9

Statistical errors, confidence intervals and limits

In Chapters 5–8, several methods for estimating properties of p.d.f.s (moments and other parameters) have been discussed along with techniques for obtaining the variance of the estimators. Up to now the topic of 'error analysis' has been limited to reporting the variances (and covariances) of estimators, or equivalently the standard deviations and correlation coefficients. This turns out to be inadequate in certain cases, and other ways of communicating the statistical uncertainty of a measurement must be found.

After reviewing in Section 9.1 what is meant by reporting the standard deviation as an estimate of statistical uncertainty, the **confidence interval** is introduced in Section 9.2. This allows for a quantitative statement about the fraction of times that such an interval would contain the true value of the parameter in a large number of repeated experiments. Confidence intervals are treated for a number of important cases in Sections 9.3 through 9.6, and are extended to the multidimensional case in Section 9.7. In Sections 9.8 and 9.9, both Bayesian and classical confidence intervals are used to estimate limits on parameters near a physically excluded region.

9.1 The standard deviation as statistical error

Suppose the result of an experiment is an estimate of a certain parameter. The variance (or equivalently its square root, the standard deviation) of the estimator is a measure of how widely the estimates would be distributed if the experiment were to be repeated many times with the same number of observations per experiment. As such, the standard deviation σ is often reported as the statistical uncertainty of a measurement, and is referred to as the **standard error**.

For example, suppose one has n observations of a random variable x and a hypothesis for the p.d.f. $f(x;\theta)$ which contains an unknown parameter θ. From the sample $x_1, \ldots, x_n$ a function $\hat{\theta}(x_1, \ldots, x_n)$ is constructed (e.g. using maximum likelihood) as an estimator for θ. Using one of the techniques discussed in Chapters 5–8 (e.g. analytic method, RCF bound, Monte Carlo, graphical) the standard deviation of $\hat{\theta}$ can be estimated. Let $\hat{\theta}_{\mathrm{obs}}$ be the value of the estimator actually observed, and $\hat{\sigma}_{\hat{\theta}}$ the estimate of its standard deviation. In reporting the measurement of θ as $\hat{\theta}_{\mathrm{obs}} \pm \hat{\sigma}_{\hat{\theta}}$ one means that repeated estimates all based on n observations of x would be distributed according to a p.d.f. $g(\hat{\theta})$ centered around some true value θ and true standard deviation $\sigma_{\hat{\theta}}$, which are estimated to be $\hat{\theta}_{\mathrm{obs}}$ and $\hat{\sigma}_{\hat{\theta}}$.

For most practical estimators, the sampling p.d.f. $g(\hat{\theta})$ becomes approximately Gaussian in the large sample limit. If more than one parameter is estimated, then the p.d.f. will become a multidimensional Gaussian characterized by a covariance matrix V. Thus by estimating the standard deviation, or for more than one parameter the covariance matrix, one effectively summarizes all of the information available about how repeated estimates would be distributed. By using the error propagation techniques of Section 1.6, the covariance matrix also gives the equivalent information, at least approximately, for functions of the estimators.

Although the 'standard deviation' definition of statistical error bars could in principle be used regardless of the form of the estimator's p.d.f. $g(\hat{\theta})$, it is not, in fact, the conventional definition if $g(\hat{\theta})$ is not Gaussian. In such cases, one usually reports confidence intervals as described in the next section; this can in general lead to asymmetric error bars. In Section 9.3 it is shown that if $g(\hat{\theta})$ is Gaussian, then the so-called 68.3% confidence interval is the same as the interval covered by $\hat{\theta}_{\mathrm{obs}} \pm \hat{\sigma}_{\hat{\theta}}$.

9.2 Classical confidence intervals (exact method)

An alternative (and often equivalent) method of reporting the statistical error of a measurement is with a confidence interval, which was first developed by Neyman [Ney37]. Suppose as above that one has n observations of a random variable x which can be used to evaluate an estimator $\hat{\theta}(x_1, \ldots, x_n)$ for a parameter θ, and that the value obtained is $\hat{\theta}_{\mathrm{obs}}$. Furthermore, suppose that by means of, say, an analytical calculation or a Monte Carlo study, one knows the p.d.f. of $\hat{\theta}$, $g(\hat{\theta};\theta)$, which contains the true value θ as a parameter. That is, the real value of θ is not known, but for a given θ, one knows what the p.d.f. of $\hat{\theta}$ would be.

Figure 9.1 shows a probability density for an estimator $\hat{\theta}$ for a particular value of the true parameter θ. From $g(\hat{\theta};\theta)$ one can determine the value u_α such that there is a fixed probability α to observe $\hat{\theta} \geq u_\alpha$, and similarly the value v_β such that there is a probability β to observe $\hat{\theta} \leq v_\beta$. The values u_α and v_β depend on the true value of θ, and are thus determined by

$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta), \tag{9.1}$$

and

$$\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)d\hat{\theta} = G(v_\beta(\theta);\theta), \tag{9.2}$$

where G is the cumulative distribution corresponding to the p.d.f. $g(\hat{\theta};\theta)$.

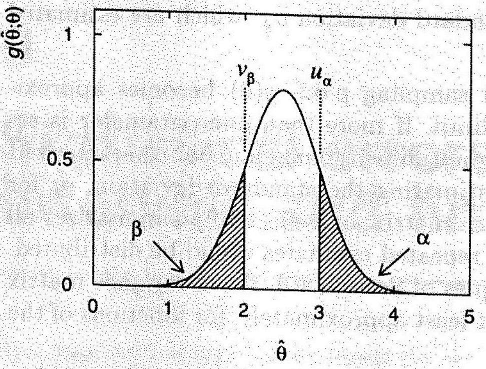

Fig. 9.1 A p.d.f. $g(\hat\theta;\theta)$ for an estimator $\hat\theta$ for a given value of the true parameter θ. The two shaded regions indicate the values of $\hat\theta \le v_\beta$, which has a probability β, and $\hat\theta \ge u_\alpha$, which has a probability α.

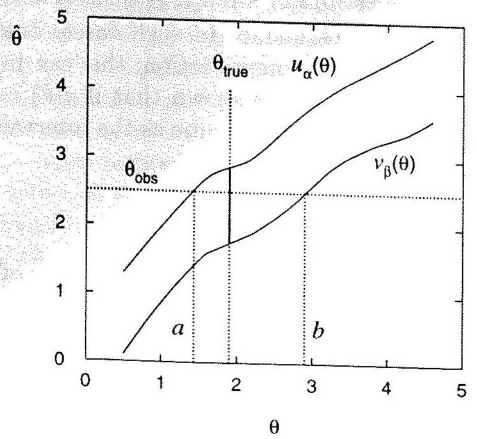

Fig. 9.2 Construction of the confidence interval $[a,b]$ given an observed value $\hat\theta_{\text{obs}}$ of the estimator $\hat\theta$ for the parameter θ (see text).

Figure 9.2 shows an example of how the functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ might appear as a function of the true value of θ. The region between the two curves is called the **confidence belt**. The probability for the estimator to be inside the belt, regardless of the value of θ, is given by

$$P(v_\beta(\theta) \le \hat\theta \le u_\alpha(\theta)) = 1 - \alpha - \beta. \tag{9.3}$$

As long as $u_\alpha(\theta)$ and $v_\beta(\theta)$ are monotonically increasing functions of θ, which in general should be the case if $\hat\theta$ is to be a good estimator for θ, one can determine the inverse functions

$$\begin{aligned} a(\hat\theta) &\equiv u_\alpha^{-1}(\hat\theta), \\ b(\hat\theta) &\equiv v_\beta^{-1}(\hat\theta). \end{aligned} \tag{9.4}$$

The inequalities

$$\begin{aligned} \hat\theta &\ge u_\alpha(\theta), \\ \hat\theta &\le v_\beta(\theta), \end{aligned} \tag{9.5}$$

then imply respectively

$$\begin{aligned} a(\hat\theta) &\ge \theta, \\ b(\hat\theta) &\le \theta. \end{aligned} \tag{9.6}$$

Equations (9.1) and (9.2) thus become

$$\begin{aligned} P(a(\hat\theta) \ge \theta) &= \alpha, \\ P(b(\hat\theta) \le \theta) &= \beta, \end{aligned} \tag{9.7}$$

or taken together,

$$P(a(\hat\theta) \le \theta \le b(\hat\theta)) = 1 - \alpha - \beta. \tag{9.8}$$

If the functions $a(\hat\theta)$ and $b(\hat\theta)$ are evaluated with the value of the estimator actually obtained in the experiment, $\hat\theta_{\text{obs}}$, then this determines two values, a and b, as illustrated in Fig. 9.2. The interval $[a,b]$ is called a **confidence interval** at a **confidence level** or **coverage probability** of $1-\alpha-\beta$. The idea behind its construction is that the coverage probability expressed by equations (9.7), and hence also (9.8), holds regardless of the true value of θ, which of course is unknown. It should be emphasized that a and b are random values, since they depend on the estimator $\hat\theta$, which is itself a function of the data. If the experiment were repeated many times, the interval $[a,b]$ would include the true value of the parameter θ in a fraction $1-\alpha-\beta$ of the experiments.

The relationship between the interval $[a,b]$ and its coverage probability $1-\alpha-\beta$ can be understood from Fig. 9.2 by considering the hypothetical true value indicated as θ_{true}. If this is the true value of θ, then $\hat\theta_{\text{obs}}$ will intersect the solid segment of the vertical line between $u_\alpha(\theta_{\text{true}})$ and $v_\beta(\theta_{\text{true}})$ with a probability of $1-\alpha-\beta$. From the figure one can see that the interval $[a,b]$ will cover θ_{true} if $\hat\theta_{\text{obs}}$ intersects this segment, and will not otherwise.

In some situations one may only be interested in a **one-sided confidence interval** or **limit**. That is, the value a represents a lower limit on the parameter θ such that $a \le \theta$ with the probability $1-\alpha$. Similarly, b represents an upper limit on θ such that $P(\theta \le b) = 1-\beta$.

Two-sided intervals (i.e. both a and b specified) are not uniquely determined by the confidence level $1-\alpha-\beta$. One often chooses, for example, $\alpha = \beta = \gamma/2$ giving a so-called **central confidence interval** with probability $1-\gamma$. Note that a central confidence interval does not necessarily mean that a and b are equidistant from the estimated value $\hat\theta$, but only that the probabilities α and β are equal.

By construction, the value a gives the hypothetical value of the true parameter θ for which a fraction α of repeated estimates $\hat\theta$ would be higher than the one actually obtained, $\hat{\theta}_{\text{obs}}$, as is illustrated in Fig. 9.3. Similarly, b is the value of θ for which a fraction β of the estimates would be lower than $\hat{\theta}_{\text{obs}}$. That is, taking $\hat{\theta}_{\text{obs}} = u_\alpha(a) = v_\beta(b)$, equations (9.1) and (9.2) become

$$\begin{aligned} \alpha &= \int_{\hat{\theta}_{\text{obs}}}^{\infty} g(\hat{\theta}; a)\, d\hat{\theta} = 1 - G(\hat{\theta}_{\text{obs}}; a), \\ \beta &= \int_{-\infty}^{\hat{\theta}_{\text{obs}}} g(\hat{\theta}; b)\, d\hat{\theta} = G(\hat{\theta}_{\text{obs}}; b). \end{aligned} \tag{9.9}$$

The previously described procedure to determine the confidence interval is thus equivalent to solving (9.9) for a and b, e.g. numerically.

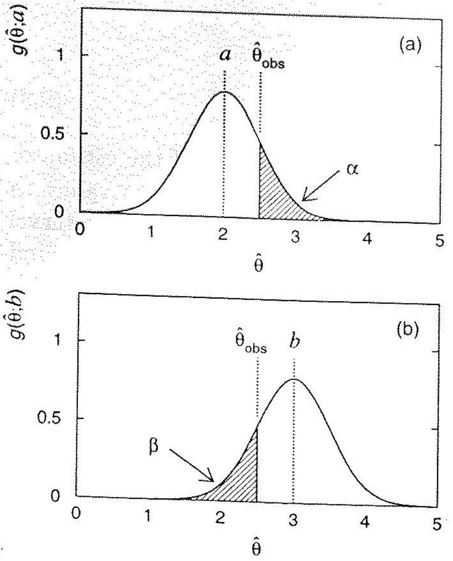

Fig. 9.3 (a) The p.d.f. $g(\hat{\theta}; a)$, where a is the lower limit of the confidence interval. If the true parameter θ were equal to a, the estimates $\hat{\theta}$ would be greater than the one actually observed $\hat{\theta}_{\text{obs}}$ with a probability α. (b) The p.d.f. $g(\hat{\theta}; b)$, where b is the upper limit of the confidence interval. If θ were equal to b, $\hat{\theta}$ would be observed less than $\hat{\theta}_{\text{obs}}$ with probability β.

Figure 9.3 also illustrates the relationship between a confidence interval and a test of goodness-of-fit, cf. Section 4.5. For example, we could test the hypothesis $\theta = a$ using $\hat{\theta}$ as a test statistic. If we define the region $\hat{\theta} \geq \hat{\theta}_{\text{obs}}$ as having equal or less agreement with the hypothesis than the result obtained (a one-sided test), then the resulting P-value of the test is α. For the confidence interval, however, the probability α is specified first, and the value a is a random quantity depending on the data. For a goodness-of-fit test, the hypothesis, here $\theta = a$, is specified and the P-value is treated as a random variable.

Note that one sometimes calls the P-value, here equal to α, the 'confidence level' of the test, whereas the one-sided confidence interval $\theta \geq a$ has a confidence level of $1 - \alpha$. That is, for a test, small α indicates a low level of confidence in the hypothesis $\theta = a$. For a confidence interval, small α indicates a *high* level of confidence that the interval $\theta \geq a$ includes the true parameter. To avoid confusion we will use the term P-value or (observed) significance level for goodness-of-fit tests, and reserve the term confidence level to mean the coverage probability of a confidence interval.

The confidence interval $[a, b]$ is often expressed by reporting the result of a measurement as $\hat{\theta}^{+d}_{-c}$, where $\hat{\theta}$ is the estimated value, and $c = \hat{\theta} - a$ and $d = b - \hat{\theta}$ are usually displayed as **error bars**. In many cases the p.d.f. $g(\hat{\theta}; \theta)$ is approximately Gaussian, so that an interval of plus or minus one standard deviation around the measured value corresponds to a central confidence interval with $1 - \gamma = 0.683$ (see Section 9.3). The 68.3% central confidence interval is usually adopted as the conventional definition for error bars even when the p.d.f. of the estimator is not Gaussian.

If, for example, the result of an experiment is reported as $\hat{\theta}^{+d}_{-c} = 5.79^{+0.32}_{-0.25}$, it is meant that if one were to construct the interval $[\hat{\theta} - c, \hat{\theta} + d]$ according to the prescription described above in a large number of similar experiments with the same number of measurements per experiment, then the interval would include the true value θ in $1 - \alpha - \beta$ of the cases. It does not mean that the probability (in the sense of relative frequency) that the true value of θ is in the fixed interval $[5.54, 6.11]$ is $1 - \alpha - \beta$. In the frequency interpretation, the true parameter θ is not a random variable and is assumed to not fluctuate from experiment to experiment. In this sense the probability that θ is in $[5.54, 6.11]$ is either 0 or 1, but we do not know which. The interval itself, however, is subject to fluctuations since it is constructed from the data.

A difficulty in constructing confidence intervals is that the p.d.f. of the estimator $g(\hat{\theta}; \theta)$, or equivalently the cumulative distribution $G(\hat{\theta}; \theta)$, must be known. An example is given in Section 10.4, where the p.d.f. for the estimator of the mean ξ of an exponential distribution is derived, and from this a confidence interval for ξ is determined. In many practical applications, estimators are Gaussian distributed (at least approximately). In this case the confidence interval can be determined easily; this is treated in detail in the next section. Even in the case of a non-Gaussian estimator, however, a simple approximate technique can be applied using the likelihood function; this is described in Section 9.6.

9.3 Confidence interval for a Gaussian distributed estimator

A simple and very important application of a confidence interval is when the distribution of $\hat{\theta}$ is Gaussian with mean θ and standard deviation $\sigma_{\hat{\theta}}$. That is, the cumulative distribution of $\hat{\theta}$ is

$$G(\hat{\theta}; \theta, \sigma_{\hat{\theta}}) = \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta}' - \theta)^2}{2\sigma_{\hat{\theta}}^2}\right) d\hat{\theta}'. \tag{9.10}$$

This is a commonly occurring situation since, according to the central limit theorem, any estimator that is a linear function of a sum of random variables becomes Gaussian in the large sample limit. We will see that for this case, the

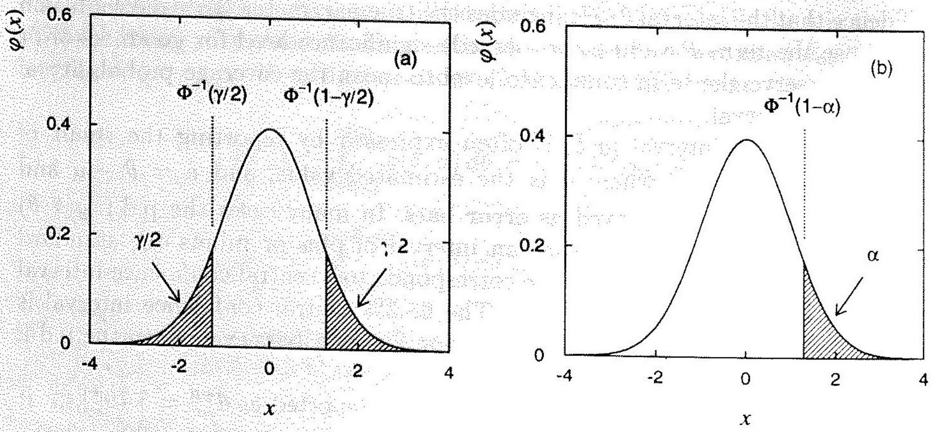

Fig. 9.4 The standard Gaussian p.d.f. $\varphi(x)$ showing the relationship between the quantiles Φ^{-1} and the confidence level for (a) a central confidence interval and (b) a one-sided confidence interval.

somewhat complicated procedure explained in the previous section results in a simple prescription for determining the confidence interval.

Suppose that the standard deviation $\sigma_{\hat{\theta}}$ is known, and that the experiment has resulted in an estimate $\hat{\theta}_{\rm obs}$. According to equations (9.9), the confidence interval $[a, b]$ is determined by solving the equations

$$\begin{aligned} \alpha &= 1 - G(\hat{\theta}_{\rm obs}; a, \sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{\rm obs} - a}{\sigma_{\hat{\theta}}}\right), \\ \beta &= G(\hat{\theta}_{\rm obs}; b, \sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{\rm obs} - b}{\sigma_{\hat{\theta}}}\right), \end{aligned} \tag{9.11}$$

for a and b, where G has been expressed using the cumulative distribution of the standard Gaussian Φ (2.26) (see also (2.27)). This gives

$$\begin{aligned} a &= \hat{\theta}_{\rm obs} - \sigma_{\hat{\theta}}\, \Phi^{-1}(1-\alpha), \\ b &= \hat{\theta}_{\rm obs} + \sigma_{\hat{\theta}}\, \Phi^{-1}(1-\beta). \end{aligned} \tag{9.12}$$

Here Φ^{-1} is the inverse function of Φ, i.e. the quantile of the standard Gaussian, and in order to make the two equations symmetric we have used $\Phi^{-1}(\beta) = -\Phi^{-1}(1-\beta)$.

The quantiles $\Phi^{-1}(1-\alpha)$ and $\Phi^{-1}(1-\beta)$ represent how far away the interval limits a and b are located with respect to the estimate $\hat{\theta}_{\rm obs}$ in units of the standard deviation $\sigma_{\hat{\theta}}$. The relationship between the quantiles of the standard Gaussian distribution and the confidence level is illustrated in Fig. 9.4(a) for central and Fig. 9.4(b) for one-sided confidence intervals.

Consider a central confidence interval with $\alpha = \beta = \gamma/2$. The confidence level $1-\gamma$ is often chosen such that the quantile is a small integer, e.g. $\Phi^{-1}(1-\gamma/2) = 1, 2, 3, \ldots$. Similarly, for one-sided intervals (limits) one often chooses a small integer for $\Phi^{-1}(1-\alpha)$. Commonly used values for both central and one-sided intervals are shown in Table 9.1. Alternatively one can choose a round number for the confidence level instead of for the quantile. Commonly used values are shown in Table 9.2. Other possible values can be obtained from [Bra92, Fro79, Dud88] or from computer routines (e.g. the routine **GAUSIN** in [CER97]).

Table 9.1 The values of the confidence level for different values of the quantile of the standard Gaussian Φ^{-1}: for central intervals (left) the quantile $\Phi^{-1}(1-\gamma/2)$ and confidence level $1-\gamma$; for one-sided intervals (right) the quantile $\Phi^{-1}(1-\alpha)$ and confidence level $1-\alpha$.

$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$	$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987

Table 9.2 The values of the quantile of the standard Gaussian Φ^{-1} for different values of the confidence level: for central intervals (left) the confidence level $1-\gamma$ and the quantile $\Phi^{-1}(1-\gamma/2)$; for one-sided intervals (right) the confidence level $1-\alpha$ and the quantile $\Phi^{-1}(1-\alpha)$.

$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

For the conventional 68.3% central confidence interval one has $\alpha = \beta = \gamma/2$, with $\Phi^{-1}(1-\gamma/2) = 1$, i.e. a '1 σ error bar'. This results in the simple prescription

$$[a, b] = [\hat{\theta}_{\rm obs} - \sigma_{\hat{\theta}}, \hat{\theta}_{\rm obs} + \sigma_{\hat{\theta}}]. \tag{9.13}$$

Thus for the case of a Gaussian distributed estimator, the 68.3% central confidence interval is given by the estimated value plus or minus one standard deviation. The final result of the measurement of θ is then simply reported as $\hat{\theta}_{\rm obs} \pm \sigma_{\hat{\theta}}$.

If the standard deviation $\sigma_{\hat{\theta}}$ is not known a priori but rather is estimated from the data, then the situation is in principle somewhat more complicated. If, for example, the estimated standard deviation $\hat{\sigma}_{\hat{\theta}}$ had been used instead of $\sigma_{\hat{\theta}}$, then it would not have been so simple to relate the cumulative distribution $G(\hat{\theta}; \theta, \hat{\sigma}_{\hat{\theta}})$ to Φ, the cumulative distribution of the standard Gaussian, since $\hat{\sigma}_{\hat{\theta}}$ depends in general on $\hat{\theta}$. In practice, however, the recipe given above can still be applied using the estimate $\hat{\sigma}_{\hat{\theta}}$ instead of $\sigma_{\hat{\theta}}$, as long as $\hat{\sigma}_{\hat{\theta}}$ is a sufficiently good approximation of the true standard deviation, e.g. for a large enough data sample. For the small sample case where $\hat{\theta}$ represents the mean of n Gaussian random variables of unknown standard deviation, the confidence interval can be determined by relating the cumulative distribution $G(\hat{\theta}; \theta, \hat{\sigma}_{\hat{\theta}})$ to Student's t distribution (see e.g. [Fro79], [Dud88] Section 10.2).

Exact determination of confidence intervals becomes more difficult if the p.d.f. of the estimator $g(\hat{\theta}; \theta)$ is not Gaussian, or worse, if it is not known analytically. For a non-Gaussian p.d.f. it is sometimes possible to transform the parameter $\theta \to \eta(\theta)$ such that the p.d.f. for the estimator $\hat{\eta}$ is approximately Gaussian. The confidence interval for the transformed parameter η can then be converted back into an interval for θ. An example of this technique is given in Section 9.5.

9.4 Confidence interval for the mean of the Poisson distribution

Along with the Gaussian distributed estimator, another commonly occurring case is where the outcome of a measurement is a Poisson variable n $(n = 0, 1, 2, \ldots)$. Recall from (2.9) that the probability to observe n is

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}, \tag{9.14}$$

and that the parameter ν is equal to the expectation value $E[n]$. The maximum likelihood estimator for ν can easily be found to be $\hat{\nu} = n$. Suppose that a single measurement has resulted in the value $\hat{\nu}_{\rm obs} = n_{\rm obs}$, and that from this we would like to construct a confidence interval for the mean ν.

For the case of a discrete variable, the procedure for determining the confidence interval described in Section 9.2 cannot be directly applied. This is because the functions $u_\alpha(\theta)$ and $v_\beta(\theta)$, which determine the confidence belt, do not exist for all values of the parameter θ. For the Poisson case, for example, we would need to find $u_\alpha(\nu)$ and $v_\beta(\nu)$ such that $P(\hat{\nu} \geq u_\alpha(\nu)) = \alpha$ and $P(\hat{\nu} \leq v_\beta(\nu)) = \beta$ for all values of the parameter ν. But if α and β are fixed, then because $\hat{\nu}$ only takes on discrete values, these equations hold in general only for particular values of ν.

A confidence interval $[a, b]$ can still be determined, however, by using equations (9.9). For the case of a discrete random variable and a parameter ν these become

$$\begin{aligned} \alpha &= P(\hat{\nu} \geq \hat{\nu}_{\rm obs}; a), \\ \beta &= P(\hat{\nu} \leq \hat{\nu}_{\rm obs}; b), \end{aligned} \tag{9.15}$$

and in particular for a Poisson variable one has

$$\begin{aligned} \alpha &= \sum_{n=n_{\rm obs}}^{\infty} f(n; a) = 1 - \sum_{n=0}^{n_{\rm obs}-1} f(n; a) = 1 - \sum_{n=0}^{n_{\rm obs}-1} \frac{a^n}{n!} e^{-a}, \\ \beta &= \sum_{n=0}^{n_{\rm obs}} f(n; b) = \sum_{n=0}^{n_{\rm obs}} \frac{b^n}{n!} e^{-b}. \end{aligned} \tag{9.16}$$

For an estimate $\hat{\nu} = n_{\rm obs}$ and given probabilities α and β, these equations can be solved numerically for a and b. Here one can use the following relation between the Poisson and χ^2 distributions,

$$\begin{aligned} \sum_{n=0}^{n_{\rm obs}} \frac{\nu^n}{n!} e^{-\nu} &= \int_{2\nu}^{\infty} f_{\chi^2}(z; n_{\rm d} = 2(n_{\rm obs} + 1))\, dz \\ &= 1 - F_{\chi^2}(2\nu; n_{\rm d} = 2(n_{\rm obs} + 1)), \end{aligned} \tag{9.17}$$

where f_{χ^2} is the χ^2 p.d.f. for $n_{\rm d}$ degrees of freedom and F_{χ^2} is the corresponding cumulative distribution. One then has

$$\begin{aligned} a &= \tfrac{1}{2} F_{\chi^2}^{-1}(\alpha; n_{\rm d} = 2n_{\rm obs}), \\ b &= \tfrac{1}{2} F_{\chi^2}^{-1}(1 - \beta; n_{\rm d} = 2(n_{\rm obs} + 1)). \end{aligned} \tag{9.18}$$

Quantiles $F_{\chi^2}^{-1}$ of the χ^2 distribution can be obtained from standard tables (e.g. in [Bra92]) or from computer routines such as `CHISIN` in [CER97]. Some values for $n_{\rm obs} = 0, \ldots, 10$ are shown in Table 9.3.

Note that the lower limit a cannot be determined if $n_{\rm obs} = 0$. Equations (9.15) say that if $\nu = a$ $(\nu = b)$, then the probability is α (β) to observe a value greater (less) than or equal to the one actually observed. Because the case of equality, $\hat{\nu} = \hat{\nu}_{\rm obs}$, is included in the inequalities (9.15), one obtains a conservatively large confidence interval, i.e.

$$\begin{aligned} P(\nu \geq a) &\geq 1 - \alpha, \\ P(\nu \leq b) &\geq 1 - \beta, \\ P(a \leq \nu \leq b) &\geq 1 - \alpha - \beta. \end{aligned} \tag{9.19}$$

An important special case is when the observed number $n_{\rm obs}$ is zero, and one is interested in establishing an upper limit b. Equation (9.15) becomes

$$\beta = \sum_{n=0}^{0} \frac{b^n e^{-b}}{n!} = e^{-b}, \tag{9.20}$$

for ρ simply by using the inverse of the transformation (9.22), i.e. $A = \tanh a$ and $B = \tanh b$.

Consider for example a sample of size $n = 20$ for which one has obtained the estimate $r = 0.5$. From equation (5.17) the standard deviation of r can be estimated as $\hat{\sigma}_r = (1 - r^2)/\sqrt{n} = 0.168$. If one were to make the incorrect approximation that r is Gaussian distributed for such a small sample, this would lead to a 68.3% central confidence interval for ρ of $[0.332, 0.668]$, or $[0.067, 0.933]$ at a confidence level of 99%. Thus since the sample correlation coefficient r is almost three times the standard error $\hat{\sigma}_r$, one might be led to the incorrect conclusion that there is significant evidence for a non-zero value of ρ, i.e. a '3 σ effect'. By using the z-transformation, however, one obtains $z = 0.549$ and $\hat{\sigma}_z = 0.243$. This corresponds to a 99% central confidence interval of $[-0.075, 1.174]$ for ζ, and $[-0.075, 0.826]$ for ρ. Thus the 99% central confidence interval includes zero.

Recall that the lower limit of the confidence interval is equal to the hypothetical value of the true parameter such that r would be observed higher than the one actually observed with the probability α. One can ask, for example, what the confidence level would be for a lower limit of zero. If we had assumed that $g(r; \rho, n)$ was Gaussian, the corresponding probability would be 0.14%. By using the z-transformation, however, the confidence level for a limit of zero is 2.3%, i.e. if ρ were zero one would obtain r greater than or equal to the one observed, $r = 0.5$, with a probability of 2.3%. The actual evidence for a non-zero correlation is therefore not nearly as strong as one would have concluded by simply using the standard error $\hat{\sigma}_r$ with the assumption that r is Gaussian.

9.6 Confidence intervals using the likelihood function or χ^2

Even in the case of a non-Gaussian estimator, the confidence interval can be determined with a simple approximate technique which makes use of the likelihood function or equivalently the χ^2 function where one has $L = \exp(-\chi^2/2)$. Consider first a maximum likelihood estimator $\hat{\theta}$ for a parameter θ in the large sample limit. In this limit it can be shown ([Stu91] Chapter 18) that the p.d.f. $g(\hat{\theta}; \theta)$ becomes Gaussian,

$$g(\hat{\theta}; \theta) = \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta} - \theta)^2}{2\sigma_{\hat{\theta}}^2} \right), \qquad (9.26)$$

centered about the true value of the parameter θ and with a standard deviation $\sigma_{\hat{\theta}}$.

One can also show that in the large sample limit the likelihood function itself becomes Gaussian in form centered about the ML estimate $\hat{\theta}$,

$$L(\theta) = L_{\max} \exp\left(\frac{-(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2} \right). \qquad (9.27)$$

From the RCF inequality (6.16), which for an ML estimator in the large sample limit becomes an equality, one obtains that $\sigma_{\hat{\theta}}$ in the likelihood function (9.27) is the same as in the p.d.f. (9.26). This has already been encountered in Section 6.7, equation (6.24), where the likelihood function was used to estimate the variance of an estimator $\hat{\theta}$. This led to a simple prescription for estimating $\sigma_{\hat{\theta}}$, since by changing the parameter θ by N standard deviations, the log-likelihood function decreases by $N^2/2$ from its maximum value,

$$\log L(\hat{\theta} \pm N\sigma_{\hat{\theta}}) = \log L_{\max} - \frac{N^2}{2}. \qquad (9.28)$$

From the results of the previous section, however, we know that for a Gaussian distributed estimator $\hat{\theta}$, the 68.3% central confidence interval can be constructed from the estimator and its estimated standard deviation $\hat{\sigma}_{\hat{\theta}}$ as $[a, b] = [\hat{\theta} - \hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \hat{\sigma}_{\hat{\theta}}]$ (or more generally according to (9.12) for a confidence level of $1 - \gamma$). The 68.3% central confidence interval is thus given by the values of θ at which the log-likelihood function decreases by 1/2 from its maximum value. (This is assuming, of course, that $\hat{\theta}$ is the ML estimator and thus corresponds to the maximum of the likelihood function.)

In fact, it can be shown that even if the likelihood function is not a Gaussian function of the parameters, the central confidence interval $[a, b] = [\hat{\theta} - c, \hat{\theta} + d]$ can still be approximated by using

$$\log L(\hat{\theta}^{+d}_{-c}) = \log L_{\max} - \frac{N^2}{2}, \qquad (9.29)$$

where $N = \Phi^{-1}(1 - \gamma/2)$ is the quantile of the standard Gaussian corresponding to the desired confidence level $1 - \gamma$. (For example, $N = 1$ for a 68.3% central confidence interval; see Table 9.1.) In the case of a least squares fit with Gaussian errors, i.e. with $\log L = -\chi^2/2$, the prescription becomes

$$\chi^2(\hat{\theta}^{+d}_{-c}) = \chi^2_{\min} + N^2. \qquad (9.30)$$

A heuristic proof that the intervals defined by equations (9.29) and (9.30) approximate the classical confidence intervals of Section 9.2 can be found in [Ead71, Fro79]. Equations (9.29) and (9.30) represent one of the most commonly used methods for estimating statistical uncertainties. One should keep in mind, however, that the correspondence with the method of Section 9.2 is only exact in the large sample limit. Several authors have recommended using the term 'likelihood interval' for an interval obtained from the likelihood function [Fro79, Hud64]. Regardless of the name, it should be kept in mind that it is interpreted here as an approximation to the classical confidence interval, i.e. a random interval constructed so as to include the true parameter value with a given probability.

As an example consider the estimator $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ for the parameter τ of an exponential distribution, as in the example of Section 6.2 (see also Section 6.7). There, the ML method was used to estimate τ given a sample of $n = 50$ measurements of an exponentially distributed random variable t. This sample was sufficiently large that the standard deviation $\sigma_{\hat{\tau}}$ could be approximated by the values of τ where the log-likelihood function decreased by 1/2 from its maximum (see Fig. 6.4). This gave $\hat{\tau} = 1.06$ and $\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$.

Figure 9.6 shows the log-likelihood function $\log L(\tau)$ as a function of τ for a sample of only $n = 5$ measurements of an exponentially distributed random variable, generated using the Monte Carlo method with the true parameter $\tau = 1$. Because of the smaller sample size the log-likelihood function is less parabolic than before.

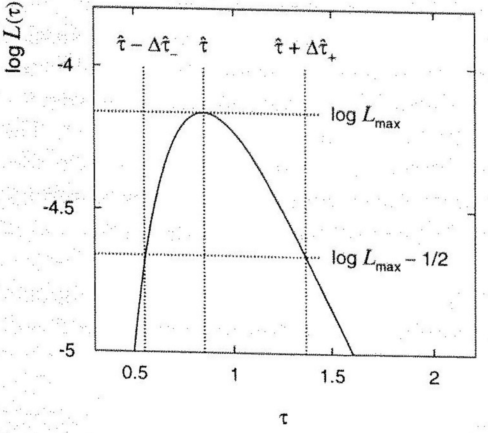

Fig. 9.6 The log-likelihood function $\log L(\tau)$ as a function of τ for a sample of $n = 5$ measurements. The interval $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ determined by $\log L(\tau) = \log L_{\max} - 1/2$ can be used to approximate the 68.3% central confidence interval.

One could still use the half-width of the interval determined by $\log L_{\max} - 1/2$ to approximate the standard deviation $\sigma_{\hat{\tau}}$, but this is not really what we want. The statistical uncertainty is better communicated by giving the confidence interval, since one then knows the probability that the interval covers the true parameter value. Furthermore, by giving a central confidence interval (and hence asymmetric errors, $\Delta\hat{\tau}_- \neq \Delta\hat{\tau}_+$), one has equal probabilities for the true parameter to be higher or lower than the interval limits. As illustrated in Fig. 9.6, the central confidence interval can be approximated by the values of τ where $\log L(\tau) = \log L_{\max} - 1/2$, which gives $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+] = [0.55, 1.37]$ or $\hat{\tau} = 0.85^{+0.52}_{-0.30}$.

In fact, the same could have been done in Section 6.7 by giving the result there as $\hat{\tau} = 1.062^{+0.165}_{-0.137}$. Whether one chooses this method or simply reports an averaged symmetric error (i.e. $\hat{\tau} = 1.06 \pm 0.15$) will depend on how accurately the statistical error needs to be given. For the case of $n = 5$ shown in Fig. 9.6, the error bars are sufficiently asymmetric that one would probably want to use the 68.3% central confidence interval and give the result as $\hat{\tau} = 0.85^{+0.52}_{-0.30}$.

9.7 Multidimensional confidence regions

In Section 9.2, a confidence interval $[a, b]$ was constructed so as to have a certain probability $1 - \gamma$ of containing a parameter θ. In order to generalize this to the case of n parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, one might attempt to find an n-dimensional confidence interval $[\mathbf{a}, \mathbf{b}]$ constructed so as to have a given probability that $a_i < \theta_i < b_i$, simultaneously for all i. This turns out to be computationally difficult, and is rarely done.

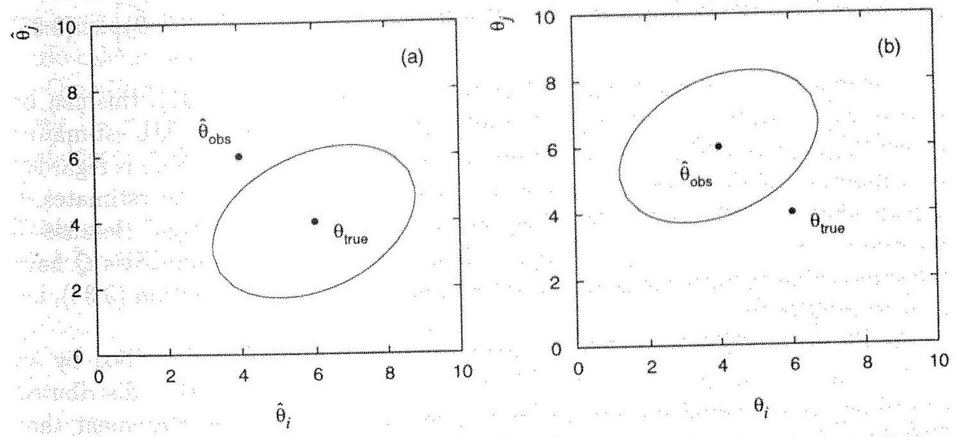

Fig. 9.7 (a) A contour of constant $g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}_{\mathrm{true}})$ (i.e. constant $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{\mathrm{true}})$) in $\hat{\boldsymbol{\theta}}$-space. (b) A contour of constant $L(\boldsymbol{\theta})$ corresponding to constant $Q(\hat{\boldsymbol{\theta}}_{\mathrm{obs}}, \boldsymbol{\theta})$ in $\boldsymbol{\theta}$-space. The values $\boldsymbol{\theta}_{\mathrm{true}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{obs}}$ represent particular constant values of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$, respectively.

It is nevertheless quite simple to construct a **confidence region** in the parameter space such that the true parameter $\boldsymbol{\theta}$ is contained within the region with a given probability (at least approximately). This region will not have the form $a_i < \theta_i < b_i$, $i = 1, \ldots, n$, but will be more complicated, approaching an n-dimensional hyperellipsoid in the large sample limit.

As in the single-parameter case, one makes use of the fact that both the joint p.d.f. for the estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ as well as the likelihood function become Gaussian in the large sample limit. That is, the joint p.d.f. of $\hat{\boldsymbol{\theta}}$ becomes

$$g(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\tfrac{1}{2} Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\right], \tag{9.31}$$

where Q is defined as

$$Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{9.32}$$

Here V^{-1} is the inverse covariance matrix and the superscript T indicates a transposed (i.e. row) vector. Contours of constant $g(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$ correspond to constant $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$. These are ellipses (or for more than two dimensions, hyperellipsoids) in $\hat{\boldsymbol{\theta}}$-space centered about the true parameters $\boldsymbol{\theta}$. Figure 9.7(a) shows a contour of constant $Q(\hat{\boldsymbol{\theta}})$, where $\boldsymbol{\theta}_{\mathrm{true}}$ represents a particular value of $\boldsymbol{\theta}$.

Also as in the one-dimensional case, one can show that the likelihood function $L(\boldsymbol{\theta})$ takes on a Gaussian form centered about the ML estimators $\hat{\boldsymbol{\theta}}$,

$$L(\boldsymbol{\theta}) = L_{\max} \exp\left[-\tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T V^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right] = L_{\max}\exp\left[-\tfrac{1}{2}\,Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})\right]. \qquad (9.33)$$

The inverse covariance matrix V^{-1} is the same here as in (9.31); this can be seen from the RCF inequality (6.19) and using the fact that the ML estimators attain the RCF bound in the large sample limit. The quantity Q here is regarded as a function of the parameters $\boldsymbol{\theta}$ which has its maximum at the estimates $\hat{\boldsymbol{\theta}}$. This is shown in Fig. 9.7(b) for $\hat{\boldsymbol{\theta}}$ equal to a particular value $\hat{\boldsymbol{\theta}}_{\text{obs}}$. Because of the symmetry between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ in the definition (9.32), the quantities Q have the same value in both the p.d.f. (9.31) and in the likelihood function (9.33), i.e. $Q(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) = Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$.

As discussed in Section 7.5, it can be shown that if $\hat{\boldsymbol{\theta}}$ is described by an n-dimensional Gaussian p.d.f. $g(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$, then the quantity $Q(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$ is distributed according to a χ^2 distribution for n degrees of freedom. The statement that $Q(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$ is less than some value Q_γ, i.e. that the estimate is within a certain distance of the true value $\boldsymbol{\theta}$, implies $Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) < Q_\gamma$, i.e. that the true value θ is within the same distance of the estimate. The two events therefore have the same probability,

$$P(Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) \leq Q_\gamma) = \int_0^{Q_\gamma} f(z;n)dz, \qquad (9.34)$$

where $f(z;n)$ is the χ^2 distribution for n degrees of freedom (equation (2.34)). The value Q_γ is chosen to correspond to a given probability content,

$$\int_0^{Q_\gamma} f(z;n)dz = 1-\gamma. \qquad (9.35)$$

That is,

$$Q_\gamma = F^{-1}(1-\gamma;n) \qquad (9.36)$$

is the quantile of order $1-\gamma$ of the χ^2 distribution. The region of $\boldsymbol{\theta}$-space defined by $Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) \leq Q_\gamma$ is called a **confidence region** with the confidence level $1-\gamma$. For a likelihood function of Gaussian form (9.33) it can be constructed by finding the values of $\boldsymbol{\theta}$ at which the log-likelihood function decreases by $Q_\gamma/2$ from its maximum value,

$$\log L(\boldsymbol{\theta}) = \log L_{\max} - \frac{Q_\gamma}{2}. \qquad (9.37)$$

As in the single-parameter case, one can still use the prescription given by (9.37) even if the likelihood function is not Gaussian, in which case the probability statement (9.34) is only approximate. For an increasing number of parameters, the approach to the Gaussian limit becomes slower as a function of the sample size, and furthermore it is difficult to quantify when a sample is large enough for (9.34) to apply. If needed, one can determine the probability that a region constructed according to (9.37) includes the true parameter by means of a Monte Carlo calculation.

Quantiles of the χ^2 distribution $Q_\gamma = F^{-1}(1-\gamma;n)$ for several confidence levels $1-\gamma$ and $n = 1,2,3,4,5$ parameters are given in Table 9.4. Values of the confidence level are shown for various values of the quantile Q_γ in Table 9.5.

Table 9.4 The values of the confidence level $1-\gamma$ for different values of Q_γ and for $n = 1,2,3,4,5$ fitted parameters.

Q_γ	$1-\gamma$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Table 9.5 The values of the quantile Q_γ for different values of the confidence level $1-\gamma$ for $n = 1,2,3,4,5$ fitted parameters.

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

For $n = 1$ the expression (9.36) for Q_γ can be shown to imply

$$\sqrt{Q_\gamma} = \Phi^{-1}(1-\gamma/2), \qquad (9.38)$$

where Φ^{-1} is the inverse function of the standard normal distribution. The procedure here thus reduces to that for a single parameter given in Section 9.6, where $N = \sqrt{Q_\gamma}$ is the half-width of the interval in standard deviations (see equations (9.28), (9.29)). The values for $n = 1$ in Tables 9.4 and 9.5 are thus related to those in Tables 9.1 and 9.2 by equation (9.38).

For increasing n, the confidence level for a given Q_γ decreases. For example, in the single-parameter case, $Q_\gamma = 1$ corresponds to $1-\gamma = 0.683$. For $n = 2$, $Q_\gamma = 1$ gives a confidence level of only 0.393, and in order to obtain $1-\gamma = 0.683$ one needs $Q_\gamma = 2.30$.

We should emphasize that, as in the single-parameter case, the confidence region $Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) \leq Q_\gamma$ is a random region in $\boldsymbol{\theta}$-space. The confidence region varies upon repetition of the experiment, since $\hat{\boldsymbol{\theta}}$ is a random variable. The true parameters, on the other hand, are unknown constants.

9.8 Limits near a physical boundary

Often the purpose of an experiment is to search for a new effect, the existence of which would imply that a certain parameter is not equal to zero. For example, one could attempt to measure the mass of the neutrino, which in the standard theory is massless. If the data yield a value of the parameter significantly different from zero, then the new effect has been discovered, and the parameter's value and a confidence interval to reflect its error are given as the result. If, on the other hand, the data result in a fitted value of the parameter that is consistent with zero, then the result of the experiment is reported by giving an upper limit on the parameter. (A similar situation occurs when absence of the new effect corresponds to a parameter being large or infinite; one then places a lower limit. For simplicity we will consider here only upper limits.)

Difficulties arise when an estimator can take on values in the excluded region. This can occur if the estimator $\hat{\theta}$ for a parameter θ is of the form $\hat{\theta} = x - y$, where both x and y are random variables, i.e. they have random measurement errors. The mass squared of a particle, for example, can be estimated by measuring independently its energy E and momentum p, and using $\widehat{m^2} = E^2 - p^2$. Although the mass squared should come out positive, measurement errors in E^2 and p^2 could result in a negative value for $\widehat{m^2}$. Then the question is how to place a limit on m^2, or more generally on a parameter θ when the estimate is in or near an excluded region.

Consider further the example of an estimator $\hat{\theta} = x - y$ where x and y are Gaussian variables with means μ_x, μ_y and variances σ_x^2, σ_y^2. One can show that the difference $\hat{\theta} = x - y$ is also a Gaussian variable with $\theta = \mu_x - \mu_y$ and $\sigma_{\hat{\theta}}^2 = \sigma_x^2 + \sigma_y^2$. (This can be shown using characteristic functions as described in Chapter 10.)

Assume that θ is known a priori to be non-negative (e.g. like the mass squared), and suppose the experiment has resulted in a value $\hat{\theta}_{\text{obs}}$ for the estimator $\hat{\theta}$. According to (9.12), the upper limit θ_{up} at a confidence level $1 - \beta$ is

$$\theta_{\text{up}} = \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}\, \Phi^{-1}(1 - \beta). \tag{9.39}$$

For the commonly used 95% confidence level one obtains from Table 9.2 the quantile $\Phi^{-1}(0.95) = 1.645$.

The interval $(-\infty, \theta_{\text{up}}]$ is constructed to include the true value θ with a probability of 95%, regardless of what θ actually is. Suppose now that the standard deviation is $\sigma_{\hat{\theta}} = 1$, and the result of the experiment is $\hat{\theta}_{\text{obs}} = -2.0$. From equation (9.39) one obtains $\theta_{\text{up}} = -0.355$ at a confidence level of 95%. Not only is $\hat{\theta}_{\text{obs}}$ in the forbidden region (as half of the estimates should be if θ is really zero) but the upper limit is below zero as well. This is not particularly unusual, and in fact is expected to happen in 5% of the experiments if the true value of θ is zero.

As far as the definition of the confidence interval is concerned, nothing fundamental has gone wrong. The interval was designed to cover the true value of θ in a certain fraction of repeated experiments, and we have obviously encountered one of those experiments where θ is not in the interval. But this is not a very satisfying result, since it was already known that θ is greater than zero (and certainly greater than $\theta_{\text{up}} = -0.355$) without having to perform the experiment.

Regardless of the upper limit, it is important to report the actual value of the estimate obtained and its standard deviation, i.e. $\hat{\theta}_{\text{obs}} \pm \sigma_{\hat{\theta}}$, even if the estimate is in the physically excluded region. In this way, the average of many experiments (e.g. as in Section 7.6) will converge to the correct value as long as the estimator is unbiased. In cases where the p.d.f. of $\hat{\theta}$ is significantly non-Gaussian, the entire likelihood function $L(\theta)$ should be given, which can be combined with that of other experiments as discussed in Section 6.12.

Nevertheless, most experimenters want to report some sort of upper limit, and in situations such as the one described above a number of techniques have been proposed (see e.g. [Hig83, Jam91]). There is unfortunately no established convention on how this should be done, and one should therefore state what procedure was used.

As a solution to the difficulties posed by an upper limit in an unphysical region, one might be tempted to simply increase the confidence level until the limit enters the allowed region. In the previous example, if we had taken a confidence level $1 - \beta = 0.99$, then from Table 9.2 one has $\Phi^{-1}(0.99) = 2.326$, giving $\theta_{\text{up}} = 0.326$. This would lead one to quote an upper limit that is smaller than the intrinsic resolution of the experiment ($\sigma_{\hat{\theta}} = 1$) at a very high confidence level of 99%, which is clearly misleading. Worse, of course, would be to adjust the confidence level to give an arbitrarily small limit, e.g. $\Phi^{-1}(0.97725) = 2.00001$, or $\theta_{\text{up}} = 10^{-5}$ at a confidence level of 97.725%!

In order to avoid this type of difficulty, a commonly used technique is to simply shift a negative estimate to zero before applying equation (9.39), i.e.

$$\theta_{\text{up}} = \max(\hat{\theta}_{\text{obs}}, 0) + \sigma_{\hat{\theta}}\, \Phi^{-1}(1 - \beta). \tag{9.40}$$

In this way the upper limit is always at least the same order of magnitude as the resolution of the experiment. If $\hat{\theta}_{\text{obs}}$ is positive, the limit coincides with that of the classical procedure. This technique has a certain intuitive appeal and is often used, but the interpretation as an interval that will cover the true parameter value with probability $1 - \beta$ no longer applies. The coverage probability is clearly greater than $1 - \beta$, since the shifted upper limit (9.40) is in all cases greater than or equal to the classical one (9.39).

Another alternative is to report an interval based on the Bayesian posterior p.d.f. $p(\theta|\mathbf{x})$. As in Section 6.13, this is obtained from Bayes' theorem,

$$p(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)\, \pi(\theta)}{\int L(\mathbf{x}|\theta')\, \pi(\theta')\, d\theta'}, \tag{9.41}$$

where $\mathbf{x}$ represents the observed data, $L(\mathbf{x}|\theta)$ is the likelihood function and $\pi(\theta)$ is the prior p.d.f. for θ. In Section 6.13, the mode of $p(\theta|\mathbf{x})$ was used as an estimator for θ, and it was shown that this coincides with the ML estimator if the prior density $\pi(\theta)$ is uniform. Here, we can use $p(\theta|\mathbf{x})$ to determine an interval $[a, b]$ such that for given probabilities α and β one has

$$\begin{aligned} \alpha &= \int_{-\infty}^{a} p(\theta|\mathbf{x})\, d\theta \\ \beta &= \int_{b}^{\infty} p(\theta|\mathbf{x})\, d\theta. \end{aligned} \tag{9.42}$$

Choosing $\alpha = \beta$ then gives a central interval, with e.g. $1 - \alpha - \beta = 68.3\%$. Another possibility is to choose α and β such that all values of $p(\theta|\mathbf{x})$ inside the interval $[a, b]$ are higher than any values outside, which implies $p(a|\mathbf{x}) = p(b|\mathbf{x})$. One can show that this gives the shortest possible interval.

One advantage of a Bayesian interval is that prior knowledge, e.g. $\theta \geq 0$, can easily be incorporated by setting the prior p.d.f. $\pi(\theta)$ to zero in the excluded region. Bayes' theorem then gives a posterior probability $p(\theta|\mathbf{x})$ with $p(\theta|\mathbf{x}) = 0$ for $\theta < 0$. The upper limit is thus determined by

$$1 - \beta = \int_{-\infty}^{\theta_{\mathrm{up}}} p(\theta|\mathbf{x}) d\theta = \frac{\int_{-\infty}^{\theta_{\mathrm{up}}} L(\mathbf{x}|\theta)\, \pi(\theta)\, d\theta}{\int_{-\infty}^{\infty} L(\mathbf{x}|\theta)\, \pi(\theta)\, d\theta}. \tag{9.43}$$

The difficulties here have already been mentioned in Section 6.13, namely that there is no unique way to specify the prior density $\pi(\theta)$. A common choice is

$$\pi(\theta) = \begin{cases} 0 & \theta < 0 \\ 1 & \theta \geq 0. \end{cases} \tag{9.44}$$

The prescription says in effect: normalize the likelihood function to unit area in the physical region, and then integrate it out to θ_{up} such that the fraction of area covered is $1 - \beta$. Although the method is simple, it has some conceptual drawbacks. For the case where one knows $\theta \geq 0$ (e.g. the neutrino mass) one does not really believe that $0 < \theta < 1$ has the same prior probability as $10^{40} < \theta < 10^{40} + 1$. Furthermore, the upper limit derived from $\pi(\theta) =$ constant is not invariant with respect to a nonlinear transformation of the parameter.

It has been argued [Jef48] that in cases where $\theta \geq 0$ but with no other prior information, one should use

$$\pi(\theta) = \begin{cases} 0 & \theta \leq 0 \\ \frac{1}{\theta} & \theta > 0. \end{cases} \tag{9.45}$$

This has the advantage that upper limits are invariant with respect to a transformation of the parameter by raising to an arbitrary power. This is equivalent to a uniform (improper) prior of the form (9.44) for $\log \theta$. For this to be usable, however, the likelihood function must go to zero for $\theta \to 0$ and $\theta \to \infty$, or else the integrals in (9.43) diverge. It is thus not applicable in a number of cases of practical interest, including the example discussed in this section. Therefore, despite its conceptual difficulties, the uniform prior density is the most commonly used choice for setting limits on parameters.

Figure 9.8 shows the upper limits at 95% confidence level derived according to the classical, shifted and Bayesian techniques as a function of $\hat{\theta}_{\mathrm{obs}} = x - y$ for $\sigma_{\hat{\theta}} = 1$. For the Bayesian limit, a prior density $\pi(\theta) =$ constant was used. The shifted and classical techniques are equal for $\hat{\theta}_{\mathrm{obs}} \geq 0$. The Bayesian limit is always positive, and is always greater than the classical limit. As $\hat{\theta}_{\mathrm{obs}}$ becomes larger than the experimental resolution $\sigma_{\hat{\theta}}$, the Bayesian and classical limits rapidly approach each other.

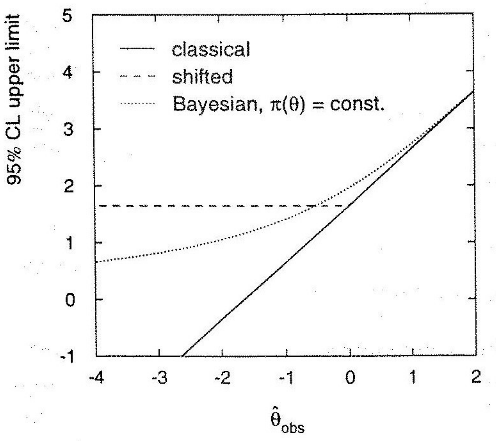

Fig. 9.8 Upper limits at 95% confidence level for the example of Section 9.8 using the classical, shifted and Bayesian techniques. The shifted and classical techniques are equal for $\hat{\theta}_{\mathrm{obs}} \geq 0$.

9.9 Upper limit on the mean of Poisson variable with background

As a final example, recall Section 9.4 where an upper limit was placed on the mean ν of a Poisson variable n. Often one is faced with a somewhat more complicated situation where the observed value of n is the sum of the desired signal events n_{s} as well as background events n_{b},

$$n = n_{\mathrm{s}} + n_{\mathrm{b}}, \tag{9.46}$$

where both n_{s} and n_{b} can be regarded as Poisson variables with means ν_{s} and ν_{b}, respectively. Suppose for the moment that the mean for the background ν_{b} is known without any uncertainty. For ν_{s} one only knows a priori that $\nu_{\mathrm{s}} \geq 0$. The goal is to construct an upper limit for the signal parameter ν_{s} given a measured value of n.

Since n is the sum of two Poisson variables, one can show that it is itself a Poisson variable, with the probability function

$$f(n;\nu_{\rm s},\nu_{\rm b}) = \frac{(\nu_{\rm s}+\nu_{\rm b})^n}{n!}\, e^{-(\nu_{\rm s}+\nu_{\rm b})}. \tag{9.47}$$

The ML estimator for $\nu_{\rm s}$ is

$$\hat{\nu}_{\rm s} = n - \nu_{\rm b}, \tag{9.48}$$

which has zero bias since $E[n] = \nu_{\rm s} + \nu_{\rm b}$. Equations (9.15), which are used to determine the confidence interval, become

$$\begin{aligned} \alpha &= P(\hat{\nu}_{\rm s} \geq \hat{\nu}_{\rm s}^{\rm obs}; \nu_{\rm s}^{\rm lo}) = \sum_{n \geq n_{\rm obs}} \frac{(\nu_{\rm s}^{\rm lo}+\nu_{\rm b})^n\, e^{-(\nu_{\rm s}^{\rm lo}+\nu_{\rm b})}}{n!}, \\ \beta &= P(\hat{\nu}_{\rm s} \leq \hat{\nu}_{\rm s}^{\rm obs}; \nu_{\rm s}^{\rm up}) = \sum_{n \leq n_{\rm obs}} \frac{(\nu_{\rm s}^{\rm up}+\nu_{\rm b})^n\, e^{-(\nu_{\rm s}^{\rm up}+\nu_{\rm b})}}{n!}. \end{aligned} \tag{9.49}$$

These can be solved numerically for the lower and upper limits $\nu_{\rm s}^{\rm lo}$ and $\nu_{\rm s}^{\rm up}$. Comparing with the case $\nu_{\rm b} = 0$, one sees that the limits from (9.49) are related to what would be obtained without background by

$$\begin{aligned} \nu_{\rm s}^{\rm lo} &= \nu_{\rm s}^{\rm lo}(\text{no background}) - \nu_{\rm b}, \\ \nu_{\rm s}^{\rm up} &= \nu_{\rm s}^{\rm up}(\text{no background}) - \nu_{\rm b}. \end{aligned} \tag{9.50}$$

The difficulties here are similar to those encountered in the previous example. The problem occurs when the total number of events observed $n_{\rm obs}$ is not large compared to the expected number of background events $\nu_{\rm b}$. Values of $\nu_{\rm s}^{\rm up}$ for $1-\beta = 0.95$ are shown in Fig. 9.9(a) as a function of the expected number of background events $\nu_{\rm b}$. For small enough $n_{\rm obs}$ and a high enough background level $\nu_{\rm b}$, a non-negative solution for $\nu_{\rm s}^{\rm up}$ does not exist. This situation can occur, of course, because of fluctuations in $n_{\rm s}$ and $n_{\rm b}$.

Because of these difficulties, the classical limit is not recommended in this case. As previously mentioned, one should always report $\hat{\nu}_{\rm s}$ and an estimate of its variance even if $\hat{\nu}_{\rm s}$ comes out negative. In this way the average of many experiments will converge to the correct value. If, in addition, one wishes to report an upper limit on $\nu_{\rm s}$, the Bayesian method can be used with, for example, a uniform prior density [Hel83]. The likelihood function is given by the probability (9.47), now regarded as a function of $\nu_{\rm s}$,

$$L(n_{\rm obs}|\nu_{\rm s}) = \frac{(\nu_{\rm s}+\nu_{\rm b})^{n_{\rm obs}}}{n_{\rm obs}!}\, e^{-(\nu_{\rm s}+\nu_{\rm b})}. \tag{9.51}$$

The posterior probability density for $\nu_{\rm s}$ is obtained as usual from Bayes' theorem,

$$p(\nu_{\rm s}|n_{\rm obs}) = \frac{L(n_{\rm obs}|\nu_{\rm s})\, \pi(\nu_{\rm s})}{\int_{-\infty}^{\infty} L(n_{\rm obs}|\nu_{\rm s}')\, \pi(\nu_{\rm s}')\, d\nu_{\rm s}'}. \tag{9.52}$$

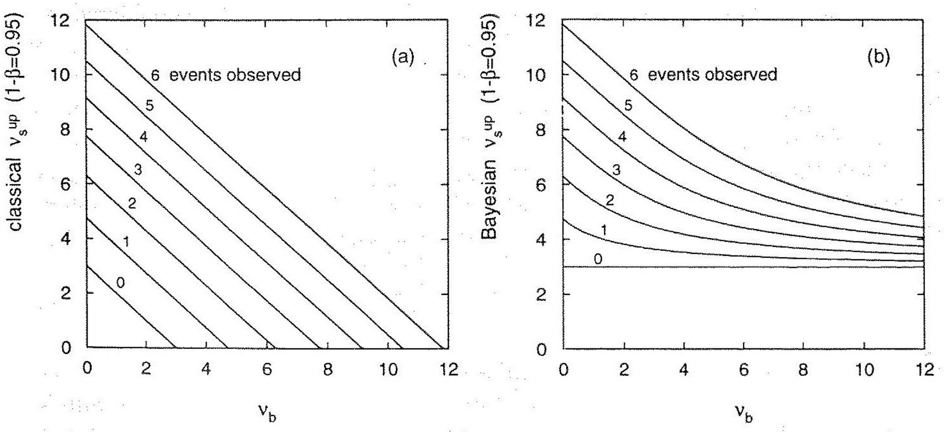

Fig. 9.9 Upper limits $\nu_{\rm s}^{\rm up}$ at a confidence level of $1-\beta = 0.95$ for different numbers of events observed $n_{\rm obs}$ and as a function of the expected number of background events $\nu_{\rm b}$. (a) The classical limit. (b) The Bayesian limit based on a uniform prior density for $\nu_{\rm s}$.

Taking $\pi(\nu_{\rm s})$ to be constant for $\nu_{\rm s} \geq 0$ and zero for $\nu_{\rm s} < 0$, the upper limit $\nu_{\rm s}^{\rm up}$ at a confidence level of $1-\beta$ is given by

$$\begin{aligned} 1-\beta &= \frac{\int_0^{\nu_{\rm s}^{\rm up}} L(n_{\rm obs}|\nu_{\rm s})\, d\nu_{\rm s}}{\int_0^{\infty} L(n_{\rm obs}|\nu_{\rm s})\, d\nu_{\rm s}} \\ &= \frac{\int_0^{\nu_{\rm s}^{\rm up}} (\nu_{\rm s}+\nu_{\rm b})^{n_{\rm obs}}\, e^{-(\nu_{\rm s}+\nu_{\rm b})}\, d\nu_{\rm s}}{\int_0^{\infty} (\nu_{\rm s}+\nu_{\rm b})^{n_{\rm obs}}\, e^{-(\nu_{\rm s}+\nu_{\rm b})}\, d\nu_{\rm s}}. \end{aligned} \tag{9.53}$$

The integrals can be related to incomplete gamma functions (see e.g. [Arf95]), or since $n_{\rm obs}$ is a positive integer, they can be solved by making the substitution $x = \nu_{\rm s} + \nu_{\rm b}$ and integrating by parts $n_{\rm obs}$ times. Equation (9.53) then becomes

$$\beta = \frac{e^{-(\nu_{\rm s}^{\rm up}+\nu_{\rm b})} \sum_{n=0}^{n_{\rm obs}} \frac{(\nu_{\rm s}^{\rm up}+\nu_{\rm b})^n}{n!}}{e^{-\nu_{\rm b}} \sum_{n=0}^{n_{\rm obs}} \frac{\nu_{\rm b}^n}{n!}}. \tag{9.54}$$

This can be solved numerically for the upper limit $\nu_{\rm s}^{\rm up}$. The upper limit as a function of $\nu_{\rm b}$ is shown in Fig. 9.9(b) for various values of $n_{\rm obs}$. For the case without background, setting $\nu_{\rm b} = 0$ gives

$$\beta = e^{-\nu_{\rm s}^{\rm up}} \sum_{n=0}^{n_{\rm obs}} \frac{(\nu_{\rm s}^{\rm up})^n}{n!}, \tag{9.55}$$

which is identical to the equation for the classical upper limit (9.16). This can be seen by comparing Figs 9.9(a) and (b). The Bayesian limit is always greater than or equal to the corresponding classical one, with the two agreeing only for $\nu_{\rm b} = 0$.

The agreement for the case without background must be considered accidental, however, since the Bayesian limit depends on the particular choice of a constant prior density $\pi(\nu_{\rm s})$. Nevertheless, the coincidence spares one the trouble of having to defend either the classical or Bayesian viewpoint, which may account for the general acceptance of the uniform prior density in this case.

Often the result of an experiment is not simply the number n of observed events, but includes in addition measured values $x_1, \ldots, x_n$ of some property of the events. Suppose the probability density for x is

$$f(x; \nu_{\rm s}, \nu_{\rm b}) = \frac{\nu_{\rm s} f_{\rm s}(x) + \nu_{\rm b} f_{\rm b}(x)}{\nu_{\rm s} + \nu_{\rm b}}, \tag{9.56}$$

where the components $f_{\rm s}(x)$ for signal and $f_{\rm b}(x)$ for background events are both assumed to be known. If these p.d.f.s have different shapes, then the values of x contain additional information on whether the observed events were signal or background. This information can be incorporated into the limit $\nu_{\rm s}$ by using the extended likelihood function,

$$\begin{aligned} L(\nu_{\rm s}) &= \frac{(\nu_{\rm s} + \nu_{\rm b})^n}{n!} e^{-(\nu_{\rm s}+\nu_{\rm b})} \prod_{i=1}^{n} \frac{\nu_{\rm s} f_{\rm s}(x_i) + \nu_{\rm b} f_{\rm b}(x_i)}{\nu_{\rm s} + \nu_{\rm b}} \\ &= \frac{e^{-(\nu_{\rm s}+\nu_{\rm b})}}{n!} \prod_{i=1}^{n} [\nu_{\rm s} f_{\rm s}(x_i) + \nu_{\rm b} f_{\rm b}(x_i)], \qquad (9.57) \end{aligned}$$

as defined in Section 6.9, or by using the corresponding formula for binned data as discussed in Section 6.10.

In the classical case, one uses the likelihood function to find the estimator $\hat{\nu}_{\rm s}$. In order to find the classical upper limit, however, one requires the p.d.f. of $\hat{\nu}_{\rm s}$. This is no longer as simple to find as before, where only the number of events was counted, and must in general be determined numerically. For example, one can perform Monte Carlo experiments using a given value of $\nu_{\rm s}$ (and the known value $\nu_{\rm b}$) to generate numbers $n_{\rm s}$ and $n_{\rm b}$ from a Poisson distribution, and corresponding x values according to $f_{\rm s}(x; \nu_{\rm s})$ and $f_{\rm b}(x; \nu_{\rm b})$. By adjusting $\nu_{\rm s}$, one can find that value for which there is a probability β to obtain $\hat{\nu}_{\rm s} \leq \hat{\nu}_{\rm s}^{\rm obs}$. Here one must still deal with the problem that the limit can turn out negative.

In the Bayesian approach, $L(\nu_{\rm s})$ is used directly in Bayes' theorem as before. Solving equation (9.53) for $\nu_{\rm s}^{\rm up}$ must in general be done numerically. This has the advantage of not requiring the sampling p.d.f. for the estimator $\hat{\nu}_{\rm s}$, in addition to the previously mentioned advantage of automatically incorporating the prior knowledge $\nu_{\rm s} \geq 0$ into the limit.

Further discussion of the issue of Bayesian versus classical limits can be found in [Hig83, Jam91, Cou95]. A technique for incorporating systematic uncertainties in the limit is given in [Cou92].

10 Characteristic functions and related examples

10.1 Definition and properties of the characteristic function

The **characteristic function** $\phi_x(k)$ for a random variable x with p.d.f. $f(x)$ is defined as the expectation value of e^{ikx},

$$\phi_x(k) = E[e^{ikx}] = \int_{-\infty}^{\infty} e^{ikx} f(x) dx. \tag{10.1}$$

This is essentially the Fourier transform of the probability density function. It is useful in proving a number of important theorems, in particular those involving sums of random variables. One can show that there is a one-to-one correspondence between the p.d.f. and the characteristic function, so that knowledge of one is equivalent to knowledge of the other. Some characteristic functions of important p.d.f.s are given in Table 10.1.

Suppose one has n independent random variables $x_1, \ldots, x_n$, with p.d.f.s $f_1(x_1), \ldots, f_n(x_n)$, and corresponding characteristic functions $\phi_1(k), \ldots, \phi_n(k)$, and consider the sum $z = \sum_i x_i$. The characteristic function $\phi_z(k)$ for z is related to those of the x_i by

$$\begin{aligned} \phi_z(k) &= \int \ldots \int \exp\left(ik \sum_{i=1}^{n} x_i \right) f_1(x_1) \ldots f_n(x_n) dx_1 \ldots dx_n \\ &= \int e^{ikx_1} f_1(x_1) dx_1 \ldots \int e^{ikx_n} f_n(x_n) dx_n \\ &= \phi_1(k) \ldots \phi_n(k). \qquad (10.2) \end{aligned}$$

That is, the characteristic function for a sum of independent random variables is given by the product of the individual characteristic functions.

The p.d.f. $f(z)$ is obtained from the inverse Fourier transform,

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_z(k)\, e^{-ikz}\, dk. \tag{10.3}$$