Part III

Other comments, examples and applications

Chapter 8

Appendix on probability and inference

8.1 Unifying role of subjective approach

I would like to give some examples to clarify what I mean by 'linguistic schizophrenia' (see Section 3.3.2). Let us consider the following:

1. probability of a '6' when tossing a die;
2. probability that the 100001st event will be accepted in the cuts of the analysis of simulated events, if I know that 91 245 out of 100 000 events[1] have already been accepted;
3. probability that a real event will be accepted in the analysis, given the knowledge of point 2, and assuming that exactly the same analysis program is used, and that the Monte Carlo describes best the physics and the detector;
4. probability that an observed track is π^+, if I have learned from the Monte Carlo that ... ;
5. probability that the Higgs mass is greater than 400 GeV;
6. probability that the 1000th decimal digit of π is 5;
7. probability of rain tomorrow;
8. probability that the US dollar will be exchanged at ≥ 2 DM before the end of 1999 (statement made in spring 1998).

Let us analyse in detail the statements.

- The evaluation of point 1 is based on considerations of physical symmetry, using the combinatorial evaluation rule. The first remark is that a convinced frequentist should abstain from assessing such a probability until he has collected statistical data on that die. Otherwise he is implicitly assuming that the frequency-based definition is not a definition, but one of the possible evaluation rules (and then the concept can only be that related to the degree of belief ...).

[1] Please note that 'event' is also used here according to HEP jargon (this is quite a case of homonymy to which one has to pay attention, but it has nothing to do with the linguistic schizophrenia I am talking about).

For those who, instead, believe that probability is only related to symmetry the answer appears to be absolutely objective: 1/6. But it is clear that one is in fact giving a very precise and objective answer to something that is not real. Instead, we should only talk about reality. This example should help to clarify the de Finetti sentence quoted in Section 2.2 (*"The classical view ... "*, in particular, *"The original sentence becomes meaningful if reversed ... "*).

- Point 2 leads to a consistent answer within the frequentistic approach, which is numerically equal to the subjective one [see, for example, (5.33) and (5.36)], whilst it has no solution in a combinatorial definition.
- Points 3 and 4 are different from point 2. The frequentistic definition is not applicable. The translation from simulated events to real events is based on beliefs, which may be as firmly based as you like, but they remain beliefs. So, although this operation is routinely carried out by every experimentalist, it is meaningful only if the probability is meant as a degree of belief and not a limit of relative frequency.
- Points 3–7 are only meaningful if probability is interpreted as a degree of belief.[2]

The unifying role of subjective probability should be clear from these examples. All those who find statements 1–7 meaningful, are implicitly using subjective probability. If not, there is nothing wrong with them, on condition that they make probabilistic statements only in those cases where their definition of probability is applicable (essentially never in real life and in research). If, however, they still insist on speaking about probability outside the condition of validity of their definition, refusing the point of view of subjective probability, they fall into the self-declared linguistic schizophrenia of which I am talking, and they generate confusion.[3]

Another very important point is the crucial role of coherence (see Section 3.3.2), which allows the exchange of the value of the probability between rational individuals: if someone tells me that he judges the probability of a given event to be 68%, then I imagine that he is as confident about it as he would be about extracting a white ball from a box which contains 100 balls, 68 of which are white. This event could be related, for example, to the result of a measurement:

$$\mu = \mu_\circ \pm \sigma(\mu)\,,$$

assuming a Gaussian model. If an experimentalist feels ready to place a 2:1 bet[4] in favour of the statement, but not a 1:2 bet against it, it means that his assessment of probability is not coherent. In other words, he is cheating, for he knows that his result will be interpreted differently from what he really believes (he has consciously overestimated the 'error bar', because he is afraid of being contradicted). If you want to know whether a result is coherent, take an interval given by 70% of the quoted uncertainty and ask the experimentalist if he is ready to place a 1:1 bet in either direction.

8.2 Frequentists and combinatorial evaluation of probability

In the previous section it was was said that frequentists should abstain from assessing probabilities if a long-run experiment has not been carried out. But frequentists do, using a sophisticated

[2] In fact, one could use the combinatorial evaluation in point 6 as well, because of the discussed cultural reasons, but not everybody is willing to speak about the probability of something which has a very precise value, although unknown.

[3] See for example Refs. [46] and [60], where it is admitted that the Bayesian approach is good for decision problems, although they stick to the frequentistic approach.

[4] This corresponds to a probability of $2/3 \approx 68\%$.

reasoning, of which perhaps not everyone is aware. I think that the best way to illustrate this reasoning is with an example of an authoritative exponent, Polya [61], who adheres to von Mises' views [62].

> *"A bag contains p balls of various colors among which there are exactly f white balls. We use this simple apparatus to produce a random mass phenomenon. We draw a ball, we look at its color and we write W if the ball is white, but we write D if it is of a different color. We put back the ball just drawn into the bag, we shuffle the balls in the bag, then we draw again one and note the color of this second ball, W or D. In proceeding so, we obtain a random sequence (...):*
>
> $$W\ D\ D\ D\ W\ D\ D\ W\ W\ D\ D\ D\ W\ W\ D\ .$$
>
> *What is the long range relative frequency of the white balls?*
>
> *Let us assume that the balls are homogeneous and exactly spherical, made of the same material and having the same radius. Their surfaces are equally smooth, and their different coloration influences only negligibly their mechanical behavior, if it has any influence at all. The person who draws the balls is blindfolded or prevented in some other manner from seeing the balls. The position of the balls in the bag varies from one drawing to the other, is unpredictable, beyond our control. Yet the permanent circumstances are well under control: the balls are all the same shape, size, and weight; they are indistinguishable by the person who draws them.*
>
> *Under such circumstances we see no reason why one ball should be preferred to another and we naturally expect that, in the long run, each ball will be drawn approximately equally often. Let us say that we have the patience to make 10000 drawings. Then we should expect that each of the p balls will appear about*
>
> $$\frac{10\,000}{p}\ times.$$
>
> *There are f white balls. Therefore, in 10 000 drawings, we expect to get white*
>
> $$f\,\frac{10\,000}{p} = 10\,000\,\frac{f}{p}\ times;$$
>
> *this is the expected frequency of the white balls. To obtain the relative frequency, we have to divide by the number of observations, or drawings, that is, 10 000. And so we are led to the statement: the long range relative frequency, or probability, of the white balls is f/p.*
>
> *The letters f and p are chosen to conform to the traditional mode of expression. As we have to draw one of the p balls, we have to choose one of p possible cases. We have good reasons (equal condition of the p balls) not to prefer any of these p possible cases to any other. If we wish that a white ball should be drawn (for example, if we are betting on white), the f white balls appear to us as favourable cases. Hence we can describe the probability f/p as the ratio of the number of favourable cases to the number of possible cases."*

The approach sketched in the above example is based on the refusal of calling probability (the intuitive concept of it) by its name. The term 'probability' is used instead for 'long-range relative frequency'. Nevertheless, the value of probability is not evaluated from the information about past frequency, but from the hypothetical long-range relative frequency, based on: a) plausible (and subjective!) reasoning on equiprobability (although not stated with this term) of the possible outcomes; b) the expectation ($\equiv$belief) that the relative frequency will be equal to the fraction of white balls in the bag.[5] The overall result is to confuse the matter, without any

[5] Sometimes this expectation is justified advocating the law of large numbers, expressed by the Bernoulli theorem. This is unacceptable, as pointed out by de Finetti: *"For those who seek to connect the notion of*

philosophical or practical advantages (compare the twisted reasoning of the above example with Hume's lucid exposure of the concept of probability and its evaluation by symmetry arguments, reported in Section 2.2).

8.3 Interpretation of conditional probability

As repeated throughout these notes, and illustrated with many examples, probability is always conditioned probability. Absolute probability makes no sense. Nevertheless, there is still something in the primer which can be misleading and that needs to be clarified, namely the so-called 'formula of conditional probability' (Section 3.4.2):

$$P(E\,|\,H) = \frac{P(E \cap H)}{P(H)} \qquad (P(H) \neq 0)\,. \tag{8.1}$$

What does it mean? Textbooks present it as a definition (a kind of 4th axiom), although very often, a few lines later in the same book, the formula $P(E \cap H) = P(E\,|\,H) \cdot P(H)$ is presented as a theorem (!).

In the subjective approach, one is allowed to talk about $P(E\,|\,H)$ independently of $P(E \cap H)$ and $P(H)$. In fact, $P(E\,|\,H)$ is just the assessment of the probability of E, under the condition that H is true. Then it cannot depend on the probability of H. It is easy to show with an example that this point of view is rather natural, whilst that of considering (8.1) as a definition is artificial. Let us take

- H = Higgs mass of 250 GeV;
- E = the decay products which are detected in a LHC detector;
- the evaluation of $P(E\,|\,H)$ is a standard PhD student task. He chooses $m_H = 250\,\mathrm{GeV}$ in the Monte Carlo and counts how many events pass the cuts (for the interpretation of this operation, see the previous section). No one would think that $P(E\,|\,H)$ must be evaluated only from $P(E \cap H)$ and $P(H)$, as the definition (8.1) would imply. Moreover, the procedure is legitimate even if we knew with certainty that the Higgs mass was below 200 GeV and, therefore, $P(H) = 0$.

In the subjective approach, (8.1) is a true theorem required by coherence. It means that although one can speak of each of the three probabilities independently of the others, once two of them have been elicited, the third is constrained. It is interesting to demonstrate the theorem to show that it has nothing to do with the kind of heuristic derivation of Section 3.4.2:

- Let us imagine a coherent bet on the conditional event $E\,|\,H$ to win a unitary amount of money ($B = 1$, as the scale factor is inessential). Remembering the meaning of conditional probability in terms of bets (see Section 3.4.2), this means that
  - we pay (with certainty) $A = P(E\,|\,H)$;
  - we win 1 if E and H are both verified (with probability $P(E \cap H)$);

probability with that of frequency, results which relate probability and frequency in some way (and especially those results like the 'law of large numbers') play a pivotal rôle, providing support for the approach and for the identification of the concepts. Logically speaking, however, one cannot escape from the dilemma posed by the fact that the same thing cannot both be assumed first as a definition and then proved as a theorem; nor can one avoid the contradiction that arises from a definition which would assume as certain something that the theorem only states to be very probable."[11]

– we get our money back (i.e. A) if H does not happen (with probability $P(\overline{H})$).

- The expected value of the 'gain' G is given by the probability of each event multiplied by the gain associated with each event:

$$\mathrm{E}(G) \;=\; 1\cdot(-P(E\,|\,H)) + P(E\cap H)\cdot 1 + P(\overline{H})\cdot P(E\,|\,H)\,,$$

where the first factors of the products on the right-hand side of the formula stand for probability, the second for the amount of money. It follows that

$$\begin{aligned}\mathrm{E}(G) &= -P(E\,|\,H) + P(E\cap H) + (1-P(H))\cdot P(E\,|\,H)\\ &= P(E\cap H) - P(E\,|\,H)\cdot P(H)\,. \end{aligned} \tag{8.2}$$

- Coherence requires the rational better to be indifferent to the direction of the bet, i.e. $\mathrm{E}(G)=0$. Applying this condition to (8.2) we obtain (8.1).

8.4 Are the beliefs in contradiction to the perceived objectivity of physics?

This is one of the most important points to be clarified since it is felt by many to be the biggest obstacle, preventing them from understanding the Bayesian approach: is there a place for beliefs in science? The usual criticism is that science must be objective and, hence, that there should be no room for subjectivity. A colleague once told me: *"I do not believe something. I assess it. This is not a matter for religion!"*

As I understand it, there are two possible ways to surmount the obstacle. The first is to try to give a more noble status of objectivity to the Bayesian approach, for example by formulating objective priors. In my opinion the main result of this attempt is to spoil the original nature of the theory, by adding dogmatic ingredients [22]. The second way consists, more simply, in recognizing that beliefs are a natural part of doing science. Admitting that they exist does not spoil the perceived objectivity of well-established science. In other words, one needs only to look closely at how frontier science makes progress, instead of seeking refuge in an idealized concept of objectivity.[6]

Clearly this discussion would require another book, and not just some side remarks, but I am confident that the reader for whom this report is intended, and who is supposed to have working experience in frontier research, is already prepared for what I am going to say. I find it hard to discuss these matters with people who presume to teach us about the way physics, and science in general, proceeds, without having the slightest direct experience of what they are talking about.

First of all, I would like to invite you to pay attention to the expressions we use in private and public discussions, and in written matter too.[7] Here are some examples:

- "I believe that ... ";
- "We have to get experience with ... ";

[6] My preferred motto on this matter is *"no one should be allowed to speak about objectivity unless he has had 10–20 years working experience in frontier science, economics, or any other applied field"*.

[7] For example, the statistician D. Berry [63] has amused himself by counting how many times Hawking uses 'belief', 'to believe', or synonyms, in his *'A brief history of time'*. The book could have been entitled *'A brief history of beliefs'*, pointed out Berry in his talk ...

- "I don't trust that guy (or that collaboration, or that procedure)";
- "Oh yes, if this has been told you by ... , then you can rely on it";
- "We have only used the calorimeter for this analysis, because we are not yet confident with the central detector";
- The evening before I had to talk about this subject, I overheard the following conversation in the CERN cafeteria:
  - Young fellow: *"I have measured the resistivity, and it turns out to be* $10^{11}\,\Omega$*"*;
  - Senior: *"No, it cannot be. Tomorrow I will make the measurement and I am sure to get the right value. ... By the way, have you considered that ... ?"*

The role of beliefs in physics has been highlighted out in a particularly efficient way by the science historian Peter Galison [37]:

> *"Experiments begin and end in a matrix of beliefs. ... beliefs in instrument type, in programs of experiment enquiry, in the trained, individual judgments about every local behaviour of pieces of apparatus."*

Then, taking as an example the discovery of the positron, he remarks:

> *"Taken out of time there is no sense to the judgment that Anderson's track 75 is a positive electron; its textbook reproduction has been denuded of the prior experience that made Anderson confident in the cloud chamber, the magnet, the optics, and the photography."*

This means that pure observation does not create, or increase, knowledge without personal inputs which are needed to elaborate the information.[8] In fact, there is nothing really objective in physics, if by objective we mean that something follows necessarily from observation, like the proof of a theorem. There are, instead, beliefs everywhere. Nevertheless, physics is objective, or at least that part of it that is at present well established, if we mean by 'objective', that a rational individual cannot avoid believing it. This is the reason why we can talk in a relaxed way about beliefs in physics without even remotely thinking that it is at the same level as the stock exchange, betting on football scores, or ... New Age. The reason is that, after centuries of experimentation, theoretical work and successful predictions, there is such a consistent network of beliefs, it has acquired the status of an objective construction: one cannot mistrust one of the elements of the network without contradicting many others. Around this solid core of objective knowledge there are fuzzy borders which correspond to areas of present investigations, where the level of intersubjectivity is still very low. Nevertheless, when one proposes a new theory or model, one has to check immediately whether it contradicts some well-established beliefs. An interesting example comes from the 1997 HERA high Q^2 events, already discussed in Section 1.9. A positive consequence of this claim was to trigger a kind of mega-exercise undertaken by many theorists, consisting of systematic cross-checks of HERA data, candidate theories, and previous experimental data. The conclusion is that the most influential physicists[9] tend not to

[8] Recently, I met an elderly physicist at the meeting of the Italian Physical Society, who was nostalgic about the good old times when we could see $\pi \to \mu \to$ e decay in emulsions, and complained that at present the sophisticated electronic experiments are based on models. It took me a while to convince him that in emulsions as well he had a model and that he was not seeing these particles either.

[9] Outstanding physicists have no reluctance in talking explicitly about beliefs. Then, paradoxically, objective science is for those who avoid the word 'belief' nothing but the set of beliefs of the influential scientists to which they believe ...

believe a possible explanation in terms of new physics [64, 65]. But this has little to do with the statistical significance of the events. It is more a question of the difficulty of inserting this evidence into what is considered to be the most likely network of beliefs.

I would like to conclude this section with a Feynman quotation [66].

> *"Some years ago I had a conversation with a layman about flying saucers - because I am scientific I know all about flying saucers! I said 'I don't think there are flying saucers'. So my antagonist said, 'Is it impossible that there are flying saucers? Can you prove that it's impossible?' 'No', I said, 'I can't prove it's impossible. It's just very unlikely'. At that he said, 'You are very unscientific. If you can't prove it impossible then how can you say that it's unlikely?' But that is the way that is scientific. It is scientific only to say what is more likely and what less likely, and not to be proving all the time the possible and impossible. To define what I mean, I might have said to him, 'Listen, I mean that from my knowledge of the world that I see around me, I think that it is much more likely that the reports of flying saucers are the results of the known irrational characteristics of terrestrial intelligence than of the unknown rational efforts of extra-terrestrial intelligence'. It is just more likely. That is all."*

8.5 Biased Bayesian estimators and Monte Carlo checks of Bayesian procedures

This problem has already been raised in Sections 5.2.2 and 5.2.3. We have seen there that the expected value of a parameter can be considered, somehow, to be analogous to the estimators[10] of the frequentistic approach. It is well known, from courses on conventional statistics, that one of the nice properties an estimator should have is that of being free of bias.

Let us consider the case of Poisson and binomial distributed observations, exactly as they have been treated in Sections 5.5.1 and 5.5.2, i.e. assuming a uniform prior. Using the typical notation of frequentistic analysis, let us indicate with θ the parameter to be inferred, with $\hat{\theta}$ its estimator.

Poisson: $\theta = \lambda$; X indicates the possible observation and $\hat{\theta}$ is the estimator in the light of X:

$$\begin{aligned} \hat{\theta} &= \mathrm{E}[\lambda \,|\, X] = X + 1\,, \\ \mathrm{E}[\hat{\theta}] &= \mathrm{E}[X+1] = \lambda + 1 \neq \lambda\,. \end{aligned} \tag{8.3}$$

The estimator is biased, but consistent (the bias become negligible when X is large).

Binomial: $\theta = p$; after n trials one may observe X favourable results, and the estimator of p is then

$$\begin{aligned} \hat{\theta} &= \mathrm{E}[p \,|\, X] = \frac{X+1}{n+2}\,, \\ \mathrm{E}[\hat{\theta}] &= \mathrm{E}\left[\frac{X+1}{n+2}\right] = \frac{n\,p+1}{n+2} \neq p\,. \end{aligned} \tag{8.4}$$

In this case as well the estimator is biased, but consistent.

[10] It is worth remembering that, in the Bayesian approach, the complete answer is given by the final distribution. The prevision ('expected value') is just a way of summarizing the result, together with the standard uncertainty. Besides motivations based on penalty rules, which we cannot discuss, a practical justification is that what matters for any further approximated analysis, are expected values and standard deviation, whose properties are used in uncertainty propagation. There is nothing wrong in providing the mode(s) of the distribution or any other quantity one finds it sensible to summarize $f(\mu)$ as well.

What does it mean? The result looks worrying at first sight, but, in reality, it is the analysis of bias that is misleading. In fact:

- the initial intent is to reconstruct at best the parameter, i.e. the true value of the physical quantity identified with it;
- the freedom from bias requires only that the expected value of the estimator should equal the value of the parameter, for a given value of the parameter,

$$\begin{aligned} \mathrm{E}[\hat{\theta}\,|\,\theta] &= \theta \qquad (\text{e.g. } \mathrm{E}[\hat{\lambda}\,|\,\lambda] = \lambda)\,, \\ (\text{i.e. } \int \hat{\theta}\, f(\hat{\theta}\,|\,\theta)\, d\hat{\theta} &= \theta\ )\,. \end{aligned} \tag{8.5}$$

But what is the true value of θ? We don't know, otherwise we would not be wasting our time trying to estimate it (always keep real situations in mind!). For this reason, our considerations cannot depend only on the fluctuations of $\hat{\theta}$ around θ, but also on the different degrees of belief of the possible values of θ. Therefore they must depend also on $f_\circ(\theta)$. For this reason, the Bayesian result is that which makes the best use[11] of the state of knowledge about θ and of the distribution of $\hat{\theta}$ for each possible value θ. This can be easily understood by going back to the examples of Section 1.7. It is also easy to see that the freedom from bias of the frequentistic approach requires $f_\circ(\theta)$ to be uniformly distributed from $-\infty$ to $+\infty$ (implicitly, as frequentists refuse the very concept of probability of θ). Essentially, whenever a parameter has a limited range, the frequentistic analysis decrees that Bayesian estimators are biased.

There is another important and subtle point related to this problem, namely that of the Monte Carlo check of Bayesian methods. Let us consider the case depicted in Fig. 1.3 and imagine making a simulation, choosing the value $\mu_\circ = 1.1$, generating many (e.g. 10 000) events, and considering three different analyses:

1. a maximum likelihood analysis;
2. a Bayesian analysis, using a flat distribution for μ;
3. a Bayesian analysis, using a distribution of μ 'of the kind' $f_\circ(\mu)$ of Fig. 1.3, assuming that we have a good idea of the kind of physics we are doing.

Which analysis will reconstruct a value closest to $\mu_\circ$? You don't really need to run the Monte Carlo to realize that the first two procedures will perform equally well, while the third one, advertised as the best in these notes, will systematically underestimate $\mu_\circ$!

Now, let us assume we have observed a value of x, for example $x = 1.1$. Which analysis would you use to infer the value of μ? Considering only the results of the Monte Carlo simulation it seems obvious that one should choose one of the first two, but certainly not the third!

This way of thinking is wrong, but unfortunately it is often used by practitioners who have no time to understand what is behind Bayesian reasoning, who perform some Monte Carlo tests, and decide that the Bayesian theorem does not work![12] The solution to this apparent paradox is simple. If you believe that μ is distributed like $f_\circ(\mu)$ of Fig. 1.3, then you should use this

[11] I refer to the steps followed in the proof of Bayes' theorem given in Section 2.7. They should convince the reader that $f(\theta\,|\,\hat{\theta})$ calculated in this way is the best we can say about θ. Some say that in the Bayesian inference the answer is the answer (I have heard this sentence from A. Smith at the Valencia-6 conference), in the sense that one can use all his best knowledge to evaluate the probability of an event, but then, whatever happens, cannot change the assessed probability, but, at most, it can — and must — be taken into account for the next assessment of a different, although analogous event.

[12] This is an actual statement I have heard by Monte Carlo-oriented HEP yuppies.

distribution in the analysis and also in the generator. Making a simulation based only on a single true value, or on a set of points with equal weight, is equivalent to assuming a flat distribution for μ and, therefore, it is not surprising that the most grounded Bayesian analysis is that which performs worst in the simple-minded frequentistic checks. It is also worth remembering that priors are not just mathematical objects to be plugged into Bayes' theorem, but must reflect prior knowledge. Any inconsistent use of them leads to paradoxical results.

8.6 Frequentistic coverage

Another prejudice toward Bayesian inference shared by practitioners who have grown up with conventional statistics is related to the so-called 'frequentistic coverage'. Since, in my opinion, this is a kind of condensate of frequentistic nonsense,[13] I avoid summarizing it in my own words, as the risk of distorting something in which I cannot see any meaning is too high. A quotation[14] taken from Ref. [68] should clarify the issue:

> *"Although particle physicists may use the words 'confidence interval' loosely, the most common meaning is still in terms of original classical concept of "coverage" which follows from the method of construction suggested in Fig. ... This concept is usually stated (too narrowly, as noted below) in terms of a hypothetical ensemble of similar experiments, each of which measures m and computes a confidence interval for m_t with say, 68% C.L. Then the classical construction guarantees that in the limit of a large ensemble, 68% of the confidence intervals contain the unknown true value m_t, i.e., they 'cover' m_t. This property, called coverage in the frequentistic sense, is the defining property of classical confidence intervals. It is important to see this property as what it is: it reflects the relative frequency with which the statement, 'm_t is in the interval (m_1, m_2)', is a true statement. The probabilistic variables in this statements are m_1 and m_2; m_t is fixed and unknown. It is equally important to see what frequentistic coverage is not: it is a not statement about the degree of belief that m_t lies within the confidence interval of a particular experiment. The whole concept of 'degree of belief' does not exist with respect to classical confidence intervals, which are cleverly (some would say devilishly) defined by a construction which keeps strictly to statements about $P(m \mid m_t)$ and never uses a probability density in the variable m_t.*
>
> *This strict classical approach can be considered to be either a virtue or a flaw, but I think that both critics and adherents commonly make a mistake in describing coverage from the narrow point of view which I described in the preceeding paragraph. As Neyman himself pointed out from the beginning, the concept of coverage is not restricted to the idea of an ensemble of hypothetical nearly-identical experiments. Classical confidence intervals have a much more powerful property: if, in an ensemble of real, different, experiments, each experiment measures whatever observables it likes, and construct a 68% C.L. confidence interval, then in the long run 68% of the confidence intervals cover the true value of their respective observables. This is directly applicable to real life, and is the real beauty of classical confidence intervals."*

I think that the reader can judge for himself whether this approach seems reasonable. From the Bayesian point of view, the full answer is provided by $P(m_t \mid m)$, to use the same notation of Ref. [68]. If this evaluation has been carried out under the requirement of coherence, from $P(m_t \mid m)$ one can evaluate a probability for m_t to lie in the interval (m_1, m_2). If this probability is 68%, in order to stick to the same value this implies:

[13] Zech says, more optimistically: *"Coverage is the magic objective of classical confidence bounds. It is an attractive property from a purely esthetic point of view but it is not obvious how to make use of this concept."*[67]

[14] The translation of the symbols is as follows: m stands for the measured quantity (x or $\hat{\theta}$ in these notes); m_t stands for the true value (μ or θ here); $P(\cdot \mid \cdot)$ for $f(\cdot \mid \cdot)$.

- one believes 68% that m_t is in that interval;
- one is ready to place a $\approx 2:1$ bet on m_t being in that interval and a $\approx 1:2$ bet on m_t being elsewhere;
- if one imagines n situations in which one has similar conditions (they could be different experiments, or simply urns containing a 68% proportion of white balls) and thinks of the relative frequency with which one expects that this statement will be true (f_n), logic applied to the basic rules of probability imply that, with the increasing n, it will become more and more improbable that f_n will differ much from 68% (Bernoulli theorem).

So, the intuitive concept of 'coverage' is naturally included in the Bayesian result and it is expressed in intuitive terms (probability of true value and expected frequency). But this result has to depend also on priors, as seen in the previous section and in many other places in this report (see, for example, Section 1.7). Talking about coverage independently of prior knowledge (as frequentists do) makes no sense, and leads to contradictions and paradoxes. Imagine, for example, an experiment operated for one hour at LEP200 and reporting zero candidate events for zirconium production in e^+e^- in the absence of expected background. I do not think that there is a single particle physicist ready to believe that, if the experiment is repeated many times, in only 68% of the cases the 68% C.L. interval [0.00, 1.29] will contain the true value of the 'Poisson signal mean', as a blind use of Table II of Ref. [60] would imply.[15] If this example seems a bit odd, I invite you to think about the many 95% C.L. lower limits on the mass of postulated particles. Do you really believe that in 95% of the cases the mass is above the limit, and in 5% of the cases below the limit? If this is the case, you would bet \$5 on a mass value below the limit, and receive \$100 if this happened to be true (you should be ready to accept the bet, since, if you believe in frequentistic coverage, you must admit that the bet is fair). But perhaps you will never accept such a bet because you believe much more than 95% that the mass is above the limit, and then the bet is not fair at all; or because you are aware of thousands of lower limits, and a particle has never shown up on the 5% side ...

8.7 Bayesian networks

In Section 8.4 I mentioned the network of beliefs which give the perceived status of objectivity to consolidated science. In fact, belief networks, also called Bayesian networks, are not only an abstract idea useful in epistemology. They represent one of the most promising applications of Bayesian inference and they have generated a renewed interest in the field of artificial intelligence, where they are used for expert systems, decision makers, etc. [69].

Although, to my knowledge, there are not yet specific HEP applications of these methods, I would like to give a rough idea of what they are and how they work, with the help of a simple example. You are visiting some friends, and, minutes after being in their house, you sneeze. You know you are allergic to pollen and to cats, but it could also be a cold. What is the cause of the sneeze? Figure 8.1 sketches the problem. There are some facts about which you are sure (the sneeze, the weather conditions and the season), but you don't know if the sneeze is a symptom of a cold or of an allergy. In particular, you don't know if there is a cat in the house.

[15] One would object that this is, more or less, the result that we could obtain making a Bayesian analysis with a uniform prior. But it was said that this prior assumes a positive attitude of the experimenters, i.e. that the experiment was planned, financed, and operated by rational people, with the hope of observing something (see Sections 5.4.3 and 5.5.2). This topic, together with the issue of reporting experimental results in a prior-free way, is discussed in detail in Ref. [25].

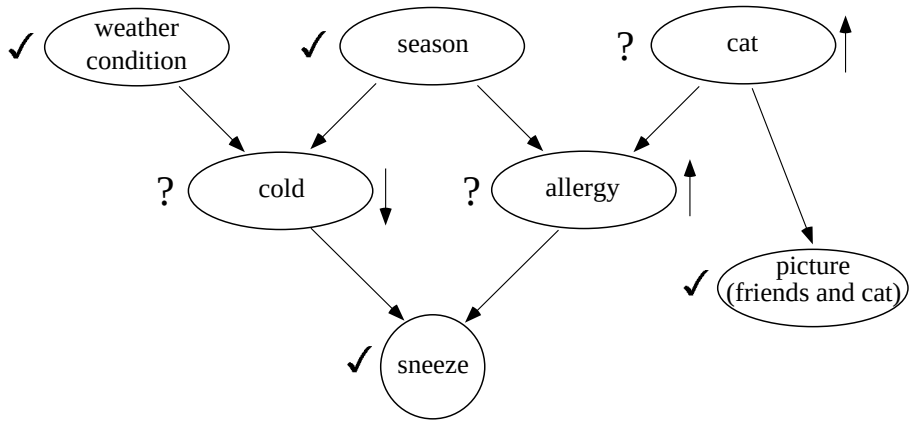

Figure 8.1: An example of belief network.

Then, you see a picture of your friend with a cat. This could be an indication that they have a cat, but it is just an indication. Nevertheless, this indication increases the probability that there is a cat around, and then the probability that the cause of the sneeze is cat's hair allergy increases, while the probability of any other potential cause decreases. If you then establish with certainty the presence of the cat, the cause of the allergy also becomes practically certain.

The idea of Bayesian networks is to build a network of causes and effects. Each event, generally speaking, can be certain or uncertain. When there is a new piece of evidence, this is transmitted to the whole network and all the beliefs are updated. The research activity in this field consists of the most efficient way of doing the calculation, using Bayesian inference, graph theory, and numerical approximations.

If one compares Bayesian networks with other ways of pursuing artificial intelligence their superiority is rather clear: they are close to the natural way of human reasoning, the initial beliefs can be those of experts (avoiding the long training needed to set up, for example, neural networks, unfeasible in practical applications), and they learn by experience as soon as they start to receive evidence [70].

8.8 Why do frequentistic hypothesis tests 'often work'?

The problem of classifying hypotheses according to their credibility is natural in the Bayesian framework. Let us recall briefly the following way of drawing conclusions about two hypotheses in the light of some data:

$$\frac{P(H_i \,|\, \text{Data})}{P(H_j \,|\, \text{Data})} \;=\; \frac{P(\text{Data} \,|\, H_i)}{P(\text{Data} \,|\, H_j)} \cdot \frac{P_\circ(H_i)}{P_\circ(H_j)} \,. \tag{8.6}$$

This form is very convenient, because:

- it is valid even if the hypotheses H_i do not form a complete class [a necessary condition if, instead, one wants to give the result in the standard form of Bayes' theorem given by formula (3.11)];
- it shows that the Bayes factor is an unbiased way of reporting the result (especially if a different initial probability could substantially change the conclusions);

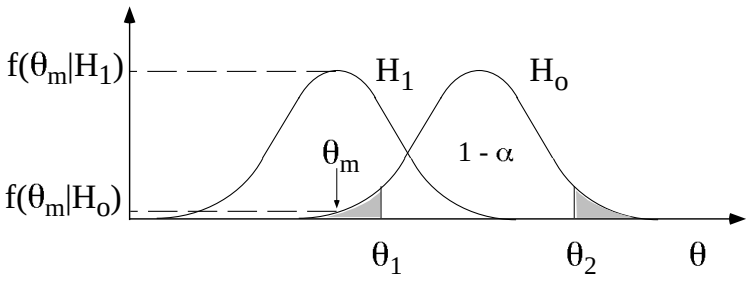

Figure 8.2: Testing a hypothesis $H_\circ$ implies that one is ready to replace it with an alternative hypothesis.

- the Bayes factor depends only on the likelihoods of observed data and not at all on unobserved data (contrary to what happens in conventional statistics, where conclusions depend on the probability of all the configurations of data in the tails of the distribution[16]). In other words, Bayes' theorem applies in the form (8.6) and not as

$$\underbrace{\frac{P(H_i\,|\,\text{Data+Tail})}{P(H_j\,|\,\text{Data+Tail})} = \frac{P(\text{Data+Tail}\,|\,H_i)}{P(\text{Data+Tail}\,|\,H_j)} \cdot \frac{P_\circ(H_i)}{P_\circ(H_j)}}_{?};$$

- testing a single hypothesis does not make sense: one may talk of the probability of the Standard Model (SM) only if one is considering an Alternative Model (AM), thus getting, for example,

$$\frac{P(\text{AM}\,|\,\text{Data})}{P(\text{SM}\,|\,\text{Data})} = \frac{P(\text{Data}\,|\,\text{AM})}{P(\text{Data}\,|\,\text{SM})} \cdot \frac{P_\circ(\text{AM})}{P_\circ(\text{SM})} :$$

  $P(\text{Data}\,|\,\text{SM})$ can be arbitrarily small, but if there is not a reasonable alternative one has only to accept the fact that some events have been observed which are very far from the expectation value;

- repeating what has been said several times, in the Bayesian scheme the conclusions depend only on observed data and on previous knowledge; in particular, they do not depend on
  - how the data have been combined;
  - data not observed and considered to be even rarer than the observed data;
  - what the experimenter was planning to do before starting to take data. (I am referring to predefined fiducial cuts and the stopping rule, which, according to the frequentistic scheme should be defined in the test protocol. Unfortunately I cannot discuss this matter here in detail and I recommend the reading of Ref. [10]).

At this point we can finally reply to the question: "why do commonly-used methods of hypothesis testing usually work?" (see Sections 1.8 and 1.9).

By reference to Fig. 8.2 (imagine for a moment the figure without the curve H_1), the argument that θ_m provides evidence against $H_\circ$ is intuitively accepted and often works, not (only) because of probabilistic considerations of θ in the light of $H_\circ$, but because it is often reasonable to imagine an alternative hypothesis H_1 that

[16] The necessity of using integrated distributions is due to the fact that the probability of observing a particular configuration is always very small, and a frequentistic test would reject the null hypotheses.

Figure 8.3: Experimental obituary (courtesy of Alvaro de Rujula [71]).

1. maximizes the likelihood $f(\theta_m \,|\, H_1)$ or, at least

$$\frac{P(\theta_m \,|\, H_1)}{P(\theta_m \,|\, H_\circ)} \gg 1\,;$$

2. has a comparable prior $[P_\circ(H_1) \approx P_\circ(H_\circ)]$, such that

$$\frac{P(H_1 \,|\, \theta_m)}{P(H_\circ \,|\, \theta_m)} = \frac{P(\theta_m \,|\, H_1)}{P(\theta_m \,|\, H_\circ)} \cdot \frac{P_\circ(H_1)}{P_\circ(H_\circ)} \quad \approx \quad \frac{P(\theta_m \,|\, H_1)}{P(\theta_m \,|\, H_\circ)} \longrightarrow \gg 1\,.$$

So, even though there is no objective or logical reason why the frequentistic scheme should work, the reason why it often does is that in many cases the test is made when one has serious doubts about the null hypothesis. But a peak appearing in the middle of a distribution, or any excess of events, is not, in itself, a hint of new physics (Fig. 8.3 is an invitation to meditation ...). My recommendations are therefore the following.

- Be very careful when drawing conclusions from χ^2 tests, '$3\,\sigma$ golden rule', and other 'bits of magic';
- Do not pay too much attention to fixed rules suggested by statistics 'experts', supervisors, and even Nobel laureates, taking also into account that
  - they usually have permanent positions and risk less than PhD students and postdocs who do most of the real work;
  - they have been 'miseducated' by the exciting experience of the glorious 1950s to 1970s: as Giorgio Salvini says, *"when I was young, and it was possible to go to sleep at night after having added within the day some important brick to the building of the*

elementary particle palace. We were certainly lucky." [72]. Especially when they were hunting for resonances, priors were very high, and the 3–4 σ rule was a good guide.

- Fluctuations exist. There are millions of frequentistic tests made every year in the world. And there is no probability theorem ensuring that the most extreme fluctuations occur to a precise Chinese student, rather than to a large HEP collaboration (this is the same reasoning of many Italians who buy national *lotteria* tickets in Rome or in motorway restaurants, because 'these tickets win more often' ...).

As a conclusion to these remarks, and to invite the reader to take with much care the assumption of equiprobability of hypothesis (a hidden assumption in many frequentistic methods), I would like to add this quotation by Poincaré [6]:

> *"To make my meaning clearer, I go back to the game of écarté mentioned before.*[17] *My adversary deals for the first time and turns up a king. What is the probability that he is a sharper? The formulae ordinarily taught give 8/9, a result which is obviously rather surprising. If we look at it closer, we see that the conclusion is arrived at as if, before sitting down at the table, I had considered that there was one chance in two that my adversary was not honest. An absurd hypothesis, because in that case I should certainly not have played with him; and this explains the absurdity of the conclusion. The function on the à priori probability was unjustified, and that is why the conclusion of the à posteriori probability led me into an inadmissible result. The importance of this preliminary convention is obvious. I shall even add that if none were made, the problem of the à posteriori probability would have no meaning. It must be always made either explicitly or tacitly."*

8.9 Frequentists and Bayesian 'sects'

Many readers may be interested in how the problem 'to Bayes or not to Bayes' is viewed by statisticians. In order to thoroughly analyse the situation, one should make a detailed study not only of the probability theory, but also of the history and sociology of statistical science. The most I can do here is to give personal impressions, certainly biased, and some references. I invite the reader to visit the statistics department in his University, browse their journals and books, and talk to people (and to judge the different theses by the logical strength of their arguments, not weighing them just by numbers ...).

8.9.1 Bayesian versus frequentistic methods

An often cited paper for a reasonably balanced discussion [46] on the subject is the article *"Why isn't everyone a Bayesian?"*, by B. Efron [73]. Key words of the paper are: *Fisherian inference*; *Frequentistic theory*; *Neyman–Pearson–Wald*; *Objectivity*. For this reason, pointing out this paper as 'balanced' is not really fair. Nevertheless, I recommend reading the article, together with the accompanying comments and the reply by the author published in the same issue of the journal (a typical practice amongst statisticians).

So, it is true that *"Fisherian and Neyman–Pearson–Wald ideas have shouldered Bayesian theory aside in statistical practice"* [73], but *"The answer is simply that statisticians do not know what the statistical paradigm says. Why should they? There are very few universities in the world with statistics departments that provides a good course on the subject."* [74] Essentially, the main point of the Efron paper is to maintain traditional methods, despite the *"disturbing*

[17] See Section 1.6.

catalog of inconsistencies" [73], and the *"powerful theoretical reasons for preferring Bayesian inference"* [73]. Moreover, perhaps not everybody who cites the Efron paper is aware of further discussions about it, like the letter in which Zellner [75] points out that one of the problems posed by Efron already had a Bayesian solution (in the Jeffreys' book [29]), that Efron admitted to knowing and even to having used [76]. As a kind of final comment on this debated paper, I would like to cite Efron's last published reply I am aware of [76]:

> *"First of all let me thank the writers for taking my article in its intended spirit: not as an attack on the Bayesian enterprise, but rather as a critique of its preoccupation with philosophical questions, to the detriment of statistical practice. Meanwhile I have received some papers, in particular one from A.F.M. Smith, which show a healthy Bayesian interest in applications, so my worries were overstated if not completely groundless."*

There are some other references which I would like to suggest if you are interested in forming your own opinion on the subject. They have also appeared in The American Statistician, where in 1997 an entire Teaching Corner section of the journal [63] was devoted to three papers presented in a round table on 'Bayesian possibilities for introductory statistics' at the 156th Annual Meeting of the American Statistical Association, held in Chicago, in August 1996. For me these articles are particularly important because I was by chance in the audience of the round table (really 'by chance'!). At the end of the presentations I was finally convinced that frequentism was dead, at least as a philosophical idea. I must say, I was persuaded by the non-arguments of the defender of frequentism even more than by the arguments of the defenders of the Bayesian approach. I report here the abstract[18] of Moore, who presented the 'reason to hesitate' to teach Bayesian statistics:

> *"The thesis of this paper is that Bayesian inference, important though it is for statisticians, is among the mainly important statistical topics that it is wise to avoid in most introductory instruction. The first reason is pragmatic (and empirical): Bayesian methods are as yet relatively little used in practice. We have an obligation to prepare students to understand the statistics they will meet in their further studies and work, not the statistics we may hope will someday replace now-standard methods. A second argument also reflects current conditions: Bayesians do not agree on standard approaches to standard problem settings. Finally, the reasoning of Bayesian inference, depending as it does on ideas of conditional probability, is quite difficult for beginners to appreciate. There is of course no easy path to a conceptual grasp of inference, but standard inference at least rests on repetition of one straightforward question, What would happen if I did this many times? "*

Even if some arguments might be valid, thinking about statisticians who make surveys in a standardized form (in fields that they rarely understand, such as medicine and agriculture), surely they do not hold in physics, even less in frontier physics. As I commented to Moore after his talk, what is important for a physicist is not "what would happen if I did this many times?", but "what am I learning by the experiment?".[19]

8.9.2 Orthodox teacher versus sharp student - a dialogue by Gabor

As a last comment about frequentistic ideas I would like to add here a nice dialogue, which was circulated via internet on 19th February 1999, with an introduction and comment by the

[18] I quote here the original abstract, which appears on page 18 of the conference abstract book.

[19] I also made other comments on the general illogicality of his arguments, which you may easily imagine by reading the abstract. For these comments I even received applause from the audience, which really surprised me, until I learned that David Moore is one of the most authoritative American statisticians: only a outsider like me would have said what I said ...

author, the statistician George Gabor [77] of Dalhousie University (Halifax, N.S., Canada). It was meant as a contribution to a discussion triggered by D.A. Berry (that of Refs. [10] and [63]) a few days before.

> *"Perhaps a Socratic exchange between an ideally sharp, i.e not easily bamboozled student (S.) of a typical introductory statistics course and his prof (P.) is the best way to illustrate what I think of the issue. The class is at the point where confidence interval (CI) for the normal mean is introduced and illustrated with a concrete example for the first time.*
>
> **P.** *...and so a 95% CI for the unknown mean is (1.2, 2.3).*
>
> **S.** *Excuse me sir, just a few minutes ago you emphasized that a CI is some kind of random interval with certain coverage properties in REPEATED trials.*
>
> **P.** *Correct.*
>
> **S.** *What, then, is the meaning of the interval above?*
>
> **P.** *Well, it is one of the many possible realizations from a collection of intervals of a certain kind.*
>
> **S.** *And can we say that the 95 collective, is somehow carried over to this particular realization?*
>
> **P.** *No, we can't. It would be worse than incorrect; it would be meaningless for the probability claim is tied to the collective.*
>
> **S.** *Your claim is then meaningless?*
>
> **P.** *No, it isn't. There is actually a way, called Bayesian statistics, to attribute a single-trial meaning to it, but that is beyond the scope of this course. However, I can assure you that there is no numerical difference between the two approaches.*
>
> **S.** *Do you mean they always agree?*
>
> **P.** *No, but in this case they do provided that you have no reason, prior to obtaining the data, to believe that the unknown mean is in any particularly narrow area.*
>
> **S.** *Fair enough. I also noticed sir that you called it 'a' CI, instead of 'the' CI. Are there others then?*
>
> **P.** *Yes, there are actually infinitely many ways to obtain CI's which all have the same coverage properties. But only the one above is a Bayesian interval (with the proviso above added, of course).*
>
> **S.** *Is Bayesian-ness the only way to justify the use of this particular one?*
>
> **P.** *No, there are other ways too, but they are complicated and they operate with concepts that draw their meaning from the collective (except the so called likelihood interval, but then this strange guy does not operate with probability at all).*
>
> **...**
>
> *It could be continued ad infinitum. Assuming sufficiently more advanced students one could come up with similar exchanges concerning practically every frequentist concept orthodoxy operates with (sampling distribution of estimates, measures of performance, the very concept of independence, etc.). The point is that orthodoxy would fail at the first opportunity had students been sufficiently sharp, open minded, and inquisitive. That we are not humiliated repeatedly by such exchanges (in my long experience not a single one has ever taken place) says more about... well, I don't quite know about what — the way the mind plays tricks with the concept of probability? The background of our students? Both?*
>
> *Ultimately then we teach the orthodoxy not only because of intellectual inertia, tradition, and the rest; but also because, like good con artists, we can get away with it. And that I find very disturbing. I must agree with Basu's dictum that nothing in orthodox statistics makes sense unless it has a Bayesian interpretation. If, as is the case, the only thing one*

can say about frequentist methods is that they work only in so far as they don't violate the likelihood principle; and if they don't (and they frequently do), they numerically agree with a Bayesian procedure with some flat prior - then we should go ahead and teach the real thing, not the substitute. (The latter, incidentally, can live only parasitically on an illicit Bayesian usage of its terms. Just ask an unsuspecting biologist how he thinks about a CI or a P-value.)

One can understand, or perhaps follow is a better word, the historical reasons orthodoxy has become the prevailing view. Now, however, we know better."

8.9.3 Subjective or objective Bayesian theory?

Once you have understood that probability and frequencies are different concepts, that probability of hypothesis is a useful and natural concept for reporting results, that Bayes' theorem is a powerful tool for updating probability and learning from data, that priors are important and pretending that they do not exist is equivalent to assuming them flat, and so on, it is difficult to then take a step back. However, it is true that there is no single shared point of view among those who, generally speaking, support the Bayesian approach. I don't pretend that I can provide an exhaustive analyse of the situation here, or to be unbiased about this matter either.

The main schools of thought are the 'subjectivists' and the 'objectivists'. The dispute may look strange to an outsider, if one thinks that both schools use probability to represent degrees of belief. Nevertheless, objectivists want to minimize the person's contribution to the inference, by introducing reference priors (for example Jeffreys' priors [29]) or other constraints, such as maximum entropy (for an overview see Refs. [19] and [78]). The motto is *"let the data speak for themselves"*. I find this subject highly confusing, and even Bernardo and Smith (Bernardo is one of the key persons behind reference priors) give the impression of contradicting themselves often on this point as, for example, when the subject of reference analysis is introduced:

"to many attracted to the formalism of the Bayesian inferential paradigm, the idea of a non-informative prior distribution, representing 'ignorance' and 'letting the data speak for themselves' has proved extremely seductive, often being regarded as synonymous with providing objective inferences. It will be clear from the general subjective perspective we have maintained throughout this volume, that we regard this search for 'objectivity' to be misguided. However, it will also be clear from our detailed development in Section 5.4 that we recognize the rather special nature and role of the concept of a 'minimal informative' prior specification - appropriately defined! In any case, the considerable body of conceptual and theoretical literature devoted to identifying 'appropriate' procedures for formulating prior representations of 'ignorance' constitutes a fascinating chapter in the history of Bayesian Statistics. In this section we shall provide an overview of some of the main directions followed in this search for a Bayesian 'Holy Grail'.[19]

In my point of view, the extreme idea along this line is represented by the Jaynes' 'robot' (*"In order to direct attention to constructive things and away from controversial irrelevance, we shall invent an imaginary being. Its brain is to be designed by us, so that it reasons according to certain defined rules. These rules will be deduced from simple desiderata which, it appears to us, would be desirable in human brains"* [79]).

As far as I understand it, I see only problems with objectivism, although I do agree on the notion of a commonly perceived objectivity, in the sense of intersubjectivity (see Section 8.4). Frankly, I find probabilistic evaluations made by a coherent subjectivist, assessed under personal

responsibility, to be more trustworthy and more objective than values obtained in a mechanical way using objective prescriptions [22].

Moving to a philosophical level deeper than this kind of angels' sex debate (see Section 3.6), there is the important issue of what an event is. All events listed in Section 8.1 (apart from that of point 4) are somehow verifiable. Perhaps one will have to wait until tomorrow, the end of 1999, or 2010, but at a certain point the event may become certain, either true or false. However, one can think about other events, examples of which have been shown in these notes, that are not verifiable, either for a question of principle, or by accident.

- The old friend could die, carrying with him the secret of whether he had been cheating, or simply lucky (Section 3.4.5).

- The particle interacts with the detector (Section 3.4.4) and continues its flight: was it really a π or a μ?

- Using our best knowledge about temperature measurement we can state that the temperature of a room at a certain instant is $21.7 \pm 0.3\,^{\circ}\mathrm{C}$ with 95% probability (Section 8.1); after the measurement the window is opened, the weather changes, the thermometer is lost: how is it possible to verify the event '$21.4 \leq T/^{\circ}\mathrm{C} \leq 22.0$'?

This problem is present every time we make a probabilistic statement about physics quantities. It is present not only when a measurand is critically time dependent (the position of a plane above the Atlantic), but also in the case of fundamental constants. In this latter case we usually believe in the progress of science and thus we hope that the quantity will be measured so well in the future that it will one day become a kind of exact value, in comparison to today's uncertainty. But it is absurd to think that one day we will be able to 'open an electron' and read on a label all its properties with an infinite number of digits. This means that for scientific applications it is convenient to enlarge the concept of an event (see Section 3.3.2), releasing the condition of verifiability.[20] At this point the normative role of the hypothetical coherent bet becomes crucial. A probability evaluation, made by an honest person well-trained in applying coherence on verifiable events, becomes, in my opinion, the only means by which degrees of belief can be exchanged among rational people. We have certainly reached a point in which the domain of physics, metaphysics and moral overlap, but it looks to me that this is exactly the way in which science advances.

It seems to me that almost all Bayesian schools support this idea of the extended meaning of an event, explicitly or tacitly (anyone who speaks about $f(\theta)$, with θ a parameter of a distribution, does it). A more radical point of view, which is very appealing from the philosophical perspective, but more difficult to apply (at least in physics), is the predictive approach (or operational subjectivism), along the lines of de Finetti's thinking. The concept of probability is strictly applied only to real observables, very precisely ('operationally') defined. The events are all associated with discrete uncertain numbers (integer or rational), in the simplest case 1 or 0 if there are only two possibilities (true or false). Having excluded non-observables, it makes no sense to speak of $f(\mu \,|\, \mathrm{data})$, but only of $f(x \,|\, \mathrm{data})$, where X stands for a future (or, in general, not yet known) observation. For the moment I prefer to stick to our 'metaphysical' true values, but I encourage anyone who is interested in this subject to read Lad's recent book [80], which also contains a very interesting philosophical and historical introduction to the subject.

[20] It is interesting to realize, in the light of this reflection, that the ISO definition of true value (*"a value compatible with the definition of a given particular quantity"*, see Sections 1.2 and 1.3) can accommodate this point of view.

8.9.4 Bayes' theorem is not all

Finally, I would like to recall that Bayes' theorem is a very important tool, but it can be used only when the scheme of prior, likelihood, and final is set up, and the distributions are properly normalized.[21] This happens very often in measurement uncertainty problems, but less frequently in other applications, such as assessing the probabilities of hypotheses. When Bayes' theorem is not applicable, conclusions may become strongly dependent on individuals and the only guidance remains the normative rule of the hypothetical coherent bet.

8.10 Solution to some problems

Here are the solutions to some of the examples of the notes.

8.10.1 AIDS test

The AIDS test problem (Example 7 of Section 1.9) is a very standard one. Let us solve it using the Bayes factor:

$$\begin{aligned}
\frac{P(\mathrm{HIV}\,|\,\mathrm{Positive})}{P(\overline{\mathrm{HIV}}\,|\,\mathrm{Positive})} &= \frac{P(\mathrm{Positive}\,|\,\mathrm{HIV})}{P(\mathrm{Positive}\,|\,\overline{\mathrm{HIV}})}\cdot\frac{P_\circ(\mathrm{HIV})}{P(\overline{\mathrm{HIV}})} \\
&= \frac{\approx 1}{0.002}\times\frac{0.1/60}{\approx 1} = 500\times\frac{1}{600} = \frac{1}{1.2} \\
P(\mathrm{HIV}\,|\,\mathrm{Positive}) &= 45.5\%\,.
\end{aligned}$$

Writing Bayes' theorem in this way helps a lot in understanding what is going on. Stated in terms of signal to noise and selectivity (see problem 1 in Section 3.4.4), we are in a situation in which the selectivity of the test is not enough for the noisy conditions. So in order to be practically sure that the patient declared 'positive' is infected, with this performance of the analysis, one needs independent tests, unless the patient belongs to high-risk classes. For example, a double independent analysis on an average person would yield

$$P(\mathrm{HIV}\,|\,\mathrm{Positive}_1\cap\mathrm{Positive}_2) = 99.76\%\,,$$

similar[22] to that obtained in the case where a physician had a 'severe doubt' (i.e. $P_\circ(\mathrm{HIV})\approx P_\circ(\overline{\mathrm{HIV}})$ that the patient could be infected:

$$P(\mathrm{HIV}\,|\,\mathrm{Positive}, P_\circ(\mathrm{HIV})\approx 0.5) = 99.80\%\,.$$

We see then that, as discussed several times (see Section 8.8), the conclusion obtained by arbitrary probability inversion is equivalent to assuming uniform priors.

[21] I have made use several times in these notes of improper distributions, i.e. such that

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x \to \infty\,,$$

but, as specified, they were always thought to be the limit of proper distributions (see, for example, Section 5.5.2).

[22] There is nothing profound in the fact that the two cases give very similar results. It is just due to the numbers of these examples (i.e. $500\approx 600$).

8.10.2 Gold/silver ring problem

The three-box problem (Section 3.4.4) seems to be intuitive for some, but not for everybody. Let us label the three boxes: A, Golden-Golden; B, Golden-Silver; C, Silver-Silver. The initial probability (i.e. before having checked the first ring) of having chosen the box A, B, or C is, by symmetry, $P_\circ(A) = P_\circ(B) = P_\circ(C) = 1/3$.

This probability is updated after the event $E =$'the first ring extracted is golden' by Bayes' theorem:

$$\begin{aligned}
P(A\,|\,E) &= \frac{P(E\,|\,A)\cdot P_\circ(A)}{P(E\,|\,A)\cdot P_\circ(A) + P(E\,|\,B)\cdot P_\circ(B) + P(E\,|\,C)\cdot P_\circ(C)} = 2/3 \\
P(B\,|\,E) &= \frac{P(E\,|\,B)\cdot P_\circ(B)}{P(E\,|\,A)\cdot P_\circ(A) + P(E\,|\,B)\cdot P_\circ(B) + P(E\,|\,C)\cdot P_\circ(C)} = 1/3 \\
P(C\,|\,E) &= \frac{P(E\,|\,C)\cdot P_\circ(C)}{P(E\,|\,A)\cdot P_\circ(A) + P(E\,|\,B)\cdot P_\circ(B) + P(E\,|\,C)\cdot P_\circ(C)} = 0\,,
\end{aligned}$$

where $P(E\,|\,A)$, $P(E\,|\,B)$ and $P(E\,|\,C)$ are, respectively, 1, 1/2 and 0.

Finally, calling $F =$'the next ring will be golden if I extract it from the same box', we have, using the probability rules:

$$\begin{aligned}
P(F\,|\,E) &= P(F|A,E)\cdot P(A|E) + P(F|B,E)\cdot P(B|E) + P(F|C,E)\cdot P(C|E) \\
&= 1\times 2/3 + 0\times 1/3 + 0\times 0 = 2/3\,.
\end{aligned}$$

Chapter 9

Further HEP applications

9.1 Poisson model: dependence on priors, combination of results and systematic effects

The inference on the parameter λ of the Poisson has been treated in Sections 5.5.2 and 5.6.5. Here we will take a look at other applications of practical interest.

9.1.1 Dependence on priors

The results of Sections 5.5.2 and 5.6.5 were obtained using a uniform prior. One may worry how much the result changes if different priors are used in the analysis. Bearing in mind the rule of coherence, we are clearly interested only in reasonable[1] priors.

In frontier physics the choice of $f_\circ(\lambda) = k$ is often not reasonable. For example, searching for monopoles, one does not believe that $\lambda = 10^6$ and $\lambda = 1$ are equally possible. Realistically, one would expect to observe, with the planned experiment and running time, $\mathcal{O}(10)$ monopoles, if they exist at all. We follow the same arguments of Section 5.4.3 (negative neutrino mass), modelling the prior beliefs of a community of rational people who have planned and run the experiment. For reasons of mathematical convenience, we model $f_\circ(\lambda)$ with an exponential, but, extrapolating the results of Section 5.4.3, it is easy to understand that the exact function is not really crucial for the final result.

The function

$$f_\circ(\lambda) = \frac{1}{10} e^{-\lambda/10}\,, \tag{9.1}$$

with

$$\begin{aligned} \mathrm{E}_\circ[\lambda] &= 10 \\ \sigma_\circ(\lambda) &= 10 \end{aligned}$$

may be well suited to the case: the highest beliefs are for small values of λ, but also values up

[1] I insist on the fact that they must be reasonable, and not just any prior. The fact that absurd priors give absurd results does not invalidate the inferential framework based on subjective probability.

to 30 or 50 would not be really surprising. We obtain the following results:

$$\begin{aligned} f(\lambda \,|\, x=0) &= \frac{e^{-\lambda}\frac{1}{10}\,e^{-\lambda/10}}{\int_0^\infty(\dots)\mathrm{d}\lambda} && (9.2)\\ &= \frac{11}{10}\,e^{-\frac{11}{10}\lambda} && (9.3)\\ \mathrm{E}[\lambda] &= 0.91 \\ P(\lambda \le 2.7) &= 95\% \\ \lambda_u &= 2.7 \text{ with } 95\% \text{ probability}\,. && (9.4) \end{aligned}$$

The result is very stable. Changing $\mathrm{E}_\circ[\lambda]$ from '∞' to 10 has only a 10% effect on the upper limit. As far as the scientific conclusions are concerned, the two limit are identical. For this reason one should not worry about using a uniform prior, and complicate one's life to model a more realistic prior.

As an exercise, we can extend this result to a generic expected value of events, still sticking to the exponential:

$$f_\circ(\lambda) = \frac{1}{\lambda_\circ}e^{-\lambda/\lambda_\circ}\,,$$

which has an expected value $\lambda_\circ$. The uniform distribution is recovered for $\lambda_\circ \to \infty$. We get:

$$\begin{aligned} f(\lambda \,|\, x=0,\lambda_\circ) &\propto e^{-\lambda}\frac{1}{\lambda_\circ}e^{-\lambda/\lambda_\circ} \\ f(\lambda \,|\, x=0,\lambda_\circ) &= (1+\lambda_\circ)e^{-\lambda\,(1+\lambda_\circ)/\lambda_\circ} \\ &= \frac{1}{\lambda_1}e^{-\lambda/\lambda_1} \\ \text{with } \frac{1}{\lambda_1} &= \frac{1}{1}+\frac{1}{\lambda_\circ} \\ F(\lambda \,|\, x=0,\lambda_\circ) &= 1-e^{-\lambda/\lambda_\circ}\,. \end{aligned}$$

The upper limit, at a probability level P_u, becomes:

$$\lambda_u = -\lambda_1\,\ln(1-P_u)\,. \qquad (9.5)$$

9.1.2 Combination of results from similar experiments

Results may be combined in a natural way making an interactive use of Bayesian inference. As a first case we assume several experiments having the same efficiency and exposure time.

- Prior knowledge:
$$f_\circ(\lambda \,|\, I_\circ)\,;$$
- Experiment 1 provides Data_1:
$$f_1(\lambda \,|\, I_\circ, \mathrm{Data}_1) \propto f(\mathrm{Data}_1 \,|\, \lambda, I_\circ)\cdot f_\circ(\lambda \,|\, I_\circ)\,;$$
- Experiment 2 provides Data_2:
$$f_2(\lambda \,|\, I_\circ, \mathrm{Data}_1 \dots) \propto f(\mathrm{Data}_2 \,|\, \lambda, I_\circ)\cdot f_1(\lambda \,|\, \dots)\,;$$
$$\Rightarrow f_2(\lambda \,|\, I_\circ, \mathrm{Data}_1, \mathrm{Data}_2)\,.$$

- Combining n similar independent experiments we get

$$\begin{aligned} f(\lambda \,|\, \underline{x}) &\propto \Pi_{i=1}^{n} f(x_i \,|\, \lambda) \cdot f_\circ(\lambda) \\ &\propto f(\underline{x} \,|\, \lambda) \cdot f_\circ(\lambda) \\ &\propto \Pi_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \cdot f_\circ(\lambda) \\ &\propto e^{-n\,\lambda}\lambda^{\sum_{i=1}^{n} x_i} f_\circ(\lambda)\,. \end{aligned} \tag{9.6}$$

  Then it is possible to evaluate expected value, standard deviation, and probability intervals.

As an exercise, let us analyse the two extreme cases, starting from a uniform prior:

$\boxed{\sum_i x_i = 0}$ if none of the n similar experiments has observed events we have

$$\begin{aligned} f(\lambda \,|\, n \text{ expts}, 0 \text{ evts}) &= n e^{-n\,\lambda} \\ F(\lambda \,|\, n \text{ expts}, 0 \text{ evts}) &= 1 - e^{-n\,\lambda} \\ \lambda_u &= -\frac{\ln(1-P_u)}{n} \quad \text{with probability } P_u\,. \end{aligned}$$

$\boxed{\sum_i x_i \text{ \textbf{“large”}}}$ If the number of observed events is large (and the prior flat), the result will be normally distributed:

$$f(\lambda) \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda)\,.$$

Then, in this case it is more practical to use maximum likelihood methods than to make integrals (see Section 2.9). From the maximum of $f(\lambda)$, in correspondence of $\lambda = \lambda_m$, we easily get:

$$\mu_\lambda = \mathrm{E}(\lambda) \approx \lambda_m = \frac{\sum_{i=1}^{n} x_i}{n}\,,$$

and from the second derivative of $\ln f(\lambda)$ around the maximum:

$$\begin{aligned} \left.\frac{\partial^2 \ln f(\lambda)}{\partial \lambda^2}\right|_{\lambda_m} &= \frac{-n^2}{\sum_{i=1}^{n} x_i} \\ \sigma_\lambda^2 &\approx -\left(\left.\frac{\partial^2 \ln f(\lambda)}{\partial \lambda}\right|_{\lambda_m}\right)^{-1} = \frac{1}{n}\frac{\sum_{i=1}^{n} x_i}{n} \\ \sigma_\lambda &\approx \frac{\sqrt{\mu_\lambda}}{\sqrt{n}}\,. \end{aligned}$$

9.1.3 Combination of results: general case

The previous case is rather artificial and can be used, at most, to combine several measurements of the same experiment repeated n times, each with the same running time. In general, experiments differ in size, efficiency, and running time. A result on λ is no longer meaningful. The quantity which is independent from these contingent factors is the rate, related to λ by

$$r = \frac{\lambda}{\epsilon\, S\, \Delta T} = \frac{\lambda}{\mathcal{L}}\,,$$

where ϵ indicates the efficiency, S the generic 'size' (either area or volume, depending on whatever is relevant for the kind of detection) and ΔT the running time: all the factors have been grouped into a generic 'integrated luminosity' $\mathcal{L}$ which quantify the effective exposure of the experiment.

As seen in the previous case, the combined result can be achieved using Bayes' theorem iteratively, but now one has to pay attention to the fact that:

- the observable is Poisson distributed, and the each experiment can infer a λ parameter;
- the result on λ must be translated[2] into a result on r.

Starting from a prior on r (e.g. a monopole flux) and going from experiment 1 to n we have

- from $f_\circ(r)$ and $\mathcal{L}_1$ we get $f_\circ(\lambda)$; then, from the data we perform the inference on λ and then on r:

$$\begin{aligned} f_\circ(r)\,\&\,\mathcal{L}_1 &\rightarrow f_{\circ_1}(\lambda) \\ \mathrm{Data}_1 &\rightarrow f_1(\lambda \,|\, \mathrm{Data}_1, f_{\circ_1}(\lambda)) \\ &\rightarrow f_1(r \,|\, \mathrm{Data}_1, \mathcal{L}_1, f_\circ(r))\,. \end{aligned}$$

- The process is iterated for the second experiment:

$$\begin{aligned} f_1(r)\,\&\,\mathcal{L}_2 &\rightarrow f_{\circ_2}(\lambda) \\ \mathrm{Data}_2 &\rightarrow f_2(\lambda \,|\, \mathrm{Data}_2, f_{\circ_2}(\lambda)) \\ &\rightarrow f_2(r \,|\, \mathrm{Data}_2, \mathcal{L}_2, f_1(r)) \\ &\rightarrow f_2(r \,|\, (\mathrm{Data}_1, \mathcal{L}_1), (\mathrm{Data}_2, \mathcal{L}_2), f_\circ(r))\,, \end{aligned}$$

- and so on for all the experiments.

Lets us see in detail the case of null observation in all experiments ($\underline{x} = \underline{0}$) , starting from a uniform distribution.

Experiment 1:

$$\begin{aligned} f_1(\lambda \,|\, x_1 = 0) &= e^{-\lambda} \\ f_1(r \,|\, x_1 = 0) &= \mathcal{L}_1 e^{-\mathcal{L}_1 r} \qquad (9.7) \\ r_{u_1} &= \frac{-\ln 0.05}{\mathcal{L}_1} \text{ at 95\% probability}\,. \qquad (9.8) \end{aligned}$$

Experiment 2:

$$\begin{aligned} f_{\circ_2} &= \frac{\mathcal{L}_1}{\mathcal{L}_2} e^{-\frac{\mathcal{L}_1}{\mathcal{L}_2}\lambda} \\ f_2(\lambda \,|\, x_2 = 0) &\propto e^{-\lambda} \frac{\mathcal{L}_1}{\mathcal{L}_2} e^{-\frac{\mathcal{L}_1}{\mathcal{L}_2}\lambda} \\ &\propto e^{-\left(1+\frac{\mathcal{L}_1}{\mathcal{L}_2}\right)\lambda} \\ f_2(r \,|\, x_1 = x_2 = 0) &= (\mathcal{L}_1 + \mathcal{L}_2)\, e^{-(\mathcal{L}_1+\mathcal{L}_2)\, r}. \end{aligned}$$

[2] This two-step inference is not really needed, but it helps to follow the inferential flow. One could think more directly of

$$f(x \,|\, r, \mathcal{L}_i) = \frac{e^{-r\,\mathcal{L}_i}(r\,\mathcal{L}_i)^x}{x!}\,.$$

When the dependence between the two quantities is not linear, a two-step inference may cause trouble: see comments in Section 9.3.3.

Experiment n:

$$f_n(r \,|\, \underline{x} = \underline{0}, f_\circ(r) = k) = \sum_i \mathcal{L}_i e^{-\sum_i \mathcal{L}_i r} \,. \tag{9.9}$$

The final result is insensitive to the data grouping. As the intuition suggests, many experiments give the same result of a single experiment with equivalent luminosity. To get the upper limit, we calculate, as usual, the cumulative distribution and require a certain probability P_u for r to be below r_u [i.e. $P_u = P(r \le r_u)$]:

$$\begin{aligned} F_n(r \,|\, \underline{x} = \underline{0}, f_\circ(r) = k) &= 1 - e^{-\sum_i \mathcal{L}_i r} \\ r_u &= \frac{-\ln(1-P_u)}{\sum_i \mathcal{L}_i} \\ \frac{1}{r_u} &= \frac{-\sum_i \mathcal{L}_i}{\ln(1-P_u)} \\ &= \sum_i \frac{-\mathcal{L}_i}{\ln(1-P_u)} \\ &= \sum_i \frac{1}{r_{u_i}} \,, \end{aligned}$$

obtaining the following rule for the combination of upper limits on rates:

$$\frac{1}{r_u} = \sum_i \frac{1}{r_{u_i}} \,. \tag{9.10}$$

We have considered here only the case in which no background is expected, but it is not difficult to take background into account, following what has been said in Section 5.6.5.

9.1.4 Including systematic effects

A last interesting case is when there are systematic errors of unknown size in the detector performance. Independently of where systematic errors may enter, the final result will be an uncertainty on $\mathcal{L}$. In the most general case, the uncertainty can be described by a probability density function:

$$f(\mathcal{L}) = f(\mathcal{L} \,|\, \text{best knowledge on experiment}) \,.$$

For simplicity we analyse here only the case of a single experiment. In the case of many experiments, we only need to iterate the Bayesian inference, as has often been shown in these notes.

Following the general lines given in Section 2.10.3, the problem can be solved by considering the conditional probability, obtaining :

$$f(r \,|\, \text{Data}) = \int f(r \,|\, \text{Data}, \mathcal{L})\, f(\mathcal{L})\, \mathrm{d}\mathcal{L} \,. \tag{9.11}$$

The case of absolutely precise knowledge of $\mathcal{L}$ is recovered when $f(\mathcal{L})$ is a Dirac delta.

Let us treat in some more detail the case of null observation ($\underline{x} = \underline{0}$). For each possible value of $\mathcal{L}$ one has an exponential of expected value $1/\mathcal{L}$ [see Eq. (9.7)]. Each of the exponentials is weighted with $f(\mathcal{L})$. This means that, if $f(\mathcal{L})$ is rather symmetrical around its barycentre (expected value), in a first approximation the more and less steep exponentials will compensate,

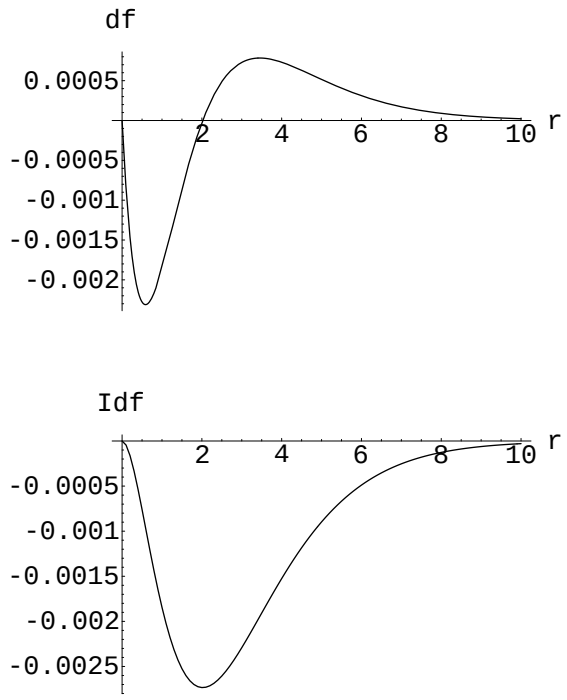

Figure 9.1: Inference on the rate of a process, with and without taking into account systematic effects: upper plot: difference between $f(r \,|\, x = 0, \mathcal{L} = 1.0 \pm 0.1)$ and $f(r \,|\, x = 0, \mathcal{L} = 1 \pm 0)$, using a normal distribution of $\mathcal{L}$; lower plot: integral of the difference, to give a direct idea of the variation of the upper limit.

and the result of integral (9.11) will be close to $f(r)$ calculated in the barycentre of $\mathcal{L}$, i.e. in its nominal value $\mathcal{L}_\circ$:

$$\begin{aligned} f(r \,|\, \text{Data}) &= \int f(r \,|\, \text{Data}, \mathcal{L})\, f(\mathcal{L})\, \mathrm{d}\mathcal{L} \approx f(r \,|\, \text{Data}, \mathcal{L}_\circ) \\ r_u \,|\, \text{Data} &\approx r_u \,|\, \text{Data}, \mathcal{L}_\circ \,. \end{aligned}$$

To make a numerical example, let us consider $\mathcal{L} = 1.0 \pm 0.1$ (arbitrary units), with $f(\mathcal{L})$ following a normal distribution. The upper plot of Fig. 9.1 shows the difference between $f(r \,|\, \text{Data})$ calculated applying Eq. (9.11) and the result obtained with the nominal value $\mathcal{L}_\circ = 1$:

$$\begin{aligned} df &= f(r \,|\, x = 0, f(\mathcal{L})\,) - f(r \,|\, x = 0, \mathcal{L} = 1.0) && (9.12) \\ &= \int f(r \,|\, x = 0, \mathcal{L})\, f(\mathcal{L})\, \mathrm{d}\mathcal{L} - e^{-r} \,. && (9.13) \end{aligned}$$

d is negative up to $r \approx 2$, indicating that systematic errors normally distributed tend to increase the upper limit. But the size of the effect is very tiny, and depends on the probability level chosen for the upper limit. This can be seen better in the lower plot of Fig. 9.1, which shows the integral of the difference of the two functions. The maximum difference is for $r \approx 2$. As far as the upper limits are concerned, we obtain (the large number of — non-significatant—digits

is only to observe the behaviour in detail):

$$\begin{aligned} r_u(x=0, \mathcal{L}=1\pm 0, \text{ at } 90\%) &= 2.304 \\ r_u(x=0, \mathcal{L}=1.0\pm 0.1, \text{ at } 90\%) &= 2.330 \\ r_u(x=0, \mathcal{L}=1\pm 0, \text{ at } 95\%) &= 2.996 \\ r_u(x=0, \mathcal{L}=1.0\pm 0.1, \text{ at } 95\%) &= 3.042\,. \end{aligned}$$

An uncertainty of 10% due to systematics produces only a 1% variation of the limits.

To simplify the calculation (and also to get a feeling of what is going on) we can use some approximations.

1. Since the dependence of the upper limit of r from $1/\mathcal{L}$ is given by

$$r_u = \frac{-\ln(1-P_u)}{\mathcal{L}}\,,$$

the upper limit averaged with the belief on $\mathcal{L}$ is given by

$$r_u = -\ln(1-P_u)\,\mathrm{E}\left[\frac{1}{\mathcal{L}}\right] = \int \frac{1}{\mathcal{L}} f(\mathcal{L})\,\mathrm{d}\mathcal{L}\,.$$

We need to solve an integral simpler than in the previous case. For the above example of $\mathcal{L}=1.0\pm 0.1$ we obtain $r_u = 2.326$ at 90% and $r_u = 3.026$ at 95%.

2. Finally, as a real rough approximation, we can take into account the small asymmetry of r_u around the value obtained at the nominal value of $\mathcal{L}$ averaging the two values of $\mathcal{L}$ at $\pm\sigma_{\mathcal{L}}$ from $\mathcal{L}_\circ$:

$$\begin{aligned} r_u &\approx \frac{-\ln(1-P_u)}{2}\left(\frac{1}{\mathcal{L}_\circ-\sigma_{\mathcal{L}}}+\frac{1}{\mathcal{L}_\circ+\sigma_{\mathcal{L}}}\right) \\ &\approx \frac{-\ln(1-P_u)}{\mathcal{L}_\circ}\left(1+\frac{\sigma^2_{\mathcal{L}}}{\mathcal{L}^2_\circ}\right)\,. \end{aligned}$$

We obtain numerically identical results to the previous approximation.

The main conclusion is that the uncertainty due to systematics plays only a second-order role, and it can be neglected for all practical purposes. A second observation is that this uncertainty increases slightly the limits if $f(\mathcal{L})$ is distributed normally, but the effect could also be negative if the $f(\mathcal{L})$ is asymmetric with positive skewness.

As a more general remark, one should not forget that the upper limit has the meaning of an uncertainty and not of a value of quantity. Therefore, as nobody really cares about an uncertainty of 10 or 20% on the uncertainty, the same is true for upper/lower limits. At the per cent level it is mere numerology (I have calculated it at the 10^{-4} level just for mathematical curiosity).

9.1.5 Is there a signal?

There is an important remark to be made on the interpretation of the result: can we conclude from an upper limit that the searched for signal does not exist? Tacitly yes. But let us take the final distribution of λ for $x=0$ (with a uniform prior and neglecting systematic effects) and let us read the result in a complementary way:

$$P(\lambda \geq \lambda_L) = e^{-\lambda_L}\,.$$

We obtain, for example:

$$\begin{aligned} P(\lambda \geq 10^{-1}) &= 90\% \\ P(\lambda \geq 10^{-2}) &= 99\% \\ \cdots &\quad \cdots \end{aligned}$$

Since $P(\lambda = 0) = 0$, it seems that we are almost sure that there is a signal, although of very small size. The solution to this apparent paradox is to remember that the analysis was done assuming that a new signal existed and that we only wanted to infer its size from the observation, under this assumption. On the other hand, from the experimental result we cannot conclude that the signal does not exist.

For the purpose of these notes, we follow the good sense of physicists who, for reasons of economy and simplicity, tend not to believe in a new signal until there is strong evidence that it exists. However, to state with a number what 'strong evidence' means is rather subjective. For a more extensive discussion about this point see Ref. [25].

9.1.6 Signal and background: a *Mathematica* example

As a last application of the Poissonian model, let us make a numerical example of a counting measurement in the presence of background. To compare full and approximative results, let us choose a number large enough for the normal approximation to be reasonable. For example, we have observed 44 counts with an expected background of 28 counts. What can we tell about the signal? We solve the problem with the following *Mathematica* code[3] applied to the formulae of Section 5.6.5 (s stands for λ_s and b for $\lambda_{B_\circ}$):

```
(*******************************************************)
ClearAll["Global‘*"]

f = (Exp[-s]*(b0+s)^x)/(x!Sum[b0^i/i!, {i, 0, x}])

x=44;
b0=28;

m = NIntegrate[s*f, {s, 0, 1E^6}]
sigma = Sqrt[NIntegrate[s^2*f, {s, 0, 1E^6}] - m^2]

Plot[f, {s, 0, 50}, AxesLabel->{s, "f"}]

fd1=D[Log[f], s];
fd2= D[fd1, s];

res=FindMinimum[-f, {s,m}];
smax = res[[2]]
sigma2=1/Sqrt[-(fd2 /. res[[2]])]
(*******************************************************)
```

The code evaluates and plots the final distribution of λ_s obtained from a uniform prior [formula (5.88)] and calculates:

- the prevision $\mathrm{E}(\lambda_S)$

$$\mathtt{m} \equiv \mathrm{E}(\lambda_S) = 17.0\,;$$

[3] If you are interested in Bayesian analysis with *Mathematica* you may take a look at Refs. [81] and [82] (I take for responsibility on the quality of the products, as I have never used them).

- the standard deviation $\sigma(\lambda_S)$:

$$\texttt{sigma} \equiv \sigma(\lambda_S) = 6.7\,;$$

- the mode λ_{S_m}:

$$\texttt{smax} \equiv \lambda_{S_m} = 16.0\,;$$

- the approximated standard deviation calculated from the shape of the final distribution around the mode (see Section 2.9):

$$\texttt{sigma2} = 6.6\,.$$

The resulting probability density function for the signal is shown in Fig. 9.2.

The approximate, but still Bayesian, reasoning to get the same result is as follows.

1. Given this status of information, the certain quantities are:
   - The average value of the background: $\lambda_{B_\circ} = 28\pm \approx 0$;
   - The observation $x = 44$ (it does not even make sense to write ± 0: 44 is 44 !).
2. Instead, we are uncertain on the parameter λ of the Poissonian distribution responsible for the observed number of counts; we can infer [see (5.59) and (5.61)]

$$\lambda \approx x \pm \sqrt{x} = 44.0 \pm 6.6\,.$$

3. Since λ is due to the contribution of the signal and background, we have, finally:

$$\lambda_S = \lambda - \lambda_{B_\circ} = 16.0 \pm 6.6\,.$$

The last evaluation is an example of how Bayesian reasoning helps, independently of explicit use of Bayes' theorem. Nevertheless, these results are still conditioned by the assumption that the signal looked for exists. In fact, Fig. 9.2 does not really prove, from a logical point of view, that the signal does exist, although the distribution seems so nicely separated from zero (see also Ref. [25]).

9.2 Unbiased results

In the Bayesian approach there is a natural way of giving results in an unbiased way, so that everyone may draw his own scientific conclusion depending on his prior beliefs. One can simply present likelihoods or, for convenience, ratios of likelihoods (Bayes' factors, see Sections 3.5 and 8.8). Some remarks are needed in order not to give the impression that, at the end of this long story, we have not just ended up at likelihood methods.

- First, without priors, the likelihoods cannot be turned into probabilities of the values of physics quantities or of probabilities of hypotheses. Even the 'mathematically harmless' uniform distribution, which gets simplified in Bayes' formula, does its important job. For this reason publishing only likelihoods does not mean publishing unbiased conclusions, but rather publishing no conclusions! Hence, one is not allowed to use this 'result' for uncertainty propagation, as it has no uncertainty meaning.

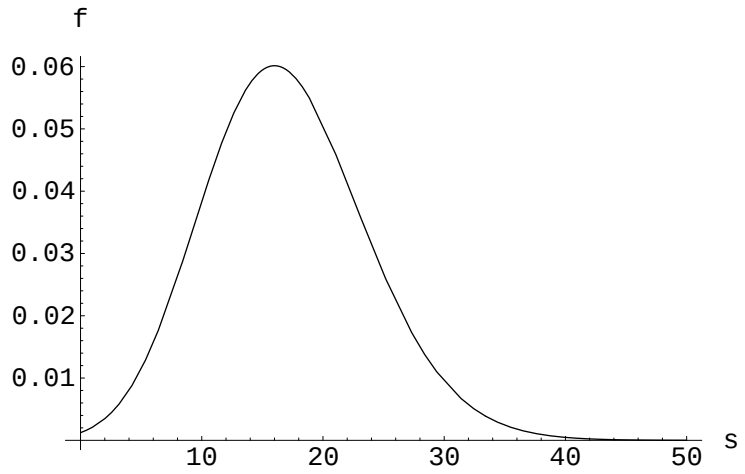

Figure 9.2: Final distribution for the signal ($s \equiv \lambda_S$), having observed 44 counts, with an expected background of 28 counts.

- This game of avoiding the priors can be done only at the final level of the analysis. Presuming priors can be avoided for each step is absurd, in the sense that the reader has no way of completely redoing the analysis by plugging in his preferred priors at all the crucial points. If the experimenter refrains to choose a prior, but, nevertheless, he goes on in the steps of the analysis, he is in fact using uniform priors in all inferences. This may be absolutely reasonable but one has to be aware of what one is doing. If instead there are reasons for using non-uniform priors in some parts of the analysis, as his past experience suggested, the experimenter should not feel guilty. There are so many subjective and even really arbitrary ingredients in a complex analysis that we must admit, if we believe somebody's results and we use his conclusions as if they were our conclusions, it is simply because we trust him. So we are confident that his knowledge of the detector and of the measurement is superior to ours and this justifies his choices. As a matter of fact, the choice of priors is insignificant compared with all the possible choices of a complicated experiment.

- The likelihoods are probabilities of observables given a particular hypothesis. Also their evaluation has subjective (and arbitrary) contributions. Sticking to the idealistic position of providing only objective data is equivalent to stagnating research.

Having clarified these points, let us look at two typical cases.

Classifying hypotheses. In the case of a discrete number of hypotheses, the proper quantities to report are the likelihoods of the data for each hypothesis

$$P(\text{data}\,|\,H_i)\,,$$

or Bayes' factor for any of the couples

$$\frac{P(\text{data}\,|\,H_i)}{P(\text{data}\,|\,H_j)}\,.$$

On the other hand, the likelihood for a given hypothesis alone, e.g. $P(\text{data}\,|\,H_\circ)$, does not help the reader to form his idea on the hypothesis, nor on alternatives (see also Section

8.8). Therefore, if a collaboration publishes experimental evidence against the Standard Model, suggesting some kind of explanation in terms of a new effect, it should report the likelihoods for both hypotheses. (See also 5th bullet of Section 8.8 in the case of Gaussian likelihood).

Values of quantities. In this case the likelihoods are summarized by the likelihood function $f(\text{data}\,|\,\mu)$. In this case one may also calculate Bayes' factors between any pair of values

$$\frac{f(\text{data}\,|\,\mu_i)}{f(\text{data}\,|\,\mu_j)}\,.$$

This can be interesting if only a discrete number of solutions are admissible.

When one publishes a likelihood function this should be clearly stated. Otherwise the temptation to turn $f(\text{data}\,|\,\mu)$ into $f(\mu\,|\,\text{data})$ is really strong. In fact, taking the example of the neutrino mass of Section 1.7, the formula

$$\frac{1}{\sqrt{2\,\pi}\,2}\,e^{-\frac{(-4-\mu)^2}{8}}$$

(with mass in eV) can easily be considered as if it were a result for μ:

$$\frac{1}{\sqrt{2\,\pi}\,2}\,e^{-\frac{[\mu-(-4)]^2}{8}}$$

and conclude that $\mu_\nu = -4 \pm 2\,\text{eV}$.

After having criticized this way of publishing the data for the second time, I try in the next section to encourage this way of presenting the result, on condition that one is well aware of what one is writing.

9.2.1 Uniform prior and fictitious quantities

Let us consider n independent data sets, or experiments, each of which gives information on the quantity μ. For each data set there is a likelihood

$$f_i(\text{data}_i\,|\,\mu)\,.$$

Each data set gives, by itself, the following information:

$$f(\mu\,|\,\text{data}_i) \quad\propto\quad f_i(\text{data}_i\,|\,\mu)\cdot f_\circ(\mu)\,. \tag{9.14}$$

The global inference is obtained using Bayes' theorem iteratively:

$$f(\mu\,|\,\bigcup_i \text{data}_i) \quad\propto\quad \Pi_i f_i(\text{data}_i\,|\,\mu)\cdot f_\circ(\mu)\,. \tag{9.15}$$

We may use, as a formal tool, a ficticious inference $\tilde{f}(\mu)$ using for each data set a uniform prior in the range $-\infty < \mu + \infty$:

$$\tilde{f}_i(\mu\,|\,\text{data}_i) \quad\propto\quad f_i(\text{data}_i\,|\,\mu)\cdot k\,. \tag{9.16}$$

This allows us to rewrite

$$f(\mu\,|\,\bigcup_i \text{data}_i) \quad\propto\quad \Pi_i \tilde{f}_i(\mu\,|\,\text{data}_i)\cdot f_\circ(\mu)\,.$$

This stratagem has the advantage that one can report 'pseudoresults' on fictitious quantities which, in the case of Gaussian likelihoods, may be combined according to the usual formula of the average with the inverse of the variances (see Section 5.4.2). They can be transformed, finally, into the physical result using the physical prior $f_\circ(\mu)$. It is important to state the procedure clearly and, if possible, to indicate the fictitious quantity with different symbols. For example, the result of the problem of Section 5.4.3 can be reported in the following way:

> "From the observed value of –5.4 eV and the knowledge of the likelihood, described by a normal distribution centred in the true value of the mass with $\sigma = 3.3\,\mathrm{eV}$ independent of the mass, we get a fictitious mass of
>
> $$\tilde{m}_\nu = -5.4 \pm 3.3\,\mathrm{eV}\,,$$
>
> where 'fictitious' indicates a hypothetical mass which could assume any real number with uniform distribution. Assuming the more physical hypothesis $f_\circ(m_\nu) \geq 0$ yields to ... (see figure ...), from which follows a 95% upper limit of 3.9 eV."

The conclusion of this section is that the uniform prior is a convenient prior for many purposes:

- it produces results very similar to those obtainable using the rational priors of those who have done the experiment, as shown in many of the examples given in these notes (see, for example, Section 5.4.3);
- it allows easy combination of data and a physics motivated prior can be added at the end;
- there is no problem of 'double counting' the same prior, as would happen if several experimenters were to use the same non-uniform prior to infer the same quantity from different data.

The problem of presenting unbiased results in frontier measurements is also discussed in Refs. [26], [25], [83] and [84].

9.3 Constraining the mass of a hypothetical new particle: analysis strategy on a toy model

As a last example of an application, let us consider a case which somehow reminds one of the current effort to reduce the uncertainty of the mass of the Higgs particle. Since I don't have access to the original data, and I don't want this exercise to be considered as any a kind of claim, I will just invent the rules of the game.[4] So, physics data and results will be imaginary, but the inferential procedure will be performed according to what I consider to be the proper way of doing it.

9.3.1 The rules of the game

The hypothetical world of this analysis is:

[4] This section is intentionally pedagogical. An analysis using the best physical assumptions can be found in Ref. [26]. Indeed, this analysis follows the strategy outlined here, with some variations introduced to match the information available in the real situation.

- three experiments (A, B and C) took data in an e^+e^- collider, at different energies and with different sensitivity to the H particle production ('H stands for hypothetical...'). The experiments reported 0 candidates and no background was expected (this is a minor approximation to simplify the formulae: we have seen how the background and its uncertainty may be treated).
- The beam energy was 0.09 and 0.1 in arbitrary units (you may think of TeV) and the kinematical factor which suppresses the production near threshold (and eventually takes into account efficiencies, tagging, etc.) is chosen somehow 'arbitrarily' to be β^3 factor, where β is the velocity of the pair produced particles.
- Cross-section and integrated luminosity are summarized into a sensitivity factor k, such that the expected number of events is

$$\lambda = k\,\beta^3 = k\,\left(1 - \frac{m^2}{E_b^2}\right)^{3/2} .$$

- We also have other pieces of information on H: two indirect determinations are characterized by a Gaussian likelihood, and each of them would allow a Gaussian determination of the mass, if one considered that this could be uniformly distributed from $-\infty$ to $+\infty$ (see Section 9.2).
- The five datasets are considered to be independent.
- The prior of the scientific community about the value of the mass has changed in recent years, due not only to negative results, but also to theoretical progress:
  - essentially, once there was uncertainty even in the order of magnitude, i.e. $f(\ln m) \approx k$, yielding $f(m) \approx 1/m$; as a conservative position, one could still stick to this position;
  - at present, many think that $\mathcal{O}(m) \approx 0.1$–$0.2$; this state of uncertainty can be modelled by a uniform distribution over the range of interest.

9.3.2 Analysis of experiment A

A has been run at a c.m. energy of 2×0.09, with sensitivity factor 20. It has observed 0 H candidate events. One can proceed in two ways, one correct and the other wrong. Let us start with the wrong one.

9.3.3 Naïve procedure

We have learned that we can infer

$$f(\lambda \,|\, x = 0) = e^{-\lambda}$$

and, from the relation $\lambda = k\,(1 - m^2/E_b^2)^{3/2}$ we get

$$f(m) = 3\,k\,\frac{m}{E_b^2}\left(1 - \frac{m^2}{E_b^2}\right)^{\frac{1}{2}} \exp\left[-k\,\left(1 - \frac{m^2}{E_b^2}\right)^{\frac{3}{2}}\right], \tag{9.17}$$

which is clearly wrong, since it goes to 0 for $m \to E_b$. Figure 9.3 shows the plot of $f(m)$, obtained with the following *Mathematica* code:

```
(*******************************************************)
ClearAll["Global`*"]

(* Experiment A has been run at beam energy 0.09, with
sensitivity factor k=20, threshold function beta^3,
and has observed 0 events.*)

ka=20
eba=0.09

v=Sqrt[1-(m/eba)^2]
lambda= ka*v^3

(* f(m) obtained from f(lambda)=Exp[-lambda] by lambda=k beta^3
   (threshold factor) using p.d.f transformation *)

fl=Exp[-lambda]
J=Abs[D[lambda, m]]
fm=fl*J

(* check normalization and plot *)

NIntegrate[fm, {m, 0, eba}]
Plot[fm, {m, 0.06, eba}, AxesLabel -> {m, f}]

(* Strange result: try to figure out the reason! *)
(*******************************************************)
```

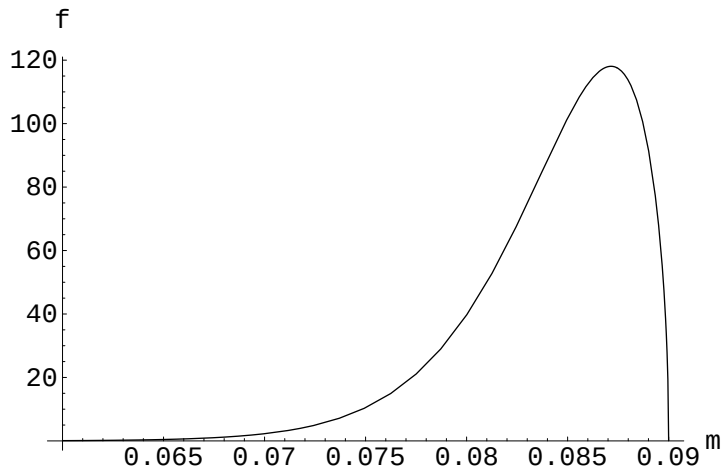

Figure 9.3: Inference on the mass of the hypothetical particle H with the information of experiment A, obtained from the intermediate inference on λ assuming a uniform prior on this quantity. The result looks very strange.

The result is against our initial rational belief and we refuse to accept it. The origin of this strange behaviour is due to the term $\sqrt{1-m^2/E_b^2}$ in (9.17) that comes directly from the Jacobian of the transformation and, indirectly from having assumed a prior uniform in λ. To solve the problem we have to change prior. But this seems to be cheating. Should not the prior come before? How can we change it after we have seen the final distribution?

We should not confuse what we think with how we model it. The intuitive prior is on the mass values, and this prior should be flat (at least at this stage of the analysis, as discussed in Section 9.2). Unfortunately, what is flat in λ is not flat in m, and vice versa. This problem has been discussed in Section 5.3. In fact, it is not really a problem of probability, but of extrapolating intuitive probability (which is at the basis of subjective probability and only deals with discrete numbers) to continuous variables. This is the price we pay for using all the mathematical tools of differential calculus. But one has to be very careful in formulating the problem. If one wants to get rid of these problems, one may discretize λ and m in a way which is consistent with to the experimental resolution. If we discretize, a flat distribution in m is mapped to a flat distribution in λ. And the problems caused by the Jacobian go away with the Jacobian itself, at the expense of some complications in computation.

9.3.4 Correct procedure

In order to solve the problem consistently with our beliefs, we have to avoid the intermediate inference[5] on λ, and write prior and likelihood directly in terms of m:

$$f(m \,|\, x=0) \propto \exp\left[-k\,\left(1-\frac{m^2}{E_b^2}\right)^{\frac{3}{2}}\right] \cdot f_\circ(m)\,, \tag{9.18}$$

with $f_\circ(m) = \textit{constant}$. Let us do it again with *Mathematica*:

```
(*********************************************************)
(* Now let's do it right: *)

lik=Exp[-lambda]
norm=NIntegrate[lik, {m, 0, eba}]

(* fa(m) is the final distribution from experiment A,
   under the condition that m < eba *)

fa=lik/norm
Plot[fa, {m, 0.06, eba}, AxesLabel -> {m, f}]
(*********************************************************)
```

The final distribution is shown in Fig. 9.4. It is now reasonable and consistent with the expectations: The values of mass which are less believable are those which could have been produced easier, given the kinematics. From $f(m \,|\, x=0)$ we can calculate several results, for example a 95% upper limit, the average and the standard deviation:

```
(*********************************************************)
NIntegrate[fa, {m, 0, 0.0782}]
ava = NIntegrate[m*fa, {m, 0, eba}]
stda = Sqrt[NIntegrate[m*fa, {m, 0, eba}] - ava^2]
(*********************************************************)
```

We get:

$$m \;>\; 0.0782 \text{ with } 95\% \text{ probability} \tag{9.19}$$

$$\mathrm{E}(m) \;=\; 0.0856 \tag{9.20}$$

$$\sigma(m) \;=\; 0.0038\,. \tag{9.21}$$

[5] A two-step inference was shown in Section 9.1.3 for the case of monopole search. There there was no problem because λ and r are linearly related.

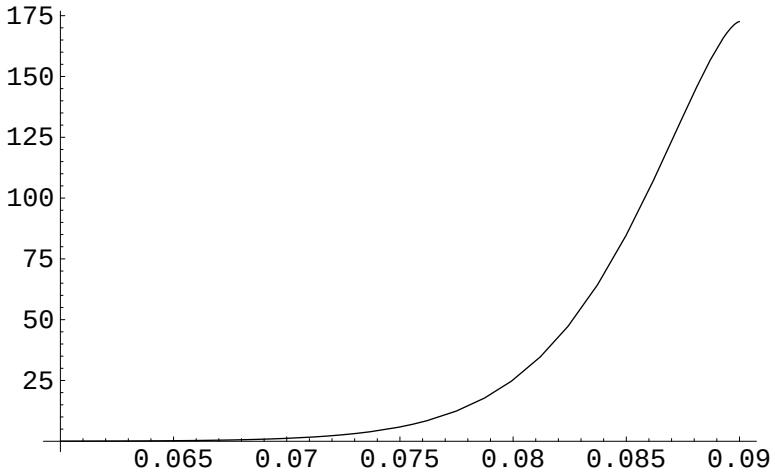

Figure 9.4: Inference on m obtained from a direct inference on m, starting from a uniform prior in this quantity.

9.3.5 Interpretation of the results

At this point we need to discuss the results achieved so far, to understand what they mean, and what they don't.

- Can we say that we believe that m is above 0.0782 with 95% probability and below with 5%? Would you bet 5 \$ that m is below 0.0782 with the hope of winning 100 \$ if it turns out to be the case?
- Can we say we believe that we have done a precise measurement of m, since it looks like $m = 0.0856 \pm 0.0038$?
- How can we say that $m = 0.0856 \pm 0.0038$ and speak about a lower bound?

In particular, the first statement gives the impression that we can say something about the mass even if H was too heavy to be produced. In general, a statement like

> *"A 95% confidence level lower bound of* $77.5\,GeV/c^2$ *is obtained for the mass of the Standard Model Higgs boson."* [85]

may be misleading, because it transmits information which is inconsistent with the experimental observation. The interpretation of the result (9.19) is limited to

$$P(m > 0.0782 \,|\, 0 \leq m \leq 0.09) \;\; = \;\; 0.95\,, \tag{9.22}$$

as may be understood intuitively and will be shown in a while (there is also the condition of the uniform prior, but at this level it is irrelevant). So, given this status of information, I could bet 1:19 that the m is below 0.0782, but only on condition that the bet is invalidated if m turns out to be greater than the beam energy (see Section 3.4.2). Otherwise, I would choose the other direction (19:1 on '$m > 0.0782$') without hesitation (and wish fervently that somebody accepts my bet ...).

What are our rational beliefs on m, on the basis of experiment A, releasing the condition $\leq m \leq E_b$? The data cannot help much because there is no experimental sensitivity, and the conclusions depend essentially on the priors.

To summarize, the result of the inference is:

$\mathbf{m < E_b}$**:** $P(m > 0.0782) = 0.95$; $m = 0.0856 \pm 0.038$, etc. ;

$\mathbf{m \geq E_b}$**:** "HIC SUNT LEONES"[6]

As a final remark on the presentation of the result, I would like to comment on the three significative digits with which the result on the 'conditional lower bound' has been given. For the sake of the exercise the mass bound has been evaluated from the condition (9.22). But does it really matter if the limit is 0.0782, rather than 0.0780, or 0.0800? As stated in Sections 5.4.3 and 9.1.1, the limits have to be considered in the same way as the uncertainty. Nobody cares if the uncertainty of the uncertainty is 10 or 20%, and nobody would redo a MACRO-like experiment to lower the monopole limit by 20%. Simply translating this argument to the case under study, it may give the impression that one significant digit would be enough (0.08), but this is not true, if we stick to presenting the result under the condition that m is smaller than E_b. In fact, what really matters, is not the absolute mass, but the mass difference with respect to the kinematical limit. If the experiment ran with infinite statistics and found 'nothing', there is no interest in providing a detailed study for the limit: it will be exactly the same as the kinematical limit. Therefore, the interesting piece of information that the experimenter should provide is how far the lower bound is from the kinematical limit, i.e. what really matters is not the absolute mass scale, but rather the mass difference. In our case we have

$$\Delta m = E_b - \textit{lower bound} = 0.0118 \rightarrow 0.012\,. \tag{9.23}$$

Two digits for this number are enough (or even only one, if the first were greater that 5) and the value of the lower bound becomes[7]

$$m > 0.078 \text{ at } 95\%, \text{ if } 0 \leq m \leq 0.09\,.$$

9.3.6 Outside the sensitivity region

With the Bayesian method it is possible to trace the point in which an unstated condition has been introduced, and how to remove it, or how to take it into account. With the form of the likelihood used in (9.18) it was implicit that m should not exceed E_b. A more physically motivated likelihood should be:

$$f(x = 0 \,|\, m) = \begin{cases} \exp\left[-k\,\left(1 - \frac{m^2}{E_b^2}\right)^{3/2}\right] & \text{if} \quad 0 \leq m \leq E_b \\ 1 & \text{if} \quad m > E_b \end{cases} \tag{9.24}$$

Taking a uniform prior, we get the following posterior:

$$f(m \,|\, x = 0) = \begin{cases} \dfrac{\exp\left[-k\left(1-\frac{m^2}{E_b^2}\right)^{3/2}\right]}{\int_0^{E_b} \exp\left[-k\left(1-\frac{m^2}{E_b^2}\right)^{3/2}\right] \mathrm{d}m + (m_{max} - E_b)} & \text{if} \quad 0 \leq m \leq E_b \\ \dfrac{m_{max} - E_b}{\int_0^{E_b} \exp\left[-k\left(1-\frac{m^2}{E_b^2}\right)^{3/2}\right] \mathrm{d}m + (m_{max} - E_b)} & \text{if} \quad m > E_b \end{cases} \tag{9.25}$$

[6] *"Here are the lions"* is what the ancient Romans used to write on the parts of their maps representing unexplored regions.

[7] Numerologists may complain that this does not correspond to exactly 95%, but the same happens when a standard uncertainty is rounded to one or two digits and the probability level calculated from the rounded number may differ a lot from the nominal 68.3% calculated from the original value. But who cares?

where $(m_{max} - E_b)$ comes from the integral $\int_{E_b}^{m_{max}} 1 \cdot \mathrm{d}m$. So, we get our solution (9.18) for $m_{max} = E_b$. In general, the probability that $m \leq E_b$ is smaller than 1 and decreases for increasing m_{max}. For the parameters of experiment A the integral in the denominator is equal to 0.0058. Therefore, if, for example, $m_{max} = 3\,E_b$

$$\begin{aligned} P(m < E_b \,|\, x = 0, m_{max} = 3\,E_b) &= 2.7\% \\ P(m < 0.078 \,|\, x = 0, m_{max} = 3\,E_b) &= 0.13\%. \end{aligned}$$

There is another reasoning which leads to the same conclusion. At $m = E_b$ the detector has zero sensitivity. For this reason, in case of null observation, this values gets the maximum degree of belief. As far as larger values are concerned, the odds ratios with respect to $m = E_b$ must be invariant, since they are not influenced by the experimental observations, i.e.

$$\frac{f(m \,|\, x = 0)}{f(m = E_b \,|\, x = 0)} = \frac{f_\circ(m)}{f_\circ(m = E_b)} \qquad (m > E_b)\,. \tag{9.26}$$

Since we are using, for the moment, a uniform distribution, the condition gives:

$$f(m \,|\, x = 0) = f(m = E_b \,|\, x = 0) \quad (m > E_b)\,. \tag{9.27}$$

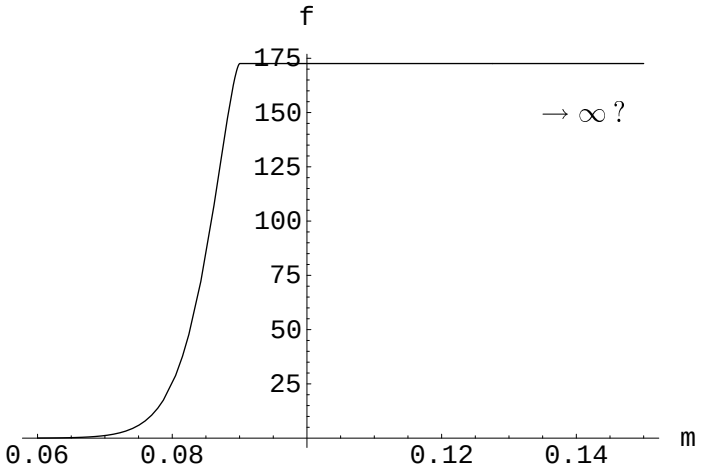

Figure 9.5: Result of the inference from experiment A, taking into account values of mass above the beam energy as well. These all have the same degree of belief and the normalization constant depends on the maximum value of m considered. Therefore the distribution is usually improper.

We easily get the result shown in Fig. 9.5 by this piece of *Mathematica* code:

```
(*******************************************************)
famax=fa/.m->eba
f2a=If[m<eba, fa, famax]

(* f2a(m) represents instead the (improper) distribution
   extended also for values larger that eba, in the light
   of a flat prior and of the Experiment A *)

Plot[f2a, {m,0.06,0.15}]
(*******************************************************)
```

The curve is extended on the right side up to a limit which cannot be determined by this experiment, it could virtually go to infinity. For this reason the ratio of probabilities

$$\frac{P(m < E_b)}{P(m \geq E_b)}$$

decreases (i.e. we tend to believe more strongly large mass values) but its exact value is not well defined. For this reason we leave the function 'open' on the right side and unnormalized. The normalization will be done when we can include other data which can provide an upper limit.

9.3.7 Including other experiments

Each of the other experiments are treated in exactly the same way. Comparing B and C it is interesting to see how the beam energy and the sensitivity factor contribute to constraining the mass. For reasons of space the plots are not shown. This is the rest of the *Mathematica* code to conclude the analysis:

```
(*******************************************************)
(* Experiment B has been run at beam energy 0.09, with
   sensitivity factor k=100,  threshold function beta^3, and
   has observed 0 events.*)

kb=100
ebb=0.09
v=Sqrt[1-(m/ebb)^2]
lambda= kb*v^3
lik=Exp[-lambda]
norm=NIntegrate[lik, {m, 0, ebb}]
fb=lik/norm
avb = NIntegrate[m*fb, {m, 0, ebb}]
Plot[fb, {m, 0.07, ebb}, PlotRange->{0, 600},
     AxesLabel -> {m, f}]

fbmax=fb/.m->ebb
f2b=If[m<ebb, fb, fbmax]
Plot[f2b, {m,0.07,0.15},  PlotRange->{0, 600},
     AxesLabel -> {m, f}]

(* The conclusions from A + B are, with and without the condition m<ebeam,
   respectively (remember that the latter is improper): *)

fcom1ab=fa*fb/NIntegrate[fa*fb, {m, 0, eba}]
avab = NIntegrate[m*fcom1ab, {m, 0, eba}]
Plot[fcom1ab, {m, 0.07, eba},  PlotRange->{0, 600},
     AxesLabel -> {m, f}]
fcom2ab=f2a*f2b

(* Experiment C has been run at beam energy 0.1, with sensitivity factor k=10,
   threshold function beta^3 and, has observed 0 events. *)

kc=10
ebc=0.1
v=Sqrt[1-(m/ebc)^2]
lambda= kc*v^3
lik=Exp[-lambda]
```

```
norm=NIntegrate[lik, {m, 0, ebc}]
fc=lik/norm
Plot[fc, {m, 0.07, ebc}, PlotRange->{0, 100},
     AxesLabel -> {m, f}]
avc = NIntegrate[m*fc, {m, 0, ebc}]
fcmax=fc/.m->(ebc-0.000001)
f2c=If[m<ebc, fc, fcmax]
Plot[f2c, {m,0.07,0.15},  PlotRange->{0, 100},
     AxesLabel -> {m, f}]
(*******************************************************)
```

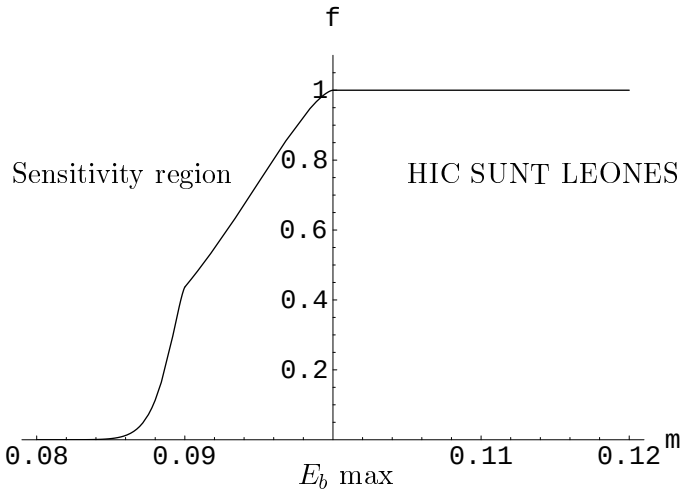

Figure 9.6: Final distribution (improper, see text) on m from experiments A, B and C. The curve has been arbitrarily rescaled to have the maximum of 1.

The combination of the result is done in the usual way, multiplying the likelihoods or the final p.d.f.'s, if these were obtained from a uniform distribution. We only see the combination of the three experiments, shown in Fig. 9.6. Finally, the indirect determinations are also included.

```
(*******************************************************)
(* Conclusions from A + B + C , with and without the condition m<ebeam,
   respectively (remember that the latter is improper): *)

fcom1abc=f2a*f2b*fc/NIntegrate[f2a*f2b*fc, {m, 0, ebc}]
avabc=NIntegrate[m*fcom1abc, {m, 0, ebc}]
Plot[fcom1abc, {m, 0.07, ebc},  PlotRange->{0, 150},
     AxesLabel -> {m, f}]
fcom2abc=f2a*f2b*f2c

(* Now we add independent determinations of m,
   deriving from normal likelihoods,
   and assuming uniform prior *)

g1=1/sigma1/(Sqrt[2*Pi])*Exp[-(m-mu1)^2/2/sigma1^2]
g2=1/sigma2/(Sqrt[2*Pi])*Exp[-(m-mu2)^2/2/sigma2^2]

mu1=0.09
sigma1=0.04
```

```
mu2=0.15
sigma2=0.04

(* The two overall (improper) priors may be a uniform,
   or 1/m, i.e.  flat in ln(m), to express initial
   uncertainty on the order of magnitude of m *)

p1=1
p2=1/m

final1=fcom2abc*g1*g2*p1/NIntegrate[fcom2abc*g1*g2*p1, {m, 0, 10}]
mean1=NIntegrate[m*final1, {m, 0, 10}]
std1=Sqrt[NIntegrate[m^2*final1, {m, 0, 10}]-mean1^2]
Plot[final1, {m, 0.0, 0.25},  PlotRange->{0, 20},
     AxesLabel -> {m, f}]

final2=fcom2abc*g1*g2*p2/NIntegrate[fcom2abc*g1*g2*p2, {m, 0, 10}]
mean2=NIntegrate[m*final2, {m, 0, 10}]
std2=Sqrt[NIntegrate[m^2*final2, {m, 0, 10}]-mean2^2]
Plot[final2, {m, 0.0, 0.25},  PlotRange->{0, 20},
     AxesLabel -> {m, f}]
(*******************************************************)
```

Finally, the two extra pieces of information enable us to constrain the mass also on the upper side and to arrive at a proper distribution (see Fig. 9.7), under the condition that H exists.

From the final distribution we can evaluate, as usual, all the quantities that we find interesting to summarize the result with a couple of numbers. For a more realistic analysis of this problem see Ref. [26].

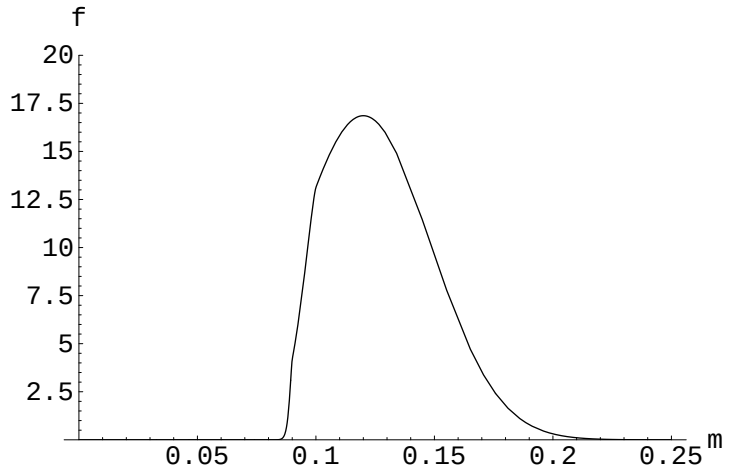

Figure 9.7: Final mass distribution using all five pieces of experimental information, and assuming uniform priors. The curve obtained from the prior $1/m$ does not differ substantially from this.