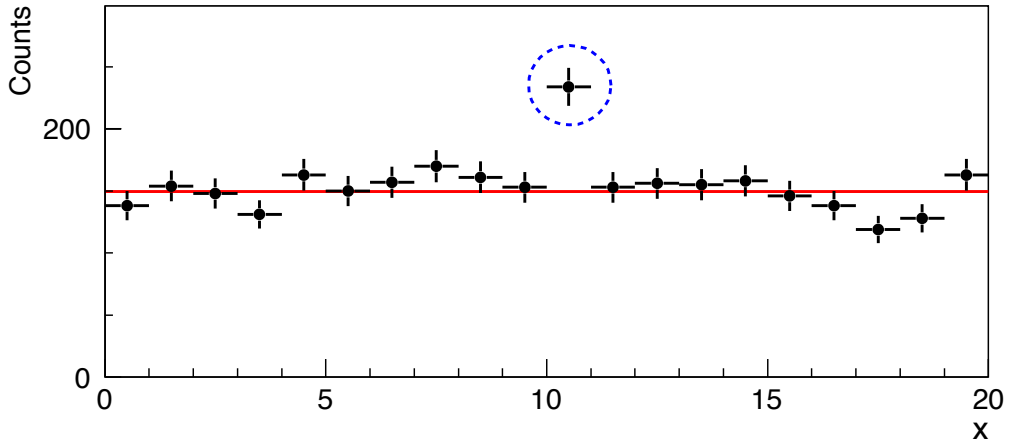


Figure 17: Histogram showing a signal s (dashed circle) above a uniform background b (full line) in 20 bins of x. In this example, the significance $s/\sqrt{b} = 6$. This significance is large enough to establish the presence of a signal by visual inspection, without a model selection analysis.

We will now identify the parameters of the hypotheses H_0 and H_1, and set the prior ranges. The total number of counts in the histogram is denoted by N_{tot}, and the number of bins by n_{bins}.

Hypothesis H_0: There is no signal and the histogram is populated by a uniform background b (counts per bin). This background rate is the only parameter of H_0, and we assume a uniform prior $p(b|H_0) = 1/\Delta b$, with Δb some constant.

Hypothesis H_1: The histogram is populated by a uniform background b, and a narrow signal s that contributes only to one bin, located at x_s. This hypothesis has three parameters with the following prior probabilities:

Parameter b : We assume, like above, a flat background rate b (counts per bin) with a uniform prior $p(b|H_1) = 1/\Delta b$.

Parameter s : When the signal significance $\mathcal{S} = s/\sqrt{b}$ is large (Fig. 17) we do not need model selection so that we can restrict ourselves to small signals $0 < s < s_{\rm max} = \mathcal{S}\sqrt{b} \approx \mathcal{S}\sqrt{N_{\rm tot}/n_{\rm bins}} \equiv \Delta s$. We assign a uniform prior $p(s|H_1) = 1/\Delta s$ which is calculated with, say, $\mathcal{S} = 6$.

Parameter x_s : The position of the signal is expected to be within the histogram range Δx with a uniform prior of $p(x_s|H_1) = 1/\Delta x$.

Under the hypothesis H_0, the likelihood of the data is given by (Poisson statistics)

$$p(\boldsymbol{n}|b, H_0) = \prod_{i=1}^{n_{\rm bins}} \frac{b^{n_i}}{n_i!}\, e^{-b}. \tag{8.9}$$

Minimisation of the negative log likelihood gives for the mode and width (inverse of the Hessian) of b:

$$b_0 = \frac{N_{\rm tot}}{n_{\rm bins}}, \qquad \sigma_0 = \frac{\sqrt{N_{\rm tot}}}{n_{\rm bins}}. \tag{8.10}$$

Under the hypothesis H_1 we have for the likelihood

$$p(\boldsymbol{n}|b, s, x_s, H_1) = \left[\prod_{i=1}^{n'_{\rm bins}} \frac{b^{n_i}}{n_i!}\, e^{-b}\right] \times \left[\frac{(b+s)^{n_s}}{n_s!}\, e^{-(b+s)}\right]. \tag{8.11}$$

The product in the first bracket gives the contribution to the likelihood of all bins that contain only background. The term in the second bracket gives the contribution from the bin that contains both signal and background. This bin is taken to be that with the largest count n_s in the histogram. Minimisation of the negative log likelihood gives

$$\begin{aligned} b_1 &= \frac{N'_{\rm tot}}{n'_{\rm bins}}, \qquad \sigma_1 = \frac{\sqrt{N'_{\rm tot}}}{n'_{\rm bins}} \\ s &= n_s - b_1, \qquad \sigma_s = \sqrt{\sigma_1^2 + n_s}, \end{aligned} \tag{8.12}$$

where n_s is the largest bin content in the histogram, $N'_{\rm tot} = N_{\rm tot} - n_s$ and $n'_{\rm bins} = n_{\rm bins} - 1$. The bin with the largest content is located at x_s to which we assign an error of $\sigma_x = w/\sqrt{12}$, where $w = \Delta x/n_{\rm bins}$ is the bin width.

Exercise 8.3: (i) Derive the results given in (8.10) and (8.12). (ii) Show that the variance of a uniform distribution in the range $a < x < b$ is given by $(b-a)^2/12$.

Setting the prior odds to unity, we find from (8.7) for the posterior odds

$$R_{\rm post} = \frac{p(H_1|\boldsymbol{n})}{p(H_0|\boldsymbol{n})} = \left[\left(\frac{b_1}{b_0}\right)^{N'_{\rm tot}} \left(\frac{b_1+s}{b_0}\right)^{n_s} e^{-n_{\rm bins}(b_1-b_0)-s}\right] \times \left[\frac{2\pi\,\sigma_1\,\sigma_s\,\sigma_x}{\sigma_0\,\Delta s\,\Delta x}\right]. \qquad (8.13)$$

Here the first term in the square brackets is the likelihood ratio, and the second term is the Occam factor.[38] Note that the background prior range Δb, which is difficult to assign, cancels in the Occam factor because b is a parameter of both H_0 and H_1.

Because the values of the ratio's in (8.13) can become quite extreme, it is convenient to express them on a decibel scale, defined by $10\log_{10}(R)$ dB. A value of 0 dB means a unit ratio (fifty-fifty odds), 3 dB is about a factor of 2, and 10 dB corresponds to a factor of 10. In the following we will accept a posterior odds of +10 dB or larger, that is, we are willing to bet on H_1 when there is an odds ratio of 10 : 1 in favour of H_1.

In Fig. 18 we show simulated data where 500 counts are distributed over 20 bins. A

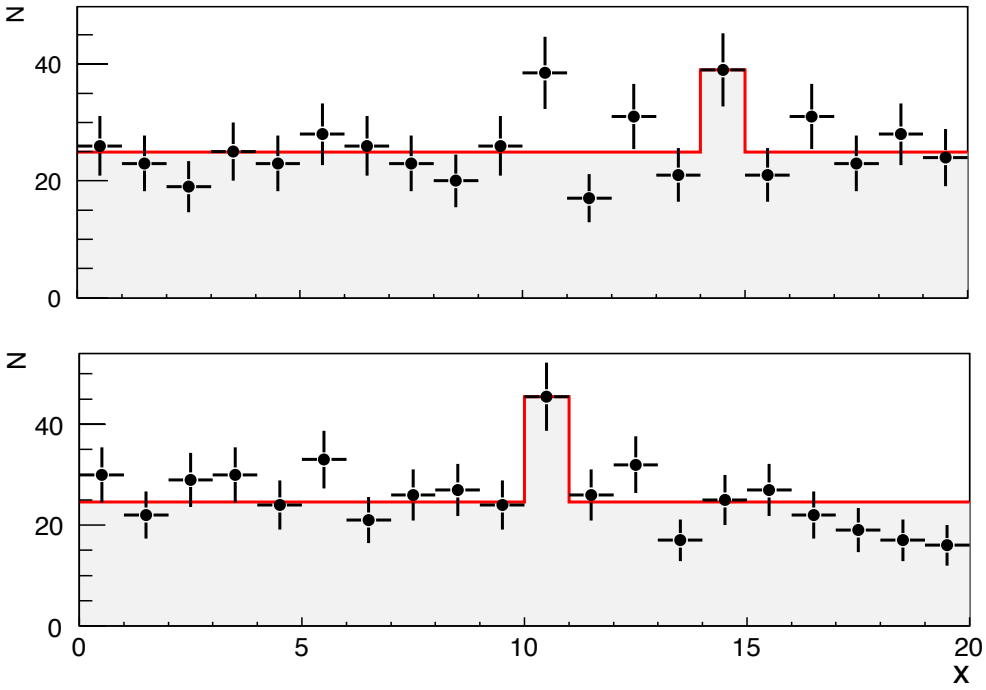


Figure 18: Signal search in a histogram, assuming that the signal contributes to one bin, superimposed on a uniform background (full curves). The top plot shows a signal with a significance $\mathcal{S} = 2.8$. The fit is rejected because the posterior odds are found to be −3.3 dB. The bottom plot shows a signal with a significance $\mathcal{S} = 4.2$. The fit is accepted because the posterior odds are found to be +12.0 dB.

signal with a strength of half the nominal background is generated in the 11$^{\rm th}$ bin; this corresponds, on average, to a significance of $\mathcal{S} = 12.5/\sqrt{25} = 2.5$. The differences between the top and bottom histograms of Fig. 18 are simply caused by random fluctuations.

[38] For simplicity we have neglected correlations between the parameters in the Occam factor; we leave it as an exercise to take correlations into account.

Both histograms were fitted to the hypotheses H_0 and H_1 and the posterior odds were calculated from (8.13). The results are listed in Table 1. The fit of the top histogram

Table 1: The signal significance, likelihood ratio, Occam factor and the posterior odds from the fits to the two histograms of Fig. 18. Also given are the χ^2 and the corresponding p-values for the fits to the hypothesis H_1, with 17 degrees of freedom. The values given in brackets are from the fit to the background-only hypothesis H_0, with 19 degrees of freedom.

	Significance	Likelihood	Occam	Posterior	χ^2	p-value
Top	2.8	+14 dB	−17 dB	−3 dB	18.2 (22.8)	0.38 (0.25)
Bottom	4.2	+29 dB	−17 dB	+12 dB	24.0 (32.7)	0.12 (0.03)

picks for the signal a random fluctuation in the 15^{th} bin ($\mathcal{S} = 2.8$). However, the hypothesis H_1 is rejected in favour of H_0 because the likelihood ratio of 14 dB is not large enough to overcome the Occam penalty factor of −17 dB. Selection on the basis of p-values (goodness of fit) would wrongly favour the hypothesis H_1 (see Table 1). The fit to the bottom histogram finds the signal at the 11^{th} bin with a significance of $\mathcal{S} = 4.2$. Here the hypothesis H_1 is accepted on grounds of the large posterior odds of +12 dB. Note that a standard cut of 5% on the p-value would also accept H_1, and reject H_0.

For other worked-out examples of model selection we refer to Loredo [Lor90], Bretthorst [Bret96], Sivia [Sivia06] and Gregory [Greg05].

9 Concluding Remarks

In these lectures we have shown how the intimate connection between probability and logic leads to a beautifully coherent and simple theory of inference. This theory is based on the Cox' desiderata of plausible inference which lead to an algebra that turns out to be the same as that of probability, based on the Kolmogorov axioms. For a Bayesian, probability is thus a measure of plausibility, and probabilities can be assigned to propositions, or hypotheses. This is in contrast to the Frequentist notion of probability as a relative frequency of occurrence. Here a probability cannot be assigned to a hypothesis since it is not a random variable. This distinction lies at the heart of the difference between the Bayesian and Frequentist approaches to data analysis.

Necessary input to Bayesian inference is prior information which enters via Bayes' theorem through a multiplication of the likelihood function by a prior probability density. The assignment of priors is still an open issue but in practice this is often not such a great concern because priors are important only when the data carry little information on the inference being conducted, or when relevant information is contained in the prior and not in the data. But prior probability assignment is clearly an important issue which we have, in these lectures, tried to trace back to Bernoulli's principle of insufficient reason and to symmetry and maximum entropy arguments.

Many examples in these lectures clearly show that Bayesian inference often amounts to straight-forward application of probability calculus, using little else but expansion, probability inversion and marginalisation. The Bayesian approach has therefore the

virtue of taking much mystery out of statistical methods while providing a simple and well-founded framework for creative solutions to statistical data analysis problems.

A Gaussian Integration

In this appendix we derive expressions for the normalisation and marginalisation integrals of a multivariate Gaussian distribution. Starting point is the well known result

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} \mathrm{d}z = \sqrt{2\pi}. \tag{A.1}$$

From this it follows immediately that

$$\int \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \lambda_i z_i^2\right) \mathrm{d}z_1 \cdots \mathrm{d}z_n = \frac{(2\pi)^{\frac{n}{2}}}{\sqrt{\lambda_1 \cdots \lambda_n}} \qquad (\lambda_i > 0).$$

Introducing the vector notation

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

we can write

$$\int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}\, \boldsymbol{z}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{z}\right) \mathrm{d}\boldsymbol{z} = \frac{(2\pi)^{\frac{n}{2}}}{\sqrt{|\boldsymbol{D}|}}, \tag{A.2}$$

where $|\boldsymbol{D}| = \lambda_1 \cdots \lambda_n$ denotes the determinant of $\boldsymbol{D}$. Next, we introduce the non-singular linear transformation $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$. The Jacobian of this transformation is $\mathrm{d}\boldsymbol{z} = |\boldsymbol{A}|\, \mathrm{d}\boldsymbol{x}$ so that (A.2) becomes

$$\int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}\, \boldsymbol{x}^{\mathrm{T}} (\boldsymbol{A}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{A}) \boldsymbol{x}\right] |\boldsymbol{A}|\, \mathrm{d}\boldsymbol{x} = \frac{(2\pi)^{\frac{n}{2}}}{\sqrt{|\boldsymbol{D}|}}. \tag{A.3}$$

Setting $\boldsymbol{H} \equiv \boldsymbol{V}^{-1} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{A}$ we find for the normalisation integral

$$\int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}\, \boldsymbol{x}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} = \int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}\, \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} = \sqrt{(2\pi)^n\, |\boldsymbol{V}|} \tag{A.4}$$

where we have used the fact that $|\boldsymbol{H}| = |\boldsymbol{V}|^{-1} = |\boldsymbol{A}|^2\, |\boldsymbol{D}|$. Note that $\boldsymbol{H}$ (and also $\boldsymbol{V}$) is, by construction, symmetric positive definite. A normalised multivariate Gaussian density with a mean $\bar{\boldsymbol{x}}$ is thus given by

$$G(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n\, |\boldsymbol{V}|}} \exp\left[-\tfrac{1}{2}\, (\boldsymbol{x} - \bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})\right]. \tag{A.5}$$

To calculate the marginal integral

$$G(x_1, \ldots, x_m) = \frac{1}{\sqrt{(2\pi)^n\, |\boldsymbol{V}|}} \int \cdots \int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}\, \boldsymbol{x}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{x}\right) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_n, \tag{A.6}$$

we partition the vector $\boldsymbol{x}$ and the matrix $\boldsymbol{H}$ into

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{H} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{R} \\ \boldsymbol{R}^{\mathrm{T}} & \boldsymbol{Q} \end{pmatrix} \tag{A.7}$$

where $\boldsymbol{p}$ contains the first m elements of $\boldsymbol{x}$ and $\boldsymbol{q}$ the last $n-m$ elements. There is no loss of generality by integrating over the last elements of $\boldsymbol{x}$ since we are free to re-arrange the vector $\boldsymbol{x}$ as we like. The integral (A.6) is solved by completing the squares, that is, the exponent is written in the form (see also the solution to Exercise 3.6 in Appendix B)

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{x} = (\boldsymbol{q}-\hat{\boldsymbol{q}})^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{q}-\hat{\boldsymbol{q}}) + C. \tag{A.8}$$

Inserting (A.7) in (A.8) and comparing the terms on the left and right-hand sides yields

$$\boldsymbol{p}^{\mathrm{T}}\boldsymbol{R} = -\hat{\boldsymbol{q}}^{\mathrm{T}}\boldsymbol{Q}, \qquad \boldsymbol{R}^{\mathrm{T}}\boldsymbol{p} = -\boldsymbol{Q}\hat{\boldsymbol{q}} \qquad \text{and} \qquad \boldsymbol{p}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{p} = \hat{\boldsymbol{q}}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{q}} + C,$$

so that we obtain for the unknown $\hat{\boldsymbol{q}}$ and C in (A.8):

$$\hat{\boldsymbol{q}} = -\boldsymbol{Q}^{-1}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{p}, \qquad \hat{\boldsymbol{q}}^{\mathrm{T}} = -\boldsymbol{p}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Q}^{-1}, \qquad C = \boldsymbol{p}^{\mathrm{T}}\left(\boldsymbol{P} - \boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)\boldsymbol{p}. \tag{A.9}$$

To simplify the expression for C we observe that

$$\boldsymbol{V} = \boldsymbol{H}^{-1} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{R} \\ \boldsymbol{R}^{\mathrm{T}} & \boldsymbol{Q} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{1} & -\boldsymbol{P}^{-1}\boldsymbol{R} \\ -\boldsymbol{Q}^{-1}\boldsymbol{R}^{\mathrm{T}} & \boldsymbol{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_{\mathrm{p}} \\ \boldsymbol{V}_{\mathrm{q}} \end{pmatrix}. \tag{A.10}$$

where

$$\boldsymbol{V}_{\mathrm{p}} = \left(\boldsymbol{P} - \boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1} \qquad \text{and} \qquad \boldsymbol{V}_{\mathrm{q}} = \left(\boldsymbol{Q} - \boldsymbol{R}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{R}\right)^{-1}. \tag{A.11}$$

The above can easily be verified by substitution into the relation $\boldsymbol{H}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{H} = \boldsymbol{1}$. Comparing equations (A.9) and (A.11) we see that

$$C = \boldsymbol{p}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{p}}^{-1}\boldsymbol{p}$$

where $\boldsymbol{V}_{\mathrm{p}}$ is the $m \times m$ sub-matrix of the original covariance matrix $\boldsymbol{V}$. Eq. (A.6) can now be written as

$$G(\boldsymbol{p}) = \frac{1}{\sqrt{(2\pi)^n\,|\boldsymbol{V}|}} \exp\left(-\tfrac{1}{2}\,\boldsymbol{p}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{p}}^{-1}\boldsymbol{p}\right) \int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}\,(\boldsymbol{q}-\hat{\boldsymbol{q}})^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{q}-\hat{\boldsymbol{q}})\right] \mathrm{d}\boldsymbol{q}. \tag{A.12}$$

According to (A.4) the integral over $\boldsymbol{q}$ evaluates to $\sqrt{(2\pi)^{(n-m)}\,|\boldsymbol{Q}|}$ so that

$$G(\boldsymbol{p}) = \frac{\sqrt{(2\pi)^{(n-m)}\,|\boldsymbol{Q}|}}{\sqrt{(2\pi)^n\,|\boldsymbol{V}|}} \exp\left(-\tfrac{1}{2}\,\boldsymbol{p}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{p}}^{-1}\boldsymbol{p}\right). \tag{A.13}$$

Because $G(\boldsymbol{p})$ is normalised to unity we find, by integrating (A.13), the following non-trivial relationship between the determinants

$$|\boldsymbol{V}| = |\boldsymbol{Q}|\,|\boldsymbol{V}_{\mathrm{p}}|\,.$$

Using this relation to eliminate $|\boldsymbol{V}|$ and $|\boldsymbol{Q}|$ in (A.13) we finally find for our marginal distribution

$$G(\boldsymbol{p}) = \frac{1}{\sqrt{(2\pi)^m\,|\boldsymbol{V}_{\mathrm{p}}|}} \exp\left(-\tfrac{1}{2}\,\boldsymbol{p}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{p}}^{-1}\boldsymbol{p}\right). \tag{A.14}$$

To summarise, integration over k out of n variables of a multivariate Gaussian distribution is calculated by deleting the corresponding elements of the random variable $\boldsymbol{x} - \bar{\boldsymbol{x}}$ together with the corresponding rows and columns of the covariance matrix $\boldsymbol{V}$. Then $\boldsymbol{x}$, $\bar{\boldsymbol{x}}$, $\boldsymbol{V}$ and n are simply replaced by $\boldsymbol{x}'$, $\bar{\boldsymbol{x}}'$, $\boldsymbol{V}'$ and $n' = n - k$ in (A.5).

B Solution to Selected Exercises

Exercise 1.1

This is because astronomers often have to draw conclusions from the observation of rare events. Bayesian inference is well suited for this since it is based solely on the evidence carried by the data (and prior information) instead of being based on hypothetical repetitions of the experiment.

Exercise 2.3

$\overline{A} = A \uparrow A, \quad A \wedge B = (A \uparrow A) \uparrow (B \uparrow B)$ and $A \vee B = (A \uparrow B) \uparrow (A \uparrow B)$.

Exercise 2.4

(i) From the truth table (2.1) it is seen that both $\overline{B}$ and $A \Rightarrow B$ are true if and only if A is false. But this is just the implication $\overline{B} \Rightarrow \overline{A}$.

(ii) From (2.1) we can derive the following truth table

$A \Rightarrow B$	A	B
1	0	0
1	0	1
1	1	1

Thus, given that the impication is true then if A is false, B can be either true or false while if A is true, then B *must* be true. Likewise if B is false, A *must* be false while if B is true, A can be either true or false. This is consistent with the conclusions drawn in the four syllogisms given in Section 2.1.

Exercise 2.5

From de Morgan's law and repeated application of the product and sum rules (2.5) and (2.6) we find

$$\begin{aligned} P(A \vee B) &= 1 - P(\overline{A \vee B}) = 1 - P(\overline{A}\,\overline{B}) \\ &= 1 - P(\overline{B}|\overline{A})P(\overline{A}) = 1 - P(\overline{A})[1 - P(B|\overline{A})] \\ &= 1 - P(\overline{A}) + P(B|\overline{A})P(\overline{A}) = P(A) + P(\overline{A}B) \\ &= P(A) + P(\overline{A}|B)P(B) = P(A) + P(B)[1 - P(A|B)] \\ &= P(A) + P(B) - P(A|B)P(B) = P(A) + P(B) - P(AB). \end{aligned}$$

Exercise 2.6

(i) When A and B are mutually exclusive then

$$P(AB|I) = P(A|BI)P(B|I) = P(B|AI)P(A|I) = 0,$$

which implies that the conditional probabilities $P(A|BI)$ and $P(B|AI)$ are zero. In this case it is meaningless to talk about conditional probability inversion.

(ii) When A and B are independent, probability inversion reduces to the statements $P(A|BI) = P(A|I)$ and $P(B|AI) = P(B|I)$, and Bayes' theorem becomes a triviality.

(iii) Two propositions A and B cannot be both exclusive and independent, unless one of them is a contradiction. This follows immediately from $P(AB|I) = 0$ (exclusive) and $P(AB|I) = P(A|I)P(B|I)$ (independent) so that either $P(A|I)$ or $P(B|I)$, or both, must be zero.

Exercise 2.7

(i) The probability for Mr. White to have AIDS is

$$P(A|T) = \frac{P(T|A)P(A)}{P(T|A)P(A) + P(T|\overline{A})P(\overline{A})} = \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.03 \times 0.99} = 0.25.$$

(ii) For full efficiency, $P(T|A) = 1$ so that

$$P(A|T) = \frac{1 \times 0.01}{1 \times 0.01 + 0.03 \times 0.99} = 0.25.$$

(iii) For zero false-positives, $P(T|\bar{A}) = 0$ so that

$$P(A|T) = \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0 \times 0.99} = 1.$$

Exercise 2.8

(i) If nobody has AIDS then $P(A) = 0$ and thus

$$P(A|T) = \frac{P(T|A) \times 0}{P(T|A) \times 0 + P(T|\bar{A}) \times 1} = 0.$$

(ii) If everybody has AIDS then $P(A) = 1$ and thus

$$P(A|T) = \frac{P(T|A) \times 1}{P(T|A) \times 1 + P(T|\bar{A}) \times 0} = 1.$$

In both these cases the posterior is thus equal to the prior independent of the likelihood $P(T|A)$.

Exercise 2.10

(i) When μ is known we can write for the posterior distribution

$$P(\pi|S,\mu) = \frac{P(S|\pi)P(\pi)}{P(S|\pi)P(\pi) + P(S|\sim\pi)P(\sim\pi)} = \frac{\varepsilon\mu}{\varepsilon\mu + \delta(1-\mu)}.$$

(ii) When μ is unknown we expand $P(\pi|S)$ in μ which gives

$$P(\pi|S) = \int_0^1 P(\pi,\mu|S)\,\mathrm{d}\mu = \int_0^1 P(\pi|S,\mu)p(\mu)\,\mathrm{d}\mu.$$

Assuming a uniform prior $p(\mu) = 1$ gives for the probability that the signal S corresponds to a pion

$$P(\pi|S) = \int_0^1 \mathrm{d}\mu \frac{\varepsilon\mu}{\varepsilon\mu + \delta(1-\mu)} = \frac{\varepsilon[\varepsilon - \delta + \delta\ln(\delta/\varepsilon)]}{(\delta-\varepsilon)^2},$$

where we have used the *Mathematica* program to evaluate the integral.

Exercise 2.11

The quantities x, x_0, d and ϑ are related by $x = x_0 + d\tan\vartheta$. With $p(\vartheta|I) = 1/\pi$ it follows that $p(x|I)$ is Cauchy distributed

$$p(x|I) = p(\vartheta|I)\left|\frac{\mathrm{d}\vartheta}{\mathrm{d}x}\right| = \frac{1}{\pi}\frac{\cos^2\vartheta}{d} = \frac{1}{\pi}\frac{d}{(x-x_0)^2+d^2}.$$

The first and second derivatives of $L = -\ln p$ are

$$\frac{\mathrm{d}L}{\mathrm{d}x} = \frac{2(x-x_0)}{(x-x_0)^2+d^2} \qquad \frac{\mathrm{d}^2L}{\mathrm{d}x^2} = -\frac{4(x-x_0)^2}{[(x-x_0)^2+d^2]^2} + \frac{2}{(x-x_0)^2+d^2}.$$

This gives for the position and width of $p(x)$

$$\frac{\mathrm{d}L(\hat{x})}{\mathrm{d}x} = 0 \quad \Rightarrow \quad \hat{x} = x_0 \qquad \frac{1}{\sigma^2} = \frac{\mathrm{d}^2L(\hat{x})}{\mathrm{d}x^2} = \frac{2}{d^2} \quad \Rightarrow \quad \sigma = \frac{d}{\sqrt{2}}.$$

Exercise 2.12

(i) The posterior distribution of the first measurement is

$$P(\pi|S_1) = \frac{P(S_1|\pi)P(\pi)}{P(S_1|\pi)P(\pi) + P(S_1|\sim\pi)P(\sim\pi)} = \frac{\varepsilon\mu}{\varepsilon\mu + \delta(1-\mu)}.$$

Using this as the prior for the second measurement we have

$$P(\pi|S_1,S_2) = \frac{P(S_2|\pi,S_1)P(\pi|S_1)}{P(S_2|\pi,S_1)P(\pi|S_1)+P(S_2|\sim\pi,S_1)P(\sim\pi|S_1)} = \cdots = \frac{\varepsilon^2\mu}{\varepsilon^2\mu+\delta^2(1-\mu)}.$$

Here we have assumed that the two measurements are *independent*, that is,

$$P(S_2|\pi,S_1) = P(S_2|\pi) = \varepsilon \qquad P(S_2|\sim\pi,S_1) = P(S_2|\sim\pi) = \delta.$$

(ii) Direct application of Bayes' theorem gives

$$P(\pi|S_1,S_2) = \frac{P(S_1,S_2|\pi)P(\pi)}{P(S_1,S_2|\pi)P(\pi)+P(S_1,S_2|\sim\pi)P(\sim\pi)} = \frac{\varepsilon^2\mu}{\varepsilon^2\mu+\delta^2(1-\mu)}.$$

Here we have again assumed that the two measurements are independent, that is,

$$P(S_1,S_2|\pi) = P(S_1|\pi)P(S_2|\pi) = \varepsilon^2 \qquad P(S_1,S_2|\sim\pi) = P(S_1|\sim\pi)P(S_2|\sim\pi) = \delta^2.$$

Both results are thus the same when we assume that the two measurements are independent.

Exercise 3.1

Because averaging is a linear operation we have

$$\begin{aligned} <\Delta x^2> &= <(x-\bar{x})^2> = <x^2> -2\bar{x}<x> + <x>^2 \\ &= <x^2> -2<x>^2 + <x>^2 = <x^2> - <x>^2 . \end{aligned}$$

Exercise 3.2

The covariance matrix can be written as

$$V_{ij} = <(x_i-\bar{x}_i)(x_j-\bar{x}_j)> = <x_ix_j> - <x_i><x_j> .$$

For independent variables the joint probability factorizes $p(x_i,x_j|I) = p(x_i|I)p(x_j|I)$ so that

$$<x_ix_j> = \int \mathrm{d}x_i\, x_i p(x_i|I) \int \mathrm{d}x_j\, x_j p(x_j|I) = <x_i><x_j> .$$

This implies that the off-diagonal elements of V_{ij} vanish.

Exercise 3.3

(i) Substituting $x = x_0$ in the Breit-Wigner formula gives for the maximum value $2/(\pi\Gamma)$. Substituting $x = x_0 \pm \Gamma/2$ gives a value of $1/(\pi\Gamma)$ which is just half the maximum.

(ii) When we substitute $x_0 = 0$ and $\Gamma = 2$, the Breit-Wigner reduces to the Gauchy distribution

$$p(x|x_0=0,\Gamma=2,I) = \frac{1}{\pi}\,\frac{1}{1+x^2}.$$

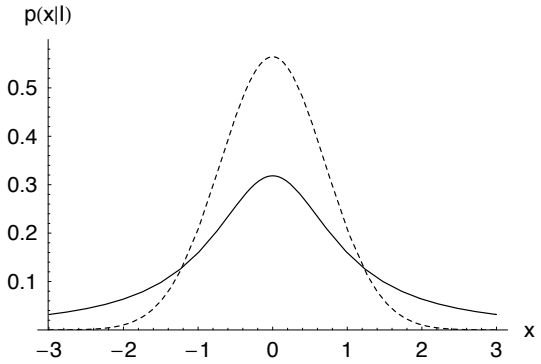
For this distribution $L = \ln\pi + \ln(1+x^2)$. The first and second derivatives of L are given by

$$\frac{\mathrm{d}L}{\mathrm{d}x} = \frac{2x}{1+x^2} \qquad \text{and} \qquad \frac{\mathrm{d}^2L}{\mathrm{d}x^2} = -\frac{4x^2}{(1+x^2)^2} + \frac{2}{1+x^2}.$$

From $\mathrm{d}L/\mathrm{d}x = 0$ we find $\hat{x} = 0$. It follows that the second derivative of L at $\hat{y}$ is 2 so that the width of the distribution is $\sigma = 1/\sqrt{2}$. This gives the Gaussian approximation

$$p(x|x_0=0,\Gamma=2,I) \approx \frac{\exp(-x^2)}{\sqrt{\pi}}.$$

This Gaussian approximation is pretty bad (dashed curve) when compared to the original Cauchy distribution (full curve).



Exercise 3.4

For $z = x + y$ and $z = xy$ we have

$$\begin{aligned} p(z|I) &= \iint \delta(z - x - y)\, f(x)g(y)\, \mathrm{d}x\mathrm{d}y = \int f(z-y)g(y)\, \mathrm{d}y \qquad \text{and} \\ p(z|I) &= \iint \delta(z - xy)\, f(x)g(y)\, \mathrm{d}x\mathrm{d}y = \iint \delta(z - w)\, f(w/y)g(y)\, \frac{\mathrm{d}w}{|y|}\mathrm{d}y = \int f(z/y)g(y)\, \frac{\mathrm{d}y}{|y|}. \end{aligned}$$

Exercise 3.5

(i) The inverse transformations are

$$x = \frac{u+v}{2} \qquad y = \frac{u-v}{2}$$

so that the determinant of the Jacobian is

$$|J| = \left| \begin{array}{cc} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{array} \right| = \left| \begin{array}{cc} 1/2 & 1/2 \\ 1/2 & -1/2 \end{array} \right| = \frac{1}{2}.$$

The joint distribution of u and v is thus

$$p(u,v) = p(x,y)\,|J| = \frac{1}{2} f(x)g(y) = \frac{1}{2} f\left(\frac{u+v}{2}\right) g\left(\frac{u-v}{2}\right).$$

Integration over v gives

$$p(u) = \int p(u,v)\, \mathrm{d}v = \frac{1}{2}\int f\left(\frac{u+v}{2}\right) g\left(\frac{u-v}{2}\right)\, \mathrm{d}v = \int f(w)g(u-w)\, \mathrm{d}w$$

which is just (3.15).

(ii) Here we have for the inverse transformation and the Jacobian determinant

$$x = \sqrt{uv} \qquad y = \sqrt{\frac{u}{v}} \qquad |J| = \left| \begin{array}{cc} v(4uv)^{-1/2} & u(4uv)^{-1/2} \\ (4uv)^{-1/2} & -u(4uv^3)^{-1/2} \end{array} \right| = \frac{1}{2v}$$

which gives for the joint distribution

$$p(u,v) = \frac{1}{2v} f\left(\sqrt{uv}\right) g\left(\sqrt{\frac{u}{v}}\right).$$

Marginalisation of v gives

$$p(u) = \int p(u,v)\, \mathrm{d}v = \int \frac{\mathrm{d}v}{2v}\, f\left(\sqrt{uv}\right) g\left(\sqrt{\frac{u}{v}}\right) = \int \frac{\mathrm{d}w}{w}\, f(w) g\left(\frac{u}{w}\right)$$

which is just (3.16).

Exercise 3.6

According to (3.15) the distribution of z is given by

$$p(z|I) = \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} \mathrm{d}x \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{z-x-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

The standard way to deal with such an integral is to 'complete the squares', that is, to write the exponent in the form $a(x-b)^2 + c$. This allows to carry out the integral

$$\int_{-\infty}^{\infty} \mathrm{d}x \exp\left\{-\tfrac{1}{2}\left[a(x-b)^2 + c\right]\right\} = \exp\left(-\tfrac{1}{2}c\right)\int_{-\infty}^{\infty} \mathrm{d}y \exp\left(-\tfrac{1}{2}ay^2\right) = \sqrt{\frac{2\pi}{a}}\exp\left(-\tfrac{1}{2}c\right).$$

Our problem is now reduced to finding the coefficients a, b and c such that the following equation holds

$$p(x-r)^2 + q(x-s)^2 = a(x-b)^2 + c$$

where the left-hand side is a generic expression for the exponent of the convolution of our two Gaussians. Since the coefficients of the powers of x must be equal at both sides of the equation we have

$$p+q=a \qquad pr+qs=ab \qquad pr^2+qs^2=ab^2+c.$$

Solving these equations for a, b and c gives

$$a = p+q \qquad b = \frac{pr+qs}{p+q} \qquad c = \frac{pq}{p+q}(s-r)^2.$$

Then substituting

$$p = 1/\sigma_1^2 \qquad q = 1/\sigma_2^2 \qquad r = \mu_1 \qquad s = z-\mu_2$$

yields the desired result

$$p(z|I) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left[-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right].$$

Exercise 3.7

For independent random variables x_i with variance σ_i^2 we have $<\Delta x_i \Delta x_j> = \sigma_i\sigma_j\delta_{ij}$.

(i) For the sum $z = \sum x_i$ we have $\partial z/\partial x_i = 1$ so that (3.20) gives

$$<\Delta z^2> = \sum_i\sum_j \sigma_i\sigma_j\delta_{ij} = \sum_i \sigma_i^2.$$

(ii) For the product $z = \prod x_i$ we have $\partial z/\partial x_i = z/x_i$ so that (3.20) gives

$$<\Delta z^2> = \sum_i\sum_j \frac{z}{x_i}\frac{z}{x_j}\sigma_i\sigma_j\delta_{ij} = z^2\sum_i\left(\frac{\sigma_i}{x_i}\right)^2.$$

Exercise 3.8

We have

$$\varepsilon = \frac{n}{N} = \frac{n}{n+m} \qquad \frac{\partial\varepsilon}{\partial n} = \frac{m}{(n+m)^2} = \frac{N-n}{N^2} \qquad \frac{\partial\varepsilon}{\partial m} = -\frac{n}{(n+m)^2} = -\frac{n}{N^2}$$

and

$$<\Delta n^2> = n \qquad <\Delta m^2> = m = N-n \qquad <\Delta n\Delta m> = 0.$$

Inserting this in (3.20) gives for the variance of ε

$$\begin{aligned} <\Delta\varepsilon^2> &= \left(\frac{\partial\varepsilon}{\partial n}\right)^2 <\Delta n^2> + \left(\frac{\partial\varepsilon}{\partial m}\right)^2 <\Delta m^2> \\ &= \left(\frac{N-n}{N^2}\right)^2 n + \left(\frac{n}{N^2}\right)^2 (N-n) = \frac{\varepsilon(1-\varepsilon)}{N}. \end{aligned}$$

Exercise 3.9

The cumulative distribution is defined by

$$c(x) = \int_x p(y|I)\,\mathrm{d}y \qquad \Rightarrow \qquad \frac{\mathrm{d}c}{\mathrm{d}x} = p(x|I) \qquad \text{and thus} \qquad p(c|I) = p(x|I)\left|\frac{\mathrm{d}x}{\mathrm{d}c}\right| = 1.$$

Exercise 3.10

Writing out (3.25) in components gives, using the definition (3.26)

$$\sum_k V_{ik}U^{\mathrm{T}}_{kj} = \sum_k U^{\mathrm{T}}_{ik}\lambda_k\delta_{kj} \quad \Rightarrow \quad \sum_k V_{ik}u^j_k = \lambda_j u^j_i \quad \Rightarrow \quad \boldsymbol{V}\boldsymbol{u}^j = \lambda_j\boldsymbol{u}^j.$$

Exercise 3.11

(i) For a symmetric matrix $\boldsymbol{V}$ and two arbitrary vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ we have

$$\boldsymbol{y}\,\boldsymbol{V}\boldsymbol{x} = \sum_{ij} y_i\,V_{ij}\,x_j = \sum_{ij} x_j V^{\mathrm{T}}_{ji}\,y_i = \sum_{ij} x_j V_{ji}\,y_i = \boldsymbol{x}\,\boldsymbol{V}\boldsymbol{y}.$$

(ii) Because $\boldsymbol{V}$ is symmetric we have

$$\boldsymbol{u}^i\,\boldsymbol{V}\boldsymbol{u}^j = \boldsymbol{u}^j\,\boldsymbol{V}\boldsymbol{u}^i \quad \Rightarrow \quad \lambda_i\boldsymbol{u}_i\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_i\boldsymbol{u}_j \quad \text{or} \quad (\lambda_i-\lambda_j)\,\boldsymbol{u}_i\boldsymbol{u}_j = 0 \quad \Rightarrow \quad \boldsymbol{u}_i\boldsymbol{u}_j = 0 \quad \text{for} \quad \lambda_i \neq \lambda_j.$$

(iii) If $\boldsymbol{V}$ is positive definite we have $\boldsymbol{u}^i\boldsymbol{V}\boldsymbol{u}^i = \lambda_i\boldsymbol{u}^i\boldsymbol{u}^i = \lambda_i > 0$.

Exercise 4.1

Expansion of $P(R_2|I)$ in the hypothesis space $\{R_1, W_1\}$ gives

$$\begin{aligned} P(R_2|I) &= P(R_2,R_1|I) + P(R_2,W_1|I) \\ &= P(R_2|R_1,I)P(R_1|I) + P(R_2|W_1,I)P(W_1|I) \\ &= \frac{R-1}{N-1}\frac{R}{N} + \frac{R}{N-1}\frac{W}{N} = \frac{R}{N}. \end{aligned}$$

Exercise 4.2

For draws with replacement we have $P(R_2|R_1,I) = P(R_2|W_1,I) = P(R_1|I) = R/N$ and $P(W_1|I) = W/N$. Inserting this in Bayes' theorem gives

$$P(R_1|R_2,I) = \frac{P(R_2|R_1,I)P(R_1|I)}{P(R_2|R_1,I)P(R_1|I) + P(R_2|W_1,I)P(W_1|I)} = \frac{R}{N}.$$

Exercise 4.3

The likelihood for N signals by N particles is $P(N|N,\varepsilon) = \varepsilon^N$. For a flat prior, the normalised posterior is $p(\varepsilon|N,N) = (N+1)\,\varepsilon^N$. The α confidence interval is calculated from

$$\int_0^a p(\varepsilon|N,N)\mathrm{d}\varepsilon = a^{N+1} = 1-\alpha \qquad \Rightarrow \qquad a = (1-\alpha)^{1/(N+1)}.$$

Putting $N=4$ and $\alpha = 0.65$ gives $a = 0.81$ so that $\varepsilon = 1^{+0}_{-0.19}$ at 65% CL.

Exercise 4.4

To calculate the likelihood $P(n|h,\mu)$, we expand it in the set $N = \{n, n+1, \ldots\}$

$$\begin{aligned} P(n|h,\mu) &= \sum_{N=n}^{\infty} P(n,N|h,\mu) = \sum_{N=n}^{\infty} P(n|N,h)\;P(N|\mu) \\ &= \sum_{N=n}^{\infty} \frac{N!}{n!(N-n)!}h^n(1-h)^{N-n}\;\frac{\mu^N}{N!}e^{-\mu} \\ &= \frac{(\mu h)^n}{n!}e^{-\mu}\sum_{N=n}^{\infty}\frac{1}{(N-n)!}(\mu-\mu h)^{N-n} = \frac{(\mu h)^n}{n!}e^{-\mu}\;e^{(\mu-\mu h)} = \frac{(\mu h)^n}{n!}e^{-\mu h} \end{aligned}$$

It is illustrative to try to get a posterior for h by inverting $P(n|h,\mu)$. Assuming a uniform prior for h and μ we obtain

$$p(h,\mu|n) = C\ P(n|h,\mu) = C\ \frac{(\mu h)^n}{n!} e^{-\mu h}.$$

Integration over μ gives

$$p(h|n) = \frac{C}{n!\,h} \int_0^\infty z^n e^{-z} \mathrm{d}z = \frac{C}{h}.$$

But this posterior does not depend on n and is also improper (not normalisable). We do not encounter here a break-down of Bayesian probability theory but, instead, a warning that n by itself does not contain enough information to properly conduct inference on h. Indeed, a little thought reveals that we need both the number of heads (n_1) and the number of tails (n_2), and also that the knowledge of n_1 tells us nothing about n_2 and *vice versa*. These numbers are thus uncorrelated so that

$$P(n_1,n_2|h,\mu) = P(n_1|h,\mu)\ P(n_2|h,\mu) = \frac{(h\mu)^{n_1}\ [(1-h)\mu]^{n_2}}{n_1!\,n_2!} e^{-\mu}.$$

Assuming a uniform prior for μ, we obtain for the posterior

$$\begin{aligned} p(h|n_1,n_2) &\propto p(h|I) \int_0^\infty p(h,\mu|n_1,n_2)\,\mathrm{d}\mu \ \propto\ p(h|I)\,h^{n_1}(1-h)^{n_2} \int_0^\infty \mu^{(n_1+n_2)} e^{-\mu} \\ &= C\,h^{n_1}\ (1-h)^{n_2}\ p(h|I), \end{aligned}$$

which is the same as (4.16) or (4.17).

Exercise 4.5

Without loss of generality we can consider marginalisation of the multinomial distribution over all but the first probability. According to (4.19) we have

$$n_2' = \sum_{i=2}^{k} n_i = N - n_1 \qquad \text{and} \qquad p_2' = \sum_{i=2}^{k} p_i = 1 - p_1.$$

Inserting this in (4.18) gives the binomial distribution

$$P(n_1, n_2' \,|\, p_1, p_2', N) = \frac{N!}{n!(N-n)!}\ p_1^{n_1}(1-p_1)^{N-n_1}.$$

Exercise 4.6

From the product rule we have

$$P(n_1,\ldots,n_k|I) = P(n_1,\ldots,n_{k-1}|n_k,I)P(n_k|I)$$

From this we find for the conditional distribution

$$\begin{aligned} P(n_1,\ldots,n_{k-1}|n_k,I) &= \frac{P(n_1,\ldots,n_k|I)}{P(n_k|I)} \\ &= \frac{N!}{n_1!\cdots n_k!} p_1^{n_1}\cdots p_{k-1}^{n_{k-1}} p_k^{n_k} \times \frac{n_k!(N-n_k)!}{N!} \frac{1}{p_k^{n_k}(1-p_k)^{N-n_k}} \\ &= \frac{(N-n_k)!}{n_1!\cdots n_{k-1}!} \left(\frac{p_1}{1-p_k}\right)^{n_1} \cdots \left(\frac{p_{k-1}}{1-p_k}\right)^{n_{k-1}}. \end{aligned}$$

Exercise 4.7

The solution to this exercise is very similar to that of Exercise 4.4.

Exercise 4.8

(i) The likelihood to observe n counts in a time window Δt is given by the Poisson distribution

$$P(n|\mu) = \frac{\mu^n}{n!}\ \exp(-\mu)$$

with $\mu = R\Delta t$ and R the average counting rate. Assuming a flat prior for $\mu \in [0, \infty]$ the posterior is

$$p(\mu|n) = C\,\frac{\mu^n}{n!}\,\exp(-\mu)$$

with C a normalisation constant which turns out to be unity: the Poisson distribution has the remarkable property that it is normalized with respect to both n and μ:

$$\sum_{n=0}^{\infty} P(n|\mu) = \sum_{n=0}^{\infty} \frac{\mu^n}{n!}\,\exp(-\mu) = 1 \qquad \text{and} \qquad \int_0^{\infty} p(\mu|n)\,\mathrm{d}\mu = \int_0^{\infty} \frac{\mu^n}{n!}\,\exp(-\mu)\,\mathrm{d}\mu = 1.$$

The mean, second moment and the variance of the posterior are

$$<\mu> = n+1, \qquad <\mu^2> = (n+1)(n+2), \qquad <\Delta\mu^2> = n+1.$$

The log posterior and the derivatives are

$$L = \text{constant} - n\ln(\mu) + \mu, \qquad \frac{\mathrm{d}L}{\mathrm{d}\mu} = 1 - \frac{n}{\mu}, \qquad \frac{\mathrm{d}^2L}{\mathrm{d}\mu^2} = \frac{n}{\mu^2}.$$

Setting the derivative to zero we find $\hat{\mu} = n$. The square root of the inverse of the Hessian at the mode gives for the width $\sigma = \sqrt{n}$.

(ii) The probability $p(\tau|R, I)\,\mathrm{d}\tau$ that the time interval between the passage of two particles is between τ and $\tau + \mathrm{d}\tau$ is given by

$$\begin{aligned} p(\tau|R,I)\,\mathrm{d}\tau &= P(\text{`no particle passes during } \tau\text{'}) \times P(\text{`one particle passes during } \mathrm{d}\tau\text{'}) \\ &= \exp(-R\tau) \times R\mathrm{d}\tau. \end{aligned}$$

Exercise 4.9

The derivatives of the characteristic function (4.26) are

$$\frac{\mathrm{d}\phi(k)}{\mathrm{d}k} = (i\mu - k\sigma^2)\exp\left(i\mu k - \frac{1}{2}\sigma^2k^2\right) \qquad \frac{\mathrm{d}^2\phi(k)}{\mathrm{d}k^2} = \left[(i\mu - k\sigma^2)^2 - \sigma^2\right]\exp\left(i\mu k - \frac{1}{2}\sigma^2k^2\right).$$

From (4.25) we find for the first and second moment

$$<x> = \frac{1}{i}\frac{\mathrm{d}\phi(0)}{\mathrm{d}k} = \mu \qquad <x^2> = \frac{1}{i^2}\frac{\mathrm{d}^2\phi(0)}{\mathrm{d}k^2} = \mu^2 + \sigma^2$$

from which it immediately follows that the variance is given by $<x^2> - <x>^2 = \sigma^2$.

Exercise 4.10

From (4.24) and (4.26) we have

$$\phi(k) = \phi_1(k)\,\phi_2(k) = \exp\left[i(\mu_1+\mu_2)k - \frac{1}{2}(\sigma_1^2+\sigma_2^2)k^2\right]$$

which is just the characteristic function of a Gauss with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

Exercise 5.2

From differentiating the logarithm of (5.14) we get

$$\begin{aligned} -\frac{\partial \ln Z}{\partial \lambda_k} &= -\frac{1}{Z}\frac{\partial Z}{\partial \lambda_k} = -\frac{1}{Z}\sum_{i=1}^{n} m_i\,\frac{\partial}{\lambda_k}\exp\left(\sum_{j=1}^{m}\lambda_j f_{ji}\right) \\ &= \frac{1}{Z}\sum_{i=1}^{n} f_{ki}\,m_i\exp\left(\sum_{j=1}^{m}\lambda_j f_{ji}\right) = \sum_{i=1}^{n} f_{ki}\,p_i = \beta_k. \end{aligned}$$

Exercise 6.1

The log posterior of (6.3) is given by

$$L = \text{Constant} + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{d_i-\mu}{\sigma}\right)^2.$$

Setting the first derivative to zero gives

$$\frac{\mathrm{d}L}{\mathrm{d}\mu} = -\sum_{i=1}^{n}\left(\frac{d_i-\mu}{\sigma^2}\right) = 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} d_i.$$

The Hessian and the square root of its inverse at $\hat{\mu}$ are

$$\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2} = \sum_{i=1}^{n}\frac{1}{\sigma^2} = \frac{n}{\sigma^2} \qquad \Rightarrow \qquad \left[\frac{\mathrm{d}^2L(\hat{\mu})}{\mathrm{d}\mu^2}\right]^{-\frac{1}{2}} = \frac{\sigma}{\sqrt{n}}.$$

Exercise 6.4

By substituting $t = \frac{1}{2}\chi^2$ we find for the average

$$<\chi^2> = \int_0^\infty \chi^2 p(\chi^2|\nu)\,\mathrm{d}\chi^2 = \frac{2}{\Gamma(\alpha)}\int_0^\infty t^\alpha e^{-t}\,\mathrm{d}t = \frac{2\Gamma(\alpha+1)}{\Gamma(\alpha)} = 2\alpha = \nu.$$

Likewise, the second moment is found to be

$$<\chi^4> = \int_0^\infty \chi^4 p(\chi^2|\nu)\,\mathrm{d}\chi^2 = \frac{4}{\Gamma(\alpha)}\int_0^\infty t^{\alpha+1} e^{-t}\,\mathrm{d}t = \frac{4\Gamma(\alpha+2)}{\Gamma(\alpha)} = 4\alpha(\alpha+1) = \nu(\nu+2).$$

Therefore the variance is

$$<\chi^4> - <\chi^2>^2 = \nu(\nu+2) - \nu^2 = 2\nu.$$

Exercise 6.6

In case of a polynomial parameterisation

$$f(x;\boldsymbol{a}) = a_1 + a_2x + a_3x^2 + a_4x^3 + \cdots$$

the basis functions are given by $f_\lambda(x) = x^{\lambda-1}$. To give an example, for a quadratic polynomial the equation (6.24) takes the form

$$\begin{pmatrix} \sum_i w_i & \sum_i w_i x_i & \sum_i w_i x_i^2 \\ \sum_i w_i x_i & \sum_i w_i x_i^2 & \sum_i w_i x_i^3 \\ \sum_i w_i x_i^2 & \sum_i w_i x_i^3 & \sum_i w_i x_i^4 \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{pmatrix} = \begin{pmatrix} \sum_i w_i d_i \\ \sum_i w_i d_i x_i \\ \sum_i w_i d_i x_i^2 \end{pmatrix}.$$

Exercise 6.7

In case $f(x;\boldsymbol{a}) = a_1$ (fit to a constant) the matrix $\boldsymbol{W}$ and the vector $\boldsymbol{b}$ are one-dimensional:

$$W = \sum_i w_i, \qquad b = \sum_i w_i d_i.$$

It then immediately follows from Eqs. (6.24) and (6.26) that

$$\hat{a}_1 = \frac{\sum_i w_i d_i}{\sum_i w_i} \pm \frac{1}{\sqrt{\sum_i w_i}}.$$

Exercise 6.8

The covariance matrix $\boldsymbol{V}'$ of $\boldsymbol{r}$ and $\boldsymbol{s}$ is given by (6.28) as

$$V'_{ij} = \sigma_i^2 \delta_{ij}, \qquad V'_{\lambda\mu} = \delta_{\lambda\mu}, \qquad V'_{i\lambda} = 0.$$

Differentiation of (6.27) gives for the derivative matrix $\boldsymbol{D}$

$$D_{ij} = \frac{\partial d_i}{\partial r_j} = \delta_{ij}, \qquad D_{i\lambda} = \frac{\partial d_i}{\partial s_\lambda} = \Delta_{i\lambda}.$$

Carrying out the matrix multiplication $\boldsymbol{V} = \boldsymbol{D}\boldsymbol{V}'\boldsymbol{D}^{\mathrm{T}}$ we find

$$\begin{aligned} V_{ij} &= \sum_k \sum_l D_{ik} V'_{kl} D^{\mathrm{T}}_{lj} + \sum_\lambda \sum_\mu D_{i\lambda} V'_{\lambda\mu} D^{\mathrm{T}}_{\mu j} \\ &= \sum_k \sum_l \delta_{ik} \sigma_k^2 \delta_{kl} \delta_{lj} + \sum_\lambda \sum_\mu \Delta_{i\lambda} \delta_{\lambda\mu} \Delta_{j\mu} \\ &= \sigma_i^2 \delta_{ij} + \sum_\lambda \Delta_{i\lambda} \Delta_{j\lambda}. \end{aligned}$$

Exercise 6.9

From (6.37) we have for the log posterior

$$L(\boldsymbol{a}, \boldsymbol{s}) = \text{Constant} + \tfrac{1}{2} \sum_{i=1}^{n} w_i \left(d_i - t_i(\boldsymbol{a}) - \sum_{\lambda=1}^{m} s_\lambda \Delta_{i\lambda} \right)^2 + \tfrac{1}{2} \sum_{\lambda=1}^{m} s_\lambda^2.$$

Setting the derivative to zero leads to

$$\frac{\partial L(\boldsymbol{a}, \boldsymbol{s})}{\partial s_\lambda} = -\sum_{i=1}^{n} w_i (d_i - t_i) \Delta_{i\lambda} + \sum_{\mu=1}^{m} s_\mu \sum_{i=1}^{n} w_i \Delta_{i\lambda} \Delta_{i\mu} + s_\lambda = 0.$$

This equation can be rewritten as

$$\sum_{\mu=1}^{m} s_\mu \left(\delta_{\lambda\mu} + \sum_{i=1}^{n} w_i \Delta_{i\lambda} \Delta_{i\mu} \right) = \sum_{i=1}^{n} w_i (d_i - t_i) \Delta_{i\lambda}.$$

But this is just the matrix equation $\sum_\mu S_{\lambda\mu} s_\mu = b_\lambda$ as given in (6.39).

Exercise 6.10

The best estimate $\hat{\mu}$ of the temperature is, according to Exercise 6.7, given by the weighted average. Setting $d_i = d$ and $w_i = 1/\sigma_i^2 = 1/\sigma^2$, we find for the average and variance of n measurements

$$\hat{\mu} = \frac{\sum w_i d_i}{\sum w_i} = d \qquad \text{and} \qquad <\Delta\hat{\mu}^2> = \frac{1}{\sum w_i} = \frac{\sigma^2}{n}.$$

1. Offsetting all data points by an amount $\pm\Delta$ gives for the best estimate $\hat{\mu}^{\pm} = d \pm \Delta$. Adding the statistical and systematic deviations in quadrature we thus find from the offset method that

$$\hat{\mu} = d \pm \sqrt{\frac{\sigma^2}{n} + \Delta^2} \qquad \Rightarrow \qquad \hat{\mu} = d \pm \Delta \qquad \text{for} \qquad n \to \infty.$$

2. The matrix $\boldsymbol{S}$ and the vector $\boldsymbol{b}$ defined in (6.39) are in this case one-dimensional. We have

$$S = 1 + n \left(\frac{\Delta}{\sigma} \right)^2, \qquad b = \frac{n(d-\mu)\Delta}{\sigma^2}, \qquad \sum w_i (d_i - t_i)^2 = \frac{n(d-\mu)^2}{\sigma^2}.$$

Inserting this in (6.42) we find, after some algebra

$$L(\mu) = \frac{1}{2}\,\frac{n(d-\mu)^2}{\sigma^2 + n\Delta^2}, \qquad \frac{\mathrm{d}L(\mu)}{\mathrm{d}\mu} = -\frac{n(d-\mu)}{\sigma^2 + n\Delta^2}, \qquad \frac{\mathrm{d}^2L(\mu)}{\mathrm{d}\mu^2} = \frac{n}{\sigma^2 + n\Delta^2}.$$

Setting the first derivative to zero gives $\hat{\mu} = d$. Equating the error as the square root of the inverse of the Hessian (second derivative at $\hat{\mu}$) obtains the same result as from the offset method:

$$\hat{\mu} = d \pm \sqrt{\frac{\sigma^2}{n} + \Delta^2} \qquad \Rightarrow \qquad \hat{\mu} = d \pm \Delta \qquad \text{for} \qquad n \to \infty.$$

We now add a second set of n measurements which do not have a systematic error Δ. The weighted average gives for the best estimate $\hat{\mu}$ and the variance of the combined data

$$\hat{\mu} = d \qquad \text{and} \qquad <\Delta\hat{\mu}^2> = \frac{\sigma^2}{2n}.$$

1. Offsetting the first set of n data points by an amount $\pm\Delta$ but leaving the second set intact gives for the best estimate $\hat{\mu}^{\pm} = d \pm \Delta/2$. Adding the statistical and systematic deviations in quadrature we thus find from the offset method that

   $$\hat{\mu} = d \pm \sqrt{\frac{\sigma^2}{2n} + \left(\frac{\Delta}{2}\right)^2} \qquad \Rightarrow \qquad \hat{\mu} = d \pm \frac{\Delta}{2} \qquad \text{for} \qquad n \to \infty.$$

   But this error is larger than if we would have considered only the second data set:

   $$\hat{\mu} = d \pm \frac{\sigma}{\sqrt{n}} \qquad \Rightarrow \qquad \hat{\mu} = d \pm 0 \qquad \text{for} \qquad n \to \infty$$

   In other words, the offset method violates the requirement that the error derived from all available data must always be smaller than that derived from a subset of the data.

2. The matrix $\boldsymbol{S}$ and the vector $\boldsymbol{b}$ defined in (6.39) are now two-dimensional but with many zero-valued elements since the systematic error Δ of the second data set is zero. We find

   $$\boldsymbol{S} = \begin{pmatrix} S & 0 \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{b} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

   where S and b are defined above. The log posterior of (6.42) is found to be

   $$L(\mu) = \frac{1}{2}\left[2n\left(\frac{d-\mu}{\sigma}\right)^2 - \boldsymbol{b}\,\boldsymbol{S}^{-1}\boldsymbol{b}\right] = \frac{1}{2}\,\frac{n(d-\mu)^2}{\sigma^2 + n\Delta^2}\left(2 + n\frac{\Delta^2}{\sigma^2}\right).$$

   Solving the equation $\mathrm{d}L(\mu)/\mathrm{d}\mu = 0$ immediately yields $\hat{\mu} = d$. The inverse of the second derivative gives an estimate of the error. After some straight forward algebra we find

   $$\hat{\mu} = d \pm \left(\frac{\sigma^2}{n} + \Delta^2\right)\left(2 + n\frac{\Delta^2}{\sigma^2}\right)^{-1}.$$

   It is seen that the error vanishes in the limit $n \to \infty$, as it should be.

Exercise 8.1

In (3.11) we state that the posterior $p(\lambda|\boldsymbol{d}, H_1)$ is Gaussian distributed in the neighborhood of the mode $\hat{\lambda}$. Indeed, that is exactly what we find inserting Eqs. (8.4), (8.5) and (8.6) in Bayes theorem,

$$p(\lambda|\boldsymbol{d}, H_1) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\lambda - \hat{\lambda}}{\sigma}\right)^2\right].$$

The approximations made in Eqs. (8.4) and (8.5) are thus consistent with those made in (3.11).

Exercise 8.2

When H_i has n free parameters $\boldsymbol{\lambda}$ and H_j has m free parameters $\boldsymbol{\mu}$, the Occam factor in (8.8) becomes the ratio of multivariate Gaussian normalisation factors, see (3.11):

$$\sqrt{\frac{(2\pi)^m \, |\boldsymbol{V}_\mu|}{(2\pi)^n \, |\boldsymbol{V}_\lambda|}}.$$

References

[Agost94] G. D'Agostini, Nucl. Instr. Meth. **A346**, 306 (1994).

[Agost99] G. D'Agostini, '*Sceptical combination of experimental results: General considerations and application to* ε'/ε', arXiv:hep-ex/9910036 (1999).

[Agost03] G. D'Agostini, '*Bayesian Inference in Processing Experimental Data—Principles and Basic Applications*', arXiv:physics/0304102 (2003).

[Agost05] G. D'Agostini, '*Fits, and especially linear fits, with errors on both axes, extra variance of the data points and other complications*', arXiv:physics/0511182 (2005).

[Alekh00] S.I. Alekhin, '*Statistical properties of estimators using the covariance matrix*', hep-ex/0005042 (2000).

[Alt06] NA49 Collab., C. Alt et al., '*Upper limit of* D^0 *production in central Pb-Pb collisions at 158A GeV*', Phys. Rev. **C73**, 034910 (2006), nucl-ex/0507031.

[Berg88] J.O. Berger and R.L. Wolpert, '*The Likelihood Principle*', Inst. Math. Stat. Lecture Notes-Monograph Series Vol 6 (1988), available from Google Books.

[Bret90] G.L. Bretthorst, '*An Introduction to Parameter Estimation using Bayesian Probability Theory*' in '*Maximum Entropy and Bayesian Methods*', ed. P.F. Fougère, Kluwer Academic Publishers (1990).

[Bret96] G.L. Bretthorst, '*An Introduction to Model Selection using Probability Theory as Logic*' in '*Maximum Entropy and Bayesian Methods*', ed. G.L. Heidbreder, Kluwer Academic Publishers (1996).

[Cox46] R.T. Cox, '*Probability, frequency and reasonable expectation*', Am. J. Phys. **14**, 1 (1946).

[Cox61] R.T. Cox, '*the Algebra of Probable Inference*', Johns Hopkins Press (1961).

[Cowan98] G. Cowan, '*Statistical Data Analysis*', Oxford University Press (1998).

[Eid04] PDG, S. Eidelman et al., Phys. Lett. **B592**, 1 (2004).

[Greg05] P. Gregory, '*Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press (2005).'

[Gull88] S.F. Gull, '*Bayesian Inductive Inference and Maximum Entropy*' in '*Maximum-Entropy and Bayesian Methods in Science and Engineering*', eds. G.J. Ericson and C.R. Smith, Vol. I, Kluwer Academic Publishers (1988).

[James00] F. James et al. eds., '*Proc. Workshop on Confidence Limits*', CERN Yellow Report 2000-005.

[James06] F. James, '*Statistical Methods in Experimental Physics*', World Scientific, Singapore (2006).

[Jay85] E.T. Jaynes, '*Bayesian Methods: General Background*', in Proc. '*Maximum Entropy and Bayesian Methods in Applied Statistics*', ed. J.H. Justice, Cambridge University Press (1985).

[Jay98] `http://omega.albany.edu:8008/JaynesBook.html` (1998).

[Jay03] E.T. Jaynes, '*Probability Theory—The Logic of Science*', Cambridge University Press (2003).

[Kass96] E. Kass and L. Wasserman, '*The Selection of Prior Distributions by Formal Rules*', J. Am. Stat. Assoc., Vol. **91**, No. 435 (1996).

[Lor90] T. Loredo, '*From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics*' in '*Maximum Entropy and Bayesian Methods*', ed. P.F. Fougère, Kluwer Academic Publishers (1990).

[Pump02] J. Pumplin et al., '*New generation of parton distributions with uncertainties from global QCD analysis*', JHEP 0207:012 (2002), hep-ph/0201195.

[Shan48] C.E. Shannon, '*The mathematical theory of communication*', Bell Systems Tech. J. **27**, 379 (1948).

[Sivia06] D.S. Sivia, '*Data Analysis—a Bayesian Tutorial*', Oxford University Press (1997); second edition with J. Skilling (2006).

[Stump02] D. Stump et al., '*Uncertainties of predictions from parton distribution functions*', Phys. Rev. **D65**, 014012 (2002), hep-ph/0101051.

[Stump09] D. Stump, private communication.

[Take96] T. Takeuchi, Prog. Theor. Phys. Suppl. **123**, 247 (1996), hep-ph/9603415.

[Zech02] G. Zech, '*Frequentist and Bayesian confidence intervals*', EPJdirect **C12**, 1 (2002).

Index