Chapter 6

Confidence intervals

This chapter develops the notion introduced in section 5 on how one defines a statistical error and extends this to look at one and two-sided intervals (see section 6.1). A special case of the one-sided interval is the upper or lower limit we can place on a hypothesised effect or process that has not been observed (see section 6.2). Each of these concepts can be interpreted in terms of a frequentist or Bayesian approach. Section **??** discusses the concept of Bayesian upper limits in the context of a fit to data, and appendix D contains tables of integrals of several common PDFs that can be used to determine confidence intervals and limits.

6.1 Two-sided intervals

The relevance of one and two-sided confidence intervals is discussed in the context of useful distributions used to represent PDFs. In particular the following sections highlight the use of Gaussian, Poisson, and binomial distributions as particular use cases of such intervals.

In general for a distribution $f(x)$ given by some variable x, we can define some region of interest in x called a ***two-sided interval*** such that $x_1 < x < x_2$. If $f(x)$ is a PDF associated with the measurement of an observable x, then the normalised area contained within that interval will correspond to the probability of obtaining a measurement of the observable x that falls within that interval, namely

$$P(x_1 < x < x_2) = \frac{\int_{x_1}^{x_2} f(x)dx}{\int_{-\infty}^{+\infty} f(x)dx}. \tag{6.1.1}$$

The denominator is one if $f(x)$ is a properly normalised PDF. The complement $\overline{P(x_1 < x < x_2)} = 1 - P(x_1 < x < x_2)$ is the probability of obtaining a measurement that falls outside this interval. If the PDF is symmetric about the mid-point in the interval, then the probability of obtaining a measurement of x above or below the interval is equal, and corresponds to exactly half of $\overline{P}$. The probability P is referred to as the ***confidence level*** (CL), or coverage of the interval.

6.2 Upper and lower limit calculations

Upper limits are ***one-sided confidence intervals***, with a given CL, where the lower integration limit has been set to $-\infty$ (or a suitable physical bound). We normally quote that some observable is constrained

to be less than $x_{\rm UL}$ at a CL of $Y\%$. This means that based on our observations, we are able to deduce a one-sided confidence interval between some lower physical limit and $x_{\rm UL}$ that contains $Y\%$ of the area of the PDF $f(x)$ representing the measurement. The corollary of this is that if the mean value and variance of our measurement are the true mean and variance of the observable under study, then $Y\%$ of the time a repeat measurement will yield a result $x < x_{\rm UL}$. The compliment of this probability is the chance that on repeating the measurement we will find a result that satisfies $x \geq x_{\rm UL}$.

The above discussion can be expressed as

$$CL = \frac{\int\limits_{-\infty}^{x_{\rm UL}} f(x)dx}{\int\limits_{-\infty}^{+\infty} f(x)dx}, \tag{6.2.1}$$

where it is assumed that it is sensible to integrate over all x; this is Eq. (**??**), but with $x_{\rm UL}$ in place of x_2.

If it is physically uninteresting or inappropriate to integrate over negative values of x, we can consider modifying the limits of integration to obtain

$$CL = \frac{\int\limits_{0}^{x_{\rm UL}} f(x)dx}{\int\limits_{0}^{+\infty} f(x)dx}, \tag{6.2.2}$$

where we have implicitly multiplied $f(x)$ by some prior $g(x) =$ constant if $x \geq 0$, otherwise $g(x) = 0$. In this approach the PDF $f(x)$ may be a simple distribution or a more sophisticated model such as a χ^2 or likelihood function (see section 8). Often the implicit step of multiplying $f(x)$ by some prior is glossed over when computing such a limit.

The purpose of placing an upper limit on an observable is to be able to state with some confidence at what level we have ascertained that the effect associated with that observable does not exist. In order for such an approach to be meaningful, we would like the chosen CL to be reliable, while at the same time to not be overly conservative. In practice we often quote upper limits with a 90% or 95% CL. For example, we perform a search for some effect X, and place an upper limit of $x_{UL} < 3\times10^{-6}$ at 90% CL on the probability for this effect. Having deduced this we are confident that in 90% of repeated measurements the central value obtained from a subsequent search should be below the obtained upper limit. It follows that in 10% of cases we may find a value that is larger than the quoted upper limit. Here one implicitly assumes that the measurement corresponds to our best estimate of the true value of X. A consequence of this is that often when searching for an effect, we can find ourselves in the situation where we observe the effect occurring with a probability greater than or equal to a previous upper limit. That previous result is not wrong − it was just statistically unlucky in comparison with the subsequent discovery.

Being able to place such constraints on physical processes is a useful way to express the fact that a sought after effect has not been found, and at the same time to incorporate the sensitivity of the measurement, taking into account measurement uncertainties. If there are significant systematic uncertainties associated with a measurement, then these should also be incorporated into an upper limit. There are more sophisticated algorithms that can be used in such cases, for example the unified approach discussed in section 6.6. A brief discussion of a more general Monte Carlo simulated data-based method for computing limits is also discussed (section 6.7).

Lower limits are the lower bound on the extent of an effect, as determined from data. An example of a situation where one might want to place a lower bound on a number is in the computation of an efficiency. If one has a limited number of events to test the efficiency of a detector, and finds that for each of those events the detector is working, then the computed efficiency would be 100%. However as no detector is perfect, it is desirable to be able to place some lower limit on the computed efficiency of the detector. This is an example

of when one would like to compute a limit for a binomial distribution (see section 6.5). Mathematically a lower limit can be expressed in a similar way to Eq. (6.2.1) where

$$CL = \frac{\int_{x_{\rm LL}}^{+\infty} f(x)dx}{\int_{-\infty}^{+\infty} f(x)dx}, \tag{6.2.3}$$

and $x_{\rm LL}$ is the lower bound on x for the desired confidence level. Where necessary one can restrict the upper limits of the above integral at a physical maximum.

6.3 Limits for a Gaussian distribution

As introduced in section 5, statistical errors are Gaussian in nature, where the Gaussian distribution is

$$G(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{6.3.1}$$

The mean value of this distribution is given by $x = \mu$, and the variance is σ^2. The fraction of the area contained between $x = \mu - \sigma$ and $x = \mu + \sigma$ (or $|z| < 1$) is 68.3% to 3 significant figures. This is a two-sided interval, as shown in Figure 6.1 in terms of the z distribution.

The Gaussian distribution is symmetric about its mean value, so it follows that given the fraction of area within $\pm 1\sigma$ is 68.3%, then the fraction of area outside of this interval is 31.7%. This is equally distributed above and below the 1σ interval. Hence there is 15.85% of the area to each side of the 1σ interval. By constructing the two-sided interval $\mu \pm \sigma$ we have divided the PDF into three distinct regions. The interpretation of the the integrals of these regions falls into the following categories (i) the probability of events occurring within the interval (ii) the probability of events occurring below the interval, and (iii) the probability of events occurring above the interval. Naturally there will be many circumstances where we only care if an event lies within or outside of an interval, however sometimes it is useful to make the distinction if an event occurs above or below and interval. If we consider an experiment where the PDF describing some data is a Gaussian distribution that has been derived from a measurement m, then when we repeat that measurement there would be the expectation that the probability of measuring a value m' within 1σ of the original measurement is 68.3%. Similarly the probability that m' differs from m by more than 1σ is 31.7%.

We can extend this concept to include asymmetric intervals, where we let one of the interval limits tend to $\pm\infty$, while the other interval remains finite. In this case, the interval will split the PDF into two distinct parts where the integrals represent (i) the probability of events occurring within the interval, and (ii) the probability of events occurring outside the interval. Figure 6.2 shows a Gaussian PDF with a one-sided interval where $-\infty < z < +1.64$ ($-\infty < x < +1.64\sigma$). The probability corresponding to this interval is

$$P = \int_{-\infty}^{+1.64} G(z)dz = \int_{-\infty}^{+1.64\sigma} G(x;\mu,\sigma)dx, \tag{6.3.2}$$

$$= 0.95, \tag{6.3.3}$$

given that a Gaussian PDF is normalised to a total integral of unity. So if we make a measurement of an observable with a mean value of μ and variance σ^2, then 95% of the time we would expect that a repeat measurement of that observable would yield a result less than $\mu + 1.64\sigma$. Alternatively we may say that $x < \mu + 1.64\sigma$ with 95% CL. The corollary of this is that 5% of the time, the repeat measurement would yield a result of $x \geq \mu + 1.64\sigma$.

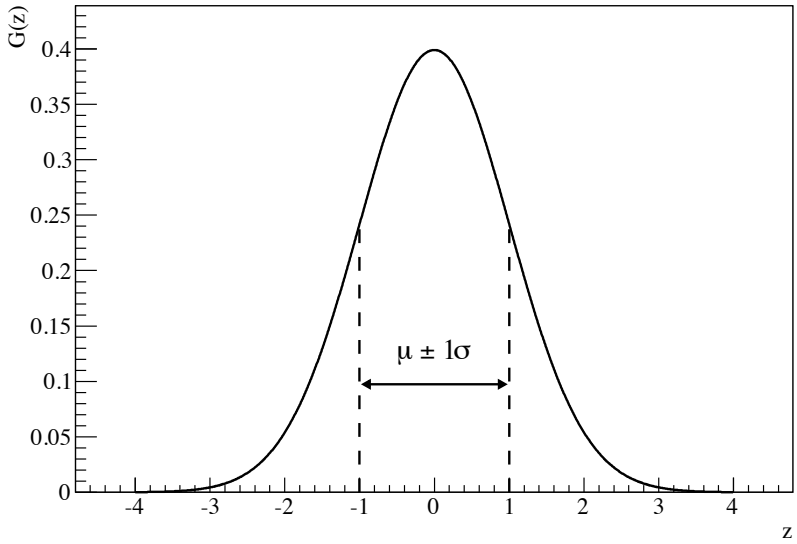

Figure 6.1: A Gaussian distribution indicating the two-sided interval corresponding to $-1 < z < +1$.

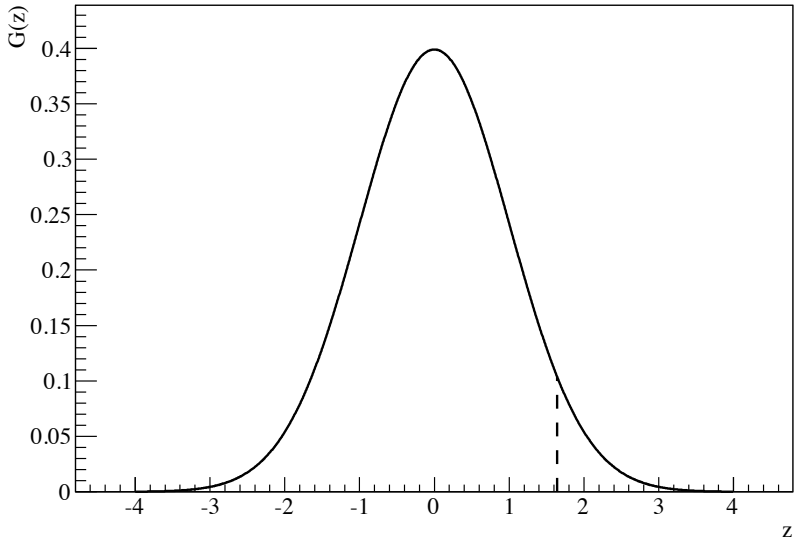

Figure 6.2: A Gaussian distribution indicating the one-sided interval corresponding to $z < 1.64$.

The importance of one and two sided intervals derived from Gaussian PDFs is the result of the observation, in the large statistics limit, that statistical errors are Gaussian in nature. Hence an understanding of Gaussian intervals provides a way of predicting what to expect when an observable is measured, as long as there is some prior indication of what the mean and variance of the distribution should be. Usually this prior indication will be in the form of a previous measurement, or an average of previous measurements (see section 5). One and two sided confidence intervals for the Gaussian probability distribution can be found in Tables D.8 and D.9.

6.4 Limits for a Poisson distribution

Rare processes are described by Poisson statistics. In the limit of small numbers, where the Gaussian approximation is invalid, it is useful to consider one and two sided intervals of a Poisson distribution to obtain intervals and upper limits. The following discussion mirrors the previous section. Two significant differences between the Poisson and Gaussian distributions are that r, unlike x, is a discrete parameter and the Poisson distribution

$$f(r,\lambda) = \frac{\lambda^r e^{-\lambda}}{r!}, \tag{6.4.1}$$

is not symmetric about its mean value λ (see Fig. 4.4), whereas a Gaussian distribution is. This fact is relevant when constructing a two-sided interval and in particular when determining the $\pm 1\sigma$ uncertainty interval on a measured observable in the limit of small statistics. Such a two-sided interval can be constructed by integrating the Poisson distribution for a given r such that the limits λ_1 and λ_2 are equally probable in order to obtain the desired CL. In doing so we naturally determine an asymmetric interval about the mean value λ. If we are measuring some quantity where we wish to express 68.3% coverage about a mean, then an asymmetric PDF such as the Poisson distribution naturally leads to an asymmetric uncertainty. As with the Gaussian case, a one-sided interval is a matter of integrating $f(r,\lambda)$ for a given observed number of events r, to obtain a limit with the desired coverage.

Figure 6.3 shows the one and two sided confidence intervals obtained for λ in a counting experiment as a function of the number of observed signal events r. The upper limit is quoted in terms of both 90% and 95% CL, as these are commonly found as the levels of coverage used in many scientific publications. The corresponding two-sided interval plot also includes the 68.3% CL contours in order to be able to enable a comparison with the Gaussian σ. One sided integral tables of the Poisson PDF can be found in appendix D. The case study described in section 6.8.2 gives an example of using a Poisson distribution to set a confidence interval. While these intervals are represented by a smooth distribution, one should note that the possible outcomes of an experiment are in terms of discrete numbers of events.

The situation encountered where one has a non-zero background component modifies the previous discussion on computing limits. For such a scenario, where one observes $N_{\rm sig}$ signal events and N_{bg} background events, both of which are distributed according to a Poisson distribution with means $\lambda_{\rm sig}$ and $\lambda_{\rm bg}$, respectively. One can show that the sum of these two components is also a Poisson distribution with a mean of $\lambda_{\rm sig} + \lambda_{\rm bg}$. Given sufficient knowledge of $\lambda_{\rm bg}$, one can proceed to compute limits on $\lambda_{\rm sig}$. This situation is discussed in Cowan (1998), which also highlights issues surrounding measurements involving large backgrounds with small numbers of observed signal events. This particular problem is also discussed in the context of the unified approach in section 6.6.

6.5 Limits for a binomial distribution

Setting a confidence interval on some binomial quantity can be done in an analogous way as described above for the Poisson distribution. The likelihood for a binomial distribution is shown in Fig 4.2, with a number of trials N, of which r are successful, and the probability of success or failure is given by p or $1-p$, respectively. Given an experiment consisting of N trials with r successes one can obtain limits on the allowed values for p by integrating the likelihood distribution in order to obtain the desired level of coverage. Figure 6.4 illustrates this for a binomial experiment with 10 trials, of which 2 were successful. The upper limit obtained on p at 95% CL is 0.47. Unlike the Poisson distribution, the binomial distribution has a discrete number of trials N and a discrete number of successes r in the outcome, so it is not possible to produce two-dimensional contours of one and two-sided limits on p scanning through r for a given N. However for a particular scenario one can perform the appropriate one and two-sided integrals of distributions, such as that shown, in order to obtain the corresponding confidence intervals.

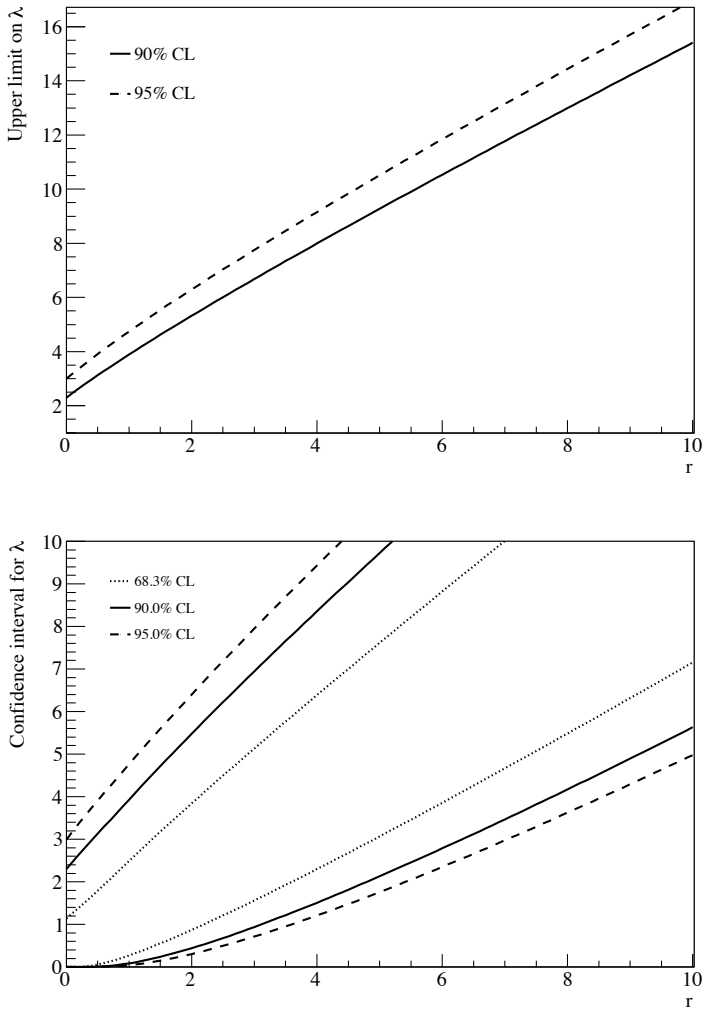

Figure 6.3: (top) upper limit and (bottom) two sided confidence intervals for a Poisson distribution of r observed events for (dotted) 68.3%, (solid) 90%, and (dashed) 95% confidence levels. The allowed region for two-sided confidence intervals is between the curves drawn for a given CL.

The concept of a binomial error was introduced in chapter 5 via measurement of detection efficiency. A pathology was highlighted when the efficiency was either 0% or 100%. For these scenarios one is either minimally or maximally efficient and the binomial error given by Eq. (5.3.2) is determined to be zero. It is now possible to revisit this situation to consider the appropriate outcome of an extreme scenario. As discussed in chapter 4, one can compute a likelihood distribution for a given outcome of a binomial experiment. This distribution (for example see Fig. 4.2) is computed for a given number of trials with a given number of successes, and from this one can determine the relative likelihood for p, the probability of success. In terms of the binomial error p is replaced by the detection efficiency ϵ and the resulting likelihood distribution $\mathcal{L}(\epsilon|N, r)$ can be used to determine a lower (or upper) limit on detection efficiency.

Example: One thousand events are recorded in order to determine the detection efficiency of a silicon detector. In each event the silicon detector produced a clear signal, indicating that there was no recorded inefficiency. Determine a lower limit on the efficiency of this device.

In order to compute the lower limit one must first determine the likelihood distribution $\mathcal{L}(\epsilon|1000, 1000) =$

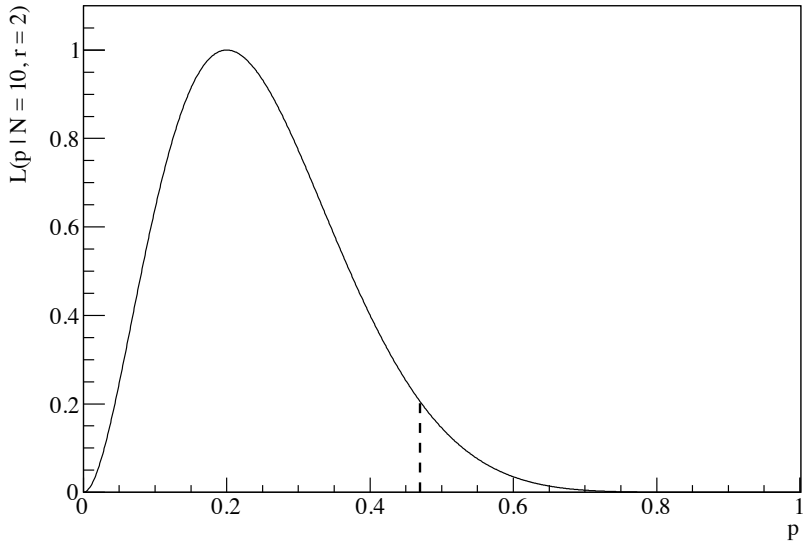

Figure 6.4: The likelihood distribution as a function of p for a binomial experiment with 10 trials, of which 2 were successful. The dashed vertical line corresponds to the 95% CL upper limit.

ϵ^{1000}. This distribution can be integrated from some lower limit $\epsilon_{\rm LL}$ to one in order to obtain the desired level of coverage. As a result of this integration the detector is found to be $> 0.9977\%$ efficient at 90% CL (see Fig. 6.5).

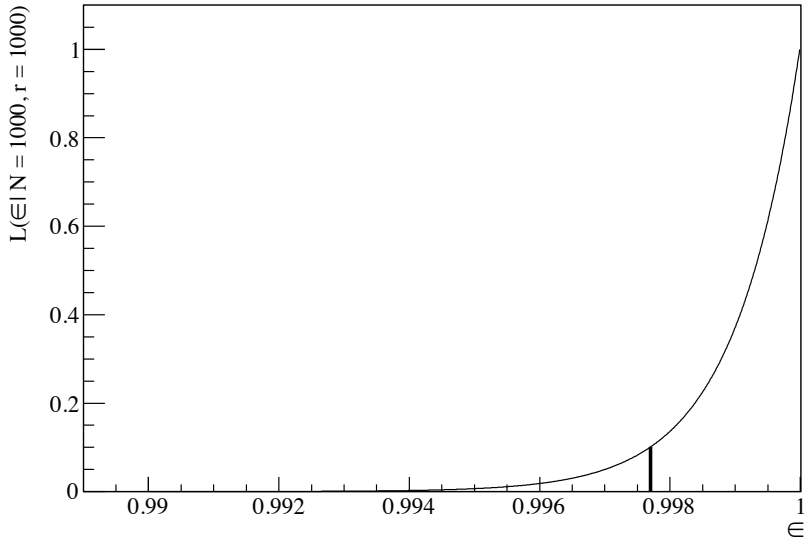

Figure 6.5: The likelihood distribution as a function of ϵ (detection efficiency) for a sample of 1000 events, all of which were detected. The vertical line corresponds to the 90% CL lower limit.

6.6 Unified approach to analysis of small signals

A problem that can be encountered when dealing with small signals is that the constructed intervals can sometimes be an empty set. To avoid such problems Feldman and Cousins (1998) proposed a so-called unified approach to obtain a frequentist confidence interval. In some fields this is simply referred to as the *Feldman-Cousins method*. In this method one constructs an interval $[\mu_1, \mu_2]$ allowed for some parameter μ given a measurement of the observable x using a likelihood ratio based ordering principle. The probability that the true value of the parameter given by μ_t is contained within the interval is denoted by the confidence level α, i.e. for some arbitrary PDF

$$P(\mu \in [\mu_1, \mu_2]) = \alpha. \tag{6.6.1}$$

The unified approach to constructing the confidence interval introduces an ordering principle, where values of μ are sequentially added to the interval in order of a likelihood ranking, stopping when the desired coverage is obtained. The likelihood ranking is obtained by normalising the probability $P(x|\mu)$ for a given value of μ by the probability $P(x|\mu_{\rm best})$ for the largest physically allowed estimate of the the parameter μ, called $\mu_{\rm best}$. For each value of μ one can compute the likelihood ratio

$$R = \frac{P(x|\mu)}{P(x|\mu_{\rm best})}. \tag{6.6.2}$$

As with other frequentist interval estimations, discrete problems will tend to have intervals that suffer from incorrect coverage, however by construction this method either has correct coverage or suffers from over coverage. Other pathologies exist for implementations of this method and are discussed in Feldman and Cousins (1998).

To illustrate how the likelihood ratio ordering algorithm works in practice, the following considers the scenario encountered when attempting to compute a confidence interval for a Poisson observable, resulting from some measurement where one has observed N events with an expectation of N_b background events. Here the parameter μ corresponds to the mean number of signal events that can be inferred from the experiment. The Poisson probability for this scenario is given by

$$P(N|\mu) = \frac{(\mu + N_b)^N e^{-(\mu+N_b)}}{N!}, \tag{6.6.3}$$

as can be seen from Eq. (B.1.13), where here $\lambda = \mu + N_b$, and $r = N$. The value of $\mu_{\rm best}$ is given by

$$\mu_{\rm best} = \max(0, N - N_b), \tag{6.6.4}$$

which is always physical, and represents the most probable value of the signal yield. For a number of observed events N less than the expected number of background, one takes the most probable value of signal $\mu_{\rm best}$ to be zero. If $N > N_b$, then $\mu_{\rm best} = N - N_b$. It follows from Eqns (6.6.2) and (6.6.3) that

$$R = \frac{(\mu + N_b)^N e^{-(\mu+N_b)}}{(\mu_{\rm best} + N_b)^N e^{-(\mu_{\rm best}+N_b)}}. \tag{6.6.5}$$

Given this information it is possible to construct a two dimensional confidence interval as a function of μ and N for a given value of N_b via a two step process. The first step is to construct one-dimensional intervals for a given μ and fixed N_b over a suitably large range of N. Having constructed an interval in N, the second step is to repeat the process for different values of μ. The set of one-dimensional intervals can be used to create a two-dimensional interval. Equation (6.6.4) sets the value of μ_{best}, which will change for each assumed value

of N, and for a given combination of μ, N, μ_{best} and N_b one can compute $P(N|\mu)$, $P(N|\mu_{best})$ and the ratio R. By ranking from most to least probable outcome and then summing up the values of N falling within the desired coverage one can obtain a confidence band in the $N-\mu$ plane. This band corresponds to an interval in N with the desired coverage for the number of observed events for some true value of μ and N_b that one would expect to find on making a measurement. While this information is useful when planning an experimental search (if one builds an experiment to look for some effect, then it is useful to know how many events might be obtained), it is of little use once you have performed a measurement and obtained some number of events $N = N_{obs}$. In this second scenario the process of building the two-dimensional interval continues as one can iterate through different possible values of μ in order to determine a set of confidence intervals. Together this set provides us with the two-dimensional confidence interval in the $N-\mu$ plane. There are two pathologies manifest as a result of this process. The first is that the desired coverage is almost never obtained − for most assumed values of μ this method will over-cover. While not ideal, this situation is an improvement over traditional approaches that can lead to both under and over coverage. Secondly the confidence interval has non-singular values due to the discrete nature of the Poisson distribution. The following example illustrates the process of using the ordering principle to construct a 90% CL interval in N for given values of μ and N_b.

Example: Consider the scenario where one expects to observe $\mu = 2$ signal events with a background $N_b = 3$ in a number counting experiment. What is the 90% CL on N for this? In other words, with 90% confidence how many events will this experiment actually observe? In order to answer this question one needs to construct a confidence band for $\mu = 2$ and $N_b = 3$, which could be used as a starting point to calculate a two-dimensional confidence region, if one so desired.

For a given value of N the value of $\mu_{\rm best}$ is given by Eq. (6.6.4), and both $P(\mu, N)$ and $P(\mu_{\rm best}, N)$ are determined using Eq. (6.6.3). Using these one can compute R as shown in Table 6.1. The most likely outcome corresponds to the case where one obtains five events based on the expectation of $\mu = 2$ and $N_b = 3$, and this forms the seed about which one constructs an interval. One can see from the table that the probability for this outcome to occur is 17.5% which is quite large, and $R = 1$ by construction.

Table 6.1: Computations used to construct the 90% CL interval in N given an expected background of three ($N_b = 3$) for an expected number of signal events $\mu = 2$. The final column indicates the cumulative coverage of the confidence level constructed by including all values of N (in order of decreasing R) from 1 down to the value of that particular entry.

N	$\mu_{\rm best}$	$P(N\|\mu)$	$P(N\|\mu_{\rm best})$	R	Rank	In interval	CL
0	0	0.0067	0.0498	0.1353	11	no	0.9863
1	0	0.0337	0.1494	0.2256	9	no	0.9615
2	0	0.0842	0.2240	0.3759	7	yes	0.8915
3	0	0.1404	0.2240	0.6266	5	yes	0.7420
4	1	0.1755	0.1954	0.8981	3	yes	0.4972
5	2	0.1755	0.1755	1.0000	1	yes	0.1755
6	3	0.1462	0.1606	0.9103	2	yes	0.3217
7	4	0.1044	0.1490	0.7010	4	yes	0.6016
8	5	0.0653	0.1396	0.4677	6	yes	0.8073
9	6	0.0363	0.1318	0.2752	8	yes	0.9278
10	7	0.0181	0.1251	0.1449	10	no	0.9796

It is interesting to note that on applying the ordering principle for this method one can see that the $N = 6$ outcome is preferred over the $N = 4$ outcome even though $P(6, 2) < P(4, 2)$. The underlying principle of this method is not to consider the probability of a given outcome in isolation, or compared against some other outcome, but to interpret a given outcome in terms of the most probable physical outcome given those same circumstances (i.e. with respect to $P(N, \mu_{\rm best})$. The 90% CL interval constructed in this example is the interval $N \in [2-9]$ events. The coverage corresponding to this interval is actually 92.78%, larger than the required 90% coverage.

6.7 Monte Carlo method

For complicated PDFs it is often convenient to simulate the measurement process using a Monte Carlo (MC) based event simulation. The MC simulation will use a sequence of pseudo-random numbers to generate simulated data based on a model, and that simulated data can be used to place limits on an observable based on an assumed input (e.g. the result of a measurement). Using such a technique one can often generate a large number of simulated data in order to test the measurement procedure. The estimate of an upper limit at some confidence level CL can be determined via

$$CL = \frac{\sum_{i=1}^{x=x_{UL}} y_i x_i}{\sum_{i=1}^{N} y_i x_i}, \tag{6.7.1}$$

where the sum is over the binned data. The x value of the i^{th} bin is x_i, and there are y_i entries in that bin. The value of the upper limit is given by $x = x_{UL}$. Often one finds that it is necessary to interpolate between two adjacent bins of data in order to obtain an estimate that corresponds to the desired confidence level instead of reporting a limit that results in under or over coverage.

A confidence level that lies between bins i and $i+1$ can be estimated assuming that the confidence level varies linearly in x across the two bins. The sum over the data up to and including the i^{th} bin is given by S_i, thus the probability contained within the interval i to $i+1$ is $S_{i+1} - S_i$. Assuming a linear approximation to the change in probability as a function of x, the fraction

$$f = \frac{CL - S_i}{S_{i+1} - S_i}, \tag{6.7.2}$$

can be used to estimate the x value of the desired confidence level, i.e.

$$x_{UL} = x_i + f(x_{i+1} - x_i). \tag{6.7.3}$$

In general this is an issue for frequentist limit calculations. One should check the coverage obtained for a limit to ensure that the difference between the coverage quoted and the target coverage is understood, and where appropriate one should adjust the derived limit to regain the correct level of coverage (see section 6.8.3), or report the actual CL obtained.

6.8 Case studies

This section discusses a few case studies that can be understood by developing some of the concepts and applying some of the techniques discussed in this chapter and earlier parts of the book.

6.8.1 Multivariate normal distribution

The multivariate normal distribution is an extension of a one-dimensional Gaussian to an n dimensional space, where in general one allows for a non-trivial covariance between any combination of pairs of dimensions. This distribution can be used to describe uncertainties (or more precisely confidence regions) obtained when measuring a set of quantities that are correlated with each other, and is often used in the bivariate case ($n = 2$) to illustrate so-called error ellipses in a two-dimensional plane. These ellipses are contours of equal

confidence level values (usually 68.3% to corresponding to a 1σ interval, or some multiple $N\sigma$). The bivariate normal distribution is obtained when allowing for x and y to be correlated and this is given by

$$P(\underline{x}) = \frac{1}{2\pi|V|^{1/2}} e^{-(\underline{x}-\underline{\mu})^T V^{-1}(\underline{x}-\underline{\mu})/2}, \tag{6.8.1}$$

where here V is a 2×2 covariance matrix, $\underline{x}$ is the column matrix with elements x and y, and $\underline{\mu}$ is the corresponding column matrix of the means of x and y. The matrices V and V^{-1} can be written in full as

$$V = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}, \tag{6.8.2}$$

and

$$V^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1/\sigma_x^2 & -\rho/\sigma_x\sigma_y \\ -\rho/\sigma_x\sigma_y & 1/\sigma_y^2 \end{pmatrix}, \tag{6.8.3}$$

where ρ is the Pearson correlation given by Eq. (3.6.5). Hence the bivariate normal distribution can be written in full as

$$\begin{aligned} P(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \\ &\exp\left[\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right]. \end{aligned} \tag{6.8.4}$$

One can see how this relates to the simplified case of two uncorrelated variables x and y, in the limit where $\sigma_{xy} = 0$ ($\rho = 0$), and both V and V^{-1} are diagonal matrices. In this case the distribution in Eq. (6.8.4) reduces to

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-(x-\mu_x)^2/2\sigma_x^2 + (y-\mu_y)^2/2\sigma_y^2}, \tag{6.8.5}$$

which follows directly from the discussion of the Gaussian distribution in section 4.4. Figure 6.6 shows 1, 2, and 3σ contours for a bivariate normal distribution where x and y are uncorrelated i.e. Eq. (6.8.5) as well as a distribution where x and y are correlated i.e. Eq. (6.8.4). In the special case of $\sigma_{xy} = 0$ and $\sigma_x^2 = \sigma_y^2 = 1$ shown in the figure, the error ellipses are concentric circles. If however there is a non-trivial correlation then these ellipses deviate from circles and the major and minor axes of the ellipses are not collinear with the horizontal and vertical axes.

Confidence regions shown in the figure have mean values $\mu_x = \mu_y = 0$. The innermost contour corresponds to a 1σ confidence level, and the successive outer contours correspond to 2σ and 3σ confidence levels, respectively. If these distributions corresponded to the PDFs related to some measurement of the parameters x and y, then a measurement within the innermost contour would be compatible with $(x, y) = (0, 0)$ at a level better than 1σ. A measurement falling within the inner two contours would be compatible with $(x, y) = (0, 0)$ at a level better than 2σ, and so on. If the parameters x and y are correlated, then the confidence level shape changes as can be seen by comparing the left and right handed plots in Figure 6.6. If one neglected the correlation between two variables then, as shown in the figure, it is possible that reported confidence regions would be quite different from the actual confidence regions.

Given a two-dimensional probability density function $P(x, y)$ representing a measurement of two observables x and y, one can integrate over one of the observables in order to obtain some marginal probability density function $P(x)$ or $P(y)$. If $P(x, y) = P(x)P(y)$ this is a straightforward procedure where one can choose to use numerical or analytic means to perform the integral depending on the functional form of either $P(x)$ or

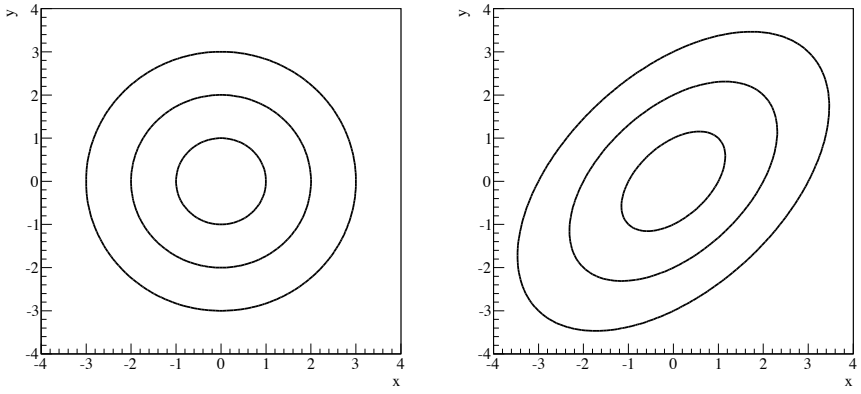

Figure 6.6: The 1, 2, and 3σ contours for bivariate normal distributions for (left) uncorrelated and (right) correlated parameters x and y. The covariance assumed for the plot on the right is $\sigma_{xy} = 0.5$ ($\rho_{xy} = 0.5$).

$P(y)$. In general for a bivariate normal distribution $\rho \neq 0$ and hence it is not possible to express $P(x, y)$ as the product of two independent PDFs. One can still project the 2D distribution onto a single dimension using numerical integration methods to integrate over the unwanted dimension. Alternatively one can perform a transformation of the data from the correlated (x, y) basis to an uncorrelated (u, v) one as discussed in section 3.6.4. After transforming the data to the (u, v) basis the bivariate normal distribution would simplify to the form given in Eq. (6.8.5).

The multivariate normal distribution is a simple extension of Eq. (6.8.1) where

$$P(\underline{x}) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} e^{-(\underline{x}-\underline{\mu})^T V^{-1} (\underline{x}-\underline{\mu})/2}, \tag{6.8.6}$$

and now the column matrices are read as having n elements, and V and V^{-1} are the $n \times n$ covariance matrix and its inverse.

6.8.2 Upper limit calculation from a search

The T2K experiment has been constructed in Japan with the goal of searching for one type of almost massless neutral particle, called a neutrino, to change into another type. In all there are three types of neutrino: the electron neutrino ν_e, the muon neutrino ν_μ, and the tau neutrino: ν_τ. The result of the first search performed at this experiment for ν_μ's to change into ν_e's is reported by Abe *et al.* (2011). The T2K Collaboration determined a background expectation of 1.5 ± 0.3 events, and observed 6 candidate $\nu_\mu \rightarrow \nu_e$ events. The the uncertainty ± 0.3 quoted here is the systematic uncertainty on the background estimation. From this result T2K conclude that the probability to observe 6 or more events given the expected level of background is 7×10^{-3} (equivalent to 2.5σ significance in terms of a Gaussian uncertainty). The details of the calculation performed by this experiment can be found in the reference, and the following provides an illustration of how one might compute this probability using both Frequentist and Bayesian methods.

A possible Frequentist method: One can consider performing an ensemble of Monte Carlo simulated pseudo-experiments, where one simulates a Poisson distribution with a mean $\lambda = 1.5$ corresponding to the mean expected background. If one does this N times, then the probability can be computed (neglecting the systematic uncertainty of 0.3 events) by determining the fraction of simulated experiments that result in six or more events appearing. As N tends to infinity, the probability obtained tends to the true value of the 6 events being the result of a background fluctuation. This situation is illustrated in Fig. 6.7, where

from 10^6 Monte Carlo simulated experiments, one finds a probability of observing 6 or more events, with an expectation of 1.5 background to be 0.45%. One can take into account the systematic uncertainty on the background in a simple way by shifting the expected background level by ± 0.3 events and repeating the ensemble of Monte Carlo simulated pseudo-experiments, or if one assumes that the quoted systematic uncertainty is Gaussian in nature, then it would be possible to convolute a Poisson distribution with $\lambda = 1.5$ events with a Gaussian distribution of width 0.3 events, and use the resulting distribution to simulate the pseudo-experiments. It should be noted that for this example, one could also simply numerically integrate the Poisson distribution given in Eq. (B.1.13) in order to obtain the corresponding probability of observing five or less events, which is discussed in section 6.4. The complement of this would provide the probability of observing six or more.

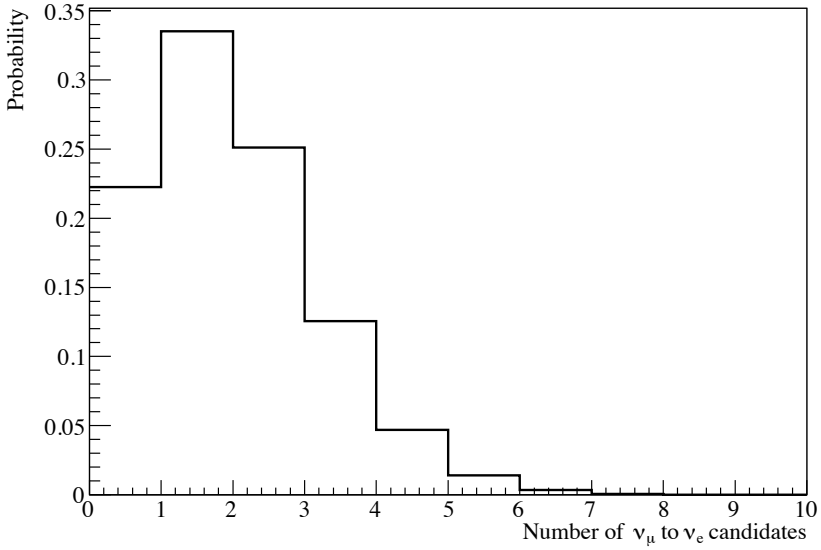

Figure 6.7: A Poisson distribution generated using 10^6 simulated experiment with a mean value of $\lambda = 1.5$.

A more robust frequentist method would be to adopt the unified approach described in section 6.6 which discusses this particular problem.

A possible Bayesian method: Given that the background events are Poisson in nature, one can compute a Bayesian probability by defining a prior $P(B)$ for the number of events observed (e.g. a uniform prior for physically allowed values of the observed number of events), and then multiplying a Poisson distribution by the prior to obtain $P(A|B)P(B)$. One can then compute the probability from the following integral ratio

$$1 - P = 1 - \frac{\int_0^6 P(A|B)P(B)dA}{\int_0^\infty P(A)dA} = \frac{\int_6^{+\infty} P(A|B)P(B)dA}{\int_0^\infty P(A)dA}, \tag{6.8.7}$$

where the lower limit is given by the physical bound of observing no events. This integral ratio depends on the mean value of λ used for the Poisson distribution. Systematic uncertainties can be included in the calculation in the same way as discussed for the Frequentist method. Having computed the probability, one should check the dependence of the result obtained on the choice of the prior probability used.

6.8.3 Coverage: when is a confidence level not a confidence level?

In order to correctly report a confidence interval one has to understand the coverage associated with that interval. If one considers a one-dimensional problem, there are three pieces of information required in order to quote an interval, (i) the confidence level (or coverage) of the interval, (ii) the lower limit, and (iii) the upper limit. By construction a Bayesian limit has the correct coverage for example if one considers the Poisson limit problem discussed above then this is implied by Eq. (6.8.7). A frequentist method however does not necessarily report the desired coverage, as can clearly be seen from the discussion of the unified approach (section 6.6). The discrete nature of the Poisson distribution when applying the unified method leads to a limit that either reports the correct coverage or a more conservative one. This is considered a virtue of the method by the authors. But if one wants to compute a given coverage, for example a 90% CL upper limit, then the discrete nature of that problem generally results in the practitioner choosing which limit to quote: do they quote a 90% CL and report a limit that has a greater than 90% coverage, or do they report the actual coverage obtained and in doing so loose the ease with which their result may be understood in comparison with other experiments? Often the solution taken is the former, with the argument that it is good to be conservative. Formally however this could lead to biased interpretation of results as you have mis-represented the precision of your experiment which is clearly not good practice.

If you are faced with the problem of trying to report a confidence interval where the coverage does not correspond to some "standard" level, e.g. 68.3%, then it is better to stipulate the coverage obtained as accurately as possible so that someone trying to interpret your result is able to do so without any ambiguity or mistake. If you fall foul of the temptation to quote a limit that under or over covers, then your result may be interpreted in a way that is not correct, and in extreme circumstances could even create some level of confusion with that particular measurement. Just as Bayesian statisticians are concerned about prior independence of the results obtained from data analysis, so frequentists should be equally concerned about the coverage of their limits or intervals.