

## Chapter 5

# Interval Estimation

In the previous chapter, we discussed the problem of point estimation, that is, finding something like best estimates for the values of unknown parameters. Simply quoting such numbers as the result of the measurement is not entirely satisfactory, however. At the very least, this lacks any idea of how precise the measurements may be. We now rectify this deficiency, with a discussion of **interval estimation**. The idea is to give a range of values, that is, an interval, for the parameter with a given probability content. The interpretation of the probability depends on whether frequentist or Bayesian approaches are being used.

The general situation can be described as follows: Suppose we sample random variables  $X$  from a probability distribution  $f(x; \theta, \nu)$ , depending on parameters  $\theta$  and  $\nu$ . The parameter space is here divided into two subsets. The first set, denoted  $\theta$ , represents parameters that are of interest to learn about. The second set, denoted  $\nu$ , are parameters needed to completely describe the sampling distribution, but which values are not of much interest. These are called **nuisance parameters**, because lack of knowledge about them can make it difficult make statements concerning  $\theta$ .

The intervals we quote may be **two-sided**, with an upper and a lower end to the range, or **one-sided** with either an upper or lower value quoted, with the other end of the range often constrained by some physical or mathematical condition. One-sided intervals are called upper or lower **limits**. A widely used means of summarizing the result of a measurement of some quantity (parameter), is to quote a two-sided 68% interval. This is in recognition that a sampling from a normal distribution has a  $\sim 68\%$  chance of falling within  $\pm 1$  standard deviation of the mean. In older works the term **probable error** sometimes appears; this corresponds to an interval with 50% probability content instead. The quoting of probable errors has gone out of style in physics applications. In the case of one-sided intervals, a probability content of 90% is often used nowadays, but sometimes 95% or 99% intervals may be given.

## 5.1 Bayesian Intervals

We will call intervals obtained according to Bayesian methodology Bayesian intervals. Often, these are referred to as confidence intervals, but this term will be reserved here for frequentist intervals, so that the distinction may be preserved. Thus, a **Bayesian interval** for a parameter  $\theta$  (or a vector of parameters) is an interval (or region) in parameter space with degree-of-belief probability  $\alpha$  that  $\theta$  is in the interval. The probability  $\alpha$  is called the (Bayesian) **confidence level**.

Conceptually, the construction of a Bayesian interval is very simple: Just find a region of parameter space such that the integral of the posterior distribution (also called the **Bayes distribution**) over that region has probability  $\alpha$ . If there are no nuisance parameters, this is a region  $R$  in:

$$\alpha = \int_R P(\theta; x) d\theta. \quad (5.1)$$

If there are nuisance parameters, simply integrate them out:

$$\alpha = \int_R d\theta \int_{\infty} d\nu P(\theta, \nu; x). \quad (5.2)$$

Integrating out a subset of the parameters is called **marginalizing**.

The Bayesian interval according to the above is ambiguous – there are many solutions for region  $R$ . It is common practice to quote the smallest such region, using an **ordering principle** on  $P$ . That is, we order  $P$  (marginalized as necessary) with respect to  $\theta$  from highest to lowest probability. Then we include regions of  $\theta$  starting at the highest probability until we accumulate a total probability  $\alpha$ . In the case that a uniform prior is used, the integration corresponds to integrating over the likelihood function, up to a fraction  $\alpha$  of its total area. See Fig. 5.1 for an illustration.

For example, consider sampling from a Poisson distribution with mean  $\theta$ . Suppose we obtain a sample result with  $n = 0$  counts, and wish to obtain a 90% confidence Bayes interval assuming a uniform prior. The likelihood function is  $e^{-\theta}$ , and monotonically decreases from its value at  $\theta = 0$ . Thus, using the ordering principle, the interval will be  $(0, \theta_u)$ , where  $\theta_u$  is obtained by solving:

$$0.9 = \int_0^{\theta_u} e^{-\theta} d\theta. \quad (5.3)$$

The result is  $\theta_u = -\ln 0.1 = 2.3$ , and the 90% Bayesian interval for  $\theta$  is  $(0, 2.3)$ .

In the interpretation stage of an analysis, Bayesian intervals may be given, as deemed useful to the consumer. However, when doing a Bayesian analysis, the prior that is being used should always be made clear – it can make a difference. Further, unless the choice of prior is well-constrained it is desirable to check sensitivity to the choice of prior by trying variations. It may be remarked that Bayesian intervals are often used when someone wants to give an upper limit, as the question then being addressed is typically of the sort: “How large do I think the true value of the parameter could be?”

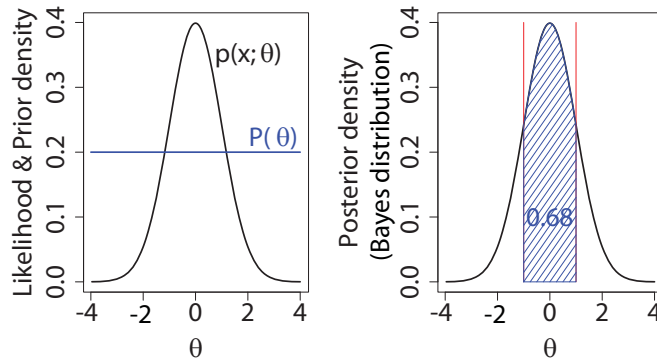


Figure 5.1: Left: Example of a likelihood function,  $L(\theta; x) = p(x; \theta)$ , and a uniform prior  $P(\theta) = \text{constant}$ . Right: The resulting posterior distribution, with the ordered 68% Bayesian interval shown.

## 5.2 Frequency Intervals

The notion of a confidence interval in the frequency sense originated with Jerzy Neyman. We quote here Neyman’s definition of a Confidence Interval (CI) [1]:

“If the functions  $\theta_\ell$  and  $\theta_u$  possess the property that, whatever be the possible value  $\vartheta_1$  of  $\theta_1$  and whatever be the values of the unknown parameters  $\theta_2, \theta_3, \dots, \theta_s$ , the probability

$$P\{\theta_\ell \leq \vartheta_1 \leq \theta_u | \vartheta_1, \theta_2, \dots, \theta_s\} \equiv \alpha, \quad (5.4)$$

then we will say that the functions  $\theta_\ell$  and  $\theta_u$  are the lower and upper confidence limits of  $\theta_1$ , corresponding to the confidence coefficient  $\alpha$ . The interval  $(\theta_\ell, \theta_u)$  is called the Confidence Interval for  $\theta_1$ .”

In this definition,  $\theta_1$  is the parameter of interest, and  $\theta_2, \dots, \theta_s$  are nuisance parameters. The reader is cautioned that some people use instead  $1 - \alpha$  for the “confidence level”. It is crucial to understand that the probability statement is about  $\theta_\ell$  and  $\theta_u$ , and not about  $\theta_1$ . The random variables are  $\theta_\ell$  and  $\theta_u$ , not  $\theta_1$ .

Unfortunately, the confidence interval is not an especially intuitive concept. For many people there is a strong mind set towards the degree-of-belief notion. After all, it is the true value of the parameter that one is ultimately concerned with. The idea of simply describing the data seems sensible, but people have trouble accepting the consequences. This problem was already encountered by Neyman, who goes on to say:

“In spite of the complete simplicity of the above definition, certain persons have difficulty in following it. These difficulties seem to be

due to what Karl Pearson (1938) used to call routine of thought. In the present case the routine was established by a century and a half of continuous work with Bayes's theorem. . ."

Physicists have the same difficulty. Indeed, physicists often re-invent Bayesian methodologies on their own. Physicists are inherently Bayesian – they want to know “the answer”. It can be quite a hurdle to embrace the frequentist methodology with its disregard for the “physical” parameter space. The confusion is enhanced by fact that the same numbers are often quoted for the two types of interval.

Let us re-formulate Neyman's definition as follows: To be clear, for the moment denote the “true” value of  $\theta$  as  $\theta_t$ , and treat  $\theta$  as a variable. In fact, the true value  $\theta_t$  could actually vary from sampling to sampling, but we may think of it as fixed for convenience. Let the dimensionality of the parameter space be denoted  $d$ . It is desired to obtain a confidence region for  $\theta$ , at the  $\alpha$  confidence level. That is, for any  $X = x$ , we wish to have a prescription for finding sets  $C_\alpha(x)$  such that:

$$\alpha = \text{Probability}[\theta_t \in C_\alpha(x)].$$

The probability statement here is over the distribution of possible  $x$  values. We introduce the method with the entire parameter space, then discuss the problem of nuisance parameters. The terms “confidence interval”, “confidence region”, and “confidence set” are considered to be synonyms.

When discussing frequentist statistics, the term **coverage** is often used to describe the probability that a particular algorithm produces an interval containing the true value. Thus, a correct 68% confidence interval “covers” the true value of the parameter in 68% of the samplings. If an algorithm has less than the desired coverage, it is said to **under cover**, and if it has greater than the desired coverage, it is said to **over cover**.

It may be observed that a confidence interval can always be found. Consider, for example, the problem of determining the mean of a normal distribution. Suppose we sample a value  $x$  from an  $N(\theta, 1)$  distribution. We could form a 68% confidence interval for  $\theta$  as follows: Throw a random number,  $r$ , uniform in  $(0, 1)$ . If  $r < 0.68$ , quote the interval  $(-\infty, \infty)$ . Otherwise, quote the null interval. This is a valid confidence interval, including the exact value of  $\theta$ , independent of  $\theta$ , with a frequency of precisely 68%. However, it is a useless exercise – it has nothing to do with the measurement! To be useful, then, we must require more from our interval estimation; we should strive for sufficiency, and perhaps the smallest interval. We could also ask for other properties, such as equal distances on each side of a point estimator.

Let's try again. Notice that 68% of the time we take a sample  $x$  from an  $N(\theta, 1)$  distribution, the value of  $x$  will be within 1 unit of  $\theta$ , see Fig. 5.2. Thus, if we quote the interval  $(x - 1, x + 1)$ , we will have a valid 68% confidence interval for  $\theta$  — the quoted interval will include  $\theta$  with a frequency of precisely 68%. Of course, for any given sample, the quoted interval either includes  $\theta$  or it doesn't. We might even know that it doesn't, e.g., if the interval is outside

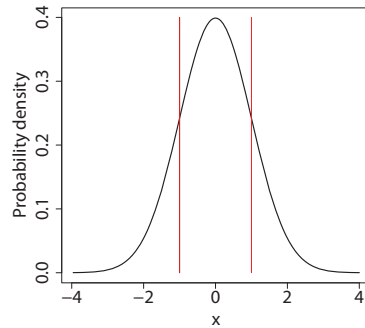


Figure 5.2: The PDF for a  $N(0, 1)$  distribution, with the region from  $-\sigma = -1$  to  $+\sigma = +1$  marked.

some “physical” boundary on allowed values of  $\theta$ . This is irrelevant to the task of describing the data! Notice that these statistics are sufficient.

### 5.2.1 The Basic Method

We construct our confidence region in the Neyman sense, using as criterion an ordering principle.

For any possible  $\theta$  (including possibly “non-physical” values), we construct a set of possible observations  $x$  for which that value of  $\theta$  will be included in the  $\alpha$  confidence region. We call this set  $S_\alpha(\theta)$ . The set  $S_\alpha(\theta)$  is defined as the smallest set for which:

$$\int_{S_\alpha(\theta)} f(x; \theta) \mu(dS) \geq \alpha, \quad (5.5)$$

where  $x \in S_\alpha(\theta)$  if  $f(x; \theta) \geq \min_{x \in S_\alpha(\theta)} f(x; \theta)$ . Given an observation  $x$ , the confidence interval  $C_\alpha(x)$  for  $\theta$  at the  $\alpha$  confidence level, is just the set of all values of  $\theta$  for which  $x \in S_\alpha(\theta)$ . By construction, this set has a probability  $\alpha$  (or more) of including the true value of  $\theta$ .

There are two technical remarks in order:

1. To be general, the measure  $\mu(dS)$  is used; this may be defined as appropriate in the cases of a continuous or discrete distribution. The presence of the inequality in Eqn. 5.5 is intended to handle the situation in which there is discreteness in the sampling space. In this case, the algorithm may err on the side of overcoverage.
2. It is possible that there will be degeneracies on sets of non-zero measure such that there is more than one solution for  $S_\alpha(\theta)$ . In this case, some other criteria will be needed to select a unique set.

**Example: Normal Distribution**

Suppose we sample a vector  $x$  of length  $n$  from an IID normal distribution with standard deviation one:

$$f(X = x; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2}, \quad (5.6)$$

where  $\theta$  is an unknown parameter. Given measurement  $x$ , we want to quote a  $\alpha = 68\%$  confidence interval for  $\theta$ . Let  $m$  be the sample mean,  $m = \frac{1}{n} \sum_{i=1}^n x_i$ . The sample mean is also distributed according to a normal distribution, with standard deviation  $\sigma_m = 1/\sqrt{n}$ . In this case,  $S_{.68}(\theta) = \{m \mid \theta - \sigma_m, \theta + \sigma_m\}$ . Then  $C_{.68} = (m - \sigma_m, m + \sigma_m)$ .

**5.2.2 Contrast with Bayesian Interval**

Let us compare the Bayesian and frequentist intervals with examples. Suppose we sample an estimator  $\hat{\theta}$  from a  $N(\theta, 1)$  distribution. Suppose further that we know that only positive values for  $\theta$  are physically possible. We wish to quote a 68% confidence interval for  $\theta$ , using the ordering principle.

In this case, the frequentist confidence interval is simple. It is given by  $(\hat{\theta} - 1, \hat{\theta} + 1)$ . This is illustrated by the thick black lines in Fig. 5.3. The thinner black line midway between these lines shows the location of the maximum likelihood or least squares estimator according to the frequentist methodology. For large negative values of this estimator,  $\hat{\theta} < -1$ , the 68% confidence interval is entirely in the non-physical regime. Nevertheless, as a description of the measurement, this is the appropriate interval to quote. It is irrelevant that the analyst knows that the quoted region does not include the true value of  $\theta$ . It is only in the frequency sense that the methodology is required to include the true value with 68% probability.

It is true that the frequentist could quote these intervals excluding  $\theta < 0$  when it is known that  $\theta$  cannot be in this region. In this case, sometimes a null interval would be quoted. While this procedure satisfies the frequency criteria, it is less informative as a description of the measurement, and hence not a favored approach.

Fig. 5.3 also shows two different Bayesian intervals as a function of the sampled  $\hat{\theta}$ . The difference between the two intervals is in the choice of prior. The Bayesian imposes a prior that excludes the unphysical region, since the degree-of-belief is zero for unphysical values. Thus, neither Bayesian interval, shown by either the solid red lines or the dashed blue lines, includes any negative values. For small values of  $\hat{\theta}$  the Bayesian intervals become one-sided. It may be observed that Bayesian intervals may be smaller or larger than frequentist intervals.

For another example, suppose that we are looking for a peak in a histogram. Assume large statistics so that the bin contents are distributed normally to a good approximation. We fit the histograms with a linear background plus a signal peak according to a  $N(0, 0.1)$  shape. Consider two examples, shown in

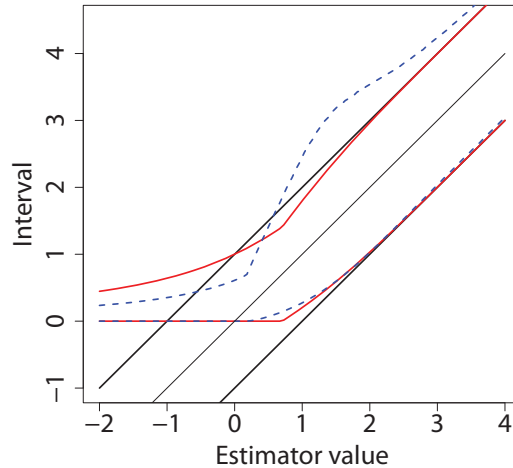


Figure 5.3: Comparison of Bayesian (with two different choices of prior) and frequentist intervals as a function of the (frequentist) MLE. Black: 68% (frequentist) confidence interval; Red: 68% Bayesian interval with a uniform prior,  $P(\theta) \propto \text{constant.}$ ; Dashed blue: 68% Bayesian interval with a square root prior,  $P(\theta) \propto \sqrt{\theta}$ .

Fig. 5.4, measuring event yields in the signal peak according to a least-squares fit. The signal is allowed to be positive or negative in the fit. We assume that on physical grounds the true signal strength can not be negative.

The results of the fits, expressed as either frequentist or Bayesian intervals, are shown in Table 5.1. Depending on the result, the two types of interval may be very similar or very different.

Table 5.1: Results of the fits to the data in Fig. 5.4. A uniform prior is assumed for the Bayesian intervals.

Interval type	Left plot	Right plot
Frequentist	$90.4 \pm 16.4$	$-34.8 \pm 16.4$
Bayesian	$90.4 \pm 16.4$	$0_{-0.0}^{+7.1}$

We'll conclude this section with some general remarks. For the description of the data, frequentist intervals are used. Typically, two-sided intervals are most useful in this context, especially if the underlying sampling distribution is approximately normal. Two-sided intervals better accommodate combining results from different experiments. Notice that there is no consideration of the “significance” of the result here, for example whether the signal strength is deemed to be greater than zero or not. If the statistics is very low and Poisson-

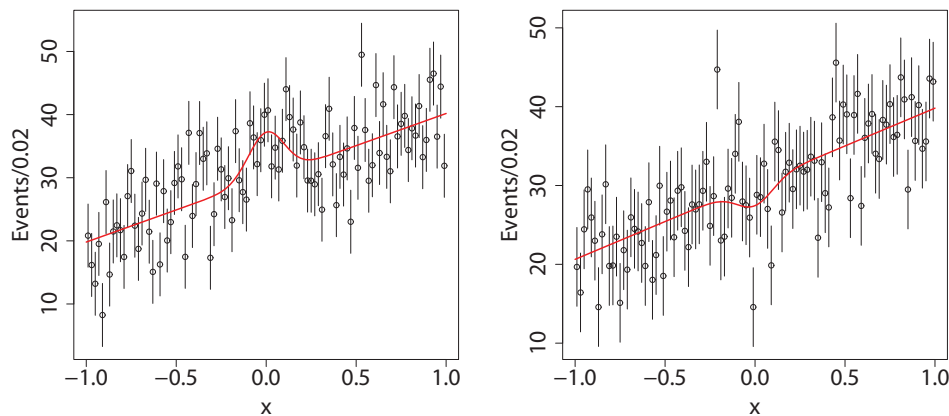


Figure 5.4: Two histograms with fits for a peak superimposed on a background. The background is linear in  $x$ , each bin sampled from a normal distribution with standard deviation of 5. The peak shape is  $N(0, 0.1)$ .

distributed, quoting a two-sided interval becomes more problematic, and it is usually a good idea to also provide the raw counts.

Optionally, we may provide an interpretation of the results, that is, a statement concerning what the true value of the parameter may be. This is the domain of Bayesian statistics. In this case, depending on the situation, a one-sided interval may be of more interest than a two-sided interval, for example, when no significant signal is observed. The choice of prior does make a difference. Common practice in physics is to use ignorance priors, in spite of the ambiguity in choosing such priors.

For a Bayesian analysis, the only ingredient required from the current measurement is the likelihood function. However, in order to perform a frequentist analysis, besides the result of the measurement, it is necessary to know the sampling distribution, up to the unknown parameter(s). The likelihood function is not enough. We will expand on and illustrate this point later.

### 5.2.3 Maximum likelihood analysis

Later we'll discuss several approaches to the construction of confidence intervals. However, it is useful to develop some intuition by first examining some more specific situations. Consider the context of a maximum likelihood analysis.

Let  $\hat{\theta}(x)$  be the maximum likelihood estimator for  $\theta$  for a sampling  $x$ . We may restate the problem in terms of the probability distribution for  $\hat{\theta}$ , motivated by the notion that the maximum likelihood statistic is a sufficient statistic for  $\theta$  if one exists.

Thus, let  $h(\hat{\theta}; \theta)$  be the probability distribution for  $\hat{\theta}$ . Let now  $S_\alpha(\theta)$  be the

smallest set for which:

$$\int_{S_\alpha(\theta)} h(\hat{\theta}; \theta) \mu(dS) \geq \alpha,$$

where  $\hat{\theta} \in S_\alpha(\theta)$  if  $h(\hat{\theta}; \theta) \geq \min_{\hat{\theta} \in S_\alpha(\theta)} h(\hat{\theta}; \theta)$ . Given an observation  $\hat{\theta}$ , the confidence interval  $C_\alpha(\hat{\theta})$  for  $\theta$  at the  $\alpha$  confidence level, is just the set of all values of  $\theta$  for which  $\hat{\theta} \in S_\alpha(\theta)$ . By construction, this set has a probability  $\alpha$  (or more) of including the true value of  $\theta$ .

Consider now the **likelihood ratio**:

$$\lambda(\theta, \hat{\theta}) = \frac{L(\theta; \hat{\theta})}{L(\hat{\theta}; \hat{\theta})}.$$

The denominator is the maximum of the likelihood for the observation  $\hat{\theta}$ . We have  $0 \leq \lambda(\theta, \hat{\theta}) \leq 1$ . For any value of  $\theta$ , we may make a table of possible results  $\hat{\theta}$  for which we will accept the value  $\theta$  with confidence level  $\alpha$ . Consider the set:

$$A_\alpha(\theta) \equiv \{\hat{\theta} | \lambda(\theta, \hat{\theta}) \geq \lambda_\alpha(\theta)\},$$

where  $\lambda_\alpha(\theta)$  is chosen such that  $A_\alpha(\theta)$  contains a probability fraction  $\alpha$  of the sample space for  $\{\hat{\theta}\}$ . That is:

$$P \left[ \lambda(\theta, \hat{\theta}) \geq \lambda_\alpha(\theta); \theta \right] \geq \alpha. \quad (5.7)$$

Notice that we are ordering on likelihood ratio  $\lambda(\theta, \hat{\theta})$ . Sometimes  $\lambda_\alpha(\theta)$  is independent of  $\theta$ .

We then use this table to construct confidence intervals as follows: Suppose we observe a result  $\hat{\theta}_s$ . We go through our table of sets  $A_\alpha(\theta)$  looking for  $\hat{\theta}_s$ . Everytime we find it, we include that value of  $\theta$  in our confidence interval. This gives a confidence interval for  $\theta$  at the  $\alpha$  confidence level. That is, the true value of  $\theta$  will be included in the interval with probability  $\alpha$ . For clarity, we'll repeat this procedure more explicitly in algorithmic form:

### The Method

The algorithm is the following:

1. Find  $\hat{\theta}$ , the value of  $\theta$  for which the likelihood is maximized.
2. For any point  $\theta^*$  in parameter space, form the statistic

$$\lambda(\theta^*, \hat{\theta}) \equiv \frac{L(\theta^*; x)}{L(\hat{\theta}; x)}.$$

3. Evaluate the probability distribution for  $\lambda$  (considering all possible experimental outcomes), under the hypothesis that  $\theta = \theta^*$ . Using this distribution, determine critical value  $\lambda_\alpha(\theta^*)$ .

4. Ask whether  $\widehat{\theta} \in A_\alpha(\theta^*)$ . It will be if the observed value of  $\lambda(\theta^*, \widehat{\theta})$  is larger than (or equal to)  $\lambda_\alpha(\theta^*)$ . If this condition is satisfied, then  $\theta^*$  is inside the confidence region; otherwise it is outside.
5. Consider all possible  $\theta^*$  to determine the entire confidence region.

In general, the analytic evaluation of the probability in step (3) is intractable. We thus usually employ the Monte Carlo method to compute this probability. In this case, steps (3)–(5) are replaced by the specific procedure:

3. Simulate many experiments with  $\theta^*$  taken as the true value(s) of the parameter(s), obtaining for each experiment the result  $x_{\text{MC}}$  and maximum likelihood estimator  $\widehat{\theta}_{\text{MC}}$ .
4. For each Monte Carlo simulated experiment, form the statistic:

$$\lambda_{\text{MC}} \equiv \frac{L(\theta^*; x_{\text{MC}})}{L(\widehat{\theta}_{\text{MC}}; x_{\text{MC}})}.$$

The critical value  $\lambda_\alpha(\theta^*)$  is estimated as the value for which a fraction  $\alpha$  of the simulated experiments have a larger value of  $\lambda_{\text{MC}}$ .

5. If  $\lambda(\theta^*, \widehat{\theta}) \geq \lambda_\alpha(\theta^*)$ , then  $\theta^*$  is inside the confidence region; otherwise it is outside.

In other words, if  $\lambda(\theta^*, \widehat{\theta})$  is larger than at least a fraction  $1 - \alpha$  of the simulated experiments, then  $\theta^*$  is inside the confidence region. That is, find the fraction of simulated experiments for which  $\lambda(\theta^*, \widehat{\theta}) \geq \lambda_{\text{MC}}$ . If this fraction is at least  $1 - \alpha$ , the point  $\theta^*$  is inside the confidence region  $\{\theta\}_{\text{CR}}(x)$ ; otherwise it is outside.

6. This procedure is repeated for many choices of  $\theta^*$  in order to map out the confidence region.

### Example: normal distribution

Let's try a simple example, a single sampling from a  $N(\theta, 1)$  distribution. The probability distribution is:

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

Given a result  $x$ , the likelihood function is:

$$L(\theta; x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

We wish to determine an  $\alpha$  confidence interval for  $\theta$ .

According to our algorithm, we first find the maximum likelihood estimator for  $\theta$ . The result is  $\hat{\theta} = x$ . Then, for any point  $\theta^*$  in parameter space, we form the statistic

$$\lambda \equiv \frac{L(\theta^*; x)}{L(\hat{\theta}; x)} \quad (5.8)$$

$$= e^{-\frac{1}{2}[(x-\theta^*)^2 - (x-\hat{\theta})^2]} \quad (5.9)$$

$$= e^{-\frac{1}{2}(x-\theta^*)^2}. \quad (5.10)$$

We need the probability distribution for  $\lambda$ . Actually, what we want is the probability that  $\lambda < \lambda_{\text{obs}}$ , assuming parameter value  $\theta^*$ . If this probability is greater than  $1 - \alpha$ , then the confidence interval includes  $\theta^*$ . This is the same as asking the related question for  $-2 \ln \lambda$ . But this probability is just the probability that the  $\chi^2$  will be greater than the observed  $\chi^2$ :

$$P = P[\chi^2 > (x - \theta^*)^2].$$

We accept all  $\theta^*$  for which  $P > 1 - \alpha$ . We shall return to this method in section 5.3

#### 5.2.4 Example: Interval Estimates for Poisson Distribution

Interval estimation for a discrete distribution is in principle the same as for a continuous distribution, but the discrete element introduces a complication. We illustrate with the Poisson distribution:

$$g(n, \theta) = \frac{\theta^n e^{-\theta}}{n!}$$

- If we are interested in obtaining an upper limit on  $\theta$ , given an observation  $n$ , at the  $\alpha$  confidence level, the “usual” technique is to solve the following equation for  $\theta_1(n)$ :

$$1 - \alpha = \sum_{k=0}^n g(k; \theta_1(n)).$$

- Similarly, the “usual” lower limit,  $\theta_0(n)$ , is defined by

$$1 - \alpha = \sum_{k=n}^{\infty} g(k; \theta_0(n)).$$

We define  $\theta_0(0) = 0$ .

These “usual” limits are graphed as a function of  $n$  for  $\alpha = 0.9$  in Fig. 5.5. Note that these are the intervals that would be computed in a Bayesian analysis. However, they also satisfy a “shortest interval” criterion, given that they are

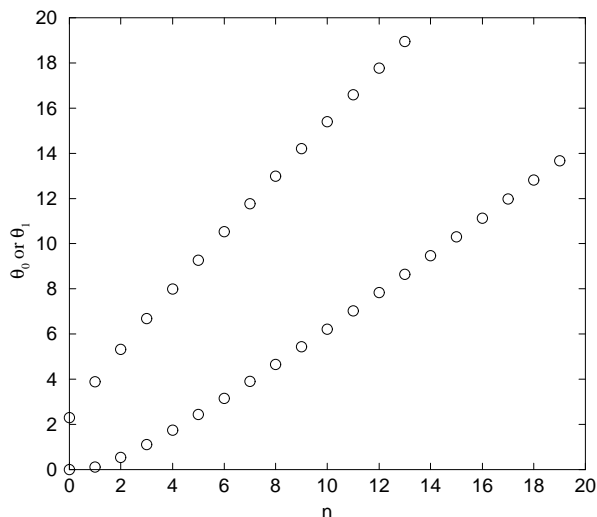


Figure 5.5: The upper and lower  $\alpha = 0.9$  intervals for Poisson sampling as a function of the observed number of counts.

one-sided intervals, that is also desirable in a frequentist analysis. Let us thus examine the coverage properties of these intervals.

To see how the results of this prescription compare with the desired confidence level, we calculate for the lower limit the (frequentist) probability that  $\theta_0 \leq \theta$ :

$$P(\theta_0 \leq \theta) = \sum_{n=0}^{\infty} g(n; \theta) P(\theta_0(n) \leq \theta) \quad (5.11)$$

$$= \sum_{n=0}^{n_0(\theta)} \frac{\theta^n e^{-\theta}}{n!}, \quad (5.12)$$

where the critical value  $n_0(\theta)$  is defined according to:

$$P(\theta_0(n) \leq \theta) = \begin{cases} 1, & n \leq n_0; \\ 0, & n > n_0. \end{cases} \quad (5.13)$$

A similar computation is performed for the upper limit,  $\theta_1(n)$ . The result is shown in Fig. 5.6.

We find that the quoted intervals do not have a constant coverage independent of  $\theta$ . However, their coverage is bounded below by  $\alpha$  for all values of  $\theta$  and hence we accept them as confidence intervals according to Eq. 5.5, although not according to Eq. 5.4. That is, these intervals are “conservative”, in the sense

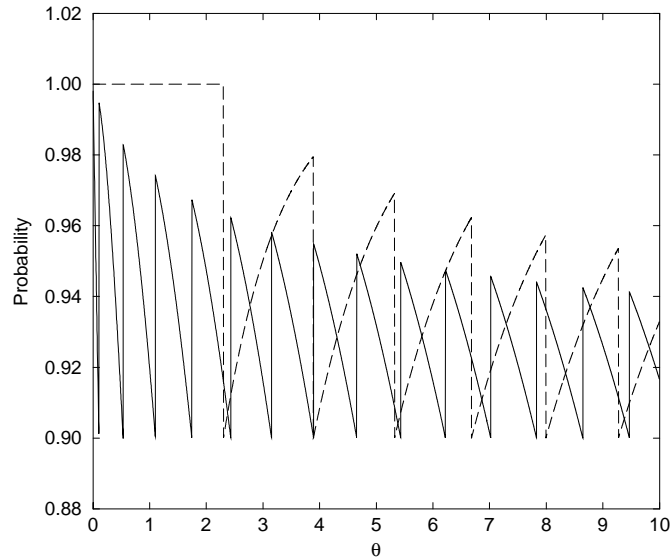


Figure 5.6: Coverage of the one-sided limits at  $\alpha = 0.9$  for the Poisson distribution, as a function of  $\theta$ . Solid curve:  $P(\theta_0(n) \leq \theta)$ . Dashed curve:  $P(\theta_1(n) \geq \theta)$ .

that they include the true value at least as often as the stated probability. The inequality is a consequence of the discreteness of the sampling distribution.

It is amusing that it is possible to quote confidence intervals with exact coverage even for discrete distributions. For example, for a discrete distribution  $G$ , the procedure, given a sample  $n$ , is as follows:

- Define the variable  $y = n + x$ , where  $x$  is sampled from a uniform distribution:

$$f(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note that  $y$  uniquely determines both  $n$  and  $x$ , hence,  $y$  is sufficient for  $\theta$ , since  $n$  is.

- Define  $G(n; \theta)$  as the probability of observing  $n$  or more:

$$G(n; \theta) = \sum_{k=n}^{\infty} g(k; \theta).$$

- Let  $y_0 = n_0 + x_0$ , for some  $x_0$  and  $n_0$ . Then

$$\begin{aligned} P\{y > y_0\} &= P\{n > n_0\} + P\{n = n_0\}P\{x > x_0\} \\ &= G(n_0 + 1; \theta) + g(n_0; \theta)(1 - x_0) \\ &= x_0 G(n_0 + 1; \theta) + (1 - x_0)G(n_0; \theta). \end{aligned}$$

We use this equation to derive exact confidence intervals for  $\theta$ :

- For a lower limit, define “critical value”  $y_c = n_c + x_c$  corresponding to probability  $1 - \alpha_\ell$  by:

$$\begin{aligned} P\{y > y_c\} &= 1 - \alpha_\ell \\ &= x_c G(n_c + 1; \theta) + (1 - x_c) G(n_c; \theta). \end{aligned}$$

- For an observation  $y = n + x$ , define  $\theta_\ell(y)$  according to:

$$1 - \alpha_\ell = xG(n + 1; \theta_\ell) + (1 - x)G(n; \theta_\ell).$$

Whenever  $y > y_c$ , then  $\theta_\ell > \theta$ , and whenever  $y < y_c$ , then  $\theta_\ell < \theta$ . Since the probability of sampling a value  $y < y_c$  is  $\alpha_\ell$ , the probability that  $\theta_\ell$  is less than  $\theta$  is  $\alpha_\ell$ . Therefore, the interval  $(\theta_\ell, \infty)$  is a  $100\alpha_\ell\%$  confidence interval for  $\theta$ .

Confidence intervals for a Poisson distribution may thus be calculated, for any observed sampling  $n$ . For example, according to the above, the prescription for a lower limit on mean  $\theta$  is:

- Define variable  $y = n + x$ , where  $x$  is sampled from a uniform distribution:

$$f(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

$y$  uniquely determines both  $n$  and  $x$ , hence,  $y$  is sufficient for  $\theta$ , since  $n$  is.

- The  $\alpha_\ell$  confidence level lower limit  $\theta_\ell$  is obtained by solving:

$$1 - \alpha_\ell = -x \frac{\theta_\ell^n e^{-\theta_\ell}}{n!} + \sum_{k=n}^{\infty} \frac{\theta_\ell^k e^{-\theta_\ell}}{k!}.$$

We may see the similarity with the “usual” method, and how that method is modified to obtain a confidence interval.

Having given this algorithm, it is important to notice that we haven’t really improved our information with it. All we have done is turn our discrete distribution into a continuous distribution by adding a random number that has nothing to do with the original measurement, and is independent of  $\theta$ . Thus, this idea is virtually unused, and over coverage is accepted for discrete sampling distributions.

### 5.2.5 Confidence Intervals from Pivotal Quantities

With the above preliminary discussion, we discuss several methods used in obtaining (frequentist) confidence intervals. We begin with a method based on “pivotal quantities”.

**Definition 5.1** *Pivotal Quantity:* Consider a sample  $X = (X_1, X_2, \dots, X_n)$  from population  $P$ , governed by parameters  $\theta$ . A function  $R(X, \theta)$  is called **pivotal** iff the distribution of  $R$  does not depend on  $\theta$ .

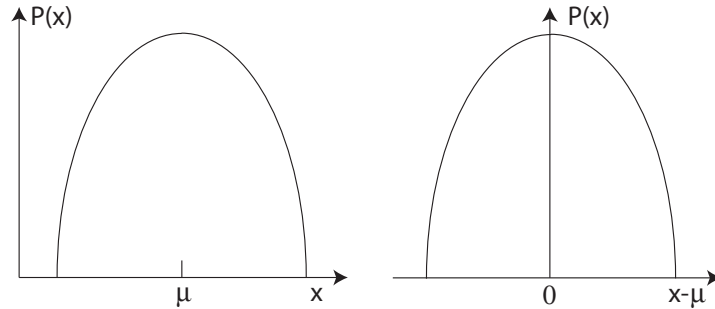


Figure 5.7: Example of a distribution with location parameter  $\mu$ .

The notion of a pivotal quantity is a generalization of the feature of a location parameter: If  $\mu$  is a location parameter for  $X$ , then the distribution of  $X - \mu$  is independent of  $\mu$  (see Fig. 5.7). Thus, a  $X - \mu$  is a pivotal quantity. However, not all pivotal quantities involve location parameters.

If a suitable pivotal quantity is known, it may be used in the calculation of confidence intervals as follows: Let  $R(X, \theta)$  be a pivotal quantity, and  $\alpha$  be a desired confidence level. Find  $c_1, c_2$  such that:

$$P[c_1 \leq R(X, \theta) \leq c_2] \geq \alpha. \quad (5.14)$$

For simplicity, we'll use " $= \alpha$ " henceforth, presuming a continuous distribution. The key point is that, since  $R$  is pivotal,  $c_1$  and  $c_2$  are constants, independent of  $\theta$ .

Now define:

$$C(X) \equiv \{\theta : c_1 \leq R(X, \theta) \leq c_2\}.$$

$C(X)$  is a confidence region with  $\alpha$  confidence level, since

$$P[\theta \in C(X)] = P[c_1 \leq R(X, \theta) \leq c_2] = \alpha.$$

Fig. 5.8 illustrates the idea. In fact, we already used this method, in our example of section 5.2.1.

### Pivotal Quantities: Example

Consider IID sampling  $X = X_1, \dots, X_n$  from a PDF of the form (e.g., a normal distribution):

$$p(x) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right). \quad (5.15)$$

The pivotal quantity method may be applied to obtaining confidence intervals for different cases, according to:

- Case I: Parameter  $\mu$  is unknown and  $\sigma$  is known. Then  $X_i - \mu$ , for any  $i$ , is pivotal. Also, the quantity  $m - \mu$  is pivotal, where  $m$  is the sample

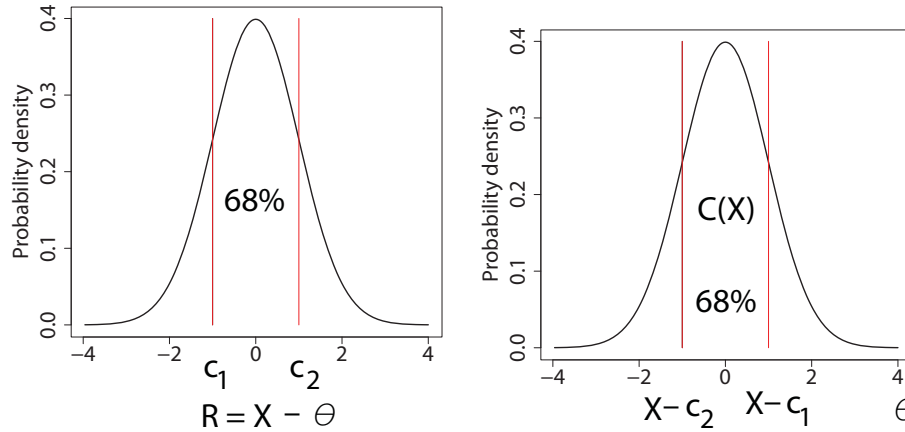


Figure 5.8: The difference between the RV and the mean is a pivotal quantity for a normal distribution. Left: Graph of  $f(X - \theta)$  as a function of  $X - \theta$ , showing the constants  $c_1$  and  $c_2$  that mark of a region with probability 68%. Right: The 68% confidence interval for parameter  $\theta$ .

mean,  $m \equiv \frac{1}{n} \sum_{i=1}^n X_i$ . As a sufficient statistic,  $m$  is a better choice for forming a confidence set for  $\mu$ .

- Case II: Both  $\mu$  and  $\sigma$  are unknown. Let  $s^2$  be the sample variance:

$$s^2 \equiv \frac{1}{n} \sum_{i=1}^n (X_i - m)^2.$$

$s/\sigma$  is a pivotal quantity, and can be used to derive a confidence set (interval) for  $\sigma$  (since  $\mu$  does not appear).

Another pivotal quantity is:

$$t(X) \equiv \frac{m - \mu}{(s/\sqrt{n})}. \quad (5.16)$$

This permits confidence intervals for  $\mu$ :

$$\{\mu : c_1 \leq t(X) \leq c_2\} = (m - c_2 s/\sqrt{n}, m - c_1 s/\sqrt{n})$$

at the  $\alpha$  confidence level, where

$$P(c_1 \leq t(X) \leq c_2) = \alpha.$$

Remark:  $t(X)$  is often called a ‘‘Studentized (Student was a pseudonym for William Gosset, used because his employer did not permit employees to publish.) statistic’’ (though it isn’t a statistic, since it depends also on unknown  $\mu$ ). In the case of a normal sampling, the distribution of  $t$  is Student’s  $t_{n-1}$ .)

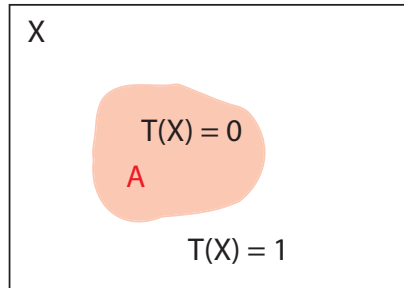


Figure 5.9: Diagram illustrating the test acceptance region, the set of values of random variable  $X$  such that  $T(X) = 0$ .

### 5.3 Confidence Intervals from Inverting Test Acceptance Regions

For any test  $T$  of a hypothesis  $H_0$  versus alternative hypothesis  $H_1$  (see chapter 6 for further discussion of hypothesis tests), we define statistic (“decision rule”)  $T(X)$  with values 0 or 1. Acceptance of  $H_0$  corresponds to  $T(X) = 0$ , and rejection to  $T(X) = 1$ , see Fig. 5.9.

The set  $A = \{x : T(x) \neq 1\}$  is called the **acceptance region**. We call  $1 - \alpha$  the **significance level** of the test if

$$1 - \alpha = P[T(X) = 1], \quad H_0 \text{ is true.} \quad (5.17)$$

That is, the significance level is the probability of rejecting  $H_0$  when  $H_0$  is true (called a **Type I error**).

Let  $T_{\theta_0}$  be a test for  $H_0 : \theta = \theta_0$  with significance level  $1 - \alpha$  and acceptance region  $A(\theta_0)$ . Let, for each  $x$ ,

$$C(x) = \{\theta : x \in A(\theta)\}. \quad (5.18)$$

Now, if  $\theta = \theta_0$ ,

$$P(X \notin A(\theta_0)) = P(T_{\theta_0} = 1) = 1 - \alpha. \quad (5.19)$$

That is, again for  $\theta = \theta_0$ ,

$$\alpha = P[X \in A(\theta_0)] = P[\theta_0 \in C(X)]. \quad (5.20)$$

This holds for all  $\theta_0$ , hence, for any  $\theta_0 = \theta$ ,

$$P[\theta \in C(X)] = \alpha. \quad (5.21)$$

That is,  $C(X)$  is a confidence region for  $\theta$ , at the  $\alpha$  confidence level.

We shall find that ratios of likelihoods can be useful as test statistics. Ordering on the likelihood ratio is often used to define acceptance regions. Hence, the likelihood ordering may be used to construct confidence sets. We have already introduced this in section 5.2.3; however, it bears repeating.

To be explicit, define the ‘‘Likelihood Ratio’’:

$$\lambda(\theta; X) \equiv \frac{L(\theta; X)}{\max_{\theta'} L(\theta'; X)}.$$

For any  $\theta = \theta_0$ , we build an acceptance region according to:

$$A(\theta_0) = \{x : T_{\theta_0}(x) \neq 1\},$$

where

$$T_{\theta_0}(x) = \begin{cases} 0 & \lambda(\theta_0; x) > \lambda_\alpha(\theta_0) \\ 1 & \lambda(\theta_0; x) < \lambda_\alpha(\theta_0), \end{cases}$$

and  $\lambda_\alpha(\theta_0)$  is determined by requiring, for  $\theta = \theta_0$ ,

$$P[\lambda(\theta_0; X) > \lambda_\alpha(\theta_0)] = \alpha.$$

For simplicity, we have here stated things implicitly assuming continuous distributions; section 5.2.3 contains appropriate modifications for possibly discrete distributions.

Let us look at a brief case study of the use of this method. The example is in a two-dimensional parameter space. Suppose we have taken data that is sensitive to mixing of  $D$  mesons, and we wish to use ordering on the likelihood ratio to construct a 95% confidence region. There are two mixing parameters to be determined:

$$\begin{aligned} x' &\equiv \frac{\Delta m}{\Gamma} \cos \delta + \frac{\Delta \Gamma}{2\Gamma} \sin \delta, \\ y' &\equiv \frac{\Delta \Gamma}{2\Gamma} \cos \delta - \frac{\Delta m}{\Gamma} \sin \delta, \end{aligned}$$

where  $\delta$  is an unknown strong phase (between Cabibbo-favored and doubly Cabibbo-suppressed amplitudes). The measurement is sensitive to  $x'^2, y'$ ;  $x'^2 < 0$  is an unphysical region of parameter space, but the maximum of the likelihood may occur at negative  $x'^2$ .

Because the parameter space is two-dimensional, we seek to draw a contour in the  $x'^2, y'$  plane enclosing a 95% confidence region. The procedure is to map the contour by generating ensembles of MC experiments at many points  $(x_0'^2, y_0')$  in  $(x'^2, y')$  space. A point is inside the contour iff the observed likelihood ratio is larger than at least 5% of MC experiments for that parameter point. Let:

$$\begin{aligned} \lambda_{\text{MC}} &= \frac{\mathcal{L}_{(x_0'^2, y_0')(\text{MC})}}{\mathcal{L}_{\text{max}}(\text{MC})} \\ \lambda_{\text{Data}} &= \frac{\mathcal{L}_{(x_0'^2, y_0')(\text{Data})}}{\mathcal{L}_{\text{max}}(\text{Data})} \end{aligned}$$

If  $P(\lambda_{\text{Data}} > \lambda_{\text{MC}}) > 0.05$ , then  $(x_0'^2, y_0')$  is inside (or on) the contour. The result of this procedure is shown in Fig. 5.10.

This method of inverting test acceptance regions is simple in principle, but can be very intensive of computer time. It also must be cautioned that this

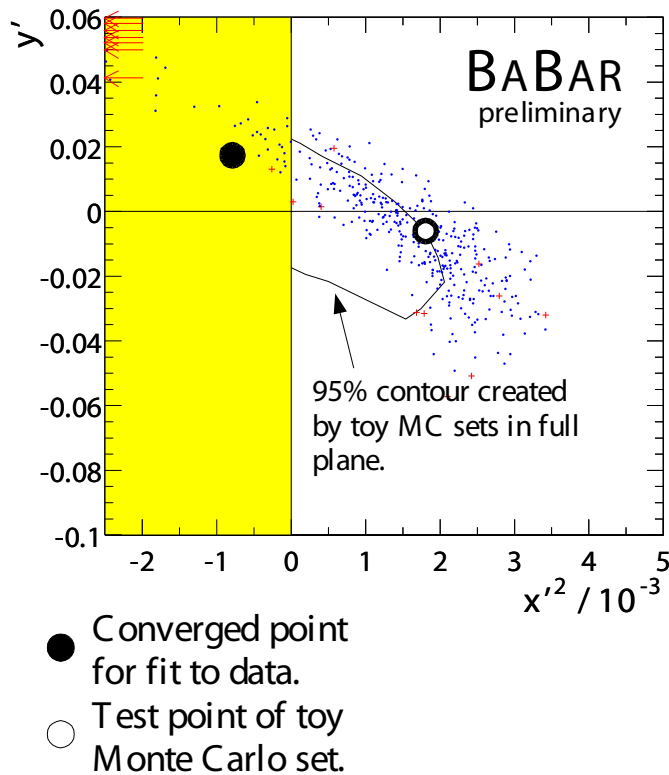


Figure 5.10: The 95% confidence contour for  $D^0$  mixing parameters [2]. The solid circle shows the location where the likelihood is maximal. The open circle shows the position of a point in parameter space used in a Monte Carlo simulation. The blue dots show the results of simulated experiments (with true parameters at the open circle) for which the likelihood ratio is greater than that in the actual experiment, while the red points show simulated experiments where the likelihood ratio is less than in the actual experiment. The yellow region is unphysical.

method may not result in intervals with all the desirable properties. To see this, consider another example, using the  $\chi^2$  statistic:  $\chi^2(x; \theta) \equiv \sum_{i=1}^n (x_i - \theta)^2$ . This statistic is distributed according to a  $\chi^2$  distribution with  $n$  degrees of freedom,  $P_n(\chi^2)$ , if  $\theta$  is known and the sampling is normal. If the sample mean is substituted for  $\theta$ , the distribution is according to a  $\chi^2$  distribution with  $n - 1$  degrees of freedom.

For example, suppose we have a sample of size  $n = 10$  from a  $N(\theta, 1)$  distribution. We wish to determine a 68% confidence interval for unknown parameter  $\theta$ . We'll compare the results of using two different methods.

For the first method, we'll use a pivotal quantity. Let  $\Delta\chi^2(\theta)$  be the difference between the  $\chi^2$  estimated at  $\theta$  and the minimum value, at  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{10} x_i$ :

$$\Delta\chi^2 \equiv \chi^2 - \chi_{\min}^2 \quad (5.22)$$

$$= \sum_{i=1}^{10} [(x_i - \theta)^2 - (x_i - \hat{\theta})^2] \quad (5.23)$$

$$= \sum_{i=1}^{10} [2x_i(\hat{\theta} - \theta) + \theta^2 - \hat{\theta}^2] \quad (5.24)$$

$$= 2 \times 10\hat{\theta}(\hat{\theta} - \theta) + 10\theta^2 - 10\hat{\theta}^2 \quad (5.25)$$

$$= 10(\hat{\theta} - \theta)^2. \quad (5.26)$$

Note that  $\hat{\theta}$  is normally distributed with mean  $\theta$  and variance  $1/10$ , and that  $\hat{\theta} - \theta$  is pivotal, hence so is  $\Delta\chi^2$ . Finding the points where  $\Delta\chi^2 = 1$  corresponds to our familiar method for finding the 68% confidence interval:

$$(\hat{\theta} - 1/\sqrt{10}, \hat{\theta} + 1/\sqrt{10}).$$

This method is illustrated in the left plot in Fig. 5.11.

Now in the second method, we'll invert a test acceptance region based on the  $\chi^2$  statistic. Consider the chi-square goodness-of-fit test for:

$$H_0 : \quad \theta = \theta_0, \quad (5.27)$$

$$H_1 : \quad \theta \neq \theta_0. \quad (5.28)$$

At the 68% significance level, we accept  $H_0$  if

$$\chi^2(\theta_0) < \chi_{\text{crit}}^2,$$

where

$$F(\chi_{\text{crit}}^2, 10) \equiv P(\chi^2 < \chi_{\text{crit}}^2, 10 \text{ DOF}) = 68\%.$$

Note that 10 DOF is used here, since  $\theta_0$  is specified.

If  $\chi_{\text{crit}}^2 > \chi_{\min}^2$ , we have the confidence interval

$$\hat{\theta} \pm \sqrt{(\chi_{\text{crit}}^2 - \chi_{\min}^2)/10},$$

and if  $\chi_{\text{crit}}^2 < \chi_{\text{min}}^2$ , we have a null confidence interval. This method is illustrated in the middle plot in Fig. 5.11.

The distributions of the lengths of the confidence intervals obtained with these two methods are shown in the right plot of Fig. 5.11. The intervals obtained using the pivotal quantity are always of the same length, reflecting the constancy of the resolution of the measurement. The intervals obtained by inverting the test acceptance region are of varying length, including often length zero. Both methods provide valid confidence intervals. However, the pivotal quantity intervals are clearly preferable here. The test acceptance intervals do a much poorer job of describing the precision of the measurement in this example.

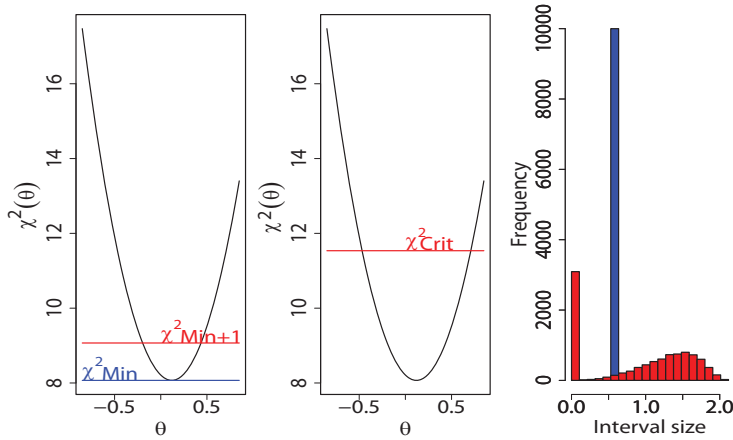


Figure 5.11: Comparison of two confidence intervals for  $\theta$  for samples of size  $n = 10$  from an  $N(\theta = 0, 1)$  distribution. Left: The pivotal quantity method; Middle: The method of inverting a test acceptance region; Right: Comparison of the distribution of the lengths of the intervals for the two methods. The blue histogram is for the pivotal quantity method. The red histogram is the method of inverting a test acceptance region.

### 5.3.1 Nuisance parameters

Let  $f(x; \theta)$  be a probability distribution with a  $d$ -dimensional parameter space. Divide this parameter space into subspaces  $\gamma_1 \equiv \theta_1, \dots, \gamma_k \equiv \theta_k$  and  $\phi_1 \equiv \theta_{k+1}, \dots, \phi_{d-k} \equiv \theta_d$ , where  $1 \leq k \leq d$ . Let  $x$  be a sampling from  $f$ . We wish to obtain a confidence interval for  $\gamma$ , at the  $\alpha$  confidence level. That is, for any observation  $x$ , we wish to have a prescription for finding sets  $R_\alpha(x)$  such that:

$$\alpha = \text{Probability}[\gamma \in R_\alpha(x)].$$

This problem, unfortunately, does not have a general solution, for  $k < d$ . We can see this as follows:

We use the same sets  $S_\alpha(\theta)$  as constructed in section 5.2.1. Given an observation  $x$ , the confidence interval  $R_\alpha(x)$  for  $\gamma$  at the  $\alpha$  confidence level, is just the set of all values of  $\gamma$  for which  $x \in S_\alpha[\theta = (\gamma, \phi)]$ . We take here the union over all values of the nuisance parameters in  $S_\alpha[\theta = (\gamma, \phi)]$  since at most one of those sets is from the true  $\theta$ , and we want to make sure that we include this set. By construction, this set has a probability of at least  $\alpha$  of including the true value of  $\gamma$ . The region may, however, very substantially overcover, depending on the problem.

Suppose we are interested in some parameters  $\mu \subset \theta$ , where  $\dim(\mu) < \dim(\theta)$ . Let  $\eta \subset \theta$  stand for the remaining “nuisance” parameters. If you can find pivotal quantities (e.g., normal distribution), then the problem is solved. Unfortunately, this is not always possible. The approach of test acceptance regions is also problematic:  $H_0$  becomes “composite” (A “simple hypothesis” is one in which the population is completely specified. A composite hypothesis is one that is not simple. See chapter 6.) , since nuisance parameters are unspecified. In general, we don’t know how to construct the acceptance region with specified significance level for

$$H_0 : \mu = \mu_0, \quad \eta \text{ unspecified.}$$

In the following sections we’ll discuss some possible approaches to dealing with the problem of nuisance parameters. Whatever methodology is used, coverage and sensitivity to nuisance parameters can be studied with simulations.

### 5.3.2 Asymptotic Inference

When it is not possible, or practical, to find an exact solution, it may be possible to base an approximate treatment on asymptotic criteria.

**Definition 5.2** Let  $X = (X_1, \dots, X_n)$  be a sample from population  $P \in \mathcal{P}$ , where  $\mathcal{P}$  is the space of possible populations. Let  $T_n(X)$  be a test for

$$H_0 : P \in \mathcal{P}_0 \tag{5.29}$$

$$H_1 : P \in \mathcal{P}_1. \tag{5.30}$$

If

$$\lim_{n \rightarrow \infty} \alpha_{T_n}(P) \leq \alpha \tag{5.31}$$

for any  $P \in \mathcal{P}_0$ , then  $\alpha$  is called an **Asymptotic Significance Level** of  $T_n$ .

**Definition 5.3** Let  $X = (X_1, \dots, X_n)$  be a sample from population  $P \in \mathcal{P}$ . Let  $\theta$  be a parameter vector for  $P$ , and let  $C(X)$  be a confidence set for  $\theta$ . If  $\liminf_n P[\theta \in C(X)] \geq \alpha$  for any  $P \in \mathcal{P}$ , then  $\alpha$  is called an **Asymptotic Significance Level** of  $C(X)$ .

**Definition 5.4** If  $\lim_{n \rightarrow \infty} P[\theta \in C(X)] = \alpha$  for any  $P \in \mathcal{P}$ , then  $C(X)$  is an **Asymptotically Correct** confidence set.

There are many possible approaches, for example, one can look for “Asymptotically Pivotal” quantities; or invert acceptance regions of “Asymptotic Tests”.

### 5.3.3 Profile Likelihood

We may put a lower bound, in the sense of coverage, (up to discreteness issues) on the confidence region for comparison, using the “profile likelihood”.

Consider likelihood  $L(\mu, \eta)$ , based on observation  $X = x$ . Let

$$L_P(\mu) = \sup_{\eta} L(\mu, \eta).$$

$L_P(\mu) = L(\mu, \eta(\mu))$  is called the **Profile Likelihood** for  $\mu$ . The “MINOS” method of error estimation, in the popular MINUIT program uses the profile likelihood.

Let  $\dim(\mu) = r$ . Consider the likelihood ratio test for  $H_0 : \mu = \mu_0$  with

$$\lambda(\mu_0) = \frac{L_P(\mu_0)}{\max_{\theta'} L(\theta')}, \quad (5.32)$$

where  $\theta = \{\mu, \eta\}$ . The set

$$C(X) = \{\mu : -2 \ln \lambda(\mu) \geq c_{\alpha}\}, \quad (5.33)$$

where  $c_{\alpha}$  is the  $\chi^2$  corresponding to the  $\alpha$  probability point of a  $\chi^2$  with  $r$  degrees of freedom, is an  $\alpha$  asymptotically correct confidence set.

### 5.3.4 Conditional Likelihood

Consider likelihood  $L(\mu, \eta)$ . Suppose  $T_{\eta}(X)$  is a sufficient statistic for  $\eta$  for any given  $\mu$ . Then conditional distribution  $f(X|T_{\eta})$  does not depend on  $\eta$ . The likelihood function corresponding to this conditional distribution is called the **Conditional Likelihood**.

Note that estimates (e.g., MLE for  $\mu$ ) based on conditional likelihood may be different than for those based on full likelihood. This eliminates the nuisance parameter problem, if it can be done without too high a price. We’ll see an example later of the use of conditional likelihoods.

### 5.3.5 Case study: Poisson sampling with nuisance parameters

Let us look at a not-uncommon situation as a case study. This is the problem of interval estimation in Poisson sampling, where the numbers of counts may be small, and where nuisance parameters exist. The parameter of interest is related to the mean of the Poisson, but obscured by the presence of a nuisance scale factor and another nuisance additive term.

To make the example more explicit, suppose we are trying to measure a branching fraction for some decay process where we count decays (events) of interest for some period of time. We observe  $n$  events. However, we must subtract an estimated background contribution of  $\hat{b} \pm \sigma_b$  events. Furthermore, the efficiency and parent sample are estimated to give a scaling factor  $\hat{f} \pm \sigma_f$ .

We wish to determine a (frequency) confidence interval for the unknown branching fraction,  $B$ .

- Assume  $n$  is sampled from a Poisson distribution with mean  $\mu = \langle n \rangle = fB + b$ .
- Assume background estimate  $\hat{b}$  is sampled from a normal distribution  $N(b, \sigma_b)$ , with  $\sigma_b$  known.
- Assume scale estimate  $\hat{f}$  is sampled from a normal distribution  $N(f, \sigma_f)$  with  $\sigma_f$  known.

The likelihood function is:

$$L(n, \hat{b}, \hat{f}; B, b, f) = \frac{\mu^n e^{-\mu}}{n!} \frac{1}{2\pi\sigma_b\sigma_f} e^{-\frac{1}{2}\left(\frac{\hat{b}-b}{\sigma_b}\right)^2 - \frac{1}{2}\left(\frac{\hat{f}-f}{\sigma_f}\right)^2}. \quad (5.34)$$

We are interested in the branching fraction  $B$ . In particular, we would like to summarize the data relevant to  $B$ , for example, in the form of a confidence interval, without dependence on the uninteresting quantities  $b$  and  $f$ . We have seen that obtaining a confidence region in all three parameters  $(B, b, f)$  is straightforward. Unfortunately quoting a confidence interval for just one of the parameters is a hard problem in general.

This is a commonly encountered problem, with many variations. There are a variety of approaches that have been used over the years, often without much justification (and often with Bayesian ingredients). In a situation such as this, it is generally desirable to at least provide  $n$ ,  $\hat{b} \pm \sigma_b$ , and  $\hat{f} \pm \sigma_f$ . This provides the consumer with sufficient information to average or interpret the data as they see fit. However, it lacks the compactness of a confidence interval, so let us explore one possible approach, using the profile likelihood. In a difficult situation such as this, it is important to check the coverage properties of the proposed methodology to see whether it is acceptable.

To carry out the approach of the profile likelihood, we vary  $B$ , and for each trial  $B$  we maximize the likelihood with respect to the nuisance parameters  $b$  and  $f$ . We compute the value of  $-2 \ln L$  and take the difference with the value computed at the maximum likelihood over all three parameters. We compare the difference in  $-2 \ln L$  with the 68% probability point of a  $\chi^2$  for one degree of freedom. That is we compare  $-2\Delta \ln L$  with 1. We summarize the algorithm as follows:

1. Write down the likelihood function in all parameters.
2. Find the global maximum.
3. Search in  $B$  parameter for where  $-\ln L$  increases from the minimum by a specified amount (e.g.,  $\Delta = 1/2$ ), re-optimizing with respect to  $f$  and  $b$ .

With the method defined, we ask: does it work? To answer this, we investigate the frequency behavior of this algorithm. For large statistics (normal distribution), we know that for  $\Delta = 1/2$  this method produces a 68% confidence

interval on  $B$ . So we really need to check how far can we trust it into the small statistics regime.

Figure 5.12 shows how the coverage depends on the value of  $\Delta \equiv \Delta \ln L$ . The coverage dependence for a normal distribution is shown for comparison. It may be seen that the Poisson sampling generally follows the trend of the normal curve, but that there can be fairly large deviations in coverage at low statistics. The discontinuities arise because of the discreteness of the Poisson distribution. The small wiggles are artifacts of the limited statistics in the simulation.

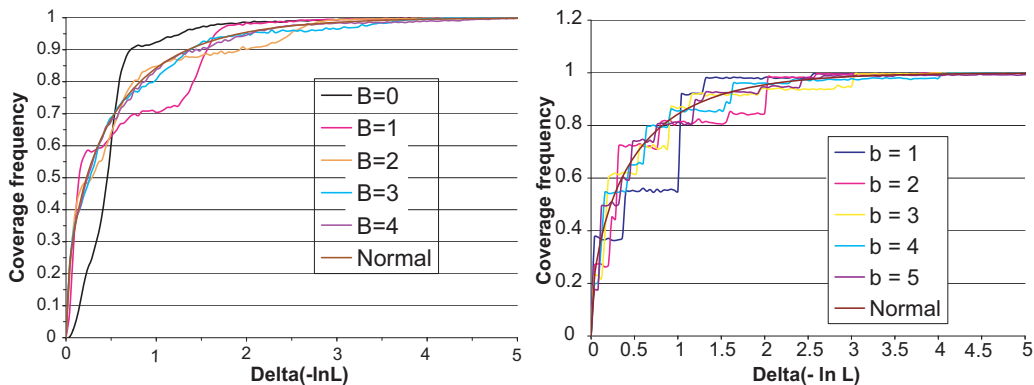


Figure 5.12: Dependence of coverage on  $\Delta \ln L$ . The curves labelled “Normal” show the coverage for a normal distribution. Left: Curves for different values of  $B$  for  $f = 1.0, \sigma_f = 0.1, b = 0.5, \sigma_b = 0.1$ . Right: Curves for different values of  $b$ , for  $B = 0, f = 1, \sigma_f = 0, \sigma_b = 0$ .

In Fig. 5.13 we look at the dependence of the coverage on the background, as well as the uncertainty in the background or the scale factor. Both plots are for a branching fraction of zero. The left plot is for  $\Delta = 1/2$ , which gives 68% confidence intervals if the distribution is normal. In the right plot,  $\Delta$  is changed to 0.8, showing that we can obtain at least 68% coverage for almost all parameter values.

Fig. 5.14 shows the dependence of the coverage on the scale factor  $f$  and its uncertainty, for an expected signal of 1 event and an expected background of 2 events. We can see how additional uncertainty in the scale factor helps the coverage improve; the same is true for uncertainty in the background (Fig. 5.13).

Finally, Fig. 5.15 shows what the intervals themselves look like, for a set of 200 experiments, as a function of the MLE for  $B$ . The MLE for  $B$  can be negative because of the background subtraction. The cluster of points around a value of -3 for the MLE is what happens when zero events occur.

We have illustrated with this example how one can check the coverage for a proposed methodology. With today’s computers, such investigations can often be performed without difficulty.

In this example, we learn that the likelihood method considered works pretty well even for rather low expected counts, for 68% confidence intervals. The

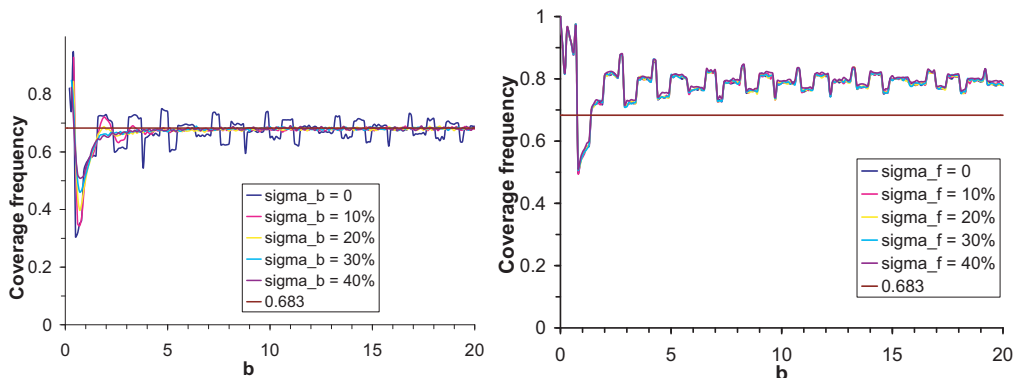


Figure 5.13: Dependence on  $b$ . Left: curves for different values of  $\sigma_b$ , for  $B = 0$ ,  $f = 1$ ,  $\sigma_f = 0$ ,  $\Delta = 1/2$ ; Right: Changing  $\Delta$  to  $\Delta = 0.8$ , for  $B = 0$ ,  $f = 1$ ,  $\sigma_b = 0$ .

choice of  $\Delta = 1/2$  is applicable to the normal distribution, hence we see the central limit theorem at work. To the extent there is uncertainty in  $b$  and  $f$ , the coverage may be improved. However, if  $\sigma_b \approx b$  or  $\sigma_f \approx f$ , we enter a regime not studied here. In that case, it is likely that the normal assumption is not valid. A possible approach at very low statistics is to choose a larger  $\Delta(-\ln L)$  if one wants to insure at least 68%, and is willing to lose some of the power in the data.

It is important to recognize that being good enough for 68% confidence interval doesn't mean good enough for a significance test (the subject of the next chapter), where we are usually concerned with the tails of the distribution.

### 5.3.6 Method of likelihood ratios

We return now more generally to the method of likelihood ratios and ask under what conditions the method works as desired. This is an area of some confusion. In particular, a commonly used likelihood ratio method for obtaining a 68% "confidence interval" is to find the parameter values where the logarithm of the likelihood function decreases by  $1/2$  from its maximum value. This is referred to as a likelihood ratio method.

This is motivated by the normal distribution. Suppose a sample is taken from a normal distribution with known standard deviation:

$$n(x; \theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - w(\theta))^2}{2\sigma^2} \right],$$

where  $w(\theta)$  is assumed to be an invertible function of  $\theta$ . The logarithm of the likelihood function is

$$\ln \mathcal{L}(\theta; x) = -\frac{[x - w(\theta)]^2}{2\sigma^2} + \text{constant}.$$

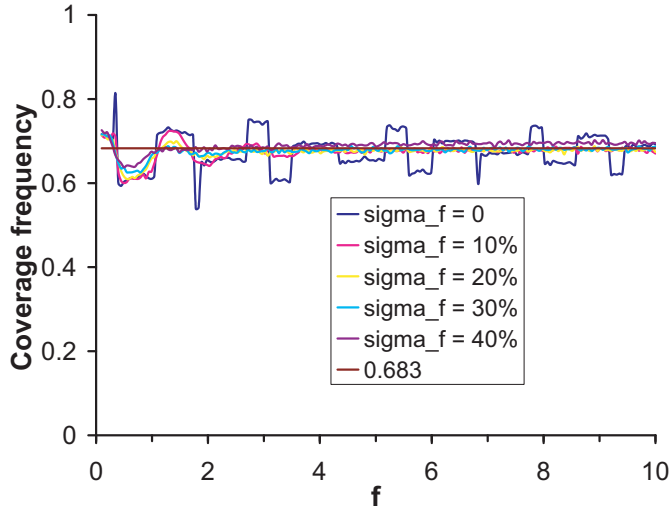


Figure 5.14: Dependence on  $f$  and  $\sigma_f$  for  $B = 1$ ,  $b = 2$ ,  $\sigma_b = 0$ ,  $\Delta = 1/2$

The maximum likelihood is at  $w(\theta^*) = x$ , assuming the solution exists. The points where the function decreases by  $1/2$  from the maximum are at  $w(\theta_{\pm}) = x \pm \sigma$ .

This corresponds to a 68% confidence interval: The probability that  $x$  will lie in the interval  $(w(\theta) - \sigma, w(\theta) + \sigma)$  is 0.68. Thus, in 68% of the times one makes a measurement (samples a value  $x$ ), the interval  $(x - \sigma, x + \sigma)$  will contain the true value of  $w(\theta)$ , and 32% of the time it will not. The probability that the interval  $(\theta_-, \theta_+)$  contains  $\theta$  is 0.68, and  $(\theta_-, \theta_+)$  is therefore a 68% confidence interval. The probability statement is about random variables  $\theta_{\pm}$ , not about  $\theta$ .

The example assumes that the data ( $x$ ) is drawn from a normal distribution. The likelihood function (as a function of  $\theta$ ), on the other hand, is not necessarily normal. As long as  $w(\theta)$  is “well-behaved” (e.g., is invertible), the above method yields a 68% confidence interval for  $\theta$ .

If data is sampled from a distribution which is at least approximately normal (as will be the case in the asymptotic regime if the central limit theorem applies), and the parameter of interest is related to the mean in a well-behaved manner, this method gives a confidence interval. It also has the merit of being relatively easy to calculate.

Now consider a simple non-normal distribution, and ask whether the method still works. For example, consider a “triangle” distribution (see Fig. 5.16:

$$f(x; \theta) = \begin{cases} 1 - |x - \theta| & \text{if } |x - \theta| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The peak of the likelihood function is  $\ln \mathcal{L}(\theta = x; x) = 0$ . Evaluating the  $\ln \mathcal{L} - 1/2$  points:

$$\ln \mathcal{L}(\theta_{\pm}; x) = \ln(1 - |x - \theta_{\pm}|) = -1/2,$$

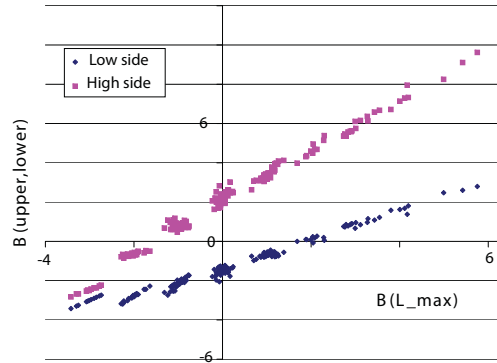


Figure 5.15: What the intervals look like, for a sample of 200 experiments, where  $\Delta = 1/2$ ,  $B = 0$ ,  $f = 1.0$ ,  $\sigma_f = 0.1$ ,  $b = 3.0$ ,  $\sigma_b = 0.1$ .

yields  $\theta_{\pm} = x \pm 0.393$ . Is this a 68% confidence interval for  $\theta$ ? That is, does this interval have a 68% probability of including  $\theta$ ? Since  $\theta_{\pm}$  are linearly related to  $x$ , this is equivalent to asking if the probability is 68% that  $x$  is in the interval  $(\theta - 0.393, \theta + 0.393)$ :

$$P(x \in (\theta - 0.393, \theta + 0.393)) = \int_{\theta - 0.393}^{\theta + 0.393} f(x; \theta) dx = 0.63,$$

which is less than 68%. Thus, this method does not give a 68% confidence interval.

A correct 68% CI can be found by evaluating:

$$\text{Prob}(x \in (x_-, x_+)) = 0.68 = \int_{x_-}^{x_+} f(x; \theta) dx.$$

This gives  $x_{\pm} = \theta \pm 0.437$  (if a symmetric interval is desired), so that the 68% CI given result  $x$  is  $(x - 0.437, x + 0.437)$ , an interval with a 68% probability of containing  $\theta$ . The basic approach thus still works, if we use the points where the likelihood falls to a fraction 0.563 of its maximum, but it is wrong to use the fraction which applies for a normal distribution.

If the normal approximation is invalid, one can simulate the experiment (or otherwise compute the probability distribution) in order to find the appropriate likelihood ratio for the desired confidence level. However, there is no guarantee that even this procedure will give a correct confidence interval, because the

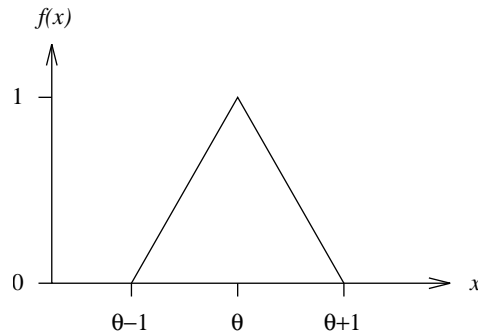


Figure 5.16: A triangle distribution, with location parameter  $\theta$ .

appropriate fraction of the maximum likelihood may depend on the value of the parameter under study. This dependence may be weak enough that the procedure is reasonable in the region of greatest interest.

If  $\theta$  is a location parameter, or a function of a location parameter, the ratio corresponding to a given confidence level will be independent of  $\theta$ , since in that case, the shape of the distribution does not depend on the parameter. This fact can be expressed in the form of a theorem:

**Theorem:** Let  $x$  be a random variable with PDF  $f(x; \theta)$ . If there exists a transformation  $x$  to  $u$ , and an invertible transformation  $\theta$  to  $\tau$ , such that  $\tau$  is a location parameter for  $u$ , then the estimation of intervals by the likelihood ratio method yields confidence intervals. Equivalently, if  $f(x; \theta)$  is of the form:

$$f(x; \theta) = g[u(x) - \tau(\theta)] \left| \frac{du}{dx} \right|,$$

then the likelihood ratio method yields confidence intervals.

If the parameter is a function of a location parameter, then the likelihood function is of the form (for some random variable  $x$ ):

$$\mathcal{L}(\theta; x) = f[x - h(\theta)].$$

Finding the points according to the appropriate ratio to the maximum of the likelihood merely corresponds to finding the points in the pdf such that  $x$  is within a region around  $h(\theta)$  with probability  $\alpha$ . Hence, the quoted interval for  $\theta$  according to this method will be a confidence interval (possibly complicated if the inverse mapping of  $h(\theta)$  is multi-valued).

### 5.3.7 Method of integrating the likelihood function

Another common method of estimating intervals involves integrating the likelihood function. For example, a 68% interval is obtained by finding an interval

which contains 68% of the area under the likelihood function, treated as a function of the parameter for a given value of the random variable.

This method is often interpreted as a Bayesian method, since it yields a Bayesian interval if the prior distribution is uniform. However, it is simply a statement of an algorithm for finding an interval, and we may ask whether it yields a confidence interval, without reference to Bayesian statistics.

This method may be motivated by considering the normal distribution, with mean  $\theta$ , and (known) standard deviation  $\sigma$ :

$$n(x; \theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \theta)^2}{2\sigma^2} \right].$$

The likelihood function, given a measurement  $x$ , is:

$$\mathcal{L}(\theta; x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \theta)^2}{2\sigma^2} \right].$$

If the likelihood function is integrated, as a function of  $\theta$ , to obtain a (symmetric) interval containing 68% of the total area,

$$0.68 = \int_{\theta_-}^{\theta_+} \mathcal{L}(\theta; x) d\theta,$$

we obtain  $\theta_{\pm} = x \pm \sigma$ . There is a 68% probability that the interval given by random variables  $(\theta_-, \theta_+)$  will contain  $\theta$ , and so this is a 68% confidence interval.

We may ask whether the method works more generally, *i.e.*, does this method always give a confidence interval? It may be easily seen that the method also works for the triangle distribution considered in section 5.3.6. However, we may demonstrate that the answer in general is “no”, with another simple example.

Consider the following modified “triangle” distribution:

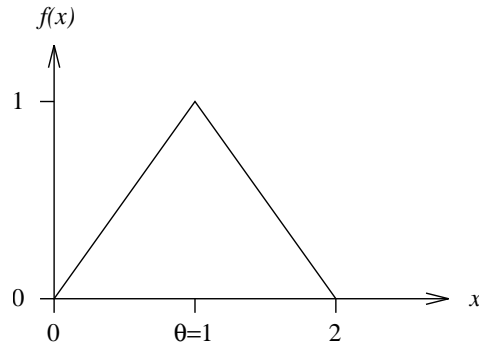
$$f(x; \theta) = \begin{cases} 1 - |x - \theta^2| & \text{if } |x - \theta^2| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

This distribution is shown for  $\theta = 1$  in Fig. 5.17. Sampling from this distribution provides no information on the sign of  $\theta$ . Hence, let  $\theta > 0$  stand for its magnitude. Suppose we wish to obtain a 50% confidence level upper limit (chosen for simplicity) on  $\theta$ , given an observation  $x$ .

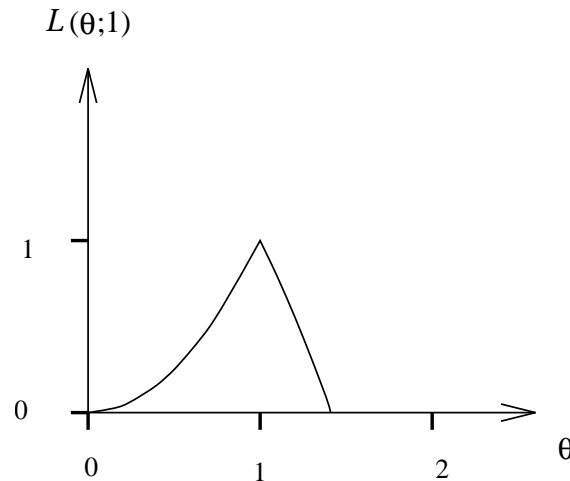
To apply the method of integrating the likelihood function, given an observation  $x$ , solve for  $u(x)$  in the equation:

$$0.5 = \int_0^{u(x)} \mathcal{L}(\theta; x) d\theta / \int_0^{\infty} \mathcal{L}(\theta; x) d\theta.$$

Does this procedure give a 50% confidence interval? That is, does  $\text{Prob}(u(x) > \theta) = 0.5$ ? If  $\theta = 1$ , 50% of the time we will observe  $x < 1$ , and 50%  $x > 1$ . Thus, the interval will be a 50% confidence interval if  $u(1) = 1$ .

Figure 5.17: Modified triangle distribution, for  $\theta = 1$ .

From the graph of the likelihood function for  $x = 1$ , see Fig. 5.18, we see that 50% of the area occurs at a value of  $\theta < 1$ . In fact,  $u(1) = 0.94$ . Integration of this likelihood function does not give an interval with a confidence level equal to the integrated area.

Figure 5.18: Likelihood function corresponding to sampling  $x = 1$  from the modified triangle distribution.

Integration to 50% of the area gives an interval which includes  $\theta = 1$  with 41% probability. This is still not a confidence interval, even at the 41% confidence level, because the probability is not independent of  $\theta$ . As  $\theta \rightarrow \infty$ , the probability  $\rightarrow 1/2$ , and as  $\theta \rightarrow 0$ , the probability  $\rightarrow 1$ .

The likelihood ratio method, with properly determined ratio (e.g., 0.563 for a 68% confidence level) does yield a confidence interval for  $\theta$  (with attention to signs, since  $\theta^2$  is not strictly invertible).

Let us address in general the necessary and sufficient conditions for the integrals of the likelihood function to yield confidence intervals. The theorem which we state is based on the following theorem from Lindley [3]:

**Theorem 5.1** (Lindley) *“The necessary and sufficient condition for the fiducial distribution of  $\theta$ , given  $x$ , to be a Bayes’ distribution is that there exist transformations of  $x$  to  $u$ , and of  $\theta$  to  $\tau$ , such that  $\tau$  is a location parameter for  $u$ .” [The fiducial distribution is:  $\phi_x(\theta) = -\partial_\theta F(x; \theta)$ . The restrictions on CDF  $F$  are that the derivative exist, that  $\lim_{\theta \rightarrow \infty} F(x; \theta) = 0$ , and  $\lim_{\theta \rightarrow -\infty} F(x; \theta) = 1$ .]”*

**Proof:** Proof: (following Lindley) Assuming the pdf  $f(x; \theta)$  exists, the Bayes’ distribution corresponding to prior  $p(\theta)$  is:

$$f(x; \theta)p(\theta)/\rho(x)$$

where

$$\rho(x) = \int_{-\infty}^{\infty} f(x; \theta)p(\theta) d\theta.$$

Thus, we wish to find the condition for the existence of a solution to

$$-\partial_\theta F(x; \theta) = [\partial_x F(x; \theta)] p(\theta)/\rho(x). \quad (5.35)$$

Since the  $F$ =constant solution is not permitted, a solution exists only if  $F$  is of the form:

$$F(x; \theta) = G(R(x) - P(\theta)),$$

where  $G$  is an arbitrary function, and  $R(x)$  and  $P(\theta)$  are the integrals of  $\rho(x)$  and  $p(\theta)$ .

If  $F$  is of the above form, the existence of a solution to Eqn. 5.35 may be demonstrated by considering the random variable  $u = R(x)$  and parameter  $\tau = P(\theta)$ . Then  $F = G(u - \tau)$ , and it can be verified by substitution that a solution to Eqn. 5.35 exists with a uniform prior in the parameter  $\tau$ :  $p(\tau)$ =constant. Since  $\tau$  is a location parameter for  $u$ , this completes the proof.

We are ready to answer the question: When does likelihood integral method give a confidence interval?

**Theorem 5.2** *Let  $f(x; \theta)$  be a continuous one-dimensional probability density for random variable  $x$ , depending on population parameter  $\theta$ . Let  $I = (a(x), b(x))$  be an interval obtained by integrating the likelihood function according to:*

$$\alpha = \frac{\int_{a(x)}^{b(x)} f(x; \theta) d\theta}{\int_{-\infty}^{+\infty} f(x; \theta) d\theta},$$

where  $0 < \alpha < 1$ . The interval  $I$  is a confidence interval if and only if the probability distribution is of the form:

$$f(x; \theta) = g[v(x) - \theta] \left| \frac{dv(x)}{dx} \right|,$$

### 5.3. CONFIDENCE INTERVALS FROM INVERTING TEST ACCEPTANCE REGIONS 107

where  $g$  and  $v$  are arbitrary functions. Equivalently, a necessary and sufficient condition for  $I$  to be a confidence interval is that there exist a transformation  $x \rightarrow v$  such that  $\theta$  is a location parameter for  $v$ .

**Proof:** The proof consists mainly in showing that this is a special case of Lindley's theorem.

Consider PDF  $f(x; \theta)$ , with cdf  $F(x; \theta)$ :  $f(x; \theta) = \partial_x F(x; \theta)$ . It will be sufficient to discuss one-sided intervals, since other intervals can be expressed as combinations of these. We wish to find a confidence interval specified by random variable  $u(x)$  such that:

$$\text{Prob}(u(x) > \theta) = \alpha.$$

That is,  $u(x)$  is a random variable which is greater than  $\theta$  with probability  $\alpha$ . Assume  $u$  exists and is invertible (hence also unique). This corresponds to a value of  $x$  which is greater than (or possibly less than, a case which may be dealt with similarly)  $x_\theta = u^{-1}(\theta)$  with probability  $\alpha$ .

If  $p(u; \theta)$  is the pdf for  $u$ , we require:

$$\int_{-\infty}^{\theta} p(u; \theta) du = 1 - \alpha,$$

or, in terms of the pdf for  $x$ :

$$\int_{-\infty}^{x_\theta} f(x; \theta) dx = F(x_\theta; \theta) = 1 - \alpha.$$

Given a sample  $x$ , use this equation by setting  $x_\theta = x$ , and solving  $F(x; u(x)) = 1 - \alpha$  for  $u(x)$ . This has the required property, since if  $x < x_\theta = u^{-1}(\theta)$ , then  $u(x) < \theta$ , and if  $x > x_\theta$ , then  $u(x) > \theta$ .

Find the condition on  $f(x; \theta)$  such that this interval is the same as the interval obtained by integrating the likelihood function. That is, seek the condition such that  $u(x) = u_b(x)$ , where:

$$\frac{\int_{-\infty}^{u_b(x)} f(x; \theta) d\theta}{\int_{-\infty}^{\infty} f(x; \theta) d\theta} = \alpha.$$

The left-hand-side is the integral of a Bayes' distribution, with prior  $p(\theta) = 1$ .

The  $u(x) = u_b(x)$  requirement is thus:

$$\int_{-\infty}^x f(x'; u) dx' = 1 - \int_{-\infty}^u f(x; \theta) d\theta / \rho(x),$$

$$\rho(x) = \int_{-\infty}^{\infty} f(x; \theta) p(\theta) d\theta.$$

Differentiating with respect to  $u$  yields

$$-\partial_u F(x; u) = f(x; u) / \rho(x). \quad (5.36)$$

Since this must be true for any  $\alpha$  we choose, hence for any  $u$  for a given  $x$ , this corresponds to the situation in Lindley's theorem with a uniform prior for the Bayes' distribution in  $\theta$ . Thus, this condition is satisfied if and only if  $F$  is of the form  $F(x; \theta) = G(\int \rho(x) dx - \theta)$ , or  $f(x; \theta) = G'(\int \rho(x) dx - \theta)\rho(x)$ . With  $v(x) = \int \rho(x) dx$  this is in the form as stated. If  $\theta$  is a location parameter for a function of  $x$ , we may verify that Eq. 5.36 holds by substitution. This completes the proof.

The integral method theorem may be stated intuitively as follows: If the parameter is a location parameter for a function of  $x$ , then the likelihood function is of the form:

$$\mathcal{L}(\theta; v(x)) = g[v(x) - \theta].$$

In this case, integrals over  $\theta$  correspond to regions of probability  $\alpha$  in  $v(x)$ , and hence in  $x$  if  $v$  is invertible.

The likelihood ratio and likelihood integral methods are distinct approaches, yielding different intervals. In the domain where the parameter is a location parameter, *i.e.*, in the domain where the integral method yields confidence intervals, the two methods are equivalent: They yield identical intervals, assuming that intervals with similar properties (e.g., upper or lower limit, or interval with smallest extent) are being sought. The ratio method continues to yield confidence intervals in some situations outside of this domain (in particular, the parameter need only be a function of a location parameter), and hence is the more general method for obtaining confidence intervals, although the determination of the appropriate ratios may not be easy.

## 5.4 Is the Likelihood Function Enough?

In Bayesian interval estimation, it is the likelihood function that appears for a given sampling result. Together with the prior distribution, this is all that is needed to obtain the posterior distribution and thence Bayesian intervals. The situation is different for frequentist intervals, as we shall demonstrate by example.

Since the likelihood ratio and likelihood integral methods both give confidence intervals for the case where  $\theta$  is the mean of a normal distribution, it is interesting to ask the following question:

- Suppose the likelihood function, as a function of  $\theta$ , is a normal function.
- Does this imply that either or both of the methods we have discussed will necessarily give confidence intervals?
- If a normal likelihood function implies that the data was sampled from a normal distribution, then this will be the case.
- However, there is no such implication, as we will demonstrate by an example.

We motivate our example: It is often suspected (though extremely difficult to prove a posteriori) that an experimental measurement is biased by some preconception of what the answer “should be”. For example, a preconception could be based on the result of another experiment, or on some theoretical prejudice. A model for such a biased experiment is that the experimenter works “hard” until he gets the expected result, and then quits. Consider a simple example of a distribution which could result from such a scenario.

Consider an experiment in which a measurement of a parameter  $\theta$  corresponds to sampling from a Gaussian distribution of standard deviation one:

$$n(x; \theta, 1)dx = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} dx.$$

Suppose the experimenter has a prejudice that  $\theta$  is greater than one. Subconsciously, he makes measurements until the sample mean,  $m = \frac{1}{n} \sum_{i=1}^n x_i$ , is greater than one, or until he becomes convinced (or tired) after a maximum of  $N$  measurements. The experimenter then uses the sample mean to estimate  $\theta$ .

For illustration, assume that  $N = 2$ . In terms of the random variables  $m$  and  $n$ , the PDF is:

$$f(m, n; \theta) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(m-\theta)^2}, & n = 1, m > 1 \\ 0, & n = 1, m < 1 \\ \frac{1}{\pi} e^{-(m-\theta)^2} \int_{-\infty}^1 e^{-(x-m)^2} dx & n = 2 \end{cases} \quad (5.37)$$

This distribution is illustrated in Fig. 5.19.

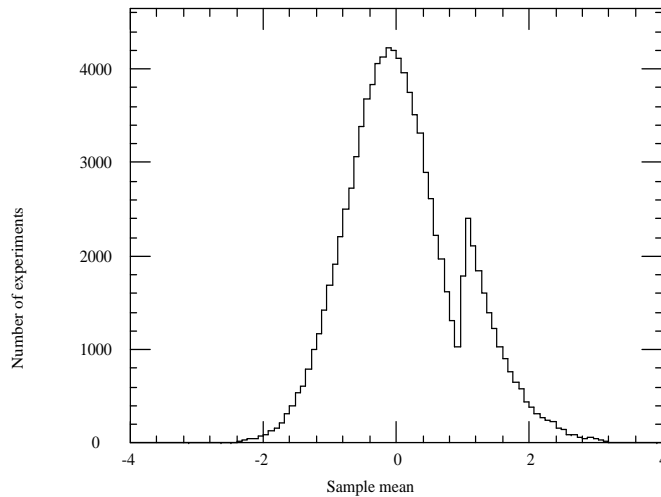


Figure 5.19: The simulated distribution for the sample mean from sampling according to Eq: 5.37, for  $\theta = 0$ .

The likelihood function, as a function of  $\theta$ , has the shape of a normal distribution, given any experimental result. The peak is at  $\theta = m$ , so  $m$  is the

maximum likelihood estimator for  $\theta$ . In spite of the normal form of the likelihood function, the sample mean is not sampled from a normal distribution. The interval defined by where the likelihood function falls by  $e^{-1/2}$  does not correspond to a 68% CI:

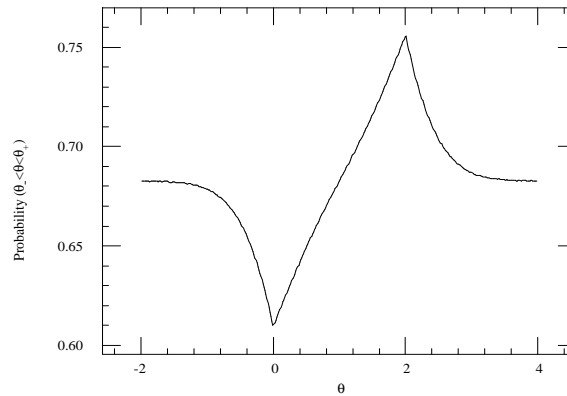


Figure 5.20: Coverage for the interval defined by  $\Delta \ln L = 1/2$  for the distribution in Eq. 5.37.

Integrals of the likelihood function correspond to particular likelihood ratios for this distribution, and hence also do not give confidence intervals. For example,  $m$  will be greater than  $\theta$  with a probability larger than 0.5. However, 50% of the area under the likelihood function always occurs at  $\theta = m$ . The interval  $(-\infty, m)$  thus obtained is not a 50% CI:

The experimenter in this scenario thinks he is taking  $n$  samples from a normal distribution, and uses one of these methods, in the knowledge that it works for a normal distribution. He gets an erroneous result because of the mistake in the distribution. If the experimenter realizes that sampling was actually from a non-normal distribution, he can do a more careful analysis by other methods to obtain more valid results. It is incorrect to argue that since each sampling is from a normal distribution, it does not matter how the number of samplings was chosen.

In contrast, the Bayesian analysis uses the likelihood function. However, The Bayesian interprets the result as a degree of confidence in true answer, hence doesn't care about coverage.

### 5.4.1 Hitting a Math Boundary

A technical issue, sometimes encountered at low statistics arises as follows: Consider a maximum likelihood fit for a set of events to some distribution, depending on parameters of interest. For example, suppose the sampling distribution is (see Fig. 5.22):

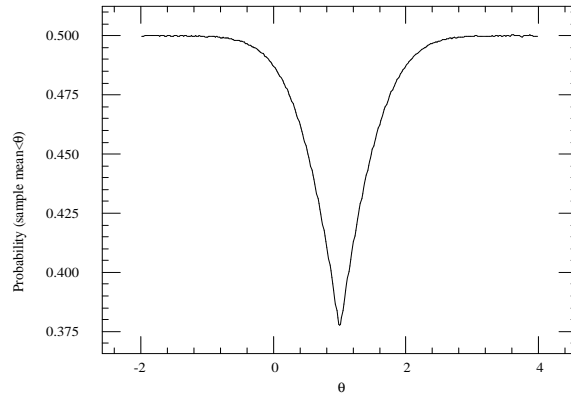


Figure 5.21: Coverage for the interval defined by the integral of the likelihood function up to 50% for the distribution in Eq. 5.37.

$$f(x; \theta) = \frac{\theta}{2} + \frac{1 - \theta}{A\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \quad x \in (-1, 1); \tag{5.38}$$

resulting in the likelihood function for a sample of size  $n$ :

$$L(\theta; \{x_i, i = 1 \dots, n\}) = \prod_{i=1}^n p(x_i; \theta). \tag{5.39}$$

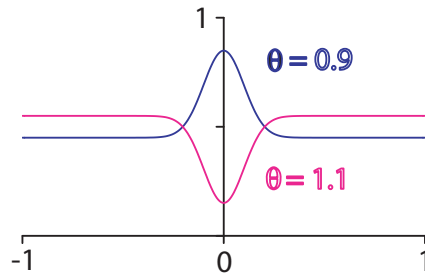


Figure 5.22: Possible PDFs given by Eq. 5.38.

The maximum with respect to  $\theta$  may be outside of region where the PDF is defined. The function  $f(x; \theta)$  may become negative in some regions of  $x$ . If there are no events in these regions, the likelihood is still “well-behaved”. However, the resulting fit, as a description of the data, will typically look poor even in the region of positive PDF, see Fig. 5.23. This is considered unacceptable.

The practical resolution to this problem is to constrain the fit to remain within bounds such that PDF is everywhere legitimate, see Fig. ???. Note that the parameters may still be “unphysical”. We’ve expressed this example as

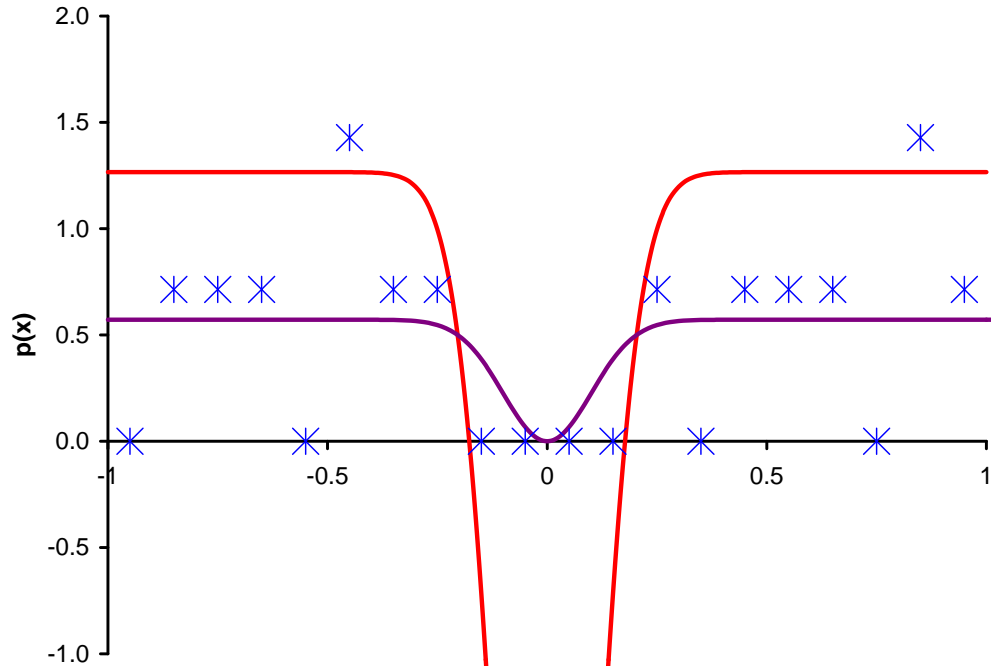


Figure 5.23: An example of an analysis based on Eq. 5.38. The points represent histogrammed data. The red curve occurs when the PDF is allowed to become negative in a maximum likelihood fit. The purple curve is the result when the fit is constrained so that the PDF is everywhere positive.

a problem in point estimation, but the same methodology applies in interval evaluation as well.

## 5.5 Displaying Errors on Poisson Statistics

A topic that sometimes arises when statistics are small is “How should I display Poisson errors in a histogram?” Exact coverage is complicated by the discreteness of the distribution. Several methods have been used. We’ll look at four approaches here, and compare their visual impressions.

Noting that  $\theta$  is the variance of a Poisson distribution of mean  $\theta$ , the use of  $\sqrt{n}$  to represent Poisson errors in a histogram is a popular method. However, the interval  $\theta_{u,\ell} = n \pm \sqrt{n}$  greatly undercovers for low mean counts. The appearance of zero error bars for  $n = 0$  also gives a visually awkward impression.

A Bayesian approach with uniform prior may be used:

$$\int_n^{\theta_u} \frac{\theta^n e^{-\theta}}{n!} d\theta = \int_{\theta_\ell}^n \frac{\theta^n e^{-\theta}}{n!} d\theta = \alpha/2,$$

unless  $\int_0^n \frac{\theta^n e^{-\theta}}{n!} d\theta < \alpha/2$ , in which case, set  $\theta_\ell = 0$ ,  $\int_n^{\theta_u} \frac{\theta^n e^{-\theta}}{n!} d\theta = \alpha$ .

An alternative Bayesian interval is obtained with:

$$\int_{\theta_u}^{\infty} \frac{\theta^n e^{-\theta}}{n!} d\theta = \int_0^{\theta_\ell} \frac{\theta^n e^{-\theta}}{n!} d\theta = (1 - \alpha)/2,$$

unless  $\theta_\ell > n$ , in which case, set  $\theta_\ell = 0$  and  $\int_n^{\theta_u} \frac{\theta^n e^{-\theta}}{n!} d\theta = \alpha$ . The  $\sqrt{n}$  and Bayesian methods are illustrated in Fig. 5.24.

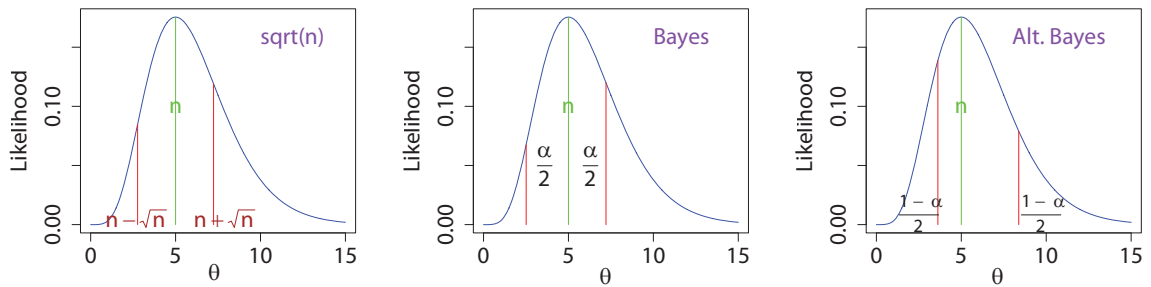


Figure 5.24: Ideas for three methods for plotting error bars on Poisson-distributed quantities. Left:  $\sqrt{n}$  error bars; Middle: A Bayesian method, integrating fractions  $\alpha/2$  on each side of the maximum likelihood; Middle: An alternative Bayesian method, integrating fractions  $(1 - \alpha)/2$  in the low and high tails of the likelihood.

A strictly frequentist approach, which may overcover but never undercovers, may also be used:

$$\sum_{k=0}^n \frac{\theta_u^k e^{-\theta_u}}{k!} = (1 - \alpha)/2, \quad \sum_{k=n}^{\infty} \frac{\theta_\ell^k e^{-\theta_\ell}}{k!} = (1 - \alpha)/2,$$

with  $\theta_\ell(0) \equiv 0$ . This is equivalent to the integrals:

$$\int_{\theta_u}^{\infty} \frac{\theta^n e^{-\theta}}{n!} d\theta = \int_0^{\theta_\ell} \frac{\theta^{n-1} e^{-\theta}}{(n-1)!} d\theta = (1 - \alpha)/2,$$

unless  $n = 0$ , in which set  $\theta_\ell = 0$  and  $\int_n^{\theta_u} \frac{\theta^n e^{-\theta}}{n!} d\theta = \alpha$ .

The intervals for given observed counts  $n$  are compared for all four approaches in the left side of Fig. 5.25. The differences are fairly subtle except at only a few counts. The right side of the figure shows the coverage frequency for the methods. As remarked above, the  $\sqrt{n}$  intervals have very bad undercoverage at low mean values. The frequentist systematically overcovers, while the two Bayesian methods fluctuate in coverage about the target value of 68%.

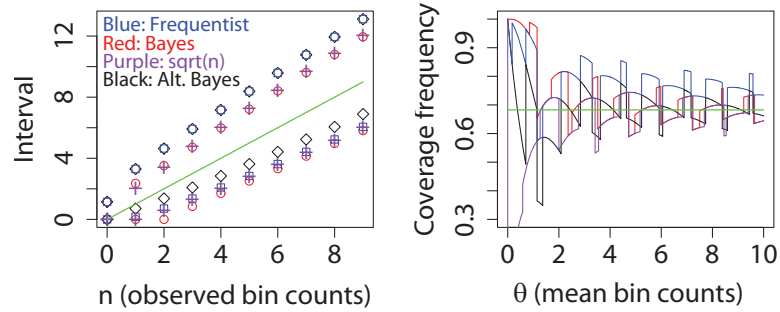


Figure 5.25: Left: The intervals for the four methods of displaying Poisson intervals, as a function of the observed number,  $n$ . Right: The coverage of the four methods of displaying Poisson intervals, as a function of parameter  $\theta$ .

Figure 5.26 shows the intervals with the likelihood function for  $n = 1$  (left) and a sample histogram using each of the methods (right). A subjective evaluation of these methods favors either the Bayesian or frequentist approaches, with little strong preference between them. My slight preference is for the Bayesian algorithm, as it doesn't have trend of inflating the errors. But you might prefer the sign of the asymmetry in the RooFit method. The  $\sqrt{n}$  error bars are disconcerting for  $n = 0$ , and the alternative Bayesian seems too lopsided.

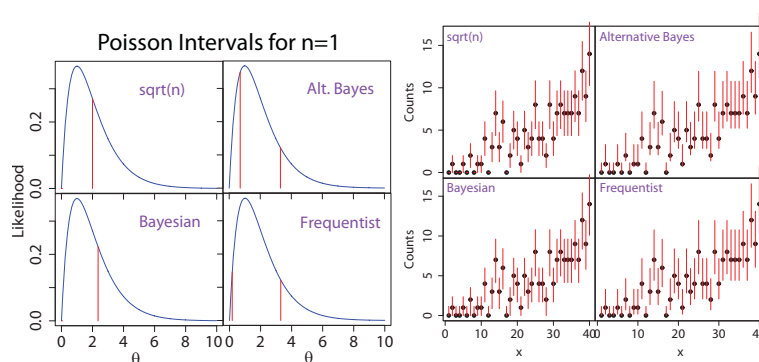


Figure 5.26: Left: The intervals obtained when  $n = 1$ , for all four methods; Right: A sample histogram showing the appearance using each method.

# Bibliography

- [1] J. Neyman, *Phil. Trans. A* **236** (1937) 333.
- [2] U. Egede, “Mixing in the  $D^0 - \bar{D}^0$  system at BaBar”, International Workshop on Frontier Science, Frascati, October 6-11, 2002, SLAC-PUB-9552, arXiv:hep-ex/0210060v1.
- [3] D. V. Lindley, “Fiducial Distributions and Bayes’ Theorem”, *Journal of the Royal Statistical Society, Series B (Methodological)*, **20** (1958) 102.