PROBABILITY AND STATISTICS
APPLIED TO HIGH-ENERGY PHYSICS

D. Drijard, W.T. Eadie
F. James, M. Roos, B. Sadoulet

Lectures given in the
Academic Training Programme of CERN
1968 - 1969
4. Determination of error on parameter estimate
   4.1 Statement of the problem
   4.2 The normal distribution
       4.2.1 Confidence region for multinormal case
   4.3 Confidence intervals (classical) for the general case
       4.3.1 Central and non-central intervals
       4.3.2 Asymptotic confidence intervals
       4.3.3 Confidence intervals from likelihood function
   4.4 Bayesian approach to confidence intervals
       4.4.1 Pathological cases
4. DETERMINATION OF ERROR ON PARAMETER ESTIMATE

4.1 Statement of the problem

We have so far done only point estimation, that is, finding the best single estimate \( \hat{t} \) (previously denoted \( \hat{\theta} \)) of a true parameter value \( \theta \). This estimate will, of course, not be infinitely precise. We have already used the variance of \( t \) as a measure of the error in the sense that it is the mean square deviation from the average value of \( t \) (or from \( \theta \) for unbiased estimators) and we have used the criterion of minimum variance to help us choose the estimator.

But we wish to make stronger statements about the relationship of \( t \) and \( \theta \). For example we would like to be able to say that with a probability \( \beta \), the true parameter value \( \theta \) lies in some range, \( t_{1\beta} \leq \theta \leq t_{2\beta} \). Such a statement, or a series of such statements with different confidence levels \( \beta \), are known as confidence statements. Clearly they tell us much more than do an estimate and a variance, for they give us information about the distribution of \( t \). As we shall see, however, there is no general agreement among statisticians, concerning the way these confidence statements should be arrived at, nor their interpretation.

Once again, we will consider the normal distribution, for it is in the problem of giving confidence intervals for the mean of a normal distribution that all methods give the same results. Moreover, these results can be justified rigorously without establishing new postulates.

4.2 The normal distribution

Consider a normal distribution with unit variance:

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}
\]

(64)

If we make one observation \( X \), it will be distributed according to (64), and if \( \mu \) is known, we can, by integrating (64), calculate the probability of \( X \) falling between two given values, say \( a \) and \( b \)
If \( \mu \) is not known, we can still calculate the probability of finding an \( X \) in some range about the unknown \( \mu \):

\[
\mathcal{P} \left( \mu - a \leq X \leq \mu + b \right) = \int_{\mu - a}^{\mu + b} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{X'}{2\sigma^2}} dX'.
\]

Rearranging the left-hand side of Eq. (66), we get

\[
(66) \quad \mathcal{P} \left( X - b \leq \mu \leq X + a \right) = \beta
\]

This is just the kind of statement we are looking for, but we must be a little careful about its interpretation. Remember that it is (classically) a probability statement about \( X \); not about \( \mu \). From the way it is written, it looks as if we had put bounds on the variation of \( \mu \), but in fact \( \mu \) is an unknown constant, and \( X \) is the random variable. Our probability statement is that the bounds \( X - b \) and \( X + a \) will include the true value \( \mu \) with a probability \( \beta \).

Probability density functions like (64) always allow us to make probability statements about the observable \( X \) if we are given the parameter \( \theta \) (in this case \( \mu \)). It happens in the case of the normal distribution, because of the symmetry of the function between \( X \) and \( \mu \), that the
inversion from a probability statement about \( X \) to a confidence statement about \( \mu \) is algebraically simple. With more complicated functions, the inversion is not so straightforward, and is not always algebraically possible. Several techniques have been proposed to deal with this problem. But first it is worth while to study in more detail the normal case since this will prove useful when considering the general problem.

4.2.1 Confidence regions for the multinormal case

From the Central Limit Theorem, the practical importance of the normal distribution is clear. In the case of many normally distributed parameter estimates, we would like to establish confidence regions with given probability content \( \beta \) in the sense of Section 4.2. Let us define the probability content of a region by an integral of the type (66).

We have already given the principal properties of the many-dimensional normal distribution in Chapter II, Section 2.3.3. The p.d.f. is:

\[
\frac{1}{(2\pi)^{n/2} |V|^{1/2}} e^{-\frac{1}{2} (x - \mu)^T V^{-1} (x - \mu)}
\]

where \( x \) is the vector of the \( n \) variables, \( \mu \) is the vector of their mean values, and \( V \) is their covariance matrix, which is the inverse of the second derivative matrix:

\[
\sqrt{\frac{1}{2} \frac{\partial^2 f}{\partial x_i \partial x_j}} \bigg|_{x = \mu} = \frac{\partial^2 f}{\partial x_i \partial x_j} \bigg|_{x = \mu}
\]

In the \( n \)-dimensional space when (68) is defined, the hypersurfaces of constant probability density are hyperellipsoids, and this is a convenient choice for the shape of the confidence regions. The equation of such a hyperellipsoid is:
where \( \alpha^2 \) is a constant defining the particular surface. The probability content of the region inside the surface (69) is

\[
\beta = P \left( \chi^2(n) \leq \alpha^2 \right),
\]

where \( \chi^2(n) \) is the usual chi-squared function with \( n \) degrees of freedom. [Note that the factor \( \frac{1}{2} \) in the exponent of (68) has been included in \( \alpha^2 \) in (69). There is always a factor 2 when going from likelihood (68) to \( \chi^2 \) (70).] Usually we wish to specify the confidence coefficient (probability content) \( \beta \) and find the \( \alpha^2 \) that defines the confidence region. In Table IV.1 we give values of \( \alpha^2 \) as a function of \( n \) and \( \beta \) allowing the determination of such regions.

These confidence regions can be drawn for the case \( n = 2 \), which is sufficiently general to allow us to visualize the general properties. This is done Fig. IV.3, where the covariance matrix \( V \) has been written:

\[
V = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix}.
\]

\( \rho \) is the correlation coefficient, satisfying \(-1 \leq \rho \leq 1\).
<table>
<thead>
<tr>
<th>No. parameters</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>1.07</td>
<td>2.70</td>
<td>3.84</td>
<td>6.63</td>
</tr>
<tr>
<td>2</td>
<td>1.39</td>
<td>2.41</td>
<td>4.61</td>
<td>5.99</td>
<td>9.21</td>
</tr>
<tr>
<td>3</td>
<td>2.37</td>
<td>3.67</td>
<td>6.25</td>
<td>7.82</td>
<td>11.36</td>
</tr>
<tr>
<td>4</td>
<td>3.36</td>
<td>4.88</td>
<td>7.70</td>
<td>9.49</td>
<td>13.28</td>
</tr>
<tr>
<td>5</td>
<td>4.35</td>
<td>6.06</td>
<td>9.24</td>
<td>11.07</td>
<td>15.09</td>
</tr>
<tr>
<td>6</td>
<td>5.35</td>
<td>7.23</td>
<td>10.65</td>
<td>12.59</td>
<td>16.81</td>
</tr>
<tr>
<td>7</td>
<td>6.35</td>
<td>8.38</td>
<td>12.02</td>
<td>14.07</td>
<td>18.49</td>
</tr>
<tr>
<td>8</td>
<td>7.34</td>
<td>9.52</td>
<td>13.36</td>
<td>15.51</td>
<td>20.09</td>
</tr>
<tr>
<td>9</td>
<td>8.34</td>
<td>10.66</td>
<td>14.68</td>
<td>16.92</td>
<td>21.67</td>
</tr>
<tr>
<td>10</td>
<td>9.34</td>
<td>11.78</td>
<td>15.99</td>
<td>18.31</td>
<td>23.21</td>
</tr>
</tbody>
</table>

*) For -ln likelihood, divide these values of $\alpha^2$ by 2.0.

**) Confidence = limit of confidence that all parameters are within this error region at the same time.
(a) parameters $X_1$ and $X_2$ uncorrelated
    independent (because normal)

(b) parameters $X_1$ and $X_2$
    positively correlated

(c) parameters $X_1$ and $X_2$
    completely correlated
    $\rho = 1$
    covariance matrix singular

Fig. IV.3
Note that the regions in Eq. (69) and in Fig. IV.3 put confidence bounds on all the parameters. This amounts to making a probability statement about the simultaneous values of all parameters. [It is for this reason that the confidence regions defined by \( \alpha^2 \) get large for large \( n \) as seen in Table IV.1]. However, these regions are not rectangular. We cannot, by this technique, set independent confidence limits on each parameter, to be satisfied simultaneously even if the parameters are independent as in Fig. IV.3a. The meaning of these confidence limits for several parameters is clear from Fig. IV.4, where, to be specific, we give the probability content for different regions when \( \alpha = 1 \) and \( \rho = 0 \). In Fig. IV.4, the inner ellipse corresponds to the confidence region (69). The rectangle circumscribed about the ellipse is the confidence region defined by:

\[
P \left[ \left( \mu_1 - \sigma_1 \leq X_1 \leq \mu_1 + \sigma_1 \right) \right.
\]

\[
\left. \text{and} \left( \mu_2 - \sigma_2 \leq X_2 \leq \mu_2 + \sigma_2 \right) \right] = \beta_2.
\]

The long horizontal band defined by the lines \( X_2 = \mu_2 \pm \sigma_2 \) is a one-parameter confidence interval, corresponding to the statement that, whatever the value of \( X_1 \),

\[
P \left[ \left( \mu_2 - \sigma_2 \leq X_2 \leq \mu_2 + \sigma_2 \right) \right] = \beta_3.
\]

From a table of the \( \chi^2 \) function, using (70), we can find the probability contents:

\[
\beta = 39.3\% \\
\beta_2 = 46.6\% = (0.683)^2 \\
\beta_3 = 68.3\%
\]

Example: Consider the estimation of the mass and width of a resonance from an observed mass spectrum. Suppose that we have obtained normally-distributed estimates (this is not necessary as we will see later) with a known (or estimated) covariance matrix. We wish to quote errors on the mass \( M \) and width \( \Gamma \), of the resonance. We may then proceed in different
Fig. IV.4

ways, depending on what we wish to establish. For example, if we want to give a probability that both $M$ and $\Gamma$ lie in some (elliptical) region, we must use (71). However, if we only want to use the data to determine a confidence region for $M$, we can use (72) which, given $\beta$, allow us to establish tighter bounds on $M$ because we say nothing about $\Gamma$. (This does not prevent one of our collaborators from doing the same thing for $\Gamma$ alone using the same data.) This means that we may give a 68% confidence region for $M$ and a 68% confidence region for $\Gamma$, but we know that the probability is only 46% that both regions include the correct values simultaneously if they are not correlated.
4.3 Confidence intervals (classical) for the general case (1 dimension)

We give here the classical method for "inverting" the probability statement of a general p.d.f. to get confidence intervals for the parameters. We must first choose the desired confidence level $\beta$. Given the p.d.f. $f(X|\theta)$ and the estimator $T(X)$ we can calculate the probability of getting a value $t$ of $T(X)$ in some range. (If this is difficult analytically, it can always be done by Monte Carlo since it is only an integration). In particular, we can find two values $t_{11}$ and $t_{12}$ such that the probability of finding a $t$ in this range is $\beta$:

$$P\left( t_{11} \leq t \leq t_{12} / \theta_{1} \right) = \beta$$

These two points are plotted in Fig. IV.5 on the $\theta$- and $t$-axes. For another value of $\theta$, say $\theta_{2}$, we can repeat this calculation to find $t_{21}$ and $t_{22}$, and so on, as shown in Fig. IV.5.
We now draw one smooth curve connecting the points \( t_{11} \) and another connecting the points \( t_{12} \). The region between the curves is known as the confidence belt. It contains probability \( \beta \) along each horizontal line.

Suppose that an experiment gives \( t = t_0 \). We can then establish a confidence interval for \( \theta \) by taking the intersection of the line \( t = t_0 \) with the confidence belt, as shown in Fig. IV.5. Notice that the confidence belt was constructed horizontally but is read vertically. In pathological cases, the confidence belt may wiggle in such a way that the resulting confidence interval consists of several disconnected pieces. Care must then be taken to assure that the results are still meaningful.

Although a rigorous justification of this technique is difficult (if not impossible) we can easily see that it is a good and reasonable method. Firstly, the interval obtained is clearly invariant under well-behaved variable transformations \( \theta \rightarrow t(\theta) \). This is a necessary property of a reasonable confidence statement.

Secondly, this technique can be justified by using the cumulative distribution function \( F(t/\theta) \). For each of a set of different values of \( \theta \), we draw \( F(t/\theta_i) \) as in the example of Fig. IV.6.
Now, given an experimental value $t_0$, we draw the vertical line shown in Fig. IV.6 and mark off a distance $\beta$ on it between points A and B. From the curves $F(t|\Theta_A)$ and $F(t|\Theta_B)$ passing through A and B, we find the values $\Theta_A$ and $\Theta_B$, which are the confidence bounds.

4.3.1 Central and non-central intervals

Having chosen the desired confidence coefficient $\beta$, we still have not determined the confidence interval uniquely. In fact, many different intervals will have the same probability content. The confidence regions of Section 4.2.1 are known as central confidence intervals since the probability content outside the interval is the same on each side. That is, $t_1$ and $t_2$ determine a central confidence interval of probability $\beta$ if

$$
\Pr(\Theta \geq t_2) = \Pr(\Theta \leq t_1) = \frac{1-\beta}{2}. \tag{73}
$$

Speaking strictly in terms of probability, there is no special reason although it is customary to quote central intervals when one can find some other interval containing the same probability. Depending on the problem, we may indeed prefer to quote only upper (or lower) confidence bounds, for example

$$
\Pr(\Theta \geq t_1) = \beta.
$$

In fact this will sometimes be unavoidable in case the confidence belt becomes vertical as in Fig. IV.7.
One should be careful in handling such cases, but once the problem is recognized, there is usually a straightforward way to treat it. In particular, the generalization from central to non-central or one-sided confidence regions should pose no conceptual difficulties.

Example:

Consider the lower confidence bounds for the mean $\lambda$ of a Poisson distribution, for instance the number of events observed in a histogram bin, when the probability for the bin is small.

Given an observation $n$, the number of events in a bin, with mean $\lambda$, has the distribution

$$P\left( N = n \right) = e^{-\lambda} \frac{\lambda^n}{n!}.$$
We want to find $\lambda$ such that

$$P\left( \frac{N > n}{\lambda} \right) = \sum_{n=n}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} = \alpha$$

Since there is a unique solution to this equation, $\lambda$ is a lower confidence bound for $\lambda$. The solution can be obtained from the fact that

$$P\left( \frac{N \leq n}{\lambda} \right) = 1 - P\left\{ \chi^2(v) < 2\lambda \right\}$$

with $v = 2(n+1)$

$$P\left( \frac{N \leq n-1}{\lambda} \right) = 1 - P\left\{ \chi^2(v) < 2\lambda \right\}$$

with $v = 2n$

$$P\left( \frac{N \geq n}{\lambda} \right) = P\left\{ \chi^2(v) < 2\lambda \mid v = 2n \right\}$$

= $\alpha$

From tables of $\chi^2$ distributions, we have

<table>
<thead>
<tr>
<th>$n$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>0.051</td>
<td>0.355</td>
<td>0.82</td>
<td>1.37</td>
<td>1.92</td>
</tr>
</tbody>
</table>

for the 5% level.

Let $\lambda = 0.051$ and $P(N < 1) = 95\%$. Then the lower confidence bound for $\lambda$ at the 5% level, given $n = 1$, is $\lambda = 0.051$.

4.3.2 Asymptotic confidence intervals

In Section 3.1.4 we have seen that $\hat{\theta}$, the maximum likelihood estimate of $\theta$ is asymptotically distributed as $N(\theta, \frac{1}{n\lambda})$, when $n$ observations have been made.

This immediately suggests a confidence interval, from the calculation used in the normal case above. The interval so obtained is $(\hat{\theta}, \bar{\theta})$ where
\[
\theta = \hat{\theta} - \frac{\lambda a/2}{\sqrt{n I(\hat{\theta})}} \tag{74}
\]

and

\[
\tilde{\theta} = \hat{\theta} + \frac{\lambda a/2}{\sqrt{n I(\hat{\theta})}}
\]

and \(\lambda a/2\) is the \(a/2\)-level point of the standard normal distribution.

However, because of the bias of the estimate \(\hat{\theta}\) and the fact that the asymptotic behaviour may be reached only slowly this method can give only a rough confidence interval.

A more accurate confidence interval (in fact, near the true value \(\theta_0\), the most accurate interval) is obtained by using the fact that \(\lambda \ln L/\lambda \theta\) is distributed with mean 0 and variance \(nI(\theta)\) for all \(n\). The approximation now used is that the distribution of

\[
\frac{1}{\sqrt{n I(\hat{\theta})}} \frac{\lambda \ln L}{\lambda \theta}
\]

is standard normal, and it has the exact mean and variance.

The interval is then

\[
\left| \frac{1}{\sqrt{n I(\hat{\theta})}} \frac{\lambda \ln L}{\lambda \theta} \right| < \frac{\lambda a}{2} \tag{75}
\]

Example:

Consider the case of observations \(X\) taken from the negative exponential distribution

\[
\phi (X | \mu) = \frac{1}{\mu} e^{-\frac{X}{\mu}}
\]
Then
\[ \ln L = -n \ln \mu - \frac{\sum X_i}{\mu} \]
and from this,
\[ \frac{\partial \ln L}{\partial \mu} = -\frac{n}{\mu} + \frac{\sum X_i}{\mu^2} \]
\[ \hat{\mu} = \bar{X} \]
\[ \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2} - 2 \frac{\sum X_i}{\mu^3} \]
\[ \Xi \left( -\frac{\partial^2 \ln L}{\partial \mu^2} \right) = \frac{n}{\mu^2} = n \mathcal{I}(\mu) \]
\[ n \mathcal{I}(\hat{\mu}) = \frac{n}{\bar{X}^2} \]

The rough confidence interval is now
\[ \bar{X} \pm Z_{\alpha/2} \frac{\bar{X}}{\sqrt{n}} \]
or
\[ \bar{X} \left( 1 - \frac{Z_{\alpha/2}}{\sqrt{n}} \right) \leq \mu \leq \bar{X} \left( 1 + \frac{Z_{\alpha/2}}{\sqrt{n}} \right) \]
The locally most accurate interval is given asymptotically by

\[
\left| \frac{\bar{x}}{\mu^2} - \frac{\bar{x}}{\mu} \right| \leq \frac{\lambda \mu}{\sqrt{n} \mu}
\]

and hence

\[
\frac{\bar{x}}{1 + \frac{\lambda d/2}{\sqrt{n}}} \leq \mu \leq \frac{\bar{x}}{1 - \frac{\lambda d/2}{\sqrt{n}}}
\]

Note that for \( \theta = 1/\mu \), both methods give the same interval.

4.3.3 Confidence intervals from the likelihood function

Consider the case of a normal distribution with unit variance, and unknown mean \( \mu \). Then the likelihood for a set of \( n \) observations is given by Eq. (III.11) with \( \sigma = 1 \).

As a function of \( \mu \), \( L(x/\mu) \) has the same belt-shaped form as the normal probability density function. The log-likelihood is a parabola in \( \mu \),

\[
\ln L (x/\mu) = \ln c - \frac{1}{2} \sum (x_i - \bar{x})^2 - \frac{n}{2} (\mu - \bar{x})^2
\]

The maximum occurs at \( \mu = \bar{x} \). In Fig. IV.8, we show the case of \( n = 1 \), after a change of origin to give \( \ln L = 0 \) at the maximum. A line drawn at \( \ln L = -0.5 \) gives the interval \( \bar{x} - 1 \leq \mu \leq \bar{x} + 1 \) [from \( (\mu - X)^2/2 = \frac{1}{2} \)].
From the properties of the normal distribution of $X$, we have

$$P[(X - \mu)^2 \leq 1] = 68.3\%$$

or

$$P[-1 \leq X - \mu \leq 1] = 68.3\%$$

or

$$P[\bar{X} - 1 \leq \mu \leq \bar{X} + 1] = 68.3\%$$

Similarly, a line drawn at $\ln L = -2$ yields the usual 95.5% confidence interval for the parameter $\mu$. As before, the probability statements are based on the properties of the random variable $\bar{X}$, not on any properties of the parameter $\mu$. In particular, the interval has the same properties, if the experiment is repeated many times. Thus 4.5% of the trials, the interval given by analysing the likelihood function will not include the true value. This is easily demonstrated by a simple Monte Carlo experiment. The reason is simply that there is a 4.5% chance of obtaining an observation further than 2 standard deviations from the true mean $\mu$ (Fig. IV.9).
A useful property is that maximum likelihood estimates of functions of parameters are the same functions of the maximum likelihood estimates of the parameters.

Thus if \( g = g(\mu) \) and \( \hat{\mu} \) is a maximum likelihood estimate of \( \mu \), then \( \hat{g} \) is a maximum likelihood estimate of \( g = g(\hat{\mu}) \).

Suppose two experiments \( E_1 \) and \( E_2 \) give rise to two sets of data \( X_1 \) and \( X_2 \), and that each experiment has one unknown parameter, \( \theta_1 \) and \( \theta_2 \), respectively. If the likelihood functions \( L_1(\theta_1 | X_1) \) and \( L_2(\theta_2 | X_2) \) are equal, then we must make the same inferences about \( \theta_2 \) as about \( \theta_1 \), as far as estimates, confidence intervals, etc., are concerned. The same conclusions apply if the two functions are proportional

\[
L_1(\theta_1 | X_1) = k L_2(\theta_2 | X_2)
\]

where \( k \) is independent of \( \theta_1 \) and \( \theta_2 \). This is called the likelihood principle.
Let us generalize now from the normal case, treated above, and suppose that \( \ln L(\theta \mid X) \) is a continuous function of \( \theta \), with only one maximum in the region of interest of \( \theta \). There exists transformations \( g(\theta, X) \), which transforms the curve \( \ln L(\theta \mid X) \) into a parabola around a function \( G \) of the observations.

\[
\ln L_g(g \mid X) = \frac{1}{2} (g - G(X))^2.
\]

Asymptotically it can be shown that one may choose \( g \) independent of \( X \). Then \( G \) is the maximum likelihood estimate of \( g \).

Now by the likelihood principle, we can make the same inferences about \( g \) as we made about the normal parameter \( \mu \). In particular, one 95.5\% confidence interval, say \( \underline{g} \leq g \leq \bar{g} \), is given for \( g \) by

\[
G(X) - 2 \leq g \leq G(X) + 2,
\]

where

\[
G(X) = \bar{g}.
\]

We now wish to transform this confidence interval for \( g \) into a corresponding confidence interval for \( \theta \). This is done more easily by considering the likelihood functions.

Because

\[
\ln L_\theta(\theta \mid X) = \ln L_g(g(\theta) \mid X)
\]

and the confidence interval \((g, \bar{g})\) in \( g \) correspond to the intersection of \( \ln L_g \), with \( \ln L = -2 \), the values of \( \theta \) corresponding to \( \underline{g}, \bar{g} \) are the intersections of \( \ln L = -2 \) with \( \ln L_\theta(\theta \mid x) \), cf. Fig. IV.10.

This is an extremely useful procedure. By using the properties of the likelihood function and maximum likelihood estimates, it is possible, without actually finding the necessary transformation, to make inferences about a parameter by transforming it in such a way that the observed likelihood is distributed normally in terms of the transformed parameter, about which appropriate inferences can be made. The inferences, from the theory of probability, are then transformed back to the original parameter. The likelihood function thus provides an elegant transformation mechanism.
Some words of caution may, however, be necessary.

(i) The transformation method is exact only to order $1/n$. It removes the primary bias term in the likelihood expansion but leaves other terms in $1/n$. The point here is that we have made the experimental likelihood distribution normal in the parameter instead of finding a transformation that would have made the theoretical distribution normal. To order $1/n$ this is the same thing, but for small samples, it is not exact.

(ii) The intervals obtained are those which are central on the transformed, normal case. This can result in very wide bounds in the original case, when non-central intervals might be more appropriate.

(iii) When the maximum of the likelihood function is not unique, the procedure no longer leads to a single continuous interval, in general. However, one can make confidence statement of the form

$$P \left( \theta_4 \leq \theta \leq \theta_2 \quad \alpha \quad \theta_3 \leq \theta \leq \theta_4 \right) = 1 - \alpha$$

i.e. probability that either
The interpretation of such a statement as a confidence interval for \( \theta \) is somewhat dubious. It might be possible to find a single continuous interval with the same probability content; such an interval would sometimes be more meaningful.

The proper conclusion is that, in such cases, a confidence interval gives only a very incomplete summary of the results.

(iv) Care must also be taken in cases where the above procedure leads to indefinite (or infinite) confidence intervals, Fig. IV.13. Again this is a case where the implied transformation is of rather pathological nature. It seems probable, that the wrong question is being asked.
concerning the parameters, or that some information concerning the model or the data has been omitted.

In these cases, an alternative transformation, leading to some other distribution than normal, should be looked for. Alternatively, the problem may require a more complex interpretation than obtained simply from a single maximum likelihood estimate, and a confidence bound.

4.4 Bayesian approach to confidence intervals

The Bayesian approach to obtaining "confidence intervals" is, formally, closely related to the use of the likelihood function. However, the interpretation and meaning of the results are totally different.

Using Bayes Theorem (Chapter II, Section 1.3.3), the posterior distribution for a parameter $\theta$, given observations $x$, is

$$P(\theta | x) = \frac{P(x | \theta) P(\theta)}{\int P(x | \theta) P(\theta) d\theta}$$

$$= k L(x | \theta) P(\theta)$$

where $L(x | \theta)$ is the likelihood function of the data $x$, and $k$ is a normalizing constant.

This distribution can be handled like any other, and in particular an interval containing a specified proportion $\beta$ of the distribution can be calculated from

$$\beta = \frac{\int^{\theta} P(\theta | x) d\theta}{\theta}$$
To interpret what is meant by the interval $(\theta - \bar{\theta})$, we must first recall that $P(\theta | x)$ summarizes our degree of belief about $\theta$. Then in giving the interval $(\theta - \bar{\theta})$ we imply that we think it unlikely (probability $= 1 - \beta$) that $\theta$ lies outside this interval. Put in another way, we are prepared to bet that $\theta$ lies in this interval at any odds better than $\beta : (1 - \beta)$ on.

Although the formulation of a prior distribution is a subjective process, it must be borne in mind that the effect of the prior distribution decreases as the number of observations increases. Also, the final decisions to be made are equally subjective, and it may be as well that the subjective knowledge which will be incorporated in the decision should also be included in the analyses providing information for the decision.

4.4.1 Pathological cases

Consider now the case where the likelihood function has an awkward (or pathological) form, as shown in Fig. IV.13. The classical approach, using the likelihood function will produce a reasonable 68.3% confidence interval, but an infinite interval for 95% confidences, although it will give a lower bound. However, if it is decided to give only a lower bound, the correct value is given by the intersection of the log-likelihood with the line $L = 1.34$, the intersections then corresponding to upper and lower 5% points, or a 10% confidence interval. The appropriate intersection can be used as a single 5% confidence bound.

The Bayesian analysis would start from a prior distribution $P(\theta)$, and using the likelihood function, obtain a posterior distribution
If we suppose $P(\theta)$ to be uniform over the range $(0,A)$, and the above likelihood to appear, then we have the following pictures.

(The area under both curves should be the same 1)
What has happened is that our knowledge about large values of $\vartheta$ is not changed, but for small values we do know more. The confidence interval $(\underline{\vartheta}, \overline{\vartheta})$ before the experiment, will change to $(\underline{\vartheta | X}, \overline{\vartheta | X})$ after observing $X$. 