Part I

Subjective probability in physics? Scientific reasoning in conditions of uncertainty

Chapter 1

Uncertainty in physics and the usual methods of handling it

"In almost all circumstances, and at all times,
we find ourselves in a state of uncertainty.
Uncertainty in every sense.
Uncertainty about actual situations, past and present . . .
Uncertainty in foresight: this would not be eliminated
or diminished even if we accepted, in its most absolute form,
the principle of determinism; in any case, this is no longer in fashion.
Uncertainty in the face of decisions: more than ever in this case . . .
Even in the field of tautology (i.e of what is true or false by mere
definition, independently of any contingent circumstances) we always
find ourselves in a state of uncertainty . . . (for instance,
of what is the seventh, or billionth, decimal place of π . . .) . . . "
(Bruno de Finetti)

1.1 Uncertainty in physics

It is fairly well accepted among physicists that any conclusion which results from a measurement is affected by a certain degree of uncertainty. Let us remember briefly the reasons which prevent us from reaching certain statements. Figure 1.1 sketches the activity of physicists (or of any other

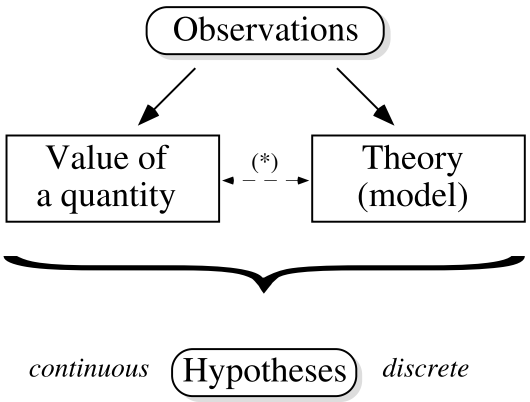

Figure 1.1: From observations to hypotheses. The link between value of a quantity and theory is a reminder that sometimes a physics quantity has meaning only within a given theory or model.

scientist). From experimental data one wishes to determine the value of a given quantity, or to establish which theory describes the observed phenomena better. Although they are often seen as separate, both tasks may be viewed as two sides of the same process: going from observations to hypotheses. In fact, they can be stated in the following terms.

A: Which values are (more) compatible with the definition of the measurand, under the condition that certain numbers have been observed on instruments (and subordinated to all the available knowledge about the instrument and the measurand)?

B: Which theory is (more) compatible with the observed phenomena (and subordinated to the credibility of the theory, based also on aesthetics and simplicity arguments)?

The only difference between the two processes is that in the first the number of hypotheses is virtually infinite (the quantities are usually supposed to assume continuous values), while in the second it is discrete and usually small.

The reasons why it is impossible to reach the ideal condition of certain knowledge, i.e. only one of the many hypotheses is considered to be true and all the others false, may be summarized in the following, well-understood, scheme.

A: As far as the determination of the value of a quantity is concerned, one says that *"uncertainty is due to measurement errors"*.

B: In the case of a theory, we can distinguish two subcases:

(B_1) The law is probabilistic, i.e. the observations are not just a logical consequence of the theory. For example, tossing a regular coin, the two sequences of heads and tails

```
hhhhhhhhhhhhhhhhhhhhhhhhh
```

```
hhttttthhhhthhtthhhththht
```

have the same probability of being observed (as any other sequence). Hence, there is no way of reaching a firm conclusion about the regularity of the coin after an observed sequence of any particular length.[1]

(B_2) The law is deterministic. But this property is only valid in principle, as can easily be understood. In fact, in all cases the actual observations also depend on many other factors external to the theory, such as initial and boundary conditions, influence factors, experimental errors, etc. All unavoidable uncertainties on these factors mean that the link between theory and observables is of a probabilistic nature in this case too.

1.2 True value, error and uncertainty

Let us start with case **A**. A first objection would be *"What does it mean that uncertainties are due to errors? Isn't this just tautology?"*. Well, the nouns 'error' and 'uncertainty', although currently used almost as synonyms, are related to different concepts. This is a first hint that in this subject there is neither uniformity of language, nor of methods. For this reason the metrological organizations have recently made great efforts to bring some order into the field [1, 2, 3, 4, 5].

[1] But after observation of the first sequence one would strongly suspect that the coin had two heads, if one had no means of directly checking the coin. The concept of probability will be used, in fact, to quantify the degree of such suspicion.

In particular, the International Organization for Standardization (ISO) has published a *"Guide to the expression of uncertainty in measurement"* [3], containing definitions, recommendations and practical examples. Consulting the 'ISO Guide' we find the following definitions.

- Uncertainty: *"a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurement."*
- Error: *"the result of a measurement minus a true value of the measurand."*

One has to note the following.

- The ISO definition of uncertainty defines the concept; as far as the operative definition is concerned, they recommend the 'standard uncertainty', i.e. the standard deviation (σ) of the possible values that the measurand may assume (each value is weighted with its 'degree of belief' in a way that will become clear later).
- It is clear that the error is usually unknown, as follows from the definition.
- The use of the article 'a' (instead of 'the') when referring to 'true value' is intentional, and rather subtle.

Also the ISO definition of true value differs from that of standard textbooks. One finds, in fact:

- true value: *"a value compatible with the definition of a given particular quantity."*

This definition may seem vague, but it is more practical and pragmatic, and of more general use, than *"the value obtained after an infinite series of measurements performed under the same conditions with an instrument not affected by systematic errors."* For instance, it holds also for quantities for which it is not easy to repeat the measurements, and even for those cases in which it makes no sense to speak about repeated measurements under the same conditions. The use of the indefinite article in conjunction with true value can be understood by considering the first item on the list in the next section.

1.3 Sources of measurement uncertainty

It is worth reporting the sources of uncertainty in measurement as listed by the ISO Guide:

"1 incomplete definition of the measurand;

2 imperfect realization of the definition of the measurand;

3 non-representative sampling – the sample measured may not represent the defined measurand;

4 inadequate knowledge of the effects of environmental conditions on the measurement, or imperfect measurement of environmental conditions;

5 personal bias in reading analogue instruments;

6 finite instrument resolution or discrimination threshold;

7 inexact values of measurement standards and reference materials;

8 inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm;

9 approximations and assumptions incorporated in the measurement method and procedure;

10 variations in repeated observations of the measurand under apparently identical conditions."

These do not need to be commented upon. Let us just give examples of the first two sources.

1. If one has to measure the gravitational acceleration g at sea level, without specifying the precise location on the earth's surface, there will be a source of uncertainty because many different — even though 'intrinsically very precise' — results are consistent with the definition of the measurand.[2]

2. The magnetic moment of a neutron is, in contrast, an unambiguous definition, but there is the experimental problem of performing experiments on isolated neutrons.

In terms of the usual jargon, one may say that sources 1–9 are related to systematic effects and 10 to 'statistical effects'. Some caution is necessary regarding the sharp separation of the sources, which is clearly somehow artificial. In particular, all sources 1–9 may contribute to 10, because they each depend upon the precise meaning of the clause *"under apparently identical conditions"* (one should talk, more precisely, about 'repeatability conditions' [3]). In other words, if the various effects change during the time of measurement, without any possibility of monitoring them, they contribute to the random error.

1.4 Usual handling of measurement uncertainties

The present situation concerning the treatment of measurement uncertainties can be summarized as follows.

- Uncertainties due to statistical errors are currently treated using the frequentistic concept of 'confidence interval', although
  - there are well known cases — of great relevance in frontier physics — in which the approach is not applicable (e.g. small number of observed events, or measurement close to the edge of the physical region);
  - the procedure is rather unnatural, and in fact the interpretation of the results is unconsciously subjective (as will be discussed later).

- There is no satisfactory theory or model to treat uncertainties due to systematic errors[3] consistently. Only *ad hoc* prescriptions can be found in the literature and in practice (*"my supervisor says ... "*): *"add them linearly"; "add them linearly if ... , else add them quadratically"; "don't add them at all"*.[4] The fashion at the moment is to add them quadratically if they are considered to be independent, or to build a covariance matrix of

[2] It is then clear that the definition of true value implying an indefinite series of measurements with ideal instrumentation gives the illusion that the true value is unique. The ISO definition, instead, takes into account the fact that measurements are performed under real conditions and can be accompanied by all the sources of uncertainty in the above list.

[3] To be more precise one should specify 'of unknown size', since an accurately assessed systematic error does not yield uncertainty, but only a correction to the raw result.

[4] By the way, it is a good and recommended practice to provide the complete list of contributions to the overall uncertainty [3]; but it is also clear that, at some stage, the producer or the user of the result has to combine the uncertainty to form his idea about the interval in which the quantity of interest is believed to lie.

statistical and systematic contribution to treat the general case. In my opinion, besides all the theoretically motivated excuses for justifying this praxis, there is simply the reluctance of experimentalists to combine linearly 10, 20 or more contributions to a global uncertainty, as the (out of fashion) 'theory' of maximum bounds would require.[5]

The problem of interpretation will be treated in the next section. For the moment, let us see why the use of standard propagation of uncertainty, namely

$$\sigma^2(Y) = \sum_i \left(\frac{\partial Y}{\partial X_i}\right)^2 \sigma^2(X_i) + \textit{correlation terms}\,, \tag{1.1}$$

is not justified (especially if contributions due to systematic effects are included). This formula is derived from the rules of probability distributions, making use of linearization (a usually reasonable approximation for routine applications). This leads to theoretical and practical problems.

- X_i and Y should have the meaning of random variables.
- In the case of systematic effects, how do we evaluate the input quantities $\sigma(X_i)$ entering in the formula in a way which is consistent with their meaning as standard deviations?
- How do we properly take into account correlations (assuming we have solved the previous questions)?

It is very interesting to go to your favourite textbook and see how 'error propagation' is introduced. You will realize that some formulae are developed for random quantities, making use of linear approximations, and then suddenly they are used for physics quantities without any justification.[6] A typical example is measuring a velocity $v \pm \sigma(v)$ from a distance $s \pm \sigma(s)$ and a time interval $t \pm \sigma(t)$. It is really a challenge to go from the uncertainty on s and t to that of v without considering s, t and v as random variables, and to avoid thinking of the final result as a probabilistic statement on the velocity. Also in this case, an intuitive interpretation conflicts with standard probability theory.

1.5 Probability of observables versus probability of true values

The criticism about the inconsistent interpretation of results may look like a philosophical quibble, but it is, in my opinion, a crucial point which needs to be clarified. Let us consider the example of n independent measurements of the same quantity under identical conditions (with n large enough to simplify the problem, and neglecting systematic effects). We can evaluate the arithmetic average $\overline{x}$ and the standard deviation σ. The result on the true value μ is

$$\mu = \overline{x} \pm \frac{\sigma}{\sqrt{n}}\,. \tag{1.2}$$

[5] And in fact, one can see that when there are only two or three contributions to the 'systematic error', there are still people who prefer to add them linearly.

[6] Some others, including some old lecture notes of mine, try to convince the reader that the propagation is applied to the observables, in a very complicated and artificial way. Then, later, as in the 'game of the three cards' proposed by professional cheaters in the street, one uses the same formulae for physics quantities, hoping that the students do not notice the logical gap.

The reader will have no difficulty in admitting that the large majority of people interpret (1.2) as if it were[7]

$$P(\overline{x} - \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%\,. \tag{1.3}$$

However, conventional statistics says only that[8]

$$P(\mu - \frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + \frac{\sigma}{\sqrt{n}}) = 68\%\,, \tag{1.4}$$

a probabilistic statement about $\overline{X}$, given μ, σ and n. Probabilistic statements concerning μ are not foreseen by the theory (*"μ is a constant of unknown value"*[9]), although this is what we are, intuitively, looking for: Having observed the *effect* $\overline{x}$ we are interested in stating something about the possible true value responsible for it. In fact, when we do an experiment, we want to increase our knowledge about μ and, consciously or not, we want to know which values are more or less probable. A statement concerning the probability that an observed value falls within a certain interval around μ is meaningless if it cannot be turned into an expression which states the quality of the knowledge about μ itself. Since the usual probability theory does not help, the probability inversion is performed intuitively. In routine cases it usually works, but there are cases in which it fails (see Section 1.7).

1.6 Probability of the causes

Generally speaking, what is missing in the usual theory of probability is the crucial concept of probability of hypotheses and, in particular, probability of causes: *"the essential problem of the experimental method"* (Poincaré):

> *"I play at écarté with a gentleman whom I know to be perfectly honest. What is the chance that he turns up the king? It is 1/8. This is a problem of the probability of effects. I play with a gentleman whom I do not know. He has dealt ten times, and he has turned the king up six times. What is the chance that he is a sharper? This is a problem in the probability of causes. It may be said that it is the essential problem of the experimental method"* [6].
> *" ... the laws are known to us by the observed effects. Trying to deduct from the effects the laws which are the causes, it is solving a problem of probability of causes"* [7].

A theory of probability which does not consider probabilities of hypothesis is unnatural and prevents transparent and consistent statements about the causes which may have produced the observed effects from being assessed.

[7] There are also those who express the result, making the trivial mistake of saying *"this means that, if I repeat the experiment a great number of times, then I will find that in roughly 68% of the cases the observed average will be in the interval $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$"*. (Besides the interpretation problem, there is a missing factor of $\sqrt{2}$ in the width of the interval ...)

[8] The capital letter to indicate the average appearing in (1.4) is used because here this symbol stands for a random variable, while in (1.3) it indicated a realization of it. For the Greek symbols this distinction is not made, but the different role should be evident from the context.

[9] It is worth noting the paradoxical inversion of role between μ, about which we are in a state of uncertainty, considered to be a constant, and the observation $\overline{x}$, which has a certain value and which is instead considered a random quantity. This distorted way of thinking produces the statements to which we are used, such as speaking of *"uncertainty (or error) on the observed number"*: If one observes 10 on a scaler, there is no uncertainty on this number, but on the quantity which we try to infer from the observation (e.g. λ of a Poisson distribution, or a rate).

1.7 Unsuitability of confidence intervals

According to the standard theory of probability, statement (1.3) is nonsense, and, in fact, good frequentistic books do not include it. They speak instead about 'confidence intervals', which have a completely different interpretation [that of (1.4)], although several books and many teachers suggest an interpretation of these intervals as if they were probabilistic statements on the true values, like (1.3). But it seems to me that it is practically impossible, even for those who are fully aware of the frequentistic theory, to avoid misleading conclusions. This opinion is well stated by Howson and Urbach in a paper to Nature [8]:

> *"The statement that such-and-such is a 95% confidence interval for μ seems objective. But what does it say? It may be imagined that a 95% confidence interval corresponds to a 0.95 probability that the unknown parameter lies in the confidence range. But in the classical approach, μ is not a random variable, and so has no probability. Nevertheless, statisticians regularly say that one can be '95% confident' that the parameter lies in the confidence interval. They never say why."*

The origin of the problem goes directly to the underlying concept of probability. The frequentistic concept of confidence interval is, in fact, a kind of artificial invention to characterize the uncertainty consistently with the frequency-based definition of probability. But, unfortunately – as a matter of fact – this attempt to classify the state of uncertainty (on the true value) trying to avoid the concept of probability of hypotheses produces misinterpretation. People tend to turn arbitrarily (1.4) into (1.3) with an intuitive reasoning that I like to paraphrase as 'the dog and the hunter': We know that a dog has a 50% probability of being 100 m from the hunter; if we observe the dog, what can we say about the hunter? The terms of the analogy are clear:

$$\begin{aligned} \text{hunter} &\leftrightarrow \text{true value} \\ \text{dog} &\leftrightarrow \text{observable}\,. \end{aligned}$$

The intuitive and reasonable answer is *"The hunter is, with 50% probability, within 100 m of the position of the dog."* But it is easy to understand that this conclusion is based on the tacit assumption that 1) the hunter can be anywhere around the dog; 2) the dog has no preferred direction of arrival at the point where we observe him. Any deviation from this simple scheme invalidates the picture on which the inversion of probability (1.4) $\rightarrow$ (1.3) is based. Let us look at some examples.

Example 1: Measurement at the edge of a physical region.

An experiment, planned to measure the electron-neutrino mass with a resolution of $\sigma = 2\,\mathrm{eV}/c^2$ (independent of the mass, for simplicity, see Fig. 1.2), finds a value of $-4\,\mathrm{eV}/c^2$ (i.e. this value comes out of the analysis of real data treated in exactly the same way as that of simulated data, for which a $2\,\mathrm{eV}/c^2$ resolution was found).
What can we say about m_ν?

$$m_\nu = -4 \pm 2\ \mathrm{eV}/c^2\ \ ?$$

$$P(-6\,\mathrm{eV}/c^2 \le m_\nu \le -2\ \mathrm{eV}/c^2) = 68\%\ \ ?$$

$$P(m_\nu \le 0\ \mathrm{eV}/c^2) = 98\%\ \ ?$$

No physicist would sign a statement which sounded like he was 98% sure of having found a negative mass!

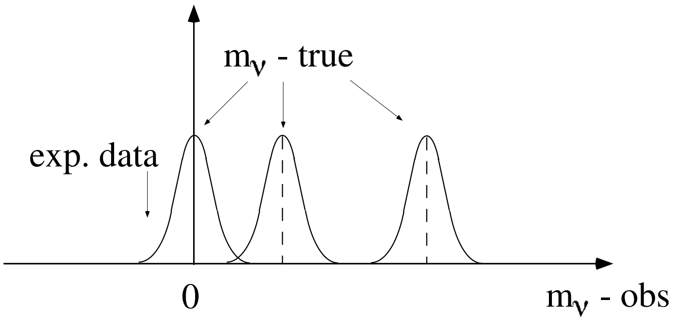

Figure 1.2: Negative neutrino mass?

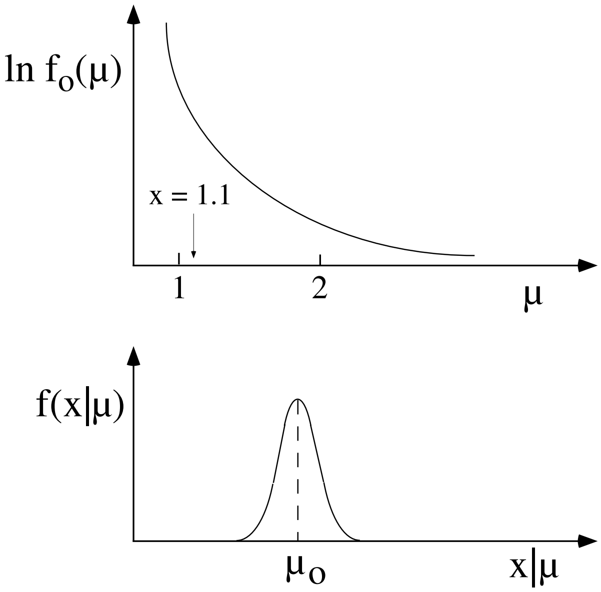

Figure 1.3: Case of highly asymmetric expectation on the physics quantity.

Example 2: Non-flat distribution of a physical quantity.
Let us take a quantity μ that we know, from previous knowledge, to be distributed as in Fig. 1.3. It may be, for example, the energy of bremsstrahlung photons or of cosmic rays. We know that an observable value X will be normally distributed around the true value μ, independently of the value of μ. We have performed a measurement and obtained $x = 1.1$, in arbitrary units. What can we say about the true value μ that has caused this observation? Also in this case the formal definition of the confidence interval does not work. Intuitively, we feel that there is more chance that μ is on the left side of (1.1) than on the right side. In the jargon of the experimentalists, *"there are more migrations from left to right than from right to left"*.

Example 3: High-momentum track in a magnetic spectrometer.
The previous examples deviate from the simple dog–hunter picture only because of an asymmetric possible position of the 'hunter'. The case of a very-high-momentum track in a central detector of a high-energy physics (HEP) experiment involves asymmetric response of a detector for almost straight tracks and non-uniform momentum distribution of charged particles produced in the collisions. Also in this case the simple inversion scheme does not

work.

To sum up the last two sections, we can say that intuitive inversion of probability

$$P(\ldots \leq \overline{X} \leq \ldots) \Longrightarrow P(\ldots \leq \mu \leq \ldots)\,, \tag{1.5}$$

besides being theoretically unjustifiable, yields results which are numerically correct only in the case of symmetric problems.

1.8 Misunderstandings caused by the standard paradigm of hypothesis tests

Similar problems of interpretation appear in the usual methods used to test hypotheses. I will briefly outline the standard procedure and then give some examples to show the kind of paradoxical conclusions that one can reach.

A frequentistic hypothesis test follows the scheme outlined below (see Fig. 1.4). [10]

1. Formulate a hypothesis $H_\circ$.
2. Choose a test variable θ of which the probability density function $f(\theta\,|\,H_\circ)$ is known (analytically or numerically) for a given $H_\circ$.
3. Choose an interval $[\theta_1, \theta_2]$ such that there is high probability that θ falls inside the interval:

$$P(\theta_1 \leq \theta \leq \theta_2) = 1 - \alpha\,, \tag{1.6}$$

   with α typically equal to 1% or 5%.
4. Perform an experiment, obtaining $\theta = \theta_m$.
5. Draw the following conclusions :
   - if $\theta_1 \leq \theta_m \leq \theta_2$ $\Longrightarrow H_\circ$ accepted;
   - otherwise $\Longrightarrow H_\circ$ rejected with a significance level α.

The usual justification for the procedure is that the probability α is so low that it is practically impossible for the test variable to fall outside the interval. Then, if this event happens, we have good reason to reject the hypothesis.

One can recognize behind this reasoning a revised version of the classical 'proof by contradiction' (see, e.g., Ref. [10]). In standard dialectics, one assumes a hypothesis to be true and looks for a logical consequence which is manifestly false in order to reject the hypothesis. The slight difference is that in the hypothesis test scheme, the false consequence is replaced by an improbable one. The argument may look convincing, but it has no grounds. In order to analyse the problem well, we need to review the logic of uncertainty. For the moment a few examples are enough to indicate that there is something troublesome behind the procedure.

[10] At present,'P-values' (or 'significance probabilities') are also *"used in place of hypothesis tests as a means of giving more information about the relationship between the data and the hypothesis than does a simple reject/do not reject decision"* [9]. They consist in giving the probability of the 'tail(s)', as also usually done in HEP, although the name 'P-values' has not yet entered our lexicon. Anyhow, they produce the same interpretation problems of the hypothesis test paradigm (see also example 8 of next section).

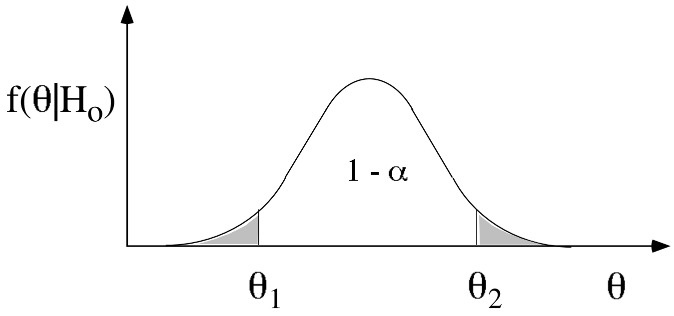

Figure 1.4: Hypothesis test scheme in the frequentistic approach.

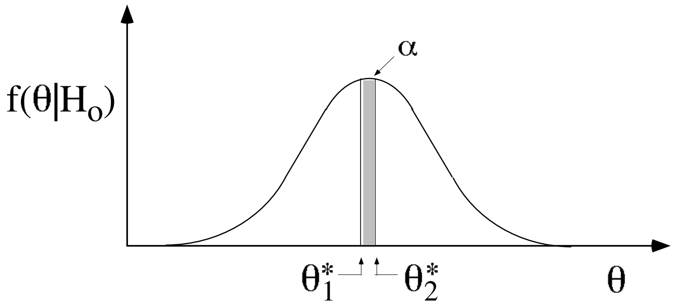

Figure 1.5: Would you accept this scheme to test hypotheses?

Example 4: Choosing the rejection region in the middle of the distribution.
Imagine choosing an interval $[\theta_1^*, \theta_2^*]$ around the expected value of θ (or around the mode) such that

$$P(\theta_1^* \le \theta \le \theta_2^*) = \alpha\,, \tag{1.7}$$

with α small (see Fig. 1.5). We can then reverse the test, and reject the hypothesis if the measured θ_m is inside the interval. This strategy is clearly unacceptable, indicating that the rejection decision cannot be based on the argument of practically impossible observations (smallness of α).

One may object that the reason is not only the small probability of the rejection region, but also its distance from the expected value. Figure 1.6 is an example against this objection. Although the situation is not as extreme as that depicted in Fig. 1.5, one would need a certain amount of courage to say that the $H_\circ$ is rejected if the test variable falls by chance in 'the bad region'.

Example 5: Has the student made a mistake?
A teacher gives to each student an individual sample of 300 random numbers, uniformly distributed between 0 and 1. The students are asked to calculate the arithmetic average. The prevision[11] of the teacher can be quantified with

$$\mathrm{E}\left[\overline{X}_{300}\right] = \frac{1}{2} \tag{1.8}$$

$$\sigma\left[\overline{X}_{300}\right] = \frac{1}{\sqrt{12}} \cdot \frac{1}{\sqrt{300}} = 0.017\,, \tag{1.9}$$

[11] By prevision I mean, following [11], a probabilistic 'prediction', which corresponds to what is usually known as expectation value (see Section 5.2.2).

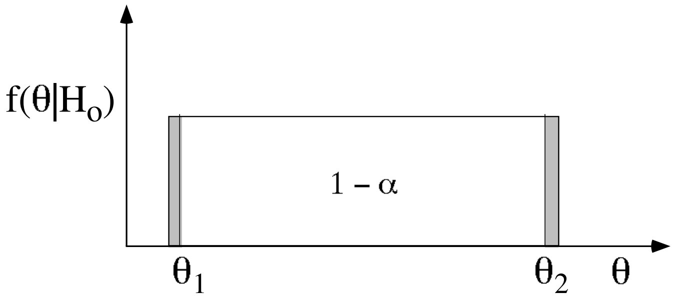

Figure 1.6: Would you accept this scheme to test hypotheses?

with the random variable $\overline{X}_{300}$ normally distributed because of the central limit theorem. This means that there is 99% probability that an average will come out in the interval $0.5 \pm (2.6 \times 0.017)$:

$$P(0.456 \leq \overline{X}_{300} \leq 0.544) = 99\% \,. \tag{1.10}$$

Imagine that a student obtains an average outside the above interval (e.g. $\overline{x} = 0.550$). The teacher may be interested in the probability that the student has made a mistake (for example, he has to decide if it is worthwhile checking the calculation in detail). Applying the standard methods one draws the conclusion that

> *"the hypothesis $H_\circ$ = 'no mistakes' is rejected at the 1% level of significance",*

i.e. one receives a precise answer to a different question. In fact, the meaning of the previous statement is simply

> *"there is only a 1% probability that the average falls outside the selected interval, if the calculations were done correctly".*

But this does not answer our natural question,[12] i.e. that concerning the probability of mistake, and not that of results far from the average if there were no mistakes. Moreover, the statement sounds as if one would be 99% sure that the student has made a mistake! This conclusion is highly misleading.

How is it possible, then, to answer the very question concerning the probability of mistakes? If you ask the students (before they take a standard course in hypothesis tests) you will hear the right answer, and it contains a crucial ingredient extraneous to the logic of hypothesis tests:

> *"It all depends on who has made the calculation!"*

In fact, if the calculation was done by a well-tested program the probability of mistake would be zero. And students know rather well their probability of making mistakes.

Example 6: A bad joke to a journal.[13]

[12] Personally, I find it is somehow impolite to give an answer to a question which is different from that asked. At least one should apologize for being unable to answer the original question. However, textbooks usually do not do this, and people get confused.

[13] Example taken from Ref. [12].

A scientific journal changes its publication policy. The editors announce that results with a significance level of 5% will no longer be accepted. Only those with a level of $\leq 1\%$ will be published. The rationale for the change, explained in an editorial, looks reasonable and it can be shared without hesitation: *"We want to publish only good results."*

1000 experimental physicists, not convinced by this severe rule, conspire against the journal. Each of them formulates a wrong physics hypothesis and performs an experiment to test it according to the accepted/rejected scheme.

Roughly 10 physicists get 1% significant results. Their papers are accepted and published. It follows that, contrary to the wishes of the editors, the first issue of the journal under the new policy contains only wrong results!

The solution to the kind of paradox raised by this example seems clear: The physicists knew with certainty that the hypotheses were wrong. So the example looks like an odd case with no practical importance. But in real life who knows in advance with certainty if a hypothesis is true or false?

1.9 Statistical significance versus probability of hypotheses

The examples in the previous section have shown the typical ways in which significance tests are misinterpreted. This kind of mistake is commonly made not only by students, but also by professional users of statistical methods. There are two different probabilities:

$P(H \mid$ "data"$)$: the probability of the hypothesis H, conditioned by the observed data. This is the probabilistic statement in which we are interested. It summarizes the status of knowledge on H, achieved in conditions of uncertainty: it might be the probability that the W mass is between 80.00 and 80.50 GeV, that the Higgs mass is below 200 GeV, or that a charged track is a π^- rather than a K^-.

$P($"data"$\mid H)$: the probability of the observables under the condition that the hypothesis H is true.[14] For example, the probability of getting two consecutive heads when tossing a regular coin, the probability that a W mass is reconstructed within 1 GeV of the true mass, or that a 2.5 GeV pion produces a $\geq 100\,\text{pC}$ signal in an electromagnetic calorimeter.

Unfortunately, conventional statistics considers only the second case. As a consequence, since the very question of interest remains unanswered, very often significance levels are incorrectly treated as if they were probabilities of the hypothesis. For example, *"H refused at 5% significance"* may be understood to mean the same as *"H has only 5% probability of being true."*

It is important to note the different consequences of the misunderstanding caused by the arbitrary probabilistic interpretation of confidence intervals and of significance levels. Measurement uncertainties on directly measured quantities obtained by confidence intervals are at least numerically correct in most routine cases, although arbitrarily interpreted. In hypothesis tests, however, the conclusions may become seriously wrong. This can be shown with the following examples.

Example 7: AIDS test.

An Italian citizen is chosen at random to undergo an AIDS test. Let us assume that the

[14] This should not be confused with the probability of the actual data, which is clearly 1, since they have been observed.

analysis used to test for HIV infection has the following performances:

$$P(Positive\,|\,HIV) \approx 1, \tag{1.11}$$
$$P(Positive\,|\,\overline{HIV}) = 0.2\%\,. \tag{1.12}$$

The analysis may declare healthy people 'Positive', even if only with a very small probability.
Let us assume that the analysis states 'Positive'. Can we say that, since the probability of an analysis error Healthy → Positive is only 0.2%, then the probability that the person is infected is 99.8%? Certainly not. If one calculates on the basis of an estimated 100 000 infected persons out of a population of 60 million, there is a 55% probability that the person is healthy![15] Some readers may be surprised to read that, in order to reach a conclusion, one needs to have an idea of how 'reasonable' the hypothesis is, independently of the data used: a mass cannot be negative; the spectrum of the true value is of a certain type; students often make mistakes; physical hypotheses happen to be incorrect; the proportion of Italians carrying the HIV virus is roughly 1 in 600. The notion of prior reasonableness of the hypothesis is fundamental to the approach we are going to present, but it is something to which physicists put up strong resistance (although in practice they often instinctively use this intuitive way of reasoning continuously and correctly). In this report I will try to show that 'priors' are rational and unavoidable, although their influence may become negligible when there is strong experimental evidence in favour of a given hypothesis.

Example 8: Probabilistic statements about the 1997 HERA high-Q^2 events.
A very instructive example of the misinterpretation of probability can be found in the statements which commented on the excess of events observed by the HERA experiments at DESY in the high-Q^2 region. For example, the official DESY statement [13] was:[16]

> *"The two HERA experiments, H1 and ZEUS, observe an excess of events above expectations at high x (or $M = \sqrt{x\,s}$), y, and Q^2. For $Q^2 > 15\,000$* GeV^2 the joint distribution has a probability of less than one per cent to come from Standard Model NC DIS processes."

Similar statements were spread out in the scientific community, and finally to the press. For example, a message circulated by INFN stated (it can be understood even in Italian)
"La probabilità che gli eventi osservati siano una fluttuazione statistica è inferiore all' 1%."
Obviously these two statements led the press (e.g. Corriere della Sera, 23 Feb. 1998) to

[15] The result will be a simple application of Bayes' theorem, which will be introduced later. A crude way to check this result is to imagine performing the test on the entire population. Then the number of persons declared Positive will be all the HIV infected plus 0.2% of the remaining population. In total 100 000 infected and 120 000 healthy persons. The general, Bayesian solution is given in Section 8.10.1

[16] One might think that the misleading meaning of that sentence was due to unfortunate wording, but this possibility is ruled out by other statements which show clearly a quite odd point of view of probabilistic matter. In fact the DESY 1998 activity report [14] insists that *"the likelihood that the data produced are the result of a statistical fluctuation ... is equivalent to that of tossing a coin and throwing seven heads or tails in a row"* (replacing 'probability' by 'likelihood' does not change the sense of the message). Then, trying to explain the meaning of a statistical fluctuation, the following example is given: *"This process can be simulated with a die. If the number of times a die is thrown is sufficiently large, the die falls equally often on all faces, i.e. all six numbers occur equally often. The probability for each face is exactly a sixth or 16.66%, assuming the die is not loaded. If the die is thrown less often, then the probability curve for the distribution of the six die values is no longer a straight line but has peaks and troughs. The probability distribution obtained by throwing the die varies about the theoretical value of 16.66% depending on how many times it is thrown."*

announce that scientists were highly confident that a great discovery was just around the corner.[17]

The experiments, on the other hand, did not mention this probability. Their published results [15] can be summarized, more or less, as *"there is a* $\lessapprox 1\%$ *probability of observing such events or rarer ones within the Standard Model".*

To sketch the flow of consecutive statements, let us indicate by SM *"the Standard Model is the only cause which can produce these events"* and by tail the *"possible observations which are rarer than the configuration of data actually observed".*

1. Experimental result: $P(data + tail \,|\, SM) \lesssim 1\%$.
2. Official statements: $P(SM \,|\, data) \lesssim 1\%$.
3. Press: $P(\overline{SM} \,|\, data) \gtrsim 99\%$, simply applying standard logic to the outcome of step 2. They deduce, correctly, that the hypothesis $\overline{SM}$ (= hint of new physics) is almost certain.

One can recognize an arbitrary inversion of probability. But now there is also something else, which is more subtle, and suspicious: *"why should we also take into account data which have not been observed?"*[18] Stated in a schematic way, it seems natural to draw conclusions on the basis of the observed data:

$$\mathbf{data} \longrightarrow P(H \,|\, data)\,,$$

although $P(H \,|\, data)$ differs from $P(data \,|\, H)$. But it appears strange that unobserved data too should play a role. Nevertheless, because of our educational background, we are so used to the inferential scheme of the kind

$$\mathbf{data} \longrightarrow P(H \,|\, data + tail)\,,$$

that we even have difficulty in understanding the meaning of this objection.[19]

Let us consider a new case, conceptually very similar, but easier to understand intuitively.

Example 9: Probability that a particular random number comes from a generator.
The value $x = 3.01$ is extracted from a Gaussian random-number generator having $\mu = 0$ and $\sigma = 1$. It is well known that

$$P(|X| > 3) = 0.27\%\,,$$

[17] One of the odd claims related to these events was on a poster of an INFN exhibition at Palazzo delle Esposizioni in Rome: *"These events are absolutely impossible within the current theory ... If they will be confirmed, it will imply that ... ."* Some friends of mine who visited the exhibition asked me what it meant that "something impossible needs to be confirmed".

[18] This is as if the conclusion from the AIDS test depended not only on $P(Positive \,|\, \overline{HIV})$ and on the prior probability of being infected, but also on the probability that this poor guy experienced events rarer than a mistaken analysis, like sitting next to Claudia Schiffer on an international flight, or winning the lottery, or being hit by a meteorite.

[19] I must admit I have fully understood this point only very recently, and I thank F. James for having asked, at the end of the CERN lectures, if I agreed with the sentence *"The probability of data not observed is irrelevant in making inferences from an experiment."* [10] I was not really ready to give a convincing reply, apart from a few intuitions, and from the trivial comment that this does not mean that we are not allowed to use MC data (strictly speaking, frequentists should not use MC data, as discussed in Section 8.1). In fact, in the lectures I did not talk about 'data+tails', but only about 'data'. This topic will be discussed again in Section 8.8.

but we cannot state that the value x has 0.27% probability of coming from that generator, or that the probability that the observation is a statistical fluctuation is 0.27%. In this case, the value comes with 100% probability from that generator, and it is at 100% a statistical fluctuation. This example helps to illustrate the logical mistake one can make in the previous examples. One may speak about the probability of the generator (let us call it A) only if another generator B is taken into account. If this is the case, the probability depends on the parameters of the generators, the observed value x and on the probability that the two generators enter the game. For example, if B has $\mu = 6.02$ and $\sigma = 1$, it is reasonable to think that

$$P(A \,|\, x = 3.01) = P(B \,|\, x = 3.01) = 0.5\,. \tag{1.13}$$

Let us imagine a variation of the example: The generation is performed according to an algorithm that chooses A or B, with a ratio of probability 10 to 1 in favour of A. The conclusions change: Given the same observed value $x = 3.01$, one would tend to infer that x is most probably due to A. It is not difficult to be convinced that, even if the value is a bit closer to the centre of generator B (for example $x = 3.3$), there will still be a tendency to attribute it to A. This natural way of reasoning is exactly what is meant by 'Bayesian', and will be illustrated in these notes.[20]. It should be noted that we are only considering the observed data ($x = 3.01$ or $x = 3.3$), and not other values which could be observed ($x \geq 3.01$, for example)

I hope these examples might at least persuade the reader to take the question of principles in probability statements seriously. Anyhow, even if we ignore philosophical aspects, there are other kinds of more technical inconsistencies in the way the standard paradigm is used to test hypotheses. These problems, which deserve extensive discussion, are effectively described in an interesting American Scientist article [10].

At this point I imagine that the reader will have a very spontaneous and legitimate objection: *"but why does this scheme of hypothesis tests usually work?"*. I will comment on this question in Section 8.8, but first we must introduce the alternative scheme for quantifying uncertainty.

[20] As an exercise, to compare the intuitive result with what we will learn later, it may be interesting to try to calculate, in the second case of the previous example ($P(A)/P(B) = 10$), the value x such that we would be in a condition of indifference (i.e. probability 50% each) with respect to the two generators.

Chapter 2

A probabilistic theory of measurement uncertainty

> *"If we were not ignorant there would be no probability, there could only be certainty. But our ignorance cannot be absolute, for then there would be no longer any probability at all. Thus the problems of probability may be classed according to the greater or less depth of our ignorance."*
> *(Henri Poincaré)*

2.1 Where to restart from?

In the light of the criticisms made in the previous chapter, it seems clear that we would be advised to completely revise the process which allows us to learn from experimental data. Paraphrasing Kant [16], one could say that (substituting the words in *italics* with those in parentheses):

> *"All metaphysicians* (physicists) *are therefore solemnly and legally suspended from their occupations till they shall have answered in a satisfactory manner the question, how are synthetic cognitions a priori possible* (is it possible to learn from observations)*?"*

Clearly this quotation must be taken in a playful way (at least as far as the invitation to suspended activities is concerned . . .). But, joking apart, the quotation is indeed more pertinent than one might initially think. In fact, Hume's criticism of the problem of induction, which interrupted the 'dogmatic slumber' of the great German philosopher, has survived the subsequent centuries.[1] We shall come back to this matter in a while.

[1]For example, it is interesting to report Einstein's opinion [17] about Hume's criticism: *"Hume saw clearly that certain concepts, as for example that of causality, cannot be deduced from the material of experience by logical methods. Kant, thoroughly convinced of the indispensability of certain concepts, took them – just as they are selected – to be necessary premises of every kind of thinking and differentiated them from concepts of empirical origin. I am convinced, however, that this differentiation is erroneous."* In the same Autobiographical Notes [17] Einstein, explaining how he came to the idea of the arbitrary character of absolute time, acknowledges that *"The type of critical reasoning which was required for the discovery of this central point was decisively furthered, in my case, especially by the reading of David Hume's and Ernst Mach's philosophical writings."* This tribute to Mach and Hume is repeated in the 'gemeinverständlich' of special relativity [18]: *"Why is it necessary to drag down from the Olympian fields of Plato the fundamental ideas of thought in natural science, and to attempt to reveal their earthly lineage? Answer: In order to free these ideas from the taboo attached to them, and thus to achieve greater freedom in the formation of ideas or concepts. It is to the immortal credit of D. Hume and E. Mach that they, above all others, introduced this critical conception."* I would like to end this parenthesis dedicated to Hume with a last citation, this time by de Finetti [11], closer to the argument of this chapter: *"In the philosophical*

In order to build a theory of measurement uncertainty which does not suffer from the problems illustrated above, we need to ground it on some kind of first principles, and derive the rest by logic. Otherwise we replace a collection of formulae and procedures handed down by tradition with another collection of cooking recipes.

We can start from two considerations.

1. In a way which is analogous to Descartes' *cogito*, the only statement with which it is difficult not to agree — in some sense the only certainty — is that (see end of Section 1.1)

   > *"the process of induction from experimental observations to statements about physics quantities (and, in general, physical hypothesis) is affected, unavoidably, by a certain degree of uncertainty".*

2. The natural concept developed by the human mind to quantify the plausibility of the statements in situations of uncertainty is that of probability.[2]

In other words we need to build a probabilistic (probabilistic and not, generically, statistic) theory of measurement uncertainty.

These two starting points seem perfectly reasonable, although the second appears to contradict the criticisms of the probabilistic interpretation of the result, raised above. However this is not really a problem, it is only a product of a distorted (i.e. different from the natural) view of the concept of probability. So, first we have to review the concept of probability. Once we have clarified this point, all the applications in measurement uncertainty will follow and there will be no need to inject *ad hoc* methods or use magic formulae, supported by authority but not by logic.

2.2 Concepts of probability

We have arrived at the point where it is necessary to define better what probability is. This is done in Section 3.3. As a general comment on the different approaches to probability, I would like, following Ref. [19], to cite de Finetti [11]:

> *"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. The numerous, different opposed attempts to put forward particular points of view which, in the opinion of their supporters, would endow Probability Theory with a 'nobler status', or a 'more scientific' character, or 'firmer' philosophical or logical foundations, have only served to generate confusion and obscurity, and to provoke well-known polemics and disagreements - even between supporters of essentially the same framework.*
>
> *The main points of view that have been put forward are as follows.*

arena, the problem of induction, its meaning, use and justification, has given rise to endless controversy, which, in the absence of an appropriate probabilistic framework, has inevitably been fruitless, leaving the major issues unresolved. It seems to me that the question was correctly formulated by Hume … and the pragmatists … However, the forces of reaction are always poised, armed with religious zeal, to defend holy obtuseness against the possibility of intelligent clarification. No sooner had Hume begun to prise apart the traditional edifice, then came poor Kant in a desperate attempt to paper over the cracks and contain the inductive argument — like its deductive counterpart — firmly within the narrow confines of the logic of certainty."

[2] Perhaps one may try to use instead fuzzy logic or something similar. I will only try to show that this way is productive and leads to a consistent theory of uncertainty which does not need continuous injections of extraneous matter. I am not interested in demonstrating the uniqueness of this solution, and all contributions on the subject are welcome.

> *The classical view is based on physical considerations of symmetry, in which one should be obliged to give the same probability to such 'symmetric' cases. But which 'symmetry'? And, in any case, why? The original sentence becomes meaningful if reversed: the symmetry is probabilistically significant, in someone's opinion, if it leads him to assign the same probabilities to such events.*
>
> *The logical view is similar, but much more superficial and irresponsible inasmuch as it is based on similarities or symmetries which no longer derive from the facts and their actual properties, but merely from sentences which describe them, and their formal structure or language.*
>
> *The frequentistic (or statistical) view presupposes that one accepts the classical view, in that it considers an event as a class of individual events, the latter being 'trials' of the former. The individual events not only have to be 'equally probable', but also 'stochastically independent' ... (these notions when applied to individual events are virtually impossible to define or explain in terms of the frequentistic interpretation). In this case, also, it is straightforward, by means of the subjective approach, to obtain, under the appropriate conditions, in perfectly valid manner, the result aimed at (but unattainable) in the statistical formulation. It suffices to make use of the notion of exchangeability. The result, which acts as a bridge connecting the new approach to the old, has often been referred to by the objectivists as "de Finetti's representation theorem'.*
>
> *It follows that all the three proposed definitions of 'objective' probability, although useless per se, turn out to be useful and good as valid auxiliary devices when included as such in the subjectivist theory."*

Also interesting is Hume's point of view on probability, where concept and evaluations are neatly separated. Note that these words were written in the middle of the 18th century [20].

> *"Though there be no such thing as Chance in the world; our ignorance of the real cause of any event has the same influence on the understanding, and begets a like species of belief or opinion.*
>
> *There is certainly a probability, which arises from a superiority of chances on any side; and according as this superiority increases, and surpasses the opposite chances, the probability receives a proportionable increase, and begets still a higher degree of belief or assent to that side, in which we discover the superiority. If a dye were marked with one figure or number of spots on four sides, and with another figure or number of spots on the two remaining sides, it would be more probable, that the former would turn up than the latter; though, if it had a thousand sides marked in the same manner, and only one side different, the probability would be much higher, and our belief or expectation of the event more steady and secure. This process of the thought or reasoning may seem trivial and obvious; but to those who consider it more narrowly, it may, perhaps, afford matter for curious speculation.*
>
> ...
>
> *Being determined by custom to transfer the past to the future, in all our inferences; where the past has been entirely regular and uniform, we expect the event with the greatest assurance, and leave no room for any contrary supposition. But where different effects have been found to follow from causes, which are to appearance exactly similar, all these various effects must occur to the mind in transferring the past to the future, and enter into our consideration, when we determine the probability of the event. Though we give the preference to that which has been found most usual, and believe that this effect will exist, we must not overlook the other effects, but must assign to each of them a particular weight and authority, in proportion as we have found it to be more or less frequent."*

2.3 Subjective probability

I would like to sketch the essential concepts related to subjective probability,[3] for the convenience of those who wish to have a short overview of the subject, discussed in detail in Part II. This should also help those who are not familiar with this approach to follow the scheme of probabilistic induction which will be presented in the next section, and the summary of the applications which will be developed in the rest of the notes.

- Essentially, one assumes that the concept of probability is primitive, i.e. close to that of common sense (said with a joke, probability is what everybody knows before going to school and continues to use afterwards, in spite of what one has been taught[4]).
- Stated in other words, probability is a measure of the degree of belief that any well-defined proposition (an event) will turn out to be true.
- Probability is related to the state of uncertainty, and not (only) to the outcome of repeated experiments.
- The value of probability ranges between 0 and 1 from events which go from false to true (see Fig. 3.1 in Section 3.3.2).
- Since the more one believes in an event the more money one is prepared to bet, the 'coherent' bet can be used to define the value of probability in an operational way (see Section 3.3.2).
- From the condition of coherence one obtains, as theorems, the basic rules of probability (usually known as axioms) and the 'formula of conditional probability' (see Sections 3.4.2 and 8.3).
- There is, in principle, an infinite number of ways to evaluate the probability, with the only condition being that they must satisfy coherence. We can use symmetry arguments, statistical data (past frequencies), Monte Carlo simulations, quantum mechanics[5] and so on. What is important is that if we get a number close to one, we are very confident that the event will happen; if the number is close to zero we are very confident that it will not happen; if $P(A) > P(B)$, then we believe in the realization of A more than in the realization of B.
- It is easy to show that the usual 'definitions' suffer from circularity[6] (Section 3.3.1), and that they can be used only in very simple and stereotypical cases. In the subjective approach they can be easily recovered as 'evaluation rules' under appropriate conditions.

[3] For an introductory and concise presentation of the subject see also Ref. [21].

[4] This remark — not completely a joke — is due to the observation that most physicists interviewed are convinced that (1.3) is legitimate, although they maintain that probability is the limit of the frequency.

[5] Without entering into the open problems of quantum mechanics, let us just say that it does not matter, from the cognitive point of view, whether one believes that the fundamental laws are intrinsically probabilistic, or whether this is just due to a limitation of our knowledge, as hidden variables *à la Einstein* would imply. If we calculate that process A has a probability of 0.9, and process B 0.4, we will believe A much more than B.

[6] Concerning the combinatorial definition, Poincaré's criticism [6] is remarkable:

> *"The definition, it will be said, is very simple. The probability of an event is the ratio of the number of cases favourable to the event to the total number of possible cases. A simple example will show how incomplete this definition is: ...*
> *... We are therefore bound to complete the definition by saying '... to the total number of possible cases, provided the cases are equally probable.' So we are compelled to define the probable by the probable. How can we know that two possible cases are equally probable? Will it be by convention? If we insert at the beginning of every problem an explicit convention, well and good! We then have*

- Subjective probability becomes the most general framework, which is valid in all practical situations and, particularly, in treating uncertainty in measurements.
- Subjective probability does not mean arbitrary[7]; on the contrary, since the normative role of coherence morally obliges a person who assesses a probability to take personal responsibility, he will try to act in the most objective way (as perceived by common sense).
- The word 'belief' can hurt those who think, naïvely, that in science there is no place for beliefs. This point will be discussed in more detail in Section 8.4. For an extensive discussion see Ref. [22].
- Objectivity is recovered if rational individuals share the same culture and the same knowledge about experimental data, as happens for most textbook physics; but one should speak, more appropriately, of intersubjectivity.
- The utility of subjective probability in measurement uncertainty has already been recognized[8] by the aforementioned ISO Guide [3], after many internal discussions (see Ref. [23] and references therein):

  > *"In contrast to this frequency-based point of view of probability an equally valid viewpoint is that probability is a measure of the degree of belief that an event will occur … Recommendation INC-1 … implicitly adopts such a viewpoint of probability."*

- In the subjective approach random variables (or, better, uncertain numbers) assume a more general meaning than that they have in the frequentistic approach: a random number is just any number in respect of which one is in a condition of uncertainty. For example:

  1. if I put a reference weight (1 kg) on a balance with digital indication to the centigramme, then the random variable is the value (in grammes) that I am expected to read (X): 1000.00, 999.95 … 1000.03 … ?
  2. if I put a weight of unknown value and I read 576.23 g, then the random value (in grammes) becomes the mass of the body (μ): 576.10, 576.12 … 576.23 … 576.50 … ?

  In the first case the random number is linked to observations, in the second to true values.

- The different values of the random variable are classified by a function $f(x)$ which quantifies the degree of belief of all the possible values of the quantity.

nothing to do but to apply the rules of arithmetic and algebra, and we complete our calculation, when our result cannot be called in question. But if we wish to make the slightest application of this result, we must prove that our convention is legitimate, and we shall find ourselves in the presence of the very difficulty we thought we had avoided."

[7] Perhaps this is the reason why Poincaré [6], despite his many brilliant intuitions, above all about the necessity of the priors (*"there are certain points which seem to be well established. To undertake the calculation of any probability, and even for that calculation to have any meaning at all, we must admit, as a point of departure, an hypothesis or convention which has always something arbitrary on it …*), concludes to *"… have set several problems, and have given no solution …"*. The coherence makes the distinction between arbitrariness and 'subjectivity' and gives a real sense to subjective probability.

[8] One should feel obliged to follow this recommendation as a metrology rule. It is however remarkable to hear that, in spite of the diffused cultural prejudices against subjective probability, the scientists of the ISO working groups have arrived at such a conclusion.

- All the formal properties of $f(x)$ are the same as in conventional statistics (average, variance, etc.).

- All probability distributions are conditioned to a given state of information: in the examples of the balance one should write, more correctly,

$$\begin{array}{rcl} f(x) & \longrightarrow & f(x \,|\, \mu = 1000.00) \\ f(\mu) & \longrightarrow & f(\mu \,|\, x = 576.23)\,. \end{array}$$

- Of particular interest is the special meaning of conditional probability within the framework of subjective probability. Also in this case this concept turns out to be very natural, and the subjective point of view solves some paradoxes of the so-called 'definition' of conditional probability (see Section 8.3).

- The subjective approach is often called Bayesian, because of the central role of Bayes' theorem, which will be introduced in Section 2.6. However, although Bayes' theorem is important, especially in scientific applications, one should not think that this is the only way to assess probabilities. Outside the well-specified conditions in which it is valid, the only guidance is that of coherence.

- Considering the result of a measurement, the entire state of uncertainty is held in $f(\mu)$; then one may calculate intervals in which we think there is a given probability to find μ, value(s) of maximum belief (mode), average, standard deviation, etc., which allow the result to be summarized with only a couple of numbers, chosen in a conventional way.

2.4 Learning from observations: the 'problem of induction'

Having briefly shown the language for treating uncertainty in a probabilistic way, it remains now to see how one builds the function $f(\mu)$ which describes the beliefs in the different possible values of the physics quantity. Before presenting the formal framework we still need a short introduction on the link between observations and hypotheses.

Every measurement is made with the purpose of increasing the knowledge of the person who performs it, and of anybody else who may be interested in it. This may be the members of a scientific community, a physician who has prescribed a certain analysis or a merchant who wants to buy a certain product. It is clear that the need to perform a measurement indicates that one is in a state of uncertainty with respect to something, e.g. a fundamental constant of physics or a theory of the Universe; the state of health of a patient; the chemical composition of a product. In all cases, the measurement has the purpose of modifying a given state of knowledge. One would be tempted to say 'acquire', instead of 'modify', the state of knowledge, thus indicating that the knowledge could be created from nothing with the act of the measurement. Instead, it is not difficult to realize that, in all cases, it is just an updating process, in the light of new facts and of some reason. Let us take the example of the measurement of the temperature in a room, using a digital thermometer — just to avoid uncertainties in the reading — and let us suppose that we get 21.7 °C. Although we may be uncertain on the tenths of a degree, there is no doubt that the measurement will have squeezed the interval of temperatures considered to be possible before the measurement: those compatible with the physiological feeling of 'comfortable environment'. According to our knowledge of the thermometer used, or of thermometers in general, there will be values of temperature in a given interval around 21.7 °C which we believe more and values outside which we believe less.[9]

It is, however, also clear that if the thermometer had indicated, for the same physiological feeling, 17.3 °C, we might think that it was not well calibrated. There would be, however, no doubt that the instrument was not working properly if it had indicated 2.5 °C!

The three cases correspond to three different degrees of modification of the knowledge. In particular, in the last case the modification is null.[10]

The process of learning from empirical observations is called induction by philosophers. Most readers will be aware that in philosophy there exists the unsolved 'problem of induction', raised by Hume. His criticism can be summarized by simply saying that induction is not justified, in the sense that observations do not lead necessarily (with the logical strength of a mathematical theorem) to certain conclusions. The probabilistic approach adopted here seems to be the only reasonable way out of such a criticism.

2.5 Beyond Popper's falsification scheme

People very often think that the only scientific method valid in physics is that of Popper's falsification scheme. There is no doubt that, if a theory is not capable of explaining experimental results, it should be rejected or modified. But, since it is impossible to demonstrate with certainty that a theory is true, it becomes impossible to decide among the infinite number of hypotheses which have not been falsified. This would produce stagnation in research. A probabilistic method allows, instead, for a scale of credibility to be provided for classifying all hypotheses taken into account (or credibility ratios between any pair of hypotheses). This is close to the natural development of science, where new investigations are made in the direction which seems the most credible, according to the state of knowledge at the moment at which the decision on how to proceed was made.

As far as the results of measurements are concerned, the falsification scheme is absolutely unsuitable. Taking it literally, one should be authorized only to check whether or not the value read on an instrument is compatible with a true value, nothing more. It is understandable then that, with this premise, one cannot go very far.

We will show that falsification is just a subcase of the Bayesian inference.

2.6 From the probability of the effects to the probability of the causes

The scheme of updating knowledge that we will use is that of Bayesian statistical inference, widely discussed in the second part of this report (in particular Sections 3.4 and 5.2). I wish to make a less formal presentation of it here, to show that there is nothing mysterious behind Bayes' theorem, and I will try to justify it in a simple way.

It is very convenient to consider true values and observed values as causes and effects (see Fig. 2.1, imagining also a continuous set of causes and many possible effects). The process of going from causes to effects it is called 'deduction'.[11] The possible values x which may be

[9] To understand the role of implicit prior knowledge, imagine someone having no scientific or technical education at all, entering a physics laboratory and reading a number on an instrument. His scientific knowledge will not improve at all, apart from the triviality that a given instrument displayed a number (not much knowledge).

[10] But also in this case we have learned something: the thermometer does not work.

[11] To be correct, the deduction we are talking about is different from the classical one. We are dealing, in fact, with probabilistic deduction, in the sense that, given a certain cause, the effect is not univocally determined.

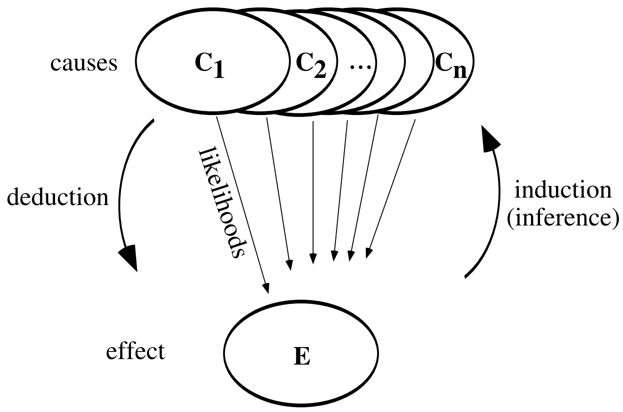

Figure 2.1: Deduction and induction.

observed are classified in belief by

$$f(x \,|\, \mu)\,.$$

This function is called 'likelihood' since it quantifies how likely it is that μ will produce any given x. It summarizes all previous knowledge on that kind of measurement (behaviour of the instruments, of influence factors, etc. – see list in Section 1.3). Often, if one deals only with random error, the $f(x \,|\, \mu)$ is a normal distribution around μ, but in principle it may have any form.

Once the likelihood is determined (we have the performance of the detector under control) we can build $f(\mu \,|\, x)$, under the hypothesis that x will be observed.[12] In order to arrive at the general formula in an heuristic way, let us consider only two values of μ. If they seem to us equally possible, it will seem natural to be in favour of the value which gives the highest likelihood that x will be observed. For example, assuming $\mu_1 = -1$, $\mu_2 = 10$, considering a normal likelihood with $\sigma = 3$, and having observed $x = 2$, one tends to believe that the observation is most likely caused by μ_1. If, on the other hand, the quantity of interest is positively defined, then μ_1 switches from most probable to impossible cause; μ_2 becomes certain. There are, in general, intermediate cases in which, because of previous knowledge (see, e.g., Fig. 1.3 and related text), one tends to believe *a priori* more in one or other of the causes. It follows that, in the light of a new observation, the degree of belief of a given value of μ will be proportional to

– the likelihood that μ will produce the observed effect;

– the degree of belief attributed to μ before the observation, quantified by $f_\circ(\mu)$.

We have finally:

$$f(\mu \,|\, x) \propto f(x \,|\, \mu) \cdot f_\circ(\mu)\,.$$

This is one of the ways to write Bayes' theorem.

[12] It is important to understand that $f(\mu \,|\, x)$ can be evaluated before one knows the observed value x. In fact, to be correct, $f(\mu \,|\, x)$ should be interpreted as beliefs of μ under the hypothesis that x is observed, and not only as beliefs of μ after x is observed. Similarly, $f(x \,|\, \mu)$ can also be built after the data have been observed, although for teaching purposes the opposite has been suggested, which corresponds to the most common case.

2.7 Bayes' theorem for uncertain quantities: derivation from a physicist's point of view

Let us show a little more formally the concepts illustrated in the previous section. This is proof of the Bayes' theorem alternative to the proof applied to events, given in Part II of these notes. It is now applied directly to uncertain (i.e. random) quantities, and it should be closer to the physicist's reasoning than the standard proof. For teaching purposes I explain it using time ordering, but this is unnecessary, as explained several times elsewhere.

- Before doing the experiment we are uncertain of the values of μ and x: we know neither the true value, nor the observed value. Generally speaking, this uncertainty is quantified by $f(x, \mu)$.

- Under the hypothesis that we observe x, we can calculate the conditional probability

$$f(\mu \,|\, x) = \frac{f(x, \mu)}{f(x)} = \frac{f(x, \mu)}{\int f(x, \mu)\, \mathrm{d}\mu}\,.$$

- Usually we don't have $f(x, \mu)$, but this can be calculated by $f(x \,|\, \mu)$ and $f(\mu)$:

$$f(x, \mu) = f(x \,|\, \mu) \cdot f(\mu)\,.$$

- If we do an experiment we need to have a good idea of the behaviour of the apparatus; therefore $f(x \,|\, \mu)$ must be a narrow distribution, and the most imprecise factor remains the knowledge about μ, quantified by $f(\mu)$, usually very broad. But it is all right that this should be so, because we want to learn about μ.

- Putting all the pieces together we get the standard formula of Bayes' theorem for uncertain quantities:

$$f(\mu \,|\, x) = \frac{f(x \,|\, \mu) \cdot f(\mu)}{\int f(x \,|\, \mu) \cdot f(\mu)\, \mathrm{d}\mu}\,.$$

The steps followed in this proof of the theorem should convince the reader that $f(\mu \,|\, x)$ calculated in this way is the best we can say about μ with the given status of information.

2.8 Afraid of 'prejudices'? Inevitability of principle and frequent practical irrelevance of the priors

Doubtless, many readers could be at a loss at having to accept that scientific conclusions may depend on prejudices about the value of a physical quantity ('prejudice' currently has a negative meaning, but in reality it simply means 'scientific judgement based on previous experience'). We shall have many opportunities to enter again into discussion about this problem, but it is important to give a general overview now and to make some firm statements on the role of priors.

- First, from a theoretical point of view, it is impossible to get rid of priors; that is if we want to calculate the probability of events of practical interest, and not just solve mathematical games.

- At a more intuitive level, it is absolutely reasonable to draw conclusions in the light of some reason, rather than in a purely automatic way.

- In routine measurements the interval of prior acceptance of the possible values is so large, compared to the width of the likelihood (seen as a function of μ), that, in practice, it is as if all values were equally possible. The prior is then absorbed into the normalization constant:

$$f(x\,|\,\mu)\cdot f_\circ(\mu) \xrightarrow[\text{prior very vague}]{} f(x\,|\,\mu). \tag{2.1}$$

- If, instead, this is not the case, it is absolutely legitimate to believe more in personal prejudices than in empirical data. This could be when one uses an instrument of which one is not very confident, or when one does for the first time measurements in a new field, or in a new kinematical domain, and so on. For example, it is easier to believe that a student has made a trivial mistake than to conceive that he has discovered a new physical effect. An interesting case is mentioned by Poincaré [6]:

> *"The impossibility of squaring the circle was shown in 1885, but before that date all geometers considered this impossibility as so 'probable' that the Académie des Sciences rejected without examination the, alas! too numerous memoirs on this subject that a few unhappy madmen sent in every year. Was the Académie wrong? Evidently not, and it knew perfectly well that by acting in this manner it did not run the least risk of stifling a discovery of moment. The Académie could not have proved that it was right, but it knew quite well that its instinct did not deceive it. If you had asked the Academicians, they would have answered: 'We have compared the probability that an unknown scientist should have found out what has been vainly sought for so long, with the probability that there is one madman the more on the earth, and the latter has appeared to us the greater.'"*

In conclusion, contrary to those who try to find objective priors which would give the Bayesian theory a nobler status of objectivity, I prefer to state explicitly the naturalness and necessity of subjective priors [22]. If rational people (e.g. physicists), under the guidance of coherency (i.e. they are honest), but each with unavoidable personal experience, have priors which are so different that they reach divergent conclusions, it just means that the data are still not sufficiently solid to allow a high degree of intersubjectivity (i.e. the subject is still in the area of active research rather than in that of consolidated scientific culture). On the other hand, the step from abstract objective rules to dogmatism is very short [22].

Turning now to the more practical aspect of presenting a result, I will give some recommendations about unbiased ways of doing this, in cases when priors are really critical (Section 9.2). Nevertheless, it should be clear that:

- since the natural conclusions should be probabilistic statements on physical quantities, someone has to turn the likelihoods into probabilities, and those who have done the experiment are usually the best candidates for doing this;

- taking the spirit of publishing unbiased results — which is in principle respectable — to extremes, one should not publish any result, but just raw data tapes.

2.9 Recovering standard methods and short-cuts to Bayesian reasoning

Before moving on to applications, it is necessary to answer an important question: *"Should one proceed by applying Bayes' theorem in every situation?"* The answer is no, and the alternative is essentially implicit in (2.1), and can be paraphrased with the example of the dog and the hunter.

We have already used this example in Section 1.7, when we were discussing the arbitrariness of probability inversion performed unconsciously by (most of)[13] those who use the scheme of confidence intervals. The same example will also be used in Section 5.2.3, when discussing the reason why Bayesian estimators appear to be distorted (a topic discussed in more detail in Section 8.5). This analogy is very important, and, in many practical applications, it allows us to bypass the explicit use of Bayes' theorem when priors do not sizably influence the result (in the case of a normal model the demonstration can be seen in Section 5.4.2).

Figure 2.2 shows how it is possible to recover standard methods from a Bayesian perspective. One sees that the crucial link is with the Maximum Likelihood Principle, which, in this approach is just a subcase (see Section 5.2.2). Then, when extra simplifying restrictions are verified, the different forms of the Least Squares are reobtained. In conclusion:

- One is allowed to use these methods if one thinks that the approximations are valid; the same happens with the usual propagation of uncertainties and of their correlations, outlined in the next section.

- One keeps the Bayesian interpretation of the results; in particular, one is allowed to talk about the probability distributions of the true values, with all the philosophical and practical advantages we have seen.

- Even if the priors are not negligible, but the final distribution is roughly normal,[14] one can evaluate the expected value and standard deviation from the shape of the distribution, as is well known:

$$\frac{\partial \ln f(\mu \,|\, x)}{\partial \mu} = 0 \quad \Rightarrow \quad E(\mu) \approx \mu_m, \tag{2.2}$$

$$-\left.\frac{\partial^2 \ln f(\mu \,|\, x)}{\partial \mu^2}\right|_{\mu_m} \quad \Rightarrow \quad \approx \frac{1}{Var(\mu)}\,, \tag{2.3}$$

  where μ_m stands for the mode of the distribution.

2.10 Evaluation of uncertainty: general scheme

Now that we have set up the framework, we can draw the general scheme to evaluate uncertainty in measurement in the most general cases. For the basic applications we will refer to the "primer" and to the appendix. For more sophisticated applications the reader is recommended to search in specialized literature.

2.10.1 Direct measurement in the absence of systematic errors

The first step consists in evaluating the uncertainty on a quantity measured directly. The most common likelihoods which describe the observed values are the Gaussian, the binomial and the Poisson distributions.

[13] Although I don't believe it, I leave open the possibility that there really is someone who has developed some special reasoning to avoid, deep in his mind, the category of the probable when figuring out the uncertainty on a true value.

[14] In case of doubt it is recommended to plot it.

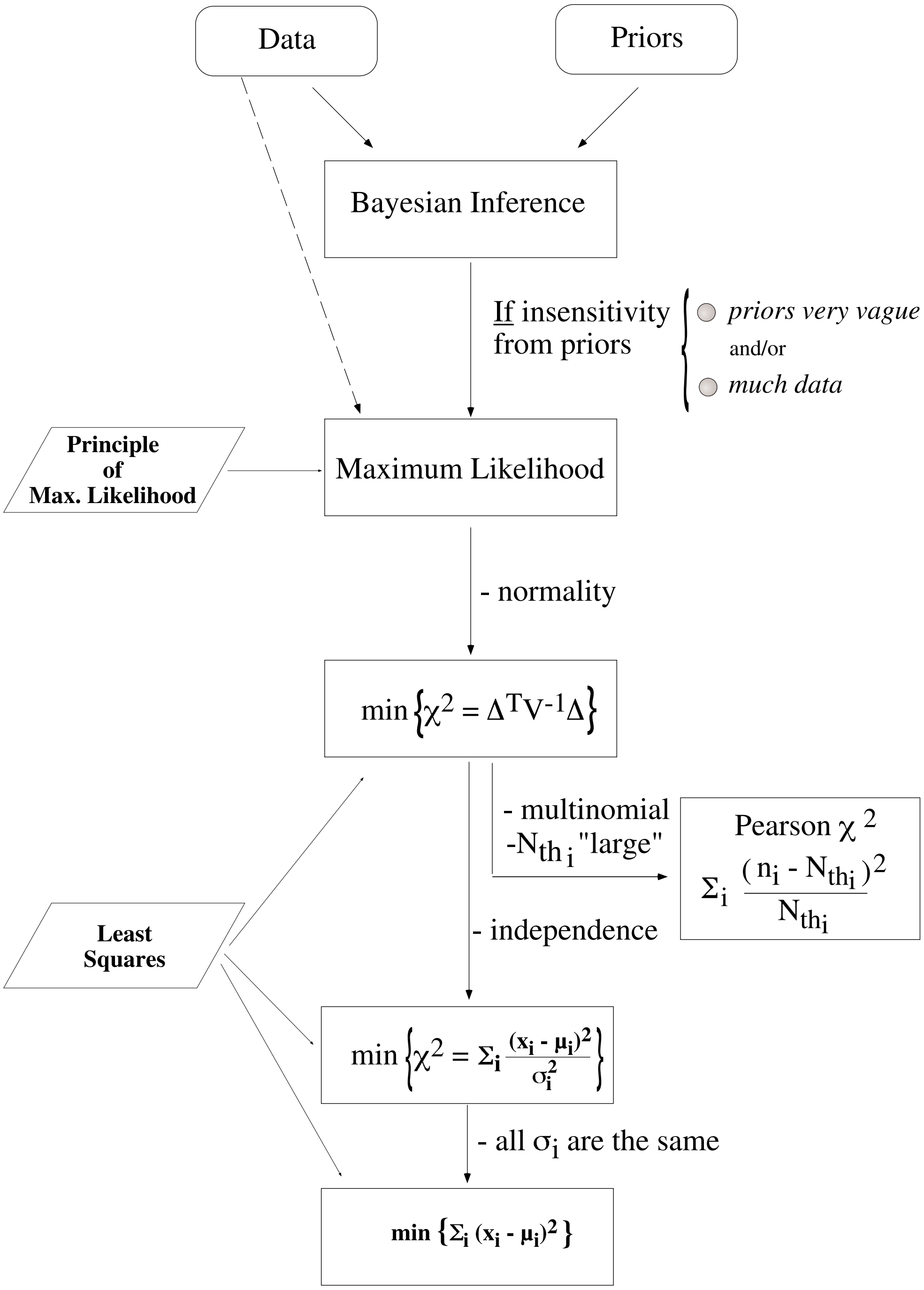

Figure 2.2: Relation between Bayesian inference and standard data analysis methods. The top-down flow shows subsequent limiting conditions. For an understanding of the relation between the 'normal' χ^2 and the Pearson χ^2 Ref. [24] is recommended.

Gaussian: This is the well-known case of 'normally' distributed errors. For simplicity, we will only consider σ independent of μ (constant r.m.s. error within the range of measurability), but there is no difficulty of principle in treating the general case. The following cases will be analysed:

- inference on μ starting from a prior much more vague than the width of the likelihood (Section 5.4.1);
- prior width comparable with that of the likelihood (Section 5.4.2): this case also describes the combination of independent measurements;
- observed values very close to, or beyond the edge of the physical region (Section 5.4.3);
- a method to give unbiased estimates will be discussed in Sections 9.2 and 9.2.1, but at the cost of having to introduce fictitious quantities.

Binomial: This distribution is important for efficiencies and, in the general case, for making inferences on unknown proportions. The cases considered include (see Section 5.5.1):

- general case with flat prior leading to the recursive Laplace formula (the problem solved originally by Bayes);
- limit to normality;
- combinations of different datasets coming from the same proportion;
- upper and lower limits when the efficiency is 0 or 1;
- comparison with Poisson approximation.

Poisson: The cases of counting experiments here considered[15] are (see Section 5.5.2):

- inference on λ starting from a flat distribution;
- upper limit in the case of null observation;
- counting measurements in the presence of a background, when its rate is well known (Sections 5.6.5 and 9.1.6);
- more complicated case of background with an uncertain rate (Section 5.6.5);
- dependence of the conclusions on the choice of experience-motivated priors (Section 9.1);
- combination of upper limits, also considering experiments of different sensitivity (Section 9.1.3).
- effect of possible systematic errors (Section 9.1.4);
- a special section will be dedicated to the lower bounds on the mass of a new hypothetical particle from counting experiments and from direct information (Section 9.3).

[15]For a general and self-contained discussion concerning the inference of the intensity of Poisson processes at the limit of the detector sensitivity, see Ref. [25].

2.10.2 Indirect measurements

The case of quantities measured indirectly is conceptually very easy, as there is nothing to 'think'. Since all values of the quantities are associated with random numbers, the uncertainty on the input quantities is propagated to that of output quantities, making use of the rules of probability. Calling μ_1, μ_2 and μ_3 the generic quantities, the inferential scheme is:

$$\begin{array}{l} f(\mu_1 \,|\, data_1) \\ f(\mu_2 \,|\, data_2) \end{array} \xrightarrow[\mu_3=g(\mu_1,\mu_2)]{} f(\mu_3 \,|\, data_1, data_2)\,. \tag{2.4}$$

The problem of going from the probability density functions (p.d.f.'s) of μ_1 and μ_2 to that of μ_3 makes use of probability calculus, which can become difficult, or impossible to do analytically, if p.d.f.'s or $g(\mu_1, \mu_2)$ are complicated mathematical functions. Anyhow, it is interesting to note that the solution to the problem is, indeed, simple, at least in principle. In fact, $f(\mu_3)$ is given, in the most general case, by

$$f(\mu_3) = \int f(\mu_1) \cdot f(\mu_2) \cdot \delta(y_3 - g(\mu_1, \mu_2))\, \mathrm{d}\mu_1 \mathrm{d}\mu_2\,, \tag{2.5}$$

where $\delta()$ is the Dirac delta function. The formula can be easily extended to many variables, or even correlations can be taken into account (one needs only to replace the product of individual p.d.f.'s by a joint p.d.f.). Equation (2.5) has a simple intuitive interpretation: the infinitesimal probability element $f(\mu_3)\, d\mu_3$ depends on 'how many' (we are dealing with infinities!) elements $d\mu_1 d\mu_2$ contribute to it, each weighed with the p.d.f. calculated in the point $\{\mu_1, \mu_2\}$. An alternative interpretation of Eq. (2.5), very useful in applications, is to think of a Monte Carlo simulation, where all possible values of μ_1 and μ_2 enter with their distributions, and correlations are properly taken into account. The histogram of μ_3 calculated from $\mu_3 = g(\mu_1, \mu_2)$ will 'tend' to $f(\mu_3)$ for a large number of generated events.[16]

In routine cases the propagation is done in an approximate way, assuming linearization of $g(\mu_1, \mu_2)$ and normal distribution of μ_3. Therefore only variances and covariances need to be calculated. The well-known error propagation formulae are recovered (Section 2.10.4), but now with a well-defined probabilistic meaning.

2.10.3 Systematic errors

Uncertainty due to systematic effects is also included in a natural way in this approach. Let us first define the notation (i is the generic index):

- $\underline{x} = \{x_1, x_2, \ldots x_{n_x}\}$ is the 'n-tuple' (vector) of observables X_i;
- $\underline{\mu} = \{\mu_1, \mu_2, \ldots \mu_{n_\mu}\}$ is the n-tuple of true values μ_i;
- $\underline{h} = \{h_1, h_2, \ldots h_{n_h}\}$ is the n-tuple of influence quantities H_i.

By influence quantities we mean:

→ all kinds of external factors which may influence the result (temperature, atmospheric pressure, etc.);

[16] As we shall see, the use of frequencies is absolutely legitimate in subjective probability, once the distinction between probability and frequency is properly made. In this case it works because of the Bernoulli theorem, which states that for a very large Monte Carlo sample *"it is very improbable that the frequency distribution will differ much from the p.d.f."* (This is the probabilistic meaning to be attributed to 'tend'.)

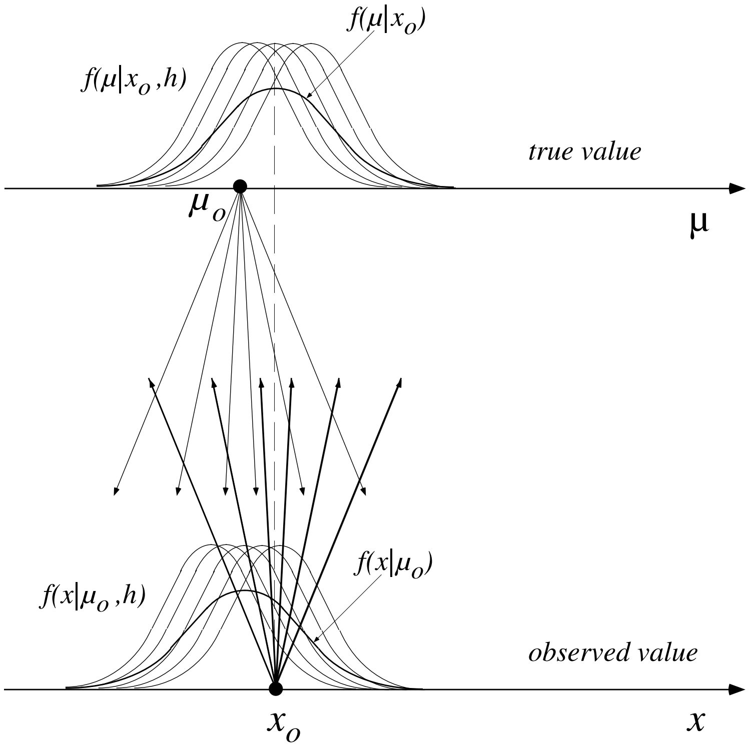

Figure 2.3: Model to handle the uncertainty due to systematic errors by the use of conditional probability.

→ all calibration constants;

→ all possible hypotheses upon which the results may depend (e.g. Monte Carlo parameters).

From a probabilistic point of view, there is no distinction between $\underline{\mu}$ and $\underline{h}$: they are all conditional hypotheses for the $\underline{x}$, i.e. causes which produce the observed effects. The difference is simply that we are interested in $\underline{\mu}$ rather than in $\underline{h}$.[17]

There are alternative ways to take into account the systematic effects in the final distribution of $\underline{\mu}$:

1. Global inference on $f(\underline{\mu}, \underline{h})$. We can use Bayes' theorem to make an inference on $\underline{\mu}$ and $\underline{h}$, as described in Section 5.2.1:

$$\underline{x} \Rightarrow f(\underline{\mu}, \underline{h} \,|\, \underline{x}) \Rightarrow f(\underline{\mu} \,|\, \underline{x})\,.$$

   This method, depending on the joint prior distribution $f_\circ(\underline{\mu}, \underline{h})$, can even model possible correlations between $\underline{\mu}$ and $\underline{h}$ (e.g. radiative correction depending on the quantity of interest).

2. Conditional inference (see Fig. 2.3). Given the observed data, one has a joint distribution

[17]For example, in the absence of random error the reading (X) of a voltmeter depends on the probed voltage (V) and on the scale offset (Z): $X = V - Z$. Therefore, the result from the observation of $X = x$ gives only a constraint between V and Z:

$$V - Z = x\,.$$

If we know Z well (within unavoidable uncertainty), then we can learn something about V. If instead the prior knowledge on V is better than that on Z we can use the measurement to calibrate the instrument.

of $\underline{\mu}$ for all possible configurations of $\underline{h}$:

$$\underline{x} \Rightarrow f(\underline{\mu}\,|\,\underline{x}, \underline{h})\,.$$

Each conditional result is reweighed with the distribution of beliefs of $\underline{h}$, using the well-known law of probability:

$$f(\underline{\mu}\,|\,\underline{x}) = \int f(\underline{\mu}\,|\,\underline{x}, \underline{h}) \cdot f(\underline{h})\,\mathrm{d}\underline{h}\,. \tag{2.6}$$

3. Propagation of uncertainties. Essentially, one applies the propagation of uncertainty, whose most general case has been illustrated in the previous section, making use of the following model: One considers a raw result on raw values $\underline{\mu}_R$ for some nominal values of the influence quantities, i.e.

$$f(\underline{\mu}_R\,|\,\underline{x}, \underline{h}_\circ)\,;$$

then (corrected) true values are obtained as a function of the raw ones and of the possible values of the influence quantities, i.e.

$$\mu_i = \mu_i(\mu_{i_R}, \underline{h})\,.$$

The three ways lead to the same result and each of them can be more or less intuitive to different people, and more less suitable for different applications. For example, the last two, which are formally equivalent, are the most intuitive for HEP experimentalists, and it is conceptually equivalent to what they do when they vary — within reasonable intervals — all Monte Carlo parameters in order to estimate the systematic errors.[18] The third form is particularly convenient to make linear expansions which lead to approximated solutions (see Section 6.1).

There is an important remark to be made. In some cases it is preferable not to 'integrate'[19] over all h's. Instead, it is better to report the result as $f(\underline{\mu}\,|\,\{h\})$, where $\{h\}$ stands for a subset of $\underline{h}$, taken at their nominal values, if:

- $\{h\}$ could be controlled better by the users of the result (for example $h_i \in \{h\}$ is a theoretical quantity on which there is work in progress);
- there is some chance of achieving a better knowledge of $\{h\}$ within the same experiment (for example h_i could be the overall calibration constant of a calorimeter);
- a discrete and small number of very different hypotheses could affect the result. For example:

$$\begin{aligned} f(\alpha_s\,|\,Th_1, \mathcal{O}(\alpha_s^2), \dots) &= \dots \\ f(\alpha_s\,|\,Th_2, \mathcal{O}(\alpha_s^2), \dots) &= \dots\,. \end{aligned}$$

  This is, in fact, the standard way in which this kind of result has been presented (apart from the inessential fact that only best values and standard deviations are given, assuming normality).

If results are presented under the condition of $\{h\}$, one should also report the derivatives of $\underline{\mu}$ with respect to the result, so that one does not have to redo the complete analysis when the influence factors are better known. A typical example in which this is usually done is the possible variation of the result due to the precise values of the charm-quark mass. A recent example in which this idea has been applied thoroughly is given in Ref. [26].

[18] But, in order to give a well-defined probabilistic meaning to the result, the variations must be performed according to $f(\underline{h})$, and not arbitrary.

[19] 'Integrate' stands for a generic term which also includes the approximate method just described.

2.10.4 Approximate methods

Of extreme practical importance are the approximate methods, which enable us not only to avoid having to use Bayes' theorem explicitly, but also to avoid working with probability distributions. In particular, propagation of uncertainty, including due to statistical effects of unknown size, is done in this way in all routine applications, as has been remarked in the previous section. These methods are discussed in Chapter 6, together with some words of caution about their uncritical use (see Sections 6.1.5, 6.2 and 6.3.2).