Practical statistics

Louis Lyons[a]
[a]Imperial College & U. of Oxford, UK

The emphasis in these lectures is very much on practical statistics, i.e. what you will be using as part of almost any data analysis procedure. The topics dealt with include a discussion of the Bayesian and frequentist approaches, a description of the statistical issues involved in searches for new physics, and explaining how covariance matrices help dealing with correlations.

This chapter should be cited as: Practical statistics, Louis Lyons, DOI: 10.23730/CYRSP-2026-001.47, in: Proceedings of the 2024 European School of High-Energy Physics, CERN Yellow Reports: School Proceedings, CERN-2026-001, DOI: 10.23730/CYRSP-2026-001, p.47.

1 Introduction

I gave three lectures. The first was on parameter determination via the likelihood and χ^2 approaches, with the latter also providing a measure of Goodness of Fit. These were slightly modified versions of similar lectures at the 2015 CLASHEP school in Ecuador [1]. The changes did not merit an update. The next lecture was about the Bayes and frequentist approaches to statistics, and also about statistical issues related to claims of discoveries of new physics. These appear as separate sections here. "Learning to love the covariance matrix" was addressed in the last lecture, and appears in Section 4.

Before we start, a general comment is in order. Our statistical procedures generally end up with us estimating something (e.g. the mass of the τ lepton with some uncertainty, the lower limit on the mass of a SUperSYmmetric particle, etc.). In addition to these **estimated** sensitivities, it is useful to quote the **expected** ones. (These can be obtained by Monte Carlo simulation of a large number of 'data' sets. Alternatively we can use a single 'data' set in which the assumed observations are exactly as predicted by the theoretical model; this is sometimes known as 'Asimov' data.) This can be used to check that the estimated sensitivity is not significantly different from the observed one. A side benefit is that if there are two competing experiments, and one has a better expected sensitivity and the other has a better observed one, each can claim to be 'better'.

I am a member of the CMS collaboration at CERN's Large Hadron Collider, and some of the examples presented here are from our results. So, apologies to

- ATLAS, who in most cases have similar analyses;
- Other High Energy Physics experiments, which also have many interesting results; and
- Astrophysics/Cosmology and other fields, for which much of the discussion here will hopefully be relevant.

I have tried to make the examples accessible to people from all fields represented at the 2024 European School of High Energy Physics.

2 Bayes and frequentist approaches to probability and statistics

These are two very different approaches to what probability is, and how statistical issues are to be addressed. Here we consider their application to parameter determination. There are also other methods of doing this, for example, the Likelihood and the Chi-squared approaches discussed in Ref. [1]. It is also

true that the frequentist and Bayesian approaches have use in other statistical procedures e.g. Hypothesis Testing (see Section 2.6) and Decision Theory (not discussed here) respectively.

The reason that it is possible to spend a lifetime measuring physical quantities (such as the mass of an object, the time difference between astrophysical events, etc.) without being aware of their differences is that in very simple cases they can give the same answer. For example, at the 90% level

$$0.21 < \text{Fraction of Dark Matter in Universe} < 0.25 \quad . \tag{1}$$

However, as we shall see at the end of section 2.3, even in this case of numerical agreement, Bayesians and frequentists differ profoundly as to the meaning of this statement.

In the above paragraph, 'very simple' means that the measurements are Gaussian distributed about the true values, and the value of the parameter can be anywhere between plus and minus infinity. However, in particle physics this is often not true, e.g. we are dealing with a small number of Poisson counts, rather than a Gaussian distribution; a possible small signal fraction cannot be negative, etc. So we have to deal with Bayesian-frequentist differences.

First, we consider their ideas about probability, and then their approaches to parameter determination.

2.1 Frequentist and Bayesian probabilities

For frequentists, probability is defined as the limit of the ratio of the number of 'successes' N_s to the number of independent trials N_t as N_t tends to infinity. For example, as can be determined either by experiment or by a simple symmetry argument, the probability of a throw of a fair dice coming out as 2 is 1/6.

Because of the need for a repeated series of trials, frequentists would not ascribe a probability to a unique event or to the value of a physical constant. This eliminates questions like:

- What is the probability that Prince William will become King of Great Britain in 2026?
- How probable is it that the masses of the top quark and Higgs boson are such that the Universe is metastable, and could decay away?
- It might have rained in London yesterday. What is the probability that it did?

In contrast, Bayesians define probability in terms of personal belief, based on an individual's experience, and so can vary from person to person. Thus Bayesians would accept that a physicist and a random man in the street might well assign different probabilities to a particular scientist winning the next Physics Nobel Prize. Similarly assigning a probability to the Hubble constant being below 70 km/sec/Mpc would be an allowed expression of a person's degree of individual Bayesian belief.

As that all sounds very vague, how does a Bayesian assign a numerical value to probability in any particular case? The answer is via a fair bet. If a Bayesian believes that a certain event occurring is 20%, he or she should be prepared to accept a bet at odds 4 to 1 for it to happen or at 1 to 4 that it does not. Any misassignment of probabilities could result in an expected loss of money.

An interesting example of Bayesian probability is provided by Bayes' portrait (see Fig. 1). Every-

one shows the same picture of Bayes simply because this is the only one that exists. But historians have cast doubt on whether this is really an image of Bayes because the clothing looks anachronistic. So we could ask "What is the probability that this is really Bayes?" This clearly is not a frequentist probability, so the question about the probability of it being Bayes is indeed a Bayesian probability.

Fig. 1: The picture of Bayes. However, there are doubts as to whether this really is Bayes.

2.2 Bayesian parameter determination

For parameter estimation, Bayesians make use of Bayes' theorem, which in general terms states that the probability $P(A;B)$ of A happening, given that B has occurred, is given by

$$P(A;B) = P(B;A) * P(A)/P(B) \quad . \tag{2}$$

This theorem is acceptable even to frequentists, provided that the probabilities involved are frequentists' probabilities. Their disapproval comes when Bayesians substitute the parameter value μ for A, and data d for B, i.e.

$$P(\mu;d) = P(d;\mu) * P(\mu)/P(d) \quad . \tag{3}$$

Here, for a given set of data $P(d;\mu)$ is the likelihood function; $P(\mu)$ is the Bayesian Prior; $P(d)$ is a constant normalising factor; and $P(\mu;d)$ is the Bayesian Posterior i.e. what we know about the parameter μ after we obtain our data.

Thus the Bayesian approach is a method of updating our knowledge about the parameter from before we performed our experiment (as encapsulated by the prior), to what we know about it after, i.e. the Posterior combines information from our data and from our prior.

Bayesians would be happy to regard the Posterior Probability as the result of their analysis for the parameter of interest. They would, however, be prepared to summarise this by a range that contained a specified fraction (e.g. 90%) of the Posterior distribution. Options shown in Fig. 2 include:

– A central interval, with 5% of the posterior below the range and a further 5% above it.

– The shortest 90% interval, which contains the highest posterior probabilities.
– An upper limit, above which there is a 10% tail.
– A lower limit, below which there is a 10% tail.

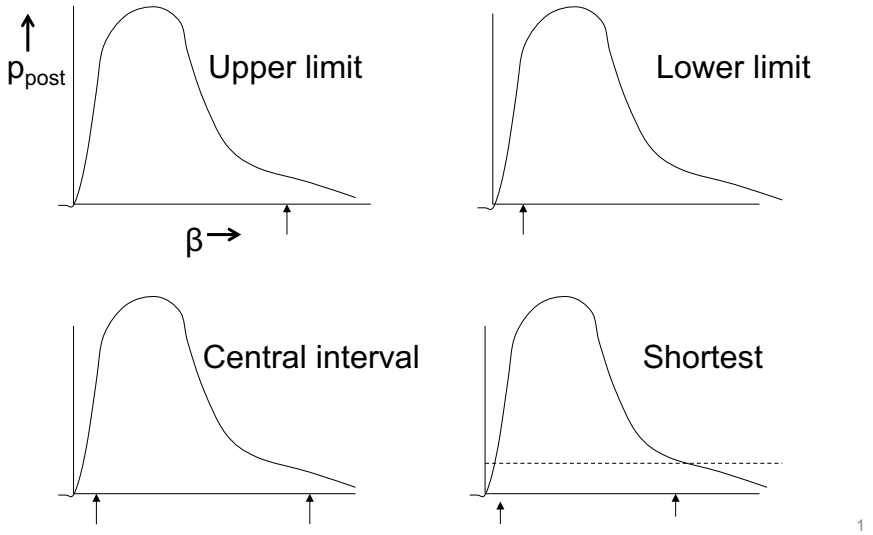


Fig. 2: Deriving various 90% intervals from a Bayesian posterior $p_{\rm post}$ for a parameter β: Upper limit, lower limit, central interval, and shortest.

Frequentists object to this Bayesian approach on two grounds. First, they would maintain that the probabilities of parameter values are not acceptable (frequentist) probabilities. Then they would object to the result (i.e. the Bayesian posterior) for the parameter of interest μ being dependent on the assumed functional form of the prior. This might be acceptable when previous measurement(s) of μ exist, but less so when you are performing a new measurement and there is no relevant earlier information. It is tempting but incorrect to think that a uniform prior $P(\mu) = \text{constant}$ would be a good way of expressing prior ignorance. This is because we have to decide whether we are ignorant about μ, or μ^2, or $\ln \mu$, etc; and priors that are constant in each of these functional forms are different from each other, and would thus yield different posteriors. It is therefore important in a Bayesian approach to perform a sensitivity analysis by using a range of plausible priors.

Bayesian posteriors will, however, be insensitive to the choice of prior when the data overpowers the prior. An example of this is the very accurate determination of the mass of the Z boson by the four experiments at CERN's LEP Collider [2]; the result had an uncertainty of 1 part in 10^5. Any reasonable prior would be essentially constant over the relevant narrow mass range, and so the posterior is insensitive to the prior.

This contrasts with searches for new physics when only an upper limit is presented. There the limit can be very sensitive to the choice of the prior [3].

2.3 Neyman construction

The Neyman construction is a frequentist procedure for finding a confidence limit for the parameter of interest, which is guaranteed to produce ranges with the correct coverage[1].

The construction produces a band on a plot of the parameter of interest μ against the data x, see Fig. 3. A specific example that I keep in mind is that the parameter of interest could be the temperature T of the fusion region at the centre of the sun, and the data could be the solar neutrino flux ϕ as estimated by a month's data from a large underground solar neutrino detector.

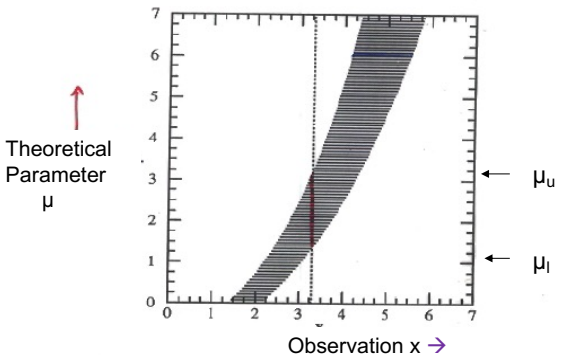


Fig. 3: Neyman Construction. To make the confidence belt (shown shaded), at a given value of μ the probability density function $p(x;\mu)$ is used to construct a region which contains the specified probability level. This is repeated for all μ. The vertical line is drawn at the observed data value. The confidence region for μ is from μ_l to μ_u, where the vertical line intersects the boundaries of the confidence belt.

The band is produced by first selecting a specific value of μ. Then we calculate the probability density for obtaining all possible values of the data; this is the probability density function (pdf). This is nontrivial to do in practice as it requires detailed knowledge of the nuclear processes within the sun; of the detector in all its aspects; what happens to neutrinos as they pass through the Sun, through space towards the Earth, and through the matter of the earth; etc. Then, for example, a 90% region of the pdf is selected. By repeating this for all possible values of μ, the confidence belt is built up.

Finally a vertical line is drawn at the actual data result x_0; the distribution as a function of μ along this line is the likelihood. The confidence interval for μ is defined as the region along this line which is within the confidence belt, i.e. μ_l to μ_u in Fig. 3. In this procedure, no prior is used.

It is important to be aware that such a confidence region is not the 90% probability region for the true value of the parameter of interest, but is the set of parameter values for which the data is likely. Also by itself it provides no information about different values of μ within the confidence region.

So the conclusion of a frequentist parameter estimation is a statement such as in Eq. (1). But it

[1] When the data is discrete, such as in a Poisson counting measurement, there are inevitably jumps in the coverage (see Problem 4 in the Appendix). In order to avoid undesirable undercoverage, it is thus necessary to have overcoverage for some values of the parameters.

is possible that Bayesians could make a similar statement, maybe even with the same numerical values. However, there is a great difference in how they interpret their statements.

For frequentists, this is a statement about the coverage of the ranges produced by an ensemble of such measurements. That is, in what fraction of these procedures would the ranges include the true value of the parameter? It is important to note that this is not a statement about the particular result you get using your data. So here the ranges (0.21 to 0.25 in Eq. (1)) are to be regarded as random variables to which the probability statement applies. In particular, frequentists do not make probability statements about physical parameters.

In contrast, Bayesians regard Eq. (1) as a statement that 90% of the Bayesian posterior for the parameter lies between 0.21 and 0.25. These values are regarded as fixed by the data we have, and Bayesians do not want to consider what might have happened if we were to repeat the experiment.

2.3.1 Feldman–Cousins approach

In the standard Neyman approach, there is freedom of how the 90% confidence region is selected; this is in analogy with the freedom of choice for Bayesian intervals described earlier. Thus it could be a central choice, where for each μ there is 5% of the pdf on either side. Alternatively, it could have all the 90% above some particular x, and nothing below; that would provide upper limits for μ. And there are more possibilities.

Feldman and Cousins (FC) exploit this arbitrariness. For each μ, they use a likelihood ratio as an ordering rule for the values of the data x to add to the selected region until it builds up to the chosen 90%. For more details on how this works, consult the original article, see Ref. [4].

Some of the advantages of the FC approach are:

- Coverage: Because it uses a Neyman construction, FC intervals are guaranteed not to undercover.
- Unified: For a given data set, the FC procedure decides whether the resulting interval will be two-sided or one-sided (i.e. just a limit). In other Neyman construction procedures, the analyser has to choose which type of interval to produce.
- Empty intervals: Other procedures can result for some data sets in no interval for the parameter of interest. This is regarded by most physicists as unfortunate. It is almost always avoided by FC intervals. Thus Fig. 4 shows the FC confidence band, as compared with the central Neyman one for a measurement x of a non-negative physical parameter μ with x being Gaussian distributed about μ with RMS = 1; the bands differ at small μ. For a measurement $x = -2$, the central Neyman approach fails to find any value of μ, but FC returns an upper limit of ~ 0.4 .
- Multi-dimensional data: The FC ordering rule makes it possible to deal with multi-dimensional data; this is often not so for other approaches.

However, computational problems make it difficult to use the Neyman construction (including the FC method) when the number of parameters of interest is more than 2.

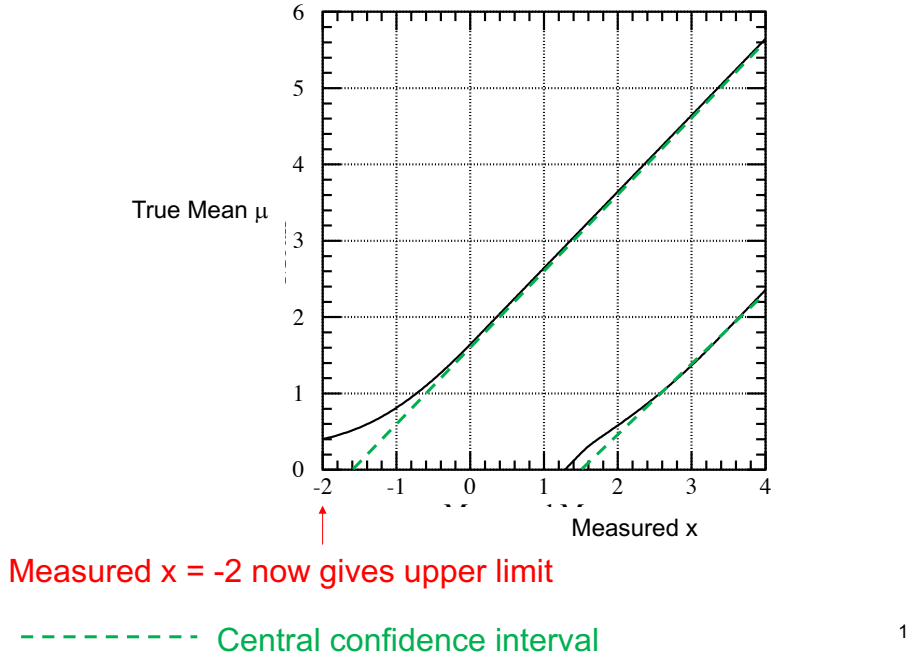


Fig. 4: Confidence bands for the Gaussian mean μ when the measurement x is Gaussian distributed about μ with variance of unity. The two black solid curves bound the 90% Feldman–Cousins region. The boundaries of the central Neyman region are instead the two green straight dashed lines.

2.4 Detailed example: Lifetime

We imagine a situation where we estimate the lifetime τ of some particle by observing the times t_i at which N of them decay. We ignore many experimental complications (e.g. background, experimental resolution, acceptance as a function of t, etc.) and assume that the pdf for t at a particular τ is

$$dy/dt = (1/\tau)\exp(-t/\tau) \quad . \tag{4}$$

2.4.1 Likelihood approach

Simple algebra yields the result that the unbinned likelihood $L(\tau)$ is given by

$$\ln L(\tau) = N(-\ln\tau - \bar{t}/\tau) \quad , \tag{5}$$

where $\bar{t}$ is the mean of the observed decay times.

It is worth noting that the unbinned likelihood depends on the data only through the **mean** of the decay times, and is independent of the **distribution** of the individual t_i. Thus a data set with all N decays occurring at the same time τ would have the identical unbinned likelihood as another data set where the N individual decay times t_i (also with mean τ) were distributed according to Eq. (4). Thus the maximum value of an unbinned likelihood is in general **not** a useful measure of Goodness of Fit.

For the remainder of this discussion about lifetimes, we consider we have just one observed event

at a decay time t_1. Then the log-likelihood is

$$\ln L(\tau) = -\ln \tau - t_1/\tau \quad , \tag{6}$$

which maximises when $\tau = t_1$. The uncertainty range for τ can be defined as the values of τ for which $\ln L = \ln L_{\rm max} - 0.5$. This yields a range for τ of $0.43t_1$ to $3.3t_1$.

With just one event, the really not recommended use of the second derivative of the log-likelihood for estimating the uncertainty on τ $[\, \sigma_\tau^2 = (-d^2 \ln L/d\tau^2)^{-1/2} \,]$ gives $\tau = t_1 \pm t_1$.

2.4.2 Frequentist approach

We next describe the frequentist approach, as exemplified by the central Neyman construction, shown in Fig. 5(b). With a single observed time t_1, at each τ the *pdf* of Eq. (4) is used to find a 68% region extending from $t = 0.174\tau$ to $t = 1.83\tau$; this results in the confidence band between the two diagonal lines in the right-hand plot. Then the vertical line at t_1 sets the 68% interval for τ as from $t_1/1.83 = 0.55t_1$ to $t_1/0.174 = 5.75t_1$.

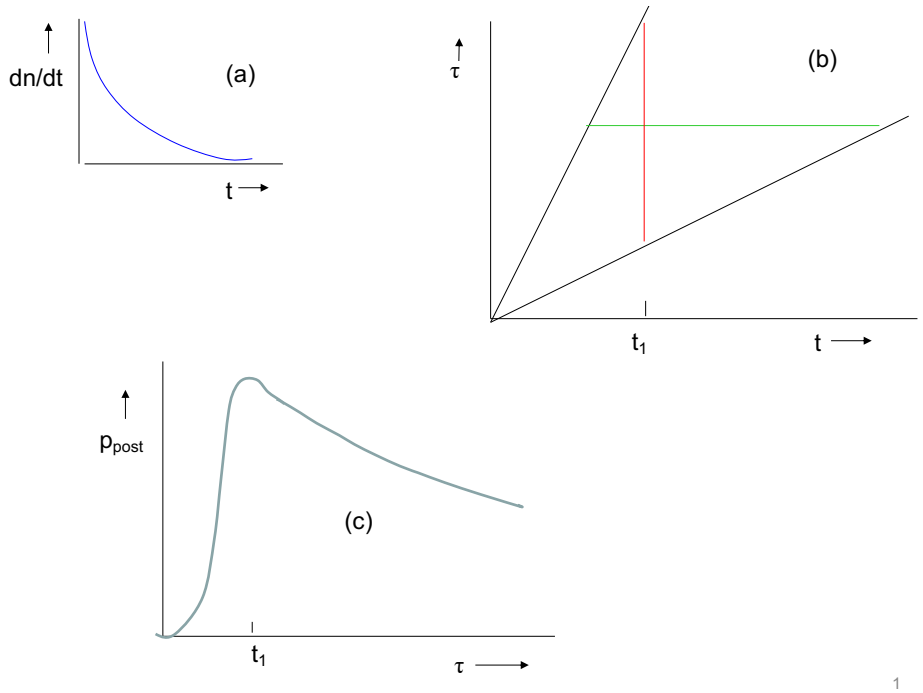


Fig. 5: Ranges on the lifetime τ, using just one observation of a decay at time t_1. (a) The *pdf* for an exponential decay. (b) The Neyman construction of the confidence band for τ is the region between the diagonal lines. (c) Sketch of the shape of the Bayesian posterior $p_{\rm post}$ using a flat prior for τ. The long tail extending to large τ results in it not being normalisable.

2.4.3 Bayesian approach

The Bayesian approach is first to calculate the likelihood and then to multiply by the choice of prior for τ. This could for example be constant[2] in τ, or in the decay rate $1/\tau$, or in $\ln \tau$, etc. These of course will yield different results.

For simplicity, we choose a prior that is flat in τ up to some large value. Then the Bayesian posterior probability will have the same shape as the likelihood function, and in principle can be used for extracting any of the possible Bayesian intervals for τ (central, shortest, upper limit, etc.)

With just one observed event at t_1 and a flat prior for τ, the Bayesian posterior is $(1/\tau)\exp(-t_1/\tau)$ (see Fig. 5(c)), and its integral over all values of τ diverges, so a range for τ cannot be calculated. This problem goes away with two or more observed events, or by using a prior that decreases at large τ.

Table 1: Comparison of methods for determining interval of lifetime τ from one observed decay at time t_1.

Method	Range
$\sigma_\tau^2 = (-d^2 \ln L/d\tau^2)^{-1/2}$	0 - $2t_1$
$\Delta(\ln L) = 0.5$	$0.43t_1$ $-3.3t_1$
Neyman construction	$0.55t_1$ - $5.75t_1$
Bayesian interval	Posterior diverges

2.5 Bayes-frequentist comparison

Table 2 compares the Bayesian and frequentist approaches.

Table 2: Comparison of Bayes and frequentist methods

	Bayesian	Frequentist
Basis of method	Bayes theorem $\rightarrow$ Posterior prob density	Uses pdf for data, for fixed param values
Meaning of probability	Degree of belief	Repeated trials
Probability for params?	Yes	Forbidden
Needs prior?	Yes	No
Choice of interval?	Yes	Yes, except for FC
Data considered	Only data you have	... + other possibilities
Likelihood Principle	Yes	No
Ensemble of experiments?	No	Yes, but often not explicit
Final statement	Posterior prob dist	Param values $\rightarrow$ data is likely
Unphysical/Empty ranges	Excluded by prior	Can occur
Systematics	Integrate over prior	Computationally hard
Coverage	Not Bayesian concept	Built in for Neyman construction
Decision making	Needs cost function	No

A cynic might conclude that Bayesians address the question everyone is interested in, by using assumptions no-one believes, in contrast to frequentists who employ impeccable logic to deal with an issue

[2] Note that a prior that is constant over an infinite range cannot be normalised. For parameter determination, this is not in general a problem, as we can replace it by a prior that is constant up to some large value $\tau_{\max}$ and then is zero. The resulting parameter will be independent of $\tau_{\max}$, provided $\tau_{\max}$ is in the region where the likelihood has become negligible.

of no interest to anyone. However, for parameter determination at the LHC, analysers are encouraged to use both approaches. If their results agree with each other, that is encouraging, but if they differ, that could be because they are answering slightly different questions, or it might be due to a bug in one (or both) analyses.

2.6 Hypothesis testing

This is where we use our data to test which of two competing hypotheses is favoured. Examples include:

- Known Standard Model (SM) Physics versus SM plus some specific version of supersymmetry.
- Does an enhancement in a mass spectrum correspond to a single peak, or to two close peaks?
- Normal mass hierarchy for neutrino masses versus inverted hierarchy.
- Does a jet of particles in an event come from a b-quark, or from one of the lighter quarks?

Here we simply mention that in particle physics, while there are some parameter-determination analyses that use a Bayesian approach, this is very unusual for Hypothesis Testing. This is because the two hypotheses usually involve different priors, whose normalisations are important. In parameter determination, there is typically only a single prior, whose absolute normalisation is irrelevant.

We continue with more on Hypothesis Testing in Section 3 below.

3 Searches for new physics

Searching for new physics beyond the Standard Model (BSM) can be an example of the statistical procedure of Hypothesis Testing (HT). Another example of HT is classification, e.g. deciding whether a jet was initiated by a b-quark or by something else. A difference between these is that for classification it is necessary to decide whether a jet comes from a b-quark or not, while with searches there is the extra option of 'no decision' between the competing hypotheses. This could be because we do not have sufficient data to distinguish between them, or because neither hypothesis is compatible with our data.

Another difference is that when searching for new physics, the criterion for rejecting the SM is very strong (see Section 3.4). In contrast, when selecting whether a jet is initiated by a b-quark, a milder criterion is used, with the resulting contamination from other quark jets being allowed for in the subsequent analysis.

In this section we discuss various statistical topics that arise in our searches for new physics; we do not deal with the classification problem here.

In the last few years, great strides have been made in statistical procedures using machine learning with deep neural networks. These have been used in many aspects of Physics analyses, but especially in searches for new physics. For more details, see the lectures by Troels Petersen. The discussions in this section apply to machine learning procedures as well.

3.1 H_0 or H_0 v H_1?

There are two hypotheses that are relevant for our searches. The null hypothesis (H_0) is that there is nothing new in our data, and that it can all be described by the SM. The alternative (H_1) is that in addition

to the SM there is some specific form of new physics e.g. a particular version of SUper SYmmetry, extra dimensions, a fourth generation, etc. In our searches we have the option of just checking H_0 or of comparing H_0 with H_1. The former is an example of Goodness of Fit (GoF), while the comparison is HT. GoF uses techniques like χ^2, Kolmogorov–Smirnov, etc., while HT commonly uses a likelihood ratio.

An advantage of using just H_0 is that it could in principle be sensitive to deviations from the SM from any type of new physics. In contrast the comparison approach is usually more sensitive if the actual form of new physics corresponds to the one specified in H_1 (see Fig. 6 for an example of this) but could be insensitive to other forms of new effects. Also it can turn out that the data gives a satisfactory GoF to both the null and the alternative hypotheses, but a Hypothesis Test of the 2 hypotheses could significantly favour one over the other.

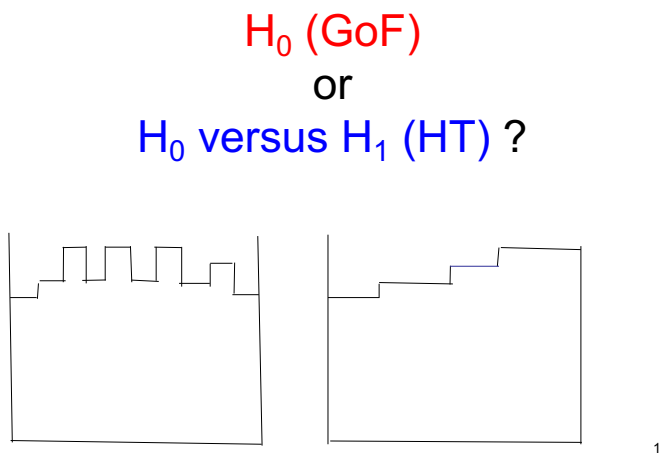


Fig. 6: Goodness of Fit versus Hypothesis Testing. We want to compare data with two hypotheses, H_0 giving a flat distribution and H_1 which has a linear rise. A χ^2 GoF test for H_0 would give identical results for the two sets of data shown (the right hand plot has the identical bin contents to the right-hand one, but shuffled in order), while a likelihood ratio HT for the right data set would favour H_1.

The Neyman–Pearson lemma [6] is relevant for the case of comparing two simple hypotheses. Here 'simple' means that the hypotheses are completely specified without any free parameters, e.g. two possible orderings of the neutrino mass states. A counter-example would involve fitting a mass spectrum with a quadratic function, with the parameters of the function being free. In particle physics, hypotheses are rarely 'simple', with the free parameters being either physically relevant (e.g. the mass of the Higgs boson), or just being nuisance parameters related to various systematic effects (e.g. jet energy scales, trigger efficiency, etc, etc.).

The lemma states that, provided the hypotheses are 'simple', the likelihood ratio $L(d;H_0)/L(d;H_1)$ is the best statistic for separating the two hypotheses, given the data d. This means that when we consider putting a cut on the data statistic for incorrectly rejecting H_0 at given probability, we will have the largest probability for accepting H_1. Even when, as usual in particle physics, the condition for this Lemma is not satisfied, the likelihood ratio may still be a useful test statistics, even if it may

not be (quite) optimal.

As always the performance of a selection procedure needs to be checked, probably by using simulation.

3.2 *p*-values

First we have to choose a data statistic t to summarise our data. In a counting experiment (e.g. a search for Dark Matter), it could be just the observed number of events n, but in more complicated cases t could be a likelihood ratio for H_0 and H_1, or a specially chosen optimal variable. We denote the value of t in our data as $t_{\rm obs}$.

Next we obtain the expected normalised distributions of t under each of the two hypotheses H_0 and H_1. With sufficient data, asymptotic formulae [5] can be used for these distributions, but Monte Carlo simulation may be needed to check that they are valid. Then the p-values p_0 and p_1 are defined as the tail areas beyond $t_{\rm obs}$. For each expected t distribution, the statistics convention is that the relevant tail is the one pointing towards the other hypothesis' t distribution, see Fig. 7(b).

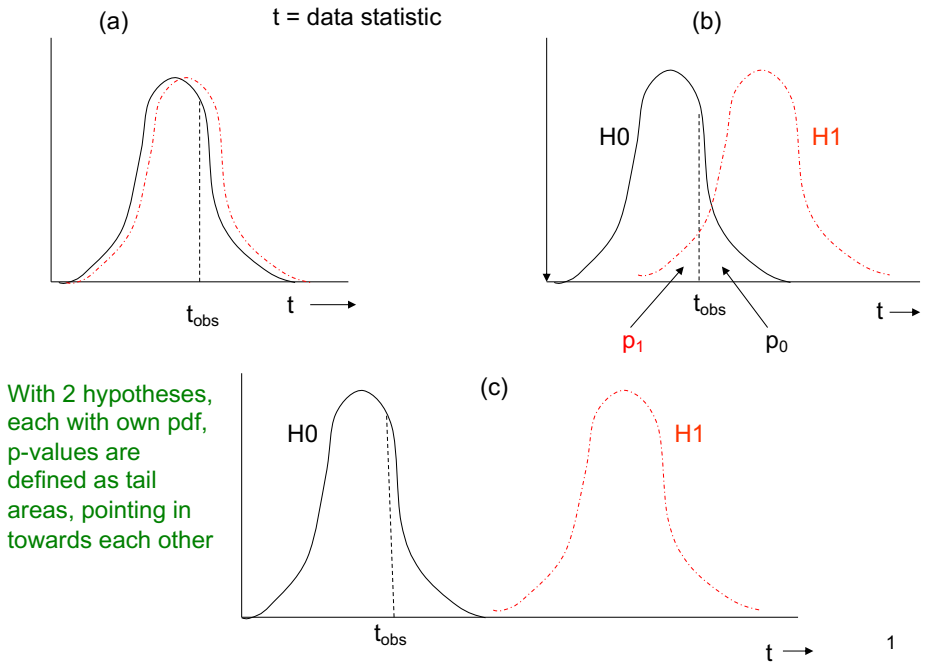


Fig. 7: *pdf*s for the data statistic t for two different hypotheses, H_0 shown in black and H_1 as dashed red curves. (a) The *pdf*s are almost identical and choosing between the hypotheses is unlikely to be possible. (b) A larger separation of the *pdf*s. With observed t being $t_{\rm obs}$, p_0 and p_1 are the tail areas beyond $t_{\rm obs}$. (c) The separation of the *pdf*s is even larger, making the choice between H_0 and H_1 easier.

A p-value indicates the degree to which our data and the relevant hypothesis are consistent. A small p-value suggests that the data may be biassed and/or the theory is incorrect. It is important to

realise it is **not** the probability of the theory being correct[3]. Apart from anything else, p-values are a frequentist concept, while the probability of a theory being true exists only as Bayesian probability.

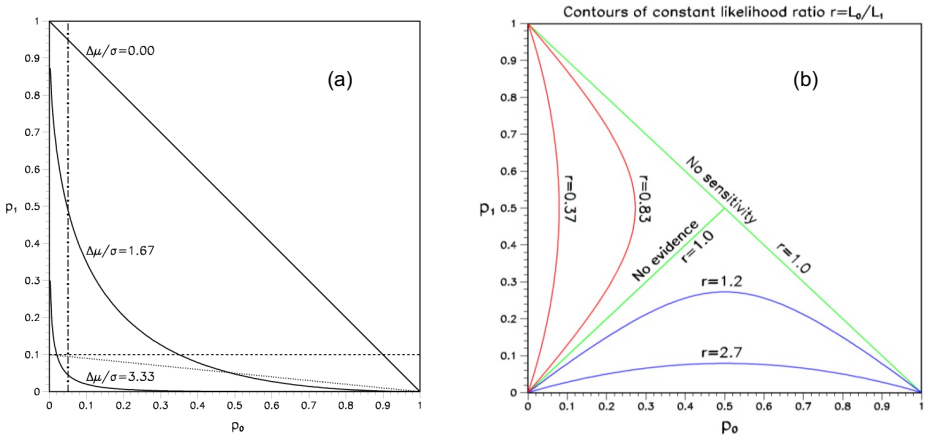


Fig. 8: p_0 versus p_1 plots. (a) The curves are possible values of (p_0, p_1) pairs for fixed values of the separation of the Gaussian pdfs. The diagonal dashed line near the bottom of the plot corresponds to $CLs = 0.1$. (b) Contours for fixed values of the likelihood ratio L_0/L_1.

Figure 8 shows plots of p_0 against p_1, the p-values for data t compared with the pdfs for two different hypotheses H_0 and H_1 assumed to be Gaussians of equal width σ and with their centres separated by $\Delta\mu$ (compare Fig. 7). The 'curves' in the left hand plot show the way the anti-correlated p_0 and p_1 vary as the data statistic t fluctuates; they are for separations $\Delta\mu/\sigma$ of 3.33, 1.67, and zero (the diagonal straight line). The plot also has horizontal and vertical straight lines showing possible cuts for excluding H_1 and H_0 respectively: the latter is shown at an unrealistically large value merely to be more visible. The diagonal dashed line corresponds to fixed CLs (see Section 3.9).

Plot (b) shows contours of the likelihood ratio L_0/L_1 in the (p_0, p_1) plane. They are very different from the contours of constant p_0 or p_1. That is, data sets that have the same p_0 can have very different L_0/L_1.

These plots provide insights on:

- CLs for exclusion of H_1.
- Punzi definition of sensitivity. As the amount of data increases, so does the separation $\Delta\mu/\sigma$ of the curves in plot (a). The Punzi sensitivity criterion is the amount of data that is sufficient for its $\Delta\mu/\sigma$ curve to lie outside the big 'no decision' square in (a) (or the slightly larger rhombus if the CLs criterion is being used for exclusion of H_1). That is, there is enough data such that whatever the result t of the experiment, either H_0 or H_1 will be excluded. In the example in the figure, the $\Delta\mu/\sigma = 3.33$ curve satisfies this condition, while the 1.67 one does not.
  The Punzi criterion is not often used in particle physics. This is primarily because it requires significantly more data than the standard procedure of needing enough data to have a 50% chance of exclusion, assuming H_0 is true.

[3]This is related to the fact that $P(A;B)$ is not the same as $P(B;A)$. An example is that the probability of being pregnant if you are female is very much smaller than the probability of being female, given that you are pregnant.

– Relation of p-values and Likelihoods, and the Jeffreys–Lindley paradox [7]. The latter draws attention to the possibility that, by using p_0, data could achieve a 5σ exclusion of H_0, while the likelihood ratio L_0/L_1 favours H_0.
– Probability of excluding H_1 when H_0 is true. This, and other probabilities, can be read off from Fig. 8(a). This is true whether the exclusion of H_1 is based on CLs or on simply p_1.

More details can be found in Ref. [8].

3.3 Look elsewhere effect

Searches for new physics often involve looking for a peak in a relevant mass distribution, see Fig. 13. We want to ensure that an observed excess is unlikely to be an upward fluctuation of the background b at that particular mass, so we calculate the p-value for obtaining an excess at least as large as the one we observed, assuming that the events there are Poisson distributed with mean b according to H_0. This is called the **local** p-value. However, usually we do not know the mass of the object we are seeking and so a fluctuation at any mass could be mistaken as evidence for a new particle. The chance of this is clearly larger than that of a fluctuation at a specific mass, and is known as a **global** p-value. This is the Look Elsewhere Effect (*LEE*). It dilutes the significance of an observed effect.

The problem is that there is no exact prescription of where 'Elsewhere' is, and to some extent it depends who you are. For a graduate student, it probably includes anywhere in the analysis they performed. But the Director-General of CERN may want to protect herself against false discovery claims for new physics arising from fluctuations in analyses from any of the experiments performed at CERN, so her *LEE* factor would be much larger.

Unfortunately there is no recommended convention for dealing with the *LEE*. It is good practice for any discovery claim to publish the local p-value, as well the global one corresponding to the reasonable mass range for the spectrum in which the peak was observed. Clearly it is important to specify what you include in your 'Elsewhere' for a global p-value.

3.4 Why 5σ for discovery?

In other fields of research, results are regarded as significant if they differ from the null hypothesis at the 5% level. Particle physics tends to use the very much more stringent requirement of $p \leq 3\ 10^{-7}$ for discovery, corresponding to a 5σ fluctation.

The motivations for this include:

– Past experience shows that previous claims of 3σ or even 4σ effects have gone away with more data.
– Given the large number of analyses performed in particle physics, and that we want to reduce false discovery claims to a minimum, it is prudent to set the bar for p for an individual analysis at a small value. This is basically LEE at a very large scale.
– Unlike statistical uncertainties, systematics are sometimes underestimated. Putting a stringent requirement on the p-value is a crude way of allowing for this.

– The current theory for particle physics is the SM, which has done an excellent job in predicting the results of a whole series of measurements, which have almost always been in agreement with the data. Before rejecting it in favour of some speculative alternative, we want to be very convinced that our data disagree with the SM. This is an example of the old adage that "Extraordinary claims require extraordinary evidence".

It is somewhat unreasonable to apply the same criterion to all analyses, but it is too complicated to have a flexible standard. Nevertheless, when it comes to the search for di-Higgs production, for example, there is no *LEE*. Furthermore there is no reason to regard it as extraordinary as it is predicted by the SM, so there is little justification for requiring 5σ[4]. Indeed non-observation of HH would be the big surprise.

3.5 Significance

It is conventional to convert a p-value into a significance z, such that the tail area beyond $z\sigma$ from the centre of a standard Gaussian is p. If the p-value was calculated for the data being either an excess or a deficit, the 2-sided tail area is relevant, while if only an excess (or only a deficit) is relevant, then the one-sided tail is used. Thus for the single tail case a p-value of 16% corresponds to $z = 1$, while $z = 5$ for $p = 3\ 10^{-7}$.

Given that there is a one-to-one relationship between z and p, the only reason for converting p to z is that the z values are easier to remember.

3.6 Wilks' theorem

In hypothesis testing, the choice between the two hypotheses can be based on the $-2\Delta \ln L$ where $\Delta \ln L$ is the difference of the ln-likelihoods of the two hypotheses; or on ΔS, the difference in the weighted sum of squares . Wilks' theorem [9] is useful for calibrating these differences.

For the theorem to apply, the first condition is that the two hypotheses are nested. This is when one of the hypotheses reduces to the other for specific choices of some of its parameters. An example would be comparing fits of a 5^{th} order polynomial or a 3^{rd} order one to our data; with the coefficients of the 4^{th} and 5^{th} order terms set equal to zero, the former reduces to the latter. A common particle physics example is seeing if a model of a SM background plus a peak is better than just the SM background. The case where the hypotheses are the Normal Mass Hierarchy for neutrino masses or the Inverted Hierarchy involves non-nested hypotheses.

For nested hypotheses, because the larger hypothesis includes the lower one as a special case, $\Delta S = S_0 - S_1$ cannot be negative, where S_0 applies to the hypothesis with the smaller number of free parameters. The theorem states that ΔS should be distributed according to the mathematical χ^2 distribution with the number of degrees of freedom equal to the difference in the number of free parameters in the two hypotheses. This would be 2 for the polynomials example earlier.

For the theorem to be applicable, the following conditions must be satisfied:

– The data should be asymptotic. This is not only for Poisson fluctuations of the contents of histogram bins to be approximately Gaussian. Also when the expected distribution is only weakly

[4] This does not imply that we can stop running the experiment once we have achieved 3σ. Once the process has been confirmed, more data will be required to study the production of di-Higgs in more detail.

dependent on a parameter, a large amount of data may be required to detect its influence. Otherwise it does not count as a free parameter. Similarly in neutrino oscillation analyses, only the product of the two parameters may be determined without a large amount of data (see Ref. [4]). This condition applies not just to Wilks' theorem but also for a weighted sum of squares being distributed as a χ^2.

- The hypotheses must be nested.
- To reduce the larger hypothesis to the smaller one, its extra parameters must be uniquely defined. This condition is true for our polynomial example, where the extra coefficients must all be zero. It is violated for comparing the SM with SM plus a peak of variable position and strength; if the strength of the peak is zero, its mass is irrelevant. This is part of the reason why the search for the Higgs boson at the LHC was conducted as a Raster Scan, with each mass being tested separately and the extra parameter being just the strength of the possible signal at that mass.
- None of the extra parameters for reducing the larger hypothesis to the smaller one should be on the boundary of its allowed region. In the above example of searching for a peak, this would be violated if the peak strength was not allowed to be negative.

Even when Wilks' theorem is not applicable (e.g. the Higgs boson's spin-parity—see Section 3.10.3), we can use Monte Carlo simulation to derive the expected distribution of our test statistic for each of the hypotheses.

3.7 Blind analysis

In looking for some new effect, there is a danger that the analyser knowingly or subconsciously might adjust the procedure to produce a more desirable result. For example, if a discovery claim is based on a small number of events above a very small background, it can make a big difference whether a couple of events should be accepted or not based on specific event features, or by a small change in the acceptance region. This can be avoided by using a 'blind' approach. This means that the procedure is chosen and frozen, without looking at the actual data. One way of achieving this is to use Monte Carlo (MC) simulation of your experiment to define the analysis procedure. But this suffers from the danger that the MC does not faithfully reproduce the real data. The aim is thus to devise a procedure which allows you to look at the data as much as possible without being aware of the final answer. There are many ways of doing this; see, for example, the review by Klein and Roodman [10].

3.8 Background systematics

As with many types of analysis, systematics can be a big issue and searches for new physics are no exception. A whole PHYSTAT meeting was devoted to systematics, and the ways in which they can be incorporated into our analyses [13].

Here we discuss just one aspect: How the uncertainty on the shape of the background affects the estimate of the strength of the Higgs signal, for example as seen in the left plot of Fig. 13.

The extraction of the p-values mentioned earlier required a functional form (e.g. exponential, polynomial, Bernstein, etc.) for the background. For any given form with a specified number of terms, the background b under the peak and its statistical uncertainty are determined. But for a different number of

terms, or for a different functional form, b would have been different, and hence there is an additional systematic uncertainty that is relevant for the p-value. The traditional way of dealing with this was to decide which models to reject because they gave worse fits to the data, and to use the remaining ones to estimate this systematic.

The discrete profiling method [14] aims to improve on this by using a method analogous to that used for continuous nuisance parameters. Figure 9 shows a series of approximate parabolae. Each one represents the fit value (χ^2 or minus twice the log-likelihood $-2\ln L$) for a fit to the data, where at each value of the parameter of interest μ, the likelihood has been profiled with respect to all the continuous nuisance parameters ν. That is

$$\ln L_{\rm Profile}(\mu) = \ln L(\mu, \nu_{\rm best}(\mu)) \quad , \tag{7}$$

where $\nu_{\rm best}(\mu)$ is the value of the nuisance parameters which maximise the likelihood at that particular μ.

Discrete profiling works similarly with respect to the different discrete functional forms, as represented by the curves of Fig. 9. The effect of this uncertainty is allowed for by choosing the lower envelope of the different curves; the total parameter range at the $\approx 68\%$ level is given by the positions at which the $-2\Delta\ln L = 1$ line intersects the envelope. Functional forms that give a poor fit will have their minima too high to make a significant contribution to the width of the interval, while those whose fit quality is medium will contribute only at larger confidence levels; and all this happens in a smooth way. For the situation shown in Fig. 9, the best fit is given by the red curve, but the uncertainty range for μ is increased to μ_1 to μ_2 because of the other functional forms; at the confidence level shown, the yellow curve has no effect.

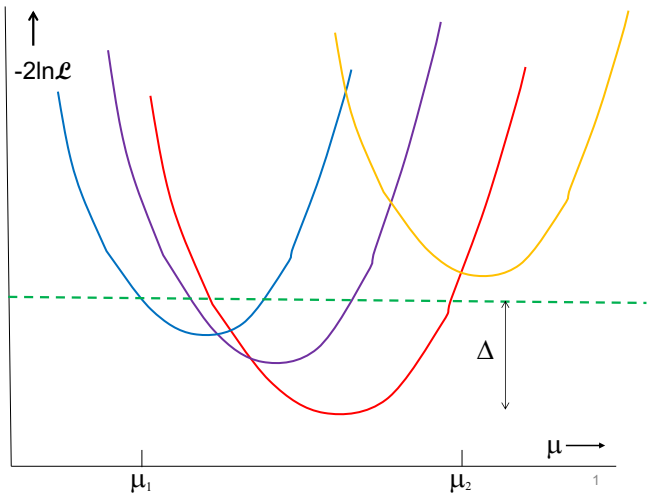


Fig. 9: Plots of $-2\ln L$ against the parameter of interest μ for different functional forms. The systematic due to the choice of different functional forms for the background is allowed for by using the lower envelope of the different curves.

3.9 Upper limits

Searches for new physics may well not result in a discovery claim. Rather than publishing nothing, upper limits can/should be set on $H1$; if the new physics had been strongly produced, we would have seen it in our data. This then constrains the allowed values of the parameters of the model.

Historically, the most famous example of an upper limit having a significant effect on the development of physics was the Michelson–Morley experiment [11]. This set an upper limit of the earth through the assumed aether which was below the speed of the earth's revolution around the sun, and the solar system's rotation in the Milky Way. This resulted in the death of the aether, and Special Relativity soon followed.

Particle physics upper limits are usually quoted at the 90% or 95% level. This is a much looser criterion than is used for discovery. This is mainly because it is far less embarrassing to have incorrectly excluded some model of new physics, than to have wrongly claimed a discovery. As Glen Cowan says, if you lose your keys at home, being 90% sure they are not in the kitchen is enough for you to move on to looking in other rooms.

A problem can arise from the weaker criterion (e.g. $p_1 < 0.05$) for excluding the alternative hypothesis H_1. Figure 7(a) shows a situation where the data is such that the pdfs of H_0 and H_1 essentially overlap, so there is no real chance of discriminating between the hypotheses. Nevertheless there is a 5% probability that the data statistic t will fluctuate down so that p_1 (and p_0 as well) is below 0.05. This could then result in H_1 being rejected by an experiment which lacks the power to do so. The common way to avoid this in particle physics is to base the exclusion instead on the poorly-named CL_s criterion [16]:

$$CLs = p_1/(1-p_0) \quad , \tag{8}$$

i.e. it is the ratio of the left hand tails of the pdfs beyond the particular data value $t_{\rm obs}$. This is guaranteed to be no smaller than p_1, and so is a conservative variant of the standard frequentist procedure. Conservatism is the price to pay for the protection offered against unjustified exclusion.

There are many methods of setting upper limits, based on frequentist, Bayesian and likelihood approaches, each with several variants; Bayesian methods involve a choice of prior for the signal strength. A comparison of several of these is in Ref. [12]. Figure 10 shows various upper limits for the signal strength s_0 in a Poisson counting experiment when the expected background b is 3.0 events, as a function of the observed event number n. The various methods produce a range of answers, which is most apparent when, because of statistical fluctuations, n is smaller than b.

Limits in particle physics are often set on the production rate of some new particle (e.g. some specific form of lepto-quark) as a function of its mass M. If there is a theoretical model for predicting its production rate as a function of the possible mass values, and there are masses where the predicted rate is above the upper limit as determined from the data, such masses are ruled out. This typically results in masses below a certain value being excluded, and so sets a lower limit on its possible mass (see Fig. 11). This is useful in restricting theories.

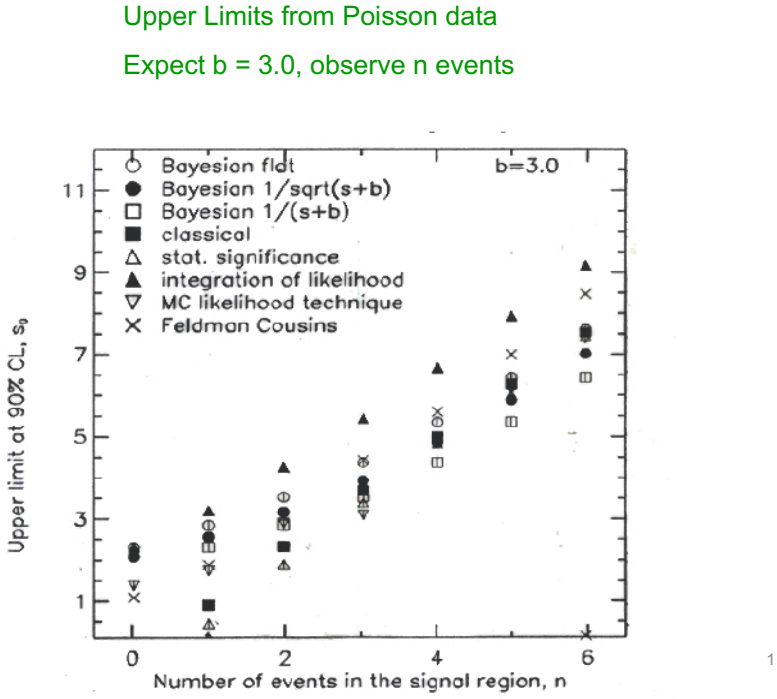


Fig. 10: The upper limits on the possible signal rate s_0 are shown for Poisson distributed data, with expected background rate $b = 3.0$ events, when n events are observed. Different methods provide a range of upper limits, which is more pronounced when $n < b$.

3.10 Example with real data

To illustrate the above ideas, we discuss briefly some real examples of analyses related to the Higgs boson. They refer to the Higgs discovery, and its mass and spin, with data collected at the time of discovery or soon after. With the larger data samples currently available, not only are these analyses now more impressive, but also many more analyses are now possible, e.g. other SM decay modes of the Higgs, searches for unusual decays, its different couplings to fermion and to boson pairs, etc. These are not discussed here; for further details see Ref. [17].

3.10.1 Higgs discovery

Figure 13 shows two mass spectra that we used as part of the discovery claim of the Higgs boson in 2012. They are from data of the CMS experiment; ATLAS had similar plots. The one on the left is for a possible Higgs decaying into two photons. There is a peak at 125 GeV, containing a largish number of events but the signal to background ratio is small. The right hand plot is for Higgs decaying to four leptons (electrons or muons) via two Z bosons. The number of events in the plot is small, but the peak at 125 GeV has a large signal to background ratio.

The Higgs hypothesis involves two parameters of interest, m_H and the strength μ of the possible Higgs signal. Rather than performing a two parameter fit to the data, a Raster Scan was used: at each mass m_H a likelihood ratio for Higgs versus no Higgs was used to extract the local p-value for the 'no Higgs' hypothesis. These are shown in Fig. 12. The global p-values for a couple of mass ranges were

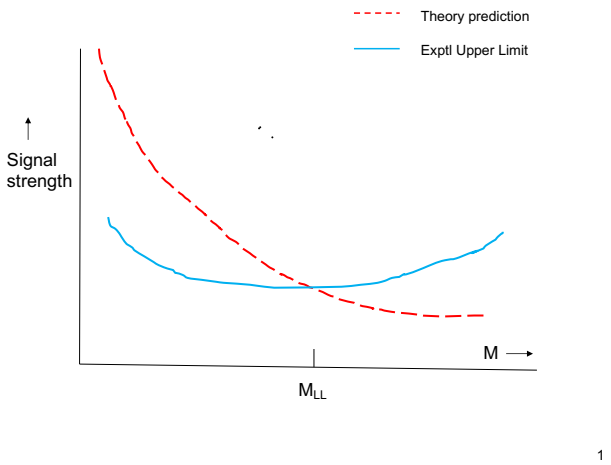


Fig. 11: The experimental upper limits on the possible signal rate as a function of signal mass M (solid blue curve), compared with the theory prediction (dashed red curve). When the prediction is above the experimental upper limit on the signal strength, the theory is ruled out, so M_{LL} is the lower limit on possible masses for the searched-for particle.

also quoted, to allow for possible fluctuations at other masses, rather than just for the observed peak position at 125 GeV.

The above procedure also included the extra uncertainty introduced by the systematic effects. This included the uncertainty resulting from the various possible choices for the functional form used to describe the background to the mass peak. At later stage of the analysis, this was dealt with by the Discrete Profiling approach—see Section 3.8.

3.10.2 Higgs mass

A very important parameter of the Higgs boson is its mass. This was determined using the information about the masses of individual events that contributed to histograms as in Fig. 13. A likelihood approach was used to extract the signal mass M_H and strength μ, with the extracted uncertainty range for M_H involving profiling over μ. For the $\gamma\gamma$ channel, the input mass spectrum had narrow bins, while an unbinned approach was used for the four-lepton channel.

As well as the statistical uncertainty, there were of course many sources of systematics to be considered. For the $\gamma\gamma$ mode, the main effect comes from the energy scale for the γ, but also the shape of the background played a role, because of the low signal to background ratio. The analysis also divided the events into categories, depending on the magnitude of the uncertainty on $m_{\gamma\gamma}$; this made use of the events with good resolution, while not ignoring the others.

For the four-lepton channel, the energy scales for electrons and muons are the most significant systematics. Although the four-lepton channel has fewer events, its better signal-to-background ratio results in the two channels having similar statistical uncertainties for the Higgs mass. Figure 14 shows $-2\Delta \ln L$ as a function of mass is shown for the two channels and their combination; and the statistical

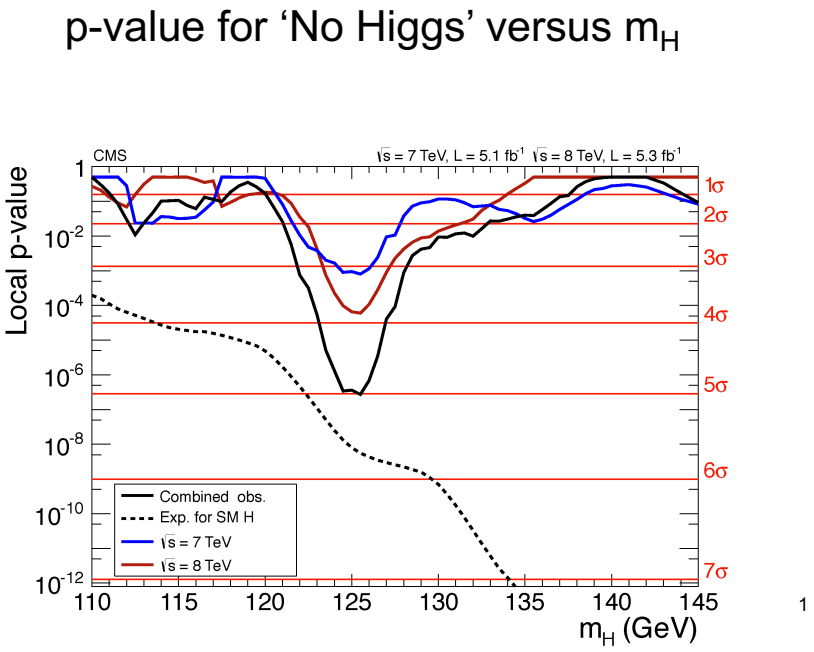


Fig. 12: Local p-values for the null hypothesis of 'No Higgs', as a function of the Higgs mass m_H. The black curve is for the combined data of 2011 and 2012. Its local significance peaks at 5σ around 125 GeV. The dashed curve shows the expected significance at m_H, assuming that the Higgs had that mass, and was produced at the rate predicted by the SM. The agreement with the observed significance around 125 GeV is regarded as satisfactory.

and the overall uncertainty for the combination. It is interesting to note that for the mass determination, a 2-D approach involving M_H and μ was used. This contrasts with the discovery analysis, where a raster scan at each mass separately was performed. These are the recommended procedures for Parameter Determination and for Hypothesis Testing respectively.

3.10.3 Higgs spin

According to the SM, the spin and parity of the Higgs is predicted to be 0^+. Other possibilities in principle are $0^-, 1^+, 1^-, 2^+$, etc. These result in different distributions for the angular variables for Higgs decays to four leptons.

Standard statistical HT involves a comparison of 2 hypotheses, but here we have many possibilities. One way of dealing with this is to treat the SM's 0^+ as the null hypothesis and the others one at a time as the alternative. Here we show the results just for the comparison of 0^+ with 0^-. The test statistic t was -2 times the log-likelihood ratio for the two different hypotheses.

Figure 15 shows the expected distribution of the test statistic $t = -2\ln(L_{0^-}/L_{0^+})$. The distributions overlap quite a bit, even though the data sample used here was larger than for the Higgs discovery in Fig. 13. The SM prediction of 0^+ gives rise to the extracted distribution of t on the right with the horizontal shaded lines. The data value of t is denoted by the arrow. It is far enough into the right hand

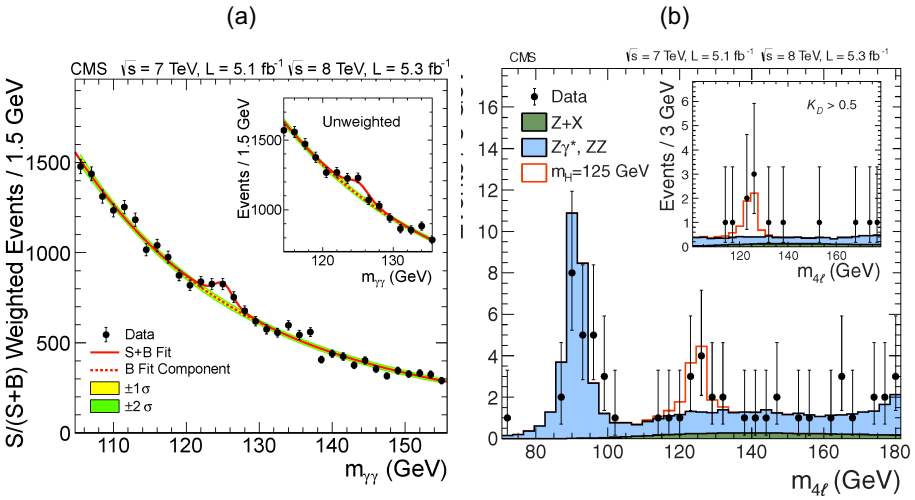


Fig. 13: Mass spectra of (a) gamma gamma and (b) ZZ decaying to four leptons. The peaks at 125 GeV on each plot are clearly visible.

tail of the 0^- distribution for CLs[5] to be below 0.05, and so 0^- is excluded.

In a sense, the ability to distinguish between these hypotheses was somewhat fortuitous; as can be seen from Fig. 15, assuming that the spin-parity is 0^+, the data (green arrow) could well have been somewhat over to the left, which would have raised the CLs value such that 0^- would not have been excluded.

4 Learning to love the covariance matrix

When we estimate the value of a single parameter of interest, we also have to provide an estimate of the uncertainty on it, e.g. the mass of the W boson is 80,360 $\pm$ 10 MeV [15]. When there are two or more parameters of interest, we have to provide not only the value and uncertainty for each, but also the correlation between these estimates. This is specified by the covariance or correlation coefficient. The aim of this section is to provide an intuitive understanding of this, and the way to deal with such correlations in our analyses.

In the one dimension case, the concept of variance applies not just to Gaussians but to any distribution, but it simplifies discussions when the distributions are Gaussians. Similarly in several dimensions involving correlations, here for simplicity we discuss the situation where the distributions are multi-dimensional Gaussians.

4.1 One-dimensional Gaussians

Before looking at the multi-dimensional Gaussian, we review briefly the properties of the one-dimensional case. The Gaussian $G(x;\mu,\sigma)$ is given by

$$G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-0.5(x-\mu)^2/\sigma^2] \quad . \tag{9}$$

[5] For excluding 0^-, CLs is defined as the ratio of the right hand tails in Fig. 15.

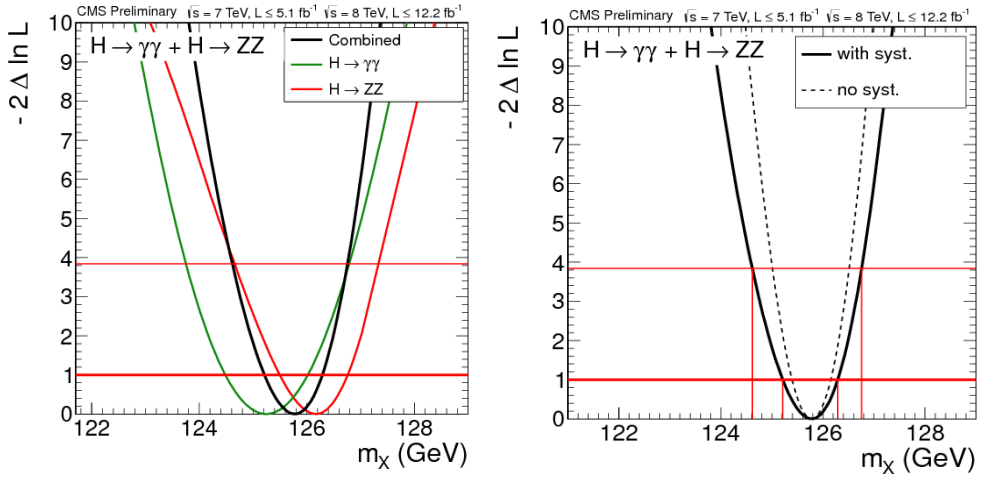


Fig. 14: Higgs mass: Plots of $-2\Delta \ln L$ versus mass. (a) The coloured curves are for the gamma gamma and the ZZ to four-lepton channels separately, and the one in black for their combination. (b) The dashed curve is for the statistical uncertainty only, while the somewhat wider solid curve includes the effect of systematics.

The factor $1/(\sqrt{2\pi}\sigma)$ ensures that $\int G dx$ is unity, and hence suitable to be a probability density distribution.

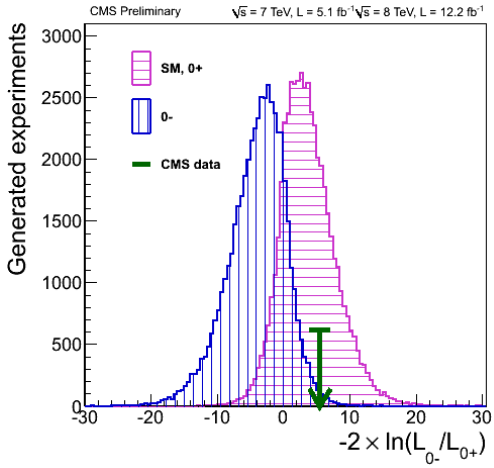
The function G clearly has a maximum at $x = \mu$, and is symmetric about it. The parameter σ determines the width of the distribution. In particular, from Eq. (9):

- The Gaussian's root mean square deviation about its mean (RMS) turns out to be σ. This explains the factor of 0.5 in the exponential.
- The value of G at $x = \mu \pm \sigma$ is $1/\sqrt{e} = 0.606$ times its maximum at $x = \mu$. So to a crude approximation, σ is the half-width at 'half'-height of the distribution of G.
- The area under the curve for G between $x = \mu - \sigma$ and $x = \mu + \sigma$ is 68% of the total area. If $G(x; \mu, \sigma)$ represents the probability density of obtaining a result x for a quantity whose true value is μ and the measurement has resolution σ, we would expect about 2/3 of the measurements to be within σ of the true value, and 1/3 to be outside that range.

Again for simplicity, we henceforth assume that the Gaussian $G(x; \mu, \sigma)$ is centred at $\mu = 0$.

We now turn to 2-dimensional Gaussians. The extension to a higher number of dimensions is straightforward.

Comparing 0^+ versus 0^- for Higgs



http://cms.web.cern.ch/news/highlights-cms-results-presented-hcp 1

Fig. 15: According to the SM, the spin and parity of the Higgs is predicted to be 0^+. Other possibilities in principle are $0^-, 1^+, 1^-, 2^+$, etc. These result in different distributions for the angular variables for Higgs decay to four leptons. The comparison here is for 0^+ with 0^-; the data, shown by the green arrow, favour 0^+.

4.2 2-dimensional Gaussians

We consider a situation where we have two measurements x and y, both Gaussian distributed and centred at zero, with widths σ_x and σ_y. i.e.

$$G_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp[-0.5\, x^2/\sigma_x^2] \quad \text{and} \quad G_y(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp[-0.5\, y^2/\sigma_y^2] \quad . \tag{10}$$

We will soon deal with the correlated case, but first we consider G_x and G_y to be uncorrelated. Then the joint distribution $G_{x,y}(x,y)$ of x and y is given simply as

$$G_{x,y}(x,y) = G_x(x) * G_y(y) = \frac{1}{2\pi\sigma_x\sigma_y} exp[-0.5\, (x^2/\sigma_x^2 + y^2/\sigma_y^2)] \quad . \tag{11}$$

Figure 17 shows contours of the function $G_{x,y}$. A simple check can be performed to show that they are consistent with x and y being uncorrelated. We imagine the contours as showing the heights at various locations on a hill, with the top of the hill at the origin. The dashed line at constant x shows a path over the hill, which avoids the summit. The highest point on the path is shown by the short arrow, and is at $y = 0$; this is also true for any path at constant x. The fact that the maximum in y for any x is independent of x is a necessary condition (albeit not sufficient) for x and y to be independent.

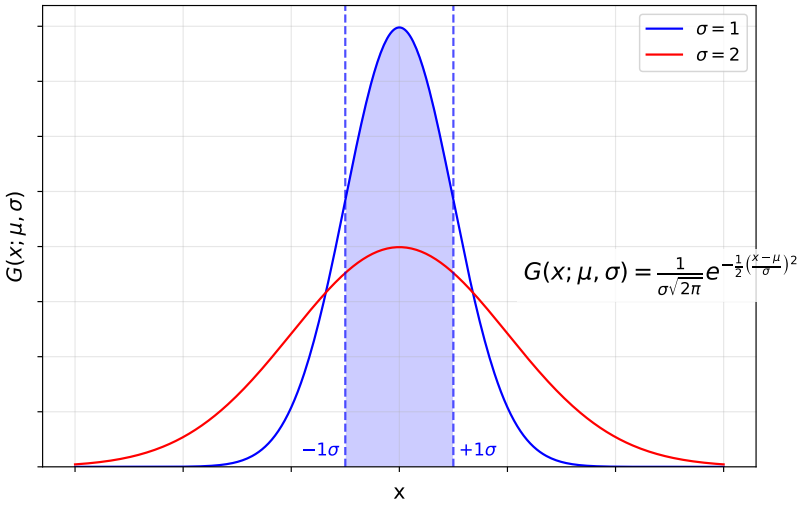


Fig. 16: Normalised 1-D Gaussian, with mean μ and standard deviation σ. The dashed curve is for the standard deviation twice as large, and hence has only half the peak height.

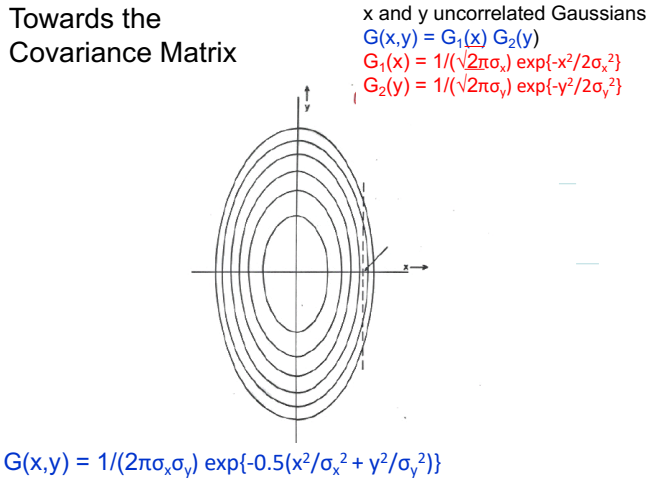


Fig. 17: 2D-Gaussian $G_{x,y}$ with G_x and G_y uncorrelated. The ellipses are contours of $G_{x,y}$ and the vertical dashed line is a path across the 'hill'. The arrow denotes the highest point on the path.

In analogy with Section 4.1, the ellipse

$$x^2/\sigma_x^2 + y^2/\sigma_y^2 = 1 \tag{12}$$

is where $G_{x,y}(x,y)$ is a factor of $\sqrt{e}$ smaller than its maximum at the origin. However in this case, the contour does **not** include 68% of the area under $G_{x,y}(x,y)$. For that, we have to use the larger contour

$$x^2/\sigma_x^2 + y^2/\sigma_y^2 = 2.3 \quad . \tag{13}$$

So care is needed in interpreting exactly what σ_x is. It corresponds to the narrower half-width in x of the 68% confidence region of Eq. (12), but which extends infinitely to large and small values of y. It is **not**

the half width of the 2-D rectangle enclosing the 68% confidence level ellipse of Eq. (13); that would be $\sqrt{2.3}\sigma_x$ (see Fig. 18).

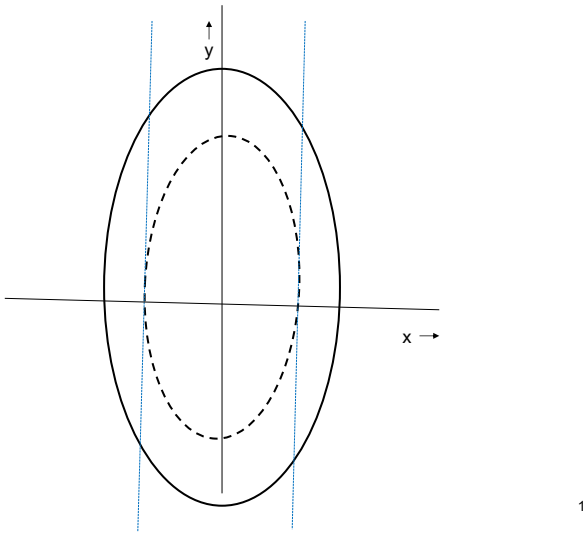


Fig. 18: Two 2-D confidence regions for x and y, both at the 68% level. The first is the region between the vertical blue lines, where y can have any value, while the second is within the larger ellipse. The smaller dashed ellipse corresponds to $2\Delta \ln L = 1$, while it is 2.3 for the larger one.

As a specific example, we choose $\sigma_x = \sqrt{2}/4 = 0.354$ and $\sigma_y = \sqrt{2}/2 = 0.707$, so that the ellipse of Eq. (12) becomes

$$8x^2 + 2y^2 = 1 \quad . \tag{14}$$

We are now ready to introduce correlations. We do this simply by rotating the axes of Fig. 17 by an angle $\theta = 30$ degrees. With respect to the new axes x' and y', the $\Delta \ln L = 1/2$ ellipse becomes

$$0.5 * (13x'^2 + 6\sqrt{3}x'y' + 7y'^2) = C \quad , \tag{15}$$

with $C = 1$. Contours in the (x', y') plane for different values of C are shown in Fig. 19. The ellipses are seen to be tilted with respect to the (x', y') axes; it is the $x'y'$ term in the above equation which is indicative of the correlation between x' and y'. The dashed path at constant x' reaches its maximum height at the point shown by the arrow, and now occurs at a value of y' that depends on x'. This is evidence of a correlation.

The next step is to write this in matrix notation. This may seem like overkill for what is a simple situation, but is going to be useful when we deal with more complicated scenarios, e.g. more dimensions, repeated use of correlated variables, combining results, etc. It enables us to deal with correlations by manipulating matrices, without having to worry about how to deal with them correctly[6]. So we now have

$$\begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} 13/2 & 3\sqrt{3}/2 \\ 3\sqrt{3}/2 & 7/2 \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = 1 \quad . \tag{16}$$

[6] As with all statistical procedures, we do need to check with a few specific cases. In this case, it is that we have the correct matrices and are using them as required.

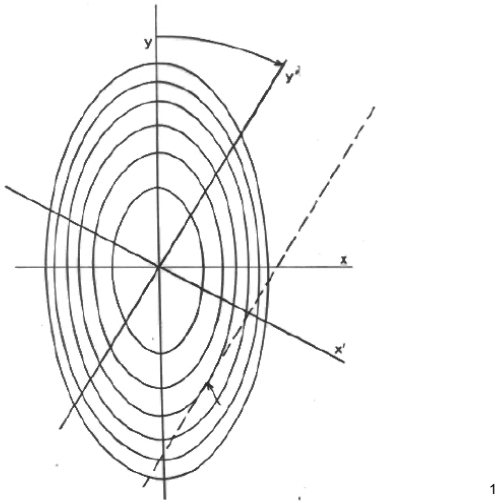


Fig. 19: The same 2-D Gaussian in x and y as in Fig. 17, but with correlation for the rotated variables x' and y'.

The 2×2 matrix above is called the inverse covariance matrix, because when we invert it, we obtain the covariance matrix:

$$\begin{bmatrix} \sigma^2_{x'} & cov(x',y') \\ cov(x',y') & \sigma^2_{y'} \end{bmatrix} = 1/32 \begin{bmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{bmatrix} \quad . \tag{17}$$

Its diagonal elements $\sigma^2_{x'}$ and $\sigma^2_{y'}$ are the variances on the new variables x' and y'. The off-diagonal element is called the covariance between x' and y', and it encapsulates information about the correlation between the variables. Its negative sign shows that the variables are anti-correlated. The covariance can be written as

$$cov(x',y') = \rho\, \sigma_{x'}\, \sigma_{y'} \quad , \tag{18}$$

where ρ is the correlation coefficient, and is confined to the range -1 to +1 inclusive.

Figure 20 demonstrates several numerical features of the inverse covariance matrix in Eq. (16) and for the covariance matrix of Eq. (17).

So far we have discussed the covariance matrix for some general variables x' and y'. From a Physics viewpoint, x' could be a parameter of interest and y' a nuisance parameter; or they could both be parameters of interest e.g. the intercept and gradient of a straight line. In either case, if we want information just about x', one approach is to 'profile' over y'. The profile likelihood $L_{prof}(x')$ is derived from the 2-D likelihood $L(x',y')$ by choosing at each x' the value of y' that maximises the likelihood for that x'. As x' varies, this consists in evaluating the likelihood at a series of points in Fig. 19 that are on a straight line including the origin and the arrowed point.

4.2.1 Understanding the covariance

For our 2 variable situation, the elements of the covariance matrix are σ_x^2, σ_y^2 and $cov(x, y)$. The first two are readily understood in terms of the uncertainties on x and on y. To give insight into the covariance term, Fig. 20 shows three ellipses with the same σ_x and the same σ_y but with different correlation coefficients ρ. (Remember that $cov(x, y) = \rho\sigma_x\sigma_y$.)

The three chosen values of ρ are:

- $\rho = 0$. This corresponds to the uncertainties on x and y being uncorrelated.
- $\rho = -0.9$. This is a strong anticorrelation. The ellipse has become quite thin, and its major axis has a negative gradient.
- $\rho = +1.0$. This is the largest value that ρ can have. The length of the minor axis has collapsed to zero, and the positive correlation is complete, so that choosing a value of x completely defines y.

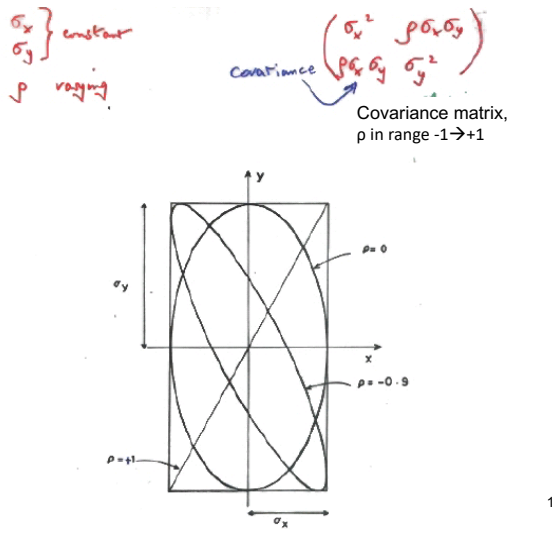


Fig. 20: Gaussians with different correlation coefficients $\rho = 0, -0.9$ and $+1$. Because they all share the same value of σ_x and also of σ_y, they fit into the same rectangle.

It is also worth noticing that all three ellipses fit into the same-sized rectangular box (see Fig. 20). This is because they share a common value of σ_x and also of σ_y.

4.3 Using the covariance matrix

4.3.1 Function of measured variables

We first deal with deriving the uncertainty on a function $f(x, y)$ of two (or more) variables, whose covariance matrix is known. The function f could be as simple as $f = x - y$, or it could be much more complicated.

The first terms in a Taylor expansion of f are

$$\delta f = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y \quad . \tag{19}$$

If we square this equation, and then take the average, we obtain

$$\overline{\delta f^2} = (\frac{\partial f}{\partial x})^2 \overline{\delta x^2} + (\frac{\partial f}{\partial y})^2 \overline{\delta y^2} + \frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\overline{\delta x \delta y} \quad . \tag{20}$$

This can be written in matrix notation as

$$\overline{\delta f^2} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial x} \end{bmatrix} \begin{bmatrix} \overline{\delta x^2} & \overline{\delta x \delta y} \\ \overline{\delta x \delta y} & \overline{\delta y^2} \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial x} \end{bmatrix} \quad . \tag{21}$$

The term on the left is the variance of f, which we are trying to determine. The elements of the vector D on the extreme right-hand side are the partial derivatives of f with respect to the variables x and y; for $f = x - y$, they would be 1 and -1. The 2×2 matrix is just the known covariance matrix C for x and y. So Eq. (21) can be written in matrix form as

$$\sigma_f^2 = \tilde{D} C D \quad . \tag{22}$$

4.3.2 Transformations of variables

Our second example involves a change of variables from p_1 and p_2 with a known covariance matrix C_p to new variables x_1 and x_2. For example, this could be from polar coordinates to Cartesians. We want to determine the covariance matrix C_x of the new variables.

Here we have just a 2-to-2 transformation, but the extension to larger dimensionality is straightforward. Indeed the procedure also works when the transformation is to a smaller dimensionality. An example of this would be having a set of 20 points in the (x, y) plane, and fitting them with a straight line $y = a + bx$ with just 2 parameters. In all cases, the aim is to derive the covariance matrix for the new quantities (e.g. a and b in the straight line example).

The procedure follows closely that of Section 4.3.1. We start with

$$\delta x_1 = \frac{\partial x_1}{\partial p_1}\delta p_1 + \frac{\partial x_1}{\partial p_2}\delta p_2 \tag{23}$$

and the analogous equation for δp_2. We then calculate the average values of δx_1^2, δx_2^2 and $\delta x_1 \delta x_2$, identify these as the elements of the covariance matrix C_x of the new variables x_1 and x_2, and write the result in matrix form as

$$C_x = \tilde{T} C_p T \quad . \tag{24}$$

Thus the covariance matrix C_x of the new variables is the old covariance matrix C_p sandwiched between the transformation matrix T and its transpose. The matrix T involves the partial derivatives of x with respect to p, as derived from the transformation equations:

$$T = \begin{bmatrix} \frac{\partial x_1}{\partial p_1} & \frac{\partial x_2}{\partial p_1} \\ \frac{\partial x_1}{\partial p_2} & \frac{\partial x_2}{\partial p_2} \end{bmatrix} \quad . \tag{25}$$

A word of caution: The diagonal elements look intuitively sensible, but the off-diagonal ones are different from each other and care is needed to get them in the correct positions.

Equation 24 for a transformation of variables is similar in structure to Eq. (22) for a function. That is because the latter can be considered as a change of variables from N items to just 1.

4.3.3 A particle physics example

A particle detector observes 'hits' at positions where charged particles have passed through its sensitive elements. The detector's magnetic field bends the charged particles along approximately helical paths. The reconstruction programme finds tracks, each of which is constructed from some of the hits in the detector.

Sometimes a short-lived uncharged particle may decay into a positive and a negative one. We wish to calculate the mass M of the unseen neutral particle using the momentum vectors of the two charged tracks of masses m_1 and m_2. For relativistic particles

$$M^2 \approx 2p_1p_2(1-\cos\theta) + m_1^2 + m_2^2 \quad , \tag{26}$$

where the ps are the magnitudes of the momenta and θ is the initial angle between them. We also want the uncertainty on the mass squared σ_{M^2}, arising from the track uncertainties as specified by their correlation matrices.

Both M and σ_{M^2} require the track parameters μ_o and covariance matrices C_o at the point of their origin. Reconstruction algorithms usually give track parameters μ_m and their covariance matrix C_m at the middle of the measured track. This is primarily because a track's direction and magnitude of its momentum are less correlated at its centre than elsewhere along the track. The transformation equations between μ_o and μ_m are known; the track's direction changes between the two locations as it bends in the magnetic field, and its momentum decreases slightly as it travels through the material of the detector. The covariance matrices are related by

$$C_o = \tilde{T}\, C_m\, T \quad , \tag{27}$$

where the transformation matrix T is obtained from the equations relating μ_o and μ_m.

From Eqs. (26) and (23), we have

$$\sigma^2_{M^2} = \tilde{D}\, C_o\, D \quad , \tag{28}$$

where D is the derivative vector derived from Eq. (26). So finally

$$\sigma^2_{M^2} = \tilde{D}\,\tilde{T}\, C_m\, T\, D \quad . \tag{29}$$

.

What this does is to provide a procedure for calculating the uncertainty on the mass, using the covariance matrices of the centre of tracks' parameters. This means that we do not have to worry about how to incorporate the correlations correctly ourselves; the matrix multiplication does this for us.

An interesting feature about the way the correlation between the momentum of the tracks and the angle between them works is that the configuration on the left of Fig. 21 will have better mass resolution

than that on the right. In the former case, slightly more curved tracks (i.e. lower momenta) will result in a larger angle θ, and this combination affects M in opposite directions (see Eq. (26)).

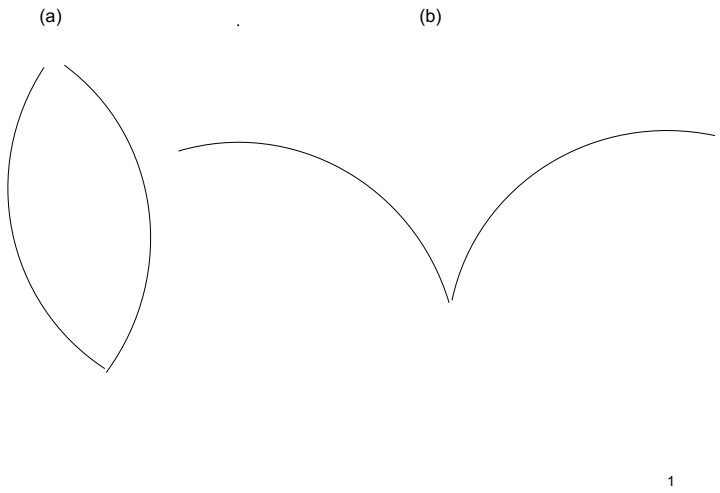


Fig. 21: Track configurations for the decays of unseen neutral particles into 2 charged ones, with different mass resolution. The one on the left will typically have better resolution.

4.4 Combining results

The most important remark about combining results is that it is much better when possible instead to combine data i.e. to perform a combined analysis of the two or more sets of data. But sometimes this is not practically possible.

Most of the discussion in this section is for combining 2 results, but is readily generalised to more. The combinations can be of results from different analyses within the same experiment, or of results from different experiments.

Combinations sometimes produce answers that are counterintuitive, and we discuss some of these.

4.4.1 Uncorrelated measurements

If we have two measurements $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ that are uncorrelated, the combined result is

$$x_{\rm comb} = w_1 x_1 + w_2 x_2, \;\; 1/\sigma^2_{\rm comb} = 1/\sigma_1^2 \; + \; 1/\sigma_2^2 \quad , \tag{30}$$

where the weights w_i are

$$w_i = \frac{1/\sigma_i^2}{1/\sigma_1^2 \; + \; 1/\sigma_2^2} \quad . \tag{31}$$

The weight of each measurement is thus proportional to the reciprocal of its variance. Also the final uncertainty $\sigma_{\rm comb}$ is smaller than the smaller of the individual uncertainties; this is the motivation for the combination.

This result can be derived by finding the value of $x_{\rm comb}$ that minimises the weighted sum of

squared discrepancies S

$$S = (x_1 - x_{\rm comb})^2/\sigma_1^2 + (x_2 - x_{\rm comb})^2/\sigma_2^2 \quad , \tag{32}$$

or by using the best linear unbiassed estimator (BLUE) approach [18].

We already have a potential paradox. Assume we are performing a Poisson counting experiment over two separate hours (e.g. the number of high-energy cosmic rays passing through our detector), and the observed numbers are 100 ± 10 and 1 ± 1. According to Eq. (30), the combined result[7] is 2 ± 1, instead of the more intuitive and correct 50.5 ± 5.0. So the question is why our first usage of Eq. (30) gives a stupid result. The answer is given later in this article.

4.4.2 Correlated single measurements

We now extend the results of the previous section to include correlations. The 2 results we now want to combine have correlation coefficient $\rho = cov(x_1, x_2)/(\sigma_1\sigma_2)$. Instead of using Eq. (32), the result for $x_{\rm comb}$ is now obtained by minimising $S_{\rm correlated}$ with respect to $x_{\rm comb}$, where

$$S_{\rm correlated} = A(x_1 - x_{\rm comb})^2 + B(x_2 - x_{\rm comb})^2 + 2C(x_1 - x_{\rm comb})(x_2 - x_{\rm comb}) \tag{33}$$

and the coefficients A, B and C are the terms of the inverse covariance matrix $\mathbf{R}$ for x_1 and x_2. In matrix notation

$$S_{\rm correlated} = \tilde{\mathbf{x}}\mathbf{R}\mathbf{x} \quad , \tag{34}$$

where $\mathbf{x}$ is the vector with elements $x_1 - x_{\rm best}$ and $x_2 - x_{\rm best}$.

Now for a slight surprise, known as Peelle's Pertinent Puzzle. In the uncorrelated case, the combined value is guaranteed to be within the range of the individual measurements, but this is not so if the correlation coefficient ρ is larger than σ_s/σ_l, where $\sigma_s(\sigma_l)$ is the smaller (larger) of the 2 uncertainties; in that case, $x_{\rm comb}$ lies beyond the measurement with the smaller uncertainty[8]. Again, a little thought shows that this is not unreasonable - see somewhere later in this article for an explanation.

However physicists are wary of extrapolation, and so when ρ is large, it is recommended to use the measurement with the smaller uncertainty as the result of the experiment, and to use the other measurement as a (hopefully satisfactory) check. This is because the amount of extrapolation for the combination is sensitive to the estimated elements of the covariance matrix.

4.4.3 Combining 2-D measurements

It is also possible to combine a pair of results, each of which consists of a pair of correlated quantities. An example of this is provided by Problem 3 in the Appendix. It involves a tracking detector consisting

[7] It is important to note that it is a crime to combine such discrepant results, punishable by being transferred from particle physics to astrology. It is necessary to identify the source of the discrepancy e.g. noise in the first hour, someone switched off the detector early in the second hour. The only reason we use such inconsistent numbers is to make the combined result obviously wrong.

[8] An interesting case is where the second measurement makes use of a subset of the data for first. The correlation coefficient then turns out to be σ_1/σ_2, and the weights are 1 and zero. i.e. the subset is ignored when it is combined with the larger one, as is sensible.

of 3 closely spaced detector planes at negative x and 3 others at positive x. In this simplified version, the tracks are just two-dimensional and straight, so $y = a + bx$. The parameters a and b are extracted from the straight line fit to the data from each set of three planes separately, resulting in correlated a and b for each. Again for simplicity we assume that the separate fits are uncorrelated with each other.

We can again use Eq. (34) to extract the best values $a_{\rm best}$ and $b_{\rm best}$, but here the matrix **R** is block diagonal 4×4, with the 2×2 diagonal blocks being the inverse covariance matrices of the parameters of the two separate straight lines, and the remaining 8 elements being zero[9]. Also the vector **x** now has 4 elements $(a_1 - a_{\rm best},\ b_1 - b_{\rm best},\ a_2 - a_{\rm best},\ b_2 - b_{\rm best})$.

Several interesting types of results are possible.

– If we have two separate measurements of our parameter pair a and b and the measurements' covariance matrices have very different correlations (e.g. opposite signs), the uncertainty on the combination can be very much smaller than the individual uncertainties. The left plot of Fig. 22 illustrates this for our tracking detector problem, where it is not only obvious but also true that the track's gradient is determined much better from the well-separated subdetectors than from the closely-spaced planes of a single sub-detector.

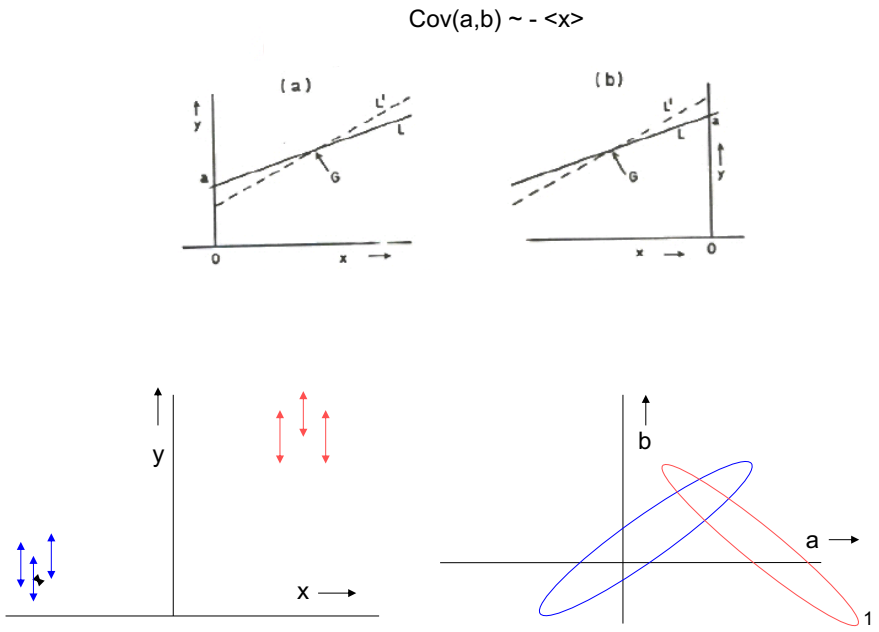


Fig. 22: A tracking detector has three planes at negative x (blue points) and 3 more at positive x (red). Straight lines $y = a + b*x$ are fitted to each of these separately, but have large uncertainties in the intercept a and gradient b. This gives rise to the uncertainty ellipses shown in the bottom right diagram. As can be seen in the top two sketches, the covariance between a and b is proportional to minus the weighted mean $\overline{x}$ of the points used in the fit, and hence have opposite signs for the blue and red lines. When these are combined, the allowed region for a and b is greatly reduced from their original uncertainties, as expected.

A Cosmology example demonstrating the same idea is shown in Fig. 23. Various methods exist

[9] If the two data sets are correlated, then the 4×4 covariance matrix is likely to have all 16 elements non-zero.

for estimating the fractions of dark energy and of matter in the Universe; the uncertainties of these fractions have different correlations in the different approaches. Each one has a large uncertainty, especially on $\Omega_{\text{dark energy}}$, but the combination determines it to much better precision.

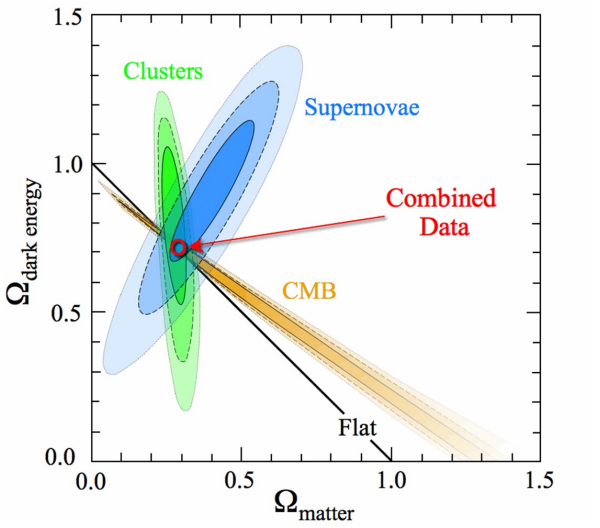


Fig. 23: A plot of the fraction of Cosmology's dark energy against the fraction of matter in the Universe, as deduced from three different types of data. Although none of these alone provides much information on the dark energy fraction, their combination does. This is because of the different correlations for each of the 3 sources of information.

– It can turn out that the combined result can have a outside the range of the individual results a_1 and a_2, and similarly for b; Fig. 24 illustrates this for our tracking example. The individual lines L_1 and L_2 both have positive gradients and negative intercept, while the gradient of the combined line L_{comb} is negative, and its intercept is positive.

– The above example basically involved parameter determination, but we can easily turn in into a Hypothesis Test, where the null hypothesis H_0 is that the gradient is positive while for the alternative H_1 it is negative. So taken individually, the first data set favours H_0 and so does the second, but the combination favours H_1.

A conceivable particle physics example could involve two independent experiments trying to choose between whether the neutrino mass hierarchy is normal or inverted. It could be that each individually favours the inverted hierarchy, but the normal one is preferred by their combination.

This seeming paradox can be made more acute with a medical example. A doctor has a choice of 2 drugs for treating patients with a particular disease. Research data showed that for patients who had asthma in their youth, drug A was better, and similarly for patients who didn't have asthma. But when someone who could not remember whether or not they had had asthma came for treatment, the doctor looked at the combined data, and found that it favoured drug B (see Table 3). This seems curious as this patient either had been or had not been an asthma sufferer, and in either case would have been prescribed drug A.

This is known as the Yule–Simpson paradox.

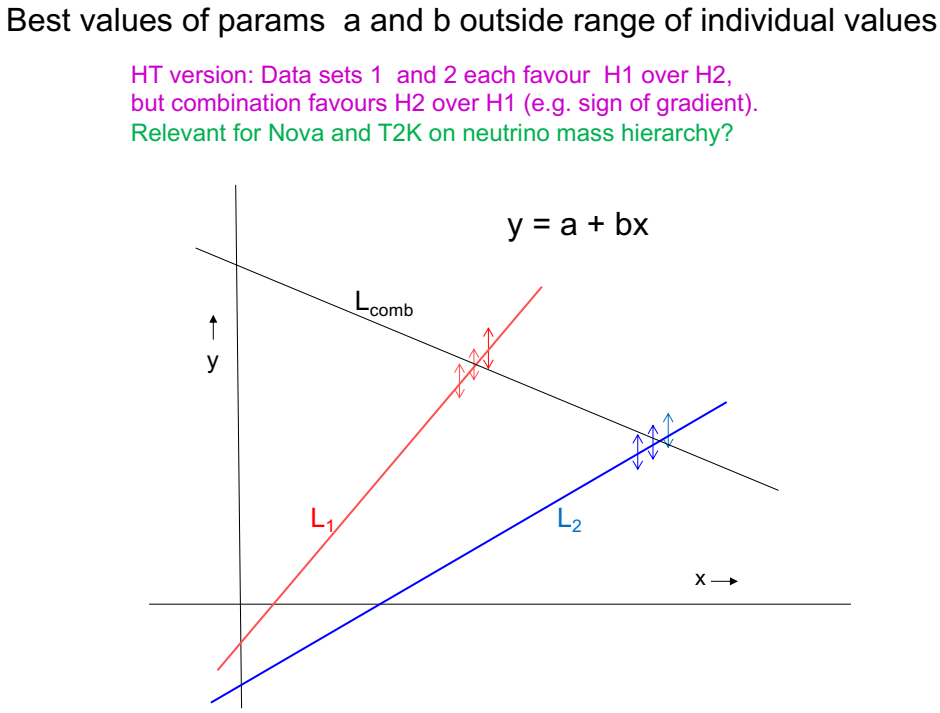


Fig. 24: Tracking example. The gradient and intercept of the combination (black line) are both outside ranges of the individual measurements (red and blue data points and lines).

Table 3: Which drug is better? Each time a drug is tested on a person, a score is assigned depending on the extent to which the drug improved their condition; higher scores correspond to a bigger improvement. Entries in the table show the ratio of the total score to the number of people in that category.

	With asthma	**Without asthma**	Combined
Drug A	400/80=5.0	180/20 = 9.0	580/100 = 5.8
Drug B	80/20 = 4.0	640/80 = 8.0	720/100 = 7.2
	Drug A better	Drug A better	Drug B better

4.5 Estimating the covariance matrix

Section 4.4 dealt with how to use covariance matrices. Here we discuss methods of deriving numerical values for elements of such matrices.

4.5.1 Assessing correlations

Just as with estimating the uncertainty on a single measured quantity, we can try to estimate the covariance matrix. It is easier to do this when the correlation coefficient ρ is zero, +1 or -1.

4.5.2 Adding individual contributions

In many situations, there will be several effects that are contributing to the covariance matrix. Assuming that these various contributions are independent, the overall covariance matrix $\mathbf{C}$ will be the sum of the

individual covariance matrices $\mathbf{C_i}$[10]. This result about adding the individual components is analogous to the situation with the variance for a single measurement being the sum of the variances from different independent sources. Even if the $\mathbf{C_i}$ are fully correlated, in general their sum will not be.

An example involves the mass of the pair-produced W bosons at LEP:

$$e^+e^- \to W^+W^- \quad . \tag{35}$$

The W can decay to a lepton and a neutrino ($l\nu$) or via 2 quark jets (jj). The final states used in the analysis were $l\nu jj$ and 4 jets. A study was made to check whether the masses M_{jj} and $M_{l\nu}$ for the two different decay modes were the same. Given that there were 2 measured quantities, the uncertainties on the masses were specified by a covariance matrix $\mathbf{C}$ to which there were several contributions, which included:

- Statistical uncertainties $\mathbf{C_s}$, where the uncertainties are uncorrelated.
- Beam energy uncertainty $\mathbf{C_b}$. This was relevant because a kinematic fit was performed to improve the mass determination. It had a different magnitude for the 2 modes but they were fully correlated.
- Colour reconnection $\mathbf{C_c}$. For the 4 jet final state, this complicated the assignment of the individual observed particles to the 4 jets, and of which pair of jets came from which W. It did not affect $M_{l\nu}$.

Thus

$$\mathbf{C} = \overbrace{\begin{bmatrix} \sigma_{s1}^2 & 0 \\ 0 & \sigma_{s2}^2 \end{bmatrix}}^{\mathbf{C}_s} + \overbrace{\begin{bmatrix} \sigma_{b1}^2 & \sigma_{b1}\sigma_{b2} \\ \sigma_{b1}\sigma_{b2} & \sigma_{b2}^2 \end{bmatrix}}^{\mathbf{C}_b} + \overbrace{\begin{bmatrix} \sigma_c^2 & 0 \\ 0 & 0 \end{bmatrix}}^{\mathbf{C}_c} + \ldots \quad . \tag{36}$$

4.5.3 Transformations

Quantities of interest for use in further analysis are often derived from some original uncorrelated measurements. An example could be measuring uncorrelated Cartesian coordinates x and y of some object, and then converting them to polar r and θ, which in general will have correlated uncertainties.

Another example is using the measured coordinates along the track of a particle in some detector, and extracting track parameters (position, direction and curvature) via a track fitting procedure. Even when the uncertainties on the original detector hits are uncorrelated, those on the track parameters are not. This is an example where the number of extracted parameters is smaller than the number of measured points.

The details of the way these situations are dealt with were described in Section 4.3.2.

[10]This result contrasts with the situation of combining different estimates of the same pair of quantities, when the final inverse covariance matrix is the sum of the individual inverse matrices. Again this is an extension of the formula $1/\sigma^2 = \Sigma 1/\sigma_i^2$, for the variance σ^2 of the combination of several independent estimates of a single quantity.

4.5.4 Repeated measurements

Just as the uncertainty on a single physical quantity can be estimated from the spread of a series of measurements (even if this may not be the best way of doing it), so too can the covariance for a pair of quantities x and y. It can be derived from the average value of $(x - \bar{x})(y - \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the mean values.

5 Conclusions

You should now have a better appreciation of some of the concepts that are central to statistical procedures for analysing your data. But a deeper understanding cannot be achieved just by listening to lectures and reading books. It is necessary to work through some problems and to engage in working on real data. If someone is asked whether they can play the violin, an unacceptable answer is "I don't know, I have never tried". Similarly with Statistics. A short list of problems is included as an appendix. These aim to illustrate specific statistical issues.

When in your analysis you come across a statistical question that is new to you, it is better first to see whether Statisticians already have a solution, before you engage in trying to find your own method of dealing with it or reinventing the wheel. Usually Statisticians' circular wheels are better than Physicists' square ones (see Fig. 25).

Fig. 25: Anyone interested in circular wheels?

Finally best wishes for your current or future analyses, and especially for its statistical aspects.

6 Appendix: Small set of problems

1) An experiment is searching for quarks of charge 2/3, which are expected to produce 4/9 the ionisation I_0 of unit charged particles. In an exposure in which 10^5 cosmic ray tracks are observed, 1 track has

its ionisation measured as $0.44I_0$. The detector is such that ionisation measurements are Gaussian distributed about their true values with standard deviation σ. Calculate the probability that this could be a statistical fluctuation on the ionisation of a unit charged particle for the following different assumptions:
a) $\sigma = 0.07I_0$ for all 10^5 tracks,
b) For 99% of the tracks $\sigma = 0.07I_0$, while for the remainder it is $0.14I_0$.

2) An experiment is determining the decay rate λ for a new particle X, whose probability density for decay at time t is proportional to $exp(-\lambda t)$. A total of nine decays are observed at decay times 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 picoseconds. Calculate the likelihood function $L(\lambda)$ at suitable values of λ (most easily done by a simple computer programme), and draw a graph of the results. Find the best estimate of λ from the maximum of the likelihood curve, and a "$\pm\sigma$" range for λ by finding the values of λ where the logarithm to the base e of the likelihood function decreases by 0.5 units from its maximum value.

3) i) A detector for tracks has 6 elements at x = -11, -10, -9, +9, +10 and +11 cm, which each measure a track's y-coordinate to an accuracy of ±1 cm. A straight line $y = a + bx$ is fitted (for example by chi-squared) to the data from the 3 elements at positive x (L1); a second line (L2) just for the data at negative x; and a third (L3) to all 6 detector elements. The inverse covariance matrix $\mathbf{R}$ for a and b has elements

$$R_{aa} = \Sigma 1/\sigma_i^2 \quad R_{bb} = \Sigma x_i^2/\sigma_i^2 \quad R_{ab} = \Sigma x_i/\sigma_i^2 \quad , \tag{37}$$

where the measurements are $y_i \pm \sigma_i$ at x_i. Evaluate the covariance matrix for a and b for each of the 3 fits. How do the uncertainties and correlations compare with what you expect?

ii) When two measurements for a pair of quantities are combined optimally, the uncertainties on the combined parameters are such that $R_c = R_1 + R_2$, where R_c is the inverse covariance matrix for the combination, and R_1 and R_2 are those for the separate measurements. Determine the covariance matrix for the combination of the parameters of L1 and L2. Explain why the uncertainties for the combination are considerably smaller than those for L1 and L2 separately.

.

4) Coverage C(μ) is a property of a statistical technique for estimating a range for a parameter μ at a confidence level α (e.g. $68\%, 90\%$ or whatever). It is the fraction of times that, in repetitions of the procedure with different data each with its own statistical fluctuations, the estimated range contains the true value μ.

In a Poisson counting experiment with n observed events, one method of obtaining a range for the Poisson parameter μ uses the estimate $n \pm \sqrt{n}$ i.e. from $n - \sqrt{n}$ to $n + \sqrt{n}$. This is supposed to have 68% coverage. Determine the actual coverage C(μ) at μ = 3.41 and 3.42 as follows: Determine for which measured values of n the nominal range from the "$n \pm \sqrt{n}$" procedure includes the specified true value, and then add up the Poisson probabilities for obtaining these measured values, again assuming the specified value of the Poisson parameter. Explain why a plot of the coverage C(μ) as a function of the Poisson parameter value μ has discontinuities.

The difference in the coverage C at the two values of μ is very similar to a specific Poisson

probability $P_{\rm Poisson}(n|\mu)$. What are the values of n and μ?

5) (a) Explain briefly the Bayesian and frequentists approaches to 'probability'.

(b) Outline how Bayesians and how frequentists would obtain 90% upper limits on the Poisson parameter μ for a counting experiment in which N events are observed.

6) An experiment is searching for a SUSY particle. With no such particle production, 100 events are expected; if the SUSY particle is produced, 110 events are expected. The experiment observes 130 events, which is 3σ above the 'No SUSY' prediction, so the p-value for the null hypothesis is 0.1%. The Lab Publicity Officer announces that we are now 99.9% certain that SUSY has been discovered.

Comment.

7) You have a histogram with 100 bins, and perform a least squares fit with a functional form that has one free parameter β. The best value of β ($\beta_{\rm best}$) results in a weighted sum of squared deviations $S(\beta_{\rm best})$ of 85. The favourite model of a Theorist friend requires the value of $\beta = \beta_{\rm Theorist}$, and she wants to know if her model is ruled out by your data. So you tell her that you have calculated $S(\beta_{\rm Theorist}) = 110$. As she has not attended my lectures, she asks you the implications of this.

You tell her that, according to the χ^2 distribution, $S = 110$ for 100 degrees of freedom is completely acceptable (as is 85 for 99 degrees of freedom), so her model is still viable.

Then you remember that the uncertainty on a parameter can be estimated by finding how far the parameter has to be moved from its best value in order to make S increase by one unit. But $S(\beta_{\rm Theorist})$ is 25 units larger than $S(\beta_{\rm best})$, and so assuming that the shape of $S(\beta)$ near the minimum is parabolic, $\beta_{\rm Theorist}$ is 5 standard deviations from its best value. The probability of this is below 1 part in a million, and so her model is ruled out.

Which answer is correct?

TAKE AWAY POINT FROM PROBLEMS:

1) Effect of mismodelling: The distribution of a measurement is assumed to be exactly Gaussian, whereas in fact there is a 1% tail. Even this small tail has an enormous effect on the result.

2) If you have never calculated a likelihood, this shows you how amazingly simple it can be. And the expected result and an approximate value of its uncertainty are easily known, so you can see whether your values are reasonable.

3) This is a simple example of combining two measurements of the same quantities. The two separate estimates of the gradient have large uncertainties but the gradient for the combination is very much smaller. This problem should help you understand why.

4) People think they understand what coverage is until they see the highly structured plots including jumps for discrete data. By actually calculating the coverage for Poisson data at the two specified values of the Poisson parameter, you should understand the origin of the jumps in coverage, and to get a better feeling for what coverage is.

Section 4.4.1 contained a combination paradox. The reason the result was wrong is because the combination formula is supposed to use the **true** uncertainties on the individual measurements, whereas we used the **estimated** ones. Since we are assuming that the rate is staying constant over the two time intervals, the weights w_1 and w_2 should be the same, rather than 0.99 and 0.01.

The explanation of why the combination of two positively correlated measurements with different uncertain ties can be outside the range of the two measured values is simple. Assume we are measuring a quantity whose true value is 100, and the measurements have uncertainties of 3 and 7. The first measurement might turn out to be somewhat above the true value at, say, 102. Because the second one has a larger uncertainty, it is likely to be further away from 100 than the first one, and because they have a large positive correlation, they probably will be on the same side of 100, e.g. at 108. Thus the true value is outside the range of the measurements, so the fact that the estimated combined value involves extrapolation is not surprising.

References

[1] L. Lyons, "Practical statistics", Proc. 2015 CERN–Latin-American School of High-Energy Physics, Eds. M. Mulders and G. Zanderighi, CERN-2016-005, (CERN, Geneva, 2016), pp. 245–270, doi:10.5170/CERN-2016-005.245.

[2] The LEP Collaborations ALEPH, DELPHI, L3 and OPAL, "Measurement of the mass of the Z boson and the energy calibration of LEP", *Phys. Lett.* **B307** (1993) 187–193, doi:10.1016/0370-2693(93)90210-9.

[3] I. Narsky, "Poisson upper limits in theory and practice", talk at Fermilab's Confidence Limits Workshop (2000). See slide 7, `https://conferences.fnal.gov/cl2k/`.

[4] G.J. Feldman and R.D. Cousins, "A unified approach to the classical statistical analysis of small signals", *Phys. Rev.* **D57** (1998) 3873, arXiv:physics/9711021v2, doi.org/10.1103/PhysRevD.57.3873

[5] G. Cowan *et al.*, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J.* **C71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0.

[6] J. Neyman and E.S. Pearson, "On the problem of the most efficient tests of statistical hypotheses", *Phil. Trans. R. Soc. Lond.* **A231** (1933) 289–337, doi:10.1098/rsta.1933.0009.

[7] D.V. Lindley, "A statistical paradox", Biometrika **44** (1957) 187–192, doi:10.1093/biomet/44.1-2.187; H. Jeffreys, "*Theory of probability*", (Oxford Univ. Press, Oxford, 1939), 3rd ed. publ. 1998, doi:doi.org/10.1093/oso/9780198503682.001.0001.

[8] L. Demortier and L. Lyons, "Testing hypotheses in particle physics: Plots of p_0 versus p_1", arXiv:1408.6123 [stat.ME].

[9] S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *Annals Math. Statist.* **9** (1938) 60–62, doi:10.1214/aoms/1177732360.

[10] R. Klein and A. Roodman, "Blind analysis in nuclear and particle science", *Ann. Rev. Nucl. Part. Sci.* **55** (2005) 141, doi:10.1146/annurev.nucl.55.090704.151521.

[11] A.A. Michelson and E.W. Morley, (1887). "On the relative motion of the Earth and the luminiferous ether", *Am. J. Sci.* **34** (1887) 333–345, doi:10.2475/ajs.s3-34.203.333.

[12] R.D. Cousins, "Why isn't every physicist a Bayesian?", *Am. J. Phys.* **63** (1995) 398–410, doi:10.1119/1.17901

[13] PHYSTAT-Systematics Workshop (2021), https://indico.cern.ch/event/1051224/.

[14] P.D. Dauncey *et al.*, "Handling uncertainties in background shapes: the discrete profiling method", JINST **10** (2015) 04015, doi:10.1088/1748-0221/10/04/P04015.

[15] ATLAS Collaboration, "Measurement of the W-boson mass and width with the ATLAS detector using proton-proton collisions at $\sqrt{s}$ = 7 TeV", *Eur. Phys. J.* **C84** (2024) 1309, doi:10.1140/epjc/s10052-024-13190-x; CMS Collaboration, "High-precision measurement of the W boson mass with the CMS experiment at the LHC", arXiv:2412.13872 [hep-ex].

[16] A.L. Read, "Presentation of search results: the CL_s technique", *J. Phys.* **G28** (2002) 2693–2704, doi:10.1088/0954-3899/28/10/313; T. Junk, "Confidence level computation for combining searches with small statistics", *Nucl. Instrum. Meth.* **A434** (1999) 435, doi:10.1016/S0168-9002(99)00498-2.

[17] ATLAS Collaboration, "A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery", *Nature* **607** (2022) 52–59, doi:10.1038/s41586-022-04893-w; CMS Collaboration, "A portrait of the Higgs boson by the CMS experiment ten years after the discovery", *Nature* **607** (2022) 60–68, doi:10.1038/s41586-022-04892-x.

[18] D. Gibaut, L. Lyons and P. Clifford, "How to combine correlated estimates of a single physical quantity", *Nucl. Instrum. Meth.* **A270** (1988) 110–117, doi:10.1016/0168-9002(88)90018-6.

ACKNOWLEDGEMENTS

I would like to thank Martijn Mulders for inviting me to give these talks at CERN's Graduate School at Peebles, and Sascha Stahl for technical help with my slides and this write-up.