Part II

Bayesian primer

- slightly reviewed version of the 1995 DESY/Rome report -

Chapter 3

Subjective probability and Bayes' theorem

"The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence."
(Bruno de Finetti)

3.1 Original abstract of the primer

Bayesian statistics is based on the subjective definition of probability as 'degree of belief' and on Bayes' theorem, the basic tool for assigning probabilities to hypotheses combining *a priori* judgements and experimental information. This was the original point of view of Bayes, Bernoulli, Gauss, Laplace, etc. and contrasts with later conventional (pseudo-)definitions of probabilities, which implicitly presuppose the concept of probability. These notes[1] show that the Bayesian approach is the natural one for data analysis in the most general sense, and for assigning uncertainties to the results of physical measurements, while at the same time resolving philosophical aspects of the problem. The approach, although little known and usually misunderstood among the high-energy physics (HEP) community, has become the standard way of reasoning in several fields of research and has recently been adopted by the international metrology organizations in their recommendations for assessing measurement uncertainty.

These notes describe a general model for treating uncertainties originating from random and systematic errors in a consistent way and include examples of applications of the model in HEP, e.g. confidence intervals in different contexts, upper/lower limits, treatment of systematic errors, hypothesis tests and unfolding.

3.2 Introduction to the primer

The purpose of a measurement is to determine the value of a physical quantity. One often speaks of the true value, an idealized concept achieved by an infinitely precise and accurate measurement, i.e. immune from errors. In practice the result of a measurement is expressed in terms of the best estimate of the true value and of a related uncertainty. Traditionally the various

[1] These notes are based on lectures given to graduate students in Rome (May 1995) and summer students at DESY (September 1995). The original report is Ref. [27]. In the present report, notes (indicated by **Note added**) are used either for clarification or to refer to those parts not contained in the original primer.

contributions to the overall uncertainty are classified in terms of 'statistical' and 'systematic' uncertainties: expressions which reflect the sources of the experimental errors (the quotation marks indicate that a different way of classifying uncertainties will be adopted here).

Statistical uncertainties arise from variations in the results of repeated observations under (apparently) identical conditions. They vanish if the number of observations becomes very large (*"the uncertainty is dominated by systematics"* is the typical expression used in this case) and can be treated — in most cases, but with some exceptions of great relevance in HEP — using conventional statistics based on the frequency-based definition of probability.

On the other hand, it is not possible to treat systematic uncertainties coherently in the frequentistic framework. Several ad hoc prescriptions for how to combine statistical and systematic uncertainties can be found in textbooks and in the literature: *"add them linearly"*; *"add them linearly if . . . , else add them quadratically"*; *"don't add them at all"*, and so on (see, e.g., Part 3 of Ref. [1]). The fashion at the moment is to add them quadratically if they are considered independent, or to build a covariance matrix of statistical and systematic uncertainties to treat general cases. These procedures are not justified by conventional statistical theory, but they are accepted because of the pragmatic good sense of physicists. For example, an experimentalist may be reluctant to add twenty or more contributions linearly to evaluate the uncertainty of a complicated measurement, or decide to treat the correlated systematic uncertainties statistically, in both cases unaware of, or simply not caring about, violating frequentistic principles.

The only way to deal with these and related problems in a consistent way is to abandon the frequentistic interpretation of probability introduced at the beginning of this century, and to recover the intuitive concept of probability as degree of belief. Stated differently, one needs to associate the idea of probability with the lack of knowledge, rather than to the outcome of repeated experiments. This has been recognized also by the International Organization for Standardization (ISO), which assumes the subjective definition of probability in its *"Guide to the expression of uncertainty in measurement"* [3].

This primer is organized as follows:

- Sections 3.3–3.6 give a general introduction to subjective probability.
- Sections 4.1–4.2 summarize some concepts and formulae concerning random variables, needed for many applications.
- Section 5.1 introduces the problem of measurement uncertainty and deals with the terminology.
- Sections 5.2–5.3 present the inferential model.
- Sections 5.4–5.6 show several physical applications of the model.
- Section 6.1 deals with the approximate methods needed when the general solution becomes complicated; in this context the ISO recommendations will be presented and discussed.
- Section 6.2 deals with uncertainty propagation. It is particularly short because, in this scheme, there is no difference between the treatment of systematic uncertainties and indirect measurements; the section simply refers to the results of Sections 5.4–6.1.
- Section 6.3 is dedicated to a detailed discussion about the covariance matrix of correlated data and the trouble it may cause.
- Section 7.1 was added as an example of a more complicated inference (multidimensional unfolding) than those treated in Sections 5.4–6.2.

3.3 Probability

3.3.1 What is probability?

The standard answers to this question are

1. the ratio of the number of favourable cases to the number of all cases;
2. the ratio of the number of times the event occurs in a test series to the total number of trials in the series.

It is very easy to show that neither of these statements can define the concept of probability:

- Definition 1 lacks the clause 'if all the cases are equally probable'. This has been done here intentionally, because people often forget it. The fact that the definition of probability makes use of the term 'probability' is clearly embarrassing. Often in textbooks the clause is replaced by 'if all the cases are equally possible', ignoring that in this context 'possible' is just a synonym of 'probable'. There is no way out. This statement does not define probability but gives, at most, a useful rule for evaluating it – assuming we know what probability is, i.e. of what we are talking about. The fact that this definition is labelled 'classical' or 'Laplace' simply shows that some authors are not aware of what the 'classicals' (Bayes, Gauss, Laplace, Bernoulli, etc.) thought about this matter. We shall call this definition 'combinatorial'.
- Definition 2 is also incomplete, since it lacks the condition that the number of trials must be very large (it goes to infinity). But this is a minor point. The crucial point is that the statement merely defines the relative frequency with which an event (a phenomenon) occurred in the past. To use frequency as a measurement of probability we have to assume that the phenomenon occurred in the past, and will occur in the future, with the same probability. But who can tell if this hypothesis is correct? Nobody: <u>we</u> have to guess in every single case. Note that, while in the first definition the assumption of equal probability was explicitly stated, the analogous clause is often missing from the second one. We shall call this definition 'frequentistic'.

We have to conclude that if we want to make use of these statements to assign a numerical value to probability, in those cases in which we judge that the clauses are satisfied, we need a better definition of probability.

3.3.2 Subjective definition of probability

So, what is probability? Consulting a good dictionary helps. Webster's states, for example, that *"probability is the quality, state, or degree of being probable"*, and then that 'probable' means *"supported by evidence strong enough to make it likely though not certain to be true"*. The concept of probable arises in reasoning when the concept of certain is not applicable. If we cannot state firmly whether an event (we use this word as a synonym for any possible statement, or proposition, relative to past, present or future) is true or false, we just say that it is possible or probable. Different events may have different levels of probability, depending whether we think that they are more likely to be true or false (see Fig. 3.1).

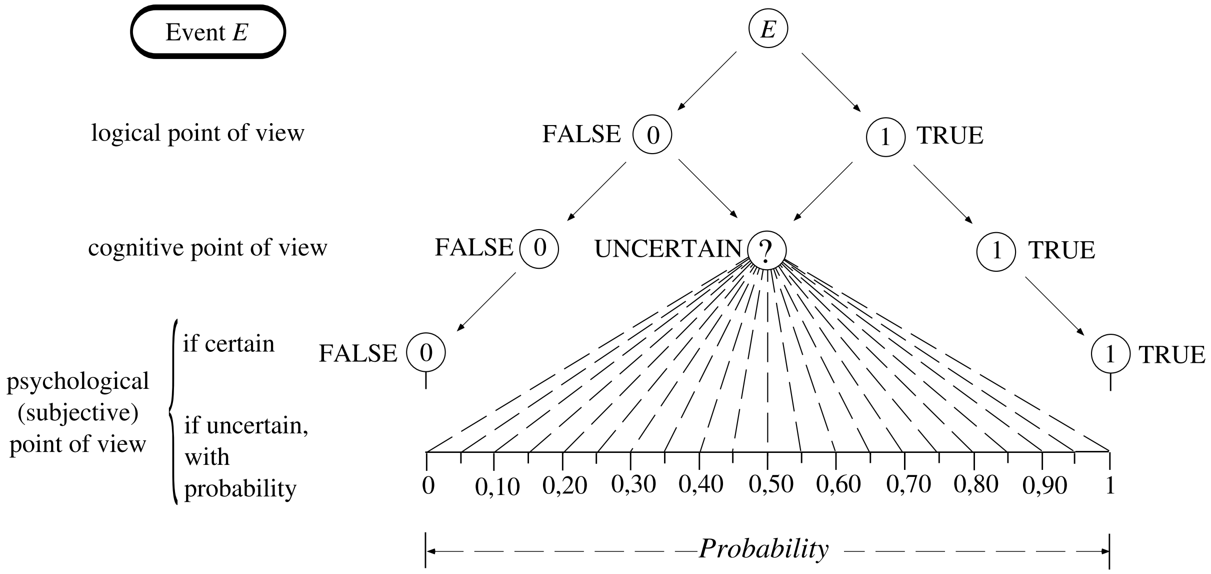

Figure 3.1: Certain and uncertain events [28].

The concept of probability is then simply

a measure of the degree of belief that an event will[2] *occur.*

This is the kind of definition that one finds in Bayesian books (see e.g. Refs. [11, 19, 29, 30, 31]) and the formulation cited here is that given in the ISO Guide [3], which we will discuss later.

At first sight this definition does not seem to be superior to the combinatorial or the frequentistic ones. At least they give some practical rules to calculate 'something'. Defining probability as degree of belief seems too vague to be of any use. We need, then, some explanation of its meaning and a tool to evaluate it with greater precision than intuitive degrees of beliefs can provide. We will look at this tool (Bayes' theorem) later. We will end this section with some explanatory remarks on the definition, but first let us discuss the advantages of this definition.

- It is natural, very general and can be applied to any imaginable event, independently of the feasibility of making an inventory of all (equally) possible and favourable cases, or of repeating the experiment under conditions of equal probability.

- It avoids the linguistic schizophrenia of having to distinguish 'scientific' probability (i.e. strictly based on 'definitions' 1 and 2 of the previous section) from 'non-scientific' probability used in everyday reasoning, including research activity (a meteorologist might feel offended to hear that evaluating the probability of rain tomorrow is not scientific).

- As far as measurements are concerned, it allows us to talk about the probability of the true value of a physical quantity, or of any scientific hypothesis. In the frequentistic frame it is only possible to talk about the probability of the outcome of an experiment, as the true value is considered to be a constant. This approach is so unnatural that most physicists speak of '95% probability that the mass of the top quark is between ... ', although they believe that the correct definition of probability is the limit of the frequency.

[2]The use of the future tense does not imply that this definition can only be applied for future events. It simply means that the statement will be proven to be true, even if it refers to the past. Think for example of the probability that it was raining in Rome on the day of the battle of Waterloo.

- It is possible to make a very general theory of uncertainty which can take into account any source of statistical or systematic error, independently of their distribution.

To get a better understanding of the subjective definition of probability let us take a look at odds in betting. The higher the degree of belief that an event will occur, the higher the amount of money A that someone (a rational better) is ready to pay in order to receive a sum of money B if the event occurs. Clearly the bet must be acceptable in both directions ('coherent' is the correct adjective), i.e. the amount of money A must be smaller or equal to B and not negative (who would accept such a bet?). The cases of $A = 0$ and $A = B$ mean that the events are considered to be false or true, respectively, and obviously it is not worth betting on certainty. They are just limit cases, and in fact they can be treated with standard logic. It seems reasonable[3] that the amount of money A that one is willing to pay grows linearly with the degree of belief. It follows that if someone thinks that the probability of the event E is p, then he will bet $A = p\,B$ to get B if the event occurs, and to lose $p\,B$ if it does not. It is easy to demonstrate that the condition of coherence implies that $0 \leq p \leq 1$.

What has gambling to do with physics? The definition of probability through betting odds has to be considered operational, although there is no need to make a bet (with whom?) each time one presents a result. It has the important role of forcing one to make an honest assessment of the value of probability that one believes. One could replace money with other forms of gratification or penalization, like the increase or the loss of scientific reputation. Moreover, the fact that this operational procedure is not to be taken literally should not be surprising. Many physical quantities are defined in a similar way. Think, for example, of the textbook definition of the electric field, and try to use it to measure $\vec{E}$ in the proximity of an electron. A nice example comes from the definition of a poisonous chemical compound: *"it would be lethal if ingested"* [4]. Clearly it is preferable to keep this operational definition at a hypothetical level, even though it is the best definition of the concept.

3.3.3 Rules of probability

The subjective definition of probability, together with the condition of coherence, requires that $0 \leq p \leq 1$. This is one of the rules which probability has to obey. It is possible, in fact, to demonstrate that coherence yields to the standard rules of probability, generally known as axioms. At this point it is worth clarifying the relationship between the axiomatic approach and the others.

- Combinatorial and frequentistic definitions give useful rules for evaluating probability, although they do not, as it is often claimed, define the concept.

- In the axiomatic approach one refrains from defining what the probability is and how to evaluate it: probability is just any real number which satisfies the axioms. It is easy to demonstrate that the probabilities evaluated using the combinatorial and the frequentistic prescriptions do in fact satisfy the axioms.

- The subjective approach to probability, together with the coherence requirement, defines what probability is and provides the rules which its evaluation must obey; these rules turn out to be the same as the axioms.

[3] This is not always true in real life as the importance of a given amount of money differs from person to person. The problem can be solved if the bet is considered *virtual*, i.e. the bet one would consider fair if one had an infinite budget.

[4] Both examples are from R. Scozzafava [32].

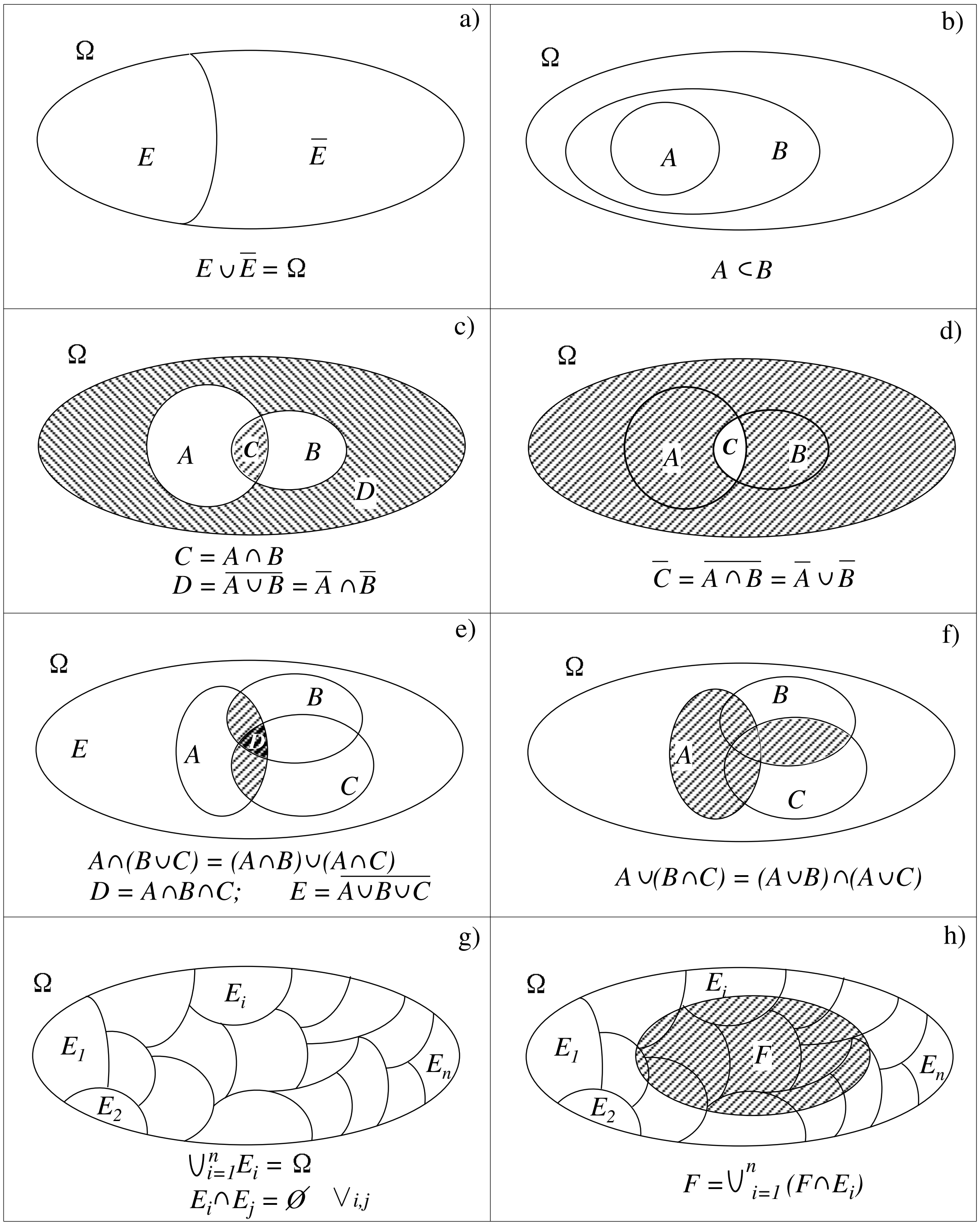

Figure 3.2: Venn diagrams and set properties.

Table 3.1: Events versus sets.

Events	Sets	
		Symbol
event	set	E
certain event	sample space	Ω
impossible event	empty set	$\emptyset$
implication	inclusion (subset)	$E_1 \subseteq E_2$
opposite event (complementary)	complementary set	$\overline{E} \quad (E \cup \overline{E} = \Omega)$
logical product ("AND")	intersection	$E_1 \cap E_2$
logical sum ("OR")	union	$E_1 \cup E_2$
incompatible events	disjoint sets	$E_1 \cap E_2 = \emptyset$
complete class	finite partition	$\begin{cases} E_i \cap E_j = \emptyset \ \ \forall\, i \neq j \\ \cup_i E_i = \Omega \end{cases}$

Since everybody is familiar with the axioms and with the analogy *events* $\Leftrightarrow$ *sets* (see Fig. 3.2 and Table 3.3.3) let us remind ourselves of the rules of probability in this form:

Axiom 1 $0 \leq P(E) \leq 1$;

Axiom 2 $P(\Omega) = 1$ (a certain event has probability 1);

Axiom 3 $P(E_1 \cup E_2) = P(E_1) + P(E_2)$, if $E_1 \cap E_2 = \emptyset$.

From the basic rules the following properties can be derived:

1: $P(E) = 1 - P(\overline{E})$;

2: $P(\emptyset) = 0$;

3: if $A \subseteq B$ then $P(A) \leq P(B)$;

4: $P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$

We also anticipate here another rule which will be discussed in Section 3.4.1:

5: $P(A \cap B) = P(A \,|\, B) \cdot P(B) = P(A) \cdot P(B \,|\, A)\,.$

3.3.4 Subjective probability and objective description of the physical world

The subjective definition of probability seems to contradict the aim of physicists to describe the laws of physics in the most objective way (whatever this means ...). This is one of the reasons why many regard the subjective definition of probability with suspicion (but probably the main reason is because we have been taught at university that probability is frequency). The main philosophical difference between this concept of probability and an objective definition that we would have liked (but which does not exist in reality) is that $P(E)$ is not an intrinsic characteristic of the event E, but depends on the state of information available to whoever evaluates $P(E)$. The ideal concept of objective probability is recovered when everybody has the

Table 3.2: Results of measurements of the gravitational constant G_N.

Institute	$G_N \left(10^{-11}\frac{\mathrm{m}^3}{\mathrm{kg}\cdot\mathrm{s}^2}\right)$	$\frac{\sigma(G_N)}{G_N}$ (ppm)	$\frac{G_N - G_N^C}{G_N^C}$ (10^{-3})
CODATA 1986 ("G_N^C")	6.6726 ± 0.0009	128	–
PTB (Germany) 1994	6.7154 ± 0.0006	83	$+6.41 \pm 0.16$
MSL (New Zealand) 1994	6.6656 ± 0.0006	95	-1.05 ± 0.16
Uni-Wuppertal (Germany) 1995	6.6685 ± 0.0007	105	-0.61 ± 0.17

same state of information. But even in this case it would be better to speak of intersubjective probability. The best way to convince ourselves about this aspect of probability is to try to ask practical questions and to evaluate the probability in specific cases, instead of seeking refuge in abstract questions. I find, in fact, that — to paraphrase a famous statement about Time — probability is objective as long as I am not asked to evaluate it. Here are some examples.

Example 1: What is the probability that a molecule of nitrogen at room temperature has a velocity between 400 and 500 m/s? The answer appears easy: Take the Maxwell distribution formula from a textbook, calculate an integral and get a number. Now let us change the question: I give you a vessel containing nitrogen and a detector capable of measuring the speed of a single molecule and you set up the apparatus (or you let a person you trust do it). Now, what is the probability that the first molecule that hits the detector has a velocity between 400 and 500 m/s? Anybody who has minimal experience (direct or indirect) of experiments would hesitate before answering. He would study the problem carefully and perform preliminary measurements and checks. Finally he would probably give not just a single number, but a range of possible numbers compatible with the formulation of the problem. Then he starts the experiment and eventually, after 10 measurements, he may form a different opinion about the outcome of the eleventh measurement.

Example 2: What is the probability that the gravitational constant G_N has a value between $6.6709 \cdot 10^{-11}$ and $6.6743 \cdot 10^{-11}$ $\mathrm{m^3kg^{-1}s^{-2}}$? Before 1994 you could have looked at the latest issue of the Particle Data Book [33] and answered that the probability was 95%. Since then — as you probably know — three new measurements of G_N have been performed [34] and we now have four numbers which do not agree with each other (see Table 3.3.4). The probability of the true value of G_N being in that range is currently dramatically decreased.

Example 3: What is the probability that the mass of the top quark, or that of any of the supersymmetric particles, is below 20 or 50 GeV/c^2? Currently it looks as if it must be zero. Ten years ago many experiments were intensively looking for these particles in those energy ranges. Because so many people where searching for them, with enormous human and capital investment, it meant that, at that time, the probability was considered rather high: high enough for fake signals to be reported as strong evidence for them.[5]

[5]We will talk later about the influence of *a priori* beliefs on the outcome of an experimental investigation.

The above examples show how the evaluation of probability is conditioned by some *a priori* (theoretical) prejudices and by some facts (experimental data). 'Absolute' probability makes no sense. Even the classical example of probability 1/2 for each of the results in tossing a coin is only acceptable if the coin is regular, it does not remain vertical (not impossible when playing on the beach), it does not fall into a manhole, etc.

The subjective point of view is expressed in a provocative way by de Finetti's [11]

"PROBABILITY DOES NOT EXIST".

3.4 Conditional probability and Bayes' theorem

3.4.1 Dependence of the probability on the state of information

If the state of information changes, the evaluation of the probability also has to be modified. For example most people would agree that the probability of a car being stolen depends on the model, age and parking site. To take an example from physics, the probability that in a HERA detector a charged particle of 1 GeV gives a certain number of ADC counts due to the energy loss in a gas detector can be evaluated in a very general way — using HEP jargon — by making a (huge) Monte Carlo simulation which takes into account all possible reactions (weighted with their cross-sections), all possible backgrounds, changing all physical and detector parameters within reasonable ranges, and also taking into account the trigger efficiency. The probability changes if one knows that the particle is a K^+: instead of very complicated Monte Carlo simulation one can just run a single particle generator. But then it changes further if one also knows the exact gas mixture, pressure, etc., up to the latest determination of the pedestal and the temperature of the ADC module.

3.4.2 Conditional probability

Although everybody knows the formula of conditional probability, it is useful to derive it here.[6] The notation is $P(E\,|\,H)$, to be read 'probability of E given H', where H stands for hypothesis. This means: the probability that E will occur under the hypothesis that H has occurred.[7]

The event $E\,|\,H$ can have three values:

TRUE: if E is TRUE <u>and</u> H is TRUE;

FALSE: if E is FALSE <u>and</u> H is TRUE;

UNDETERMINED: if H is FALSE; in this case we are merely uninterested in what happens to E. In terms of betting, the bet is invalidated and none loses or gains.

Then $P(E)$ can be written $P(E\,|\,\Omega)$, to state explicitly that it is the probability of E whatever happens to the rest of the world (Ω means all possible events). We realize immediately that this condition is really too vague and nobody would bet a penny on a such a statement. The reason for usually writing $P(E)$ is that many conditions are implicitly, and reasonably, assumed

[6]**Note added**: for a further discussion about the meaning of 'the formula of conditional probability' see Section 8.3.

[7]$P(E\,|\,H)$ should not be confused with $P(E \cap H)$, 'the probability that both events occur'. For example $P(E \cap H)$ can be very small, but nevertheless $P(E\,|\,H)$ very high. Think of the limit case

$$P(H) \equiv P(H \cap H) \le P(H\,|\,H) = 1\ :$$

'H given H' is a certain event no matter how small $P(H)$ is, even if $P(H) = 0$ (in the sense of Section 4.1.2).

in most circumstances. In the classical problems of coins and dice, for example, one assumes that they are regular. In the example of the energy loss, it was implicit (obvious) that the high voltage was on (at which voltage?) and that HERA was running (under which condition?). But one has to take care: many riddles are based on the fact that one tries to find a solution which is valid under more strict conditions than those explicitly stated in the question[35], and many people make bad business deals by signing contracts in which what was obvious was not explicitly stated.

In order to derive the formula of conditional probability let us assume for a moment that it is reasonable to talk about absolute probability $P(E) = P(E\,|\,\Omega)$, and let us rewrite

$$\begin{aligned} P(E) \equiv P(E\,|\,\Omega) &\underset{\mathbf{a}}{=} P(E\cap\Omega) \\ &\underset{\mathbf{b}}{=} P\left(E\cap(H\cup\overline{H})\right) \\ &\underset{\mathbf{c}}{=} P\left((E\cap H)\cup(E\cap\overline{H})\right) \\ &\underset{\mathbf{d}}{=} P(E\cap H)+P(E\cap\overline{H})\,, \end{aligned} \tag{3.1}$$

where the result has been achieved through the following steps:

(a) E implies Ω (i.e. $E\subseteq\Omega$) and hence $E\cap\Omega = E$;

(b) the complementary events H and $\overline{H}$ make a finite partition of Ω, i.e. $H\cup\overline{H}=\Omega$;

(c) distributive property;

(d) axiom 3.

The final result of (3.1) is very simple: $P(E)$ is equal to the probability that E occurs and H also occurs, plus the probability that E occurs but H does not occur. To obtain $P(E\,|\,H)$ we just get rid of the subset of E which does not contain H (i.e. $E\cap\overline{H}$) and renormalize the probability dividing by $P(H)$, assumed to be different from zero. This guarantees that if $E = H$ then $P(H\,|\,H) = 1$. We get, finally, the well-known formula

$$P(E\,|\,H) = \frac{P(E\cap H)}{P(H)} \qquad [P(H)\neq 0]\,. \tag{3.2}$$

In the most general (and realistic) case, where both E and H are conditioned by the occurrence of a third event $H_\circ$, the formula becomes

$$P(E\,|\,H,H_\circ) = \frac{P\left(E\cap(H\,|\,H_\circ)\right)}{P(H\,|\,H_\circ)} \qquad [P(H\,|\,H_\circ)\neq 0]\,. \tag{3.3}$$

Usually we shall make use of (3.2) (which means $H_\circ = \Omega$) assuming that Ω has been properly chosen. We should also remember that (3.2) can be resolved with respect to $P(E\cap H)$, obtaining the well-known

$$P(E\cap H) = P(E\,|\,H)P(H)\,, \tag{3.4}$$

and by symmetry

$$P(E\cap H) = P(H\,|\,E)P(E)\,. \tag{3.5}$$

We remind that two events are called independent if

$$P(E \cap H) = P(E)P(H)\,. \tag{3.6}$$

This is equivalent to saying that $P(E\,|\,H) = P(E)$ and $P(H\,|\,E) = P(H)$, i.e. the knowledge that one event has occurred does not change the probability of the other. If $P(E\,|\,H) \neq P(E)$ then the events E and H are correlated. In particular:

- if $P(E\,|\,H) > P(E)$ then E and H are positively correlated;
- if $P(E\,|\,H) < P(E)$ then E and H are negatively correlated.

3.4.3 Bayes' theorem

Let us think of all the possible, mutually exclusive, hypotheses H_i which could condition the event E. The problem here is the inverse of the previous one: what is the probability of H_i under the hypothesis that E has occurred? For example, what is the probability that a charged particle which went in a certain direction and has lost between 100 and 120 keV in the detector is a μ, π, K, or p? Our event E is 'energy loss between 100 and 120 keV', and H_i are the four 'particle hypotheses'. This example sketches the basic problem for any kind of measurement: having observed an effect, to assess the probability of each of the causes which could have produced it. This intellectual process is called inference, and it will be discussed in Section 5.2.

In order to calculate $P(H_i\,|\,E)$ let us rewrite the joint probability $P(H_i \cap E)$, making use of (3.4–3.5), in two different ways:

$$P(H_i\,|\,E)P(E) = P(E\,|\,H_i)P(H_i)\,, \tag{3.7}$$

obtaining

$$\boxed{P(H_i\,|\,E) = \frac{P(E\,|\,H_i)P(H_i)}{P(E)}\,,} \tag{3.8}$$

or

$$\boxed{\frac{P(H_i\,|\,E)}{P(H_i)} = \frac{P(E\,|\,H_i)}{P(E)}\,.} \tag{3.9}$$

Since the hypotheses H_i are mutually exclusive (i.e. $H_i \cap H_j = \emptyset$, $\forall\, i,j$) and exhaustive (i.e. $\bigcup_i H_i = \Omega$), E can be written as $\bigcup_i E \cap H_i$, the union of the intersections of E with each of the hypotheses H_i. It follows that

$$\begin{aligned} P(E)\,[\equiv P(E\cap\Omega)] &= P\left(\bigcup_i (E \cap H_i)\right) \\ &= \sum_i P(E \cap H_i) \\ &= \sum_i P(E\,|\,H_i)P(H_i)\,, \end{aligned} \tag{3.10}$$

where we have made use of (3.4) again in the last step. It is then possible to rewrite (3.8) as

$$\boxed{P(H_i\,|\,E) = \frac{P(E\,|\,H_i)P(H_i)}{\sum_j P(E\,|\,H_j)P(H_j)}\,.} \tag{3.11}$$

This is the standard form by which Bayes' theorem is known. (3.8) and (3.9) are also different ways of writing it. As the denominator of (3.11) is nothing but a normalization factor, such that $\sum_i P(H_i\,|\,E) = 1$, the formula (3.11) can be written as

$$\boxed{P(H_i\,|\,E) \propto P(E\,|\,H_i)P(H_i)\,.} \tag{3.12}$$

Factorizing $P(H_i)$ in (3.11), and explicitly writing that all the events were already conditioned by $H_\circ$, we can rewrite the formula as

$$\boxed{P(H_i\,|\,E, H_\circ) = \alpha P(H_i\,|\,H_\circ)\,,} \tag{3.13}$$

with

$$\alpha = \frac{P(E\,|\,H_i, H_\circ)}{\sum_i P(E\,|\,H_i, H_\circ)P(H_i\,|\,H_\circ)}\,. \tag{3.14}$$

These five ways of rewriting the same formula simply reflect the importance that we shall give to this simple theorem. They stress different aspects of the same concept.

- (3.11) is the standard way of writing it, although some prefer (3.8).
- (3.9) indicates that $P(H_i)$ is altered by the condition E with the same ratio with which $P(E)$ is altered by the condition H_i.
- (3.12) is the simplest and the most intuitive way to formulate the theorem: 'The probability of H_i given E is proportional to the initial probability of H_i times the probability of E given H_i.'
- (3.13–3.14) show explicitly how the probability of a certain hypothesis is updated when the state of information changes:

  $\boxed{P(H_i\,|\,H_\circ)}$ [also indicated as $P_\circ(H_i)$] is the initial, or *a priori*, probability (or simply 'prior') of H_i, i.e. the probability of this hypothesis with the state of information available before the knowledge that E has occurred;

  $\boxed{P(H_i\,|\,E, H_\circ)}$ [or simply $P(H_i\,|\,E)$] is the final, or *a posteriori*, probability of H_i after[8] the new information.

  $\boxed{P(E\,|\,H_i, H_\circ)}$ [or simply $P(E\,|\,H_i)$] is called likelihood.

To better understand the terms 'initial', 'final' and 'likelihood', let us formulate the problem in a way closer to the physicist's mentality, referring to causes and effects: the causes could be all the physical sources which may produce a certain observable (the effect). Using our example of the dE/dx measurement again, the causes are all the possible charged particles which can pass through the detector; the effect is the amount of observed ionization; the likelihoods are the probabilities that each of the particles give that amount of ionization. Note that in this example we have fixed all the other sources of influence: physics process, HERA running conditions, gas mixture, high voltage, track direction, etc. This is our $H_\circ$. The problem immediately gets rather complicated (all real cases, apart from tossing coins and dice, are complicated!). The real inference would be of the kind

$$P(H_i\,|\,E, H_\circ) \propto P(E\,|\,H_i, H_\circ)P(H_i\,|\,H_\circ)\,. \tag{3.15}$$

[8]Note that 'before' and 'after' do not really necessarily imply time ordering, but only the consideration or not of the new piece of information.

For each state $H_\circ$ (the set of all the possible values of the influence parameters) one gets a different result for the final probability. So, instead of getting a single number for the final probability we have a distribution of values. This spread will result in a large uncertainty of $P(H_i \,|\, E)$. This is what every physicist knows: if the calibration constants of the detector and the physics process are not under control, the systematic errors are large and the result is of poor quality.[9]

3.4.4 Conventional use of Bayes' theorem

Bayes' theorem follows directly from the rules of probability, and it can be used in any kind of approach. Let us take an example:

Problem 1: A particle detector has a μ identification efficiency of 95%, and a probability of identifying a π as a μ of 2%. If a particle is identified as a μ, then a trigger is fired. Knowing that the particle beam is a mixture of 90% π and 10% μ, what is the probability that a trigger is really fired by a μ? What is the signal-to-noise (S/N) ratio?

Solution: The two hypotheses (causes) which could condition the event (effect) T (= trigger fired) are μ and π. They are incompatible (clearly) and exhaustive (90% + 10% = 100%). Then:

$$\begin{aligned} P(\mu \,|\, T) &= \frac{P(T \,|\, \mu) P_\circ(\mu)}{P(T \,|\, \mu) P_\circ(\mu) + P(T \,|\, \pi) P_\circ(\pi)} \\ &= \\ &= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.02 \times 0.9} = 0.84\,, \end{aligned} \tag{3.16}$$

and $P(\pi \,|\, T) = 0.16$.

The S/N ratio is $P(\mu \,|\, T)/P(\pi \,|\, T) = 5.3$. It is interesting to rewrite the general expression of the S/N ratio if the effect E is observed as

$$S/N = \frac{P(S \,|\, E)}{P(N \,|\, E)} = \frac{P(E \,|\, S)}{P(E \,|\, N)} \cdot \frac{P_\circ(S)}{P_\circ(N)}\,. \tag{3.17}$$

This formula explicitly shows that when there are noisy conditions,

$$P_\circ(S) \ll P_\circ(N),$$

the experiment must be very selective,

$$P(E \,|\, S) \gg P(E \,|\, N),$$

in order to have a decent S/N ratio.
(How does S/N change if the particle has to be identified by two independent detectors in order to give the trigger? Try it yourself, the answer is $S/N = 251$.)

[9] Formally, the influence of the uncertainty about $H_\circ$ on $P(H_i)$ can be seen in the following way. Indicating by $H_{\circ_j}$ all possible configurations of $H_\circ$, we get from the rules of probability:

$$P(H_i \,|\, E) = \sum_j P(H_i \,|\, E, H_{\circ_j}) P(H_{\circ_j})\,.$$

Problem 2: Three boxes contain two rings each, but in one of them they are both gold, in the second both silver, and in the third one of each type. You have the choice of randomly extracting a ring from one of the boxes, the content of which is unknown to you. You look at the extracted ring, and you then have the possibility of extracting a second ring, again from any of the three boxes. Let us assume the first ring you extract is a gold one. Is it then preferable to extract the second one from the same or from a different box?

Solution: Choosing the same box you have a 2/3 probability of getting a second gold ring. (Try to apply the theorem, or help yourself with intuition; the solution is given in Section 8.10.)

The difference between the two problems, from the conventional statistics point of view, seems to be the following. In the frequentistic approach only the first problem is meaningful, since the probabilities entering in the problem are evaluated from experimental frequencies. In a pure combinatorial approach only the second problem has a solution. Nevertheless, the question is a little more subtle. What is, for example, the meaning of the 84% probability obtained as the solution of the first problem? It is no longer a ratio between the number of occurrencies of the event and the number of experimental trials. Therefore, strictly speaking, it is not a probability according to the frequentistic 'definition'. It is easy to understand that the only consistent way to interpret such a result is to consider it as the degree of belief that the particle was a muon. The same is true for the solution of the second problem.

In conclusion, although the rules of probability are the same in the different approaches (and therefore also in Bayes' theorem), only in the subjective approach are the results of the calculations consistent at every step with the definition of probability.

3.4.5 Bayesian statistics: learning by experience

The advantage of the Bayesian approach (leaving aside the 'little philosophical detail' of trying to define what probability is) is that one may talk about the probability of any kind of event, as already emphasized. Moreover, the procedure of updating the probability with increasing information is very similar to that followed by the mental processes of rational people.[10] Let us consider a few examples of 'Bayesian use' of Bayes' theorem.

Example 1: Imagine some persons listening to a common friend having a phone conversation with an unknown person X_i, and who are trying to guess who X_i is. Depending on the knowledge they have about the friend, on the language spoken, on the tone of voice, on the subject of conversation, etc., they will attribute some probability to several possible persons. As the conversation goes on they begin to consider some possible candidates for X_i, discarding others, then hesitating perhaps only between a couple of possibilities, until the state of information I is such that they are practically sure of the identity of X_i. This experience has happened to most of us, and it is not difficult to recognize the Bayesian scheme:

$$P(X_i \,|\, I, I_\circ) \propto P(I \,|\, X_i, I_\circ) P(X_i \,|\, I_\circ)\,. \tag{3.18}$$

We have put the initial state of information $I_\circ$ explicitly in (3.18) to remind us that likelihoods and initial probabilities depend on it. If we know nothing about the person, the final probabilities will be very vague, i.e. for many persons X_i the probability will be different from zero, without necessarily favouring any particular person.

[10] **Note added**: Ref. [36] shows an interesting investigations on the relation between perception and Bayesian inference.

Example 2: A person X meets an old friend F in a pub. F proposes that the drinks should be payed for by whichever of the two extracts the card of lower value from a pack (according to some rule which is of no interest to us). X accepts and F wins. This situation happens again in the following days and it is always X who has to pay. What is the probability that F has become a cheat, as the number of consecutive wins n increases?

The two hypotheses are: cheat (C) and honest (H). $P_\circ(C)$ is low because F is an old friend, but certainly not zero: let us assume 5%. To make the problem simpler let us make the approximation that a cheat always wins (not very clever...): $P(W_n \mid C) = 1$. The probability of winning if he is honest is, instead, given by the rules of probability assuming that the chance of winning at each trial is 1/2 (why not?, we shall come back to this point later): $P(W_n \mid H) = 2^{-n}$. The result

$$P(C \mid W_n) = \frac{P(W_n \mid C) \cdot P_\circ(C)}{P(W_n \mid C) \cdot P_\circ(C) + P(W_n \mid H) \cdot P_\circ(H)} = \tag{3.19}$$

$$= \frac{1 \cdot P_\circ(C)}{1 \cdot P_\circ(C) + 2^{-n} \cdot P_\circ(H)} \tag{3.20}$$

is shown in the following table.

n	$P(C \mid W_n)$ (%)	$P(H \mid W_n)$ (%)
0	5.0	95.0
1	9.5	90.5
2	17.4	82.6
3	29.4	70.6
4	45.7	54.3
5	62.7	37.3
6	77.1	22.9
...	...	...

Naturally, as F continues to win the suspicion of X increases. It is important to make two remarks.

- The answer is always probabilistic. X can never reach absolute certainty that F is a cheat, unless he catches F cheating, or F confesses to having cheated. This is coherent with the fact that we are dealing with random events and with the fact that any sequence of outcomes has the same probability (although there is only one possibility over 2^n in which F is always luckier). Making use of $P(C \mid W_n)$, X can make a decision about the next action to take:
  - continue the game, with probability $P(C \mid W_n)$ of losing with certainty the next time too;
  - refuse to play further, with probability $P(H \mid W_n)$ of offending the innocent friend.
- If $P_\circ(C) = 0$ the final probability will always remain zero: if X fully trusts F, then he has just to record the occurrence of an *a priori* rare event when n becomes large.

To better follow the process of updating the probability when new experimental data become available, according to the Bayesian scheme

> *"the final probability of the present inference is the initial probability of the next one."*

Let us call $P(C\,|\,W_{n-1})$ the probability assigned after the previous win. The iterative application of the Bayes formula yields

$$P(C\,|\,W_n) = \frac{P(W\,|\,C)\cdot P(C\,|\,W_{n-1})}{P(W\,|\,C)\cdot P(C\,|\,W_{n-1}) + P(W\,|\,H)\cdot P(H\,|\,W_{n-1})} \tag{3.21}$$

$$= \frac{1\cdot P(C\,|\,W_{n-1})}{1\cdot P(C\,|\,W_{n-1}) + \frac{1}{2}\cdot P(H\,|\,W_{n-1})}\,, \tag{3.22}$$

where $P(W\,|\,C) = 1$ and $P(W\,|\,H) = 1/2$ are the probabilities of each win. The interesting result is that exactly the same values of $P(C\,|\,W_n)$ of (3.20) are obtained (try to believe it!).

It is also instructive to see the dependence of the final probability on the initial probabilities, for a given number of wins n.

$P_\circ(C)$	$P(C\,|\,W_n)$ (%)			
	$n=5$	$n=10$	$n=15$	$n=20$
1 %	24	91	99.7	99.99
5 %	63	98	99.94	99.998
50 %	97	99.90	99.997	99.9999

As the number of experimental observations increases the conclusions no longer depend, practically, on the initial assumptions. This is a crucial point in the Bayesian scheme and it will be discussed in more detail later.

3.5 Hypothesis test (discrete case)

Although in conventional statistics books this argument is usually dealt with in one of the later chapters, in the Bayesian approach it is so natural that it is in fact the first application, as we have seen in the above examples. We summarize here the procedure:

- probabilities are attributed to the different hypotheses using initial probabilities and experimental data (via the likelihood);
- the person who makes the inference — or the 'user' — will make a decision for which he is fully responsible.

If one needs to compare two hypotheses, as in the example of the S/N calculation, the ratio of the final probabilities can be taken as a quantitative result of the test. Let us rewrite the S/N formula (3.17) in the most general case:

$$\frac{P(H_1\,|\,E,H_\circ)}{P(H_2\,|\,E,H_\circ)} = \frac{P(E\,|\,H_1,H_\circ)}{P(E\,|\,H_2,H_\circ)}\cdot\frac{P(H_1\,|\,H_\circ)}{P(H_2\,|\,H_\circ)}\,, \tag{3.23}$$

where again we have reminded ourselves of the existence of $H_\circ$. The ratio depends on the product of two terms: the ratio of the priors and the ratio of the likelihoods. When there is absolutely no reason for choosing between the two hypotheses the prior ratio is 1 and the decision depends only on the other term, called 'the Bayes factor'. If one firmly believes in either hypothesis, the Bayes factor is of minor importance, unless it is zero or infinite (i.e. one and only one of the likelihoods is vanishing). Perhaps this is disappointing for those who expected objective certainty from a probability theory, but this is in the nature of things.

3.6 Choice of the initial probabilities (discrete case)

3.6.1 General criteria

The dependence of Bayesian inferences on initial probability is considered by opponents as the fatal flaw in the theory. But this criticism is less severe than one might think at first sight.[11] In fact:

- It is impossible to construct a theory of uncertainty which is not affected by this 'illness'. Those methods which are advertised as being 'objective' tend in reality to hide the hypotheses on which they are grounded. A typical example is the maximum likelihood method, of which we will talk later.
- As the amount of information increases the dependence on initial prejudices diminishes.
- When the amount of information is very limited, or completely lacking, there is nothing to be ashamed of if the inference is dominated by *a priori* assumptions.

It is well known to all experienced physicists that conclusions drawn from an experimental result (and sometimes even the result itself!) often depend on prejudices about the phenomenon under study. Some examples:

- When doing quick checks on a device, a single measurement is usually performed if the value is 'what it should be', but if it is not then many measurements tend to be made.
- Results are sometimes influenced by previous results or by theoretical predictions. See for example Fig. 3.3 taken from the Particle Data Book [33]. The interesting book *"How experiments end"* [37] discusses, among others, the issue of when experimentalists are happy with the result and stop correcting for the systematics.
- Slight deviations from the background might be interpreted as a signal (e.g. as for the first claim of discovery of the top quark in spring 1994), while larger signals are viewed with suspicion if they are unwanted by the physics establishment.[12]
- Experiments are planned and financed according to the prejudices of the moment.[13]

These comments are not intended to justify unscrupulous behaviour or sloppy analysis. They are intended, instead, to remind us — if need be — that scientific research is ruled by subjectivity much more than outsiders imagine. The transition from subjectivity to objectivity begins when there is a large consensus among the most influential people about how to interpret the results.[14]

[11] **Note added**: for an extensive discussion about priors see Ref. [22].

[12] A case, concerning the search for electron compositeness in $e^+ e^-$ collisions, is discussed in Ref. [38].

[13] For a recent delightful report, see Ref. [39].

[14] *"A theory needs to be confirmed by experiments. But it is also true that an experimental result needs to be confirmed by a theory."* This sentence expresses clearly — though paradoxically — the idea that it is difficult to accept a result which is not rationally justified. An example of results not confirmed by the theory are the R measurements in deep-inelastic scattering shown in Fig. 3.4. Given the conflict in this situation, physicists tend to believe more in QCD and use the 'low-x' extrapolations (of what?) to correct the data for the unknown values of R.

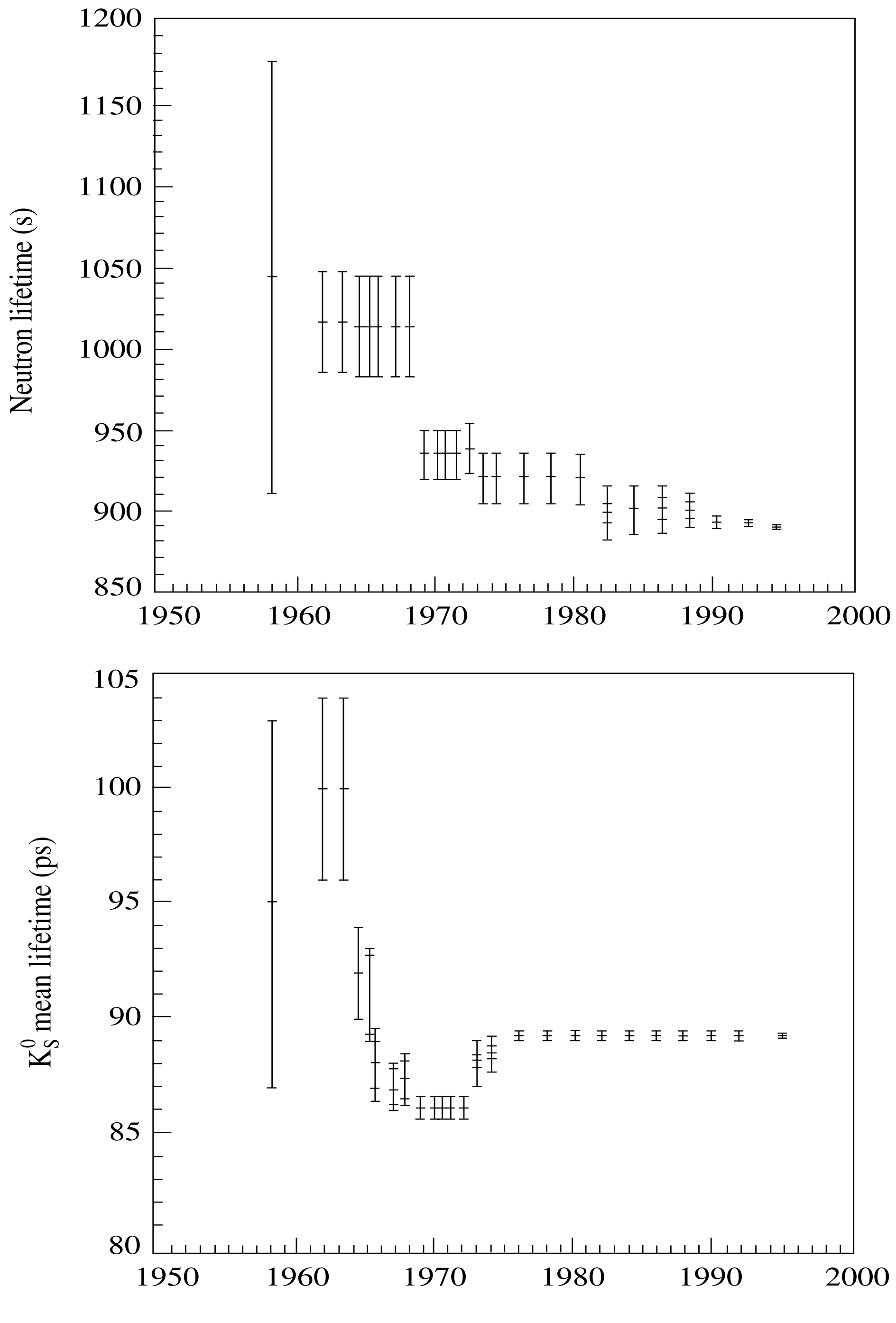

Figure 3.3: Results on two physical quantities as a function of the publication date.

In this context, the subjective approach to statistical inference at least teaches us that every assumption must be stated clearly and all available information which could influence conclusions must be weighed with the maximum attempt at objectivity.[15]

What are the rules for choosing the 'right' initial probabilities? As one can imagine, this is an open and debated question among scientists and philosophers. My personal point of view is that one should avoid pedantic discussion of the matter, because the idea of universally true priors reminds me terribly of the famous 'angels' sex' debates.

If I had to give recommendations, they would be the following.

- The *a priori* probability should be chosen in the same spirit as the rational person who places a bet, seeking to minimize the risk of losing.
- General principles — like those that we will discuss in a while — may help, but since it may be difficult to apply elegant theoretical ideas in all practical situations, in many circumstances the guess of the expert can be relied on for guidance.
- To avoid using as prior the results of other experiments dealing with the same open problem, otherwise correlations between the results would prevent all comparison between the experiments and thus the detection of any systematic errors. I find that this point is generally overlooked by statisticians.

3.6.2 Insufficient reason and maximum entropy

The first and most famous criterion for choosing initial probabilities is the simple 'Principle of Insufficient Reason' (or 'Indifference Principle'): If there is no reason to prefer one hypothesis over alternatives, simply attribute the same probability to all of them. This was stated as a principle by Laplace[16] in contrast to Leibnitz' famous 'Principle of Sufficient Reason', which, in simple words, states that 'nothing happens without a reason'. The indifference principle applied to coin and die tossing, to card games or to other simple and symmetric problems, yields to the well-known rule of probability evaluation that we have called combinatorial. Since it is impossible not to agree with this point of view, in the cases for which one judges that it does apply, the combinatorial definition of probability is recovered in the Bayesian approach if the word 'definition' is simply replaced by 'evaluation rule'. We have in fact already used this reasoning in previous examples.

A modern and more sophisticated version of the Indifference Principle is the Maximum Entropy Principle. The information entropy function of n mutually exclusive events, to each of which a probability p_i is assigned, is defined as [40]

$$H(p_1, p_2, \ldots p_n) = -K \sum_{i=1}^{n} p_i \ln p_i \,, \tag{3.24}$$

with K a positive constant. The principle states that *"in making inferences on the basis of partial information we must use that probability distribution which has the maximum entropy subject to whatever is known"* [41]. Note that, in this case, 'entropy' is synonymous with 'uncertainty' [41]. One can show that, in the case of absolute ignorance about the events E_i, the maximization of

[15] It may look paradoxical, but, due to the normative role of the coherent bet, subjective assessments are more objective than using, without direct responsibility, someone else's formulae. For example, even the knowledge that somebody else has a different evaluation of the probability is new information which must be taken into account.

[16] It may help in understanding Laplace's approach if we consider that he called the theory of probability *"good sense turned into calculation."*

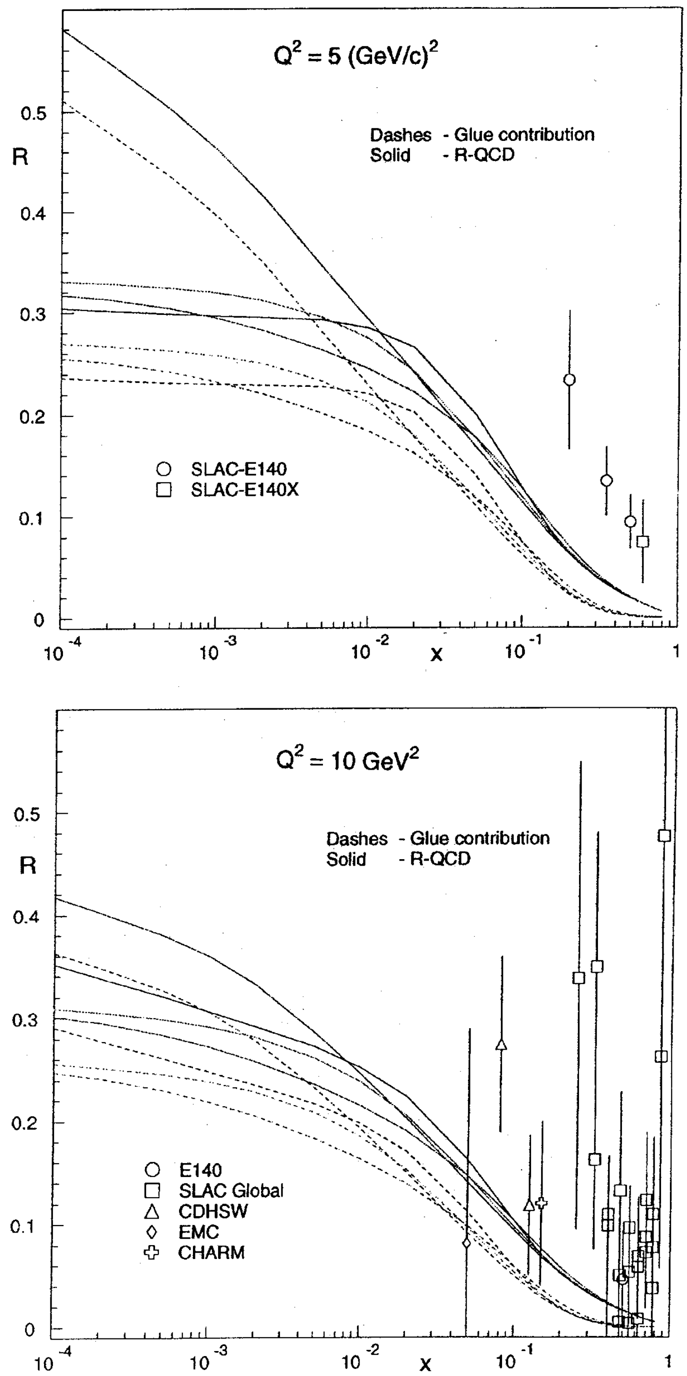

Figure 3.4: $R = \sigma_L/\sigma_T$ as a function of the deep-inelastic scattering variable x as measured by experiments and as predicted by QCD.

the information uncertainty, with the constraint that $\sum_{i=1}^{n} p_i = 1$, yields the classical $p_i = 1/n$ (any other result would have been worrying ...).

Although this principle is sometimes used in combination with the Bayes formula for inferences (also applied to measurement uncertainty, see Ref. [23]), it will not be used for applications in these notes. Those who are interested in entropy, both in information and in probability theory can find a clear introduction in Ref. [42].

Chapter 4

Distributions (a concise reminder)

4.1 Random variables

In the discussion which follows I will assume that the reader is familiar with random variables, distributions, probability density functions, and expected values, as well as with the most frequently used distributions. This section is only intended as a summary of concepts and as a presentation of the notation used in the subsequent sections.

4.1.1 Discrete variables

Uncertain numbers are numbers in respect of which we are in a condition of uncertainty. They can be the number associated with the outcome of a die, to the number which will be read on a scale when a measurement is performed, or to the numerical value of a physics quantity. In the following, we will call uncertain numbers also random variables, to come close to what physicists are used to, but one should not think, then, that random variables are only associated with the outcomes of repeated experiments. Stated simply, to define a random variable X means to find a rule which allows a real number to be related univocally (but not necessarily biunivocal) to an event (E). One could write this expression $X(E)$. Discrete variables assume a countable range, finite or not. We shall indicate the variable with X and its numerical realization with x; and differently from other notations, the symbol x (in place of n or k) is also used for discrete variables.

Here is a list of definitions, properties and notations.

Probability function.

To each possible value of X we associate a degree of belief:

$$f(x) = P(X = x)\,. \tag{4.1}$$

$f(x)$, being a probability, must satisfy the following properties:

$$0 \leq f(x_i) \leq 1\,, \tag{4.2}$$

$$P(X = x_i \cup X = x_j) = f(x_i) + f(x_j)\,, \tag{4.3}$$

$$\sum_i f(x_i) = 1\,. \tag{4.4}$$

Cumulative distribution function.

$$F(x_k) \equiv P(X \leq x_k) = \sum_{x_i \leq x_k} f(x_i)\,. \tag{4.5}$$

Properties:

$$
\begin{aligned}
&F(-\infty) = 0, && (4.6)\\
&F(+\infty) = 1, && (4.7)\\
&F(x_i) - F(x_{i-1}) = f(x_i), && (4.8)\\
&\lim_{\epsilon \to o} F(x+\epsilon) = F(x) \qquad (\textit{right side continuity})\,. && (4.9)
\end{aligned}
$$

Expected value (mean).

$$\mu \equiv \mathrm{E}[X] = \sum_i x_i f(x_i)\,. \tag{4.10}$$

In general, given a function $g(X)$ of X,

$$\mathrm{E}[g(X)] = \sum_i g(x_i) f(x_i)\,. \tag{4.11}$$

$\mathrm{E}[\cdot]$ is a linear operator:

$$\mathrm{E}[aX+b] = a\mathrm{E}[X] + b\,. \tag{4.12}$$

Variance and standard deviation.

Variance:

$$\sigma^2 \equiv \mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mu^2\,. \tag{4.13}$$

Standard deviation:

$$\sigma = \sqrt{\sigma^2}\,. \tag{4.14}$$

Transformation properties:

$$
\begin{aligned}
\mathrm{Var}(aX+b) &= a^2\,\mathrm{Var}(X)\,, && (4.15)\\
\sigma(aX+b) &= |a|\,\sigma(X)\,. && (4.16)
\end{aligned}
$$

Binomial distribution.

$X \sim \mathcal{B}_{n,p}$ (hereafter '$\sim$' stands for 'follows'); $\mathcal{B}_{n,p}$ indicates a binomial with parameters n (integer) and p (real):

$$f(x\,|\,\mathcal{B}_{n,p}) = \frac{n!}{(n-x)!\,x!}\,p^x\,(1-p)^{n-x}\,, \qquad \left\{\begin{array}{l} n = 1, 2, \ldots, \infty \\ 0 \le p \le 1 \\ x = 0, 1, \ldots, n \end{array}\right. . \tag{4.17}$$

Expected value, standard deviation and variation coefficient:

$$
\begin{aligned}
\mu &= n\,p, && (4.18)\\
\sigma &= \sqrt{n\,p\,(1-p)}, && (4.19)\\
v \equiv \frac{\sigma}{\mu} &= \frac{\sqrt{n\,p\,(1-p)}}{n\,p} \propto \frac{1}{\sqrt{n}}\,. && (4.20)
\end{aligned}
$$

$1-p$ is often indicated by q.

Poisson distribution.

$X \sim \mathcal{P}_\lambda$:

$$f(x\,|\,\mathcal{P}_\lambda) = \frac{\lambda^x}{x!}\,e^{-\lambda} \qquad \left\{ \begin{array}{l} 0 < \lambda < \infty \\ x = 0, 1, \dots, \infty \end{array} \right. . \tag{4.21}$$

(x is an integer, λ is real.)
Expected value, standard deviation and variation coefficient:

$$\mu = \lambda, \tag{4.22}$$
$$\sigma = \sqrt{\lambda}, \tag{4.23}$$
$$v = \frac{1}{\sqrt{\lambda}}. \tag{4.24}$$

Binomial → Poisson.

$$\mathcal{B}_{n,p} \xrightarrow[\substack{n \to \text{`}\infty\text{'} \\ p \to \text{`}0\text{'} \\ (\lambda = n\,p)}]{} P_\lambda$$

4.1.2 Continuous variables: probability density function

Moving from discrete to continuous variables there are the usual problems with infinite possibilities, similar to those found in Zeno's 'Achilles and the tortoise' paradox. In both cases the answer is given by infinitesimal calculus. But some comments are needed:

- The probability of each of the realizations of X is zero ($P(X = x) = 0$); but this does not mean that each value is impossible, otherwise it would be impossible to get any result.
- Although all values x have zero probability, one usually assigns different degrees of belief to them, quantified by the probability density function $f(x)$. Writing $f(x_1) > f(x_2)$, for example, indicates that our degree of belief in x_1 is greater than that in x_2.
- The probability that a random variable lies inside a finite interval, for example $P(a \leq X \leq b)$, is instead finite. If the distance between a and b becomes infinitesimal, then the probability becomes infinitesimal too. If all the values of X have the same degree of belief (and not only equal numerical probability $P(x) = 0$) the infinitesimal probability is simply proportional to the infinitesimal interval $\mathrm{d}P = k\,\mathrm{d}x$. In the general case the ratio between two infinitesimal probabilities around two different points will be equal to the ratio of the degrees of belief in the points (this argument implies the continuity of $f(x)$ on either side of the values). It follows that $\mathrm{d}P = f(x)\,\mathrm{d}x$ and then

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x\,. \tag{4.25}$$

- $f(x)$ has a dimension inverse to that of the random variable.

After this short introduction, here is a list of definitions, properties and notations:

Cumulative distribution function.

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'\,, \tag{4.26}$$

or

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}. \tag{4.27}$$

Properties of $f(x)$ and $F(x)$.

- $f(x) \ge 0$,
- $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$,
- $0 \le F(x) \le 1$,
- $P(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x = \int_{-\infty}^{b} f(x)\,\mathrm{d}x - \int_{-\infty}^{a} f(x)\,\mathrm{d}x = F(b) - F(a)$,
- if $x_2 > x_1$ then $F(x_2) \ge F(x_1)$,
- $\lim_{x\to-\infty} F(x) = 0$,
$\lim_{x\to+\infty} F(x) = 1$.

Expected value.

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x\, f(x)\,\mathrm{d}x, \tag{4.28}$$

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f(x)\,\mathrm{d}x. \tag{4.29}$$

Uniform distribution. [1]

$X \sim \mathcal{K}(a,b)$:

$$f(x\,|\,\mathcal{K}(a,b)) = \frac{1}{b-a} \qquad (a \le x \le b), \tag{4.30}$$

$$F(x\,|\,\mathcal{K}(a,b)) = \frac{x-a}{b-a}\,. \tag{4.31}$$

Expected value and standard deviation:

$$\mu = \frac{a+b}{2}, \tag{4.32}$$

$$\sigma = \frac{b-a}{\sqrt{12}}\,. \tag{4.33}$$

[1] The symbols of the following distributions have the parameters within parentheses to indicate that the variables are continuous.

Normal (Gaussian) distribution.
$X \sim \mathcal{N}(\mu, \sigma)$:

$$f(x\,|\,\mathcal{N}(\mu,\sigma)) = \frac{1}{\sqrt{2\,\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \left\{ \begin{array}{l} -\infty < \mu < +\infty \\ 0 < \sigma < \infty \\ -\infty < x < +\infty \end{array} \right. , \tag{4.34}$$

where μ and σ (both real) are the expected value and standard deviation,[2] respectively.

Standard normal distribution.

The particular normal distribution of mean 0 and standard deviation 1, usually indicated by Z:

$$Z \sim \mathcal{N}(0,1)\,. \tag{4.35}$$

Exponential distribution.

$T \sim \mathcal{E}(\tau)$:

$$f(t\,|\,\mathcal{E}(\tau)) = \frac{1}{\tau}\, e^{-t/\tau} \qquad \left\{ \begin{array}{l} 0 \le \tau < \infty \\ 0 \le t < \infty \end{array} \right. \tag{4.36}$$

$$F(t\,|\,\mathcal{E}(\tau)) = 1 - e^{-t/\tau}. \tag{4.37}$$

We use the symbol t instead of x because this distribution will be applied to the time domain.

Survival probability:

$$P(T > t) = 1 - F(t\,|\,\mathcal{E}(\tau)) = e^{-t/\tau}. \tag{4.38}$$

Expected value and standard deviation:

$$\mu = \tau \tag{4.39}$$

$$\sigma = \tau. \tag{4.40}$$

The real parameter τ has the physical meaning of lifetime.

Poisson $\leftrightarrow$ Exponential.

If X (= number of counts during the time Δt) is Poisson distributed then T (= interval of time to wait — starting from any instant — before the first count is recorded) is exponentially distributed:

$$X \sim f(x\,|\,\mathcal{P}_\lambda) \qquad \Longleftrightarrow \qquad T \sim f(x\,|\,\mathcal{E}(\tau)) \tag{4.41}$$

$$(\tau = \tfrac{\Delta T}{\lambda}) \quad . \tag{4.42}$$

[2] Mathematicians and statisticians prefer to take σ^2, instead of σ, as second parameter of the normal distribution. Here the standard deviation is preferred, since it is homogeneous to μ and it has a more immediate physical interpretation. So, one has to pay attention to be sure about the meaning of expressions like $\mathcal{N}(0.5, 0.8)$.

4.1.3 Distribution of several random variables

We only consider the case of two continuous variables (X and Y). The extension to more variables is straightforward. The infinitesimal element of probability is $\mathrm{d}F(x,y) = f(x,y)\,\mathrm{d}x\,\mathrm{d}y$, and the probability density function

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}\,. \tag{4.43}$$

The probability of finding the variable inside a certain area A is

$$\iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{4.44}$$

Marginal distributions.

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y, \tag{4.45}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\,. \tag{4.46}$$

The subscripts X and Y indicate that $f_X(x)$ and $f_Y(y)$ are functions only of X and Y, respectively (to avoid fooling around with different symbols to indicate the generic function), but in most cases we will drop the subscripts if the context helps in resolving ambiguities.

Conditional distributions.

$$f_X(x\,|\,y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int f(x,y)\,\mathrm{d}x}, \tag{4.47}$$

$$f_Y(y\,|\,x) = \frac{f(x,y)}{f_X(x)}, \tag{4.48}$$

$$f(x,y) = f_X(x\,|\,y)\,f_Y(y) \tag{4.49}$$

$$= f_Y(y\,|\,x)\,f_X(x)\,. \tag{4.50}$$

Independent random variables.

$$f(x,y) = f_X(x) f_Y(y) \tag{4.51}$$

(it implies $f_X(x\,|\,y) = f_X(x)$ and $f_Y(y\,|\,x) = f_Y(y)$.)

Bayes' theorem for continuous random variables.

$$\boxed{f(h\,|\,e) = \frac{f(e\,|\,h)\,f_h(h)}{\int f(e\,|\,h)\,f_h(h)\,\mathrm{d}h}\,.} \tag{4.52}$$

(**Note added**: see proof in Section 2.7.)

Expected value.

$$\mu_X = \mathrm{E}[X] = \iint_{-\infty}^{+\infty} x\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{4.53}$$

$$= \int_{-\infty}^{+\infty} x\,f_X(x)\,\mathrm{d}x\,, \tag{4.54}$$

and analogously for Y. In general

$$\mathrm{E}[g(X,Y)] = \iint_{-\infty}^{+\infty} g(x,y)\, f(x,y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{4.55}$$

Variance.

$$\sigma_X^2 = \mathrm{E}[X^2] - \mathrm{E}^2[X]\,, \tag{4.56}$$

and analogously for Y.

Covariance.

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{E}\left[\,(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])\,\right] && (4.57)\\
&= \mathrm{E}[XY] - \mathrm{E}[X] \cdot \mathrm{E}[Y]\,. && (4.58)
\end{aligned}$$

If X and Y are independent, then $\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$ and hence $\mathrm{Cov}(X,Y) = 0$ (the opposite is true only if X, $Y \sim \mathcal{N}(\cdot)$).

Correlation coefficient.

$$\begin{aligned}
\rho(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} && (4.59)\\
&= \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}\,. && (4.60)
\end{aligned}$$

$$(-1 \le \rho \le 1)$$

Linear combinations of random variables.

If $Y = \sum_i c_i X_i$, with c_i real, then

$$\begin{aligned}
\mu_Y = \mathrm{E}[Y] &= \sum_i c_i\,\mathrm{E}[X_i] = \sum_i c_i\,\mu_i, && (4.61)\\
\sigma_Y^2 = \mathrm{Var}(Y) &= \sum_i c_i^2\,\mathrm{Var}(X_i) + 2\sum_{i<j} c_i\,c_j\,\mathrm{Cov}(X_i,X_j) && (4.62)\\
&= \sum_i c_i^2\,\mathrm{Var}(X_i) + \sum_{i\ne j} c_i\,c_j\,\mathrm{Cov}(X_i,X_j) && (4.63)\\
&= \sum_i c_i^2\,\sigma_i^2 + \sum_{i\ne j} \rho_{ij}\,c_i\,c_j\,\sigma_i\,\sigma_j && (4.64)\\
&= \sum_{ij} \rho_{ij}\,c_i\,c_j\,\sigma_i\,\sigma_j && (4.65)\\
&= \sum_{ij} c_i\,c_j\,\sigma_{ij}\,. && (4.66)
\end{aligned}$$

σ_Y^2 has been written in different ways, with increasing levels of compactness, that can be found in the literature. In particular, (4.65) and (4.66) use the notations $\sigma_{ij} \equiv \mathrm{Cov}(X_i,X_j) = \rho_{ij}\,\sigma_i\,\sigma_j$ and $\sigma_{ii} = \sigma_i^2$, and the fact that, by definition, $\rho_{ii} = 1$.

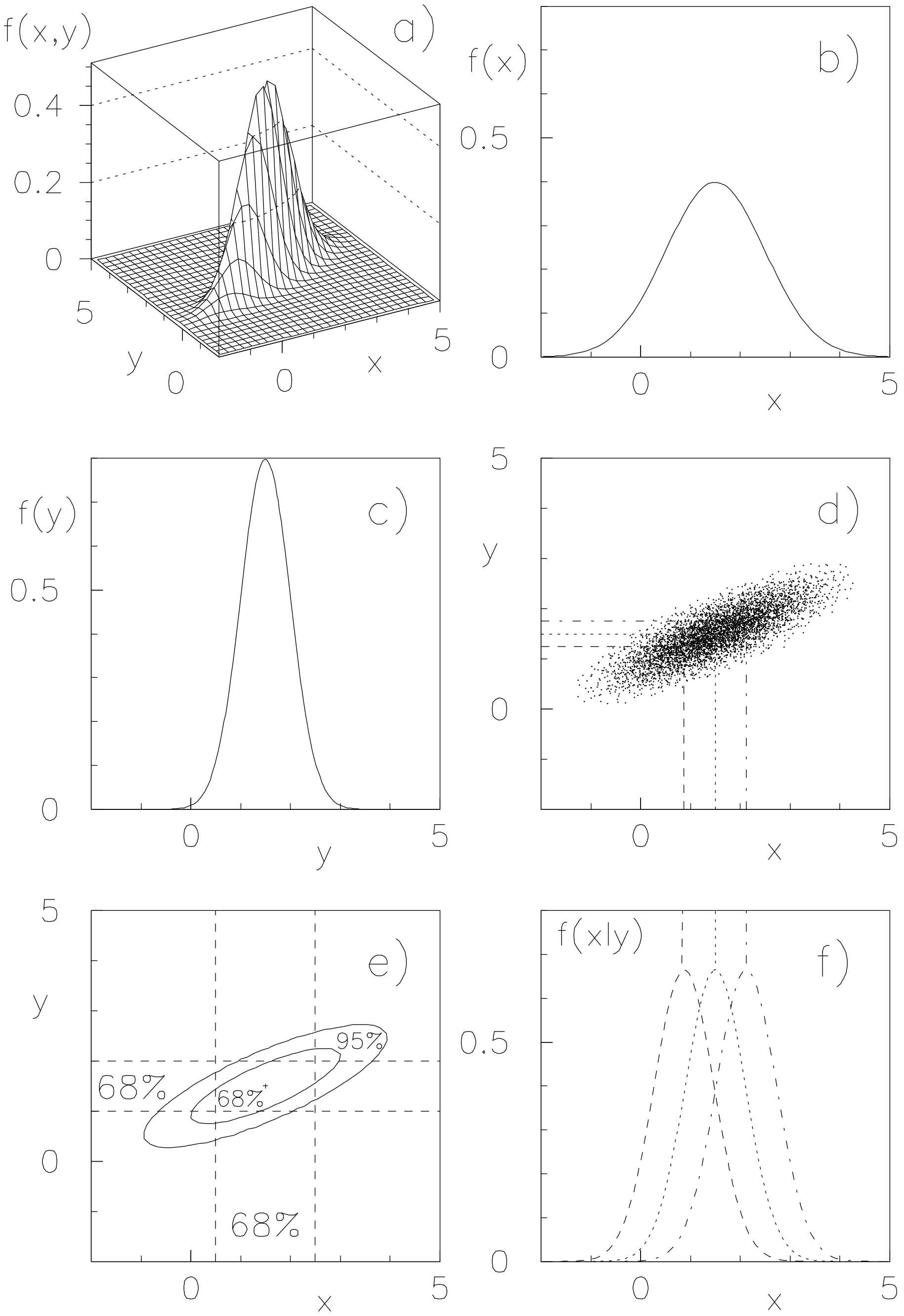

Figure 4.1: Example of bivariate normal distribution.

Bivariate normal distribution.

Joint probability density function of X and Y with correlation coefficient ρ (see Fig. 4.1):

$$\begin{aligned} f(x,y) \;=\; & \frac{1}{2\,\pi\,\sigma_x\,\sigma_y\,\sqrt{1-\rho^2}}\,\cdot \\ & \exp\left\{-\frac{1}{2\,(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}-2\,\rho\,\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\,\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}. \end{aligned} \tag{4.67}$$

Marginal distributions:

$$X \sim \mathcal{N}(\mu_x,\sigma_x), \tag{4.68}$$

$$Y \sim \mathcal{N}(\mu_y,\sigma_y)\,. \tag{4.69}$$

Conditional distribution:

$$f(y\,|\,x_\circ) = \frac{1}{\sqrt{2\,\pi}\,\sigma_y\,\sqrt{1-\rho^2}}\exp\left[-\frac{\left(y-\left[\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}\,(x_\circ-\mu_x)\right]\right)^2}{2\,\sigma_y^2\,(1-\rho^2)}\right], \tag{4.70}$$

i.e.

$$Y_{|x_\circ} \sim \mathcal{N}\left(\mu_y+\rho\,\frac{\sigma_y}{\sigma_x}\,(x_\circ-\mu_x)\,,\,\sigma_y\sqrt{1-\rho^2}\right). \tag{4.71}$$

The condition $X = x_\circ$ squeezes the standard deviation and shifts the mean of Y.

4.2 Central limit theorem

4.2.1 Terms and role

The well-known central limit theorem plays a crucial role in statistics and justifies the enormous importance that the normal distribution has in many practical applications (this is why it appears on 10 DM notes).

We have reminded ourselves in (4.61)–(4.62) of the expression of the mean and variance of a linear combination of random variables,

$$Y = \sum_{i=1}^{n} c_i X_i\,,$$

in the most general case, which includes correlated variables ($\rho_{ij} \neq 0$). In the case of independent variables the variance is given by the simpler, and better known, expression

$$\sigma_Y^2 = \sum_{i=1}^{n} c_i^2\,\sigma_i^2 \qquad (\rho_{ij}=0,\; i\neq j)\,. \tag{4.72}$$

This is a very general statement, valid for any number and kind of variables (with the obvious clause that all σ_i must be finite), but it does not give any information about the probability distribution of Y. Even if all X_i follow the same distributions $f(x)$, $f(y)$ is different from $f(x)$, with some exceptions, one of these being the normal.

The central limit theorem states that the distribution of a linear combination Y will be approximately normal if the variables X_i are independent and σ_Y^2 is much larger than any single component $c_i^2\sigma_i^2$ from a non-normally distributed X_i. The last condition is just to guarantee that there is no single random variable which dominates the fluctuations. The accuracy of the approximation improves as the number of variables n increases (the theorem says "when $n \to \infty$"):

$$n \to \infty \Longrightarrow Y \sim \mathcal{N}\left(\sum_{i=1}^{n} c_i\,\mathrm{E}(X_i), \left(\sum_{i=1}^{n} c_i^2\,\sigma_i^2\right)^{\frac{1}{2}}\right). \tag{4.73}$$

The proof of the theorem can be found in standard textbooks. For practical purposes, and if one is not very interested in the detailed behaviour of the tails, n equal to 2 or 3 may already give a satisfactory approximation, especially if the X_i exhibits a Gaussian-like shape. See, for example, Fig. 4.2, where samples of 10 000 events have been simulated, starting from a uniform distribution and from a crazy square-wave distribution. The latter, depicting a kind of worst practical case, shows that, already for $n = 20$ the distribution of the sum is practically normal. In the case of the uniform distribution $n = 3$ already gives an acceptable approximation as far as probability intervals of one or two standard deviations from the mean value are concerned. The figure also shows that, starting from a triangular distribution (obtained in the example from the sum of two uniform distributed variables), $n = 2$ is already sufficient (The sum of two triangular distributed variables is equivalent to the sum of four uniform distributed variables.)

4.2.2 Distribution of a sample average

As first application of the theorem, let us remind ourselves that a sample average $\overline{X}_n$ of n independent variables,

$$\overline{X}_n = \sum_{i=1}^{n} \frac{1}{n}\,X_i, \tag{4.74}$$

is normally distributed, since it is a linear combination of n variables X_i, with $c_i = 1/n$. Then,

$$\overline{X}_n \sim \mathcal{N}(\mu_{\overline{X}_n}, \sigma_{\overline{X}_n}), \tag{4.75}$$

$$\mu_{\overline{X}_n} = \sum_{i=1}^{n} \frac{1}{n}\,\mu = \mu, \tag{4.76}$$

$$\sigma^2_{\overline{X}_n} = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}, \tag{4.77}$$

$$\sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}. \tag{4.78}$$

This result, we repeat, is independent of the distribution of X and is already approximately valid for small values of n.

4.2.3 Normal approximation of the binomial and of the Poisson distribution

Another important application of the theorem is that the binomial and the Poisson distribution can be approximated, for large numbers, by a normal distribution. This is a general result, valid for all distributions which have the reproductive property under the sum. Distributions of this kind are the binomial, the Poisson and the χ^2. Let us go into more detail:

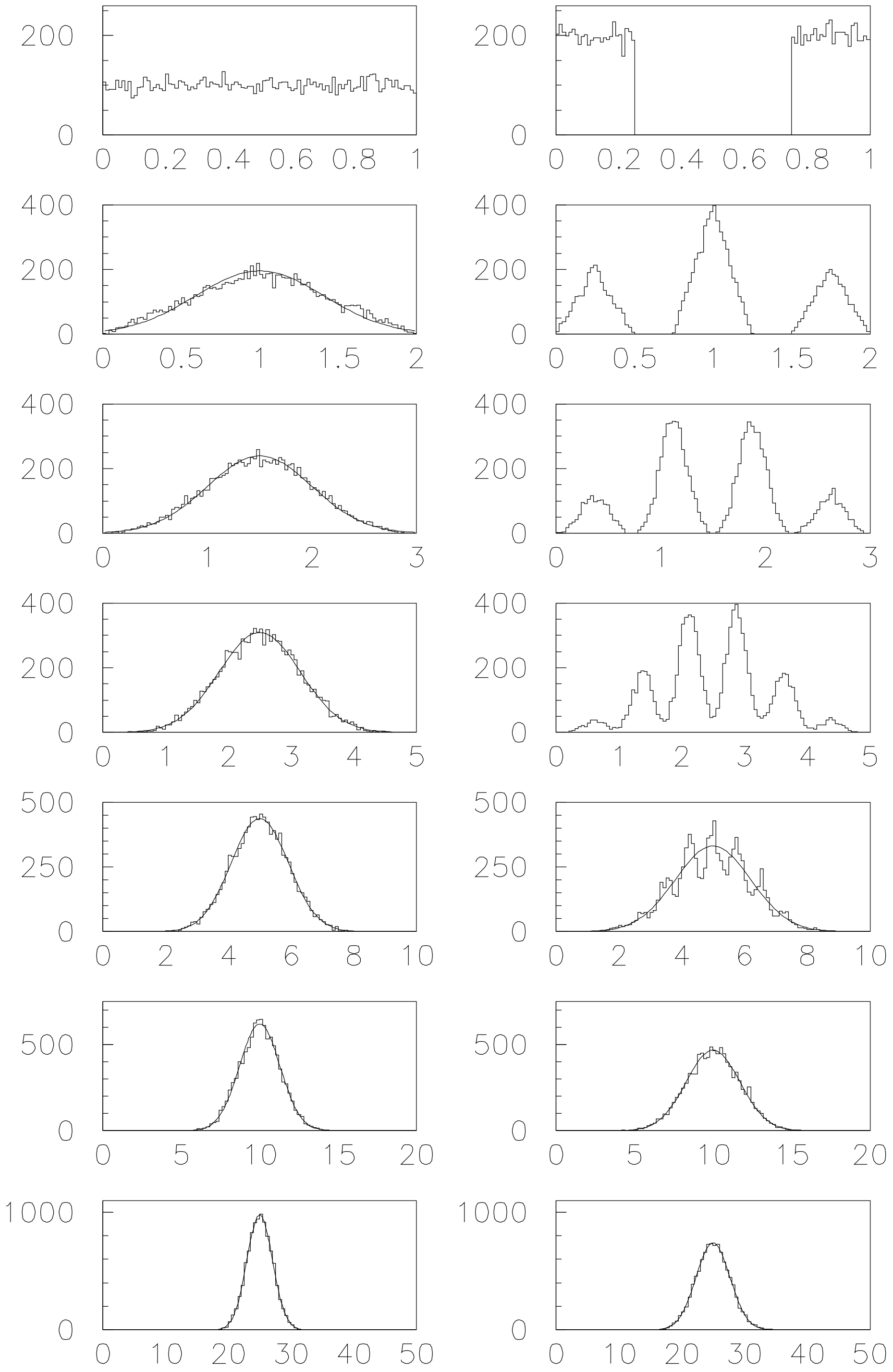

Figure 4.2: Central limit theorem at work: The sum of n variables, for two different distributions, is shown. The values of n (top to bottom) are 1, 2, 3, 5, 10, 20, 50.

$\boxed{\mathcal{B}_{n,p} \rightarrow \mathcal{N}\left(n\,p, \sqrt{n\,p\,(1-p)}\right)}$ The reproductive property of the binomial states that if X_1, X_2, $\ldots$, X_m are m independent variables, each following a binomial distribution of parameter n_i and p, then their sum $Y = \sum_i X_i$ also follows a binomial distribution with parameters $n = \sum_i n_i$ and p. It is easy to be convinced of this property without any mathematics. Just think of what happens if one tosses bunches of three, of five and of ten coins, and then one considers the global result: a binomial with a large n can then always be seen as a sum of many binomials with smaller n_i. The application of the central limit theorem is straightforward, apart from deciding when the convergence is acceptable. The parameters on which one has to base a judgment are in this case $\mu = n\,p$ and the complementary quantity $\mu^c = n\,(1-p) = n - \mu$. If they are both $\gtrsim 10$ then the approximation starts to be reasonable.

$\boxed{\mathcal{P}_\lambda \rightarrow \mathcal{N}\left(\lambda, \sqrt{\lambda}\right)}$ The same argument holds for the Poisson distribution. In this case the approximation starts to be reasonable when $\mu = \lambda \gtrsim 10$.

4.2.4 Normal distribution of measurement errors

The central limit theorem is also important to justify why in many cases the distribution followed by the measured values around their average is approximately normal. Often, in fact, the random experimental error e, which causes the fluctuations of the measured values around the unknown true value of the physical quantity, can be seen as an incoherent sum of smaller contributions e_i:

$$e = \sum_i e_i\,, \tag{4.79}$$

each contribution having a distribution which satisfies the conditions of the central limit theorem.

4.2.5 Caution

Following this commercial in favour of the miraculous properties of the central limit theorem, some words of caution are in order.

- Although I have tried to convince the reader that the convergence is rather fast in the cases of practical interest, the theorem only states that the asymptotic Gaussian distribution is reached for $n \rightarrow \infty$. As an example of very slow convergence, let us imagine 10^9 independent variables described by a Poisson distribution of $\lambda_i = 10^{-9}$: their sum is still far from a Gaussian.
- Sometimes the conditions of the theorem are not satisfied.
  - A single component dominates the fluctuation of the sum: a typical case is the well-known Landau distribution; systematic errors may also have the same effect on the global error.
  - The condition of independence is lost if systematic errors affect a set of measurements, or if there is coherent noise.
- The tails of the distributions do exist and they are not always Gaussian! Moreover, realizations of a random variable several standard deviations away from the mean are possible. And they show up without notice!

Chapter 5

Bayesian inference applied to measurements

"... these problems are classified as probability of the causes, and are the most interesting of all from their scientific applications".

"An effect may be produced by the cause a or by the cause b. The effect has just been observed. We ask the probability that it is due to the cause a. This is an à posteriori probability of cause. But I could not calculate it, if a convention more or less justified did not tell me in advance what is the à priori probability for the cause a to come into play. I mean the probability of this event to some one who had not observed the effect."
(Henri Poincaré)

5.1 Measurement errors and measurement uncertainty

One might assume that the concepts of error and uncertainty are well enough known to be not worth discussing. Nevertheless a few comments are needed (although for more details the DIN [1] and ISO [3, 4] recommendations should be consulted).

- The first concerns the terminology. In fact the words error and uncertainty are currently used almost as synonyms:

  – 'error' to mean both error and uncertainty (but nobody says 'Heisenberg Error Principle');
  – 'uncertainty' only for the uncertainty.

  'Usually' we understand what each is talking about, but a more precise use of these nouns would really help. This is strongly called for by the DIN [1] and ISO [3, 4] recommendations. They state in fact that

  – error is *"the result of a measurement minus a true value of the measurand"* – it follows that the error is usually unkown;
  – uncertainty is a *"parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand"*;

- Within the HEP community there is an established practice for reporting the final uncertainty of a measurement in the form of standard deviation. This is also recommended by the mentioned standards. However, this should be done at each step of the analysis, instead of estimating maximum error bounds and using them as standard deviation in the error propagation.

- The process of measurement is a complex one and it is difficult to disentangle the different contributions which cause the total error. In particular, the active role of the experimentalist is sometimes overlooked. For this reason it is often incorrect to quote the (nominal) uncertainty due to the instrument as if it were the uncertainty of the measurement.

5.2 Statistical inference

5.2.1 Bayesian inference

In the Bayesian framework the inference is performed by calculating the final distribution of the random variable associated with the true values of the physical quantities from all available information. Let us call $\underline{x} = \{x_1, x_2, \dots, x_n\}$ the n-tuple (vector) of observables, $\underline{\mu} = \{\mu_1, \mu_2, \dots, \mu_n\}$ the n-tuple of the true values of the physical quantities of interest, and $\underline{h} = \{h_1, h_2, \dots, h_n\}$ the n-tuple of all the possible realizations of the influence variables H_i. The term "influence variable" is used here with an extended meaning, to indicate not only external factors which could influence the result (temperature, atmospheric pressure, and so on) but also any possible calibration constant and any source of systematic errors. In fact the distinction between $\underline{\mu}$ and $\underline{h}$ is artificial, since they are all conditional hypotheses. We separate them simply because at the end we will marginalize the final joint distribution functions with respect to $\underline{\mu}$, integrating the joint distribution with respect to the other hypotheses considered as influence variables.

The likelihood of the sample $\underline{x}$ being produced from $\underline{h}$ and $\underline{\mu}$ and the initial probability are

$$f(\underline{x}\,|\,\underline{\mu}, \underline{h}, H_\circ)$$

and

$$f_\circ(\underline{\mu}, \underline{h}) = f(\underline{\mu}, \underline{h}\,|\,H_\circ)\,, \tag{5.1}$$

respectively. $H_\circ$ is intended to remind us, yet again, that likelihoods and priors — and hence conclusions — depend on all explicit and implicit assumptions within the problem, and in particular on the parametric functions used to model priors and likelihoods. To simplify the formulae, $H_\circ$ will no longer be written explicitly.

Using the Bayes formula for multidimensional continuous distributions [an extension of (4.52)] we obtain the most general formula of inference,

$$f(\underline{\mu}, \underline{h}\,|\,\underline{x}) = \frac{f(\underline{x}\,|\,\underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})}{\int f(\underline{x}\,|\,\underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})\, \mathrm{d}\underline{\mu}\, \mathrm{d}\underline{h}}\,, \tag{5.2}$$

yielding the joint distribution of all conditional variables $\underline{\mu}$ and $\underline{h}$ which are responsible for the observed sample $\underline{x}$. To obtain the final distribution of $\underline{\mu}$ one has to integrate (5.2) over all possible values of $\underline{h}$, obtaining

$$\boxed{f(\underline{\mu}\,|\,\underline{x}) = \frac{\int f(\underline{x}\,|\,\underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})\, \mathrm{d}\underline{h}}{\int f(\underline{x}\,|\,\underline{\mu}, \underline{h})\, f_\circ(\underline{\mu}, \underline{h})\, \mathrm{d}\underline{\mu}\, \mathrm{d}\underline{h}}\,.} \tag{5.3}$$

Apart from the technical problem of evaluating the integrals, if need be numerically or using Monte Carlo methods,[1] (5.3) represents the most general form of hypothetical inductive inference. The word 'hypothetical' reminds us of $H_\circ$.

When all the sources of influence are under control, i.e. they can be assumed to take a precise value, the initial distribution can be factorized by a $f_\circ(\underline{\mu})$ and a Dirac $\delta(\underline{h} - \underline{h}_\circ)$, obtaining the much simpler formula

$$\begin{aligned} f(\underline{\mu}\,|\,\underline{x}) &= \frac{\int f(\underline{x}\,|\,\underline{\mu},\underline{h})\, f_\circ(\underline{\mu})\, \delta(\underline{h}-\underline{h}_\circ)\,\mathrm{d}\underline{h}}{\int f(\underline{x}\,|\,\underline{\mu},\underline{h})\, f_\circ(\underline{\mu})\, \delta(\underline{h}-\underline{h}_\circ)\,\mathrm{d}\underline{\mu}\,\mathrm{d}\underline{h}} \\ &= \frac{f(\underline{x}\,|\,\underline{\mu},\underline{h}_\circ)\, f_\circ(\underline{\mu})}{\int f(\underline{x}\,|\,\underline{\mu},\underline{h}_\circ)\, f_\circ(\underline{\mu})\,\mathrm{d}\underline{\mu}}\,. \end{aligned} \tag{5.4}$$

Even if formulae (5.3)–(5.4) look complicated because of the multidimensional integration and of the continuous nature of $\underline{\mu}$, conceptually they are identical to the example of the $\mathrm{d}E/\mathrm{d}x$ measurement discussed in Section 3.4.3.

The final probability density function provides the most complete and detailed information about the unknown quantities, but sometimes (almost always …) one is not interested in full knowledge of $f(\underline{\mu})$, but just in a few numbers which summarize at best the position and the width of the distribution (for example when publishing the result in a journal in the most compact way). The most natural quantities for this purpose are the expected value and the variance, or the standard deviation. Then the Bayesian best estimate of a physical quantity is:

$$\widehat{\mu}_i = \mathrm{E}[\mu_i] = \int \mu_i\, f(\underline{\mu}\,|\,\underline{x})\,\mathrm{d}\underline{\mu}, \tag{5.5}$$

$$\sigma^2_{\mu_i} \equiv \mathrm{Var}(\mu_i) = \mathrm{E}[\mu_i^2] - \mathrm{E}^2[\mu_i]. \tag{5.6}$$

When many true values are inferred from the same data the numbers which synthesize the result are not only the expected values and variances, but also the covariances, which give at least the correlation coefficients between the variables:

$$\rho_{ij} \equiv \rho(\mu_i,\mu_j) = \frac{\mathrm{Cov}(\mu_i,\mu_j)}{\sigma_{\mu_i}\,\sigma_{\mu_j}}\,. \tag{5.7}$$

In the following sections we will deal in most cases with only one value to infer:

$$f(\mu\,|\,\underline{x}) = \ldots\,. \tag{5.8}$$

5.2.2 Bayesian inference and maximum likelihood

We have already said that the dependence of the final probabilities on the initial ones gets weaker as the amount of experimental information increases. Without going into mathematical complications (the proof of this statement can be found for example in Ref. [29]) this simply means that, asymptotically, whatever $f_\circ(\mu)$ one puts in (5.4), $f(\mu\,|\,\underline{x})$ is unaffected. This happens when the width of $f_\circ(\mu)$ is much larger than that of the likelihood, when the latter is considered as a mathematical function of μ. Therefore $f_\circ(\mu)$ acts as a constant in the region of μ where the likelihood is significantly different from 0. This is equivalent to dropping $f_\circ(\mu)$ from (5.4). This results in

$$f(\mu\,|\,\underline{x}) \approx \frac{f(\underline{x}\,|\,\mu,\underline{h}_\circ)}{\int f(\underline{x}\,|\,\mu,\underline{h}_\circ)\,\mathrm{d}\mu}\,. \tag{5.9}$$

[1] This is conceptually what experimentalists do when they change all the parameters of the Monte Carlo simulation in order to estimate the systematic error.

Since the denominator of the Bayes formula has the technical role of properly normalizing the probability density function, the result can be written in the simple form

$$f(\mu\,|\,\underline{x}) \propto f(\underline{x}\,|\,\mu, \underline{h}_\circ)\text{``} \equiv \mathcal{L}(\mu; \underline{x}, \underline{h}_\circ)\text{''}\,. \tag{5.10}$$

Asymptotically the final probability is just the (normalized) likelihood! The notation $\mathcal{L}$ is that used in the maximum likelihood literature (note that, not only does f become $\mathcal{L}$, but also "$\,|\,$" has been replaced by ";": $\mathcal{L}$ has no probabilistic interpretation, when referring to μ, in conventional statistics.)

If the mean value of $f(\mu\,|\,\underline{x})$ coincides with the value for which $f(\mu\,|\,\underline{x})$ has a maximum, we obtain the maximum likelihood method. This does not mean that the Bayesian methods are 'blessed' because of this achievement, and hence they can be used only in those cases where they provide the same results. It is the other way round: The maximum likelihood method gets justified when all the limiting conditions of the approach ($\rightarrow$ insensitivity of the result from the initial probability $\rightarrow$ large number of events) are satisfied.

Even if in this asymptotic limit the two approaches yield the same numerical results, there are differences in their interpretation:

- The likelihood, after proper normalization, has a probabilistic meaning for Bayesians but not for frequentists; so Bayesians can say that the probability that μ is in a certain interval is, for example, 68%, while this statement is blasphemous for a frequentist (the true value is a constant from his point of view).

- Frequentists prefer to choose $\widehat{\mu}_L$, the value which maximizes the likelihood, as estimator. For Bayesians, on the other hand, the expected value $\widehat{\mu}_B = \mathrm{E}[\mu]$ (also called the prevision) is more appropriate. This is justified by the fact that the assumption of the $\mathrm{E}[\mu]$ as best estimate of μ minimizes the risk of a bet (always keep the bet in mind!). For example, if the final distribution is exponential with parameter τ (let us think for a moment of particle decays) the maximum likelihood method would recommend betting on the value $t = 0$, whereas the Bayesian approach suggests the value $t = \tau$. If the terms of the bet are 'whoever gets closest wins', what is the best strategy? And then, what is the best strategy if the terms are 'whoever gets the exact value wins'? But now think of the probability of getting the exact value and of the probability of getting closest.

5.2.3 The dog, the hunter and the biased Bayesian estimators

One of the most important tests to judge the quality of an estimator is whether or not it is correct (not biased). Maximum likelihood estimators are usually correct, while Bayesian estimators — analysed within the maximum likelihood framework — often are not. This could be considered a weak point; however the Bayes estimators are simply naturally consistent with the state of information before new data become available. In the maximum likelihood method, on the other hand, it is not clear what the assumptions are.

Let us take an example which shows the logic of frequentistic inference and why the use of reasonable prior distributions yields results which that frame classifies as distorted. Imagine meeting a hunting dog in the country. Let us assume we know that there is a 50% probability of finding the dog within a radius of 100 m centred on the position of the hunter (this is our likelihood). Where is the hunter? He is with 50% probability within a radius of 100 m around the position of the dog, with equal probability in all directions: Obvious! This is exactly the logic scheme used in the frequentistic approach to build confidence regions from the estimator (the dog in this example). This however assumes that the hunter can be anywhere in the country.

But now let us change the state of information: the dog is by a river; the dog has collected a duck and runs in a certain direction; the dog is sleeping; the dog is in a field surrounded by a fence through which he can pass without problems, but the hunter cannot. Given any new condition the conclusion changes. Some of the new conditions change our likelihood, but some others only influence the initial distribution. For example, the case of the dog in an enclosure inaccessible to the hunter is exactly the problem encountered when measuring a quantity close to the edge of its physical region, which is quite common in frontier research.

5.3 Choice of the initial probability density function

5.3.1 Difference with respect to the discrete case

The title of this section is similar to that of Section 3.6, but the problem and the conclusions will be different. There we said that the Indifference Principle (or, in its refined modern version, the Maximum Entropy Principle) was a good choice. Here there are problems with infinities and with the fact that it is possible to map an infinite number of points contained in a finite region onto an infinite number of points contained in a larger or smaller finite region. This changes the probability density function. If, moreover, the transformation from one set of variables to the other is not linear (see, e.g., Fig. 5.1), what is uniform in one variable (X) is not uniform in another variable (e.g. $Y = X^2$). This problem does not exist in the case of discrete variables, since if $X = x_i$ has a probability $f(x_i)$ then $Y = x_i^2$ has the same probability. A different way of stating the problem is that the Jacobian of the transformation squeezes or stretches the metrics, changing the probability density function.

We will not enter into the open discussion about the optimal choice of the distribution. Essentially we shall use the uniform distribution, being careful to employ the variable which seems most appropriate for the problem, but <u>you</u> may disagree — surely with good reason — if <u>you</u> have a different kind of experiment in mind.

The same problem is also present, but well hidden, in the maximum likelihood method. For example, it is possible to demonstrate that, in the case of normally distributed likelihoods, a uniform distribution of the mean μ is implicitly assumed (see Section 5.4). There is nothing wrong with this, but one should be aware of it.

5.3.2 Bertrand paradox and angels' sex

A good example to help understand the problems outlined in the previous section is the so-called Bertrand paradox.

Problem: Given a circle of radius R and a chord drawn randomly on it, what is the probability that the length L of the chord is smaller than R?

Solution 1: Choose randomly two points on the circumference and draw a chord between them: $\Rightarrow P(L < R) = 1/3 = 0.33$.

Solution 2: Choose a straight line passing through the centre of the circle; then draw a second line, orthogonal to the first, and which intersects it inside the circle at a random distance from the centre: $\Rightarrow P(L < R) = 1 - \sqrt{3}/2 = 0.13$.

Solution 3: Choose randomly a point inside the circle and draw a straight line orthogonal to the radius that passes through the chosen point $\Rightarrow P(L < R) = 1/4 = 0.25$.

Your solution: … … …?

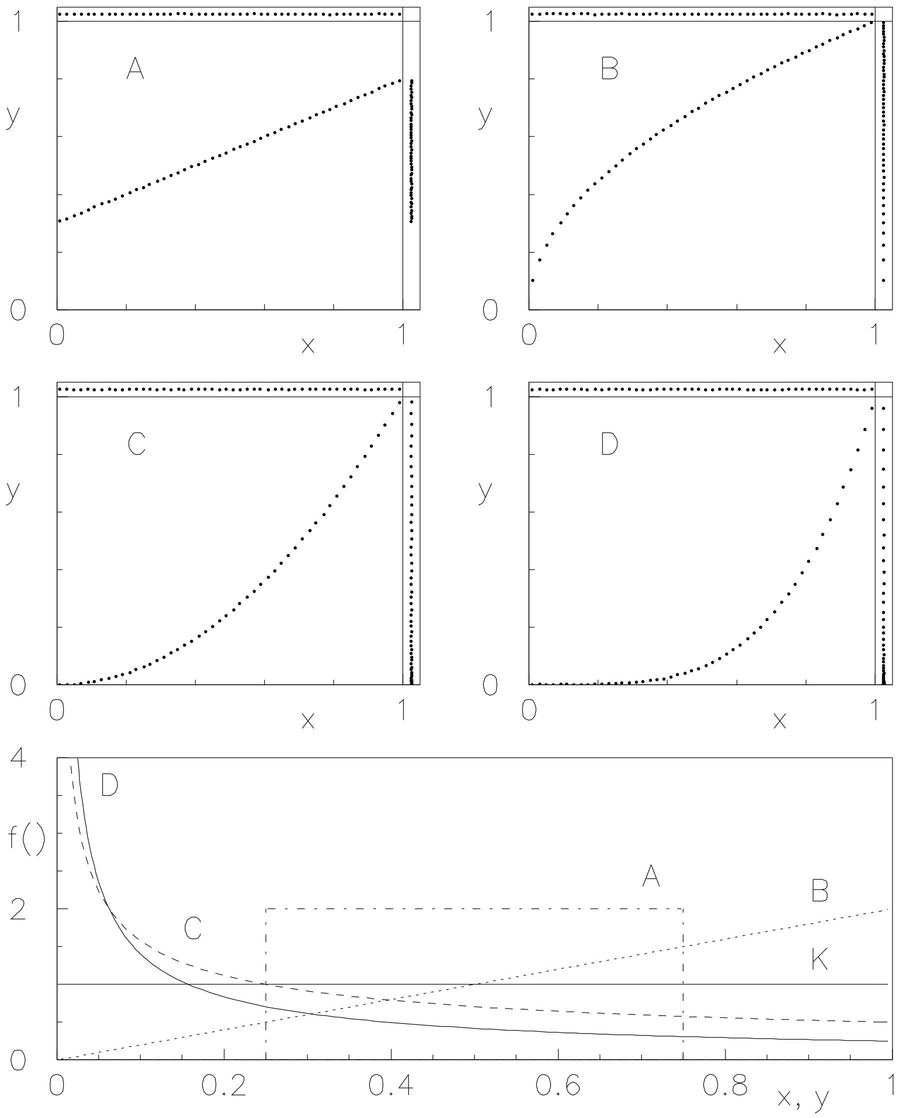

Figure 5.1: Examples of variable changes starting from a uniform distribution (K): A) $Y = 0.5\,X + 0.25$; B) $Y = \sqrt{X}$; C) $Y = X^2$; D) $Y = X^4$.

Question: What is the origin of the paradox?

Answer: The problem does not specify how to randomly choose the chord. The three solutions take a uniform distribution: along the circumference; along the radius; and inside the circle. What is uniform in one variable is not uniform in the others!

Question: Which is the right solution?

In principle you may imagine an infinite number of different solutions. From a physicist's viewpoint any attempt to answer this question is a waste of time. The reason why the paradox has been compared to the Byzantine discussions about the sex of angels is that there are indeed people arguing about it. For example, there is a school of thought which insists that Solution 2 is the right one.

In fact this kind of paradox, together with abuse of the Indifference Principle to assess, for example, the probability that the sun will rise tomorrow morning, threw a shadow over Bayesian methods at the end of the last century. The maximum likelihood method, which does not make explicit use of prior distributions, was then seen as a valid solution to the problem. But in reality the ambiguity of the proper metrics on which the initial distribution is uniform has an equivalent in the arbitrariness of the variable used in the likelihood function. In the end, what was criticized when it was stated explicitly in the Bayes formula is accepted passively when it is hidden in the maximum likelihood method.

5.4 Normally distributed observables

5.4.1 Final distribution, prevision and credibility intervals of the true value

The first application of the Bayesian inference will be that of a normally distributed quantity. Let us take a data sample $\underline{q}$ of n_1 measurements, of which we calculate the average $\overline{q}_{n_1}$. In our formalism $\overline{q}_{n_1}$ is a realization of the random variable $\overline{Q}_{n_1}$. Let us assume we know the standard deviation σ of the variable Q, either because n_1 is very large and can be estimated accurately from the sample or because it was known *a priori* (We are not going to discuss in these notes the case of small samples and unknown variance.)[2] The property of the average (see Section 4.2.2) tells us that the likelihood $f(\overline{Q}_{n_1}\,|\,\mu,\sigma)$ is Gaussian:

$$\overline{Q}_{n_1} \sim \mathcal{N}(\mu, \sigma/\sqrt{n_1}). \tag{5.11}$$

To simplify the following notation, let us call x_1 this average and σ_1 the standard deviation of the average:

$$\begin{aligned} x_1 &= \overline{q}_{n_1}, & (5.12)\\ \sigma_1 &= \sigma/\sqrt{n_1}\,. & (5.13)\end{aligned}$$

We then apply (5.4) and get

$$f(\mu\,|\,x_1, \mathcal{N}(\cdot,\sigma_1)) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\, f_\circ(\mu)}{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\, f_\circ(\mu)\,\mathrm{d}\mu}\,. \tag{5.14}$$

At this point we have to make a choice for $f_\circ(\mu)$. A reasonable choice is to take, as a first guess, a uniform distribution defined over a large interval which includes x_1. It is not really important how large the interval is, for a few σ_1 away from x_1 the integrand at the denominator tends to zero because of the Gaussian function. What is important is that a constant $f_\circ(\mu)$ can be simplified in (5.14), obtaining

$$f(\mu\,|\,x_1, \mathcal{N}(\cdot,\sigma_1)) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\,\mathrm{d}\mu}\,. \tag{5.15}$$

The integral in the denominator is equal to unity, since integrating with respect to μ is equivalent to integrating with respect to x_1. The final result is then

$$f(\mu) = f(\mu\,|\,x_1, \mathcal{N}(\cdot,\sigma_1)) = \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, e^{-\frac{(\mu-x_1)^2}{2\,\sigma_1^2}} \quad : \tag{5.16}$$

[2]**Note added:** for criticisms about the standard treatment of the small-sample problem, see Ref. [22].

- the true value is normally distributed around x_1;
- its best estimate (prevision) is $\mathrm{E}[\mu] = x_1$;
- its variance is $\sigma_\mu = \sigma_1$;
- the confidence intervals, or 'credibility intervals', in which there is a certain probability of finding the true value are easily calculable:

Probability level (confidence level) (%)	Credibility interval (confidence interval)
68.3	$x_1 \pm \sigma_1$
90.0	$x_1 \pm 1.65\sigma_1$
95.0	$x_1 \pm 1.96\sigma_1$
99.0	$x_1 \pm 2.58\sigma_1$
99.73	$x_1 \pm 3\sigma_1$

5.4.2 Combination of several measurements

Let us imagine making a second set of measurements of the physical quantity, which we assume unchanged from the previous set of measurements. How will our knowledge of μ change after this new information? Let us call $x_2 = \overline{q}_{n_2}$ and $\sigma_2 = \sigma'/\sqrt{n_2}$ the new average and standard deviation of the average (σ' may be different from σ of the sample of n_1 measurements), respectively. Applying Bayes' theorem a second time we now have to use as initial distribution the final probability of the previous inference:

$$f(\mu \,|\, x_1, \sigma_1, x_2, \sigma_2, \mathcal{N}) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma_2}\, e^{-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}}\, f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1))}{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_2}\, e^{-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}}\, f(\mu \,|\, x_1, \mathcal{N}(\cdot, \sigma_1))\, \mathrm{d}\mu}\,. \tag{5.17}$$

The integral is not as simple as the previous one, but still feasible analytically. The final result is

$$f(\mu \,|\, x_1, \sigma_1, x_2, \sigma_2, \mathcal{N}) = \frac{1}{\sqrt{2\,\pi}\,\sigma_A}\, e^{-\frac{(\mu - x_A)^2}{2\,\sigma_A^2}}\,, \tag{5.18}$$

where

$$x_A = \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}\,, \tag{5.19}$$

$$\frac{1}{\sigma_A^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\,. \tag{5.20}$$

One recognizes the famous formula of the weighted average with the inverse of the variances, usually obtained from maximum likelihood. There are some comments to be made.

- Bayes' theorem updates the knowledge about μ in an automatic and natural way.

- If $\sigma_1 \gg \sigma_2$ (and x_1 is not too far from x_2) the final result is only determined by the second sample of measurements. This suggests that an alternative vague *a priori* distribution can be, instead of uniform, a Gaussian with a large enough variance and a reasonable mean.

- The combination of the samples requires a subjective judgement that the two samples are really coming from the same true value μ. We will not discuss this point in these notes,[3] but a hint on how to proceed is to take the inference on the difference of two measurements, D, as explained at the end of Section 5.6.1 and judge yourself whether $D = 0$ is consistent with the probability density function of D.

5.4.3 Measurements close to the edge of the physical region

A case which has essentially no solution in the maximum likelihood approach is when a measurement is performed at the edge of the physical region and the measured value comes out very close to it, or even on the unphysical region. Let us take a numerical example.

Problem: An experiment is planned to measure the (electron) neutrino mass. The simulations show that the mass resolution is $3.3\,\mathrm{eV}/c^2$, largely independent of the mass value, and that the measured mass is normally distributed around the true mass.[4] The mass value which results from the analysis procedure,[5] and corrected for all known systematic effects, is $x = -5.41\,\mathrm{eV}/c^2$. What have we learned about the neutrino mass?

Solution: Our *a priori* value of the mass is that it is positive and not too large (otherwise it would already have been measured in other experiments). One can take any vague distribution which assigns a probability density function between 0 and 20 or 30 eV/c^2. In fact, if an experiment having a resolution of $\sigma = 3.3\,\mathrm{eV}/c^2$ has been planned and financed by rational people, with the hope of finding evidence of non-negligible mass, it means that the mass was thought to be in that range. If there is no reason to prefer one of the values in that interval a uniform distribution can be used, for example

$$f_{\circ K}(m) = k = 1/30 \qquad (0 \le m \le 30)\,. \tag{5.21}$$

Otherwise, if one thinks there is a greater chance of the mass having small rather than high values, a prior which reflects such an assumption could be chosen, for example a half normal with $\sigma_\circ = 10\,\mathrm{eV}$

$$f_{\circ N}(m) = \frac{2}{\sqrt{2\,\pi}\,\sigma_\circ}\,\exp\left[-\frac{m^2}{2\,\sigma_\circ^2}\right] \qquad (m \ge 0)\,, \tag{5.22}$$

or a triangular distribution

$$f_{\circ T}(m) = \frac{1}{450}\,(30 - x) \qquad (0 \le m \le 30)\,. \tag{5.23}$$

[3]**Note added**: as is easy to imagine, the problem of the 'outliers' should be treated with care, and surely avoiding automatic prescriptions. Some hints can be found in Refs. [43] and [44], and references therein.

[4]In reality, often m^2 rather than m is normally distributed. In this case the terms of the problem change and a new solution should be worked out, following the trace indicated in this example.

[5]We consider detector and analysis machinery as a black box, no matter how complicated it is, and treat the numerical outcome as a result of a direct measurement [1].

Let us consider for simplicity the uniform distribution

$$f(m\,|\,x, f_{\circ K}) = \frac{\frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp\left[-\frac{(m-x)^2}{2\,\sigma^2}\right] k}{\int_0^{30}\frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp\left[-\frac{(m-x)^2}{2\,\sigma^2}\right] k\,\mathrm{d}\mu} \tag{5.24}$$

$$= \frac{19.8}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(m-x)^2}{2\sigma^2}\right] \qquad (0 \le m \le 30)\,. \tag{5.25}$$

The value which has the highest degree of belief is $m = 0$, but $f(m)$ is non-vanishing up to $30\,\mathrm{eV}/c^2$ (even if very small). We can define an interval, starting from $m = 0$, in which we believe that m should have a certain probability. For example this level of probability can be 95%. One has to find the value $m_\circ$ for which the cumulative function $F(m_\circ)$ equals 0.95. This value of m is called the upper limit (or upper bound). The result is

$$m < 3.9\,\mathrm{eV}/c^2 \quad \text{at } 0.95\% \text{ probability}\,. \tag{5.26}$$

If we had assumed the other initial distributions the limit would have been in both cases

$$m < 3.7\,\mathrm{eV}/c^2 \quad \text{at } 0.95\% \text{ probability}\,, \tag{5.27}$$

practically the same (especially if compared with the experimental resolution of $3.3\,\mathrm{eV}/c^2$).

Comment: Let us assume an *a priori* function sharply peaked at zero and see what happens. For example it could be of the kind

$$f_{\circ S}(m) \propto \frac{1}{m}\,. \tag{5.28}$$

To avoid singularities in the integral, let us take a power of m slightly greater than -1, for example -0.99, and let us limit its domain to 30, getting

$$f_{\circ S}(m) = \frac{0.01 \cdot 30^{0.01}}{m^{0.99}}\,. \tag{5.29}$$

The upper limit becomes

$$m < 0.006\,\mathrm{eV}/c^2 \quad \text{at } \ 0.95\% \text{ probability}\,. \tag{5.30}$$

Any experienced physicist would find this result ridiculous. The upper limit is less than 0.2% of the experimental resolution; rather like expecting to resolve objects having dimensions smaller than a micron with a design ruler! Note instead that in the previous examples the limit was always of the order of magnitude of the experimental resolution σ. As $f_{\circ S}(m)$ becomes more and more peaked at zero (power of $x \to 1$) the limit gets smaller and smaller. This means that, asymptotically, the degree of belief that $m = 0$ is so high that whatever you measure you will conclude that $m = 0$: you could use the measurement to calibrate the apparatus! This means that this choice of initial distribution was unreasonable.

Instead, priors motivated by the positive attitude of the researchers are much more stable, and even when the observation is very negative the result is stable, and one always gets a limit of the order of the experimental resolution. Anyhow, it is also clear that when x is several σ below zero one starts to suspect that something is wrong with the experiment, which formally corresponds to doubts about the likelihood itself.

5.5 Counting experiments

5.5.1 Binomially distributed observables

Let us assume we have performed n trials and obtained x favourable events. What is the probability of the next event? This situation happens frequently when measuring efficiencies, branching ratios, etc. Stated more generally, one tries to infer the constant and unknown probability[6] of an event occurring.

Where we can assume that the probability is constant and the observed number of favourable events are binomially distributed, the unknown quantity to be measured is the parameter p of the binomial. Using Bayes' theorem we get

$$\begin{aligned} f(p\,|\,x,n,\mathcal{B}) &= \frac{f(x\,|\,\mathcal{B}_{n,p})\,f_\circ(p)}{\int_0^1 f(x\,|\,\mathcal{B}_{n,p})\,f_\circ(p)\,\mathrm{d}p} \\ &= \frac{\frac{n!}{(n-x)!\,x!}\,p^x\,(1-p)^{n-x}\,f_\circ(p)}{\int_0^1 \frac{n!}{(n-x)!\,x!}\,p^x\,(1-p)^{n-x}\,f_\circ(p)\,\mathrm{d}p} \\ &= \frac{p^x\,(1-p)^{n-x}}{\int_0^1 p^x\,(1-p)^{n-x}\,\mathrm{d}p}\,, \end{aligned} \tag{5.31}$$

where an initial uniform distribution has been assumed. The final distribution is known to statisticians as Beta distribution since the integral at the denominator is the special function called β, defined also for real values of x and n (technically this is a Beta with parameters $a = x+1$ and $b = n-x+1$). In our case these two numbers are integer and the integral becomes equal to $x!\,(n-x)!/(n+1)!$. We then get

$$f(p\,|\,x,n,\mathcal{B}) = \frac{(n+1)!}{x!\,(n-x)!}\,p^x\,(1-p)^{n-x}\,, \tag{5.32}$$

some examples of which are shown in Fig. 5.2.

Expected value and the variance of this distribution are:

[6] This concept, which is very close to the physicist's mentality, is not correct from the probabilistic — cognitive — point of view. According to the Bayesian scheme, in fact, the probability changes with the new observations. The final inference of p, however, does not depend on the particular sequence yielding x successes over n trials. This can be seen in the next table where $f_n(p)$ is given as a function of the number of trials n, for the three sequences which give two successes (indicated by 1) in three trials [the use of (5.32) is anticipated]:

	Sequence		
n	011	101	110
0	1	1	1
1	$2\,(1-p)$	$2\,p$	$2\,p$
2	$6\,p\,(1-p)$	$6\,p\,(1-p)$	$3\,p^2$
3	$12\,p^2\,(1-p)$	$12\,p^2\,(1-p)$	$12\,p^2\,(1-p)$

This important result, related to the concept of exchangeability, allows a physicist who is reluctant to give up the concept of unknown constant probability to see the problem from his point of view, ensuring that the same numerical result is obtained.

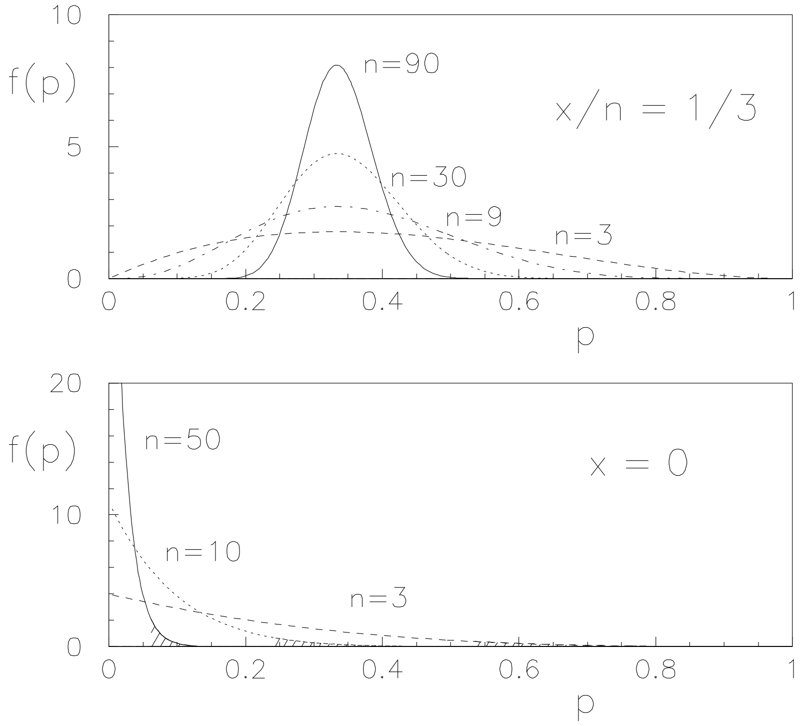

Figure 5.2: Probability density function of the binomial parameter p, having observed x successes in n trials.

$$\begin{aligned}
\mathrm{E}[p] &= \frac{x+1}{n+2}\,, & (5.33)\\
\mathrm{Var}(p) &= \frac{(x+1)(n-x+1)}{(n+3)(n+2)^2} & (5.34)\\
&= \frac{x+1}{n+2}\left(\frac{n+2}{n+2}-\frac{x+1}{n+2}\right)\frac{1}{n+3}\\
&= \mathrm{E}[p]\;(1-\mathrm{E}[p])\;\frac{1}{n+3}\,. & (5.35)
\end{aligned}$$

The value of p for which $f(p)$ has the maximum is instead $p_m = x/n$. The expression E[p] gives the prevision of the probability for the $(n+1)$-th event occurring and is called the recursive Laplace formula, or Laplace's rule of succession.

When x and n become large, and $0 \ll x \ll n$, $f(p)$ has the following asymptotic properties:

$$\begin{aligned}
\mathrm{E}[p] &\approx p_m = \frac{x}{n}\,, & (5.36)\\
\mathrm{Var}(p) &\approx \frac{x}{n}\left(1-\frac{x}{n}\right)\frac{1}{n} = \frac{p_m\,(1-p_m)}{n}\,, & (5.37)\\
\sigma_p &\approx \sqrt{\frac{p_m\,(1-p_m)}{n}}\,, & (5.38)\\
p &\sim \mathcal{N}(p_m,\sigma_p)\,. & (5.39)
\end{aligned}$$

Under these conditions the frequentistic definition (evaluation rule!) of probability (x/n) is recovered.

Let us see two particular situations: when $x = 0$ and $x = n$. In these cases one gives the result as upper or lower limits, respectively. Let us sketch the solutions:

- $\underline{x = n}$:

$$f(n \,|\, \mathcal{B}_{n,p}) = p^n, \tag{5.40}$$

$$f(p \,|\, x = n, \mathcal{B}) = \frac{p^n}{\int_0^1 p^n \, \mathrm{d}p} = (n+1)\, p^n, \tag{5.41}$$

$$F(p \,|\, x = n, \mathcal{B}) = p^{n+1}\,. \tag{5.42}$$

To get the 95% lower bound (limit):

$$F(p_\circ \,|\, x = n, \mathcal{B}) = 0.05\,,$$

$$p_\circ = \sqrt[n+1]{0.05}\,. \tag{5.43}$$

An increasing number of trials n constrain more and more p around 1.

- $\underline{x = 0}$:

$$f(0 \,|\, \mathcal{B}_{n,p}) = (1-p)^n, \tag{5.44}$$

$$f(p \,|\, x = 0, n, \mathcal{B}) = \frac{(1-p)^n}{\int_0^1 (1-p)^n \, \mathrm{d}p} = (n+1)\,(1-p)^n, \tag{5.45}$$

$$F(p \,|\, x = 0, n, \mathcal{B}) = 1 - (1-p)^{n+1}\,. \tag{5.46}$$

To get the 95% upper bound (limit):

$$F(p_\circ \,|\, x = 0, n, \mathcal{B}) = 0.95,$$

$$p_\circ = 1 - \sqrt[n+1]{0.05}\,. \tag{5.47}$$

The following table shows the 95% probability limits as a function of n. The Poisson approximation, to be discussed in the next section, is also shown.

	Probability level = 95%		
n	$x = n$	$x = 0$	
	binomial	binomial	Poisson approx. ($p_\circ = 3/n$)
3	$p \geq 0.47$	$p \leq 0.53$	$p \leq 1$
5	$p \geq 0.61$	$p \leq 0.39$	$p \leq 0.6$
10	$p \geq 0.76$	$p \leq 0.24$	$p \leq 0.3$
50	$p \geq 0.94$	$p \leq 0.057$	$p \leq 0.06$
100	$p \geq 0.97$	$p \leq 0.029$	$p \leq 0.03$
1000	$p \geq 0.997$	$p \leq 0.003$	$p \leq 0.003$

To show in this simple case how $f(p)$ is updated by the new information, let us imagine we have performed two experiments. The results are $x_1 = n_1$ and $x_2 = n_2$, respectively. Obviously the global information is equivalent to $x = x_1 + x_2$ and $n = n_1 + n_2$, with $x = n$. We then get

$$f(p\,|\,x=n,\mathcal{B}) = (n+1)\,p^n = (n_1+n_2+1)\,p^{n_1+n_2}\,. \tag{5.48}$$

A different way of proceeding would have been to calculate the final distribution from the information $x_1 = n_1$,

$$f(p\,|\,x_1=n_1,\mathcal{B}) = (n_1+1)\,p^{n_1}\,, \tag{5.49}$$

and feed it as initial distribution to the next inference:

$$\begin{aligned} f(p\,|\,x_1=n_1,x_2=n_2,\mathcal{B}) &= \frac{p^{n_2} f(p\,|\,x_1=n_1,\mathcal{B})}{\int_0^1 p^{n_2} f(p\,|\,x_1=n_1,\mathcal{B})\,\mathrm{d}p} && (5.50)\\ &= \frac{p^{n_2}(n_1+1)\,p^{n_1}}{\int_0^1 p^{n_2}(n_1+1)\,p^{n_1}\,\mathrm{d}p} && (5.51)\\ &= (n_1+n_2+1)\,p^{n_1+n_2}\,, && (5.52)\end{aligned}$$

getting the same result.

5.5.2 Poisson distributed quantities

As is well known, the typical application of the Poisson distribution is in counting experiments such as source activity, cross-sections, etc. The unknown parameter to be inferred is λ. Applying the Bayes formula we get

$$f(\lambda\,|\,x,\mathcal{P}) = \frac{\frac{\lambda^x\,e^{-\lambda}}{x!}\,f_\circ(\lambda)}{\int_0^\infty \frac{\lambda^x\,e^{-\lambda}}{x!}\,f_\circ(\lambda)\,\mathrm{d}\lambda}\,. \tag{5.53}$$

Assuming[7] $f_\circ(\lambda)$ constant up to a certain $\lambda_{max} \gg x$ and making the integral by parts we obtain

$$\begin{aligned} f(\lambda\,|\,x,\mathcal{P}) &= \frac{\lambda^x\,e^{-\lambda}}{x!}\,, && (5.54)\\ F(\lambda\,|\,x,\mathcal{P}) &= 1 - e^{-\lambda}\left(\sum_{n=0}^{x}\frac{\lambda^n}{n!}\right), && (5.55)\end{aligned}$$

where the last result has been obtained by integrating (5.54) also by parts. Figure 5.3 shows how to build the credibility intervals, given a certain measured number of counts x. Figure 5.4 shows some numerical examples.

$f(\lambda)$ has the following properties.

- Expected value, variance, and value of maximum probability are

$$\begin{aligned} \mathrm{E}[\lambda] &= x+1, && (5.56)\\ \mathrm{Var}(\lambda) &= x+1, && (5.57)\\ \lambda_m &= x\,. && (5.58)\end{aligned}$$

[7] There is a school of thought according to which the most appropriate function is $f_\circ(\lambda) \propto 1/\lambda$. If you think that it is reasonable for your problem, it may be a good prior. Claiming that this is 'the truth' is one of the many claims of the angels' sex determinations. For didactical purposes a uniform distribution is more than enough. Some comments about the $1/\lambda$ prescription will be given when discussing the particular case $x = 0$.
Note added: criticisms concerning so-called 'reference priors' can be found in Ref.[22].

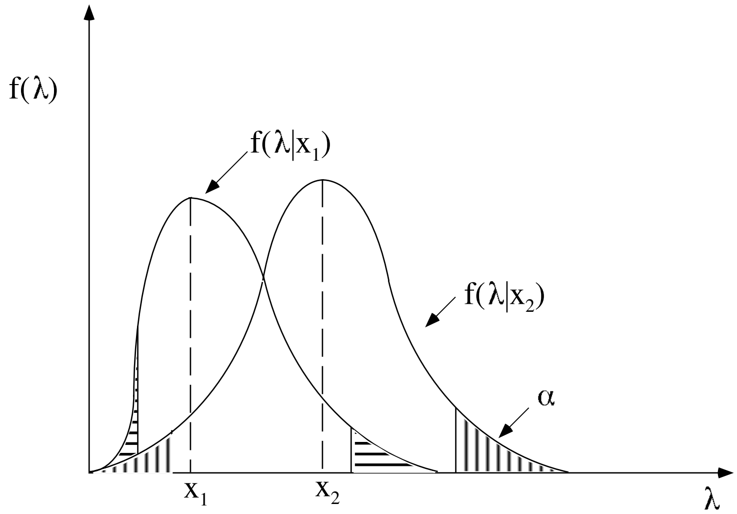

Figure 5.3: Poisson parameter λ inferred from an observed number x of counts.

The fact that the best estimate of λ in the Bayesian sense is not the intuitive value x but $x+1$ should neither surprise, nor disappoint us. According to the initial distribution used there are always more possible values of λ on the right side than on the left side of x, and they pull the distribution to their side; the full information is always given by $f(\lambda)$ and the use of the mean is just a rough approximation; the difference from the desired intuitive value x in units of the standard deviation goes as $1/\sqrt{x+1}$ and becomes immediately negligible.

- When x becomes large we get

$$\begin{aligned} \mathrm{E}[\lambda] &\approx \lambda_m = x\,, & (5.59)\\ \mathrm{Var}(\lambda) &\approx \lambda_m = x\,, & (5.60)\\ \sigma_\lambda &\approx \sqrt{x}\,, & (5.61)\\ \lambda &\sim \mathcal{N}(x, \sqrt{x})\,. & (5.62) \end{aligned}$$

Equation (5.61) is one of the most familar formulae used by physicists to assess the uncertainty of a measurement, although it is sometimes misused.

Let us conclude with a special case: $x = 0$ (see Fig. 5.5). As one might imagine, the inference is highly sensitive to the initial distribution. Let us assume that the experiment was planned with the hope of observing something, i.e. that it could detect a handful of events within its lifetime. With this hypothesis one may use any vague prior function not strongly peaked at zero. We have already come across a similar case in Section 5.4.3, concerning the upper limit of the neutrino mass. There it was shown that reasonable hypotheses based on the positive attitude of the experimentalist are almost equivalent and that they give results consistent with detector performances. Let us use then the uniform distribution

$$\begin{aligned} f(\lambda\,|\,x=0,\mathcal{P}) &= e^{-\lambda}, & (5.63)\\ F(\lambda\,|\,x=0,\mathcal{P}) &= 1-e^{-\lambda}, & (5.64)\\ \lambda &< 3 \text{ at } 95\% \text{ probability}\,. & (5.65) \end{aligned}$$

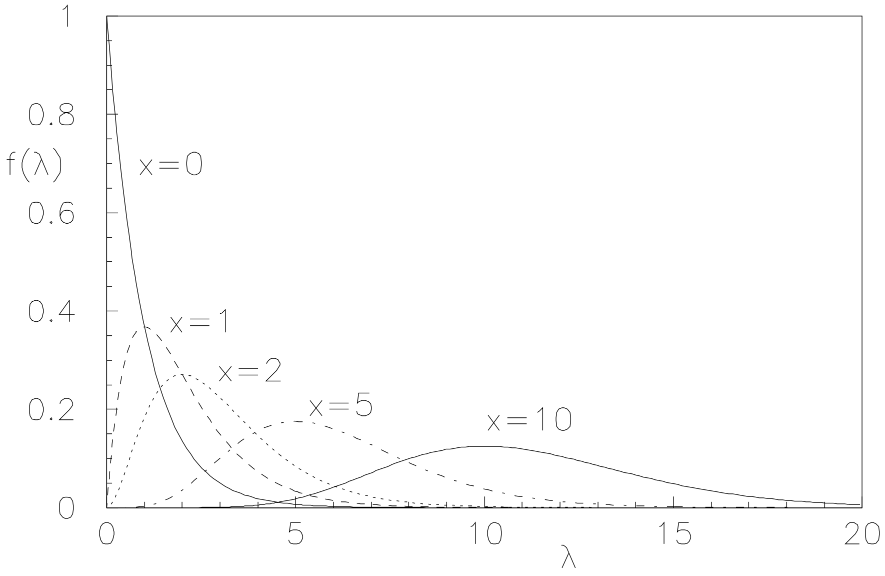

Figure 5.4: Examples of $f(\lambda\,|\,x_i)$.

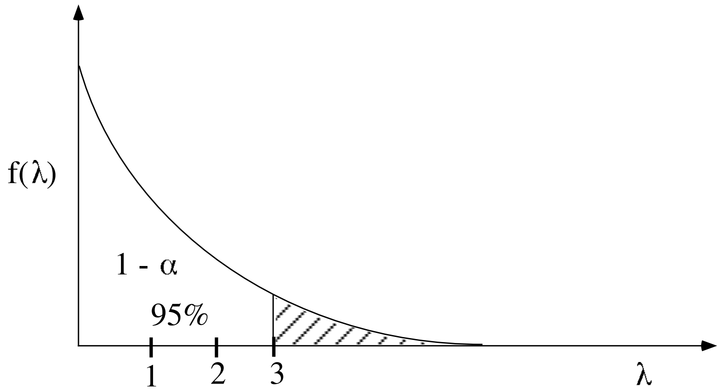

Figure 5.5: Upper limit to λ having observed 0 events.

5.6 Uncertainty due to systematic errors of unknown size

5.6.1 Example: uncertainty of the instrument scale offset

In our scheme any quantity of influence of which we do not know the exact value is a source of systematic error. It will change the final distribution of μ and hence its uncertainty. We have already discussed the most general case in Section 5.2.1. Let us make a simple application making a small variation to the example in Section 5.4.1: the 'zero' of the instrument is not known exactly, owing to calibration uncertainty. This can be parametrized assuming that its true value Z is normally distributed around 0 (i.e. the calibration was properly done!) with a standard deviation σ_Z. Since, most probably, the true value of μ is independent of the true value of Z, the initial joint probability density function can be written as the product of the marginal ones:

$$f_\circ(\mu, z) = f_\circ(\mu)\, f_\circ(z) = k\, \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right]. \tag{5.66}$$

Also the likelihood changes with respect to (5.11):

$$f(x_1 \,|\, \mu, z) = \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right]. \tag{5.67}$$

Putting all the pieces together and making use of (5.3) we finally get

$$f(\mu \,|\, x_1, \ldots, f_\circ(z)) = \frac{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \mathrm{d}z}{\iint \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\, \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z}\, \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \mathrm{d}\mu\, \mathrm{d}z}\,.$$

Integrating[8] we get

$$f(\mu) = f(\mu \,|\, x_1, \ldots, f_\circ(z)) = \frac{1}{\sqrt{2\,\pi}\,\sqrt{\sigma_1^2+\sigma_Z^2}}\, \exp\left[-\frac{(\mu-x_1)^2}{2\,(\sigma_1^2+\sigma_Z^2)}\right]. \tag{5.68}$$

The result is that $f(\mu)$ is still a Gaussian, but with a larger variance. The global standard uncertainty is the quadratic combination of that due to the statistical fluctuation of the data sample and the uncertainty due to the imperfect knowledge of the systematic effect:

$$\sigma_{tot}^2 = \sigma_1^2 + \sigma_Z^2\,. \tag{5.69}$$

This result is well known, although there are still some old-fashioned recipes which require different combinations of the contributions to be performed.

It must be noted that in this framework it makes no sense to speak of statistical and systematic uncertainties, as if they were of a different nature. They have the same probabilistic nature: $\overline{Q}_{n_1}$ is around μ with a standard deviation σ_1, and Z is around 0 with standard deviation σ_Z. What distinguishes the two components is how the knowledge of the uncertainty is gained: in one case (σ_1) from repeated measurements on the physics quantity of interest; in the second case (σ_Z) the evaluation was done by somebody else (the constructor of the instrument), or in a previous experiment, or guessed from the knowledge of the detector, or by simulation, etc. This is the reason why the ISO Guide [3] prefers the generic names 'type A' and 'type B' for the two kinds of contribution to global uncertainty. In particular, the name 'systematic uncertainty' should be avoided, while it is correct to speak about 'uncertainty due to a systematic effect'.

5.6.2 Correction for known systematic errors

It is easy to be convinced that if our prior knowledge about Z was of the kind

$$Z \sim \mathcal{N}(z_\circ, \sigma_Z) \tag{5.70}$$

[8] It may help to know that

$$\int_{-\infty}^{+\infty} \exp\left[b\,x - \frac{x^2}{a^2}\right] \mathrm{d}x = \sqrt{a^2\,\pi}\, \exp\left[\frac{a^2\,b^2}{4}\right].$$

the result would have been

$$\mu \sim \mathcal{N}\left(x_1 - z_\circ, \sqrt{\sigma_1^2 + \sigma_Z^2}\right), \tag{5.71}$$

i.e. one has first to correct the result for the best value of the systematic error and then include in the global uncertainty a term due to imperfect knowledge about it. This is a well-known and practised procedure, although there are still people who confuse $z_\circ$ with its uncertainty.

5.6.3 Measuring two quantities with the same instrument having an uncertainty of the scale offset

Let us take an example which is a little more complicated (at least from the mathematical point of view) but conceptually very simple and also very common in laboratory practice. We measure two physical quantities with the same instrument, assumed to have an uncertainty on the 'zero', modelled with a normal distribution as in the previous sections. For each of the quantities we collect a sample of data under the same conditions, which means that the unknown offset error does not change from one set of measurements to the other. Calling μ_1 and μ_2 the true values, x_1 and x_2 the sample averages, σ_1 and σ_2 the average's standard deviations, and Z the true value of the zero, the initial probability density and the likelihood are

$$f_\circ(\mu_1, \mu_2, z) = f_\circ(\mu_1)\, f_\circ(\mu_2)\, f_\circ(z) = k\, \frac{1}{\sqrt{2\,\pi}\,\sigma_Z} \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \tag{5.72}$$

and

$$\begin{aligned} f(x_1, x_2 \,|\, \mu_1, \mu_2, z) &= \frac{1}{\sqrt{2\,\pi}\,\sigma_1} \exp\left[-\frac{(x_1 - \mu_1 - z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_2} \exp\left[-\frac{(x_2 - \mu_2 - z)^2}{2\,\sigma_2^2}\right] \\ &= \frac{1}{2\,\pi\,\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{(x_1 - \mu_1 - z)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2 - z)^2}{\sigma_2^2}\right)\right], \end{aligned} \tag{5.73}$$

respectively. The result of the inference is now the joint probability density function of μ_1 and μ_2:

$$f(\mu_1, \mu_2 \,|\, x_1, x_2, \sigma_1, \sigma_2, f_\circ(z)) = \frac{\int f(x_1, x_2 \,|\, \mu_1, \mu_2, z)\, f_\circ(\mu_1, \mu_2, z)\mathrm{d}z}{\iiint f(x_1, x_2 \,|\, \mu_1, \mu_2, z)\, f_\circ(\mu_1, \mu_2, z)\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2\,\mathrm{d}z}, \tag{5.74}$$

where expansion of the functions has been omitted for the sake of clarity. Integrating we get

$$\begin{aligned} f(\mu_1, \mu_2) &= \frac{1}{2\,\pi\,\sqrt{\sigma_1^2 + \sigma_Z^2}\,\sqrt{\sigma_2^2 + \sigma_Z^2}\,\sqrt{1 - \rho^2}} \\ &\quad \exp\left\{-\frac{1}{2\,(1 - \rho^2)}\left[\frac{(\mu_1 - x_1)^2}{\sigma_1^2 + \sigma_Z^2} - 2\,\rho\,\frac{(\mu_1 - x_1)(\mu_2 - x_2)}{\sqrt{\sigma_1^2 + \sigma_Z^2}\,\sqrt{\sigma_2^2 + \sigma_Z^2}} + \frac{(\mu_2 - x_2)^2}{\sigma_2^2 + \sigma_Z^2}\right]\right\}. \end{aligned} \tag{5.75}$$

where

$$\rho = \frac{\sigma_Z^2}{\sqrt{\sigma_1^2 + \sigma_Z^2}\,\sqrt{\sigma_2^2 + \sigma_Z^2}}. \tag{5.76}$$

If σ_Z vanishes then (5.75) has the simpler expression

$$f(\mu_1,\mu_2) \xrightarrow[\sigma_Z \to 0]{} [\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\,\exp\left[-\frac{(\mu_1-x_1)^2}{2\,\sigma_1^2}\right]\,\frac{1}{\sqrt{2\,\pi}\,\sigma_2}\,\exp\left[-\frac{(\mu_2-x_2)^2}{2\,\sigma_2^2}\right]]$$

i.e. if there is no uncertainty on the offset calibration then the joint density function $f(\mu_1,\mu_2)$ is equal to the product of two independent normal functions, i.e. μ_1 and μ_2 are independent. In the general case we have to conclude the following.

- The effect of the common uncertainty σ_Z makes the two values correlated, since they are affected by a common unknown systematic error; the correlation coefficient is always non-negative ($\rho \geq 0$), as intuitively expected from the definition of systematic error.

- The joint density function is a multinormal distribution of parameters x_1, $\sigma_{\mu_1} = \sqrt{\sigma_1^2+\sigma_Z^2}$, x_2, $\sigma_{\mu_2} = \sqrt{\sigma_2^2+\sigma_Z^2}$, and ρ (see example of Fig. 4.1).

- The marginal distributions are still normal:

$$\mu_1 \sim \mathcal{N}\left(x_1, \sqrt{\sigma_1^2+\sigma_Z^2}\right), \tag{5.77}$$

$$\mu_2 \sim \mathcal{N}\left(x_2, \sqrt{\sigma_2^2+\sigma_Z^2}\right). \tag{5.78}$$

- The covariance between μ_1 and μ_2 is

$$\begin{aligned} \mathrm{Cov}(\mu_1,\mu_2) &= \rho\,\sigma_{\mu_1}\sigma_{\mu_2} \\ &= \rho\,\sqrt{\sigma_1^2+\sigma_Z^2}\,\sqrt{\sigma_2^2+\sigma_Z^2} = \sigma_Z^2\,. \end{aligned} \tag{5.79}$$

- The distribution of any function $g(\mu_1,\mu_2)$ can be calculated using the standard methods of probability theory. For example, one can demonstrate that the sum $S = \mu_1+\mu_2$ and the difference $D = \mu_1 - \mu_2$ are also normally distributed (see also the introductory discussion to the central limit theorem and Section 6.3 for the calculation of averages and standard deviations):

$$S \sim \mathcal{N}\left(x_1+x_2, \sqrt{\sigma_1^2+\sigma_2^2+(2\,\sigma_Z)^2}\right), \tag{5.80}$$

$$D \sim \mathcal{N}\left(x_1-x_2, \sqrt{\sigma_1^2+\sigma_2^2}\right). \tag{5.81}$$

  The result can be interpreted in the following way.

  - The uncertainty on the difference does not depend on the common offset uncertainty: whatever the value of the true zero is, it cancels in differences.
  - In the sum, instead, the effect of the common uncertainty is somewhat amplified since it enters 'in phase' in the global uncertainty of each of the quantities.

5.6.4 Indirect calibration

Let us use the result of the previous section to solve another typical problem of measurements. Suppose that after (or before, it doesn't matter) we have done the measurements of x_1 and x_2 and we have the final result, summarized in (5.75), we know the exact value of μ_1 (for example we perform the measurement on a reference). Let us call it μ_1°. Will this information provide a better knowledge of μ_2? In principle yes: the difference between x_1 and μ_1° defines the systematic error (the true value of the zero Z). This error can then be subtracted from x_2 to get a corrected value. Also the overall uncertainty of μ_2 should change, intuitively it should decrease, since we are adding new information. But its value doesn't seem to be obvious, since the logical link between μ_1° and μ_2 is $\mu_1^\circ \rightarrow Z \rightarrow \mu_2$.

The problem can be solved exactly using the concept of conditional probability density function $f(\mu_2 \,|\, \mu_1^\circ)$ [see (4.70)–(4.71)]. We get

$$\mu_{2\,|\,\mu_1^\circ} \sim \mathcal{N}\left(x_2 + \frac{\sigma_Z^2}{\sigma_1^2+\sigma_Z^2}\,(\mu_1^\circ - x_1),\ \sqrt{\sigma_2^2 + \left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_Z^2}\right)^{-1}}\right). \tag{5.82}$$

The best value of μ_2 is shifted by an amount Δ, with respect to the measured value x_2, which is not exactly $x_1 - \mu_1^\circ$, as was naïvely guessed, and the uncertainty depends on σ_2, σ_Z and σ_1. It is easy to be convinced that the exact result is more reasonable than the (suggested) first guess. Let us rewrite Δ in two different ways:

$$\Delta = \frac{\sigma_Z^2}{\sigma_1^2+\sigma_Z^2}\,(\mu_1^\circ - x_1) \tag{5.83}$$

$$= \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_Z^2}}\left[\frac{1}{\sigma_1^2}\cdot(x_1-\mu_1^\circ) + \frac{1}{\sigma_Z^2}\cdot 0\right]. \tag{5.84}$$

- Equation (5.83) shows that one has to apply the correction $x_1 - \mu_1^\circ$ only if $\sigma_1 = 0$. If instead $\sigma_Z = 0$ there is no correction to be applied, since the instrument is perfectly calibrated. If $\sigma_1 \approx \sigma_Z$ the correction is half of the measured difference between x_1 and μ_1°.

- Equation (5.84) shows explicitly what is going on and why the result is consistent with the way we have modelled the uncertainties. In fact we have performed two independent calibrations: one of the offset and one of μ_1. The best estimate of the true value of the zero Z is the weighted average of the two measured offsets.

- The new uncertainty of μ_2 [see (5.82)] is a combination of σ_2 and the uncertainty of the weighted average of the two offsets. Its value is smaller than it would be with only one calibration and, obviously, larger than that due to the sampling fluctuations alone:

$$\sigma_2 \le \sqrt{\sigma_2^2 + \frac{\sigma_1^2\,\sigma_Z^2}{\sigma_1^2+\sigma_Z^2}} \le \sqrt{\sigma_2^2+\sigma_Z^2}\,. \tag{5.85}$$

5.6.5 Counting measurements in the presence of background

As an example of a different kind of systematic effect, let us think of counting experiments in the presence of background. For example we are searching for a new particle, we make some selection cuts and count x events. But we also expect an average number of background events $\lambda_{B_\circ} \pm \sigma_B$, where σ_B is the standard uncertainty of $\lambda_{B_\circ}$, not to be confused with $\sqrt{\lambda_{B_\circ}}$. What can we say about λ_S, the true value of the average number associated with the signal? First we will treat the case in which the determination of the expected number of background events is well known ($\sigma_B/\lambda_{B_\circ} \ll 1$), and then the general case.

$\boxed{\sigma_B/\lambda_{B_\circ} \ll 1}$: The true value of the sum of signal and background is $\lambda = \lambda_S + \lambda_{B_\circ}$. The likelihood is

$$P(x\,|\,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}\,. \tag{5.86}$$

Applying Bayes' theorem we have

$$f(\lambda_S\,|\,x, \lambda_{B_\circ}) \;=\; \frac{e^{-(\lambda_{B_\circ}+\lambda_S)}\,(\lambda_{B_\circ}+\lambda_S)^x\,f_\circ(\lambda_S)}{\int_0^\infty e^{-(\lambda_{B_\circ}+\lambda_S)}\,(\lambda_{B_\circ}+\lambda_S)^x\,f_\circ(\lambda_S)\,\mathrm{d}\lambda_S}\,. \tag{5.87}$$

Choosing again $f_\circ(\lambda_S)$ uniform (in a reasonable interval) this gets simplified. The integral in the denominator can be calculated easily by parts and the final result is

$$f(\lambda_S\,|\,x, \lambda_{B_\circ}) \;=\; \frac{e^{-\lambda_S}\,(\lambda_{B_\circ}+\lambda_S)^x}{x!\,\sum_{n=0}^{x}\frac{\lambda_{B_\circ}^n}{n!}}\,, \tag{5.88}$$

$$F(\lambda_S\,|\,x, \lambda_{B_\circ}) \;=\; 1 - \frac{e^{-\lambda_S}\,\sum_{n=0}^{x}\frac{(\lambda_{B_\circ}+\lambda_S)^n}{n!}}{\sum_{n=0}^{x}\frac{\lambda_{B_\circ}^n}{n!}}\,. \tag{5.89}$$

From (5.88) and (5.89) it is possible to calculate in the usual way the best estimate and the credibility intervals of λ_S. Two particular cases are of interest:

- If $\lambda_{B_\circ} = 0$ then formulae (5.54) and (5.55) are recovered. In such a case one measured count is enough to claim for a signal (if somebody is willing to believe that really $\lambda_{B_\circ} = 0$ without any uncertainty ...).
- If $x = 0$ then

$$f(\lambda\,|\,x, \lambda_{B_\circ}) = e^{-\lambda_S}\,, \tag{5.90}$$

  independently of $\lambda_{B_\circ}$. This result is not really obvious.

$\boxed{\textbf{Any } g(\lambda_{B_\circ})}$: In the general case, the true value of the average number of background events λ_B is unknown. We only known that it is distributed around $\lambda_{B_\circ}$ with standard deviation σ_B and probability density function $g(\lambda_B)$, not necessarily a Gaussian. What changes with respect to the previous case is the initial distribution, now a joint function of λ_S and of λ_B. Assuming λ_B and λ_S independent the prior density function is

$$f_\circ(\lambda_S, \lambda_B) = f_\circ(\lambda_S)\,g_\circ(\lambda_B)\,. \tag{5.91}$$

We leave $f_\circ$ in the form of a joint distribution to indicate that the result we shall get is the most general for this kind of problem. The likelihood, on the other hand, remains the same as in the previous example. The inference of λ_S is done in the usual way, applying Bayes' theorem and marginalizing with respect to λ_S:

$$f(\lambda_S\,|\,x, g_\circ(\lambda_B)) = \frac{\int e^{-(\lambda_B+\lambda_S)}\,(\lambda_B+\lambda_S)^x\,f_\circ(\lambda_S,\lambda_B)\,\mathrm{d}\lambda_\mathrm{B}}{\iint e^{-(\lambda_B+\lambda_S)}\,(\lambda_B+\lambda_S)^x\,f_\circ(\lambda_S,\lambda_B)\,\mathrm{d}\lambda_S\,\mathrm{d}\lambda_B}\,. \tag{5.92}$$

The previous case [formula (5.88)] is recovered if the only value allowed for λ_B is $\lambda_{B_\circ}$ and $f_\circ(\lambda_S)$ is uniform:

$$f_\circ(\lambda_S, \lambda_B) = k\,\delta(\lambda_B - \lambda_{B_\circ})\,. \tag{5.93}$$

Chapter 6

Bypassing Bayes' theorem for routine applications

"Let us consider a dimensionless mass,
suspended from an inextensible massless wire,
free to oscillate without friction ... "
(Any textbook)

6.1 Approximate methods

6.1.1 Linearization

We have seen in the above examples how to use the general formula (5.3) for practical applications. Unfortunately, when the problem becomes more complicated one starts facing integration problems. For this reason approximate methods are generally used. We will derive the approximation rules consistently with the approach followed in these notes and then the resulting formulae will be compared with the ISO recommendations. To do this, let us neglect for a while all quantities of influence which could produce unknown systematic errors. In this case (5.3) can be replaced by (5.4), which can be further simplified if we remember that correlations between the results are originated by unknown systematic errors. In the absence of these, the joint distribution of all quantities $\underline{\mu}$ is simply the product of marginal ones:

$$f_{R_i}(\underline{\mu_i}) = \prod_i f_{R_i}(\mu_i)\,, \tag{6.1}$$

with

$$f_{R_i}(\mu_i) = f_{R_i}(\mu_i\,|\,x_i, \underline{h}_\circ) = \frac{f(x_i\,|\,\mu_i, \underline{h}_\circ)\, f_\circ(\mu_i)}{\int f(x_i\,|\,\mu_i, \underline{h}_\circ)\, f_\circ(\mu_i)\,\mathrm{d}\mu_\mathrm{i}}\,. \tag{6.2}$$

The symbol $f_{R_i}(\mu_i)$ indicates that we are dealing with raw values[1] evaluated at $\underline{h} = \underline{h}_\circ$. Since for any variation of $\underline{h}$ the inferred values of μ_i will change, it is convenient to name with the same subscript R the quantity obtained for $\underline{h}_\circ$:

$$f_{R_i}(\mu_i) \longrightarrow f_{R_i}(\mu_{R_i})\,. \tag{6.3}$$

[1]The choice of the adjective 'raw' will become clearer later on. The subscript R is also meant to represent 'random', in the sense that only sampling effects are considered at the moment.

Let us indicate with $\widehat{\mu}_{R_i}$ and σ_{R_i} the best estimates and the standard uncertainty of the raw values:

$$\begin{aligned} \widehat{\mu}_{R_i} &= \mathrm{E}[\mu_{R_i}]\,, & (6.4)\\ \sigma^2_{R_i} &= \mathrm{Var}(\mu_{R_i})\,. & (6.5) \end{aligned}$$

For any possible configuration of conditioning hypotheses $\underline{h}$, corrected values μ_i are obtained:

$$\mu_i = \mu_{R_i} + g_i(\underline{h})\,. \tag{6.6}$$

The function which relates the corrected value to the raw value and to the systematic effects has been denoted by g_i so as not to be confused with a probability density function. Expanding (6.6) in series around $\underline{h}_\circ$ we finally arrive at the expression which will allow us to make the approximated evaluations of uncertainties:

$$\boxed{\mu_i = \mu_{R_i} + \sum_l \frac{\partial g_i}{\partial h_l}\,(h_l - h_{\circ_l}) + \ldots} \tag{6.7}$$

(All derivatives are evaluated at $\{\widehat{\mu}_{R_i}, \underline{h}_\circ\}$. To simplify the notation a similar convention will be used in the following formulae.)

Neglecting the terms of the expansion above the first order, and taking the expected values, we get

$$\begin{aligned} \widehat{\mu}_i &= \mathrm{E}[\mu_i] \\ &\approx \widehat{\mu}_{R_i}\,; & (6.8)\\ \sigma^2_{\mu_i} &= \mathrm{E}\left[(\mu_i - \mathrm{E}[\mu_i])^2\right] \\ &\approx \sigma^2_{R_i} + \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)^2 \sigma^2_{h_l} \\ &\quad \left\{ + 2\sum_{l<m} \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_i}{\partial h_m}\right) \rho_{lm}\,\sigma_{h_l}\,\sigma_{h_m} \right\}\,; & (6.9)\\ \mathrm{Cov}(\mu_i, \mu_j) &= \mathrm{E}\left[\,(\mu_i - \mathrm{E}[\mu_i])(\mu_j - \mathrm{E}[\mu_j])\,\right] \\ &\approx \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_j}{\partial h_l}\right) \sigma^2_{h_l} \\ &\quad \left\{ + 2\sum_{l<m} \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_j}{\partial h_m}\right) \rho_{lm}\,\sigma_{h_l}\,\sigma_{h_m} \right\}\,. & (6.10) \end{aligned}$$

The terms included within $\{\cdot\}$ vanish if the unknown systematic errors are uncorrelated, and the formulae become simpler. Unfortunately, very often this is not the case, as when several calibration constants are simultaneously obtained from a fit (for example, in most linear fits slope and intercept have a correlation coefficient close to -0.9).

Sometimes the expansion (6.7) is not performed around the best values of $\underline{h}$ but around their nominal values, in the sense that the correction for the known value of the systematic errors has not yet been applied (see Section 5.6.2). In this case (6.7) should be replaced by

$$\mu_i = \mu_{R_i} + \sum_l \frac{\partial g_i}{\partial h_l}\,(h_l - h_{N_l}) + \ldots\,, \tag{6.11}$$

where the subscript N stands for nominal. The best value of μ_i is then

$$\begin{aligned}
\widehat{\mu}_i &= \mathrm{E}[\mu_i] \\
&\approx \widehat{\mu}_{R_i} + \mathrm{E}\left[\sum_l \frac{\partial g_i}{\partial h_l}(h_l - h_{N_l})\right] \\
&= \widehat{\mu}_{R_i} + \sum_l \delta\mu_{i_l}\,. \qquad (6.12)
\end{aligned}$$

(6.9) and (6.10) instead remain valid, with the condition that the derivative is calculated at $\underline{h}_N$. If $\rho_{lm} = 0$, it is possible to rewrite (6.9) and (6.10) in the following way, which is very convenient for practical applications:

$$\begin{aligned}
\sigma^2_{\mu_i} &\approx \sigma^2_{R_i} + \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)^2 \sigma^2_{h_l} \qquad (6.13)\\
&= \sigma^2_{R_i} + \sum_l u^2_{i_l}\,; \qquad (6.14)\\
\mathrm{Cov}(\mu_i,\mu_j) &\approx \sum_l \left(\frac{\partial g_i}{\partial h_l}\right)\left(\frac{\partial g_j}{\partial h_l}\right)\sigma^2_{h_l} \qquad (6.15)\\
&= \sum_l s_{ij_l}\left|\frac{\partial g_i}{\partial h_l}\right|\sigma_{h_l}\left|\frac{\partial g_j}{\partial h_l}\right|\sigma_{h_l} \qquad (6.16)\\
&= \sum_l s_{ij_l}\, u_{i_l}\, u_{j_l} \qquad (6.17)\\
&= \sum_l \mathrm{Cov}_l(\mu_i,\mu_j)\,. \qquad (6.18)
\end{aligned}$$

u_{i_l} is the component of the standard uncertainty due to effect h_l. s_{ij_l} is equal to the product of signs of the derivatives, which takes into account whether the uncertainties are positively or negatively correlated.

To summarize, when systematic effects are not correlated with each other, the following quantities are needed to evaluate the corrected result, the combined uncertainties and the correlations:

- the raw $\widehat{\mu}_{R_i}$ and σ_{R_i};
- the best estimates of the corrections $\delta\mu_{i_l}$ for each systematic effect h_l;
- the best estimate of the standard deviation u_{i_l} due to the imperfect knowledge of the systematic effect;
- for any pair $\{\mu_i, \mu_j\}$ the sign of the correlation s_{ij_l} due to the effect h_l.

In HEP applications it is frequently the case that the derivatives appearing in (6.12)–(6.16) cannot be calculated directly, as for example when h_l are parameters of a simulation program, or acceptance cuts. Then variations of $\underline{\mu}_i$ are usually studied by varying a particular h_l within a reasonable interval, holding the other influence quantities at the nominal value. $\delta\mu_{i_l}$ and u_{i_l} are calculated from the interval $\pm\Delta^{\pm}_i$ of variation of the true value for a given variation $\pm\Delta^{\pm}_{h_l}$ of h_l and from the probabilistic meaning of the intervals (i.e. from the assumed distribution of the true value). This empirical procedure for determining $\delta\mu_{i_l}$ and u_{i_l} has the advantage that it can take into account nonlinear effects [45], since it directly measures the difference $\widehat{\mu}_i - \widehat{\mu}_{R_i}$ for a given difference $h_l - h_{N_l}$.

Some examples are given in Section 6.1.4, and two typical experimental applications will be discussed in more detail in Section 6.3.

6.1.2 BIPM and ISO recommendations

In this section we compare the results obtained in the previous section with the recommendations of the Bureau International des Poids et Mesures (BIPM) and the International Organization for Standardization (ISO) on the expression of experimental uncertainty (Refs. [2, 3]).

1. *"The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:*

   A: *those which are evaluated by statistical methods;*

   B: *those which are evaluated by other means.*

   There is not always a simple correspondence between the classification into categories A or B and the previously used classification into 'random' and 'systematic' uncertainties. The term 'systematic uncertainty' can be misleading and should be avoided.

   The detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical result."

   Essentially the first recommendation states that all uncertainties can be treated probabilistically. The distinction between types A and B is subtle and can be misleading if one thinks of statistical methods as synonymous with probabilistic methods, as is currently the case in HEP. Here 'statistical' has the classical meaning of repeated measurements.

2. *"The components in category A are characterized by the estimated variances s_i^2 (or the estimated "standard deviations" s_i) and the number of degrees of freedom ν_i. Where appropriate, the covariances should be given."*

   The estimated variances correspond to $\sigma_{R_i}^2$ of the previous section. The degrees of freedom are related to small samples and to the Student t-distribution. The problem of small samples is not discussed in these notes, but clearly this recommendation is a relic of frequentistic methods. With the approach followed in these notes there is no need to talk about degrees of freedom, since the Bayesian inference defines the final probability function $f(\mu)$ completely.[2]

3. *"The components in category B should be characterized by quantities u_j^2, which may be considered as approximations to the corresponding variances, the existence of which is assumed. The quantities u_j^2 may be treated like variances and the quantities u_j like standard deviations. Where appropriate, the covariances should be treated in a similar way."*

   Clearly, this recommendation is meaningful only in a Bayesian framework.

4. *"The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of 'standard deviations'."*

   This is what we have found in (6.9) and (6.10).

[2] **Note added**: for criticisms about the standard treatment of the small-sample problem see Ref. [22].

5. *"If, for particular applications, it is necessary to multiply the combined uncertainty by a factor to obtain an overall uncertainty, the multiplying factor used must always be stated."*

   This last recommendation states once more that the uncertainty is by default the standard deviation of the true value distribution. Any other quantity calculated to obtain a credibility interval with a certain probability level should be clearly stated.

To summarize, the following are the basic ingredients of the BIPM/ISO recommendations.

subjective definition of probability: it allows variances to be assigned conceptually to any physical quantity which has an uncertain value;

uncertainty as standard deviation:

- it is standard;
- the rule of combination (4.62)–(4.66) applies to standard deviations and not to confidence intervals;

combined standard uncertainty: it is obtained by the usual formula of error propagation and it makes use of variances, covariances and first derivatives;

central limit theorem: it makes, under proper conditions, the true value normally distributed if one has several sources of uncertainty.

Consultation of the ISO Guide [3] is recommended for further explanations about the justification of the standards, for the description of evaluation procedures, and for examples. I would just like to end this section with some examples of the evaluation of type B uncertainties and with some words of caution concerning the use of approximations and of linearization.

6.1.3 Evaluation of type B uncertainties

The ISO Guide states that

> *"For estimate x_i of an input quantity[3] X_i that has not been obtained from repeated observations, the ... standard uncertainty u_i is evaluated by scientific judgement based on all the available information on the possible variability of X_i. The pool of information may include*
>
> - *previous measurement data;*
> - *experience with or general knowledge of the behaviour and properties of relevant materials and instruments;*
> - *manufacturer's specifications;*
> - *data provided in calibration and other certificates;*
> - *uncertainties assigned to reference data taken from handbooks."*

6.1.4 Examples of type B uncertainties

1. Previous measurements of other particular quantities, performed in similar conditions, have provided a repeatability standard deviation[4] of σ_r:

$$u = \sigma_r \,.$$

[3] By 'input quantity' the ISO Guide means any of the contributions h_l or μ_{R_i} which enter into (6.9) and (6.10).
[4] This example shows a type B uncertainty originated by random errors.

2. A manufacturer's calibration certificate states that the uncertainty, defined as k standard deviations, is $\pm\Delta$:
$$u = \frac{\Delta}{k}\,.$$

3. A result is reported in a publication as $\overline{x} \pm \Delta$, stating that the average has been performed on four measurements and the uncertainty is a 95% confidence interval. One has to conclude that the confidence interval has been calculated using the Student t-distribution:
$$u = \frac{\Delta}{3.18}\,.$$

4. A manufacturer's specification states that the error on a quantity should not exceed Δ. With this limited information one has to assume a uniform distribution:
$$u = \frac{2\,\Delta}{\sqrt{12}} = \frac{\Delta}{\sqrt{3}}\,.$$

5. A physical parameter of a Monte Carlo is believed to lie in the interval of $\pm\Delta$ around its best value, but not with uniform distribution: the degree of belief that the parameter is at the centre is higher than the degree of belief that it is at the edges of the interval. With this information a triangular distribution can be reasonably assumed:
$$u = \frac{\Delta}{\sqrt{6}}\,.$$
Note that the coefficient in front of Δ changes from the 0.58 of the previous example to the 0.41 of this. If the interval $\pm\Delta$ were a 3σ interval then the coefficient would have been equal to 0.33. These variations — to be considered extreme — are smaller than the statistical fluctuations of empirical standard deviations estimated from ≈ 10 measurements. This shows that one should not be worried that the type B uncertainties are less accurate than type A, especially if one tries to model the distribution of the physical quantity honestly.

6. The absolute energy calibration of an electromagnetic calorimeter module is not known exactly and is estimated to be between the nominal one and +10%. The statistical error is known by test beam measurements to be $18\%/\sqrt{E/\mathrm{GeV}}$. What is the uncertainty on the energy measurement of an electron which has apparently released 30 GeV?
   - There is no type A uncertainty, since only one measurement has been performed.
   - The energy has to be corrected for the best estimate of the calibration constant: +5%, with a relative uncertainty of $18\%/\sqrt{31.5}$ due to sampling (the statistical error):
   $$E = 31.5 \pm 1.0\,\mathrm{GeV}\,.$$
   - Then one has to take into account the uncertainty due to absolute energy scale calibration:
     - assuming a uniform distribution of the true calibration constant, $u = 31.5 \times 0.1/\sqrt{12} = 0.9\,\mathrm{GeV}$:
     $$E = 31.5 \pm 1.3\,\mathrm{GeV}\,;$$
     - assuming, more reasonably, a triangular distribution, $u = 31.5 \times 0.05/\sqrt{6} = 0.6\,\mathrm{GeV}$,
     $$E = 31.5 \pm 1.2\,\mathrm{GeV}\,.$$

- Instead, interpreting the maximum deviation from the nominal calibration as uncertainty (see comment at the end of Section 5.6.2),

$$E = 30.0 \pm 1.0 \pm 3.0\,\mathrm{GeV} \rightarrow E = 30.0 \pm 3.2\,\mathrm{GeV}\,.$$

As already mentioned earlier in these notes, while reasonable assumptions (in this case the first two) give consistent results, this is not true if one makes inconsistent use of the information just for the sake of giving safe uncertainties.

7. **Note added**: the original version of the primer contained at this point a more realistic and slightly more complicated example, which requires, instead, a next-to-linear treatment [45], which was not included in the notes, neither is it in this new version. Therefore, I prefer to skip this example in order to avoid confusion.

6.1.5 Caveat concerning the blind use of approximate methods

The mathematical apparatus of variances and covariances of (6.9) and (6.10) is often seen as the most complete description of uncertainty and in most cases used blindly in further uncertainty calculations. It must be clear, however, that this is just an approximation based on linearization. If the function which relates the corrected value to the raw value and the systematic effects is not linear then the linearization may cause trouble. An interesting case is discussed in Section 6.3.

There is another problem which may arise from the simultaneous use of Bayesian estimators and approximate methods. Let us introduce the problem with an example.

Example 1: 1000 independent measurements of the efficiency of a detector have been performed (or 1000 measurements of branching ratio, if you prefer). Each measurement was carried out on a base of 100 events and each time 10 favourable events were observed (this is obviously strange — though not impossible — but it simplifies the calculations). The result of each measurement will be [see (5.33)–(5.35)]:

$$\widehat{\epsilon}_i = \frac{10+1}{100+2} = 0.1078, \tag{6.19}$$

$$\sigma(\epsilon_i) = \sqrt{\frac{11 \times 91}{103 \times 102^2}} = 0.031\,. \tag{6.20}$$

Combining the 1000 results using the standard weighted average procedure gives

$$\epsilon = 0.1078 \pm 0.0010\,. \tag{6.21}$$

Alternatively, taking the complete set of results to be equivalent to 100 000 trials with 10 000 favourable events, the combined result is

$$\epsilon' = 0.10001 \pm 0.0009 \tag{6.22}$$

(the same as if one had used Bayes' theorem iteratively to infer $f(\epsilon)$ from the the partial 1000 results). The conclusions are in disagreement and the first result is clearly mistaken (the solution will be given after the following example).

The same problem arises in the case of inference of the Poisson distribution parameter λ and, in general, whenever $f(\mu)$ is not symmetrical around $\mathrm{E}[\mu]$.

Example 2: Imagine an experiment running continuously for one year, searching for monopoles and identifying none. The consistency with zero can be stated either quoting $\mathrm{E}[\lambda] = 1$ and $\sigma_\lambda = 1$, or a 95% upper limit $\lambda < 3$. In terms of rate (number of monopoles per day) the result would be either $\mathrm{E}[r] = 2.7 \cdot 10^{-3}$, $\sigma(r) = 2.7 \cdot 10^{-3}$, or an upper limit $r < 8.2 \cdot 10^{-3}$. It is easy to show that, if we take the 365 results for each of the running days and combine them using the standard weighted average, we get $r = 1.00 \pm 0.05$ monopoles per day![5] This absurdity is not caused by the Bayesian method, but by the abuse of standard rules for combining the results (the weighted average formulae (5.19) and (5.20) are derived from the normal distribution hypothesis). Using Bayesian inference would have led to a consistent and reasonable result no matter how the 365 days of running had been subdivided for partial analysis.

This suggests that in some cases it could be preferable to present the result also providing the mode of μ (p_m and λ_m of Sections 5.5.1 and 5.5.2). This way of presenting the results is similar to that suggested by the maximum likelihood approach, with the difference that for $f(\mu)$ one should take the final probability density function and not simply the likelihood. Since it is practically impossible to summarize the outcome of an inference in only two numbers (best value and uncertainty), in case of non-normality of the $f(\mu)$, more information about $f(\mu)$ should be given.

6.2 Indirect measurements

Conceptually this is a very simple task in the Bayesian framework, whereas the frequentistic one requires a lot of gymnastics, going back and forth from the logical level of true values to the logical level of estimators. If one accepts that the true values are just random variables,[6] then, calling Y a function of other quantities X, each having a probability density function $f(x)$, the probability density function of Y $f(y)$ can be calculated with the standard formulae which follow from the rules probability. Note that in the approach presented in these notes uncertainties due to systematic effects are treated in the same way as indirect measurements. It is worth repeating that there is no conceptual distinction between various components of the measurement uncertainty. When approximations are sufficient, formulae (6.9) and (6.10) can be used.

Let us take an example for which the linearization does not give the right result.

Example: The speed of a proton is measured with a time-of-flight system. Find the 68, 95 and 99% probability intervals for the energy, knowing that $\beta = v/c = 0.9971$, and that distance and time have been measured with a 0.2% accuracy.

The relation

$$E = \frac{mc^2}{\sqrt{1-\beta^2}}$$

is strongly nonlinear. The results given by the approximated method and the correct one are shown in the table below.

[5]**Note added:** this is exactly the presumed paradox reported by the 1998 issue of the PDG [46] as an argument against Bayesian statistics (Section 29.6.2, p. 175: *"If Bayesian estimates are averaged, they do not converge to the true value, since they have all been forced to be positive."*)

[6]To make the formalism lighter, let us call both the random variable associated with the quantity and the quantity itself by the same name X_i (instead of μ_{x_i}).

Probability (%)	Linearization E (GeV)	Correct result E (GeV)
68	$6.4 \leq E \leq 18$	$8.8 \leq E \leq 64$
95	$0.7 \leq E \leq 24$	$7.2 \leq E < \infty$
99	$0. \leq E \leq 28$	$6.6 \leq E < \infty$

6.3 Covariance matrix of experimental results

This section, based on Ref. [47], shows once more practical rules to build the covariance matrix associated with experimental data with correlated uncertainty (see also Sections 5.6.3 and 6.1.1), treating explicitly also the case of normalization uncertainty. Then it will be shown that, in this case, the covariance matrix evaluated in this way produces biased χ^2 fits.

6.3.1 Building the covariance matrix of experimental data

In physics applications, it is rarely the case that the covariance between the best estimates of two physical quantities,[7] each given by the arithmetic average of direct measurements ($x_i = \overline{X_i} = \frac{1}{n}\sum_{k=1}^{n} X_{ik}$), can be evaluated from the sample covariance[8] of the two averages:

$$\mathrm{Cov}(x_i, x_j) = \frac{1}{n\,(n-1)} \sum_{k=1}^{n} (X_{ik} - \overline{X}_i)(X_{jk} - \overline{X}_j)\ . \tag{6.23}$$

More frequent is the well-understood case in which the physical quantities are obtained as a result of a χ^2 minimization, and the terms of the inverse of the covariance matrix are related to the curvature of χ^2 at its minimum:

$$\left(V^{-1}\right)_{ij} = \frac{1}{2}\left.\frac{\partial^2\chi^2}{\partial X_i \partial X_j}\right|_{x_i, x_j}\ . \tag{6.24}$$

In most cases one determines independent values of physical quantities with the same detector, and the correlation between them originates from the detector calibration uncertainties. Frequentistically, the use of (6.23) in this case would correspond to having a sample of detectors, each of which is used to perform a measurement of all the physical quantities.

A way of building the covariance matrix from the direct measurements is to consider the original measurements and the calibration constants as a common set of independent and uncorrelated measurements, and then to calculate corrected values that take into account the calibration constants. The variance/covariance propagation will automatically provide the full covariance matrix of the set of results. Let us derive it for two cases that occur frequently, and then proceed to the general case.

[7] In this section the symbol X_i will indicate the variable associated to the i-th physical quantity and X_{ik} its k-th direct measurement; x_i the best estimate of its value, obtained by an average over many direct measurements or indirect measurements, σ_i the standard deviation, and y_i the value corrected for the calibration constants. The weighted average of several x_i will be denoted by $\overline{x}$.

[8] **Note added:** The '$n-1$' at the denominator of (6.23) is for the same reason as the '$n-1$' of the sample standard deviation. Although I do not agree with the rationale behind it, this formula can be considered a kind of standard and, anyhow, replacing '$n-1$' by 'n' has no effect in normal applications. As already said, in these notes I will not discuss the small-sample problem; anyone who is interested in my worries concerning default formulae for small samples, as well as Student t-distribution may have a look at Ref. [22].

Offset uncertainty

Let $x_i \pm \sigma_i$ be the $i = 1 \ldots n$ results of independent measurements and $\mathbf{V}_X$ the (diagonal) covariance matrix. Let us assume that they are all affected by the same calibration constant c, having a standard uncertainty σ_c. The corrected results are then $y_i = x_i + c$. We can assume, for simplicity, that the most probable value of c is 0, i.e. the detector is well calibrated. One has to consider the calibration constant as the physical quantity X_{n+1}, whose best estimate is $x_{n+1} = 0$. A term $V_{X_{n+1,n+1}} = \sigma_c^2$ must be added to the covariance matrix.

The covariance matrix of the corrected results is given by the transformation

$$\mathbf{V}_Y = \mathbf{M}\mathbf{V}_X\mathbf{M}^T\,, \tag{6.25}$$

where $M_{ij} = \left.\frac{\partial Y_i}{\partial X_j}\right|_{x_j}$. The elements of $\mathbf{V}_Y$ are given by

$$V_{Y_{kl}} = \sum_{ij} \left.\frac{\partial Y_k}{\partial X_i}\right|_{x_i} \left.\frac{\partial Y_l}{\partial X_j}\right|_{x_j} V_{X_{ij}}\,. \tag{6.26}$$

In this case we get

$$\begin{aligned}
\sigma^2(Y_i) &= \sigma_i^2 + \sigma_c^2, && (6.27)\\
\mathrm{Cov}(Y_i, Y_j) &= \sigma_c^2 \qquad\qquad (i \neq j), && (6.28)\\
\rho_{ij} &= \frac{\sigma_c^2}{\sqrt{\sigma_i^2+\sigma_c^2}\sqrt{\sigma_j^2+\sigma_c^2}} && (6.29)\\
&= \frac{1}{\sqrt{1+\left(\frac{\sigma_i}{\sigma_c}\right)^2}\sqrt{1+\left(\frac{\sigma_j}{\sigma_c}\right)^2}}\,, && (6.30)
\end{aligned}$$

reobtaining the results of Section 5.6.3. The total uncertainty on the single measurement is given by the combination in quadrature of the individual and the common standard uncertainties, and all the covariances are equal to σ_c^2. To verify, in a simple case, that the result is reasonable, let us consider only two independent quantities X_1 and X_2, and a calibration constant $X_3 = c$, having an expected value equal to zero. From these we can calculate the correlated quantities Y_1 and Y_2 and finally their sum ($S \equiv Z_1$) and difference ($D \equiv Z_2$). The results are

$$\mathbf{V}_Y = \begin{pmatrix} \sigma_1^2+\sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2+\sigma_c^2 \end{pmatrix}, \tag{6.31}$$

$$\mathbf{V}_Z = \begin{pmatrix} \sigma_1^2+\sigma_2^2+4\,\sigma_c^2 & \sigma_1^2-\sigma_2^2 \\ \sigma_1^2-\sigma_2^2 & \sigma_1^2+\sigma_2^2 \end{pmatrix}. \tag{6.32}$$

It follows that

$$\begin{aligned}
\sigma^2(S) &= \sigma_1^2+\sigma_2^2+(2\,\sigma_c)^2, && (6.33)\\
\sigma^2(D) &= \sigma_1^2+\sigma_2^2\,, && (6.34)
\end{aligned}$$

as intuitively expected.

Normalization uncertainty

Let us consider now the case where the calibration constant is the scale factor f, known with a standard uncertainty σ_f. Also in this case, for simplicity and without losing generality, let us suppose that the most probable value of f is 1. Then $X_{n+1} = f$, i.e. $x_{n+1} = 1$, and $V_{X_{n+1,n+1}} = \sigma_f^2$. Then

$$\sigma^2(Y_i) = \sigma_i^2 + \sigma_f^2\, x_i^2\,, \tag{6.35}$$

$$\mathrm{Cov}(Y_i, Y_j) = \sigma_f^2\, x_i\, x_j \qquad (i \neq j)\,, \tag{6.36}$$

$$\rho_{ij} = \frac{x_i\, x_j}{\sqrt{x_i^2 + \frac{\sigma_i^2}{\sigma_f^2}}\sqrt{x_j^2 + \frac{\sigma_j^2}{\sigma_f^2}}}\,, \tag{6.37}$$

$$|\rho_{ij}| = \frac{1}{\sqrt{1 + \left(\frac{\sigma_i}{\sigma_f\, x_i}\right)^2}\sqrt{1 + \left(\frac{\sigma_j}{\sigma_f\, x_j}\right)^2}}\,. \tag{6.38}$$

To verify the results let us consider two independent measurements X_1 and X_2; let us calculate the correlated quantities Y_1 and Y_2, and finally their product ($P \equiv Z_1$) and their ratio ($R \equiv Z_2$):

$$\mathbf{V}_Y = \begin{pmatrix} \sigma_1^2 + \sigma_f^2\, x_1^2 & \sigma_f^2\, x_1\, x_2 \\ \sigma_f^2\, x_1\, x_2 & \sigma_2^2 + \sigma_f^2\, x_2 \end{pmatrix}, \tag{6.39}$$

$$\mathbf{V}_Z = \begin{pmatrix} \sigma_1^2\, x_2^2 + \sigma_2^2\, x_1^2 + 4\,\sigma_f^2\, x_1^2\, x_2^2 & \sigma_1^2 - \sigma_2^2\, \frac{x_1^2}{x_2^2} \\ \sigma_1^2 - \sigma_2^2\, \frac{x_1^2}{x_2^2} & \frac{\sigma_1^2}{x_2^2} + \sigma_2^2\, \frac{x_1^2}{x_2^4} \end{pmatrix}. \tag{6.40}$$

It follows that

$$\sigma^2(P) = \sigma_1^2\, x_2^2 + \sigma_2^2\, x_1^2 + (2\,\sigma_f\, x_1\, x_2)^2\,, \tag{6.41}$$

$$\sigma^2(R) = \frac{\sigma_1^2}{x_2^2} + \sigma_2^2\, \frac{x_1^2}{x_2^4}\,. \tag{6.42}$$

Just as an unknown common offset error cancels in differences and is enhanced in sums, an unknown normalization error has a similar effect on the ratio and the product. It is also interesting to calculate the standard uncertainty of a difference in the case of a normalization error:

$$\sigma^2(D) = \sigma_1^2 + \sigma_2^2 + \sigma_f^2\, (x_1 - x_2)^2\,. \tag{6.43}$$

The contribution from an unknown normalization error vanishes if the two values are equal.

General case

Let us assume there are n independently measured values x_i and m calibration constants c_j with their covariance matrix $\mathbf{V}_c$. The latter can also be theoretical parameters influencing the data, and moreover they may be correlated, as usually happens if, for example, they are parameters of a calibration fit. We can then include the c_j in the vector that contains the measurements and $\mathbf{V}_c$ in the covariance matrix $\mathbf{V}_X$:

$$\underline{x} = \left(\begin{array}{c} x_1 \\ \vdots \\ x_n \\ c_1 \\ \vdots \\ c_m \end{array} \right), \qquad \mathbf{V}_X = \left(\begin{array}{cccc|c} \sigma_1^2 & 0 & \cdots & 0 & \\ 0 & \sigma_2^2 & \cdots & 0 & \\ \cdots & \cdots & \cdots & \cdots & \mathbf{0} \\ 0 & 0 & \cdots & \sigma_n^2 & \\ \hline & & \mathbf{0} & & \mathbf{V}_c \end{array} \right). \tag{6.44}$$

The corrected quantities are obtained from the most general function

$$Y_i = Y_i(X_i, \underline{c}) \qquad\qquad (i = 1, 2, \dots, n)\,, \tag{6.45}$$

and the covariance matrix $\mathbf{V}_Y$ from the covariance propagation $\mathbf{V}_Y = \mathbf{M}\mathbf{V}_X\mathbf{M}^T$.

As a frequently encountered example, we can think of several normalization constants, each affecting a subsample of the data – as is the case where each of several detectors measures a set of physical quantities. Let us consider just three quantities (X_i) and three uncorrelated normalization standard uncertainties (σ_{f_j}), the first common to X_1 and X_2, the second to X_2 and X_3 and the third to all three. We get the following covariance matrix:

$$\left(\begin{array}{ccc} \sigma_1^2 + \left(\sigma_{f_1}^2 + \sigma_{f_3}^2\right) x_1^2 & \left(\sigma_{f_1}^2 + \sigma_{f_3}^2\right) x_1\, x_2 & \sigma_{f_3}^2\, x_1\, x_3 \\ \left(\sigma_{f_1}^2 + \sigma_{f_3}^2\right) x_1\, x_2 & \sigma_2^2 + \left(\sigma_{f_1}^2 + \sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_2^2 & \left(\sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_2\, x_3 \\ \sigma_{f_3}^2\, x_1\, x_3 & \left(\sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_2\, x_3 & \sigma_3^2 + \left(\sigma_{f_2}^2 + \sigma_{f_3}^2\right) x_3^2 \end{array} \right). \tag{6.46}$$

6.3.2 Use and misuse of the covariance matrix to fit correlated data

Best estimate of the true value from two correlated values.

Once the covariance matrix is built one uses it in a χ^2 fit to get the parameters of a function. The quantity to be minimized is χ^2, defined as

$$\chi^2 = \underline{\Delta}^T\, \mathbf{V}^{-1}\, \underline{\Delta}\,, \tag{6.47}$$

where $\underline{\Delta}$ is the vector of the differences between the theoretical and the experimental values. Let us consider the simple case in which two results of the same physical quantity are available, and the individual and the common standard uncertainty are known. The best estimate of the true value of the physical quantity is then obtained by fitting the constant $Y = k$ through the data points. In this simple case the χ^2 minimization can be performed easily. We will consider the two cases of offset and normalization uncertainty. As before, we assume that the detector is well calibrated, i.e. the most probable value of the calibration constant is, respectively for the two cases, 0 and 1, and hence $y_i = x_i$.

Offset uncertainty

Let $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ be the two measured values, and σ_c the common standard uncertainty:

$$\begin{aligned} \chi^2 \;&=\; \frac{1}{D}\,\left[(x_1 - k)^2\,(\sigma_2^2 + \sigma_c^2) + (x_2 - k)^2\,(\sigma_1^2 + \sigma_c^2) \right. \\ &\qquad \left. -2\,(x_1 - k)\,(x_2 - k)\,\sigma_c^2\right]\,, \end{aligned} \tag{6.48}$$

where $D = \sigma_1^2\,\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)\,\sigma_c^2$ is the determinant of the covariance matrix.

Minimizing χ^2 and using the second derivative calculated at the minimum we obtain the best value of k and its standard deviation:

$$\widehat{k} = \frac{x_1\,\sigma_2^2 + x_2\,\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad (=\overline{x}), \tag{6.49}$$

$$\sigma^2(\widehat{k}) = \frac{\sigma_1^2\,\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \sigma_c^2\,. \tag{6.50}$$

The most probable value of the physical quantity is exactly that which one obtains from the average $\overline{x}$ weighted with the inverse of the individual variances. Its overall uncertainty is the quadratic sum of the standard deviation of the weighted average and the common one. The result coincides with the simple expectation.

Normalization uncertainty

Let $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ be the two measured values, and σ_f the common standard uncertainty on the scale:

$$\begin{aligned}\chi^2 = \frac{1}{D}\,\big[&(x_1-k)^2\,(\sigma_2^2 + x_2^2\,\sigma_f^2) + (x_2-k)^2\,(\sigma_1^2 + x_1^2\,\sigma_f^2)\\ &-2\cdot(x_1-k)\cdot(x_2-k)\cdot x_1\cdot x_2\cdot\sigma_f^2\big]\,,\end{aligned} \tag{6.51}$$

where $D = \sigma_1^2\,\sigma_2^2 + (x_1^2\,\sigma_2^2 + x_2^2\,\sigma_1^2)\,\sigma_f^2$. We obtain in this case the following result:

$$\widehat{k} = \frac{x_1\,\sigma_2^2 + x_2\,\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1-x_2)^2\,\sigma_f^2}, \tag{6.52}$$

$$\sigma^2(\widehat{k}) = \frac{\sigma_1^2\,\sigma_2^2 + (x_1^2\,\sigma_2^2 + x_2^2\,\sigma_1^2)\,\sigma_f^2}{\sigma_1^2 + \sigma_2^2 + (x_1-x_2)^2\,\sigma_f^2}\,. \tag{6.53}$$

With respect to the previous case, $\widehat{k}$ has a new term $(x_1-x_2)^2\,\sigma_f^2$ in the denominator. As long as this is negligible with respect to the individual variances we still get the weighted average $\overline{x}$, otherwise a smaller value is obtained. Calling r the ratio between $\widehat{k}$ and $\overline{x}$, we obtain

$$r = \frac{\widehat{k}}{\overline{x}} = \frac{1}{1 + \frac{(x_1-x_2)^2}{\sigma_1^2+\sigma_2^2}\,\sigma_f^2}\,. \tag{6.54}$$

Written in this way, one can see that the deviation from the simple average value depends on the compatibility of the two values and on the normalization uncertainty. This can be understood in the following way: as soon as the two values are in some disagreement, the fit starts to vary the normalization factor (in a hidden way) and to squeeze the scale by an amount allowed by σ_f, in order to minimize the χ^2. The reason the fit prefers normalization factors smaller than 1 under these conditions lies in the standard formalism of the covariance propagation, where only first derivatives are considered. This implies that the individual standard deviations are not rescaled by lowering the normalization factor, but the points get closer.

Example 1. Consider the results of two measurements, 8.0 and 8.5, having 2% individual and 10% common normalization uncertainty. Assuming that the two measurements refer to the same physical quantity, the best estimate of its true value can be obtained by fitting the points to a constant. Minimizing χ^2 with $\mathbf{V}$ estimated empirically by the data, as explained in the previous section, one obtains a value of 7.87 ± 0.81, which is surprising to say the least, since the most probable result is outside the interval determined by the two measured values.

Example 2. A real life case of this strange effect which occurred during the global analysis of the R ratio in e^+e^- performed by The CELLO Collaboration [48], is shown in Fig. 6.1. The data points represent the averages in energy bins of the results of the PETRA and PEP experiments. They are all correlated and the bars show the total uncertainty (see Refs. [48] and [49] for details). In particular, at the intermediate stage of the analysis shown in the figure, an overall 1% systematic error due theoretical uncertainties was included in the covariance matrix. The R values above 36 GeV show the first hint of the rise of the e^+e^- cross-section due to the Z° pole. At that time it was very interesting to prove that the observation was not just a statistical fluctuation. In order to test this, the R measurements were fitted with a theoretical function having no Z° contributions, using only data below a certain energy. It was expected that a fast increase of χ^2 per number of degrees of freedom ν would be observed above 36 GeV, indicating that a theoretical prediction without Z° would be inadequate for describing the high-energy data. The surprising result was a repulsion (see Fig. 6.1) between the experimental data and the fit: Including the high-energy points with larger R a lower curve was obtained, while χ^2/ν remained almost constant.

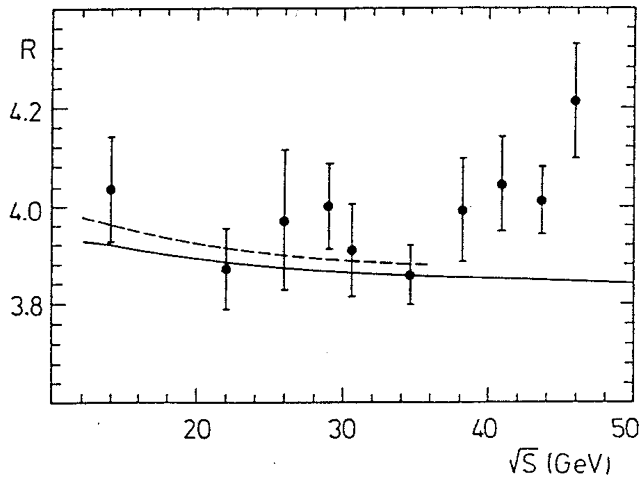

Figure 6.1: R measurements from PETRA and PEP experiments with the best fits of QED+QCD to all the data (full line) and only below 36 GeV (dashed line). All data points are correlated (see text).

To see the source of this effect more explicitly let us consider an alternative way often used to take the normalization uncertainty into account. A scale factor f, by which all data points are multiplied, is introduced to the expression of the χ^2:

$$\chi^2_A = \frac{(f\,x_1 - k)^2}{(f\,\sigma_1)^2} + \frac{(f\,x_2 - k)^2}{(f\,\sigma_2)^2} + \frac{(f-1)^2}{\sigma_f^2}\,. \tag{6.55}$$

Let us also consider the same expression when the individual standard deviations are not rescaled:

$$\chi^2_B = \frac{(f\,x_1 - k)^2}{\sigma_1^2} + \frac{(f\,x_2 - k)^2}{\sigma_2^2} + \frac{(f-1)^2}{\sigma_f^2}\,. \tag{6.56}$$

The use of χ^2_A always gives the result $\widehat{k} = \overline{x}$, because the term $(f-1)^2/\sigma_f^2$ is harmless[9] as far as the value of the minimum χ^2 and the determination on $\widehat{k}$ are concerned. Its only influence is on $\sigma(\widehat{k})$, which turns out to be equal to quadratic combination of the weighted average standard deviation with $\sigma_f\,\overline{x}$, the normalization uncertainty on the average. This result corresponds to the usual one when the normalization factor in the definition of χ^2 is not included, and the overall uncertainty is added at the end.

Instead, the use of χ^2_B is equivalent to the covariance matrix: The same values of the minimum χ^2, of $\widehat{k}$ and of $\sigma(\widehat{k})$ are obtained, and $\widehat{f}$ at the minimum turns out to be exactly the r ratio defined above. This demonstrates that the effect happens when the data values are rescaled independently of their standard uncertainties. The effect can become huge if the data show mutual disagreement. The equality of the results obtained with χ^2_B with those obtained with the covariance matrix allows us to study, in a simpler way, the behaviour of r $(= \widehat{f})$ when an arbitrary number of data points are analysed. The fitted value of the normalization factor is

$$\widehat{f} = \frac{1}{1 + \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_i^2}\,\sigma_f^2}\,. \tag{6.57}$$

If the values of x_i are consistent with a common true value it can be shown that the expected value of $\widehat{f}$ is

$$\langle \widehat{f} \rangle = \frac{1}{1 + (n-1)\,\sigma_f^2}\,. \tag{6.58}$$

Hence, there is a bias on the result when for a non-vanishing σ_f a large number of data points are fitted. In particular, the fit on average produces a bias larger than the normalization uncertainty itself if $\sigma_f > 1/(n-1)$. One can also see that $\sigma^2(\widehat{k})$ and the minimum of χ^2 obtained with the covariance matrix or with χ^2_B are smaller by the same factor r than those obtained with χ^2_A.

Peelle's Pertinent Puzzle

To summarize, when there is an overall uncertainty due to an unknown systematic error and the covariance matrix is used to define χ^2, the behaviour of the fit depends on whether the uncertainty is on the offset or on the scale. In the first case the best estimates of the function parameters are exactly those obtained without overall uncertainty, and only the parameters' standard deviations are affected. In the case of unknown normalization errors, biased results can be obtained. The size of the bias depends on the fitted function, on the magnitude of the overall uncertainty and on the number of data points.

It has also been shown that this bias comes from the linearization performed in the usual covariance propagation. This means that, even though the use of the covariance matrix can be

[9] This can be seen by rewriting (6.55) as

$$\frac{(x_1 - k/f)^2}{\sigma_1^2} + \frac{(x_2 - k/f)^2}{\sigma_2^2} + \frac{(f-1)^2}{\sigma_f^2}\,.$$

For any f, the first two terms determine the value of k, and the third one binds f to 1.

very useful in analysing the data in a compact way using available computer algorithms, care is required if there is one large normalization uncertainty which affects all the data.

The effect discussed above has also been observed independently by R.W. Peelle and reported the year after the analysis of the CELLO data [48]. The problem has been extensively discussed among the community of nuclear physicists, where it is currently known as 'Peelle's Pertinent Puzzle' [50].

Recent cases in HEP in which this effect has been found to have biased the result are discussed in Refs. [51, 52].

Note added: the solution outlined here is taken from Ref. [47], and it has to be considered an *ad hoc* solution. The general (of course Bayesian) solution to the χ^2 paradox has been worked out recently [53], and it will be published in a forthcoming paper.

Chapter 7

Bayesian unfolding

"Now we see but a poor reflection as in a mirror ... "
"Now I know in part ... "
(1 Cor.)

7.1 Problem and typical solutions

In any experiment the distribution of the measured observables differs from that of the corresponding true physical quantities due to physics and detector effects. For example, one may be interested in measuring the variables x and Q^2 in deep-inelastic scattering events. In such a case one is able to build statistical estimators which in principle have a physical meaning similar to the true quantities, but which have a non-vanishing variance and are also distorted due to QED and QCD radiative corrections, parton fragmentation, particle decay and limited detector performances. The aim of the experimentalist is to unfold the observed distribution from all these distortions so as to extract the true distribution (see also Refs. [54] and [55]). This requires a satisfactory knowledge of the overall effect of the distortions on the true physical quantity.

When dealing with only one physical variable the usual method for handling this problem is the so-called 'bin-to-bin' correction: one evaluates a generalized efficiency (it may even be larger than unity) by calculating the ratio between the number of events falling in a certain bin of the reconstructed variable and the number of events in the same bin of the true variable with a Monte Carlo simulation. This efficiency is then used to estimate the number of true events from the number of events observed in that bin. Clearly this method requires the same subdivision in bins of the true and the experimental variable and hence it cannot take into account large migrations of events from one bin to the others. Moreover it neglects the unavoidable correlations between adjacent bins. This approximation is valid only if the amount of migration is negligible and if the standard deviation of the smearing is smaller than the bin size.

An attempt to solve the problem of migrations is sometimes made by building a matrix which connects the number of events generated in one bin to the number of events observed in the other bins. This matrix is then inverted and applied to the measured distribution. This immediately produces inversion problems if the matrix is singular. On the other hand, there is no reason from a probabilistic point of view why the inverse matrix should exist. This can easily be seen by taking the example of two bins of the true quantity both of which have the same probability of being observed in each of the bins of the measured quantity. It follows that treating probability distributions as vectors in space is not correct, even in principle. Moreover the method is not able to handle large statistical fluctuations even if the matrix can be inverted (if we have, for example, a very large number of events with which to estimate its elements and we choose the binning in such a way as to make the matrix not singular). The easiest way to see this is to think of the unavoidable negative terms of the inverse of the matrix which in some extreme cases may yield negative numbers of unfolded events. Quite apart from these theoretical reservations, the actual experience of those who have used this method is rather discouraging, the results being highly unstable.

7.2 Bayes' theorem stated in terms of causes and effects

Let us state Bayes' theorem in terms of several independent causes ($C_i,\ i = 1, 2, \dots, n_C$) which can produce one effect (E). For example, if we consider deep-inelastic scattering events, the effect E can be the observation of an event in a cell of the measured quantities $\{\Delta Q^2_{meas}, \Delta x_{meas}\}$. The causes C_i are then all the possible cells of the true values $\{\Delta Q^2_{true}, \Delta x_{true}\}_i$. Let us assume we know the initial probability of the causes $P(C_i)$ and the conditional probability that the i-th cause will produce the effect $P(E \,|\, C_i)$. The Bayes formula is then

$$P(C_i \,|\, E) = \frac{P(E \,|\, C_i)\, P(C_i)}{\sum_{l=1}^{n_C} P(E \,|\, C_l)\, P(C_l)}\,. \tag{7.1}$$

$P(C_i \,|\, E)$ depends on the initial probability of the causes. If one has no better prejudice concerning $P(C_i)$ the process of inference can be started from a uniform distribution.

The final distribution depends also on $P(E \,|\, C_i)$. These probabilities must be calculated or estimated with Monte Carlo methods. One has to keep in mind that, in contrast to $P(C_i)$, these probabilities are not updated by the observations. So if there are ambiguities concerning the choice of $P(E \,|\, C_i)$ one has to try them all in order to evaluate their systematic effects on the results.

7.3 Unfolding an experimental distribution

If one observes $n(E)$ events with effect E, the expected number of events assignable to each of the causes is

$$\widehat{n}(C_i) = n(E)\, P(C_i \,|\, E)\,. \tag{7.2}$$

As the outcome of a measurement one has several possible effects E_j ($j = 1, 2, \dots, n_E$) for a given cause C_i. For each of them the Bayes formula (7.1) holds, and $P(C_i \,|\, E_j)$ can be evaluated. Let us write (7.1) again in the case of n_E possible effects,[1] indicating the initial probability of the causes with $P_\circ(C_i)$:

$$P(C_i \,|\, E_j) = \frac{P(E_j \,|\, C_i)\, P_\circ(C_i)}{\sum_{l=1}^{n_C} P(E_j \,|\, C_l)\, P_\circ(C_l)}\,. \tag{7.3}$$

One should note the following.

- $\sum_{i=1}^{n_C} P_\circ(C_i) = 1$, as usual. Note that if the probability of a cause is initially set to zero it can never change, i.e. if a cause does not exist it cannot be invented.

[1] The broadening of the distribution due to the smearing suggests a choice of n_E larger than n_C. It is worth mentioning that there is no need to reject events where a measured quantity has a value outside the range allowed for the physical quantity. For example, in the case of deep-inelastic scattering events, cells with $x_{meas} > 1$ or $Q^2_{meas} < 0$ give information about the true distribution too.

- $\sum_{i=1}^{n_C} P(C_i \,|\, E_j) = 1$. This normalization condition, mathematically trivial since it comes directly from (7.3), indicates that each effect must come from one or more of the causes under examination. This means that if the observables also contain a non-negligible amount of background, this needs to be included among the causes.
- $0 \le \epsilon_i \equiv \sum_{j=1}^{n_E} P(E_j \,|\, C_i) \le 1$. There is no need for each cause to produce at least one of the effects. ϵ_i gives the efficiency of finding the cause C_i in any of the possible effects.

After N_{obs} experimental observations one obtains a distribution of frequencies $\underline{n}(E) \equiv \{n(E_1), n(E_2), \ldots, n(E_{n_E})\}$. The expected number of events to be assigned to each of the causes (taking into account only the observed events) can be calculated by applying (7.2) to each effect:

$$\widehat{n}(C_i)|_{obs} \;=\; \sum_{j=1}^{n_E} n(E_j)\, P(C_i \,|\, E_j)\,. \tag{7.4}$$

When inefficiency[2] is also brought into the picture, the best estimate of the true number of events becomes

$$\widehat{n}(C_i) \;=\; \frac{1}{\epsilon_i}\sum_{j=1}^{n_E} n(E_j)\, P(C_i \,|\, E_j) \qquad \epsilon_i \neq 0\,. \tag{7.5}$$

From these unfolded events we can estimate the true total number of events, the final probabilities of the causes and the overall efficiency:

$$\begin{aligned} \widehat{N}_{true} &= \sum_{i=1}^{n_C} \widehat{n}(C_i)\,, \\ \widehat{P}(C_i) \equiv P(C_i \,|\, \underline{n}(E)) &= \frac{\widehat{n}(C_i)}{\widehat{N}_{true}}\,, \\ \widehat{\epsilon} &= \frac{N_{obs}}{\widehat{N}_{true}}\,. \end{aligned}$$

If the initial distribution $\underline{P_\circ}(C)$ is not consistent with the data, it will not agree with the final distribution $\underline{\widehat{P}}(C)$. The closer the initial distribution is to the true distribution, the better the agreement is. For simulated data one can easily verify that the distribution $\underline{\widehat{P}}(C)$ lies between $\underline{P_\circ}(C)$ and the true one. This suggests proceeding iteratively. Figure 7.1 shows an example of a bidimensional distribution unfolding.

More details about iteration strategy, evaluation of uncertainty, etc. can be found in Ref. [56]. I would just like to comment on an obvious criticism that may be made: ‘the iterative procedure is against the Bayesian spirit, since the same data are used many times for the same inference’. In principle the objection is valid, but in practice this technique is a trick to give to the experimental data a weight (an importance) larger than that of the priors. A more rigorous procedure which took into account uncertainties and correlations of the initial distribution would have been much more complicated. An attempt of this kind can be found in Ref. [57]. Examples of unfolding procedures performed with non-Bayesian methods are described in Refs. [54] and [55].

Note added: A recent book by Cowan [58] contains an interesting chapter on unfolding. More sophisticated methods for, generally speaking, image reconstruction can be found in Ref. [59] and references therein.

[2] If $\epsilon_i = 0$ then $\widehat{n}(C_i)$ will be set to zero, since the experiment is not sensitive to the cause C_i.

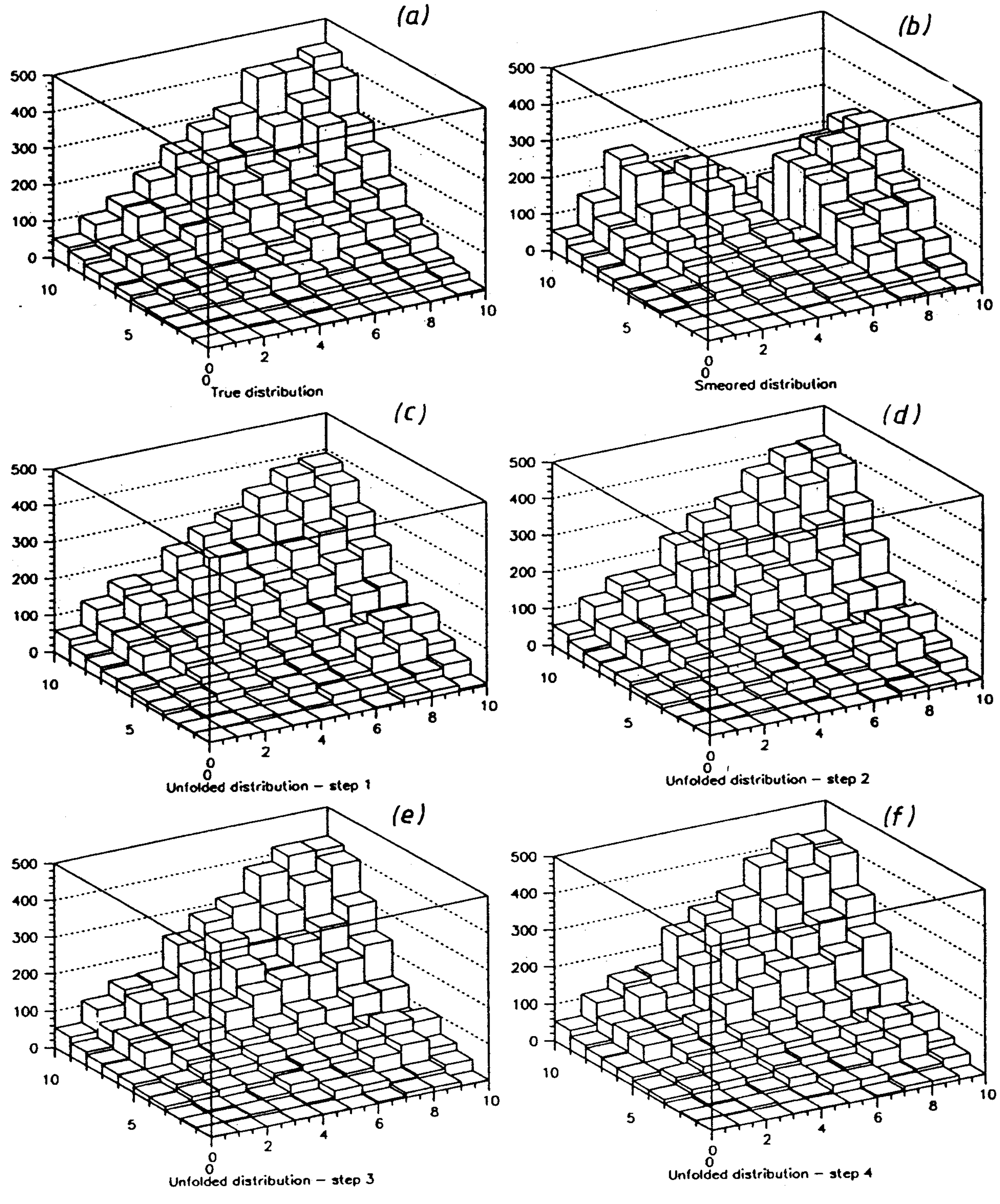

Figure 7.1: Example of two-dimensional unfolding: true distribution (a), smeared distribution (b) and results after the first four steps [(c) to (f)].