The fiducial argument

To obtain a distribution representing the inferred uncertain knowledge about the parameter directly from the data without access to any prior distribution was the glorious goal that Fisher succeeded in reaching with his fiducial distribution for a scalar parameter, but for vector parameters he got into trouble. He regarded the fiducial distribution as an ordinary probability distribution subject to the usual Kolmogorovian laws for sigma-additive probability measures. Fisher did not develop any mathematical theory of fiducial probability, but chose to illustrate his thoughts by examples. It was soon found that even in Fisher’s examples with more than one parameter, there were no unique fiducial distributions, and inconsistencies and paradoxes were identified. After revisiting Fisher’s ideas over 1930–1935, which underlie our confidence distributions, we summarise and discuss the big debate over the fiducial argument, which died out only after Fisher’s death in 1962, leaving the statistical community to regard it as badly flawed and Fisher’s biggest blunder. The chapter ends with three attempts at saving the fiducial argument. Despite the potential problems with multivariate fiducial distributions, their marginals are often exact or approximate confidence distributions.

6.1 The initial argument

Fisher introduced the fiducial argument in 1930. In the introduction to Fisher (1930) he indicates that he had solved a very important problem that had escaped “the most eminent men of their time” since Bayes introduced his theorem and the method of inverse probability was established by Laplace. Later he stated that he really had a solution to “more than 150 years of disputation between the pros and cons of inverse probability [that] had left the subject only more befogged by doubt and frustration” (Fisher discussing Neyman in Neyman [1934, p. 617]). This was a view he held throughout his life.

Fisher saw the fiducial probability as a probability measure over the parameter inherited from the probability model of the sampling experiment and the observed data. His early interpretation of fiducial probability was very similar to Neyman’s coverage probability for intervals. Later he thought of fiducial distributions as a representation of the information in the data as seen on the background of the statistical model. It would then serve as an epistemic probability distribution for rational people. This is close to the Bayesian understanding of a posterior distribution, at least for people agreeing on the prior distribution. Fisher’s fiducial probability was, however, obtained by direct inference, without any use of a prior distribution.
The fiducial argument

Fisher gave no reason for the term *fiducial*, but used it to emphasise its importance. When discussing Neyman (Neyman, 1934, p. 617) he reserved himself against Neyman’s term (confidence) coefficient because he regarded it as an ordinary probability that is termed “*fiducial*” to show that it was a probability inferred by the fiducial method of reasoning, then unfamiliar, and not by the classical method of *inverse* probability. He actually kept regarding fiducial inference as the jewel in his crown of the “ideas and nomenclature” for which he was responsible (Fisher, 1956, p. 77).

The fiducial argument grew out of the t-distribution of Student (1908) and the work of the Rothamsted scientist E. J. Maskell in developing intervals based on the t-distribution. Fisher was also influenced by H. Hotelling and M. Ezekiel and their work on regression analysis; see Aldrich (2000) for further discussion.

As we shall see, fiducial probability fell subject to a deep controversy that developed after 1935 and lasted for some 30 years, with Fisher becoming more and more isolated in his views. From around 1970, articles on fiducial probability have mainly been on the history of the concept and the controversy, and the fiducial argument itself has been virtually absent from statistical textbooks and curricula from then on; also, fiducial probability in two or more dimensions was virtually dropped from the statistical toolbox – for good reasons. With the new millennium various authors are eager to investigate fiducial and related distributions, however, as discussed in our Preface and in Chapter 1, from different perspectives and in different settings. Hannig (2009) and Hannig and Lee (2009) found that the fiducial argument is useful in certain practical contexts, and a certain generalisation of the fiducial construction is given in Hannig (2009).

The initial paper, Fisher (1930), was well received. Neyman (1934, p. 563) understood the fiducial distribution as a method of constructing confidence intervals, a problem that “has been sought by the greatest minds since the work of Bayes 150 years ago . . . The present solution means, I think not less than a revolution in the theory of statistics”.

The initial form of Fisher’s fiducial argument (Fisher, 1930) was as follows:

There are, however, certain cases in which statements in terms of probability can be made with respect to the parameters of the population. One illustration may be given before considering in what ways its logical content differs from the corresponding statement of a probability inferred from known *a priori* probabilities. In many cases the random sampling distribution of a statistic, \( T \), calculable directly from the observations, is expressible solely in terms of a single parameter, of which \( T \) is the estimate found by the method of maximum likelihood. If \( T \) is a statistic of continuous variation, and \( P \) the probability that \( T \) should be less than any specified value, we have then a relation of the form

\[
P = F(T, \theta).
\]

If now we give to \( P \) any particular value such as 0.95, we have a relationship between the statistic \( T \) and the parameter \( \theta \), such that \( T \) is the 95 per cent. value corresponding to a given \( \theta \), and this relationship implies the perfectly objective fact that in 5 per cent. of samples \( T \) will exceed the 95 per cent. value corresponding to the actual value of \( \theta \) in the population from which it is drawn. To any value of \( T \) there will moreover be usually a particular value of \( \theta \) to which it bears this relationship; we may call this the “*fiducial* 5 per cent. value of \( \theta \)” corresponding to a given \( T \). If, as usually if not always happens, \( T \) increases with \( \theta \) for all possible values, we may express the relationship by saying that the true value of \( \theta \) will be less than the fiducial 5 per cent. value corresponding to the observed value of \( T \) in exactly 5 trials in 100. By constructing a table of corresponding values, we may know as soon as \( T \) is calculated what is the fiducial 5 per cent. value of \( \theta \), and that the true value of \( \theta \) will be less than this value in just 5 per cent. of trials. This then is a definite probability statement.
6.1 The initial argument

about the unknown parameter \( \theta \), which is true irrespective of any assumption as to its \textit{a priori} distribution.

Writing \( p = 1 - P \), the fiducial \( p \)-quantile for the observed value \( T_{\text{obs}} \) is the solution for \( \theta \) to

\[
p = 1 - F(T_{\text{obs}}, \theta).
\]

This is exactly what we call the confidence distribution. From the fiducial distribution function Fisher (1930) wrote up the fiducial density

\[
-\frac{\partial}{\partial \theta} F(T_{\text{obs}}, \theta),
\]

and he noted the difference between this and the likelihood function \( \frac{\partial}{\partial \theta} F(T_{\text{obs}}, \theta) \).

As an example Fisher (1930) considered the correlation coefficient for a binormal sample of size \( n = 4 \) based on the distribution for the empirical correlation he had developed (Fisher, 1915), and he gave a table of the fiducial 5% quantile by the observed correlation coefficient. By symmetry this yields 90% confidence intervals for the correlation coefficient for empirical correlations in normal samples of size 4.

Neyman (1934) understood that the fiducial distribution is a tool for obtaining confidence intervals. For him the probability involved is a coverage probability, that is, \( p \) is the probability that the fiducial 5% quantile shall exceed the true value ex ante. After the data are observed Neyman said that \( p \) is a level of confidence. He did not call it a probability \textit{ex post}. As noted previously, Fisher did not like the term confidence. He regarded his fiducial probabilities as ordinary probabilities, but evidently as epistemic probabilities. His seminal paper in 1930 was actually an attack on inverse probability, which definitely was understood as epistemic probability at the time, and he offered fiducial probability as an alternative to the Bayesian posterior probability that he criticised because it is based on an often unfounded flat prior.

In 1931 Jeffreys found that \( h^{-1} \) could be used as a noninformative prior for the precision \( h = \sigma^{-1}/\sqrt{2} \) in a normal one-sample model. He argued that with only one observation \( x \) there should be no information on \( h \), and the prior density \( p(h) \) should thus equal the posterior density for \( h \) given \( x \), that is, it should satisfy

\[
p(h) = \int_{-\infty}^{\infty} p(h) \frac{h}{\sqrt{\pi}} e^{-h^2(x-\mu)^2} \, dx.
\]

His prior satisfies the equation, and Jeffreys proved its uniqueness by an argument that Fisher (1933) attacked. Instead of the Bayesian posterior for \( \sigma \) from a sample of size \( n \) and Jeffreys’ prior, Fisher proposed the fiducial distribution based on the \( \chi^2 \) distributed pivot \( (n-1)s^2/\sigma^2 \) considered in Section 3.5. Fisher’s fiducial distribution coincides with Jeffreys’ posterior, but they differ in their logical content, to cite Fisher: “the fiducial probability refers to the unconditional random sampling from the normal distribution, while the Bayesian posterior is conditional on the observed sample”.

The statistic \( T \) that Fisher had in mind in 1930 was the maximum likelihood estimator for \( \theta \). The fiducial distribution would then be unique in the simple one-parametric models possessing a sufficient statistic \( T \), stochastically monotone in \( \theta \). In Fisher (1935) the use of sufficient statistics is emphasised. The standard t-pivot to be used on a normal sample could be modified to \( T = (\bar{X} - \mu) / \sigma / \tilde{\sigma} \) by replacing the empirical standard deviation by another estimate \( \tilde{\sigma} \), say the mean absolute deviation. This leads to a different distribution that Fisher
The fiducial argument
did not call fiducial because it does not make full use of the data. In previous papers Fisher had only worked out one-dimensional fiducial distributions. But in 1935 he considered an observed normal sample of size \(n\) with mean \(\bar{x}\) and standard deviation \(s\), and asked what is the joint predictive fiducial distribution for estimates \(
abla\hat{\mu}\) and \(\hat{\sigma}\) based on a future sample of size \(m\) from the same normal population, added to the existing data. He considered the two pivots \(T = (\bar{x} - \hat{\mu})/\hat{\sigma}\) and \(Z = s/\hat{\sigma}\) where \(\hat{\sigma}_{n+m}\) is the combined estimate of \(\sigma\). Fisher knew the joint distribution of \(T\) and \(Z\), and obtained from this a joint predictive fiducial distribution. By letting \(m \to \infty\) a joint fiducial distribution for \(\mu\) and \(\sigma\) is obtained. By integrating out \(\mu\) the marginal fiducial density for \(\sigma\) is obtained. It is identical to what the pivot \(s/\sigma\) yields. Correspondingly the conditional fiducial distribution for \(\mu\) given \(\sigma\) is the \(N(\bar{x}, \sigma^2/n)\), as expected.

Fisher was encouraged by this finding and indicated that simultaneous fiducial distributions can be found from the joint distribution of pivots for the individual parameters, “Hence the generality of the method in univariate cases” (Fisher, 1935, p. 395). With this supposed insight Fisher proceeded to develop a fiducial distribution for the difference in expectation, \(\hat{\delta} = \mu_1 - \mu_2\) for two normal populations possibly with different variances. The data are here one sample from each population. This is the so-called Behrens–Fisher problem. Fisher’s fiducial distribution for this problem soon came under attack.

Fisher never developed a formal theory of fiducial probability, except for his introduction of the concept for one-dimensional parameters in 1930. His style was to present and develop examples. These are mostly concerned with mean and variance parameters in samples from one or two normal distributions. The most famous, and the most discussed, is his solution to the Behrens–Fisher problem. Along with his fiducial distributions for the correlation coefficient in a bivariate normal distribution, he also found fiducial distributions for the parameters of the multinormal distribution. Savage (1976) gives references to each of these examples.

6.2 The controversy
Bartlett (1936) was the first to point out problems with Fisher’s fiducial argument. He found that in the Behrens–Fisher problem the marginal fiducial distribution for the difference in expectations \(\hat{\delta} = \mu_1 - \mu_2\) leads to tests with inappropriate significance levels. The fiducial distribution for \(\hat{\delta}\) found by Fisher (1935) is the convolution \(\int f_\delta(m) f_\mu(\hat{\delta} + m) dm\) where \(f_\mu\) is the t-fiducial distribution for \(\mu\). The problem identified by Bartlett is that there is no pivot yielding this distribution. There is simply no family of sets \(D_\delta(\alpha)\) in the outcome space of the sampling experiment such that \(F(d|S) \leq \alpha \iff S \in D_\delta(\alpha)\) and \(P_\theta(D_\delta(\alpha)) = \alpha\) for all parameter vectors with \(\delta\) fixed at \(\delta = d\) (Linnik, 1963). The joint fiducial density \(f_\mu(\mu_1) f_\mu(\mu_2)\) represents a genuine fiducial distribution over regions of a particular form in the two-dimensional plane in the sense that two-dimensional intervals formed by crossing two one-dimensional confidence intervals are confidence intervals. But deriving fiducial distributions for transformed parameters \(\theta = \theta(\mu_1, \mu_2)\) by ordinary calculus will in general go wrong in the sense that there is no corresponding pivot.

Example 6.1 Derived distributions need not be fiducial, even in one dimension
The problem just pointed to occurs not only in higher dimensions. Consider the following simple example. There is one observation from \(N(\mu, 1)\), say \(y\). From the fiducial normal
distribution with density \( f(\mu) = \phi(\mu - y) \) the density \( g(\theta) = \phi(\theta - y) + \phi(-\theta - y) \) for \( \theta = |\mu| \) is obtained. If a pivot for \( \theta \) exists it must be based on \( V = |Y| \) because the sign of \( Y \) is ancillary for \( \theta \) and \( V \) is exhaustive, that is, \( (V, \text{sgn}(Y)) \) is sufficient. The distribution function for \( V, F(v, \theta) = \Phi(v - \theta) - \Phi(-v - \theta) \), is decreasing from \( \Phi(v) - \Phi(-v) \) to zero. This is the unique pivot based on \( V \). Because the range is shorter than the unit interval, \( -\partial F(v, \theta)/\partial \theta = \phi(v - \theta) - \phi(-v - \theta) \) is not a proper fiducial density. More importantly, it differs from \( g \) obtained from the fiducial density for \( \mu \). With \( C(\theta) = 1 - \Phi(v - \theta) + \Phi(-v - \theta) \) and \( C(0) = 0 \) we have a confidence distribution with point mass at zero. See also Exercise 6.2.

Fisher might not have been too bothered by these problems. From 1935 and onwards he lost interest in the confidence interpretation of his fiducial probability. He understood it rather as an epistemic probability of the parameter obtained by usual probability calculus from what he thought were unique joint fiducial distributions obtained by his step-by-step construction of conditionally unique pivots.

**Example 6.2 The step-by-step method**

Fisher’s step-by-step prescription was modelled on his derivation of the t-fiducial for \( \mu \) based on a sample from the \( N(\mu, \sigma^2) \), mentioned previously. That is, because the empirical variance \( s^2 \) is sufficient for \( \sigma^2 \) and ancillary for \( \mu \), while the mean \( \bar{x} \) is sufficient and exhaustive (\( (\bar{x}, s^2) \) is sufficient), the first step is to obtain a fiducial distribution for \( \sigma \) based on the \( \chi^2 \) pivot, and in the second step to obtain a conditional fiducial distribution for \( \mu \) given \( \sigma \) from the normal pivot. This gives a joint fiducial distribution that Fisher integrated to obtain his marginal fiducial distributions. Fisher claimed that this method, if applicable, also for more steps and in other models than the normal, leads to a unique joint fiducial distributions.

Another route than the preceding to the fiducial distribution for \( \delta = \mu_1 - \mu_2 \) is as follows. A proper fiducial distribution \( f(\tau | s_1, s_2) \) is available for the ratio of the two scale parameters \( \tau = s_1 / s_2 \) by the F-distribution. The ratio of the empirical variances is ancillary for the mean parameters. If \( \tau \) were known a proper fiducial distribution \( f(\tau | S, \tau) \) based on the sufficient statistic \( S (\bar{x}_i, s_i \) for \( i = 1, 2 \) from the two normal samples would be available. Fisher’s solution by the step-by-step method is \( f(\delta | S) = \int f(\delta | S, \tau) f(\tau | s_1, s_2, \delta) \, d\tau \).

Bartlett (1939) attacked again, and claimed that this joint fiducial distribution for \( (\mu, \sigma) \) from a sample from a normal distribution does not lead to a valid fiducial distribution for \( \theta = \mu + \sigma \). He noted correctly that Fisher had not proven that joint fiducial distributions could be handled by ordinary probability calculus to obtain the fiducial distribution for derived parameters. Fisher (1941) had demonstrated that the method works in a few cases, and apparently thought this sufficient to be sure of its general validity. Fisher then struck back at Bartlett, and showed that his fiducial distribution for \( \theta \) indeed is valid in the original sense of confidence. Bartlett was unlucky in his choice of example, and Fisher was lucky. Pedersen (1978) found that actually the only type of univariate parameter for which integration of the joint fiducial leads to a valid marginal fiducial distribution are linear parameters \( a\mu + b\sigma \). For all other derived parameters Fisher’s construction leads to intervals with wrong coverage probabilities.

Does Fisher’s step-by-step method (Fisher, 1956, pp. 159–162, 169–171) lead to unique joint fiducial distributions? There is often more than one step-by-step path to the joint distribution. In the bivariate case with observations \( x \) and \( y \) and with parameters...
The fiducial argument

α and β, Fisher’s requirement is that the joint density factors as follows: $f(x,y,α,β) = f(x,α)f(y|x,α,β)$. Brillinger (1962) considers the slightly artificial model with joint distribution function

$$F(x,y) = (1 - e^{-αx})(1 - e^{-βy})(1 - γ e^{-φx-μy})$$

where $φ = α(1 + e^{-α}), μ = β(1 + e^{-β})$ and $|γ| ≤ \frac{1}{4}$ is a given constant. The marginal distributions are respectively $F(x) = 1 - e^{-αx}$ and $F(y) = 1 - e^{-βy}$. Starting the step-by-step constructions with $x$, the fiducial density for $α$ is $xe^{αx}$. The conditional distribution of $y$ given $x$ has distribution function

$$F(y|x) = (1 - e^{-βy})[1 - γ e^{-φx-μy}[1 - φ(e^{αx} - 1)/α]].$$

By differentiation with respect to $β$ the conditional fiducial density for $β$ is obtained. Because $y$ given $x$ is exhaustive Fisher’s condition is satisfied and the joint fiducial density is obtained by multiplication. The resulting joint fiducial density is not symmetric in $(x,α)$ and $(y,β)$, to be expected from the symmetry in the model. This asymmetry reflects the fact that a different joint fiducial density results from the step-by-step construction starting with $y$. Brillinger notes that the variant of the step-by-step method based on the factorisation $f(x,y,α,β) = f_1(x,α,β)f_2(y|x,β)$, suggested by Quenoille (1958), is also possible in his example. This method does not guarantee uniqueness of the joint fiducial distribution any more than Fisher’s method.

Dempster (1963) found another and less artificial example in which two different step-by-step constructions are possible and where their resulting joint fiducial distributions differ. Consider Fisher’s original example with a normal sample of size $n$ with mean $\bar{y}$ and empirical variance $s^2$. Fisher started with a fiducial distribution on $σ$ based on the $χ^2$ pivot. One could also have started with a fiducial distribution on $θ = μ/σ$. The statistic $T = Y/S$ has actually a noncentral t-distribution that is monotone in $θ$. Given $T$ and $θ$, $S/σ$ has a distribution depending only on $T$ and $θ$. A conditional fiducial distribution is thus available from the pivot $S/σ$, and a joint fiducial distribution for $(θ,σ)$ is obtained. By transforming to parameter $(μ,σ)$ and integrating out $σ$ the marginal fiducial distribution on $μ$ is also in this case found to be a t-distribution, but now with $n - 2$ rather than $n - 1$ degrees of freedom.

For a sample from the $p$-variate normal distribution $N_p(μ,Σ)$, Fisher (1954) found a joint fiducial distribution for $μ$ and $Σ$ resulting in the natural t-distributions by integration for any linear parameter $a'μ$ for coefficient vectors $a$. For $p = 2$ and for the special case of the sample correlation being zero, Dempster (1963) was able to use the step-by-step method and found a joint fiducial distribution that differs from the one Fisher found. Thus the joint fiducial distribution depends on the step-by-step sequence chosen, and Fisher’s claim of uniqueness does not hold water.

Fisher went wrong on his fiducial distributions. His struggle to save the fiducial argument was, however, successful in the sense that concepts and methods of conditional inference were developed (Savage, 1976). Also, despite the shortcomings of joint fiducial distributions, they might be useful because their marginals often are good approximate confidence distributions; see, for example, the constructions for the meta-analysis setup in Section 13.3.
6.3 Paradoxes

Fisher died in 1962 and could not respond to the demonstrated nonuniqueness of his step-by-step method based on sufficiency/ancillarity and conditioning. It is possible that he would have argued that Brillinger’s example was artificial and that Dempster chose a cumbersome sequence of steps in a case where a natural sequence is at hand. How he would be able to defend his position against Dempster’s inconsistency result for multivariate normal samples is hard to say. Fisher’s position was to his very end that fiducial distributions are unique when found by a natural step-by-step method, and that ordinary probability calculus applies.

Example 6.3 The length problem (Example 1.5 continued)

In addition to nonuniqueness and inconsistencies, other paradoxical results were found. The following example is problematic not only for the fiducial method. Let $Y_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, n$ be independent. Each $\mu_i$ has the fiducial distribution $N(y_i, 1)$, and the joint fiducial density is obtained as the product of the marginal densities. What is then the fiducial distribution for $\theta = n^{-1} \sum_{i=1}^n \mu_i^2$? Let $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ and $\theta_{\text{fid}}$ be the stochastic variable representing the fiducial distribution. Integrating out the joint fiducial density, the marginal fiducial may be represented as $\theta_{\text{fid}} \sim (1/n) \chi^2_n(\theta)$. The moments of the noncentral $\chi^2$ are known, and we find the fiducial mean to be $E(\theta_{\text{fid}} | y_1, \ldots, y_n) = \bar{Y} + 1$ while the fiducial variance is $(4\bar{Y} + 2)/n$. The fiducial distribution is thus concentrated about $\bar{Y} + 1$, and more so the larger $n$ is. There is accordingly a considerable positive bias in the fiducial distribution, eventually with all mass at $\bar{Y} + 1$ as $n$ increases. Stein (1959) develops this so-called length problem in some more detail; see also Example 1.5.

Neyman’s confidence interval method might also lead to paradoxical results. Take the simplest possible example of an observation from $N(\mu, 1)$, say $y$. A one-sided confidence interval, open to the right, of degree 0.95, is required for $\theta = |\mu|$. In Example 6.1 we found $C(\theta) = 1 - \Phi(\{|y| - \theta\})$ as confidence distribution, with pointmass $C(0) = 2(1 - \Phi(|y|))$ at zero, and the left limit of the interval is the 5% confidence quantile. In case $\Phi(|y|) - \Phi(-|y|) \leq 0.95$ the interval is $(0, \infty)$. In that case the degree of confidence is only 95% of something we are certain about.

Confidence intervals should be constructed before observations are made. Their virtue is in controlling the coverage probability in repeated applications. Consider the following artificial example. Two independent observations are made from the discrete distribution $P(Y = \theta + i, \theta) = 1/3$ for $i = -1, 0, 1$. The interval $[Y_{(1)}, Y_{(2)}]$ (the observed range) covers $\theta$ with probability 7/9. This is the degree of confidence. With probability 2/9 the confidence interval comes out as $[\theta - 1, \theta + 1]$. This interval of length 2 has only 7/9, confidence, but for these data we are certain that $\theta = \bar{y}$. To have only 7/9 confidence in something we know for sure is silly. Fisher was reluctant to use his fiducial argument when the distribution is discrete. In the simple discrete model with only one observation $\theta - Y$ is a pivot with a uniform distribution over $-1, 0, 1$. The fiducial distribution is thus uniform over $y - 1, y, y + 1$. With two observations the range $R = Y_{(2)} - Y_{(1)}$ is ancillary. Conditioning on $R$ we get for $y = Y_{(1)}$ the preceding uniform three-point fiducial distribution for $R = 0$; for $R = 1$ the fiducial probabilities are $\frac{1}{2}$ for $y$ and $y + 1$, while the fiducial probability...
The fiducial argument

is 1 for \( \theta = y + 1 \) when \( R = 2 \). The same result is obtained from the joint distribution of \((Y_1, Y_2)\), without conditioning. The Bayesian posterior agrees with the fiducial distribution when the prior is flat. This and other examples are discussed by Kalbfleisch and Sprott (2006).

There are many parallel paradoxes. Ball et al. (2002), considered various models where this paradox occurs, among them a model for the epidemic spread of a disease. The epidemic starts with \( a = 1 \) infected and \( n \) susceptible individuals. The population is large, and individuals make contacts and are infectious independently and according to the same process, at least early on in the epidemic when the epidemic may be modelled as a branching process. When an individual is infected he or she will potentially infect other individuals over a period \( T_I \) with a known distribution. Given \( T_I \), the number of individuals infected by an infected individual is thus \( R_0 = n\beta E T_I \), the basic reproduction number for which confidence intervals are required. The temporal passage of the epidemic is often difficult to observe, but the total size \( T \) is observable. This is the total number infected people over the whole epidemic. When \( R_0 < 1 \) the epidemic fades away, while it has a nonzero probability of taking off when \( R_0 > 1 \). Ball et al. (2002) found approximate equitailed confidence intervals for \( R_0 \) based on \( T \) by inverting two-sided tests of \( H_0 : R_0 = r \) based on the likelihood function. The upper end of the confidence interval of (conservative) degree \( 1 - \alpha \) is found to be \( 2/\alpha \) unless \( T_{\text{obs}} \) is not close to 0 or \( n \). Such intermediate observations are highly unlikely for higher values of \( R_0 \), but they lead to unsatisfactory confidence intervals. It is also found that for some unlikely values of \( T_{\text{obs}} \) the resulting confidence interval at say degree 80% is empty. The parameter \( R_0 \) has a finite value, and to be 80% confident that it has no finite value is clearly paradoxical.

Empty confidence intervals occur in various situations, for example, when estimating the common mean from several independent normal samples. Let \( (\hat{\mu}_i - \mu)/S_i \) be independent t-pivots with \( v_i \) degrees of freedom, and let \( c_i \) be its \( 1 - \frac{1}{2}(1 - \alpha)^{1/k} \) quantile \( i = 1, \ldots, k \). Then each interval \( \hat{\mu}_i \pm S_i c_i \) contains \( \mu \) with probability \((1 - \alpha)^{1/k}\) and the intersection

\[
I_\alpha = \left[ \max_{1 \leq i \leq k} (\hat{\mu}_i - S_i c_i), \min_{1 \leq i \leq k} (\hat{\mu}_i + S_i c_i) \right]
\]

covers \( \mu \) with probability \( 1 - \alpha \). When for some \( i \), \( \hat{\mu}_i \) happens to be either large or small relative to the other estimates, and \( S_i \) happens to be small, the \( I_\alpha \) will be empty. Jordan and Krishnamoorthy (1996) found this paradoxical result to occur with probability of about 0.02 when the degree of confidence is 0.95.

There are other types of unusual phenomena encountered via confidence intervals and distributions. These include situations with multimodality (where, say, a 95% confidence region might consist of three separate intervals) and limited confidence, as for the Fieller problem of ratios of normal means, discussed in Section 4.6. Neyman (1934) was happy with the fiducial distribution because it led to confidence intervals. For confidence intervals there are other paradoxes in addition to those discussed in the preceding. Many of these go away, however, when confidence distributions are considered, rather than confidence intervals at fixed degree.
6.4 Fiducial distributions and Bayesian posteriors

As noted, Jeffreys found that $h^{-1}$ could be used as a noninformative prior for the precision $h = \sigma^{-1}/\sqrt{2}$ in a normal model. For a sample from $N(0, \sigma^2)$ this leads to the fiducial distribution based on the $\chi^2$ pivot. Also, a flat prior on $\mu$ leads to the normal fiducial distribution ($\sigma$ known). In the first case the sufficient statistic $S^2 = \sum Y_i^2$ satisfies a location model on the log scale, that is, $\log S^2 - \log \sigma^2$ has a log chi-squared distribution. In the second case we have a location model on the original scale. In both cases the posterior distribution based on the Jeffreys prior agrees with the fiducial distribution.

The general location scale model was considered by Fisher (1934) and by Pitman (1939). For $Y_1, \ldots, Y_n$ independent and continuously distributed with density $\sigma^{-1}f(\sigma^{-1}(y - \mu))$ where $f$ is known, Pitman used invariance to obtain the fiducial density for $(\mu, \sigma)$,

$$
\pi_n(\mu, \sigma) \propto \frac{1}{\sigma} \prod_{i=1}^{n} \left[ \frac{1}{\sigma} f\left( \frac{y_i - \mu}{\sigma} \right) \right].
$$

This is the Bayesian posterior based on the Jeffreys improper prior density $p(\mu, \sigma) = 1/\sigma$. The meaning of this joint fiducial distribution is that it allows confidence regions for the Jeffreys prior is identical to the fiducial distribution. Let the sufficient statistic $T$ could be transformed to a one-dimensional location model the Bayesian posterior based on all measurable sets.

The configuration statistic $A = (Y_{(2)} - Y_{(1)}, \ldots, Y_{(n)} - Y_{(n-1)})$ based on the order statistic is ancillary for $\theta$. Let $T$ be some measure of location, say $Y_1$, such that $(T, A)$ is sufficient. Then the conditional density for $T$ given $A$ must be proportional to the likelihood $L(\theta) = \prod f(y_i - \theta)$; see Barndorff-Nielsen and Cox (1994, p. 36). This conditional density must have the location form $g(t - \theta|a)$ with $g$ determined by $f$, making $T - \theta$ a conditional pivot given the ancillary statistic. Thus $L(\theta)/\int L(\theta')\,d\theta'$ is the fiducial density. This agrees with the Bayesian posterior with a flat prior. Fiducial distributions are invariant to monotone transformations. This is also the case for Bayesian posteriors when the prior is correspondingly transformed, that is, a Jeffreys prior is used. The result of Lindley is thus proved.

Lindley (1958) considered models with only one parameter. He proved that if the model could be transformed to a one-dimensional location model the Bayesian posterior based on the Jeffreys prior is identical to the fiducial distribution. Let the sufficient statistic $T$ have distribution function $F(t - \theta)$. The posterior based on the constant Jeffreys prior is then the density $f(t - \theta)$, and the fiducial density is also $-\partial F(t - \theta)/\partial \theta = f(t - \theta)$. This result is also obtained when no one-dimensional sufficient statistic exists. Let $Y_i$ be i.i.d. with density $f(y - \theta)$, known up to location. The configuration statistic $A = (Y_{(2)} - Y_{(1)}, \ldots, Y_{(n)} - Y_{(n-1)})$ based on the order statistic is ancillary for $\theta$. Let $T$ be some measure of location, say $Y_1$, such that $(T, A)$ is sufficient. Then the conditional density for $T$ given $A$ must be proportional to the likelihood $L(\theta) = \prod f(y_i - \theta)$; see Barndorff-Nielsen and Cox (1994, p. 36). This conditional density must have the location form $g(t - \theta|a)$ with $g$ determined by $f$, making $T - \theta$ a conditional pivot given the ancillary statistic. Thus $L(\theta)/\int L(\theta')\,d\theta'$ is the fiducial density. This agrees with the Bayesian posterior with a flat prior. Fiducial distributions are invariant to monotone transformations. This is also the case for Bayesian posteriors when the prior is correspondingly transformed, that is, a Jeffreys prior is used. The result of Lindley is thus proved.

Lindley also proved the converse. We will rephrase his result slightly: If case $A$ is ancillary for the scalar parameter $\theta$ and the scalar statistic $T$ is exhaustive, making $(A, T)$ sufficient, and the conditional distribution for $T \mid A$ yields the fiducial distribution with density $-\partial F_{T \mid A}(t, \theta, a)/\partial \theta$, then there is a Bayesian posterior that agrees with the fiducial distribution if and only if the conditional distribution for $T \mid A$ may be transformed to a location model $F_{T \mid A}(t, \theta, a) = F_{T \mid A}(R(t) - \rho(\theta), a)$ for some monotone transformations $R$ and $\rho$. In this case the prior behind the Bayesian posterior will be the Jeffreys prior. This is proved by noting that the likelihood factors, as $L(\theta)\,\text{data} = h(a)f_{T \mid A}(t, \theta, a)$. For a prior density $p$ the posterior density will be based on $p(\theta)f_{T \mid A}(t, \theta, a)/\int p(\theta')f_{T \mid A}(t, \theta', a)\,d\theta'$. 

"Schweder-Book" — 2015/10/21 — 17:46 — page 193 — #213
194

The fiducial argument

Both the Bayesian posterior distribution and the fiducial distribution are thus based on
the conditional distribution given \( a \), and we may therefore simplify notation and write
\[ f(t, \theta) = f_{T|A}(t, v, a) \] correspondingly for the conditional distribution function. If \( p \)
is a prior making the fiducial density agree with the Bayesian posterior we have the partial
differential equation

\[
- \frac{\partial}{\partial \theta} F(t, \theta) = \frac{p(\theta)}{r(t)} \frac{\partial}{\partial t} F(t, \theta)
\]

where \( r(t) \) is the normalising constant \( \int p(v)f(t, v) \text{d}v \). A solution of this equation must
have the form \( F(t, \theta) = H(R(t) - P(\theta)) \) for some function \( H \), where \( p(\theta) = \partial P(\theta)/\partial \theta \)
and \( \partial R(t)/\partial t \). For this solution to be a distribution function being monotone in both \( t \) and \( \theta \), \( H \)
must be a distribution function and \( R \) and \( P \) must both be monotone. The conditional model
can thus be transformed to a location model, and the prior must be of Jeffreys’ type. The
converse follows directly from \( F(t, \theta) = H(R(t) - P(\theta)) \) by differentiation.

Brillinger (1962) demonstrates by an example that Lindley’s result cannot be extended to
\( \theta \) being two-dimensional.

6.5 Coherence by restricting the range: Invariance or irrelevance?

Fisher made an enormous impact on statistics in the previous century. His ‘ideas and
nomenclature’ made statistics a science and changed the thinking regarding both statistical
theory and practice. Despite the well-documented and serious shortcomings of what Fisher
sketched as a theory of fiducial probability, some attempts have been made in developing
what Fisher regarded as the crown of his achievement into a coherent theory. Most
statisticians came to the conclusion that the fiducial argument simply was wrong, and turned
their backs on that part of Fisher’s work. But some picked up the challenge posed by its
demonstrated weaknesses, perhaps inspired by the depths of the challenge and by admiration
of Fisher’s imagination and great contributions to statistics in general. We will have a look
at three distinct lines of arguments aimed at restricting or specifying the fiducial argument
to obtain a coherent theory void of contradictions.

The first line of argument was drawn up by Pitman (1939). He developed the joint
fiducial distribution for the location and scale parameters \( \mu \) and \( \sigma \) from a random sample
\( x = (x_1, \ldots, x_n) \) with density known up to location and scale; see earlier. He found that
the regions of the half-plane that could be assigned fiducial probability is a much smaller
class than the Borel sets. Pitman relied on the original meaning of fiducial probability
that Fisher entertained in the first five years of the fiducial argument, namely that the
fiducial distribution should be a confidence distribution over a specified family of regions.
Pitman (1957) returned to this when commenting on Fisher (1956). For a region \( D \)
to be a confidence region Pitman requires \( D \) to be bounded by curves whose equations are
of the form \( \psi((x - \mu)/\sigma) \) for functions \( \psi \) that are continuous almost everywhere in
the \( n \)-dimensional space. Pitman (1957) notes that the two-dimensional fiducial density
cannot be integrated freely to find valid marginal distributions. The regions of integration
must satisfy the restrictions he specified. Hampel (2006) distinguishes between aleatory
probabilities and epistemic probabilities, and holds that the fiducial argument in Pitman’s
wrapping is the road to epistemic probabilities.
6.5 Coherence by restricting the range: Invariance or irrelevance?

Instead of concentrating on where the fiducial argument goes wrong Fraser (1961a) asked why it goes right in many cases. He observed as Pitman (1939) did that invariance is a key to understand fiducial probability, and he kept to location-scale models. Consider the one-sample normal problem reduced to the sufficient statistic \((\bar{x}, s)\). In the terminology of Fraser (1961a), \([\bar{x}, s]\) is the transformation in sample domain that brings the reference point \((0, 1)\) into the sample point \((\bar{x}, s) = [\bar{x}, s](0, 1)\). A parallel group of location-scale transformations exists in the parameter domain. If \((\bar{x}, s)\) has parameter \((\mu, \sigma)\) then the transformation \([\mu, \sigma]^{-1}\) brings the parameter to the reference point \((0, 1)\). The sample space, the parameter space and the two spaces of transformations are isometric; they are simply the two-dimensional half-plane. The two groups of location-scale transformations operate in the same way and can be joined,

\[
\begin{bmatrix} \mu \\ \sigma \end{bmatrix}^{-1} [\bar{x}, s] = \begin{bmatrix} -\frac{x}{\sigma} \\ 1 \end{bmatrix} [\bar{x}, s] = \begin{bmatrix} \bar{x} - \mu \\ s \sigma \end{bmatrix}.
\]

This transformation is regarded to act on sample points and brings the reference point \((0, 1)\) into the point \(( (\bar{x} - \mu)/\sigma, s/\sigma)\), which is recognised as the two-dimensional pivot. By a simple argument Fraser (1961a) demonstrates that this pivot is the only pivot that is invariant to the group of location-scale transformations in the normal model; see also Fraser (1961b). Fraser (1966) terms such fiducial distributions obtained from invariant pivots structural distributions, and develops his theory further. The structural (or fiducial) distribution with density \(s(\mu, \sigma)\) in the normal model is what is called consistent with respect to conditioning. That is, with \(V = \chi^2/(n - 1)\) and \(Z\) random variables carrying the scaled chi-distribution and the standard normal distribution respectively, the structural distribution is expressed as

\[
\begin{aligned}
\mu &= \bar{x} + \frac{Z}{\sqrt{n}} \\
\sigma &= \frac{s}{\sqrt{n}}.
\end{aligned}
\]

For a given sample \(V = s/\sigma\) will be determined whenever \(\sigma\) is known. This knowledge is brought to the distribution for the location parameter which becomes \(\mu = \bar{x} + Z\sigma/\sqrt{n}\), which is understood as the conditional structural distribution given \(\sigma\). This is identical to the structural distribution for \(\mu\) based on the group of location transformations in case \(\sigma\) is known. The same consistency is seen when conditioning on \(\mu\). This consistency in conditioning is generalised to cases with a transformation group \(H\) and a subgroup \(G\) acting on parameter points \(\theta\) and sample points \(x\). The theory of structural probability was extended in Fraser (1968) to normal and nonnormal linear models. In the general location-scale model the structural distribution is found as the fiducial distribution found by Pitman (1939), which is the Bayesian posterior with the Jeffreys prior. The interpretation of structural probability is not linked to coverage probability. It is simply the distribution in sample space that is carried over to the parameter space by transformation. The structural distribution can be obtained only when the data space and the parameter space somehow can be regarded as isometric, and there is a group of transformations on this space.

By around 1965 most statisticians had come to the conclusion that the fiducial argument is fatally flawed. The frequentists rejected the notion of a probability distribution on a parameter that by definition is fixed and not subject to a random experiment, and the Bayesians had no use for fiducial distributions; they had enough with their priors and posteriors. In addition, Fisher’s claims of uniqueness had been found not to hold water, and
his claim that ordinary probability calculus applies to fiducial distributions had been found to lead to paradoxes and inconsistencies. In 1965 Bartlett suggested that Fisher’s theory of fiducial probability should be abandoned. In reviewing Hacking’s, *Logic of Statistical Inference* (1965), Bartlett (1966, p. 632) says, however

...his attempt at a rationale for this concept is sufficiently original for me to refer to the resulting theory as the Fisher–Hacking theory of fiducial inference, and consider whether it leads me to modify my recent suggestion (Bartlett, 1965) that the Fisher theory be abandoned.

Ian Hacking is a philosopher and logician with an interest in the history and theory of probability and statistics. Is the fiducial argument saved by a nonstatistician?

Hacking has two principles. The frequency principle is that the support for a proposition should be the probability of the proposition being true, when nothing more bearing on the proposition than this probability is known. This “seems so universally to be accepted that it is hardly ever stated” (ibid., p. 135). The other principle is one of irrelevance. Irrelevaxant information should not alter the support for a proposition. Take the utterly simple case of one observation $X = \theta + V$ where the ‘measurement error’ $V$ is uniformly distributed over the three points $-1, 0, +1$. By the frequency principle all three values of $V$ are equally supported. The fact that $X = x$ is observed is irrelevant for $V$. Therefore, the support distribution for $\theta$ is $\frac{1}{3}$ on each of the tree points $x - 1, x, x + 1$. This is the fiducial distribution. In the one-dimensional case of an observation $X$ having continuous distribution $F(x, \theta)$ such that $U = F(X, \theta)$ is a pivot uniformly distributed over the unit interval, Hacking specifies irrelevance to mean that the likelihood function is ‘uniquely determined’ by the model (p. 155). That is, there are functions $g$ and $h$ such that $dF(x, \theta)/dx = h(F(x, \theta))g(x)$. All solutions of this equation are of the form $F(x, \theta) = H(R(x) - P(\theta))$ where $H$ is a distribution function and $R$ and $P$ are both monotone. It is thus only in situations that can be transformed to the location model that the principle of irrelevance is satisfied.

Lindley (1958) found that this is precisely the class of models for which there is a Bayesian prior that yields as posterior the same distribution as the fiducial distribution. This distribution must thus obey ordinary probability calculus, and Hacking concludes that at least in the one-dimensional case the fiducial argument, restricted by the principle of irrelevance, leads to a unique distribution, and is free of inconsistencies. Hacking (1965, p. 154) conjectures that the same is true for the many-parameter case “but this has not yet been proved”. Something could be said, however. One-parametric fiducial distributions respecting the irrelevance condition based on mutually independent data can be combined by multiplication. The joint fiducial distribution will then respect the principle of irrelevance, and should obey ordinary probability calculus. When this joint fiducial distribution is the joint normal considered in Example 6.3, the fiducial distribution for $\|\mu\|^2 = \sum_{i=1}^{p} \mu_i^2$ obtained by integrating the joint distribution leads to a badly biased marginal distribution. Fisher suggested the step-by-step method of building up a joint fiducial distribution, say for $\theta = (\theta_1, \theta_2)$. If now $\theta_1$ is based on $A$ which is ancillary for $\theta_2$, and $\theta_2$ in turn is based on $T$ conditional on $A = a$, the principle of irrelevance will typically not apply because $a$ is likelihood-informative of $\theta_2$. The step-by-step method would, however, work if, after transformations possibly depending on both $\theta_1$ and $a$, the pivot has the location form, $F_{T|A}(t | a, \theta) = H(R(t, a, \theta) - P(\theta_2, a, \theta))$ and the principle of irrelevance applies.
6.6 Generalised fiducial inference

Harris and Harding (1984) found that the inconsistency discussed by Dempster (1963) in the one-sample normal case where the alternative first step is to use the noncentral t-pivot for $\mu/\sigma$ is resolved by the principle of irrelevance. The noncentral t-pivot cannot in fact be transformed to location form, and Fisher’s step-by-step method seems natural and unique. The example of two dependent exponential variates of Brillinger (1962) is also resolved. The two observations $X$ and $Y$ are symmetric in the distribution, and Brillinger found that the result of the step-by-step method depends on whether $X$ or $Y$ should be used in the first step. The conditional distribution $X|Y$, however, cannot be transformed to a location model, and the second step is thus not possible when the principle of irrelevance rules. Sharma (1980) gives an example of a one-parameter discrete model with the data not affecting the (relative) likelihood of two different pivots, and thus being irrelevant to both, but where they yield different fiducial distributions. A complete resolution of Hacking’s conjecture is still not available, as far as we know.

6.6 Generalised fiducial inference

Recall first the original fiducial argument of Fisher (1930). When $T$ has distribution function $F(t, \theta)$ then $F(T, \theta) \sim U$, that is, a uniformly distributed pivotal quantity. The pivot could equally well have been formulated as $T = G(\theta, U) = F^{-1}(U, \theta)$, where $F^{-1}$ is the quantile function for $T$. For each $u \in (0,1)$ and each possible observation, $t \in \mathcal{T}$, the pivot $G$ has a unique inverse $Q(t, u)$ solving the equation $t = G(Q(t, u), u)$. This is true for any classical pivot $G$ when $T$ and $U$ have dimension 1. Then $\theta^* = Q(t, U)$ is the stochastic variable carrying the fiducial distribution, which also is the confidence distribution with distribution function $C(\theta) = 1 - F(t, \theta)$ when $T$ is stochastic increasing in $\theta$. Here $Q$ is the fiducial quantile function for $\theta$ as discussed by Fisher (1930).

Hannig (2009) generalises the fiducial argument to cases where the pivotal equations might have multiple solutions. The pivot $T = G(\theta, U)$ reflects the structure of the data-generating process, and is called the structural equation. The inverse $Q$ will now be set-valued. We will call it the solution set,

$$Q(t, u) = \{\theta: t = G(\theta, u), t \in \mathcal{T}, u \in \mathcal{U}\},$$

where $\mathcal{T}$ is the outcome space for $T$ and $\mathcal{U}$ is that for $U$.

To handle the difficulties associated with a solution set having more than one member, Hannig (2009) introduces a randomising variable $V(Q(t, u))$ on $Q(t, u)$, selecting one member randomly according to its distribution. For each solution set with multiple members, a randomising distribution must be chosen. The idea is that $V(Q(t, U))$ shall carry the fiducial distribution for the observed data $t$. Since $t$ indeed has resulted from the data-generating process, as modelled by the structural equations, we know that the stochastic solution set is nonempty, $Q(t, U) \neq \emptyset$.

**Definition 6.1 (Generalised fiducial distribution)** The generalised fiducial distribution (Hannig, 2009), here called the fiducial distribution, is the conditional distribution of

$$\theta^* = V(Q(t, U)) | \{Q(t, U) \neq \emptyset\).$$
As was demonstrated by example, Fisher’s fiducial distributions in higher dimensions are not necessarily unique. This is of course also the case for the more general fiducial distribution. We discuss this problem in the text that follows. But despite its nonuniqueness a fiducial distribution might have good frequentist properties, and might even be an exact or approximate confidence distribution. The fiducial argument might actually be seen as a tool for model selection and for obtaining confidence distributions.

Hannig and Lee (2009) considered wavelet regression. To select the set of nonvanishing coefficients to prevent overfitting they augmented the set of structural equations with structural equations not affecting the distribution of the data. They obtained a high-dimensional (generalised) fiducial distribution for the coefficients. From these complex fiducial distributions they obtained approximate confidence distributions for derived parameters. These distributions were found to have asymptotically correct coverage properties, and by simulations they were found to be about equally as good as the best competitors they considered. Wandler and Hannig (2012) also used a complex fiducial distribution to find a new and promising procedure of multiple comparison in the classical situation of \( k \) normal samples, some of which might come from the same population.

To illustrate how to construct a generalised fiducial distribution, consider the following two examples taken from Hannig (2009).

**Example 6.4 Two normal data points**

Suppose \( X_i \sim N(\mu, 1) \) for \( i = 1, 2 \) are independent. Without reducing to the minimal sufficient statistic \( X \), the structural equation is

\[
X = (X_1, X_2) = G(\mu, U) = (\mu + U_1, \mu + U_2)
\]

where \( U_i \) are independent standard normal variables. Let \( x = (x_1, x_2) \) be the observed data. The solution set \( Q(x, u) \) is empty when \( x_1 - x_2 \neq u_1 - u_2 \) and \( Q(x, u) = x_1 - u_1 = x_2 - u_2 \) when \( x_1 - x_2 = u_1 - u_2 \). The solution set is thus either empty or single valued, and there is no need for randomisation distributions. The fiducial distribution is represented by the stochastic variable \( x_1 - U_1 \mid \{x_1 - x_2 = U_1 - U_2\} \). The conditional distribution of \( U_1 \mid \{D = U_1 - U_2\} \) is normal with mean \( D/2 \) and variance \( \frac{1}{2} \). The generalised fiducial distribution is thus normal with mean \( x_1 - (x_1 - x_2)/2 \) = \( \bar{x} \) and variance \( \frac{1}{2} \).

**Example 6.5 Bernoulli trials**

Let \( X = (X_1, \ldots, X_n) \) be a set of Bernoulli trials with success probability \( p \). The goal is to find a fiducial distribution for \( p \). The sufficient statistic is \( S = \sum_{i=1}^{n} X_i \). Let the structural equation be formulated in \( S \) by way of the order statistic \( U_{(i)} \) for \( i = 1, \ldots, n \) of \( n \) independent uniformly distributed (over the unit interval) variables. With also \( U_{(0)} = 0 \) and \( U_{(n+1)} = 1 \), the structural equation is

\[
S = \sum_{s=0}^{n} sI\{U_{(s)} < p \leq U_{(s+1)}\}.
\]

The solution sets are simply \( Q(s, u) = (u_{(s)}, u_{(s+1)}) \). These solution sets are not single valued, and randomisation is necessary. That is, for each interval \( (u, v] \) in the unit interval a distribution with this interval as support must be chosen. Disregarding that the intervals are open to the left, let the distribution \( V(u, v] \) be \( u \) with probability \( \frac{1}{2} \) and \( v \) with probability
6.6 Generalised fiducial inference

The fiducial distribution is then that of $DU(\omega) + (1 - D)U(\omega + 1)$, where $D$ is an independent Bernoulli variable with success probability $\frac{1}{2}$. The density of the individual order statistics is

$$f_s(u) = \frac{n!}{(s-1)!(n-s)!} u^{s-1} (1 - u)^{n-s} \quad \text{for } 0 < u < 1.$$  

The fiducial density is consequently

$$f_p^* (p) = \frac{1}{2} f_s(p) + \frac{1}{2} f_{s+1}(p).$$

This fiducial distribution is identical to the confidence distribution obtained by half-correction; cf. Example 3.5. Its distribution function is $C(p) = 1 - B(s; n, p) + \frac{1}{2} b(s; n, p)$, where $B$ and $b$ are the binomial cumulative probability and point probability, respectively. Hannig (2009) also chose other randomising distributions $V(u, v)$: uniform over $(u, v)$, various beta distributions over $(u, v)$ and the midpoint of $(u, v)$. His extensive simulations indicate that the $f_p^*(p)$ found previously, or found by simple half-correction, is performing slightly better than the other choices. The coverage probability is actually closer to the nominal level for confidence intervals. This choice of $V(u, v)$ maximises the variance among the possible randomising distributions. This might be the reason for $f_p^*(p)$ to be superior.

Hannig (2009) developed a formula for the fiducial density in the general case of the solution set being either single valued or being empty. The density is of the form

$$f_p^*(\theta) = \frac{L(\theta, x) J(x, \theta)}{\int_\theta L(\theta', x) J(x, \theta') d\theta'}$$

where $L(\theta, x) = f_s(x, \theta)$ is the likelihood function. The weighting function $J$ depends on the structural equations in an explicit way. If $J(x, \theta) = g(\theta) h(x)$, the fiducial distribution is the Bayesian posterior under prior $g$. This happens only in special cases.

Fisher proposed to find his joint fiducial distribution by first reducing the data to a minimal sufficient statistic, and then to apply his step-by-step method, using a one-dimensional pivot in each step. Brillinger (1962), and also Dempster (1963), found that there could be several different step-by-step paths in the same model, as discussed earlier. When the sufficient statistic has the same dimension as the parameter vector, and the model admits recursive structural equations with one new parameter entering in each equation, the step-by-step method can be used. There might, however, be several different but equivalent recursive systems of structural equations in a problem. The fiducial distribution will typically depend on the system. This is one source of nonuniqueness of fiducial distributions in higher dimensions.

The (generalised) fiducial distribution is, however, nonunique also for other reasons. As was seen for the fiducial distribution for the binomial $p$, a choice of randomisation distribution $V(Q(x, u))$ had to be made when the solution set $Q(x, u)$ has multiple points. Different choices lead to different fiducial distributions.

The most disturbing source of nonuniqueness is illustrated in the simple Example 6.4, where the confidence distribution $\mu^* \sim N(\bar{x}, 1/2)$ was found for the mean $\mu$ based on a random sample of size 2 from $N(\mu, 1)$. If, however, the condition $Q(x, U) \neq \emptyset$ is formulated as $\exp(x_1 - U_1) = \exp(x_2 - U_2)$ rather than $x_1 - U_1 = x_2 - U_2$, standard
calculations lead to the fiducial distribution $N(\bar{x} - \frac{1}{2}, \frac{1}{2})$. The root of this nonuniqueness is that it was necessary to condition on the event $Q(x, U) \neq \emptyset$, which has probability zero. The conditional density, calculated by the traditional formula $f_{X \mid Y}(x \mid y) = f_{X, Y}(x, y) / f_Y(y)$, depends on how this event is formulated, and can actually be virtually everything. This source of nonuniqueness is known as Borel’s paradox; see Schweder and Hjort (1996) and Hannig (2009).

Another problem in fiducial inference is that marginalisation by integration of a joint fiducial distribution might not lead to a fiducial distribution of the derived parameter. That is, it might not be possible to find a set of reduced structural equation $X_r = G(\chi, U_r)$ that leads to the marginal distribution for $\chi = k(\theta)$ as its fiducial distribution. This problem was discussed in Section 6.3. And even when the joint fiducial distribution has good frequentist properties, the marginal distribution might be bad.

Despite the nonuniqueness of fiducial distributions, and also the problem with marginalisation, fiducial distributions might be highly useful. Hannig and co-workers used their fiducial machinery to develop statistical methods for which they investigated frequentist properties. Asymptotically they found the fiducial distributions to be correct confidence distributions to the first order in regular models. They also asked whether their fiducial distributions are approximate confidence distributions for limited data, and how good the approximation is. The methods they have derived in fairly complex models are demonstrated to be good, and it is to be expected that fiducial analysis will be useful also in the future in developing approximate confidence distributions.

6.7 Further remarks

Fisher saw the fiducial argument as the jewel in the crown of the “ideas and nomenclature” for which he was responsible (Fisher, 1956, p. 77). To him the fiducial distribution provides in precise and interpretable terms the information about the parameter that is obtained from the data in view of the model. Kalbfleisch and Sprott (2006) say that “the fiducial distribution of $\theta$ is not a distribution of belief, but a summary of the information about $\theta$ provided by the observations which one is free to believe or disbelieve”. Fisher would probably have agreed to the first part of the statement, but he would have added that the rational mind would then have the fiducial distribution as knowledge, and in that sense regard it as his or her distribution of belief. For one-parameter models this is also our view, but conditioned on the model which also might be surrounded by uncertainty. Partly to distance the concepts and methods presented in the present book from all the difficulties associated with fiducial distributions in higher dimensions we use the term confidence distribution. Confidence distributions are usually one-dimensional. With multidimensional parameters, marginalisations of a joint fiducial distribution, if available, are possible (approximate) confidence distributions. So are also marginals of Bayesian posteriors. Other ways to reduce dimensionality are by profiling the likelihood, conditioning or by other means, and to infer one confidence distribution for each parameter of interest at lower dimension, usually of dimension one. By analysis via bootstrapping or other means, we suggest that confidence distributions, or rather the statistics on which they are based, be corrected for bias; cf. Chapter 7.
6.8 Notes on the literature

If joint confidence distributions are required we suggest going a bit further than Pitman (1957) and specify in advance the class of regions, typically a nested family, to which confidence can be attached. These issues are discussed in Chapter 9. Joint fiducial distributions will serve the purpose when their restrictions are observed, the restrictions being the family of joint confidence sets for which they provide coverage. How to operate on fiducial distributions to obtain derived distributions that are fiducial is, however, an open problem.

Fisher’s aim was to identify the fiducial distribution, obtained from the data and given the model. The fiducial distribution should thus be unique, and it should be subject to the rules of ordinary probability calculus. He suggested various conditions for this to take place, but as seen via counterexamples reviewed earlier, Fisher failed to make his programme work in two or more dimensions. Our goals are different. Via methods developed in our other chapters we aim for a good confidence distribution when several may exist. We are also satisfied with interpreting the confidence distribution as a distribution of confidence over a special family of regions. This attitude will not necessarily bring universal agreement over what can be inferred from a given set of data, but hopefully it can be agreed that the inference obtained is reasonable and well argued, if not always unique. Also, in complex models where no exact confidence distribution stands out, such a distribution is acceptable when it has nearly correct null properties, that is, the implied confidence regions have approximately the right coverage probabilities.

6.8 Notes on the literature

The use of flat priors to represent lack of information came under attack in the mid-nineteenth century; see Boole (1854). Fisher was aware already from 1912 that a flat prior for \( \theta \) yields a nonflat prior for \( \psi = g(\theta) \) for a curved transformation \( g \); see Aldrich (2000), who presents the background of Fisher’s original fiducial theory. Fisher (1930) refers to graphs in Ezekiel (1930), and he understood apparently his original fiducial distributions in confidence distribution terms. Fiducial probability was put forward as an alternative to the problematic inverse probability, and was understood as epistemic probability by Fisher. The word ‘fiducial’ stems from the Latin fiducialis, related also to fiduci (trust) and to fidere (to trust).

Zabell (1992) gives a most readable presentation of Fisher and the fiducial argument. He cites Fisher’s central papers on fiducial probability and adds contextual information from Fisher’s correspondence. See also Lehmann (2011, chapter 6) for further discussion, highlighting the running conflicts with Neyman. The literature on fiducial probability is extensive. Most papers are from between 1935 and 1965, the time period when the controversy over fiducial probability boiled. Zabell’s paper gives a good guide to this literature. Many prominent statisticians have written about Fisher and his contributions, including fiducial probability, often in the format ‘On rereading Fisher’; see, for example, Savage (1976) and Efron (1998).

Fisher claimed his fiducial distribution to be ordinary probability subject to the common probability theory. But it was amply demonstrated that the distribution for a derived parameter of dimension 1, obtained by integrating a joint fiducial density, need not be a fiducial distribution. There might actually not exist any pivot behind it. In addition, joint
The fiducial argument

Fiducial distributions obtained by the step-by-step method are not unique even when based on sufficient statistics. Despite these important shortcomings Fisher defended his method to his very end. The fiducial argument, however, fell into disrespect and neglect. Davison (2001, p. 14) notes, “A few subsequent attempts have been made to resurrect fiducialism, but it now seems largely of historical importance, particularly in view of its restricted range of applicability when set alongside models of current interest.”

In this chapter we have focussed on the historical development of fiducial inference and what we may tentatively describe as the first and second waves of research work and interest in fiducial related themes – the first wave starting with Fisher (1930), followed by critical contributions from Neyman, Bartlett, Pitman and others, partly countered by Fisher again, and so on; and the second initiated by Fraser (1961a,b) leading to structural inference, and so on. We would argue that methodological statistics is experiencing a third fiducial wave, the ongoing general resurrection of Fisher’s fiducial theory to a new life. This is associated with and represented by Efron (1998), Schweder and Hjort (2002, 2013a), Hannig et al. (2006), Hannig (2009), Taraldsen and Lindqvist (2013), Xie and Singh (2013), Veroneze and Melilli (2015) and indeed other work reported on in other chapters in the present book, broadening the fiducial scope towards distributions of confidence and modernising the tools.

Exercises

6.1 Behrens–Fisher: In Scheffé’s opinion (Scheffé, 1970), the most frequent problem in applied statistics is to compare the means of two populations, most likely having different variances. When the two populations are normally distributed with independent estimates of means and variances the problem is named the Behrens–Fisher problem; see Section 6.2. The convolution of the joint fiducial distribution suggested by Fisher (1935) is known to be not a strictly correct fiducial distribution, but it might approximately be a confidence distribution. To investigate the quality of the approximation, construct a simulation to estimate the distribution of the proposed confidence distribution evaluated at the true difference $\delta = \mu_2 - \mu_1$, say, for sample sizes 5 and 10 and for a few values of the ratio $\lambda = \sigma_1/\sigma_2$ between $\frac{1}{2}$ and 2. You might do this by for each replicate drawing a sample from the fiducial distribution of each mean and thus a sample of $\delta_{\text{CD}}$. What is of interest is the frequency of $\delta_{\text{CD}} \leq \delta$. Note that the Bayesian will arrive at Fisher’s solution when priors are used to give t-distributed posteriors for each mean, and that evaluating how close the posterior for $\delta$ is to being a confidence distribution might be done in the same way.

6.2 The length problem in one dimension: There is one observation $y$ from the distribution with distribution function $F(y, \mu) = G(y - \mu)$. Find the fiducial distribution for $\mu$ and for $\theta = |\mu|$ when $G$ is known. Compare the fiducial distribution for $\theta$ to the distribution obtained from the pivot $\Phi(|y| - \theta) - \Phi(-|y| - \theta)$ in case $G = \Phi$. The latter yields a confidence distribution with a point mass at the origin. Let $\theta(\alpha, |y|)$ be the confidence $\alpha$-quantile. Show the confidence property $P_\theta(\theta \leq \theta(\alpha, |y|)) = \alpha$ despite the point mass in the confidence distribution. Pedersen (1978) considers symmetric intervals obtained from the fiducial distribution for $\mu$. For given $\alpha$ let the interval be $(-m, m)$ where $\alpha = \Phi(m - y) - \Phi(-m - y)$ to give fiducial probability $\alpha$. The solution to this equation is $m(y, \alpha)$. Pedersen shows that $P_\mu(-m(X, \alpha) \leq \mu \leq m(X, \alpha)) > \alpha$ for every $\mu$. The fiducial distribution is thus unable to provide symmetric confidence intervals for $\mu$. The fiducial distribution for $\theta$ is in general $F_{\text{fid}}(\theta, y) = G(y + \theta) - G(y - \theta)$. Show that $F_{\text{fid}}(\theta, Y)$ is not a uniform pivot, which agrees with Pedersen’s result.
6.3 **Step-by-step example:** Consider the simple normal regression model \( Y \sim N(\alpha + \beta x, \sigma^2) \). First let \( \sigma \) be known. Show that the joint fiducial distribution \( h(\alpha, \beta) \) found by the step-by-step method is the same whether the first step is for \( \alpha \) or for \( \beta \), and that it also is the same as that found from the joint linear normal pivot. What if \( \sigma \) is unknown?

6.4 **Step-by-step, counterexample:** Consider the joint exponential model of Brillinger (1962) mentioned in Section 6.2. Calculate the conditional fiducial density for \( h(\beta | \alpha) \) based on the conditional distribution \( Y | X \). Confirm that the joint fiducial density \( h(\beta | \alpha)h(\alpha) \) is asymmetric in its two arguments, showing that the step-by-step method would have led to a different joint fiducial distribution if the first step had been for \( \beta \).

6.5 **Conditional location model:** Let \( \theta \) be a scalar parameter of interest and \( \alpha \) a nuisance parameter, possibly a vector. The likelihood factors in \( A \) which is ancillary for \( \theta \) and \( T \) which is exhaustive so that \( (T, A) \) is sufficient and the likelihood is \( L(\theta, A | \text{data}) = h(\alpha, a)f_{T | A}(t, \theta, a) \) where both \( h \) and \( f \) are densities. For fixed \( t \) and \( a, \) the cumulative conditional \( F_{T | A}(t, \theta, a) \) is monotone and differentiable in \( \theta \). Give the fiducial distribution for \( \theta \). Compare this distribution to a Bayesian posterior based on a chosen prior. Will Lindley’s result hold in this case – that the fiducial distribution agrees with one of the Bayesian distributions if and only if the conditional distribution of \( T | A \) can be transformed to a location model, that is, \( F_{T | A}(t, \theta, a) = F_{T | A}(R(t) - \rho(\theta), a) \) for some monotone transformations \( R \) and \( \rho \)?

6.6 **Confidence distribution in the epidemic model:** For the epidemics model considered by Ball et al. (2002) let the time an individual is infectious be exponentially distributed with mean \( \gamma^{-1} \), during which time contacts are made with susceptible individuals as a Poisson process with intensity \( \beta n(t) \) where \( n(t) \) is the number of susceptible people at time \( t \). Use the following equations to calculate numerically the cumulative distribution for \( T \) for a range of values of \( R_0 = n/\rho \) where \( \rho = \gamma/\beta \) and \( n = 119 \). Use this to find a confidence distribution for \( T_{obs} = 29 \), which was observed in a local smallpox epidemic in Nigeria. Compare your result with what Ball et al. (2002) found, namely that at 80% confidence their confidence interval is empty. The distribution of \( T \) is found to be \( P(T = k) = \rho h_{k-1,1} \) where for

\[
E = \{(i,j): i = 1, \ldots, n; j = 2, \ldots, n+1; 1 \leq i + j \leq n + 1\}
\]

the \( h_{ij} \) for \((i,j) \in E\) satisfy

\[
(i + 1)h_{i+1,j-1} - (i + \rho)h_{ij} + \rho h_{i,j+1} = 0 \quad \text{for} \ (i,j) \in E \setminus (n, 1)
\]

and \( h_{ij} = 0 \), for \((i,j) \not\in E\).

6.7 **Borel’s paradox:** Consider Example 6.4, where the generalised fiducial distribution for \( \mu \), based on \( n = 2 \) observations from \( N(\mu, 1) \), was found to be \( N(\bar{x}, \frac{1}{2}) \). It was necessary to condition on the event \( Q(x, U) \neq \emptyset \). Instead of formulating this as \( x_1 - U_1 = x_2 - U_2 \), use the equivalent formulation \( Y_2 = \exp(x_1 - U_1) - \exp(x_2 - U_1) = 0 \). With \( Y_1 = x_1 - U_1 \), find the joint density of \((Y_1, Y_2)\), and show by standard calculations that the fiducial distribution now is \( N(\bar{x} - \frac{1}{2}, \frac{1}{2}) \).
Improved approximations for confidence distributions

Previous chapters have developed concepts and methodology pertaining to confidence distributions and related inference procedures. Some of these methods take the form of generally applicable recipes, via log-likelihood profiles, deviances and first-order large-sample approximations to the distribution of estimators of the focus estimands in question. Sometimes these recipes are too coarse and are in need of modification and perfection, however, which is the topic of the present chapter. We discuss methods based on mean and bias corrected deviance curves, t-bootstrapping, a certain acceleration and bias correction method, approximations via expansions, prepivoting and modified likelihood profiles. The extent to which these methods lead to improvements is also briefly illustrated and discussed.

7.1 Introduction

Uniformly most powerful exact inference is in the presence of nuisance parameters available only in regular exponential models for continuous data and other models with Neyman structure, as discussed and exemplified in Chapter 5. Exact confidence distributions exist in a wider class of models, but need not be canonical. The estimate of location based on the Wilcoxon statistic, for example, has an exact known distribution in the location model where only symmetry is assumed; see Section 11.4. In more complex models, the statistic on which to base the confidence distribution might be chosen on various grounds: the structure of the likelihood function, perceived robustness, asymptotic properties, computational feasibility, perspective and tradition of the study. In the given model, with finite data, it might be difficult to obtain an exact confidence distribution based on the chosen statistic. As we shall see there are various techniques available for obtaining approximate confidence distributions and confidence likelihoods, however, improving on the first-order ones worked with in Chapters 3–4.

Bootstrapping, simulation and asymptotics are useful tools in calculating approximate confidence distributions and in characterising their power properties. When an estimator, often the maximum likelihood estimator of the interest parameter, is used as the statistic on which the confidence distribution is based, bootstrapping provides an estimate of the sampling distribution of the statistic. This empirical sampling distribution can be turned into an approximate confidence distribution in several ways, which we address in the text that follows.

It is useful to start with the perhaps simplest and most widely used method, associated with the delta method for obtaining approximate confidence intervals; cf. results and
7.2 From first-order to second-order approximations

In a sample of size \( n \), suppose the estimator \( \hat{\theta} \) has an approximate multinormal distribution centred at \( \theta \) and with covariance matrix of the form \( \hat{\Sigma}/n \), so that

\[
\hat{\Sigma}^{-1/2} n(\hat{\theta} - \theta) \to_d N_p(0, I),
\]

with \( p \) the dimension of \( \theta \). Using the delta method, the confidence distribution for a parameter \( \psi = h(\theta) \) is based on linearising \( h \) at \( \hat{\theta} \), and yields

\[
C_{\delta}(\psi) = \Phi\left(\sqrt{n}(\psi - \hat{\psi})/\hat{\kappa}\right)
\] (7.1)

in terms of the cumulative standard normal. The variance estimate for \( \hat{\psi} \) is \( \hat{\kappa}^2/n \), with \( \hat{\kappa}^2 = \hat{g}^T \hat{\Sigma} \hat{g} \), where \( \hat{g} \) is the gradient of \( h \) evaluated at \( \hat{\theta} \). Again, this estimate of the confidence distribution is to be displayed post data with \( \psi \) equal to its observed value \( \psi_{\text{obs}} \).

This confidence distribution is known to be first order unbiased under weak conditions. That \( C_{\delta}(\psi) \) is first order unbiased means that the coverage probabilities converge at the rate \( 1/\sqrt{n} \), or that \( C_{\delta}(\psi_{\text{true}}) \) converges in distribution to the uniform distribution at the rate \( \sqrt{n} \). Note also that the confidence density as estimated via the delta method, say \( c_{\delta}(\psi) \), is simply the normal density \( N(\psi, \hat{\kappa}^2/n) \). The additivity of the asymptotically normal pivot implies that the confidence likelihood is Gaussian and actually identical to the confidence density \( c_{\delta}(\psi) \).

In earlier chapters we also used the deviance-based \( \text{cc}(\psi) = \Gamma_1(D_n(\psi)) \), and mentioned that this confidence curve tends to be more accurate than the delta method. The deviance method is, however, tied to the use of the maximum likelihood estimator in a given parametric model. It is to be noted that the delta method is considerably more general, and works as long as there is an adequate and approximately normal estimator of the focus parameter under study, also in semi- and nonparametric situations; cf. Chapter 11. More refined methods for obtaining confidence distributions are developed in the text that follows, and we give recipes both for parametric and nonparametric cases.

At the heart of the first-order large-sample calculus, extensively utilised in our earlier chapters, are Theorems 2.2 and 2.4 about the normal and chi-squared limits of the distributions of respectively the maximum likelihood estimator distribution and the deviance statistic. These again essentially rest on the quadratic approximation to the log-likelihood function near its optimum, viz. Lemma 2.1 and Remark 2.5. Section 7.2 examines these quadratic approximations more carefully, via expansions to the third and fourth order. The point here is partly to make it clear that several different approaches may be taken to modify and repair the two-term approximations. In this sense that section does not end with clear recipes, but helps setting the stage for further methods, which are then worked with in the following sections.

7.2 From first-order to second-order approximations

When examining and developing likelihood-based methods for inference and confidence distributions in earlier chapters, a key tool was the canonical random function

\[
A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0),
\] (7.2)

cf. the ‘master lemma’ Lemma 2.1 and Theorems 2.2 and 2.4 of Chapter 2. Its maximiser is

\[
Z_n = Z_n(\theta_0) = \arg\max(A_n) = \sqrt{n}(\hat{\theta} - \theta_0)
\]
Improved approximations for confidence distributions

and the deviance statistic is twice its maximum,\[ D_n = D_n(\theta_0) = 2 \max A_n = 2(\ell_n(\hat{\theta}) - \ell_n(\theta_0)). \]

We were able to derive very fruitful large-sample approximations to the distributional aspects of \( Z_n \) and \( D_n \) from the basic convergence property of \( A_n \):
\[ A_n(s) = s'U_n - \frac{1}{2}s'J_n s + r_n(s) \rightarrow_d A(s) = s'U - \frac{1}{2}s'Js, \]
where \( U_n = n^{-1/2}\ell'_n(\theta_0) \rightarrow_d U \sim N_p(0, J) \) and \( J_n = -n^{-1}\ell''_n(\theta_0) \rightarrow_d J \), writing again \( J \) for the information matrix \( J(\theta_0) \). Indeed, there is joint convergence in distribution
\[ (Z_n, D_n) = (\arg\max(A_n), 2 \max A_n) \rightarrow_d (\arg\max(A), 2 \max A) = (J^{-1}U, U'J^{-1}U). \]

In their turn, these results led to general recipes for confidence distributions, as discussed in Section 3.4.

To reach modified though necessarily more elaborate statements, also the next two terms in the Taylor expansion can be taken into account. Let us for simplicity of presentation concentrate on the one-dimensional case, where we may write
\[ A_n(s) = sU_n - \frac{1}{2}s'Js^2 + \frac{1}{6}(K_n/\sqrt{n})s^3 + \frac{1}{24}(L_n/n)s^4 + \text{smaller order terms}, \]
where
\[ K_n = n^{-1}\ell^{(3)}_n(\theta_0) = n^{-1}\sum_{i=1}^n \frac{\partial^3\log(f(Y_i, \theta_0))}{\partial \theta^3} \rightarrow \mu K, \]
\[ L_n = n^{-1}\ell^{(4)}_n(\theta_0) = n^{-1}\sum_{i=1}^n \frac{\partial^4\log(f(Y_i, \theta_0))}{\partial \theta^4} \rightarrow \mu L, \]
say, assumed finite. It is useful to think in terms of the first- and second-order approximations to \( A_n \), say
\[ A_{n,1}(s) = sU_n - \frac{1}{2}s'Js^2, \]
\[ A_{n,2}(s) = sU_n - \frac{1}{2}s'Js^2 + \frac{1}{6}(K_n/\sqrt{n})s^3 + \frac{1}{24}(L_n/n)s^4. \]
An illustration is provided in Figure 7.1; see Example 7.1.

The first-order approximations to respectively the maximiser and twice the maximum of \( A_n \) are
\[ \arg\max(A_{n,1}) = s_0 = U_n/J_n \quad \text{and} \quad 2 \max(A_{n,1}) = U_n^2/J_n; \quad (7.3) \]
these have as we know distributional limits \( U/J \sim N(0, J) \) and \( U^2/J \sim \chi^2_d \). To give an analysis precise also to the next level of approximation, consider \( A_n(s) \) evaluated for arguments \( s = s_0 + w/\sqrt{n} \) close to \( s_0 \). We find
\[ A_n(s_0 + w/\sqrt{n}) = s_0U_n - \frac{1}{2}s_0'Js_0^2 \]
\[ + n^{-1/2}(U_n w - Js_0 w + \frac{1}{6}K_n s_0^3) \]
\[ + n^{-1}(-\frac{1}{2}s_0'w^2 + \frac{1}{2}K_n s_0^2 w + \frac{1}{24}L_n s_0^4) + \ldots \]
7.2 From first-order to second-order approximations

Figure 7.1 The full curve is the $A_n$ random function of (7.2) based on a sample of size $n = 10$ from the exponential distribution with true parameter $\theta_0 = 3.00$, with observed $\bar{y}_{\text{obs}} = 0.259$. The first-order (quadratic) and second-order (quartic) approximations are shown as dashed and dotted lines, respectively. The maximisers are tagged (2.722 for $A_n$, corresponding to the maximum likelihood estimate on the $s = \sqrt{n}(\theta - \theta_0)$ scale; then 2.116 and 2.649 for the first- and second-order approximations). The maximum of the second-order approximation is 0.290, coming much closer to the maximum 0.293 of $A_n$ than does the maximum 0.249 of the first-order approximation.

Noting that the $U_n w - J_n s_0 w$ disappears, we find that this expression is maximised for $w$ equal to $w_0 = \frac{1}{2}(K_n/J_n^2)U_n^2$.

From this we learn first that

$$Z_n = \sqrt{n}(\hat{\theta} - \theta_0) = s_0 + w_0/\sqrt{n} = U_n/J_n + \frac{1}{2}(K_n/J_n^2)U_n^2/\sqrt{n},$$

and in particular that $Z_n \overset{d}{=} N(0,J) + O_p(n^{-1/2})$. It also leads to an informative bias expression of the type $E Z_n = b/\sqrt{n}$, with $b = \frac{1}{2}K/J^2$, which with mild further assumptions translates to a $O(1/n)$ bias expression for the maximum likelihood estimator,

$$E \hat{\theta} = \theta_0 + b/n + o(1/n).$$

(7.4)

Second, for twice the attained maximum value, we have

$$D_n \df U_n^2/J_n + \frac{1}{2}n^{-1/2}(K_n/J_n^2)U_n^3 + n^{-1} \left( \frac{1}{4}(K_n^2/J_n^4) + \frac{1}{12} (L_n/J_n^4) \right) U_n^4.$$  (7.5)

Here some algebraic efforts lead to

$$E U_n^2 = J, \quad E U_n^3 = n^{-1/2} \gamma_3, \quad E U_n^4 = 3J^2 + O(1/n),$$
Improved approximations for confidence distributions

in terms of the score skewness \( \gamma_3 = \text{Eu}(Y, \theta_0)^3 \). Hence we have an explicit correction to the mean of the deviance:

\[
ED_n = J/J_n + c/n,
\]

where \( c = \frac{1}{2}(K/J^3)\gamma_3 + \frac{1}{2}K^2/J^3 + \frac{1}{4}L/J^3 \). Note that when the parametric family used is of the exponential class type (see Chapter 8), one has \( J_n = J, K_n = K, L_n = L \), valid for each \( n \).

Example 7.1 Likelihood approximations for an exponential sample

Suppose \( Y_1, \ldots, Y_n \) are a sample from the exponential distribution with density \( f(y; \theta) = \theta \exp(-\theta y) \) and score function \( u(y; \theta) = 1/\theta - y \). Then the log-likelihood function is \( \ell_n(\theta) = n(\log \theta - \theta \bar{Y}) \), in terms of the data average \( \bar{Y} \), and

\[
A_n(s) = n[\log(\theta_0 + s/\sqrt{n}) - \log \theta_0] - \sqrt{n}s \bar{Y}.
\]

In the preceding notation concerning the third and fourth derivatives we have \( K_n = 2/\theta_0^3 \) and \( L_n = -6/\theta_0^5 \); cf. Exercise 7.1. Figure 7.1 displays \( A_n \) along with the first- and second-order approximations

\[
A_{n,1}(s) = \sqrt{n}(1/\theta_0 - \bar{Y}) - \frac{1}{2}s^2/\theta_0^3,
\]

\[
A_{n,2}(s) = \sqrt{n}(1/\theta_0 - \bar{Y}) - \frac{1}{2}s^2/\theta_0^3 + \frac{1}{3}s^3/(\theta_0^3 \sqrt{n}) - \frac{1}{4}s^4/(\theta_0^4 n),
\]

in a situation with \( \theta_0 = 3.00 \) and observed mean \( \bar{y}_{\text{obs}} = 0.259 \) over \( n = 10 \) observations. The second-order approximation works rather better than the first-order one, both in terms of argmax position (so \( \text{argmax}(A_{n,2}) \) is much closer to \( \sqrt{n}(\hat{\theta} - \theta_0) \) than is \( U_n/J_n \)) and in terms of achieved maxima (so \( 2\max(A_{n,2}) \) is much closer to the real deviance than is \( U_n^2/J_n \)).

In this situation \( \hat{\theta} = 1/\bar{Y} \) may be expressed as \( \theta_0/V_n \), where \( V_n \sim \chi_2^2/(2n) \); see Exercise 7.1. Hence

\[
E\hat{\theta} = \theta_0 \frac{2n}{2n - 2} = \theta_0 \frac{1}{1 - 1/n} \approx \theta_0(1 + 1/n),
\]

which fits with (7.4), where \( b = \frac{1}{2}K/J^3 \) here is found to be equal to \( \theta_0 \). Furthermore, the exact deviance may be expressed as \( D_n = 2n(V_n - 1 - \log V_n) \), with mean value

\[
ED_n = 2n[\log n - \psi(n)] = 1 + \frac{1}{b} + O(1/n^2),
\]

using a formula for the mean of a log-chi-squared and an expansion formula for the digamma function \( \psi(\cdot) \). This nicely matches result (7.6), since appropriate checking yields \( c = 1/6 \) in the present situation.

7.3 Pivot tuning

Outside a somewhat limited list of standard models and focus parameters one cannot expect there to be easily available and perfectly valid pivots \( \text{piv}(\psi, Y) \). Quite often one would, however, be able to argue that simple normalised ratios of the type \( t_n = (\bar{\psi} - \psi)/K \) ought to converge to a normal distribution. Such studentised ratios hence become pivots for large samples, after which the (3.13) method applies; the delta method version (7.1) would be a
7.3 Pivot tuning

further approximation for this scheme, using limiting normality rather than the finite-sample distribution for $t_n$. The convergence to normality with a variance independent of the position in the parameter space could be slow, however, and there is a need to develop fine-tuning methods. Stochastic simulation offers a venue for working with such questions.

The object is to employ simulation for approximating the confidence distribution associated with a given statistic $T$ related to the $\psi$ under focus. Along with such a $T$ one might start with an (approximate) ancillary statistic $A$, and simulate $(T, A)$ for a number of values of the interest parameter $\psi$ and the nuisance parameter $\chi$. The hope is that the conditional distribution of $T$ given $A$ is independent of the nuisance parameter, or very nearly so. Such questions can be addressed by applying regression methods to the simulated data. The regression might have the format

$$m(\psi) - T = \tau(\psi)V$$

(7.7)

where $V$ is a scaled residual. Then $\text{piv}(T, \psi) = [m(\psi) - T]/\tau(\psi)$ becomes a pivot, say with cumulative distribution function $G_n$ and density $g_n$, and the associated confidence distribution becomes

$$C_n(\psi) = G_n([m(\psi) - T]/\tau(\psi)).$$

(7.8)

Both the regression function $m(\psi)$ and scaling function $\tau(\psi)$ might depend on the ancillary statistic, and the extent to which this is the case may be sorted out by analysing simulation output.

**Example 7.2 The normal correlation coefficient**

Consider the usual correlation coefficient

$$r_n = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)^{1/2}}$$

(7.9)

based on independent binormal pairs $(X_i, Y_i)$. A reasonable approximation to its distribution, and hence to a confidence distribution of the type (7.1) for the underlying population correlation $\rho$, follows from a piece of large-sample theory which says that $\sqrt{n}(\hat{\rho} - \rho) \rightarrow_d N(0, \kappa^2)$, with a certain formula for $\kappa$. Suppose, however, that you do not know or recall this formula and that you do not have the time just now to deduce it mathematically (via the multidimensional central limit theorem, the delta method and formulae for multinormal cross-moments). Then a simple pivot tuning exercise is to simulate say 1000 copies of $r_n$, for each $\rho$ on a grid, and then run a simple linear regression on the outcomes $sd(r_n)$ with covariates, say, $\rho, \rho^2, \rho^3, \rho^4$. Then you quickly enough discover that indeed $\kappa = 1 - \rho^2$, cf. Exercise 7.3. Even if this had not been mathematically correct (it is, see Exercise A.2), using this simulation-based recipe for fine-tuning the pivot would easily give a version of (7.8) that would work quite accurately.

The delta method mechanism may be used to find the transformation $\hat{\zeta} = h(r_n)$ which is variance stabilising, as this requires $h'(\rho)\kappa(\rho)$ being constant. The solution is Fisher’s zeta transform, $\zeta = h(\rho) = \frac{1}{2} \log((1 + \rho)/(1 - \rho))$, which via (7.1) gives the simple approximate confidence distribution $\Phi(\sqrt{n}(\zeta(\rho) - \zeta))$, with $\hat{\zeta} = \frac{1}{2} \log((1 + r_n)/(1 - r_n))$. One may also use simulation output to find that $1/(n - 3)$ is a better approximation to the variance of $\hat{\zeta}$ than is $1/n$, leading to the pivot tuned $C^*(\rho) = \Phi((n - 3)^{1/2}[\zeta(\rho) - \zeta])$. ■
Improved approximations for confidence distributions

Example 7.3 Poisson parameter
Let $T$ be Poisson with mean $\psi$. The half-corrected cumulative confidence distribution function is

$$C(\psi) = 1 - \sum_{j=0}^{t_{\text{obs}}} \exp(-\psi) \psi^j / j! + \frac{1}{2} \exp(-\psi) \psi^{t_{\text{obs}}} / t_{\text{obs}}!.$$

Here $V = 2(\sqrt{\psi} - \sqrt{T})$ is approximately $N(0, 1)$ and is accordingly approximately a pivot for moderate to large $\psi$. From a simulation experiment, one finds that the distribution of $V$ is slightly skewed, and has a bit longer tails than the normal. By a little trial and error, one finds that $\exp(V/1000)$ is closely Student distributed with $\text{df} = 30$. With $Q_{30}$ being the upper quantile function of this distribution and $t_{30}$ the density, the log-likelihood is approximately

$$\ell_s(\psi) = \log t_{30}(Q_{30}(C(\psi))).$$

Examples are easily made to illustrate that the $\ell_s(\psi)$ log-likelihood quite closely approximates the real Poisson log-likelihood $\ell(\psi) = t_{\text{obs}} - \psi + t_{\text{obs}} \log(\psi/t_{\text{obs}})$. Our point here is to illustrate the approximation technique; when the exact likelihood is available it is of course preferable.

The confidence density depends on the parametrisation. By reparametrisation, the likelihood can be brought to be proportional to the confidence density. This parametrisation might have additional advantages. Let $L(\psi)$ be the likelihood and $c(\psi)$ the confidence density for the chosen parametrisation, both assumed positive over the support of the confidence distribution. The quotient $J(\psi) = L(\psi)/c(\psi)$ has an increasing integral $\mu(\psi)$, with $(\partial/\partial \psi) \mu = J$, and the confidence density of $\mu = \mu(\psi)$ is $L(\mu(\psi))$. There is thus always a parametrisation that makes the observed likelihood proportional to the observed confidence density. When the likelihood is based on a pivot of the form $\mu(\psi) - T$, the likelihood in $\mu = \mu(\psi)$ is proportional to the confidence density of $\mu$.

Example 7.4 The exponential rate parameter (Example 3.1 continued)
Let $\hat{\psi}/\psi$ be standard exponentially distributed. Taking the logarithm, the pivot is brought on translation form, and $\mu(\psi) = \log \psi$. The confidence density is thus $c(\mu) \propto L(\mu) = \exp(\hat{\mu} - \mu - \exp(\hat{\mu} - \mu))$. The log-likelihood has a more normal-like shape in the $\mu$ parametrisation than in the canonical parameter $\psi$.

When the likelihood equals the confidence density, the pivot is in broad generality of the translation type. The cumulative confidence distribution function is then of translation type, with $C(\mu) = F(\mu - \hat{\mu})$, and so is the likelihood, $L(\mu) = c(\mu) = f(\mu - \hat{\mu})$.

7.4 Bartlett corrections for the deviance
Consider a parametric model indexed by $\theta$ with a suitable focus parameter $\psi = a(\theta)$. A quite fruitful general recipe for constructing confidence curves for such a $\psi$ is via the $\chi_1^2$ approximation to the distribution of the deviance; cf. the discussion in Section 3.4 and various applications in previous chapters. The accuracy of the $cc(\psi) = \Gamma(1)(D_n(\psi))$ recipe depends on how well the implied approximation

$$P_0[cc(\psi) \leq \alpha] = P_0[D_n(\psi) \leq \Gamma^{-1}(\alpha)] \approx \alpha$$

works.
works. This might depend not only on the sample size \( n \), but also on the position \( \theta \) in the parameter space and the probability level \( \alpha \). The second-order theory developed in Section 7.2 indicates that quite generally, \( \text{E}_\theta D_n(\psi) = 1 + c(\theta)/n + O(1/n^2) \); for a suitable \( c(\theta) \), see (7.6). This holds up also in multiparameter situations, in wide generality; see, for example, Barndorff-Nielsen and Cox (1994, Chapter 4) and Brazzale et al. (2007, Chapter 2).

This suggests employing the modification \( D_n(\psi)/(1 + \epsilon) \) to come closer to the \( \chi^2_1 \) distribution, where \( 1 + \epsilon \) is the mean of \( D_n(\psi) \), perhaps computed by simulating a high number of replicates of \( D_n(\psi) \) at the required position \( \theta \) in the parameter space (as the \( \epsilon \) may depend on \( \theta \)). The theory secures that \( \epsilon \) will be close to zero for large \( n \), and even that \( n\{\text{E}_\theta D_n(\psi) − 1\} \) will converge to the constant \( c(\theta) \) in the preceding, offering another recipe for calculating \( \epsilon \). This type of modification, dividing by the right constant to get the mean right, is called ‘Bartlett correction’ in the statistical literature, and usually carried out in a context of statistical testing, to get accurate critical values and \( p \)-values. In our framework the method leads to what we term Bartlett corrected deviance confidence curves,

\[
\text{cc}^*(\psi) = \Gamma_1\left(\frac{D_n(\psi)}{\text{E}_\theta D_n(\psi)}\right) = \Gamma_1\left(\frac{D_n(\psi)}{1 + \epsilon}\right).
\]  

(7.10)

The point is that \( D_n(\psi) \) has a distribution function better approximated by \( K_n(x) = \Gamma_1(x/(1 + \epsilon)) \) than by \( \Gamma_1(x) \), and this produces (7.10).

As a reasonably simple illustration, consider again the situation of Example 7.1 with independent positive observations from the exponential density \( \exp(-\theta y) \). The deviance statistic is \( D_n(\theta) = 2n(\log\hat{\theta}/\theta - 1 + \hat{\theta}/\bar{\theta}) \), equal in distribution to \( D_n = 2n(V_n - 1 - \log V_n) \), where \( V_n \sim \chi^2_{2n}/(2n) \). We saw in the course of that example that \( \text{E}D_n = 1 + \frac{1}{6}/n + O(1/n^2) \).

The distribution of \( D_n/(1 + \frac{1}{6}/n) \) is seen via an easy simulation exercise to be rather closer to the \( \chi^2_1 \) than \( D_n \) itself, for small \( n \); see Exercise 7.1.

Bolt from heaven and extreme value statistics

Like the proverbial bolt from heaven, Usain Bolt burst onto the world scene on 31 May 2008, when he sprinted the 100 m in a world record time of 9.72 seconds (beating Asafa Powell’s previous record by 0.02 seconds; cf. Bolt [2013, p. 135]). Bolt was 21 years old and had previously focussed on the 200-m and 400-m races. His personal record on the 100-m race before the 2008 season was a mere 10.03 seconds. But how surprising was this world record by itself?

We translate this question into a statistical one as follows. We first track down and log each 100-m race in this world clocked at 10.00 or better, from the eight seasons before Bolt’s first record race, that is, from 2000 to 2007. Excluding races with too much tailwind, and also all races from athletes later on having been officially found to have used illegal performance enhancing substances, we have found a total of \( n = 195 \) such races, ranging from Asafa Powell’s previous world record 9.74 to 10.00. For convenience we translate these race results \( r_i \) to \( y_i = 10.005 - r_i \), to more easily be able to utilise the vast literature on extreme value statistics in its standard form and notation (Hary, 1960). The question is then how likely or unlikely it is, as seen from the start of 2008 after having watched 195 thrilling sub-10.005 races over eight years, that the world in the course of 2008 should experience a
Improved approximations for confidence distributions

race of 9.72 or better. This is $p = P(\max(Y'_1, \ldots, Y'_N) \geq w_0)$, with $w_0 = 10.005 - 9.72 = 0.285$, where $Y'_1, \ldots, Y'_N$ correspond to one year’s worth of $N$ world-class races (in the category of 10.005 or better). It turns out that we can model these top races using a parametric family and then find a formula for $p$, for which we shall find an estimate and establish a confidence distribution.

The distribution in question is

$$G(y, a, \sigma) = 1 - (1 - ay/\sigma)^{1/a}, \quad (7.11)$$

as dictated by general extreme value theory for values above a threshold; see, for example, Embrechts et al. (1997). This leads to a log-likelihood function

$$\ell_n(a, \sigma) = \sum_{i=1}^n \{-\log \sigma + \left(\frac{1}{a} - 1\right) \log (1 - ay_i/\sigma)\}.$$ 

Maximum likelihood parameter estimates for $a$ and $\sigma$ are 0.1821 (0.699) and 0.0745 (0.0074) (with approximate standard errors in parentheses). The model fits the sprint data well; see Figure 7.2. If the $N$ above is a Poisson variable with parameter $\lambda$, then a calculation gives

$$p = p(a, \sigma) = P(\max(Y'_1, \ldots, Y'_N) \geq w)$$
$$= 1 - \exp\{-\lambda (1 - aw/\sigma)^{1/a}\}, \quad (7.12)$$

see Exercise 7.7.

**Figure 7.2** The left panel displays the 195 fastest races recorded during the eight years 2000 to 2007 (better than or equal to Armin Hary), along with Bolt’s 9.72 seconds from May 2008; Bolt is not among those contributing to data points before 2008. The right panel gives their empirical distribution, along with the distribution function fitted from the extreme value distribution (7.11).
7.4 Bartlett corrections for the deviance

We are now on familiar ground, as we may use the deviance statistic

\[ D_n(p_0) = 2(\ell_n(\hat{\alpha}, \hat{\sigma}) - \ell_{n, \text{prof}}(p_0)), \]

involving the profile log-likelihood function \[ \ell_{n, \text{prof}}(p_0) = \max \{ \ell_n(a, \sigma) : p(a, \sigma) = p_0 \}. \] For \( \lambda \) it makes sense to use \( 195/8 = 24.375 \), the annual rate of top class races. By general likelihood theory, \( D_n(p_0) \) is approximately a \( \chi^2_1 \), under the true \( (a, \sigma) \), leading to the usual \( cc(p) = \Gamma_1(D_n(p)) \). Using Bartlett correction, we find for sample size \( n = 195 \) via simulation that \( D_n(p_0) \approx d \chi^2_1/1.07 \), and we also find that this approximation is good in a decent neighbourhood around the maximum likelihood estimates in the \( (a, \sigma) \) parameter space. This leads to the modified confidence curve

\[ cc^*(p) = \Gamma_1(D_n(p)/1.07). \]

The two curves are in fact very similar and are displayed in Figure 7.3. The point estimate for the surprise probability \( p \) is 3.5%, but the confidence distribution is skewed; a 90% confidence interval for \( p \) is [0.0%, 18.9%].

Of course these calculations are based on a set of assumptions whose validity might be queried. We do treat the 195 races as independent, even when it comes to multiple results from the same athlete. Perhaps more crucially, we are also assuming that the world of top sprinters was in reasonable equilibrium, over the years from 2000 to 2007, without any detectable trends in top performance. Overall we find these assumptions not unreasonable (see Figure 7.2), and hence have basic trust in the usefulness of the confidence distribution calculations we just carried out.

![Figure 7.3](image-url)  
**Figure 7.3** Confidence curves for the surprise probability \( p \) of (7.12) associated with Bolt’s 9.72-second race from May 2008, based on the deviance (solid line) and Bartlett corrected deviance (dashed line). The point estimate is 0.035 and a 90% confidence interval is [0, 0.189].
We might translate this kind of analysis to a surprise score of $q = 1 - p$ (in this case the probability, as of January 2008, that there will be no world record set in the course of that year of 9.72 or better). The surprise score associated with Bolt’s 31 May 2008 performance is thus 96.5%, with a confidence interval of [82.1%, 100%]. One may use the developed machinery to quantify the level or surprise for new records, and hence arguably to rank records of different events in terms of uniqueness, and so forth. Figure 7.4 provides surprise scores, along with surprise confidence scores, associated with the three greatest races of 2008: Tyson Gay’s 9.77 from June, Bolt’s 9.72 from May and finally Bolt’s 9.69 from August, where the surprise score is a beamonesque 99.2%. For further details, consult Exercise 7.7.

### 7.5 Median-bias correction

Bias correction is often desirable in point estimation. This is so also when converting a deviance function $D$ to a confidence curve. To illustrate that bias correction might be desirable, consider a sample of size $n$ from the uniform distribution over $(0, \psi)$. The maximum likelihood estimator is the largest observation $Y_{(n)}$, which by pivoting yields the cumulative confidence $C(\psi) = 1 - (Y_{(n)}/\psi)^n$ for $\psi \geq Y_{(n)}$. The deviance is $D(\psi) = 2n \log(\psi/Y_{(n)})$, which has distribution function $F(d) = P_{\psi}(D \leq d) = 1 - \exp(-\frac{1}{2}d)$, free of $\psi$. The confidence curve obtained by exact conversion, cf. Section 10.3, is thus $cc_{\psi}(\psi) = 1 - (Y_{(n)}/\psi)^n$ for $\psi \geq Y_{(n)}$. This confidence curve yields $\alpha$-level confidence intervals $Y_{(n)}[1,(1 - \alpha)^{-1/n}]$. The probability of missing $\psi$ to the left is zero, while
the probability of missing to the right is $\alpha$. The confidence curve is tail-asymmetric to the extreme, as opposed to the confidence curve $cc(\psi) = |2C(\psi) - 1|$ which is tail symmetric.

Let $\hat{\psi}$ be the maximum likelihood estimator and let $m(\psi)$ be its median, $P_\psi(\hat{\psi} \leq m(\psi)) = \frac{1}{2}$. When $m(\psi) \neq \psi$ the likelihood is median-biased in the sense that the confidence curve obtained from the deviance $cc(\psi) = F_\psi(D(\psi))$ yields confidence intervals with different probabilities of missing the true value to the right and to the left. That is,

$$P_\psi(cc(\psi) \geq \alpha, \hat{\psi} < \psi) \neq P_\psi(cc(\psi) \geq \alpha, \hat{\psi} > \psi).$$

This tail-asymmetry is always true for small values of $\alpha$, and appears generally to be the case. By bias-correcting the deviance function before conversion the resulting confidence curve is nearly tail-symmetric.

**Definition 7.1 (Median-bias corrected deviance)** The median-bias corrected deviance is $D(m(\psi))$, with distribution $F_{m(\psi)}$, resulting in the confidence curve $cc_{bc}(\psi) = F_{m(\psi)}(D(m(\psi)))$ by exact conversion (cf. Section 10.3).

When an exact confidence distribution $C$ exists by pivoting a sufficient statistic, $cc_{bc}$ is not only nearly tail-symmetric, but is also nearly equal to the tail-symmetric confidence curve $|2C(\psi) - 1|$.

De Blasi and Schweder (2015) show that the convergence to tail symmetry is of the third order in a wide class of one-parametric models, and also for the normal one-way analysis of variance model with fixed sample size within groups but with different group means $\mu_i$. Here the group means are nuisance parameters, and bias correction converges to tail symmetry at the third order of number of groups.

Consider first the normal transformation model for an i.i.d. sample of size $n$. There is an estimator $\hat{\psi}$, which need not be the maximum likelihood one, with bias and kurtosis of the order $O(1/n)$, and standard deviation and skewness of the order $O(1/\sqrt{n})$. There is an increasing transformation $g$ and constants $z_0$ and $a$ such that for $\hat{\phi} = g(\hat{\psi})$ and $\phi = g(\psi)$,

$$\hat{\phi} = \phi + (1 + a\phi)(Z - z_0) \text{ with } Z \sim N(0,1).$$

In this model there is a pivot based on $\hat{\psi}$ yielding a confidence distribution $C$ and a tail-symmetric confidence curve $cc(\psi) = |2C(\psi) - 1|$. The third-order convergence to tail-symmetry of the bias-corrected confidence curve is formulated in Theorem 7.2.

The other class of one-parametric models to consider is the exponential family where the sample of size $n$ is drawn from a distribution with probability density $f(y, \psi) = \exp\{\psi S(y) - k(\psi)\}$. This is the family considered in Section 5.5, where an optimal confidence distribution $C$ was found by conditioning, in the multiparameter exponential family. There shall be no nuisance parameters now, and no conditioning. The maximum likelihood estimator $\hat{\psi}$ has median $m(\psi)$. The deviance is $D(\psi)$, yielding the bias corrected confidence curve $cc_{bc}(\psi) = F_{m(\psi)}(D(m(\psi)))$. There is also here third-order convergence to the tail-symmetric confidence curve $cc(\psi) = |2C(\psi) - 1|$. Using the notation $O_p(n^{-k})$ for asymptotic remainders, which means $n^kO_p(n^{-k})$ being bounded in probability, De Blasi and Schweder (2015) show the following.
Improved approximations for confidence distributions

Theorem 7.2 (Third-order accuracy of approximation) In the normal transformation model (7.13), and in the one-parametric exponential family, for \( \sqrt{n}(\psi - \hat{\psi}) = O_p(1) \), one has \( \text{cc}_{bc}(\psi) = \text{cc}(\psi) + O_p(n^{-3/2}) \).

Median-bias correction might also be applied to deviances from profile likelihoods when there are nuisance parameters in the model but the median of \( \hat{\psi} \) is free of nuisance parameters. This is, for example, the case in the Neyman–Scott problem (see Section 4.4) of one-way analysis of variance for normal data, where \( n \geq 2 \) is the intragroup sample size for the \( g \) groups, all with variance \( \sigma^2 \) but with different and unknown group means \( \mu_i \). Despite the presence of the \( g \) nuisance parameters, the bias-corrected confidence curve \( \text{cc}_{bc} \) converges fast to the tail-symmetric confidence curve \( |2C(\sigma) - 1| \) obtained by pivoting the maximum likelihood estimator \( \hat{\sigma} \). De Blasi and Schweder (2015) reach a result similar to that of Theorem 7.2, and show that if \( \sqrt{n}(\sigma - \hat{\sigma}) = O_p(1) \), then \( \text{cc}_{bc}(\sigma) = |2C(\sigma) - 1| + O_p(g^{-3/2}) \) regardless of the group size \( m \).

Example 7.5 The normal variance

Let \( Y_1, \ldots, Y_n \) be i.i.d. \( N(\mu, \sigma^2) \). This is the Neyman–Scott model with \( g = 1 \) group. The profile deviance \( D_n \) for \( \sigma \) is \( n(\hat{\sigma}^2/\sigma^2 - 1 - \log(\hat{\sigma}^2/\sigma^2)) \), where \( \hat{\sigma} \) is the maximum likelihood estimator. The median of \( \hat{\sigma}^2 \) is \( \sigma^2 \Gamma_{n-1}(-\frac{1}{2})/n \), where again \( \Gamma_{n-1} \) is the cumulative distribution function of the \( \chi^2 \) with \( n - 1 \) degrees of freedom. Let \( \Gamma_{\nu} \) be the cumulative \( \chi^2 \) distribution at \( \nu \) degrees of freedom. The median of \( \hat{\sigma}^2 \) is \( m(\sigma^2) = \sigma^2 \Gamma_{n-1}(-\frac{1}{2})/n \). Let \( s_1(d) < s_2(d) \) be the two solutions of the equation \( 2\log s + s^2 - 1 = d \). The distribution function for \( D(\sigma) \) at the true value is \( F(d) = \Gamma_{n-1}(ns_1(d/n)^{-2}) - \Gamma_{n-1}(ns_2(d/n)^{-2}) \) independent of \( \sigma \) and the nuisance parameter \( \mu \). Figure 7.5 displays confidence curves for \( \sigma \) based on a small simulated data

![Figure 7.5](image-url)  

Confidence curves for \( \sigma \) for a simulated dataset of size \( n = 5 \) from the \( N(\mu, \sigma^2) \). Solid line: cc from the \( \chi^2 \) pivot; dotted line, almost on top of the solid line: cc from exact conversion of the profile deviance, with median-bias correction; dashed line: cc from exact conversion of the profile deviance. The maximum likelihood estimate is \( \hat{\sigma} = 0.26 \).
7.6 The t-bootstrap and abc-bootstrap method

For a suitable increasing transformation of \( \psi \) and \( \hat{\psi} \) to \( \gamma = h(\psi) \) and \( \hat{\gamma} = h(\hat{\psi}) \), suppose

\[
t = (\gamma - \hat{\gamma}) / \hat{\tau}
\]

(7.14)

where \( \hat{\tau} \) is a scale estimate. Let \( R \) be the distribution function of \( t \), by assumption approximately independent of underlying parameters \((\psi, \chi)\). The approximate confidence distribution for \( \gamma \) is thus

\[
C_0(\gamma) = R((\gamma - \hat{\gamma}) / \hat{\tau})
\]

yielding in its turn

\[
C(\psi) = R((h(\psi) - h(\hat{\psi}))/\hat{\tau})
\]

(7.15)

This \( t \)-bootstrap method applies even when \( t \) is not a perfect pivot, but is especially successful when it is, because \( t^* \) then has exactly the same distribution \( R \) as \( t \). Note that the method automatically takes care of bias and asymmetry in \( R \), and that it therefore aims at being more precise than the delta method considered previously, which corresponds to zero bias and a normal \( R \). The problem is that an educated guess is required for a successful pivotal transformation \( h \) and the scale estimate \( \hat{\tau} \). Such might be facilitated and aided by regression analysis of the bootstrapped statistics. Furthermore, the \( t \)-bootstrap interval is not invariant under monotone transformations. The following method is not hampered by this shortcoming.

Efron (1987) introduced acceleration and bias-corrected bootstrap percentile intervals and showed that these have several desirable aspects regarding accuracy and parameter invariance. Here we exploit some of these ideas, but ‘turn them around’ to construct accurate bootstrap-based approximations to confidence distributions and confidence likelihoods.

Suppose that on some transformed scale, from \( \psi \) and \( \hat{\psi} \) to \( \gamma = h(\psi) \) and \( \hat{\gamma} = h(\hat{\psi}) \), one has

\[
(\gamma - \hat{\gamma}) / (1 + a\gamma) - b \sim N(0, 1)
\]

(7.16)

to a very good approximation, for suitable constants \( a \) (for acceleration) and \( b \) (for bias). This is the normal transformation model of Efron (1987) considered in (7.13). Both population parameters \( a \) and \( b \) tend to be small; in typical setups with \( n \) observations, their sizes will be \( O(n^{-1/2}) \). Assuming \( a\hat{\gamma} > -1 \), the pivot in (7.16) is increasing in \( \gamma \) and

\[
C(\gamma) = \Phi((\gamma - \hat{\gamma}) / (1 + a\gamma) - b)
\]

is the confidence distribution for \( \gamma \). Thus

\[
C(\psi) = \Phi(h(\psi) - h(\hat{\psi}) / (1 + ah(\psi) - b))
\]

(7.17)
Improved approximations for confidence distributions

is the resulting confidence distribution for \( \psi \). This constitutes a good approximation to the real confidence distribution under assumption (7.16). It requires \( h \) to be known, however, as well as \( a \) and \( b \).

To come around this, look at bootstrapped versions \( \hat{\psi}^* = h(\hat{\psi}) \) from the estimated parametric model. If assumption (7.16) holds uniformly in a neighbourhood of the true parameters, then also

\[
(\hat{\psi}^* - \hat{\psi})/(1 + a\hat{\psi}) \sim N(-b, 1)
\]

with good precision. Hence the bootstrap distribution may be expressed as

\[
\hat{G}(t) = P_*(\hat{\psi}^* \leq t) = P_*(\hat{\psi}^* \leq h(t)) = \Phi\left(\frac{h(t) - \hat{\psi}}{1 + a\hat{\psi}} + b\right),
\]

which yields \( h(t) = (1 + a\hat{\psi})\{\Phi^{-1}(\hat{G}(t)) - b\} + \hat{\psi} \). Substitution in (7.17) is seen to give the abc formula

\[
\hat{G}_{abc}(\psi) = \Phi\left(\frac{\Phi^{-1}(\hat{G}(\psi)) - b}{1 + a(\Phi^{-1}(\hat{G}(\psi)) - b)} - b\right),
\]

(7.18)

since \( \Phi^{-1}(\hat{G}(\psi)) = b \). Note that an approximation \( \hat{G}_{abc}(\psi) \) to the confidence density emerges too, by evaluating the derivative of \( G_{abc} \). This may sometimes be done analytically, in cases where \( \hat{G}(\psi) \) can be found in a closed form, or may be carried out numerically.

The confidence abc-likelihood is from (7.16) equal to \( L(\gamma) = \phi((\gamma - \hat{\gamma})/(1 + a\gamma))/ (1 + a\gamma) \), which yields the log-likelihood

\[
\ell_{abc}(\psi) = -\frac{1}{2}(\Phi^{-1}(\hat{G}_{abc}(\psi)))^2 - \log[1 + a(\Phi^{-1}(\hat{G}(\psi)) - b)],
\]

(7.19)

because the unknown proportionality factor \( 1 + a\hat{\gamma} \) appearing in \( h(t) \) is a constant proportionality factor in \( L_{abc}(h(\psi)) \).

It remains to specify \( a \) and \( b \). The bias parameter \( b \) is found from \( \hat{G}(\psi) = \Phi(b) \), as noted earlier. The acceleration parameter \( a \) is found as \( a = \frac{1}{n} \text{skew} \), where there are several ways in which to calculate or approximate the skewness parameter in question. Extensive discussions may be found in Efron (1987), Efron and Tibshirani (1993, Chapters 14 and 22) and in Davison and Hinkley (1997, Chapter 5) for situations with i.i.d. data. One option is via the jackknife method, which gives parameter estimates \( \hat{\psi}_{(i)} \) computed by leaving out data point \( i \), and use

\[
a = \frac{1}{n} \frac{1}{n^{-1/2} \text{skew}} \{\hat{\psi}_{(1)} - \hat{\psi}, \ldots, \hat{\psi}_{(n)} - \hat{\psi}_{(n)}\}.
\]

Here \( \hat{\psi}_{(i)} \) is the mean of the \( n \) jackknife estimates. Another option for parametric families is to compute the skewness of the logarithmic derivative of the likelihood, at the parameter point estimate, inside the least favourable parametric subfamily; see again Efron (1987) for more details.

Note that when \( a \) and \( b \) are close to zero, the abc-confidence distribution becomes identical to the bootstrap distribution itself. Thus (7.18) provides a second-order nonlinear correction of shift and scale to the immediate bootstrap distribution. We also point out that results similar in nature to (7.19) have been derived in Pawitan (2000) in a somewhat different context of empirical likelihood approximations.
7.7 Saddlepoint approximations and the magic formula

Example 7.6 Ratio of scale parameters (Example 3.3 continued)

Consider again the parameter $\psi = \sigma_2^2 / \sigma_1^2$ of Example 3.3. The exact confidence distribution was derived there and is equal to $C(\psi) = 1 - K(\hat{\psi}^2 / \psi^2)$, with $K = K_{\nu_1, \nu_2}$, the Fisher distribution. We shall examine how successful the abc apparatus is for approximating the $C(\psi)$ and its confidence density $c(\psi)$.

In this situation, bootstrapping from the estimated parametric model leads to $\hat{\psi}^* = \hat{\sigma}_2^* / \hat{\sigma}_1^*$ of the form $\hat{\psi}^* F^{1/2}$, where $F$ has degrees of freedom $\nu_2$ and $\nu_1$. Hence the bootstrap distribution is $\hat{G}(t) = K(t^2 / \hat{\psi}^2)$, and $\hat{G}(\hat{\psi}) = \Phi(b)$ determines $b$. The acceleration constant can be computed exactly by examining the log-derivative of the density of $\hat{\psi}$, which from $\hat{\psi} = \psi F^{1/2}$ is equal to

$$p(r, \psi) = k(r^2 / \psi^2)^2 r / \psi^3.$$  

With a little work the log-derivative can be expressed as

$$\frac{1}{\psi} \left\{ -\nu_2 + (v_1 + v_2) \frac{(v_2/v_1) \hat{\psi}^2 / \psi^2}{1 + (v_2/v_1) \hat{\psi}^2 / \psi^2} \right\} = \psi \left\{ \frac{\text{Beta}(\frac{1}{2}, \frac{v_1}{2}, \frac{v_2}{2})}{v_1 + v_2} - \frac{v_2}{v_1 + v_2} \right\}.$$  

Here $\text{Beta}(a, b)$ denotes a Beta distribution with density proportional to $x^{a-1}(1-x)^{b-1}$. Calculating the three first moments of the Beta gives a formula for its skewness and hence for $a$:

$$a = \frac{1}{6} \text{SKEW} = \frac{1}{6} \frac{1 - 2p}{\sqrt{p(1-p)}} \frac{k+1}{k+2},$$  

with $k = (v_1 + v_2) / 2$ and $p = v_2 / (v_1 + v_2)$. (Using the preceding jackknife formula, or relatives directly based on simulated bootstrap estimates, obviates the need for algebraic derivations, but would give a good approximation only to the $a$ parameter for which we here found the exact value.)

Trying out the abc machinery shows that $\hat{C}_{abc}(\psi)$ is amazingly close to $C(\psi)$, even when the degrees of freedom numbers are low and imbalanced; the agreement is even more perfect when $v_1$ and $v_2$ are more balanced or when they become larger. The same holds for the densities $\hat{c}_{abc}(\psi)$ and $c(\psi)$; see Figure 7.6.  

7.7 Saddlepoint approximations and the magic formula

There is a quite large class of successful approximations to the distribution of both maximum likelihood estimators and the deviance statistic associated with second-order modifications to the log-likelihood function. The methods we shall briefly report on in this section are those related to mathematical techniques variously called Laplace approximations, saddlepoint approximations, exponential tilting, and so on. We shall focus on just a few of these, among them Barndorff-Nielsen’s ‘magic formula’ and consequent modified profile log-likelihoods. These again lead to fruitful constructions for confidence distributions and confidence curves.

Consider a parametric family indexed by some $p$-dimensional $\theta$, giving rise to a log-likelihood function $\ell(\theta) = \log L(\theta)$ and the maximum likelihood estimator $\hat{\theta}$. Two of the basic results from the first-order theory of Chapter 2 may be expressed as follows; see
Figure 7.6 Exact confidence density (full line) along with its abc-estimated version (dotted line), for parameter $\psi = \sigma_2/\sigma_1$ with 5 and 10 degrees of freedom. The parameter estimate in this illustration is $\hat{\psi} = 0.50$.

Sections 2.2 and 2.4:

\[
\ell(\theta) - \ell(\hat{\theta}) = -\frac{1}{2}(\theta - \hat{\theta})^t J(\theta - \hat{\theta}),
\]

\[
p(\hat{\theta}, \theta) = (2\pi)^{-p/2} |J(\hat{\theta})|^{1/2} \exp\left\{-\frac{1}{2}(\theta - \hat{\theta})^t J(\hat{\theta} - \theta)\right\}.
\]

Here $J(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^t$, with $\hat{J} = J(\hat{\theta})$ the observed information matrix, and $p(\cdot, \theta)$ is the probability density of $\hat{\theta}$ under $\theta$. The notation is kept generic here, and for various applications there would naturally be a sample size $n$ involved too, with $\hat{J}$ proportional to $n$.

Taking the two approximations above together, and normalising, we have

\[
p^*(\hat{\theta}, \theta) = k(\theta)(2\pi)^{-p/2} [\hat{J}(\hat{\theta})]^{1/2} \exp\{\ell(\theta) - \ell(\hat{\theta})\},
\]

(7.20)

with $k(\theta)$ the constant required to make this integrate to 1. We will in fact often have $k(\theta)$ close to 1.

This 'p-star formula' is a version of the so-called magic formula invented by Barndorff-Nielsen (1983) and subsequently extended and refined by himself and several others; see Barndorff-Nielsen (1986), Barndorff-Nielsen and Cox (1989, 1994) and Brazzale et al. (2007) for further discussion. It is instructive to work out the details of some concrete cases, in which the exact distribution of the maximum likelihood estimator may be written down, to see how well the $p^*$ formula works; see Exercise 7.8. We note that for the purposes of finding potential improvements on the $\chi^2_1$ approximation for the deviance statistic, the emphasis is on how the formula varies with $\theta$ for observed $\hat{\theta}$ and less with the probability distribution of $\hat{\theta}$ for fixed parameter. Here (7.20) leads to the modified log-likelihood
expression \( \ell(\theta) - \ell(\hat{\theta}) + \log k(\theta) \) and hence to the modified deviance
\[
D_\text{m}(\theta) = 2(\ell(\hat{\theta}) - \ell(\theta) + \log k(\hat{\theta}) - \log k(\theta))
\]
\[
= D(\theta) + 2[\log k(\hat{\theta}) - \log k(\theta)].
\]

The next stage in the development is to examine the case of a focus parameter \( \psi = a(\theta) \), with the intention of using magic formulae as earlier to work out an approximation for the conditional distribution of the relevant maximum likelihood component \( \hat{\psi} \) given the rest of the information in \( \hat{\theta} \). This will then lead to a modified profile log-likelihood function and consequentially a modified deviance statistic and modified confidence constructions. Let us structure the model parameter as \( \theta = (\psi, \chi) \), via reparameterisation if necessary, with \( \chi \) being \((p - 1)\)-dimensional nuisances. Using the aforementioned scheme we may first approximate the distribution of the maximum likelihood estimator \( \hat{\chi}(\psi) \) for given \( \psi \), that is, the estimator one needs to compute when calculating the profiled log-likelihood. This takes the form
\[
p(\hat{\chi}(\psi), \theta) = (2\pi)^{-(p-1)/2} |J_{11}(\psi, \hat{\chi}(\psi))|^{1/2} \frac{L(\theta)}{L(\psi, \hat{\chi}(\psi))} |\mathcal{J}(\psi)|.
\]
Here the \( p \times p \) matrix \( \mathcal{J}(\psi, \chi) \) is decomposed into blocks \( J_{00}, J_{01}, J_{10}, J_{11} \), with \( J_{11}(\psi, \hat{\chi}(\psi)) \) being the observed information matrix of size \((p - 1) \times (p - 1)\), computed for given value of \( \psi \). Also, \( \mathcal{J}(\psi) \) is the Jacobian of the transformation, \( \partial \hat{\chi}(\psi)/\partial \chi \), evaluated at \( \hat{\chi}(\psi) \) for the given \( \psi \).

Then on using \( p(\hat{\psi}, \hat{\chi}, \theta)/p(\hat{\chi}(\psi), \theta) \) and simplifying, we find that the conditional density of \( \hat{\psi} \), given \( \hat{\chi} \), at position \( \theta = (\psi, \chi) \) in the parameter space, is of the form
\[
p(\hat{\psi} \mid \hat{\chi}(\psi), \theta) \propto L(\psi, \hat{\chi}(\psi)) |J_{11}(\psi, \hat{\chi}(\psi))|^{1/2} |\mathcal{J}(\psi)|^{-1}.
\]
This yields finally a formula for the log-density of our focus estimator \( \hat{\psi} \), conditioning on the supplementary information in \( \hat{\chi} \), namely
\[
\ell_{\text{BN}}(\psi) = \ell_{\text{prof}}(\psi) + \frac{1}{2} \log |J_{11}(\psi, \hat{\chi}(\psi))| - \log |\mathcal{J}(\psi)|. \quad (7.21)
\]
One may also demonstrate that this is appropriately close to the marginal log-likelihood for \( \hat{\psi} \); see, for example, Barndorff-Nielsen and Cox (1994, chapter 8). This is Barndorff-Nielsen’s modified profile log-likelihood. Some notes are in order.

1. The preceding brief treatment may look partly different, and perhaps simpler on the page, compared to how this is presented in other technical literature. This is partly because we indirectly downplay the use of this machinery for producing approximations to the distribution of the maximum likelihood estimator; for our purposes we are more interested in using the apparatus for computing log-likelihood profiles and hence confidence distributions for given datasets. In other words, there are important terms somehow glossed over in the ‘proportional to’ sign above, which have to do with the former aspect (the distribution of the estimator, for given parameter values) but not the latter (producing log-likelihood profiles and confidence curves).

2. Under various regular setups, with a sample of \( n \) independent observations, the three terms of (7.21) are of sizes \( O(n), O(p \log n), O(1) \), respectively. In particular, the dominant term here is still the profile log-likelihood function we have used in various situations elsewhere in the book, and for large \( n \) the differences disappear.
Improved approximations for confidence distributions

3. Under some conditions, the $J(\psi)$ matrix has a determinant close to 1, making the last term disappear. One such class of situations is when there is parameter orthogonality, in the sense of Cox and Reid (1987). Such parameter orthogonality is always achievable, through reparameterisation, if $\psi$ or $\chi$ or both are scalars.

4. Often the precise numerical evaluation of the $|J(\psi)|$ term is the most difficult step when implementing our version of the Barndorff-Nielsen scheme given earlier. Various careful illustrations, also of closely related methods and with applications to analysis of real datasets, are offered in Brazzale et al. (2007).

5. Importantly, the modified profile log-likelihood $\ell_{BN}$ above lends itself nicely to the construction of accurate confidence distributions for the focus parameter $\psi$. The first method is

$$C_{BN}(\psi) = P_{\psi,\chi} (\hat{\psi} \geq \hat{\psi}_{obs}),$$

computed via the Barndorff-Nielsen approximation $p(\hat{\psi}|\hat{\chi},\theta)$ above, carried out at the maximum likelihood position $\hat{\chi}_{obs}$ for $\chi$. The second and often easier method uses a certain refinement of the directed likelihood root statistic $r(\psi) = \text{sgn}(\psi - \hat{\psi})D(\psi)^{1/2}$. The refinement in question takes the form

$$r^*(\psi) = r(\psi) + r(\psi)^{-1} \log \frac{u(\psi)}{r(\psi)},$$

with a certain $u(\psi)$ related to (7.21); explicit expressions are given and illustrated in Brazzale et al. (2007, chapter 2). The point is that $r^*(\psi)$ typically has a distribution closer to the standard normal than does $r(\psi)$. This leads to a magic formula version for the confidence curve:

$$cc_{BN}(\psi) = \Gamma_1(\psi^2).$$

7.8 Approximations to the gold standard in two test cases

As has been made clear in this chapter, there is a somewhat bewildering variety of methods in the statistician’s toolbox, all aiming at making better approximations than those afforded by the simpler first-order large-sample theory for likelihood inference, with recipes also for confidence distributions and confidence curves. Rather different correction and modification techniques may all succeed in getting successfully ‘to the next order’, and we ought not to expect one of these methods to be uniformly better than the others in terms of performance and accuracy. More delicate studies are required, for given classes of situations, to sort out the pros and potential cons of the various methods.

In this section we are content to illustrate just some of the methods for two situations where there is a clear ‘gold standard’, that is, a best confidence distribution that the different approximation methods aim at matching. The two situations we have selected for examination share the feature that the gold standard in question is actually hard to get at, making it particularly relevant to study how well the more automatic recipes of this chapter actually pan out. Further illustrations and details may be found in the Exercises section. We also offer a few concluding remarks in the following section, pertaining to general comparison issues.
7.8 Approximations to the gold standard in two test cases

Mean-to-variance ratio

For a random sample \( Y_1, \ldots, Y_n \) from the normal \( \mathcal{N}(\mu, \sigma^2) \), consider the mean-to-variance ratio parameter \( \lambda = \mu / \sigma^2 \). Because the normal density may be written

\[
\phi\left( \frac{y - \mu}{\sigma} \right) = \frac{1}{\sigma} \exp\left( -\frac{(y - \mu)^2}{2\sigma^2} \right),
\]

the likelihood function is

\[
s^{-n} \exp\left( \lambda \sum_{i=1}^{n} y_i - \frac{1}{2} \sum_{i=1}^{n} y_i^2 + \frac{1}{2} n \mu^2 / \sigma^2 \right).
\]

Using the optimality theory of Section 5.5, therefore, there is a unique optimal confidence distribution for \( \lambda \), of the form

\[
C_\alpha^* (\lambda) = \text{P}_\lambda \{ U \geq u_{\text{obs}} \mid W = w_{\text{obs}} \},
\]

where \( U = \sqrt{n} \bar{Y} \) and \( W = \sum_{i=1}^{n} Y_i^2 \).

To make this operational we need to find the conditional distribution involved. To this end write \( V = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = W - U^2 \). The distribution of \( (U, V) \) is

\[
\phi\left( \frac{u - \sqrt{n} \mu}{\sigma} \right) \frac{1}{\sigma} g_m\left( \frac{v}{\sigma^2} \right) \frac{1}{\sigma^2},
\]

where \( g_m \) is the density of the \( \chi_m^2 \), and \( m = n - 1 \). Hence \( (U, V) \) has density

\[
\phi\left( \frac{u - \sqrt{n} \mu}{\sigma} \right) \frac{1}{\sigma} g_m\left( \frac{w - u^2}{\sigma^2} \right) \frac{1}{\sigma^2} \propto \exp(\sqrt{n} \lambda u) (w - u^2)^{m/2-1} \exp(-\frac{1}{2} w / \sigma^2),
\]

from which it follows that \( U \mid w_{\text{obs}} \) has density

\[
h(u \mid w_{\text{obs}}) \propto \exp(\sqrt{n} \lambda u) (w_{\text{obs}} - u^2)^{m/2-1}
\]

for \( |u| \leq w_{\text{obs}}^{1/2} \). Hence

\[
C_\alpha^* (\lambda) = \frac{\int_{u_{\text{obs}}}^{\sqrt{n} \lambda w_{\text{obs}}} \exp(\sqrt{n} \lambda u) (w_{\text{obs}} - u^2)^{m/2-1} \, du}{\int_{-w_{\text{obs}}^{1/2}}^{\sqrt{n} \lambda w_{\text{obs}}} \exp(\sqrt{n} \lambda u) (w_{\text{obs}} - u^2)^{m/2-1} \, du} = \frac{\int_{-1}^{1} \exp(\sqrt{n} \lambda w_{\text{obs}}^{1/2} x) (1 - x^2)^{m/2-1} \, dx}{\int_{-1}^{1} \exp(\sqrt{n} \lambda w_{\text{obs}}^{1/2} x) (1 - x^2)^{m/2-1} \, dx}.
\] (7.22)

Arriving at this formula took some effort, as does its practical computation via numerical integration, so also from this perspective it is of interest to see how various different more automatic approximation strategies work.

Among the simpler of these approximations is the direct first-order one based on the normal approximation for the distribution of the maximum likelihood estimator \( \hat{\lambda} = \mu / \bar{Y}^2 \). We may show that \( \hat{\lambda} \) is approximately a \( \mathcal{N}(\lambda, \kappa^2 / n) \), with \( \kappa^2 = (1 / \sigma^2)(1 + 2 \mu^2 / \sigma^2) \); see Exercise 7.2. This gives rise to the approximation \( \Phi(\sqrt{n} (\lambda - \hat{\lambda}) / \kappa) \). This works for moderate to large \( n \), but not well for smaller sample sizes, also because it turns out that \( \hat{\lambda} \) has a
noticeable positive bias; in Exercise 7.2 we learn that $E \hat{\lambda} = \lambda n/(n - 3)$, which, for example, means 60% overshooting in the illustration that follows.

Usually the deviance curve yields better approximations. Via details placed in Exercise 7.2, involving the profile log-likelihood $\ell_n(\mu, \sigma) = \max\{\ell_n(\mu, \sigma) : \mu/\sigma^2 = \lambda\}$, one may show that $2(\ell_{n, prof}(\hat{\lambda}) - \ell_{n, prof}(\lambda))$ can be expressed as

$$D_n(\lambda) = n(-1 - \log \hat{\sigma}^2 + \log \hat{x} + w_n/\hat{x} - 2\lambda \hat{\mu} + \lambda^2 \hat{x}),$$

Here $\hat{x} = \hat{\sigma}^2(\lambda)$ is the maximiser of $\ell_n(\mu, \sigma)$ over $\sigma^2$ for fixed value of $\lambda$,

$$\hat{x} = \hat{\sigma}^2(\lambda) = (-1 + (1 + 4\lambda^2w_n)^{1/2})/(2\lambda^2).$$

The direct and Bartlett corrected confidence curves are then

$$cc(\lambda) = \Gamma_1(D_n(\lambda)) \quad \text{and} \quad cc_{\text{bart}}(\lambda) = \Gamma(D_n(\lambda)/(1 + \varepsilon)), \quad (7.23)$$

with $1 + \varepsilon$ the mean of $D_n(\lambda)$, as discussed in Section 7.4. Note that the deviance may be computed via general numerical optimisation algorithms, that is, without knowing or using the precise formula for $D_n(\lambda)$, but when it comes to the Bartlett correction, where $1 + \varepsilon$ needs to be found via simulations, such an explicit formula is of course helpful. The same is a fortiori true when one wishes to compare performance of different approximation schemes in a simulation study, as might be a good learning exercise for the present case of the normal mean-to-variance ratio parameter.

**Example 7.7 Mean-to-variance ratio for a balancing test**

How difficult is it to maintain your balance while concentrating, and will it be more difficult as you grow older? Teasdale et al. (1993) examined aspects of this question, exposing younger and elderly subjects to a certain balancing test. For this illustration we shall consider the ‘sway values’ 25, 21, 17, 15, 14, 14, 22, 17 (measured in millimetres), pertaining to eight younger persons going through the balancing test in question; the $\lambda$ analysis we briefly report on here may then be used to help understand in which ways the balancing skills of the elderly differ from those of younger people. The maximum likelihood estimate is $\hat{\lambda} = 1.241$ (and is biased upwards, as mentioned earlier; see Exercise 7.2).

The optimal curve to match is that of $cc(\lambda) = |2C(\lambda) - 1|$ obtained from (7.22). Figure 7.7 displays this curve (using numerical integration), along with some of the many approximations available. The simplest is $\Phi(\sqrt{n}\hat{\lambda} - \hat{\lambda}/\hat{\kappa})$, from the large-sample approximation to a normal with variance $\kappa^2 = (1 + 2\mu^2/\sigma^2)(1/\sigma^2)$. It does not work well here, for $n = 8$, and is not displayed in the figure. We do, however, portray those of (7.23). These are the two seen with median confidence point equal to $\hat{\lambda} = 1.241$, and where Bartletting helps correcting the shape, but not its main position. The method based on $t$-bootstrapping and $t^* = \sqrt{n}(\hat{\lambda}^* - \hat{\lambda})/\hat{\kappa}^*$ fares rather better.

**The normal correlation coefficient**

We return to the normal correlation coefficient $\rho$ and its estimator $r_n$ of (7.9); cf. the earlier discussion in Example 7.2. Here there is a canonical but somewhat complicated confidence
7.8 Approximations to the gold standard in two test cases

Figure 7.7 Exact optimal confidence curve for the $\lambda = \mu / \sigma^2$ parameter (solid line), along with three approximations, for the $n = 8$ data points of Example 7.7. The one with median estimate close to the optimal one is from t-bootstrapping. The two landing to the right are those using the deviance and Bartlett corrected deviance.

The density $f_n(r, \rho)$ for $r_n$ in question is seen to depend on $\rho$ only, not on the other four parameters of the binormal distribution, as $r_n$ is invariant with respect to location and scale changes; cf. details in Exercise 7.4. The point is now that the density $f_n(r, \rho)$ is nonstandard and a bit awkward to implement, hence making it relevant and interesting to see the extent to which $C^*(\rho)$ can be easily approximated via the recipes of this chapter. Indeed, see, for example, Anderson (1958, chapter 4),

$$f_n(r, \rho) = \frac{(1 - \rho^2)^{(n-1)/2}(1 - r^2)^{(n-4)/2}}{(1 - \rho r)^{(n-3)/2}} {}_2F_1\left(\frac{1}{2}, \frac{1}{2}, n - \frac{1}{2}, \frac{1}{2}(1 + \rho r)\right),$$

with $a_n = (2\pi)^{-1/2}(n-2)\Gamma(n-1)/\Gamma(n-1/2)$, and where ${}_2F_1$ is the so-called hypergeometric function of the regular kind,

$${}_2F_1(a, b, c, z) = \sum_{m=0}^{\infty} \frac{a \cdot b(m) \cdot z^m}{c(m) \cdot m!},$$

with $a(m) = a(a+1) \cdots (a+m-1) = \Gamma(a+m)/\Gamma(a)$.

To find the deviance function, with further Bartletting, we need to work with the profile log-likelihood function for $\rho$. The five-parameter binormal has a mean vector $\xi = (\xi_1, \xi_2)$
Improved approximations for confidence distributions

and a two by two covariance matrix \( \Sigma \), with diagonal \((\sigma_1^2, \sigma_2^2)\) and off-diagonal element \(\rho \sigma_1 \sigma_2\). Its log-density is

\[
-\frac{1}{2} \log |\Sigma| - \frac{1}{2} (z - \xi)^\top \Sigma^{-1} (z - \xi) - \log(2\pi),
\]

leading to a log-likelihood of the form

\[
\ell_n(\xi_1, \xi_2, \sigma_1, \sigma_2, \rho) = -\frac{1}{2} n [\log |\Sigma| + \text{Tr}(\Sigma^{-1} S_n(\xi))] .
\]

Here

\[
S_n(\xi) = n^{-1} \sum_{i=1}^n (Z_i - \xi)(Z_i - \xi) = S_n + (\bar{Z} - \xi)(\bar{Z} - \xi)^\top,
\]

in terms of

\[
\bar{Z} = \left( \begin{array}{c} \bar{X} \\ \bar{Y} \end{array} \right) \quad \text{and} \quad S_n = \left( \begin{array}{cc} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{array} \right).
\]

Here \(s_x^2, s_y^2, s_{xy}\) are the averages of \((X_i - \bar{X})^2, (Y_i - \bar{Y})^2, (X_i - \bar{X})(Y_i - \bar{Y})\), so that \(r_n = \frac{s_{xy}}{s_x s_y}\). One finds that \(\ell_n\), regardless of the \(\Sigma\) matrix, is maximised for \(\hat{\xi} = \bar{Z}\), i.e. for \(\hat{\xi}_1 = \bar{X}\) and \(\hat{\xi}_2 = \bar{Y}\). For studying the profile log-likelihood function

\[
\ell_{n, \text{prof}}(\rho) = \max_{\sigma_1, \sigma_2} \ell_n(\hat{\xi}_1, \hat{\xi}_2, \sigma_1, \sigma_2, \rho) = \ell_n(\bar{X}, \bar{Y}, \hat{\sigma}_1(\rho), \hat{\sigma}_2(\rho), \rho),
\]

the next step is to find the maximum likelihood estimators \(\hat{\sigma}_1(\rho)\) and \(\hat{\sigma}_2(\rho)\) that maximise \(\ell_n\) for given value of \(\rho\).

Using

\[
|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho^2) \quad \text{and} \quad \text{Tr}(\Sigma^{-1} S_n) = \frac{1}{1 - \rho^2} \left( \frac{s_x^2}{\sigma_1^2} + \frac{s_y^2}{\sigma_2^2} - 2\rho \frac{s_{xy}}{\sigma_1 \sigma_2} \right),
\]

one finds an expression for \(Q_n(\sigma_1, \sigma_2, \rho) = \log |\Sigma| + \text{Tr}(\Sigma^{-1} S_n)\) which is minimised when its partial derivatives with respect to \(\sigma_1\) and \(\sigma_2\) are equal to zero. These two equations may be solved to give

\[
\hat{\sigma}_1(\rho) = s_x \left( \frac{1 - \rho \hat{\rho}}{1 - \rho^2} \right)^{1/2} \quad \text{and} \quad \hat{\sigma}_2(\rho) = s_y \left( \frac{1 - \rho \hat{\rho}}{1 - \rho^2} \right)^{1/2};
\]

cf. Exercise 7.4, which with a bit of further algebraic work leads to

\[
\ell_{n, \text{prof}}(\rho) = -\frac{1}{2} n \left[ \log s_x^2 + \log s_y^2 + 2 \log(1 - \rho \hat{\rho}) - \log(1 - \rho^2) + 2 \right]
\]

and to the deviance formula

\[
D_n(\rho) = n \left[ 2 \log(1 - \rho \hat{\rho}) - \log(1 - \rho^2) - \log(1 - \hat{\rho}^2) \right]. \quad (7.25)
\]

Note that the log-profile function is indeed maximised over \(\rho\) at the value \(\hat{\rho}\), matching the fact that the observed correlation coefficient (7.9) is the maximum likelihood estimator. It is also noteworthy that the deviance depends on the data precisely via \(\hat{\rho}\), but not of any other aspects of the data.
7.9 Further remarks

Figure 7.8 For \( n = 8 \) and \( \hat{\rho} = 0.534 \), the left panel displays the optimal confidence curve (solid line) with approximations based on deviance (dashed line) and Bartlett modified deviance (dotted line). The right panel gives the optimal confidence density and the approximation based on t-bootstrapping.

Example 7.8 Spurious correlations for babies

Certain pairs of measurements \( x_1 \) and \( x_2 \) were taken on small children in connection with a study at a university hospital in Hong Kong, at age level 3, 12 and 24 months. For this little illustration we single out the \( n = 8 \) children at age 24 months, for which the empirical correlation was \( r_n = 0.534 \); cf. Exercise 7.6, with a fuller story reported on in Example 13.4. For the Bartlett correction method, we find the mean of \( D_n(\rho) \) to be 1.445, when computed at the estimate, and it is not changing much in a reasonable neighbourhood of \( \rho \) values. This leads to \( \Gamma_1(D_n(\rho)/1.445) \) as one of the competitors for confidence distribution accuracy; see Figure 7.8 (left panel). Here Bartletting helps significantly, and the result is closer to the exact curve than with several other approximation schemes. Also \( t \)-bootstrapping does well, as indirectly reported on in the right panel of the figure.

7.9 Further remarks

The delta method and the abc method remove bias by transforming the quantile function of the otherwise biased normal confidence distribution, \( \Phi(\psi - \hat{\psi}) \). The delta method simply corrects the scale of the quantile function, while the abc method applies a shift and a nonlinear scale change to remove bias both due to the non-linearity in \( \psi \) as a function of the basic parameter \( \theta \) as well as the effect on the asymptotic variance when the basic parameter is changed. The \( t \)-bootstrap method would have good theoretical properties in
Improved approximations for confidence distributions

cases where the \( \hat{\psi} \) estimator is a smooth function of sample averages. Theorems delineating suitable second-order correctness aspects of both the abc and the t-bootstrap methods above can be formulated and proved, with necessary assumptions having to do with the quality of approximations involved in (7.14) and (7.16). Methods of proof would, for example, involve Edgeworth or Cornish–Fisher expansion arguments; see, for example, Hall (1992). Such could also be used to add corrections to the delta method (7.1).

There are still other methods of theoretical and practical interest for computing approximate confidence distributions; cf. the broad literature on constructing accurate confidence intervals. One approach would be via analytic approximations to the endpoints of the abc interval, under suitable assumptions; the arguments would be akin to those found in DiCiccio and Efron (1996) and Davison and Hinkley (1997, chapter 5) regarding ‘approximate bootstrap confidence intervals’. Another approach would be via modified profile log-likelihoods, following work by Barndorff-Nielsen and others; see Barndorff-Nielsen and Cox (1994, chapters 6 and 7), Barndorff-Nielsen and Wood (1998) and Tocquet (2001). Clearly more work and further illustrations are needed to better sort out which methods have the best potential for accuracy and transparency in different situations. At any rate the abc method (7.18) appears quite generally useful and precise as part of the combined data for that study.

7.10 Notes on the literature

Theoretical and applied statistics are blessed with the existence of formidably fruitful large-sample theorems, saying and implying in various terms and forms that a great many of the estimators we work with have approximately normal distributions, with variances that may be well estimated from data, as long as sample size dominates parameter dimension. These results are relatives or generalisations of central limit and Lindeberg type theorems, along with linearisation, Taylor expansions, the delta method and Wilks theorems; cf. Chapter 2. This leads to the world of normal and chi-squared approximations, and specifically to the approximate confidence distribution of the two chief types developed and used in in earlier chapters, say \( C_n(\psi) = \Phi((\psi - \hat{\psi}_{obs})/\hat{\kappa}_{obs}) \) and \( cc_n(\psi) = \Gamma_1(D_n(\psi)) \) associated respectively with approximate normality of the estimator and approximate \( \chi^2 \)-ness of the deviance statistic.

Assessing accuracy of these proposed and implied approximations also has a long tradition in probability and mathematical statistics, leading to expressions or approximations for error terms, which in turn may be utilised to find more accurate approximations to distributions and inference methods. Early contributions in such directions include Edgeworth expansions, exhibiting the \( 1/\sqrt{n} \) and \( 1/n \) terms following the cumulative normal when considering the distribution function of the sample mean, and Cornish–Fisher expansions, which work with quantiles expressed via moments and cumulants. See Edgeworth (1909) and Cornish and Fisher (1938) to trace the original ideas and, for example, Barndorff-Nielsen and Cox (1979, 1989, 1994) and Hall (1992) for general overviews, extensions and examples. Bootstrapping may also be used in various ways to assess and repair for nonnormality. Approaches include acceleration and bias correction; see Efron (1987) and Efron and Tibshirani (1993), which has inspired our abc method of Schweder and Hjort (2003), cf. Section 7.6, and t-bootstrap, cf. Hall (1988, 1992), which lends
Exercises

7.1 **Likelihood details with the exponential distribution:** Here we go through various details pertaining to the study of the log-likelihood function when observations \( Y_1, \ldots, Y_n \) are i.i.d. from the exponential density \( \theta \exp(-\theta y) \); cf. Example 7.1.

(a) Under the true value \( \theta_0 \), show that \( 2\theta_0 Y_i \sim \chi^2 \), and use this to show that the maximum likelihood estimator \( \hat{\theta} = \frac{1}{\bar{Y}} \) may be expressed as \( \theta_0 / V_n \), where \( V_n \sim \chi^2 / (2n) \).

(b) Show that \( \sqrt{n}(V_n - 1) \to_d N(0,1) \).

(c) Show that the deviance statistic \( D_n(\theta_0) \) may be expressed as \( D_n = 2n(V_n - 1 - \log V_n) \) (in particular it is a pivot, with a distribution independent of \( \theta_0 \)). Show also that its first-order approximation is \( D_n = n(V_n - 1) \), which converges to \( \chi^2_1 \), as per the general theory.

(d) For a \( \chi^2_a \) with \( a \) degrees of freedom, show that
\[
\begin{align*}
E(\chi^2_a - a) &= 2a, \\
E(\chi^2_a - a)^3 &= 8a, \\
E(\chi^2_a - a)^4 &= 12a^2 + 48a,
\end{align*}
\]
and that these lead to skewness \((8/a)^{1/2}\) and excess kurtosis \(12/a\).

(e) Show that \( E \log(\chi^2_a) = \log 2 + \psi(a) \), where \( \psi(x) = \partial \log \Gamma(x) / \partial x \) is the digamma function. One has \( \psi(x) = \log x - \frac{1}{2} / x - \frac{1}{12} / x^2 + O(1/x^4) \). Show that \( E D_n = 2n[\log n - \psi(n)] = 1 + \frac{1}{e} / n + O(1/n^2) \).
Improved approximations for confidence distributions

(f) Show via simulations for, say, \( n = 6 \) that \( D_\nu/(1 + \frac{1}{6}/n) \) is closer to the \( \chi^2 \) in distribution than is \( D_\nu \).

7.2 Confidence for mean to variance ratio: The following relates to details and calculations for \( \lambda = \mu/\sigma^2 \) inference in the normal family; see Section 7.8.

(a) Show that \( \sqrt{n}(\hat{\lambda} - \hat{\lambda}) \rightarrow_d N(0, \kappa^2) \) with \( \kappa^2 = (1/\sigma^2)(1 + 2\mu^2/\sigma^2) \), and derive from this a normal-based approximation confidence distribution for \( \hat{\lambda} \). Show also that the maximum likelihood estimator is biased upwards, with \( E(\hat{\lambda}^*)/(n/(n-3)) \hat{\lambda} \). Check if using \( \Phi((\lambda - \lambda^*)/\kappa) \) comes closer to the optimal confidence distribution (7.22) with the adjusted estimator \( \hat{\lambda}^* = (1 - 3/n)\hat{\lambda} \) than with the direct maximum likelihood estimator.

(b) Find a formula for the profile log-likelihood \( \ell_{\text{prof}}(\hat{\lambda}) \) by maximising over \((\mu, \sigma)\) with \( \mu/\sigma^2 = \lambda \), and then find a formula for \( \nu_\sigma(\hat{\lambda}) \).

(c) Investigate \( t \)-bootstrapping of \( \sqrt{n}(\hat{\lambda} - \lambda)/\kappa \) as a method for approximating the gold standard (7.22) for small sample sizes.

7.3 Pivot tuning for the correlation coefficient: For binormal data, the empirical correlation coefficient \( \hat{\rho}_n \) has the property that \( Z_n = \sqrt{n}(\hat{\rho}_n - \rho) \) tends to a normal distribution \( N(0, \kappa^2) \); see Exercise A.2 in the Appendix.

(a) Suppose you do not remember or have time to find the formula for \( \kappa \). Set up a simple simulation experiment to read off the standard deviation of \( Z_n \), and use regression to fit this to, say, \( a_0 + a_1\rho + a_2\rho^2 + a_3\rho^3 + a_4\rho^4 \). Deduce that \( \kappa \) must be equal to or close to \( 1 - \rho^2 \).

(b) Then check Fisher’s variance stabilising zeta transform, \( \zeta = h(\rho) = \frac{1}{4} \log((1 + \rho)/(1 - \rho)) \), and verify that \( \sqrt{n}(\hat{\zeta} - \zeta) \rightarrow_d N(0, 1) \). Show also that the variance of \( \hat{\zeta} = h(\rho_n) \) is better approximated by \( 1/(n - 3) \) than with \( 1/n \). Derive the consequent modified confidence distribution for \( \rho \).

7.4 The binormal correlation coefficient: Here we work through some of the details pertaining to the binormal correlation coefficient situation studied in Example 7.2.

(a) Show that the maximum likelihood estimators of \( \zeta_1, \zeta_2 \) indeed are equal to \( \bar{X}, \bar{Y} \).

(b) Note initially that the Pearson correlation coefficient \( \hat{\rho} \) of (7.9) is invariant under location and scale changes, and hence has a distribution with density say \( f_\rho(r, \rho) \) that depends on the parameters only via \( \rho \). To see this clearly, write \( X_i = \xi_i + \sigma_1 \delta_i \) and \( Y_i = \xi_i + \sigma_2 \epsilon_i \), where \((\delta_i, \epsilon_i)^T \sim N(0, \Sigma_0) \) and \( \Sigma_0 \) has unit variances and correlation \( \rho \). Then show that \( \hat{\rho} \) based on the \((X_i, Y_i) \) is identical to \( \hat{\rho} \) based on the \((\xi_i, \epsilon_i) \).

(c) Consider the function \( Q_\rho(\Sigma) = \log|\Sigma| + \text{Tr}(\Sigma^{-1} S_0) \) for fixed \( \rho \). Solve the two equations corresponding to setting the partial derivatives of \( Q_\rho \) with respect to \( \sigma_1 \) and \( \sigma_2 \) to zero, to find

\[
\hat{\sigma}_1(\rho) = s_x \left( 1 - \rho \hat{\rho} \right)^{1/2} \quad \text{and} \quad \hat{\sigma}_2(\rho) = s_y \left( 1 - \rho \hat{\rho} \right)^{1/2}.
\]

(d) Show also that

\[
\min_{\sigma_1, \sigma_2} Q_\rho(\Sigma) = \log|\hat{\Sigma}(\rho)| + \text{Tr}(\hat{\Sigma}(\rho)^{-1} S_0) = 2 \log s_x + 2 \log s_y + 2 \log(1 - \rho \hat{\rho}) - \log(1 - \rho^2) + 2,
\]

and that this leads to the deviance formula (7.25).
7.5 More correlation: We have seen in Example 7.2 that for the usual empirical correlation coefficient \( r_n \), and under binormal circumstances, we have \( \sqrt{n}(r_n - \rho) \to_d N(0, \kappa^2) \) with \( \kappa = 1 - \rho^2 \). Consider now the specialised model where means and variances are known, which then may be taken to be zero and one.

(a) Write down the log-density and the score function, that is, the derivative of the log-density. Find the Fisher information \( J(\rho) \), and deduce that \( \sqrt{n}(\hat{\rho} - \rho) \to_d N(0, \kappa_0^2) \), where \( \kappa_0 = (1 - \rho^2)/(1 + \rho^2)^{1/2} \). Discuss the extent to which precision is increased as a consequence of knowing the means and variances.

(b) Find a formula for the deviance statistic \( D_n(\rho) \). For \( n = 10 \), use simulation to find the mean and variance of \( D_n(\rho) \) numerically, as a function of \( \rho \).

(c) For \( n = 10 \), investigate briefly how the Bartlett correction method works, compared to the simpler direct deviance method for constructing a confidence curve.

7.6 Spurious correlations: In Example 7.8 we computed the optimal confidence distribution for a binormal correlation coefficient \( \rho \), for sample size \( n = 8 \) and observed correlation \( r = 0.534 \), along with a few approximations. Do this also for the cases \( n = 7, r = 0.181 \) and \( n = 8, r = 0.202 \). Also attempt to try out other correction methods from this chapter, like that of the median-bias correction. (For this illustration and its context, note Remark 13.1.)

7.7 Bolt from heaven: From the CLP book website, access the sprint data on all \( n = 195 \) 100-m runs over the eight years 2000 to 2007, and fit the extreme-value model (7.11). Verify the \( \Gamma_1 = r_n \) along with a few approximations. Do this also for the cases \( \rho \) and variance of \( D_n(\rho) \).

7.8 Magic at work: Consider the setup of Example 5.10, involving data \( y_1, \ldots, y_n \) from the gamma distribution with parameters \( (a, b) \), and with attention on the shape parameter \( a \). The optimal confidence distribution \( C^*(a) \) is given in that example; the task here is to investigate some of its approximations. A concrete test case is that of having only five data points, these being \( 0.20, 0.45, 0.78, 1.28, 2.28 \), as in Brazzale et al. (2007, p. 13).

(a) Show that the profile log-likelihood for \( a \) may be expressed as

\[
\ell_{a,\text{prof}}(a) = n[a \log(a/\bar{y}) - \log \Gamma(a) + (a - 1)\bar{u} - a],
\]

where \( \bar{y} \) and \( \bar{u} \) are the sample means and variances, respectively.
Improved approximations for confidence distributions

with $\bar{y}$ and $\bar{y}$ the averages of $\log y_i$ and $y_i$, respectively. Investigate the distribution of the deviance statistic $D_n(a) = 2\{\ell_{\text{prof}}(G) - \ell_{\text{prof}}(a)\}$ via simulations, assessing in particular how close its mean and variance are to those of the $\chi^2_1$.

(b) For samples of size $n = 5, 10, 25$, perhaps sampled from $(a_0, b_0) = (1.7, 1.7)$ to match the type of the preceding simple data example, compare the optimal confidence distribution $C^*(a)$ with $\Gamma_1(D_n(a))$ and the Bartlett-corrected $\Gamma_1(D_n(a)/(1 + \epsilon_n))$, where $1 + \epsilon_n$ is the mean of $D_n(a)$ under the model.
Exponential families and generalised linear models

Here we apply methods and machinery developed in earlier chapters to the broad exponential family of models, where the log-density takes a linear form. We also examine regression models of the generalised linear model type, in fact along with some further generalisations of these. Some of our optimality results developed in Chapter 5 concerning certain confidence distributions apply particularly fruitfully to classes of these log-linear family models, via computational techniques or approximations. Applications of our approaches include also somewhat nonstandard situations, including the Strauss model and Ising–Potts models for point patterns and images. We also discuss extensions to what we call generalised linear-linear models, where covariates are seen as potentially influential on more than one parameter at a time in a vehicle model.

8.1 The exponential family

Chapter 2 provided the general background and apparatus for handling inference for very broad classes of parametric models, highlighting the general properties and usefulness of likelihood inference, normal and chi-squared approximations, and so on. This chapter looks more closely at certain subclasses of distributions where the general theory a fortiori continues to hold but where one often can do more, for example, in terms of accurate and efficient inference. These models in question are essentially those in which the log-densities are linear combinations of parameters. Examining such cases leads to the so-called exponential family for independent observations and by extension to generalised linear models for regression situations. We shall in particular see later that the optimality theorems for certain constructions of confidence distributions developed in Chapter 5 apply to these models.

The exponential family formula for the probability density, considered in Section 5.5, takes the form

\[ f(y, \theta) = \exp\{ h_1(\theta) T_1(y) + \cdots + h_p(\theta) T_p(y) + m(y) - k(\theta) \}, \]  

where \( \theta = (\theta_1, \ldots, \theta_p) \) is some \( p \)-dimensional parameter and \( T_1, \ldots, T_p \) are functions of \( y \). As the notation indicates, the \( m \) function must be free of model parameters, and the \( k(\theta) \) is the normalising function required to make \( f(y, \theta) \) integrate to 1. Also, the sample space for \( y \) must be the same for all parameters. The \( y \) here may be multidimensional.
Two simple examples are as follows. When \( Y \) is binomial \( (n, p) \), then its point density may be written
\[
f(y; p) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{(p/(1-p))^y}{y!} \left( \frac{p}{1-p} \right)^n = \exp \left\{ y \log(p) + \log \binom{n}{y} - n \log(1-p) \right\},
\]
with \( h(p) = \log(p/(1-p)) \). This is of the required form, with \( T(y) = y \). Second, when \( Y \) is a normal \((\mu, \sigma^2)\), then its density is
\[
f(y; \mu, \sigma) = \frac{1}{(2\pi)\sigma^2} \exp \left\{ - \frac{1}{2} \frac{(y-\mu)^2}{\sigma^2} \right\}
= \exp \left\{ \frac{\mu}{\sigma^2} - \frac{1}{2} \frac{y^2}{\sigma^2} - \frac{1}{2} \frac{\mu^2}{\sigma^2} - \log \sigma - \frac{1}{2} \log(2\pi) \right\}.
\]
This is again of the form (8.1), with \( h_1(\mu, \sigma) = \mu/\sigma^2 \) and \( h_2(\mu, \sigma) = -\frac{1}{2}(1/\sigma^2) \), and with \( T_1(y) = y \) and \( T_2(y) = y^2 \). In fact a long list of standard distributions belong to the exponential class; further examples and illustrations are found in Exercise 8.1.

Suppose now that independent observations \( Y_1, \ldots, Y_n \) are available from a model of form (8.1). Then their joint density takes the form
\[
\exp \left\{ nh_1(\theta_1) T_1 + \cdots + nh_p(\theta_p) T_p + \sum_{i=1}^n m(y_i) - nk(\theta) \right\},
\]
which is of the same general form. In particular, the averages
\[
\bar{T}_1 = n^{-1} \sum_{i=1}^n T_1(y_i), \ldots, \bar{T}_p = n^{-1} \sum_{i=1}^n T_p(y_i)
\]
form a sufficient set of statistics, of the same low dimension regardless of sample size; see Section 5.4.

For a model of type (8.1) it is sometimes convenient to reparameterise from \( \theta_1, \ldots, \theta_p \) to the so-called canonical parameters \( \eta_1 = h_1(\theta), \ldots, \eta_p = h_p(\theta) \), defined up to multiplicative constants. The log-likelihood function is then of the simpler form
\[
\ell_s(\eta) = n[\eta_1 \bar{T}_1 + \cdots + \eta_p \bar{T}_p - k_0(\eta)],
\]  
(8.2)
say. An important fact is that \( k_0(\eta) \) is convex, making the log-likelihood function concave. For this and other useful facts related to the generalised linear models, see Exercise 8.2. The original parameter is identified if the map \( \theta \to \eta \) is one-to-one.

Thus the canonical parameter for the binomial model is \( \eta = \log(p/(1-p)) \), which is the inverse of the logistic transform \( p = \exp(\eta)/(1 + \exp(\eta)) \). Also, for the normal model, the canonical parameters are \( \eta_1 = \mu/\sigma^2 \) and \( \eta_2 = 1/\sigma^2 \). The point of this reparameterisation is partly one of mathematical convenience, where certain theorems apply more easily for the \( \eta_j \) than for the original \( \theta_j \); also, it turns out to be easier for some purposes to build regression models where the \( \eta_j \) rather than the \( \theta_j \) are expressed as linear combinations of
covariates and regression coefficients. This is the topic of generalised linear models (see Sections 8.4 and 8.5), but we start our investigations in simpler models without covariates in Sections 8.2 and 8.3.

In the following sections we shall work with the apparatus of confidence distributions and so forth, for certain types of exponential family models. For various further details, discussion and illustrations pertaining directly to the exponential family class; see, for example, Barndorff-Nielsen and Cox (1994), Bickel and Doksum (2001) and Young and Smith (2005).

8.2 Applications

The exponential family lends itself to the use of optimality results of Chapter 5. We demonstrate this in the text that follows, first for a few models without covariates and then for regression models.

In the typical applications of Theorem 5.11 there is a sample of independent observations \(Y_1, \ldots, Y_n\) from the exponential family in question, say of the form

\[
f(y, \psi, \chi) = \exp\left\{ \psi S(y) + \chi_1 A_1(y) + \cdots + \chi_p A_p(y) - k(\psi, \chi_1, \ldots, \chi_p) + m(y) \right\}
\]

of parameter dimension \(p + 1\), and with \(\psi\) the focus parameter. In such a case the likelihood for the full dataset is proportional to

\[
\exp\left\{ \psi S_n + \chi_1 A_{n,1} + \cdots + \chi_p A_{n,p} - nk(\psi, \chi_1, \ldots, \chi_p) \right\}, \quad (8.3)
\]

in terms of the vector of sufficient statistics

\[
(S_n, A_{n,1}, \ldots, A_{n,p}) = \left( \sum_{i=1}^{n} S(Y_i), \sum_{i=1}^{n} A_1(Y_i), \ldots, \sum_{i=1}^{n} A_p(Y_i) \right).
\]

An important result about exponential families is that the distribution of this vector follows the same form, that is, has a density proportional to the expression of (8.3); see, for example, Young and Smith (2005, chapter 5). Thus Theorem 5.11 applies, yielding the optimal confidence distribution

\[
C_n(\psi) = P_{\psi}\{ S_n \geq S_{n,\text{obs}} \mid A_{n,1} = a_{n,1,\text{obs}}, \ldots, A_{n,p} = a_{n,p,\text{obs}} \}. \quad (8.4)
\]

Example 8.1 Odds ratio

Consider two independent binomials, say

\[
Y_0 \sim \text{Bin}(m_0, p_0) \quad \text{and} \quad Y_1 \sim \text{Bin}(m_1, p_1),
\]

perhaps associated with ‘control’ and ‘treatment’, respectively. The two probabilities may be compared in various ways, including estimation of and inference for \(p_1 - p_0\), \(p_1 / p_0\), arcsin(\(\sqrt{p_1}\)) − arcsin(\(\sqrt{p_0}\)), and so on. In medical and biostatistical literature, the customary way is via the logistic transforms

\[
p_0 = \frac{\exp(\theta)}{1 + \exp(\theta)} \quad \text{and} \quad p_1 = \frac{\exp(\theta + \psi)}{1 + \exp(\theta + \psi)}.
\]
Exponential families and generalised linear models

with \( \psi \) the log-odds difference. The natural focus is then the odds ratio

\[
OR = \rho = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \exp(\psi),
\]

for which we shall now find the optimal confidence distribution.

The joint distribution of \((Y_0, Y_1)\) may be written

\[
L = \binom{m_0}{y_0} p_0^{y_0} (1-p_0)^{m_0-y_0} \binom{m_1}{y_1} p_1^{y_1} (1-p_1)^{m_1-y_1}
\]

which in its turn leads to the log-likelihood function

\[
\ell = y_1 \log \rho + z \theta - m_0 \log(1 - \exp(\theta)) - m_1 \log(1 + \exp(\theta) \rho),
\]

where \( Z = Y_0 + Y_1 \). By Theorem 5.11 there is an optimal confidence distribution for \( \psi = \log \rho \) and hence by transformation for \( \rho \):

\[
C(\rho) = P_{\rho} \{ Y_1 > y_{1, obs} | z \} + \frac{1}{2} P_{\rho} \{ Y_1 = y_{1, obs} | z \}.
\]

To evaluate this we need the distribution of \( Y_1 \) given \( Z = z \), and find

\[
g(y_1 | z) = \binom{m_0}{z-y_1} \binom{m_1}{y_1} \rho^{y_1} / \sum_{y_1'=0}^z \binom{m_0}{z-y_1'} \binom{m_1}{y_1'} \rho^{y_1'}
\]

for \( y_1 = 0, 1, \ldots, \min(z, m_1) \), the so-called noncentral hypergeometric distribution. The case of equal probabilities is \( \rho = 1 \), which gives the ordinary hypergeometric distribution for \( Y_1 \) given \( z \). We may now compute and display \( C(\rho) \) of (8.6), using the \( g(y_1 | z) \) probabilities for each \( \rho \) over a grid.

Figure 8.1 displays the optimal confidence curve \( cc(\rho) = |1 - 2 C(\rho)| \) (solid line) for a case of \((y_0, y_1) = (1, 2)\) with \((m_0, m_1) = (15, 15)\), along with an approximation derived from the often-used

\[
\hat{\psi} \sim N(\log \rho, \hat{\kappa}^2)
\]

where \( \hat{\kappa}^2 = \frac{1}{Y_0} + \frac{1}{m_0 - Y_0} + \frac{1}{1} + \frac{1}{m_1 - Y_1} \).

For this formula and some other details, see Exercise 8.4. For an extended analysis of multiple two-by-two tables with a common odds ratio, consider the billion-dollar story of Section 14.7.

In Examples 5.8 and 5.9 and for a somewhat slim catalogue of other exponential family situations, like the one just handled, one manages to find the required conditional distribution involved in (8.4) explicitly. Typically this is not possible, however, and one needs to resort to suitable approximation tricks. One route is that of large-sample approximations, a theme that was pursued in Chapter 7, and of which the three general recipes described in Section 3.4 may be seen to be special cases. Another route is via stochastic simulation, where one attempts to sample realisations of \( S_n \) from the appropriate
8.2 Applications

Figure 8.1 Optimal confidence curve (solid line) for the odds ratio (8.5), along with a normal approximation described in Example 8.1 (dashed line), with observed \((y_0, y_1) = (1,2)\) and binomial sample sizes \((m_0, m_1) = (15,15)\). The vertical line at \(\rho = 1\) corresponds to \(p_0 = p_1\).

distribution conditional on the observed \(A_{n,j} = a_{n,j,\text{obs}}\). If one succeeds in running such a scheme, one may approximate the optimal confidence distribution as

\[
C^*_n(\psi) = \frac{1}{B} \sum_{j=1}^B I\{S^*_n, j \geq s_{n,\text{obs}}\},
\]

say, where the \(S^*_n, j\) are generated realisations from the appropriate conditional distribution, and with a suitably large value \(B\) of such realisations. To achieve monotonicity in the final \(C^*_n(\psi)\) to be displayed, the use of a common seed in the simulations yielding (8.7) may work. Otherwise, the curve may be suitably postmonotonised, via methods of isotonic regression (see Barlow et al., 1972), which is an easy task in \(\text{R}\) via the \texttt{isoreg} algorithm; cf. De Leeuw et al. (2009). If smooth output is deemed important one may also achieve this, in various ways, for example, by fitting the simulation output of (8.7) to say a mixture of two Beta cumulatives.

Example 8.2 Optimal confidence distributions for Beta parameters

Health Assessment Questionnaires (HAQ) data pertaining to the health and well-being of patients are routinely collected in many hospitals, in particular in connection with treatment and research on rheumatism. The data to be considered here are of a modified HAQ type that are recorded as near continuous measurements on a scale equal to the interval \([1,4]\). Our illustration concerns a sample of \(n = 799\) such modified HAQ data taken from an investigation at Ullevål University Hospital at Oslo; see Claeskens and Hjort (2008, example 3.7) for further background. Figure 8.2 displays a histogram and two densities for these data, when normalised to \([0,1]\), from fitting two different parametric models. The first
Figure 8.2 Histogram of \( n = 799 \) MHAQ observations (Example 8.2), normalised to lie inside the unit interval, along with densities fitted via maximum likelihood from the Beta density \( f(y, \hat{a}, \hat{b}) = f(y, 1.460, 4.349) \) (solid curve), and a three-parameter extension discussed in Example 8.7 (dashed line).

(solid line), to be discussed now, is the fitted Beta density

\[
f(y, a, b) = \frac{\Gamma(a + b)}{\Gamma(a) \Gamma(b)} y^{a-1} (1 - y)^{b-1}
\]

after the linear transformation that brings the data into the unit interval; the second model (corresponding to the dotted line in the figure) is discussed later, in Example 8.7. The Beta density fit is not perfect, but we shall nevertheless take up the challenge of providing the optimal confidence distributions for \( a \) and \( b \), assuming that the Beta model is adequate. The log-likelihood \( \ell_n(a, b) \) takes the form

\[
(a - 1) \sum_{i=1}^{n} U(Y_i) + (b - 1) \sum_{i=1}^{n} V(Y_i) + n \log \Gamma(a + b) - n \log \Gamma(a) - n \log \Gamma(b),
\]

with \( U(y) = \log y \) and \( V(y) = \log(1 - y) \), and observed sufficient statistics

\[
U_n = \sum_{i=1}^{n} U(Y_i) = u_{n, obs} = -1336.002,
\]

\[
V_n = \sum_{i=1}^{n} V(Y_i) = v_{n, obs} = -255.978.
\]

From Theorem 5.11 it is clear that the optimal confidence distributions are

\[
C_{n,1}(a) = P_a \{ U_n \geq -1336.002 \mid V_n = -255.978 \},
\]

\[
C_{n,2}(b) = P_b \{ V_n \geq -255.978 \mid U_n = -1336.002 \}.
\]
8.2 Applications

The trouble is, however, that these conditional distributions are complicated and cannot be computed easily. We resort to simulating the distribution of \( U_n | (V_n = v_{n,\text{obs}}) \) for each \( a \), and of \( V_n | (U_n = u_{n,\text{obs}}) \) for each \( b \). These are also difficult tasks, for which we shall employ a version of a method devised and discussed in Lindqvist and Taraldsen (2006, 2007). For a fixed value \( a_0 \) of \( a \), \( V_n \) is sufficient for \( b \), so the distribution of \( (Y_1, \ldots, Y_n) \) given \( V_n = v_{n,\text{obs}} \) does not depend on the free parameter \( b \) (only on the fixed \( a_0 \)). Write \( y_i = F_{a_0}^{-1}(w_i) \), where the \( w_i \) are i.i.d. and uniform. These form a sample from the Beta \( \{a_0, b\} \), if \( b \) is given. The idea is to choose \( b \) so as to match the requirement \( V_n = \sum_{i=1}^n \log\{1 - F_{a_0}^{-1}(w_i)\} = v_{n,\text{obs}} \). Solving this equation (which needs to be carried out numerically) leads to \( b^* = b^*(v_{n,\text{obs}}) \), say. In turn this yields a sample of

\[
y^*_i = y^*_i(v_{n,\text{obs}}) = F_{a_0}^{-1}(w_i)
\]

which indeed matches the \( V_n = v_{n,\text{obs}} \) requirement. Lindqvist and Taraldsen (2006) prove that this ‘solving the equation’ trick leads to a correct solution; the \( (y^*_1, \ldots, y^*_n) \) really come from the required conditional distribution of \( (Y_1, \ldots, Y_n) \) given \( V_n = v_{n,\text{obs}} \). This makes it possible for us to simulate say realisations of \( U^*_n = \sum_{i=1}^n U(y^*_i) \) from the correct distribution of \( U_n | (V_n = v_{n,\text{obs}}) \), for the given \( a_0 \), and to read off \( C_{n,1}(a_0) \) in the manner of (8.7). A similar recipe may be written up and executed for \( C_{n,2}(b_0) \), for each fixed \( b_0 \). This procedure led to Figure 8.3, displaying these simulation-based evaluations of the exact optimal confidence distributions, employing \( B = 1000 \) simulations for each fixed parameter value and exposing the resulting \( C_n(a) \) and \( C^*_n(b) \) curves to isotonic regression.

Figure 8.3 Optimal confidence distributions for \( a \) and \( b \) of the Beta distribution, for the \( n = 799 \) MHAQ observations of Figure 8.2, computed via simulation of the conditional distributions of respectively \( U_n | (V_n = -255.978) \) and \( V_n | (U_n = -1336.002) \). We have used (8.7) with \( B = 1000 \) simulations for each of the parameter points \( a \) and \( b \) at which \( C_{n,1}(a) \) and \( C^*_n(b) \) were computed.
We point out here that carrying out confidence analysis using log-likelihood profiling and the chi-squared transform on the deviances is considerably simpler, regarding programming and computer time, and actually gives curves very similar to these more expensive optimal ones in Figure 8.3. This is partly due to the high sample size, but for smaller \( n \) the optimal curves will not be so well approximated by the deviance-based ones.

**Confidence inference in a multinomial recapture model**

Assuming captures to be stochastically independent between occasions and letting all individuals having the same capture probability \( p_t \) on occasion \( t \), we have the multinomial multiple-capture model of Darroch (1958). With \( A_t \) the number of unique captures on occasion \( t \), and \( Y \) the number of unique individuals sampled over say \( k \) occasions, the likelihood is

\[
L(N, p_1, \ldots, p_k) \propto \binom{N}{Y} \prod_{t=1}^{k} p_t^{A_t} (1 - p_t)^{N - A_t}.
\]

This factors in \( Y \) and \( A = (A_1, \ldots, A_k) \), which emerges as a conditioning statistic in inference for \( N \). For fixed \( N \), \( A \) is in fact sufficient and complete for the nuisance parameter \((p_1, \ldots, p_k)\). The confidence distribution for \( N \) based on \( Y \) in the conditional model given \( A \) is therefore suggested by the arguments that led to Theorem 5.11. With half-correction due to discreteness, the cumulative confidence distribution function is

\[
C(N) = P_N\{Y > y_{\text{obs}} \mid A = a_{\text{obs}}\} + \frac{1}{2} P_N\{Y = y_{\text{obs}} \mid A = a_{\text{obs}}\}.
\]

With \( h(s, n, N, S) \) being the hypergeometric probability of having \( s \) special individuals in a random sample of size \( n \) from a population of size \( N \) with \( S \) special, the conditional probability of \( Y = y \) given \( A = a \) is

\[
\sum_{r_1=0}^{y_1} \sum_{r_2=0}^{y_2-r_1} \cdots \sum_{r_k=0}^{y_k-r_{k-1}} \prod_{t=2}^{k} h(r_t, a_t, N, \sum_{i=1}^{t-1} (a_i - r_i)). \tag{8.8}
\]

The present approach assumes the population to be closed and homogeneous with respect to capturing, which is independent over capturing occasions.

**Example 8.3 Bowhead whales in Alaska**

In the summers and autumns of 1985 and 1986, photographs were taken of bowhead whales north of Alaska (see da Silva et al., 2000; Schweder, 2003). We shall be concerned mainly with the immature component of the population that had natural marks on their bodies. The numbers of identified individuals in photographs taken on each of four sampling occasions were 15, 32, 9, 11, respectively, and with \( Y = 61 \) unique individuals in the pooled set.

The confidence distribution for number of immature whales is

\[
C(N) = P_N\{Y > 61 \mid A = (15,32,9,11)\} + \frac{1}{2} P_N\{Y = 61 \mid A = (15,32,9,11)\},
\]

calculated according to the conditional distribution (8.8). The conditional probability provides a conditional likelihood for \( N \), viz. \( L(N) = P_N\{Y = 61\} \). The likelihood happens to
be very close to the likelihood calculated from \( C(N) \), via the inverse chi-squared transform. This agreement is due to the underlying conditional pivot. With \( \hat{N} \) being the conditional maximum likelihood estimate and \( Z \) the standard normal variate, the pivot is approximately \( \hat{N}^{-1/2} - bN^{-1/2} \approx aZ \). The coefficients \( b \) and \( \sigma \) and the power \(-\frac{1}{2}\) are estimated from q–q plots of the confidence distribution associated with (8.8) calculated from the observed data. To an amazing accuracy, we find \( C(N) \approx \Phi(5.134 - 87.307 N^{-1/2}) \) over the relevant range \((120,700)\) for \( N \), with confidence 98%. The natural parameter is \( \mu(N) = 1/N^{1/2} \). Owing to the nonlinearity in the natural parameter, the likelihood is different from the confidence density (taking \( N \) to be continuous); in this case the difference is actually substantial. For some details, see Exercise 8.5. The same picture emerges for mature whales; see Schweder (2003).

### 8.3 A bivariate Poisson model

We have watched women’s handball matches from two grand tournaments – the 33 matches from the Olympic Games held in Athens, Greece, in August 2004, with 10 participating nations, and where Denmark won; and then the 48 matches from the European Championships held in Hungary, in December 2004, with 16 participating nations, and where Norway won. The results of these 81 matches are displayed in Figure 8.4. The distribution of goals scored per match and per team is reasonably well modelled by a Poisson distribution, with an average value of around 27. Our concern here is the spectator’s feeling that matches so often are quite close races, closer than they ought to have been if the number of goals scored by the two teams had been statistically independent. This motivates the construction of a suitable bivariate model, encapsulating such dependence.

Because our intention is to reach statements about the typical state of affairs for top handball matches we need our analyses to avoid being overly influenced by a few special

![Figure 8.4](image-url) Results of 81 handball matches from three top international tournaments. The two matches resulting in 41–18 and 13–33 are viewed as statistical outliers and are removed from the final analyses.
match results, so we choose to eliminate two spectacular matches from further consideration: the Norway versus Slovenia 41–18 match and the Greece versus China 13–33 match. The following analyses pertain to the \( n = 79 \) remaining ‘cleaner’ matches, consisting of results \((Y_{i,1}, Y_{i,2})\) for \( i = 1, \ldots, n \), with a total of 4260 goals (translating to an average of \( 4260/(2 \cdot 79) = 26.92 \) goals per team per match).

There are several ways to test the assumption of Poisson independence. When \( Y \sim \text{Pois} (\lambda) \), \( 2\sqrt{Y} \) is approximately a \( N(2\sqrt{\lambda}, 1) \), at least for \( \lambda \) values of size 15 and more; see Exercise 8.6. Under Poisson independence, therefore, the variables

\[
A_i = \sqrt{2}(\sqrt{Y_{i,1}} + \sqrt{Y_{i,2}}) \quad \text{and} \quad B_i = \sqrt{2}(\sqrt{Y_{i,1}} - \sqrt{Y_{i,2}})
\]

ought to be independent and approximately normal with unit standard deviation (the \( A_i \) with mean \((8\lambda)^{1/2}\) and the \( B_i \) with mean zero). Figure 8.5 is indeed consistent with the independence part, but indicates underdispersion regarding the \( B_i \). The observed sum of squares \( \sum_{i=1}^n B_i^2 = 2 \sum_{i=1}^n (\sqrt{Y_{i,1}} - \sqrt{Y_{i,2}})^2 = 56.171 \) is on the too small side of the \( \chi^2_{79} \) distribution, with a p-value of 0.024. This is the more noticeable in that the \( B_i \) cannot be expected to have fully identical mean values, that is, \( \sum_{i=1}^n B_i^2 \) will also have a portion of undermodelled variability on board.

There are several ways of constructing extensions of the Poisson model to allow for interaction; see, for example, the models and methods developed in Exercise 8.11. Those models may only reflect positive correlations, however, but for various phenomena one also wishes to allow negative dependencies, that is, with independence as an inner point in the parameter space in question. Here we shall consider the following model allowing for dependence,

\[
f_0(y_1, y_2) = k_0(\gamma, \lambda)^{-1} \frac{\lambda^{y_1} \lambda^{y_2}}{y_1! \ y_2!} \exp\{-\gamma (\sqrt{y_1} - \sqrt{y_2})^2\} \quad \text{for} \quad \gamma > 0 \quad \text{and} \quad \lambda > 0 \quad \text{for all} \quad y_1, y_2.
\]
8.3 A bivariate Poisson model

For \( y_1, y_2 = 0, 1, 2, \ldots \), with \( k_0(\gamma, \lambda) \) the appropriate infinite double sum that makes \( f_0 \) a probability density function. Here a positive \( \gamma \) corresponds to outcomes \((Y_1, Y_2)\) being squeezed tighter together with higher probability than under Poisson independence, which corresponds to \( \gamma = 0 \). Also negative values of \( \gamma \) are allowed, leading to ‘repulsion’ with \( Y_1 \) and \( Y_2 \) pushed away from each other with higher probability, as opposed to ‘attraction’. We may conveniently reparametrise the model as

\[
f(y_1, y_2, \gamma, \phi) = k(\gamma, \phi) \frac{\exp\{2\gamma \sqrt{y_1 y_2} + \phi (y_1 + y_2)\}}{y_1! y_2!},
\]

with \( \phi = \log \lambda - \gamma \) and

\[
k(\gamma, \phi) = \sum_{y_1, y_2} \frac{\exp\{2\gamma \sqrt{y_1 y_2} + \phi (y_1 + y_2)\}}{y_1! y_2!}.
\]

We have \( k(0, \phi) = \exp\{2\exp(\phi)\} \) but there is no closed form formula for the normalising constant for nonzero values of \( \gamma \). The log-likelihood function becomes

\[
\ell_n(\gamma, \phi) = n(\gamma \bar{U} + \phi \bar{V} - \log k(\gamma, \phi)),
\]

with

\[
\bar{U}_n = n^{-1} \sum_{i=1}^n 2\sqrt{Y_{i,1} Y_{i,2}} \quad \text{and} \quad \bar{V}_n = n^{-1} \sum_{i=1}^n (Y_{i,1} + Y_{i,2}),
\]

observed here to be respectively 53.5685 and 53.9240. Here we focus on the dependence parameter \( \gamma \), and the profile log-likelihood function is computed as

\[
\ell_{n, \text{prof}}(\gamma) = \max_{\text{all } \phi} \ell_n(\gamma, \phi) = n(\gamma \bar{U} + \hat{\phi}_\gamma \bar{V} - \log k(\gamma, \hat{\phi}_\gamma)),
\]

say, with \( \hat{\phi}_\gamma \) the maximiser for given value of \( \gamma \). This leads to the usual deviance curve and consequent confidence curve

\[
cc_n(\gamma) = \Gamma_1(D_n(\gamma)) \quad \text{with} \quad D_n(\psi) = 2[\ell_{n, \text{prof}}(\hat{\gamma}) - \ell_{n, \text{prof}}(\gamma)],
\]

displayed as part of Figure 8.6. There is clear indication of ‘attraction’ between \( Y_1 \) and \( Y_2 \) in the handball data, with maximum likelihood estimate 0.418 and 95% confidence interval [0.026, 0.980] for \( \gamma \). The maximum likelihood estimate of \( \phi = \log \lambda - \gamma \) is 2.876, which transforms back to \( \hat{\lambda} = 26.961 \), on the scale of goals scored per match per team.

The preceding already provides useful and relevant information concerning the degree to which handball matches tend to be tight and correlated, and serves as good illustration of our first-order large-sample calculus. We may, however, do better because the model is on exponential form where the theory of Section 5.5 tells us that the optimal confidence distribution involves the conditional distribution of \( \bar{U} \) given \( \bar{V} \), in fact taking the precise form

\[
C(\gamma) = P_{\gamma}(\bar{U} > 53.5685 \mid \bar{V} = 53.9240),
\]

now dropping the sample size index \( n \) from the notation.
Figure 8.6 Two confidence curves for the $\gamma$ parameter of the bivariate Poisson model, for the handball data, based respectively on the $\chi^2$ transform of the deviance (solid curve) and isotonised version of the MCMC scheme for the optimal confidence curve (dashed line). The 95\% interval $[0.026,0.890]$ based on the deviance is indicated, excluding the Poisson independence value zero.

This optimal confidence distribution is difficult to compute directly because the distribution of $\bar{U}$ given $\bar{V}$ is complicated. We may, however, construct a suitable Markov chain Monte Carlo (MCMC) apparatus to sample a high number of realisations from $\bar{U}$, given the observed value of $\bar{V}$, for each value of $\gamma$, as we now explain. The ideas behind these constructions are quite general and may be used for various similar occasions.

We start out writing the probability distribution of a single match result $z = (y_1, y_2)$ in the form of

$$f(z, \gamma, \phi) = \exp\{\gamma U(z) + \phi V(z) + m(z)\}/k(\gamma, \phi),$$

with

$$U(z) = 2\sqrt{y_1y_2}, \quad V(z) = y_1 + y_2, \quad m(z) = -\log(y_1!) - \log(y_2!).$$

For likelihood purposes we do not need to be bothered with the particularities of $m(z)$, but that function too comes into play when we work with the conditional distribution of $\sum_{i=1}^n U(Z_i)$ given $\sum_{i=1}^n V(Z_i) = v$. The full distribution of $n = 79$ match results is

$$f(z_1, \ldots, z_n, \gamma, \phi) = \exp\left\{\gamma \sum_{i=1}^n U(z_i) + \phi \sum_{i=1}^n V(z_i) + \sum_{i=1}^n m(z_i)\right\}/k(\gamma, \phi)^n,$$

from which we find the conditional distribution

$$g_{\gamma}(z_1, \ldots, z_n | v) = \frac{\exp\{\gamma \sum_{i=1}^n U(z_i) + \phi v + \sum_{i=1}^n m(z_i)\}/k(\gamma, \phi)^n}{\sum_{B_{\gamma}(v)} \exp\{\gamma \sum_{i=1}^n U(z_i) + \phi v + \sum_{i=1}^n m(z_i)\}/k(\gamma, \phi)^n} = \frac{\exp\{\gamma \sum_{i=1}^n U(z_i) + \sum_{i=1}^n m(z_i)\}}{\sum_{B_{\gamma}(v)} \exp\{\gamma \sum_{i=1}^n U(z_i) + \sum_{i=1}^n m(z_i)\}}.$$
Here $B_n(v)$ is the quite large set of $(z_1,\ldots,z_n)$ combinations for which $\sum_{i=1}^n V(z_i) = v$. We note that the conditional distribution indeed depends on $\gamma$ but not $\phi$, as per the general theory for exponential families, and that the normalisation constant $k(\gamma,\phi)^n$ has dropped out. What we need to do is to set up a simulation scheme generating realisations $(z_1,\ldots,z_n)$ from the one-parameter distribution

$$g_\gamma(z_1,\ldots,z_n | v) \propto \exp\left\{\gamma \sum_{i=1}^n U(z_i) + \sum_{i=1}^n m(z_i)\right\},$$

in the constrained set of values where $\sum_{i=1}^n V(z_i) = v$. In the present context this means simulating handball match results over $n = 79$ matches for which the total number of goals scored is and remains $v = 4260$. For each such simulated realisation $(z_1,\ldots,z_n)$ we record the value of $n^{-1} \sum_{i=1}^n U(z_i)$, using these to track the required distribution and to read off

$$C^*(\gamma) = P_1\left\{n^{-1} \sum_{i=1}^n U(z_i) \geq 53.5684 \mid \sum_{i=1}^n V(z_i) = 4260\right\}$$

$$= \frac{1}{B} \sum_{s=1}^B \left\{n^{-1} \sum_{i=1}^n U(z_{is}) \geq 53.5684\right\}$$

for a suitably high number of simulated realisations $B$ from the Markov chain.

The MCMC scheme we implemented here is of the Metropolis–Hastings variety (see, e.g., Gilks et al., 1996) and consists of a ‘proposal’ step followed by an ‘acceptance or rejection’ step. The proposal is to choose a single match at random and then change the current $z_i^{\text{old}} = (y_{i,1}, y_{i,2})$ into $z_i^{\text{prop}}$ equal to $(y_{i,1} + 1, y_{i,2} - 1)$ or $(y_{i,1} - 1, y_{i,2} + 1)$, with equal probabilities $\frac{1}{2}, \frac{1}{2}$, while keeping the other $n - 1$ match results unchanged. This symmetric proposal is then accepted with probability

$$p_{\text{accept}} = \min\left(1, g_\gamma(z_1^{\text{prop}}, \ldots, z_n^{\text{prop}} | v) \bigg/ g_\gamma(z_1^{\text{old}}, \ldots, z_n^{\text{old}} | v)\right) = \min(1, \exp(\Delta)),$$

say, where

$$\Delta = \gamma \sum_{i=1}^n \{U(z_i^{\text{prop}}) - U(z_i^{\text{old}})\} + \sum_{i=1}^n \{m(z_i^{\text{prop}}) - m(z_i^{\text{old}})\}$$

$$= \gamma \{U(z_i_0^{\text{prop}}) - U(z_i_0^{\text{old}})\} + m(z_i_0^{\text{prop}}) - m(z_i_0^{\text{old}}),$$

writing $i_0$ for the randomly selected match in question. The practical point is that such a chain is easy to implement and run because both the proposal draw and the acceptance probabilities are easy and quick. The chains mixed well, with suitably high acceptance rate. For good measure we ran these chains a full million of steps, for each $\gamma$ value in $\{-0.10, -0.09, \ldots, 1.09, 1.10\}$, to have a decently stable simulation-based estimate at each application. After more than a hundred million computer draws and computations, then, we reached Figure 8.6. It displays both the $cc_0(\gamma)$ based on the deviance, which is relatively easy to produce, and the harder to get $cc^*(\gamma) = |1 - 2C^*(\gamma)|$ from the isotonised version of $C^*(\gamma)$ given earlier. For smaller samples there would be a more noticeable difference
between the optimal $cc^*$ and the $\chi^2_1$ approximation leading to $cc_d$. The 95% confidence interval is $[0.026, 0.890]$, excluding the independent Poisson case of zero, demonstrating again the presence of the attraction parameter for women's handball.

8.4 Generalised linear models

Linear regression is by far the most widely used method in empirical economics, and in many other fields. It has been extended to logistic regression and several other forms of generalisations. The generic model behind these extensions is the exponential family with the linear regression structure is placed in the mean parameter, or a known function thereof. These generalised linear models (GLMs) were introduced by Nelder and Wedderburn (1972) and are discussed in many textbooks, for example, McCullagh and Nelder (1989). They are also widely used, and excellent software is available in R and other statistical systems.

Consider a regression setup with pairs $(x_i, Y_i)$, and where the task is to assess how covariate vector $x_i$ influences the mean response $E(Y_i | x_i)$ and more generally the full distribution of $Y_i | x_i$. For a GLM, there is a chosen monotone and smooth link function $g(\cdot)$ such that

$$g(E(Y_i | x_i)) = x_i^T \beta$$

for $i = 1, \ldots, n$, in this fashion mapping the mean response to a linear predictor defined in terms of regression coefficients $\beta_1, \ldots, \beta_p$. For traditional linear regression, the link function is the identity. We assume that $Y_1, \ldots, Y_n$ are independent with density function belonging to the exponential
family form
\[ f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \]
with the sample space for \( y \) the same in each case and not depending on the unknown parameters. The functions \( a(\phi), b(\theta) \) and \( c(y, \phi) \) are fully specified. The \( b(\theta) \) function plays a vital role because its derivatives yield the mean and variance function, while \( a(\phi) \) is a scale parameter.

From the two first Bartlett identities (cf. Exercise 2.2), concerning moment properties of the first and second log-derivatives of the density, follow
\[ \mathbb{E}(Y_i | x_i) = b'(\theta_i) \quad \text{and} \quad \text{Var}(Y_i | x_i) = a(\phi)b''(\theta_i); \]

\[ \text{cf. (2.3).} \]

The regression is consequently directly related to the mean response through \( b' \).

When the link function is chosen to make \( g(b'(z)) = z \), yielding \( \theta = x\beta \), the link is called canonical. In several important models, such as the binomial model for logistic or probit regression, and the Poisson model, the scale parameter is the constant \( a(\phi) = 1 \). In other models the scale depends on the free parameter \( \phi \), say in overdispersed logistic regression.

A list of important special cases is as follows.

1. Consider a nonlinear normal regression setup where \( Y_i \) is normal \((\theta_i, \sigma^2)\), with \( \theta_i = r(x_i^\beta) \) for some specified function \( r(\eta) \). Then \( b(\theta) = \frac{1}{2}\theta^2 \). The ordinary linear normal model corresponds to \( r(\eta) = \eta \) with weights \( v_i = 1 \).

2. Assume the \( Y_i \) are Poisson with parameters \( \zeta_i = \exp(x_i^\beta) \). This is Poisson regression with canonical link function. Then \( b(\theta) = \exp(\theta) \).

3. Next let \( Y_i \) be binomial \((m_i, p_i)\), with \( p_i = H(x_i^\beta) \), for a suitable distribution function \( H \). This is again a GLM with \( b(\theta_i) = m_i \log(1 + \exp(\theta_i)) \). This can be used for probit regression, for example, where \( p_i = \Phi(\eta_i) \) with the cumulative standard normal.

4. Suppose positive observations \( Y_i \) are modelled with Gamma distributions \((c_i, d_i)\), where the shape parameter \( c \) is fixed but \( d_i = \exp(x_i^\beta) \). We use the parametrisation where the mean of \( Y_i \) is \( \zeta_i = c/d_i \). The \( J_\eta \) matrix is proportional to the sample variance matrix of the covariate vectors. Suppose on the other hand that \( Y_i \) is taken to be Gamma with parameters \((c_i, d_i)\), this time with \( d \) fixed and flexible \( c_i = \exp(\eta_i) \), with again \( \eta_i = x_i^\beta \). This actually corresponds to a GLM in terms of the log \( Y_i \).

**Example 8.4 Logistic and probit regression**

Stock and Watson (2012, chapter 11) illustrate logistic regression on the Boston Home Mortgage Disclosure Act data to see whether credit institutions were racially biased when granting or denying applications for a home mortgage. In addition to intercept and the dichotomous race variable \( x_i \) (black/white) characterising the applicant there were several \( p - 2 \) there were several other explanatory variables. The binary response variable \( Y \) is 1 if the application was accepted and 0 if denied, which is a Bernoulli variable with success probability \( p = EY = g^{-1}(x_i^\beta) \). With the link function \( g(p) = \log[p/(1-p)] \) the regression is said to be logistic. In economics, the probit link \( g(p) = \Phi^{-1}(p) \) is popular. It connects the mean parameter \( p \) and the linear regression by the normal distribution. In both cases unit scale is assumed, \( a(\phi) = 1 \). The \( n = 2380 \) observations of \( Y_i \) are assumed independent, given their covariate vectors \( x_i \). There might, however, be unobserved heterogeneity in this case,
say when some unrecorded aspects of the applicant or office clerk in the credit institution is influencing the outcome of an application. Then, overdispersed logistic regression might be appropriate, say with $\alpha(\phi) = \exp(\phi)$.

In addition to logistic regression and other models based on the Bernoulli variation model, there are many other GLMs. For the command `glm` in R, the following variation models are ready made, together with their canonical links: binomial (link = logit); gaussian (link = identity); Gamma (link = inverse); inverse.gaussian (link = $1/\mu^2$); poisson (link = log); quasibinomial (link = logit); quasipoisson (link = log). The two latter are models for overdispersed binomial/logistic or Poisson/log-regression models.

In regression applications with many covariates available it is often fruitful to let the data contribute to deciding on which of these to include in a good final model, perhaps via model selection criteria like the AIC, BIC, and FIC briefly described in Section 2.6. This might cause difficulties for proper inference, as output from the selected model, in terms of estimators, p-values and confidence intervals, ought to take into account that there is an initial random element at play, that of selecting the model in the first place. Such inference after model selection issues are discussed in, for example, Claeskens and Hjort (2008, chapters 7 and 8), Efron (2014) and Hjort (2014).

Suppose in suitable generality that $\beta$ is a focus parameter, with estimator $\hat{\beta}$ and with $\hat{\kappa}$ an estimator of its standard deviation, constructed after having selected an appropriate model. The problem is that even though a construction like $t = (\beta - \hat{\beta})/\hat{\kappa}$ might have a well-understood distribution (typically close to a standard normal, for moderate $n$), conditional on the model selected, its full distribution might be much more complicated when the model selection step is factored in. One among several strategies is to carry out bootstrapping to get hold of the distribution of $t_{\text{boot}} = (\hat{\beta} - \hat{\beta}_{\text{boot}})/\hat{\kappa}_{\text{boot}}$, where the model selection step is carried out inside each bootstrap replication leading to $\hat{\beta}_{\text{boot}}$ and $\hat{\kappa}_{\text{boot}}$. It is important here that the $\beta$ in question is a well-defined parameter with clear interpretation across the various candidate models considered, not merely a Greek letter present in different model formulae. Also, the $\hat{\beta}$ estimator in such a context is really a mixture over different possibilities, say $\sum_{j=1}^{m} w(j) \hat{\beta}(j)$, where $\hat{\beta}(j)$ is the form of the estimator if model $j$ among the model candidates is being used, and $w(j)$ an indicator function for that model being selected. In view of this it is to be expected that the full distribution of $t$, to be estimated via the bootstrap distribution of $t_{\text{boot}}$, is a nonlinear mixture over many normal-like components; cf. again Claeskens and Hjort (2008, chapter 7) for results and discussion.

**Example 8.5 Blubber thickness in Antarctic minke whales**

Have southern minke whales become leaner as their abundance increased towards its maximum? The Antarctic krill is very abundant in the Southern Ocean and is the main prey for baleen whales. Humpbacks, blue whales and fin whales are large baleen whales that were severely depleted in the twentieth century by whalers. Laws (1977) suggested the ‘krill surplus hypothesis’ saying that the depletion of the large baleen whales left big amounts of krill available for other predators such the southern minke whale, which is a smaller baleen whale. From about 1970, when southern commercial whaling dwindled out and the large whales were decimated to very low abundances, the southern minke whale has increased in abundance until it peaked around year 2000. To study the fate of the southern minke whale,
about 400 individuals were caught yearly in the first period of Japanese scientific whaling operation over the 18 years 1987/88 until 2004/05. Konishi et al. (2008) and Konishi and Walløe (2015) have studied how thick the blubber was at a point on the body where blubber thickness is known to vary much. To avoid the difficulties associated with early growth, lactation and pregnancy, we shall only look at mature males.

The data for \( n = 2956 \) mature males consist of the response variable blubber thickness \( B \) and the following six covariates: date of capture (with December 1 = 1), diatom adhesion (ordinal scale 0 (none) to 4 (entire body)), longitude, latitude, year (starting from 1987/88 = 1), and body length (meter). Diatoms are algae mostly present at high latitudes. Diatom adhesion thus reflects how long the whale has been far south where krill abundance is highest. To select a model we used stepwise regression by way of \texttt{step} in R, using the Bayesian information criterion BIC as the object function. The selected model is identical to the maximal model we considered, namely a model linear in all the covariates plus the interaction longitude × latitude. Because the prime question is how blubber thickness varied over the years, year was required in all models considered. The regression coefficient for year is estimated as \( \hat{\beta} = -0.016 \) (se = 0.003), with a 95% t-interval \((-0.021, -0.010)\). These results could be shown as a confidence distribution or a confidence curve. Are they valid? If the selected model is the correct one, the results should be in order because the \( n \) is large. But we have selected the model, and might thus be a bit worried; see Exercise 8.6. To check we bootstrapped the data 1000 times and carried out the same model selection on each bootstrap sample. The bootstrap distribution of the t-statistic \( t_{\text{boot}} = (\hat{\beta} - \hat{\beta}_{\text{boot}})/\hat{\kappa}_{\text{boot}} \) for \( \beta \), the parameter of interest, turned out to be very nearly normally distributed. The results are thus validated. We are saved by having a large and informative sample! The result indicates that the krill surplus was reduced and the minke whale was hitting its carrying capacity over the study period.

8.5 Gamma regression models

The Gamma distribution with parameters \((a, b)\) in its most used parametrisation has density \( \{b^a/\Gamma(a)\}y^{a-1} \exp(-by) \) on the positive halfline, with mean \( a/b \) and standard deviation \( a/b^{1/2} \) (both parameters \( a \) and \( b \) are positive). We shall consider regression models for data of the type \((x_i, Y_i)\), with \( x_i \) some covariate vector, and where \( Y_i \) is gamma distributed with parameters influenced by \( x_i \).

Consider first the case of

\[
Y_i \sim \text{Gamma}(\lambda_i, \nu) \quad \text{where} \quad \lambda_i = x_i^T \beta
\]

for \( i = 1, \ldots, n \). The log-likelihood may be expressed in log-exponential form as

\[
\ell_n(\beta, \nu) = \sum_{i=1}^{n} \left\{ x_i^T \beta \log \nu - \log \Gamma(x_i^T \beta) + (x_i^T \beta - 1) \log y_i - \nu y_i \right\},
\]

\[
= n \beta_1 V_1 + \cdots + n \beta_p V_p - n \nu U + \sum_{i=1}^{n} \left\{ x_i^T \beta \log \nu - \log \Gamma(x_i^T \beta) \right\},
\]
Exponential families and generalised linear models

in terms of the sufficient statistics

\[ U = n^{-1} \sum_{i=1}^{n} y_i \quad \text{and} \quad V_j = n^{-1} \sum_{i=1}^{n} x_{ij} \log y_i \quad \text{for } j = 1, \ldots, p. \]

The proper parameter region for \( \beta \) is the subset of \( p \)-vectors for which \( x' \beta > 0 \) for all \( x \) in the support of the covariate distribution.

There are optimal confidence distributions for each of the regression coefficients, as per the general theory. For \( \beta_p \), associated with covariate no. \( p \), we have

\[ C^*_{p}(\beta_p) = P_{\beta_p}[V_p \gtrless v_{p,\text{obs}} \mid U = u_{\text{obs}}, V_j = v_{j,\text{obs}} \text{ for } j = 1, \ldots, p - 1]. \]  

(8.12)

This conditional distribution is quite complicated but may be worked with either via approximations, as per methods of Chapter 7, or evaluated via simulation methods. In fact the methods of Lindqvist and Taraldsen (2006, 2007), briefly outlined and then utilised in Example 8.2, may be put to use. For each candidate value of \( \beta_p = \beta_{p,0} \) one draws a high enough number \( V^*_p \) from the appropriate conditional distribution of (8.12) as follows. Start with a sample \( w_1, \ldots, w_p \) from the uniform distribution on the unit interval and consider finding a proper transformation

\[ y^*_p = y^*_1(\beta_1, \ldots, \beta_{p-1}, v) = G^{-1}(w_i, \lambda_i, v) \]

from the inverse cumulative distribution of the Gamma\((\lambda, v)\), with \( \lambda_i = \lambda_0 + x_i \beta_1 + \cdots + x_i \beta_{p-1} + x_i \beta_{p,0} \). The trick is to fine-tune the \( p \) parameters \( \beta_1, \ldots, \beta_{p-1}, v \) for this particular occasion, so as to match

\[ n^{-1} \sum_{i=1}^{n} y^*_i = u_{\text{obs}} \quad \text{and} \quad n^{-1} \sum_{i=1}^{n} x_{ij} \log y^*_i = v_j \text{ for } j = 1, \ldots, p - 1. \]

This involves the task of solving \( p \) nonlinear equations with \( p \) unknowns. When the solution has been found, one finally uses \( y^*_p \) constructed as in the preceding, with these solutions inserted, still using the same \( w_1, \ldots, w_p \), and calculates \( V^*_p = n^{-1} \sum_{i=1}^{n} x_{ip} \log y^*_i \). This is repeated a high number of times, yielding in the end a simulation approximation to \( C^*_p(\beta_{p,0}) \) of (8.12). That this works properly follows by an extension of arguments in Lindqvist and Taraldsen (2006), as the construction shows the conditional distribution depends on \( \beta_p \) but not the other parameters.

**Example 8.6 Leukaemia and white blood cells**

Table 8.1 gives survival time in weeks \( y \) for 17 patients along with their white blood cell counts \( x \) at the time of diagnosis. These data have been discussed and analysed in Feigl and Zelen (1965), Cox and Snell (1989) and Brazzale et al. (2007). The model favoured in these references is to take the observations \( Y_i \) as exponential with mean \( \lambda_i = \exp(\beta_0 + \beta_1(x_i - \bar{x})) \); in particular, each patient is then seen as having a constant hazard rate \( 1/\lambda_i \). We shall instead work with the gamma regression model

\[ Y_i \sim \text{Gamma}(\beta_0 + \beta_1(x_i - \bar{x}), v), \]

which incidentally also provides a better fit in terms, for example, of the Akaike information criterion (AIC) score; cf. various details in Exercise 8.12. Figure 8.8 displays hazard rates
8.5 Gamma regression models

Table 8.1. Survival times $y$ in weeks, along with associated $\log_{10}$ white blood cell count $x$, for 17 patients; see Example 8.6

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$x$</th>
<th>$y$</th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.36</td>
<td>65</td>
<td>4.00</td>
<td>121</td>
<td>4.54</td>
<td>22</td>
</tr>
<tr>
<td>2.88</td>
<td>156</td>
<td>4.23</td>
<td>4</td>
<td>5.00</td>
<td>1</td>
</tr>
<tr>
<td>3.63</td>
<td>100</td>
<td>3.73</td>
<td>39</td>
<td>5.00</td>
<td>1</td>
</tr>
<tr>
<td>3.41</td>
<td>134</td>
<td>3.85</td>
<td>143</td>
<td>4.72</td>
<td>5</td>
</tr>
<tr>
<td>3.78</td>
<td>16</td>
<td>3.97</td>
<td>56</td>
<td>5.00</td>
<td>65</td>
</tr>
<tr>
<td>4.02</td>
<td>108</td>
<td>4.51</td>
<td>26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.8 Estimated hazard rates for three patients, having respectively white cell blood counts 2.88, 4.00, 5.00, via the gamma regression model of Example 8.6. The three horizontal lines represent estimates of these hazard rates using the proportional hazards model used by Cox and Snell (1989) and Brazzale et al. (2007) for these data.

for three types of patients, using both the gamma regression model and the constant hazard model; checks of model fit, in addition to the AIC scores, indicate that the nonconstant hazards tell a more truthful story. The three statistics

$$U = n^{-1} \sum_{i=1}^{n} Y_i, \quad V_0 = n^{-1} \sum_{i=1}^{n} \log Y_i, \quad V_1 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}) \log Y_i$$

are found to be equal to 1.2013, $-0.5895$, $-0.6934$, respectively. For the natural focus parameter $\beta_1$, which quantifies the influence of the white cell blood count on the hazard rate, the optimal confidence distribution is

$$C^*(\beta_1) = P_{\beta_1} \{ V_1 \geq -0.6934 \mid U = 1.2013, V_0 = -0.5895 \}.$$
Figure 8.9 Optimal confidence distribution function (left) and confidence curve (right) for the coefficient parameter $\beta_1$ in the gamma regression model of Example 8.6 (solid line), along with the usual first-order approximations (dashed line). The 0.05, 0.50, 0.95 confidence points are $-2.410, -1.265, -0.521$ by the optimal construction, but $-2.502, -1.489, -0.477$ with the usual approximation.

This is computed via the simulation scheme just outlined, and produces the confidence distribution and confidence curve in Figure 8.9.

8.6 Flexible exponential and generalised linear models

As conveyed also elsewhere in our text, the ease with which statisticians can programme, profile and optimise log-likelihood functions makes it relatively easy to try out new models, perhaps outside the mainstream repertoire. There might be a certain tendency among statisticians to stick to the traditional when it comes to model building, looking for off-the-shelf methods rather than exploring new venues. That going beyond, say, the GLM apparatus is often fruitful and not too hard is also exemplified in Section 8.8, with what we call generalised linear-linear models, and where we also learn that such more elaborate models often fit the data better. The GLM and GLLM machineries might be further extended to build in parametric models for the link functions, rather than taking these to be default choices.

It is important to realise that the structure of the exponential family and generalised linear (and linear-linear) models nevertheless offers a fair amount of flexibility in that one may put in different covariates and basis functions as building blocks. Thus one may go for built-for-purpose models, for the given occasion, through the devices of adding in more components in a linear structure or building a model around given summary statistics. We shall illustrate such ideas in examples that follow.
8.6 Flexible exponential and generalised linear models

Suppose in general terms that independent observations $Y_1, \ldots, Y_n$ stem from a density of the form $\exp[\theta^T(y) - k(\theta) + m(y)]$, for suitable $T(y) = (T_1(y), \ldots, T_p(y))^t$ and associated parameters $\theta = (\theta_1, \ldots, \theta_p)$. The log-likelihood function is $\ell_n(\theta) = n(\theta^T \hat{Y}_{\text{obs}} - k(\theta))$, in terms of the average $\hat{Y}_{\text{obs}}$ of $T(Y_1), \ldots, T(Y_n)$. The $\ell_n$ function is seen to be concave, and the maximum likelihood estimate is the unique solution to $\nabla \ell_n(\hat{\theta}) = \hat{Y}_{\text{obs}} - \nabla k(\hat{\theta})$; cf. details of Exercise 8.2. Also, the Fisher information matrix is $I(\theta) = \nabla^2 k(\theta)/\nabla \theta \nabla \theta^t$, and the traditional observed information formula $I_n = -\nabla \ell_n(\hat{\theta})/\nabla \theta$ associated with the Hessian matrix of the log-likelihood function becomes $nI(\hat{\theta})$. Under model conditions we have the familiar approximation $N_p(\theta; n^{-1}J(\theta)^{-1})$ to the distribution of $\hat{\theta}$, leading to simple-to-use approximations also to the optimal confidence distributions discussed in Section 8.2 and indeed elsewhere in this chapter.

The point we wish to convey now is the flexibility this model construction offers, by using or adding in different components $T_i(y)$. For illustration, consider building a model for a smooth density on the unit interval, via

$$f(y; \theta) = \exp[\theta_1 y + \cdots + \theta_p y^p - k_p(\theta_1, \ldots, \theta_p)],$$

(8.13)

where $k_p(\theta_1, \ldots, \theta_p) = \log \{\int_0^1 \exp[\theta_1 y + \cdots + \theta_p y^p] \, dy\}$. The model might be fitted using maximum likelihood, for each model order $p$. As any continuous function on the unit interval may be uniformly well approximated with polynomials (the Stone–Weierstrass approximation theorem), this scheme may also be considered a bridge from the parametric to the nonparametric, and perhaps with a model selection criterion like the AIC or focussed information criterion (FIC) of Section 2.6 to choose the right order for the given sample size.

Note here that the averages $\bar{T}_1, \ldots, \bar{T}_p$ are statistically sufficient (see Sections 2.3 and 5.4), so knowledge of these observed averages alone makes it possible to estimate the density and also to reach all relevant inference statements. For the density model (8.13), this would amount to basing all inference on the first $p$ sample moments $n^{-1}\sum_{i=1}^n Y_i^j$. This points to the general useful idea of constructing an exponential model from what is conceived of as good summary statistics. If statistics, say, $A_n, B_n, C_n$ are computed from data $Y_1, \ldots, Y_n$, and these are thought to contain the essential information, then this invites building the vehicle model $\exp[aA_n + bB_n + cC_n - k_n(a, b, c) + m_n(y_1, \ldots, y_n)]$. If in addition the $A_n, B_n, C_n$ are formed by summing over $A(Y_i), B(Y_i), C(Y_i)$, say, then this corresponds to the basic model $\exp[aA(y) + bB(y) + cC(y) - k(a, b, c) + m(y)]$ for the individual $Y_i$. Rather than starting with the model (8.13), therefore, we might construct it from the assumption that the sample means $n^{-1}\sum_{i=1}^n Y_i^j$ up to order $p$ provide the essential information about the data and their distribution. This is a perhaps modest but fruitful twist on the use of statistical sufficiency, and also ought to have useful consequences in the world of big data.

When using such building blocks in exploratory modus one should be careful regarding the assessment of estimation uncertainty, as the usual maximum likelihood methods have as an operating assumption that the model is correct. As this easily might not be the case, one may use the model-robust variance assessment calculus associated with the sandwich matrix, with $\text{Var} \hat{\theta} \approx J^{-1} K J^{-1}/n$; see the discussion in Section 2.6 and result (2.23), with more details in Section A.5. Here $J$ is estimated as earlier, that is, $\hat{J} = J(\hat{\theta})$, whereas $K$ is estimated using $\hat{K}$, the empirical $p \times p$ variance matrix of the $n$ vectors $T(y_i)$. 
Example 8.7 Extending a model by exponential tilting

Consider the modified health assessment questionnaire (MHAQ) data of Example 8.2, featuring \( n = 799 \) measurements suitably transformed to the unit interval. In that example we used the Beta density to model the underlying MHAQ density, but it provided only a modestly good fit. Also, fitting the Beta density using maximum likelihood (or Bayes) risks being not robust in that estimates rely on high precision of measurements close to zero and one.

An option here is to start from scratch, so to speak, and fit the model (8.13) for different model orders \( p \); see Exercise 8.13. Here we instead pursue a different but related idea, which is that of extending any given parametric model via exponential tilting. Starting with any suitable \( f_0(y, \theta) \), like the Beta, we form

\[
f(y, \theta, \gamma) = f_0(y, \theta) \exp\{\gamma^T T(y) - k(\theta, \gamma)\},
\]

where \( k(\theta, \gamma) = \log\left[ \int f_0(y, \theta) \exp\{\gamma^T T(y)\} \, dy \right] \). If \( \gamma \) is close to zero, in which case \( k(\theta, \gamma) \) is also close to zero, the simpler model with only \( \theta \) is good enough. Here \( T(y) \) is one or more functions meant to relate to other aspects of the underlying density than the basic functions involved in \( \log f_0(y, \theta) \). The Beta model is log-linear in \( \log y \) and \( \log(1 - y) \), and the preceding may be used with, for example, \( y, y^2, y^3 \) as supplementary building blocks. Figure 8.2 displays the result of fitting the simplest of these models to the 799 MHAQ measurements, using the three-parameter model \( f(y, a, b, \gamma) \) from such a construction, with \( T(y) = y \). This actually leads to a vast improvement over the Beta model, as judged from the AIC scores. Inspection shows that standard errors computed under the three-parameter model conditions are essentially similar to those computed via the sandwich matrix method.

Note also that the power theorem 5.11 gives a recipe for computing the optimal confidence distribution for \( \gamma \), a parameter that can be seen as giving the degree of departure from the starting model.

Example 8.8 Unfairness of the cross-country ski-sprint event

Ski sprinting is the Formula One event of modern cross-country skiing competitions. The skiers need to not only maintain fabulous speeds throughout challenging courses, for example, an average pace of 30 km/h in a highly varying terrain over 1.5 km, but also to develop technical and tactical skills regarding how to pass competitors on different types of stretches, and so forth. Crucially, the six top athletes allowed to enter the final are also required to run through these demanding courses a total of four times, over a relatively short time period, passing from the prologue to the quarterfinal and the semifinal stages before the final. The operating assumption underlying the layout and execution of the tournament rules is of course the idealistic one that these six remaining skiers start the final on an equal footing. Table 8.2 might indicate otherwise. It appears to show that skiers who have come from the first of the two semifinals (A) do better than those who have come from the second of the two (B).

Data from such a single event cannot convincingly prove that such an unfairness exists, but we have gone to the trouble of locating and transcribing results of the preceding type for a total of 57 top events from the seasons 2014, 2013, 2012, 2011, 2010, including Olympics, World Championships and World Cup events. There are always six skiers in the final, with at least the two best from the two semifinals, whereas the last two places are given to the
8.6 Flexible exponential and generalised linear models

Table 8.2. Results from the Olympic cross-country ski-sprint finals, Sochi, 2014, for men and for women, with A and B indicating that the skier earned his or her place in the final from respectively the first and the second of the two semifinals. See Example 8.8.

<table>
<thead>
<tr>
<th></th>
<th>O.V. Hattestad</th>
<th>T. Peterson</th>
<th>E. Jönsson</th>
<th>A. Gloersen</th>
<th>S. Ustiugov</th>
<th>M. Hellner</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T. Peterson</td>
<td>A</td>
<td>V. Fabjan</td>
<td>A</td>
<td>I. Ingemarsdotter</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>E. Jönsson</td>
<td>B</td>
<td>I.F. Østberg</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>A. Gloersen</td>
<td>A</td>
<td>A.U. Jacobsen</td>
<td>A</td>
<td>S. Caldwell</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>S. Ustiugov</td>
<td>B</td>
<td>V. Fabjan</td>
<td>A</td>
<td>S. Caldwell</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>M. Hellner</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

‘lucky losers’ with the fastest times. Thus the finals are of type (4,2), (3,3), (2,4), where (4,2) refers to cases with four A and two B skiers (like for Table 8.2), and so forth. For each final, we record the ranks 1-2-3-4-5-6 (for these competitions, the time achieved is of no importance, only the ranking matters), and compute

\[ Z_A = \sum X_i \quad \text{and} \quad Z_B = \sum Y_j, \]

with \( X_i \) and \( Y_j \) denoting the ranks for the A and B skiers. For the Sochi finals, for example,

\[ Z_A = 1 + 2 + 4 + 6 = 13 \quad \text{and} \quad Z_B = 3 + 5 = 8, \]

\[ Z'_A = 1 + 3 + 4 + 5 = 13 \quad \text{and} \quad Z'_B = 2 + 6 = 8 \]

for the men and the women. These are the rank-sums associated with tests called Wilcoxon or Mann–Whitney in the literature, but each such \( Z_A \) or \( Z_B \) cannot alone prove much. What we do instead is to sum up all these Wilcoxon statistics, over all 57 events, taking also care to note how many of these are of the (4,2), (3,3), (2,4) types:

\[ Z^*_A = \sum_{k=1}^{57} Z_{A,k} = \sum_{\text{two A}} Z_{A,j} + \sum_{\text{three A}} Z_{A,j} + \sum_{\text{four A}} Z_{A,j}. \]

For the men we find \( Z^*_A = 615 \), which turns out to be far too low compared to the null distribution where the \( \binom{6}{2} = 15 \) places for \( (X_1, X_2) \) are equally likely for the (2,4) type, the \( \binom{6}{3} = 20 \) places for \( (X_1, X_2, X_3) \) equally likely for the (3,3) type, and so forth; the p-value in question is 0.0006 (computed here via simulation from the precise null distribution, but also via its accurate normal approximation). The corresponding calculation for the women gives a less clear result, with a p-value equal to 0.092. Thus we already know that there is a drastic unfairness mechanism at work for the men, favouring those who come from the first of the two semifinals, but the situation for the women is not fully decided.

The preceding analysis relates to a statistical test of a natural null hypothesis, and has yielded a clear and drastic answer to the basic question of whether there is unfairness or not at play. That analysis does not tell us more, however, for example, when it comes to the probability that the winner is an A skier, or that the gold and silver will both be awarded a B skier, and so forth. For estimates and inference for such questions we need a probability model. An exponential type model that turns out to fit the data well is the following:

\[
\begin{align*}
\hat{f}_2(\theta)(i,j) &= P(X_1 = i, X_2 = j) = e^{\theta(i+j)} / K_2(\theta), \\
\hat{f}_3(\theta)(i,j,k) &= P(X_1 = i, X_2 = j, X_3 = k) = e^{\theta(i+j+k)} / K_3(\theta), \\
\hat{f}_4(\theta)(i,j,k,l) &= P(X_1 = i, X_2 = j, X_3 = k, X_4 = l) = e^{\theta(i+j+k+l)} / K_4(\theta).
\end{align*}
\]
Figure 8.10 Confidence curves $cc(\theta)$ for the unfairness parameter $\theta$, for men (solid curve, left) and for women (dashed curve, right). The 95% intervals $[-0.381, -0.131]$ and $[-0.231, 0.025]$ are indicated, with point estimates $-0.254, -0.093$, respectively. The parameter value $\theta = 0$ corresponds to fairness.

Here $K_{24}(\theta), K_{33}(\theta), K_{42}(\theta)$ are sums over respectively 15, 20, 15 terms, made to secure the $f_{24}(i,j,\theta)$ sum to one, and so forth. Olympic fairness corresponds to $\theta = 0$, whereas a negative value of the parameter means lower ranks (closer to the medals) for A skiers, and vice versa. The log-likelihood function is a sum over 50 terms, but can easily be handled numerically. Figure 8.10 depicts the optimal confidence curves $cc(\theta)$ for the men and the women. The point estimates are $-0.254$ and $-0.093$, respectively, with 95% intervals $[-0.381, -0.131]$ and $[-0.231, 0.025]$. The situation for the men is dramatically imbalanced, on the Olympic scale of things, with an estimated probability of 65% of the winner being an A skier. The underlying reasons for this imbalance relate to the fact that skiers coming from the first semifinal have more minutes of rest. With the current layout for these competitions, this difference in resting time, between the semifinal and the final, has more impact on the men than the women. Our results have been communicated to the leaders of the International Ski Federation (FIS, Fédération Internationale de Ski) and the organisers of the World Cup events.

8.7 Strauss, Ising, Potts, Gibbs

There is a large literature on statistical models and related inferential methods for point patterns, perhaps along with stochastic marks attached to these points; see, for example, Ripley (1977), Ripley (1988), Hjort and Omre (1994), Möller and Waagepetersen (2003) and Diggle (2013). One particularly well-studied class is that of the pairwise interaction processes, where the probability distribution can be expressed in terms of functions.
involving one or two such points at a time. Many such models are of the exponential type, for which the general theory developed here applies.

Here we shall consider the Strauss model; cf. Strauss (1975) and Kelly and Ripley (1976), which has probability density taking the form

$$f(x, \beta, \rho) = \alpha \beta^N(x) \rho^M(x),$$  \hspace{1cm} (8.15)

where $x$ is a collection of points in some region $W$ of interest, comprising a total of $N(x)$ points, and with $M(x)$ counting the number of pairs with interdistance at most $r$, a certain threshold parameter most often taken to be known. Above $\alpha = \alpha(\beta, \rho)$ is the normalising constant making all probabilities sum to 1, typically too hard to actually compute. The conditional version of this model, where $N(x)$ is taken fixed at some $n$, has density

$$f(x, \rho) = c(\rho) \rho^M(x),$$  \hspace{1cm} (8.16)

Here $\rho \leq 1$ is a parameter dictating the degree of spatial inhibition; a low value of $\rho$ gives low probability to high $M(x)$, and leads to points tending to stay away from neighbours, whereas higher and less unfriendly values of $\rho$ to a greater degree invites points coming closer. The maximum value of $\rho = 1$ corresponds to the uniform distribution of $n$ points over the region in question.

The model (8.16) may also be written in exponential class form, as $f_0(x, a) = k_0(a) \exp(-aM(x))$, via $\rho = \exp(-a)$. Taking the derivative of its logarithm one sees first that $E_\rho M(X) = \zeta_0(a) = \partial \log k_0(a)/\partial a$, and second that the maximum likelihood estimate may be defined as the point $\hat{a}$ at which $M_{\text{obs}} = \zeta_0(a)$; see Exercise 8.15. This is also valid on the original parameter scale $\rho$, so one way to compute the maximum likelihood estimate $\hat{\rho}$ is to equate $M_{\text{obs}}$ to its expectation $\xi(\rho) = E_\rho M(X)$. A practical way of obtaining this estimate numerically is then to simulate a large number of Strauss patterns for each candidate value of $\rho$, reading off the estimated $\xi(\rho)$ curve, perhaps after isotonisation or smoothing, and then solving for $\xi(\rho) = M_{\text{obs}}$; see Example 8.9 for an illustration of this. There are indeed more efficient ways in which to compute the maximum likelihood estimate, for example, via the Robbins–Monro stochastic approximation algorithm as developed for this purpose by Moyeed and Baddeley (1991). The point we presently wish to emphasise is that such simulations of model-based point patterns, for each candidate value of $\rho$, enables us to compute the guaranteed optimal confidence distribution, via

$$C^*(\rho) = P_\rho \{M(X) > M_{\text{obs}}\} + \frac{1}{2} P_\rho \{M(X) = M_{\text{obs}}\}. \hspace{1cm} (8.17)$$

The associated confidence intervals that may be read off from this curve are all optimal in the sense of Theorem 5.11 (with half-correction due to discreteness).

For our illustrations we have been content to construct a somewhat brute force direct Metropolis–Hastings type generator. More elegant and efficient methods are actually available, for example, in the R package spatstat (Baddeley and Turner, 2005); we wish, however, to demonstrate that when encountering new situations and new models for which there is no easy package to use the modern statistician may still successfully accomplish the simulation task by setting up a reasonable MCMC. In our simpler solution we start off with some pattern of points $x_i = (u_i, v_i)$ in the region of interest, for example, normalised to be the
unit square, and then consider updating $x$ to say $x'$ where one of the points of $x$ is picked at random, and replaced with $x'_i = (u'_i, v'_i)$, drawn from the uniform distribution. This proposed change is then accepted with probability

$$P_{\text{accept}} = \min \left\{ 1, \frac{f(x', \rho)}{f(x, \rho)} \right\} = \min \left\{ 1, \rho^{M(x') - M(x)} \right\},$$

involving the computation of the change $M(x') - M(x)$; the new pattern is $x'$ if accepted but remains unchanged and equal to $x$ otherwise. This defines a Markov chain of point patterns with steady incremental changes and is guaranteed to converge to its equilibrium distribution, namely that of $f(x, \rho)$, as per general MCMC theory, see, for example, Hastings (1970) and Gilks et al. (1996). The idea is to carry out this simulation scheme (or any other more clever and optimised version) for each candidate parameter value $\rho$ and then read off $C^*(\rho)$ of (8.17).

**Example 8.9 Confidence for Strauss**

To illustrate their stochastic approximation algorithm for finding the maximum likelihood estimate for given point patterns, Moyeed and Baddeley (1991) examined a particular case with $n = 25$ points in the unit square, with neighbour distance parameter set to $r = 0.15$, and with the number of pairs with interdistance smaller than $r$ observed to be $M_{\text{obs}} = 4$. Figure 8.11 displays the curve $\xi(\rho) = E_\rho M(X)$, as computed via simulations, where we have used $10^4$ simulated Strauss realisations for each $\rho$ after a burn-in phase of $10^3$. The figure gives both the direct pointwise simulation based estimate of $\xi(\rho)$, as a somewhat wiggly curve, and the isotonised version. We may read off the maximum likelihood estimate, when $M_{\text{obs}} = 4$, is 0.099.

![Figure 8.11](image-url)
Figure 8.12 For the case of \( n = 25 \) points distributed according to the Strauss model (8.16) with \( r = 0.15 \) (see Example 8.9), the figure shows the optimal confidence distribution (left) along with the ensuing optimal confidence curve (right). We have simulated \( 10^5 \) Strauss point patterns after burn-in, for each of value of \( \rho \).

Figure 8.13 A two-colour chain \( x_1, \ldots, x_n \), with \( n = 100 \) (with ‘black’ broad tags and ‘white’ slim tags). It has \( A(x) = \sum_{i=1}^n I[x_i = 1] = 42 \) and \( B(x) = \sum_{i=2}^n I[x_{i-1} = x_i] = 85 \), the sufficient statistics for the Ising type model (8.18).

estimate as \( \hat{\rho} = 0.099 \), agreeing with what was found by Moyeed and Baddeley (1991). For Figure 8.12 we rather more laboriously simulated as many as \( 10^5 \) Strauss patterns for each \( \rho \), again with \( 10^3 \) as burn-in, to display the isotonised version of \( C^*(\rho) \) and the associated optimal confidence curve \( cc^*(\rho) \). Confidence intervals are read off from these curves, for example, the 90% interval \([0.032, 0.261]\) (Moyeed and Baddeley (1991) did not attempt inference apart from computing the estimate). Note that using probability theorems about the approximate normality of estimators in such models, for point pattern data observed over an increasing area, cannot help us here, where \( M(x) \) is one of a small set of numbers, and where \( M_{\text{obs}} = 4 \).

Example 8.10 Confidence for Ising

Let us consider building and analysing models for two-colour chains, say \( x = (x_1, \ldots, x_n) \), where \( x_i \) is one (black) or zero (white). Figure 8.13 gives one such chain, with \( n = 100 \) positions, or picture elements (called ‘pixels’ in dimension two). A probability model for such images takes the form

\[
f(x, \alpha, \beta) \propto \exp[\alpha A(x) + \beta B(x)],
\]

(8.18)
where $A(x) = \sum_{i=1}^{n} I(x_i = 1)$ accounts for the relative blackness and $B(x) = \sum_{i=2}^{n} I(x_{i-1} = x_i)$ is associated with the degree of spatial continuity. This can actually be reformulated in more familiar Markov chain language, giving rise to a certain transition probability matrix of $P_{ij} = P(x_i = j | x_{i-1} = i)$ for $i,j = 0,1$, but we choose formulation (8.18) because it generalises better to the important Ising type models for two- and higher-dimensional images. The image portrayed in Figure 8.13 has $A(x) = A_{obs} = 42$ and $B(x) = B_{obs} = 85$.

We seek the confidence distribution for $\beta$, for the Ising model (8.18). It again follows from our power theorem 5.11 that the optimal confidence distribution takes the form

$$C(\beta) = P_{\beta} \{ B(X) > 85 | A(X) = 42 \} + \frac{1}{2} P_{\beta} \{ B(X) = 85 | A(X) = 42 \}.$$  

To compute this we generate for each value of $\beta$ a high number $10^5$ of images $x$ from the proper conditional distribution

$$g(x, \beta) = P_{\beta} \{ X = x | A(x) = A_{obs} \} = \frac{\exp\{ \beta B(x) \}}{\sum_{x' : A(x') = A_{obs}} \exp\{ \beta B(x') \}}.$$  

We do this by implementing a Metropolis type MCMC of images set up such that its equilibrium distribution is precisely that of $g(x, \beta)$. This is achieved by moving from image $x$ to a proposal image $x'$ by switching one ‘0’, randomly selected among the zeroes, with one ‘1’, randomly selected among the ones, and then accepting this proposal with probability

$$\min\{ 1, \frac{g(x', \beta)}{g(x, \beta)} \} = \min(1, \exp[\beta (B(x') - B(x))]).$$  

The result of this computational scheme, after running such chains of chains with a long enough burn-in phase and then for a long enough time after that, for each $\beta$, is given in Figure 8.14. For that figure we have basically used (8.7) followed by isotonic regression. The median confidence estimate of $\beta$ is 1.808.

The type of analysis given here can be carried through for two-dimensional Ising type models too, with essentially the same setup, also with more parameters at play than two, but at the expense of more bookkeeping when setting up programmes and making simulation schemes efficient.

### 8.8 Generalised linear-linear models

The GLM machinery is a statistical success story, having extended the concepts and methods of classical multiple linear regression to situations with other types of outcomes. Whereas this provides a practical and reasonably flexible methodology, along with easily available software, modern statisticians should not be restrained by the inherent limitations of the strict GLM. Using the generalised models typically involves exploring and analysing how one of the parameters of a familiar probability model depends on certain covariates, via linear combinations. The classical case amounts to the normal model with the mean parameter taken as a linear function of covariates but with the standard deviation parameter kept fixed, and a second common example is the gamma distribution with one of the parameters a linear function of covariates, perhaps via a link function, but not both of the parameters; cf. Section 8.5.
8.8 Generalised linear-linear models

Importantly, one may similarly explore and analyse how covariates influence both parameters of such vehicle models, and one may do so without needing much more in our methodological and practical toolboxes, as long as one can programme the log-likelihood function and use numerical optimisation. In general terms, if the vehicle model is of the form \( f(y,a,b) \), say, then one may fit models of the type \( Y_i \sim f(y_i,a_i,b_i) \), with \( a_i = h_1(x_i\alpha) \) and \( b_i = h_2(z_i\beta) \) via two inverse link functions \( h_1 \) and \( h_2 \). Here the \( z_i \) may be either equal to \( x_i \), or a transformation or a subset of these. We may call such models generalised linear-linear models. The way forward proceeds as per the general likelihood and confidence distribution theory of Chapters 3 and 4, starting with the log-likelihood function

\[
\ell_n(\alpha,\beta) = \sum_{i=1}^{n} \log f(y_i,a_i,b_i) = \sum_{i=1}^{n} \log f(y_i,h_1(x_i^{'}\alpha),h_2(z_i^{'}\beta)).
\]  

(8.19)

One may find confidence distributions for each individual \( \alpha_j \) and \( \beta_k \) regression coefficient, and also for other focus parameters. One might of course also attempt more complex models with three or more parameters for the vehicle model, though one may run into overfitting problems if the same covariates are used to tentatively explain too many aspects of the data. To guard against this one ought to apply model selection criteria of the AIC, BIC, FIC type; see Section 2.6.

Example 8.11 Heteroscedastic regression

Classical linear regression assumes not only that the mean response is linear in the covariates, but that the variance is constant across all covariate vectors. Whereas the first condition is often acceptable, also because the linear function is the dominant term in a
Taylor expansion of a nonlinear function, the condition of variance homoscedasticity is sometimes in trouble. A useful model is then

\[ Y_i = x_i^T \beta + \sigma \exp(\phi_i z_i) \varepsilon_i \quad \text{for } i = 1, \ldots, n, \]

where the \( \varepsilon_i \) are i.i.d. and standard normal. It typically helps both numerical optimisation and statistical interpretation to have the \( z_i \) scaled such that

\[ \text{skew} \bar{z} = n^{-1} \sum_{i=1}^n z_i = 0, \]

in which case \( \sigma \) is the standard deviation for a data point where \( z_i \) is in the middle of its covariate range. The log-likelihood function is

\[ \sum_{i=1}^n \left\{ -\log \sigma - \phi_i z_i - \frac{1}{2} \frac{(y_i - x_i^T \beta)^2}{\sigma^2 \exp(2\phi_i z_i)} \right\} = -n \log \sigma - \frac{1}{2} \frac{1}{\sigma^2} \sum_{i=1}^n \frac{(y_i - x_i^T \beta)^2}{\exp(2\phi_i z_i)}, \]

from which one may find profiled estimates of \( \sigma \) and \( \beta \) for fixed \( \phi \), even in explicit form, and then optimise over \( \phi \). In particular setting up confidence distributions for the \( \phi \) parameters is not difficult, via either the direct normal approximation of the maximum likelihood estimator \( \hat{\phi} \) or the more careful \( \chi^2 \) connection for the deviance \( D_n(\phi) \).

Figure 8.15 displays the 5000 m and 10,000 m personal best times for the best 250 speedskaters of the world, as given by the Adelskalenderen (per October 2014). Denoting these \( x_i \) and \( y_i \), one may use ordinary linear regression to see how the 5k time may be used to predict the 10k time. This leads in particular to easily computed prediction quantiles \( \hat{a}_0 + \hat{b}_0 x_i + \hat{\sigma}_0 \Phi^{-1}(\alpha) \), at level \( \alpha \), with the 0.05 and 0.95 quantile curves given as straight lines in the figure; here \( \hat{a}_0, \hat{b}_0, \hat{\sigma}_0 \) are the usual estimates associated with linear regression.

The situation is clearly heteroscedastic, however, so we have fitted the preceding model with standard deviations \( \sigma \exp(\phi_1 z_i + \phi_2 z_i^2) \), where \( z_i = (x_i - \bar{x})/s_x \) is \( x_i \) normalised to have

\[ 15:10 \]
\[ 14:30 \]
\[ 13:30 \]
\[ 12:30 \]
\[ 6:00 \]
\[ 6:10 \]
\[ 6:20 \]
\[ 6:30 \]
\[ 6:40 \]
\[ 6:50 \]
\[ 7:00 \]

**Figure 8.15** Adelskalenderen data, with the personal bests on the 5k and 10k distances for the 250 best skaters (as of October 2014), with 0.05 and 0.95 quantile curves for the 10k time as a function of 5k time, with two models. The straight dotted lines are from ordinary linear regression and the curved solid lines use the heteroscedastic model.
mean zero and standard deviation one. Both heteroscedasticity parameters are significantly
different from zero and necessary for explaining the data well; we find that the confidence
distributions for $\phi_1$ and $\phi_2$ are well approximated by respectively $N(0.1070,0.0454^2)$ and
$N(-0.0881,0.0326^2)$. That the homoscedastic model is too simple is in particular noticeable
when it comes to understanding the distribution and prediction of $y$ from $x$ for the lower
values of $x$, say for 5k times below 6:10, close to the world record level. Here the quantile
curves
$$\hat{q}(x) = \hat{a} + \hat{b}x + \hat{c} \exp(\hat{\phi}_1 z(x) + \hat{\phi}_2 z(x)^2) \Phi^{-1}(a)$$
suit the purpose much better, featuring the maximum likelihood estimates in the
heteroscedastic five-parameter model, and with $z(x) = (x - \bar{x})/\sigma_x$. See Exercise 8.16 for
details.

**Example 8.12 Doubly linear Gamma regression**

Consider survival data of the regression type $(t_i, x_i, \delta_i)$, discussed, for example, in
Section 4.7; in particular, $\delta_i$ is the indicator for noncensoring. Assume the time to event
in question follows a Gamma distribution but with both of its parameters $(a, b)$ influenced
by the covariates. With $a_i = \exp(x_i \alpha)$ and $b_i = \exp(z_i \beta)$ we need a variation of (8.19) to take
censoring into account, so we are led to

$$\ell_n(a, \beta) = \sum_{\delta_i=0} \log g(y_i, a_i, b_i) + \sum_{\delta_i=0} \log[1 - G(y_i, a_i, b_i)]$$

$$= \sum_{i=1}^n \left[ \delta_i \log g(y_i, \exp(x_i \alpha), \exp(z_i \beta)) + (1 - \delta_i) \log[1 - G(y_i, \exp(x_i \alpha), \exp(z_i \beta))] \right],$$

in which $g(y, a, b)$ and $G(y, a, b)$ are the density and cumulative distribution function for the
Gamma$(a, b)$. Here $z_i$ may be $x_i$ or a subset or transformation thereof.

We illustrate this apparatus with a dataset pertaining to primary biliary cirrhosis (an
autoimmune disease of the liver), involving 312 patients, and with the time to event in
question being time from registration at the clinic to death or liver transplantation, here
measured in years. There is about 60% censoring here ($\delta_i = 0$), in that the event described
took place inside the time frame of the statistical survey ($\delta_i = 1$) for 125 of the 312 patients.
The covariates included in our analysis are $x_1$, 1 if a certain D-penicillamine drug is used,
2 if not; $x_2$, the age at registration, ranging from 26 to 79; and $x_3$, serum bilirubin in mg/dl
(ranging from 0.3 to 28.0). The model we fit takes the real times to event (i.e., without
censoring) to be Gamma$(a_i, b_i)$ with

$$a_i = \exp(a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3}) \quad \text{and} \quad b_i = \exp(b_0 + b_3 x_{i3}).$$

This model does rather better than other models we checked out in terms of AIC and FIC
scores. Parameter estimates for the four plus two parameters are $1.9455 (0.2794)$, $0.0002$
(0.0828), $-0.2096 (0.0399)$, $-0.6711 (0.1928)$, and $-1.7981 (0.1946)$, $0.6184 (0.1874)$,
with standard errors in parentheses. Regression coefficients $a_2, a_3, \beta_3$ are very significantly
present. The time to event declines drastically with increased $x_3$ and also with $x_2$. Various
associated parameters of interest may now be estimated and reported on via confidence
distributions, such as the median time to event for a given patient with his or her given
Figure 8.16  Quantiles 0.05, 0.50, 0.95 for the time from registration to either death or operation, in years, as a function of serum bilirubin, for a 50-year-old person using the D-penicillamine drug; see Example 8.12.

covariates; see Exercise 8.17 for some details. For Figure 8.16 we display the 0.05, 0.50, 0.95 quantiles of the time to event distribution for a 50-year-old patient using the drug in question, as a function of his or her level of $x_3$. These curves are better estimated using this doubly linear gamma model than with the simpler gamma regression models of Section 8.5.

8.9 Notes on the literature
The exponential family model goes back to Darmois (1935), Koopman (1936) and Pitman (1936), who saw that the common mathematical features of several different parametric models, including moment structure and sufficiency, could fruitfully put these in the same framework. Modern accounts include Johansen (1979), Brown (1986) and Barndorff-Nielsen (2014). The origin of the generalised linear model class is Nelder and Wedderburn (1972), concurrent with the emergence of effective algorithms for handling parameter estimation and variance matrices for such cases. The GLIM package (Generalized Linear Interactive Modelling), utilising these generalised iterative reweighted least squares (girls) ideas, was developed by the Royal Statistical Society’s Working Party, chaired by J. A. Nelder, and launched in 1974. Nelder continued to contribute significantly to the GLM and related topics, extending the repertoire of regression models which could be handled by available software. The standard reference for GLMs remains McCullagh and Nelder (1989). Useful generalisations of these methods to various multivariate models are covered by Fahrmeier and Tutz (1994), and important extensions in another direction, that of the so-called generalised linear mixed models, are handled and discussed in Breslow and Clayton (1993) and Skrondal and Rabe-Hesketh (2004). The so-called hierarchical
likelihood for certain kinds of linear model structures has been developed in Lee and Nelder (1996) and further in Lee et al. (2006).

Extending given parametric start models via exponential tilting, as in (8.14) and illustrated in Example 8.7, is a fairly general idea. This is useful for checking adequacy of the start model, via optimal confidence distributions for the extension parameters, and also for building bridges from the parametric to the nonparametric, by adding in more components with growing sample size. Such schemes are investigated in Mogstad (2013). These estimators take the form $f(y) = f(y, \hat{\theta}) \hat{P}(y)$, that is, a parametric start times a correction factor, also worked with in nonparametric contexts by Hjort and Glad (1995) and Hjort and Jones (1996).

It is perhaps fair to point out that from the perspectives of a modern statistician, constraining one’s models to the GLM type is an unnecessary restriction, owing to the fully available and relatively easy to use optimisation algorithms, such as the `nlm` in R. This and related highly efficient algorithms, including `nlminb` and `optim`, using methods refined from those developed and summarised in, for example, Dennis and Schnabel (1983), may be utilised to find maximum likelihood and minimum divergence type estimates for even quite complex models. The practical point is that as long as one can programme the model’s log-likelihood function, then such algorithms manage to find both the maximum likelihood estimate and the associated Hessian matrix, and the methodological point is that inference and confidence methods as developed in this and previous chapters are then in business. This is also why what we called generalised linear-linear models in Section 8.8 are relatively painless in implementation and application; see Claeskens and Hjort (2008) for more on the heteroscedastic model used in Example 8.11 as well as for other instances of models with similar ‘GLLM’ characteristics. When one moves out of the cleaner GLM terrain one risks losing log-concavity of the likelihood functions, and our power theorem 5.11 does not easily apply, but these inconveniences are often minor. Because these flexible tools encourage statisticians to investigate several models for a given dataset, model selection methods will typically be needed to avoid overfitting; see Section 2.6 and Claeskens and Hjort (2008) for a fuller account.

Programming the log-likelihood function is impractical in situations in which the normalisation constant is too hard to compute, as for the situations treated in Section 8.7. We would then typically resort to simulations, perhaps via MCMC, to find both maximum likelihood estimates and full confidence distributions. For more about point process models, of which the Strauss model of that section is an example, see Ripley (1981), Cressie (1993), Möller and Waagepetersen (2003) and Diggle (2013). For analysing spatial patterns using the R package Spatstat, see Baddeley and Turner (2005), with a comprehensive treatment in Baddeley et al. (2015). Image analysis models from statistical perspectives are treated in, for example, Ripley (1981, 1988) and Cressie (1993).

Exercises

8.1 **Exponential family:** For each of the models listed below, verify that it belongs to the exponential class and exhibit appropriate functions $T(y)$, $m(y)$, $k(\theta)$ of formula (8.1). You may consult Wikipedia or other open sources for yet further examples.

(a) $Y$ is binomial $(n, p)$. 

---
Exponential families and generalised linear models

(b) \( Y \) is normal \( (\mu, \sigma^2) \), (i) with \( \sigma \) known and \( \mu \) unknown; (ii) with \( \mu \) known and \( \sigma \) unknown; (iii) with both parameters unknown.

(c) \( Y \) is Poisson \( \lambda \). More generally, let \( Y \) be the vector \( Y_1, \ldots, Y_a \), where \( Y_i \sim \text{Pois}(w_i \theta) \), with known values \( w_i \). Show that (8.1) still holds.

(d) \( Y \) is Beta\( (a, b) \), with the first, the second, or both parameters unknown.

(e) \( Y \) is the result from a single throw of a die, with \( f(i, \theta) \propto \exp(\theta(i - 3.5)) \).

(f) \( Y \) is Gamma\( (a, b) \), with the first, the second, or both parameters unknown.

8.2 Moments for an exponential family: Suppose \( Y \) has probability density function of the exponential form \( f(y, \theta) = \exp[\theta^T (y - k(\theta)) + m(y)] \), for appropriate \( T(y) = (T_1(y), \ldots, T_p(y))^T \) and \( m(y) \), and with \( k(\theta) = k(\theta_1, \ldots, \theta_p) \) the required function to make \( f \) integrate to 1. As is clear from Sections 8.1 and 8.2, there is a long list of familiar models being special cases of such a setup.

(a) Show that \( \mathbb{E}_\theta T(Y) = k'(\theta) = \partial k(\theta)/\partial \theta \).

(b) Show that the Fisher information matrix can be expressed as

\[
J(\theta) = \text{Var}_\theta T(Y) = k''(\theta) = \partial^2 k(\theta)/\partial \theta \partial \theta^T.
\]

In particular, show that \( k(\theta) \) is convex over its natural parameter space.

(c) Show that the moment-generating function \( M(s) = \mathbb{E}_\theta [s^T T(Y)] \) for \( T(Y) \) can be expressed as \( \exp[k(\theta + s) - k(\theta)] \). For the case of \( \theta \) being one-dimensional, use this to find an expression for the skewness of \( T(Y) \).

8.3 Maximum likelihood via simulation: Assume independent observations \( Y_1, \ldots, Y_n \) follow a density of the form \( \exp[\theta^T (y - k(\theta)) + m(y)] \) studied above. For the following we have in mind situations where the normalising constant \( k(\theta) \) is hard to get by, whereas data may be simulated for any given \( \theta \), for example, via MCMC.

(a) Show that the log-likelihood function is \( \ell_n(\theta) = n[\theta^T \bar{T} - k(\theta)] \), with \( \bar{T} = n^{-1} \sum_{i=1}^n T(Y_i) \).

Show that the maximum likelihood estimator \( \hat{\theta} \) is the solution to the equation \( \zeta(\hat{\theta}) = \bar{T}_\text{obs} \), where \( \zeta(\theta) = k'(\theta) \).

Explain how this may be used for finding \( \hat{\theta} = \zeta^{-1}(\bar{T}_\text{obs}) \) via simulation.

(b) Also give a method for estimating the observed Fisher information \( nJ(\hat{\theta}) = nk''(\hat{\theta}) \) via simulation, and hence what is required for first-order inference for \( \theta \), via \( \hat{\sigma}^2 = n\theta^{-1}J(\theta)^{-1} \).

(c) Assume the data follow the normal \( (\theta, 1) \) with \( n = 10 \) and \( \bar{Y}_\text{obs} = 3.333 \). Pretend you do not know the usual mathematics associated with likelihood theory for this model, but use the simulation methods indicated earlier to arrive at both the maximum likelihood estimate, with its standard deviation, and an approximate confidence distribution.

8.4 Odds ratios: Here we look at odds and ratios of odds, equivalent to log-odds differences, in the traditional case where the probabilities involved are not too close to zero or one.

(a) Suppose \( Y \) is binomial \( (m, p) \), where we know that \( \sqrt{m(\hat{p} - \tilde{p})} \rightarrow N(0, p(1-p)) \), for \( \hat{p} = Y/m \). Show that

\[
\sqrt{m} \left( \log \frac{\hat{p}}{1-\hat{p}} - \log \frac{p}{1-p} \right) \rightarrow_d N(0, \frac{p(1-p)}{(p(1-p))^2}).
\]

Conclude from this that \( \chi^2 = (mp)^{-1} \left[ m(1-p) \right]^{-1} \) is the approximate variance for the observed log-odds \( \log(\hat{p}/(1-\hat{p})) \), with natural estimate \( 1/Y + 1/(m - Y) \).

(b) For a two-by-two table with \( Y_0 \sim \text{Bin}(m_0, p_0) \) and \( Y \sim \text{Bin}(m_1, p_1) \), as in Example 8.1, consider the odds ratio and its estimator

\[
\text{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \exp(\psi) \quad \text{and} \quad \tilde{\text{OR}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \exp(\hat{\psi}).
\]
with \( \psi \) and \( \hat{\psi} \) the log-odds and its estimator. Show that \( \log \hat{\text{OR}} = \hat{\psi} \) is approximately normal with variance naturally estimated by

\[
\hat{\tau}^2 = 1/Y_0 + 1/(m_0 - Y_0) + 1/Y_1 + 1/(m_1 - Y_1).
\]

Show that the approximate confidence distribution \( \Phi((\psi - \hat{\psi})/\hat{\tau}) \) is asymptotically correct, and explain how this also yields an approximate confidence distribution for OR.

(c) Do the calculations for Example 8.1 and reproduce a version of Figure 8.1, also with other binomial sample sizes \((m_0, m_1)\) and results \(y_0, y_1)\).

8.5 Bowhead whales: The conditional confidence distribution based on (8.4) is optimal, save for discreteness. Compare this optimal confidence distribution based on the data of Example 8.3 with the confidence distribution based on an approximate normal pivot based on the transformed parameter \( N^{-1/2} \). The coefficients of the pivot are found by running the regression \( \Phi^{-1}(C(N)) \sim N^{-1/2} \). You should find the approximate confidence distribution to be \( C(N) \approx \Phi(5.601 - 86.96N^{-1/2}) \). Plot the confidence curve for both the optimal confidence distribution and its approximation in the same diagram. Do the same for confidence density and confidence deviance.

8.6 Post-selection issues in regression: When faced with many potential covariates in data to be analysed by least squares linear regression, as in, for example, Example 8.5, it might be advisable to select a regression model by way of a search algorithm, say the powerful search algorithm \texttt{step} in R guided by Akaike’s information criterion (AIC). The selected model is fitted to the data. Are the reported standard errors, confidence intervals and \( p \)-values correct? Assume that the true model is within the reach of the search, and assume that it satisfies the standard assumption of linearity and normality. Simulate data with say \( n = 50 \) units and \( p = 10 \) potential correlated covariates. Assume that one of the covariates is of special interest, and is part of the true model which would have an \( R^2 \) around 0.85. Let \( \hat{\beta}_1 \) be its regression coefficient, and investigate the distribution of the confidence distribution at the true parameter, \( C(\hat{\beta}_1) = F_{n-p-1}((\beta_1 - \hat{\beta}_1)/\text{se}_1) \), where the point estimate and its standard error \( \text{se}_1 \) are reported by \texttt{step} for the selected model, and where the focussed covariate is required to be in the selected model. Here \( F_{n-p-1} \) is the appropriate \( t \)-distribution, and \( B = 100 \) repeated simulations should be enough. What pattern do you see?

8.7 Optimal confidence distributions for Gamma parameters: Assume independent observations \( Y_1, \ldots, Y_n \) stem from the gamma distribution with parameters \((a, b)\), with density \( f(y; a, b) = \left(b^a/\Gamma(a)\right)y^{a-1}\exp(-by)\) on the positive halfline. Here \( a \) and \( b \) are unknown positive parameters.

(a) Show that \((U_n, V_n) = (n^{-1}\sum_{i=1}^{n} Y_i, n^{-1} \sum_{i=1}^{n} \log Y_i)\) is sufficient and complete.

(b) Access the dataset pertaining to the birthweight of 189 children, along with various covariates for their mothers, from the CLP book website www.mn.uio.no/math/english/services/knowledge/CLP/; cf. Example 2.4. Fit these 189 birthweights to the Gamma distribution, using kilograms as the scale; cf. Figure 8.17.

(c) Use profile log-likelihoods and deviances to exhibit confidence distributions and confidence curves for the two parameters of this gamma distribution.

(d) Modify the methods outlined for the beta distribution in Example 8.2 to supplement your results from (c) with the optimally constructed confidence distribution and confidence curve, for these two parameters.

(e) What are the appropriate modifications of these confidence distributions when it is acknowledged that the Gamma distribution may not be fully correct?
Figure 8.17  Fitted birthweight distribution density, for smokers (left) and nonsmokers (right), for mothers of average weight, via the gamma regression model outlined in Exercise 8.8; cf. also Figure 2.2 and Example 2.4.

(f) In fact the accompanying Exercise 8.8 works under the assumption that a bigger gamma regression model is correct, say \( y_i \sim \text{Gamma}(x_i \beta, \nu) \). Show from this assumption that the birthweights \( y \) themselves cannot follow a precise Gamma distribution marginally (only approximately).

8.8 Optimal confidence distribution for Gamma regression: For the dataset on 189 birthweights and covariates for their mothers in Exercise 8.7, consider regression models for \((y, x_1, x_2)\), where \( y \) is birthweight (in kg), \( x_1 \) is the mother’s weight before pregnancy, and \( x_2 \) is an indicator for the mother being a smoker or not.

(a) Fit the gamma regression model where \( y_i \sim \text{Gamma}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \nu) \) for \( i = 1, \ldots, n \), and compute also standard errors for the estimates of the four model parameters. How do you judge the impact on smoking on birthweight? Reproduce Figure 8.17, which displays the fitted gamma distributions for a smoking and a nonsmoking mother, both of average weight.

(b) Compute the profile log-likelihood function and deviance-based confidence distribution and confidence curve for \( \beta_2 \). Comment on what you find.

(c) Use the methods of Section 8.5 to compute and exhibit the optimal confidence distribution and confidence curve for \( \beta_2 \). Comment on how close this optimal distribution is to that reached by the deviance based method.

8.9 Bivariate handball: Here we tend to various points related to the bivariate Poisson distribution of Section 8.3.

(a) For \( Y \sim \text{Pois}(\lambda) \), consider \( Z = 2\sqrt{Y} - 2\sqrt{\lambda} \). Show that \( Z \) tends to the standard normal as \( \lambda \) increases.

(b) For different values of \( \lambda \), investigate \( \log M(t, \lambda) \) numerically, where \( M(t) = \text{E} \exp(tZ) \) is the moment-generating function of \( Z \); the point is that \( \log M(t, \lambda) \) ought to be close to \( \frac{1}{2} t^2 \) for the normal approximation of (a) to be good.
Exercises 269

(c) Show that the bivariate Poisson model (8.9) is well defined for all values of \( \gamma \), including negative ones.

(d) Simulate handball match results from the estimated bivariate model (8.9), with \( \lambda = 26.961 \) and \( \gamma = 0.418 \). Compute the correlation between scores in such a match.

(e) Speculate about covariates that may be used to model and predict outcomes of given handball matches, for example, via regression modelling of \( \lambda \) and \( \gamma \).

8.10 Bivariate football: Bivariate football is different from bivariate handball in several respects. Access and organise as required in your computer the dataset on a total of 64 + 31 + 64 = 254 matches from five grand European and World tournaments from the CLP book website.

(a) Fit the (8.9) model \( f_0(y_1, y_2; \lambda, \gamma) \) of Section 8.3 to these data. Compute also standard errors for the two model parameter estimates. How do you judge the sign and size of the interaction parameter?

(b) Use profile deviances to compute and display confidence distributions and confidence curves for the \( \lambda \) and \( \gamma \) parameters, and also for the probability that a football match ends in a 0–0. Discuss whether it is advisable to remove the Germany vs. Saudi Arabia 8–0 result of the 2002 World Championships from your analysis.

(c) Use the methods of Section 8.3, appropriately modified, to also produce the optimal confidence distribution and confidence curve for the Poisson interaction parameter \( \gamma \).

8.11 A bivariate Poisson process: The extended Poisson model invented and worked with in Section 8.3 for pairs of count variables can reflect both positive and negative correlations. A simpler bivariate Poisson model construction is the following, which, however, can accommodate only nonnegative correlations.

(a) Let \( X = X_0 + Z \) and \( Y = Y_0 + Z \), where \( X_0, Y_0, Z \) are independent Poisson variables with means \( (1 - a)\lambda, (1 - a)\lambda, a\lambda \), where \( a \in [0, 1] \) is a mixing parameter. The idea here is that \( X \) and \( Y \) are observed, but not the underlying components \( X_0, Y_0, Z \). Show that \( X \) and \( Y \) are both Poisson with parameter \( \lambda \), and that the correlation between them is \( a \). This is a natural model for positively correlated Poisson variables of the same rate.

(b) Find an expression for the joint probability distribution, starting from

\[
 f(x, y; \lambda, a) = P[X = x, Y = y] = \sum_{z \leq \min(x, y)} p(x - z, (1 - a)\lambda)p(y - z, (1 - a)\lambda)p(z, a\lambda),
\]

writing \( p(x, \mu) = \exp(-\mu)\mu^x/x! \) for the Poisson point probability function for parameter \( \mu \). Fit this two-parameter model to the 79 handball match results of Section 8.3, via maximum likelihood. Find also the approximate confidence distribution for \( a \) (with a point mass at zero corresponding to independence), using the method of deviance from Section 3.4. You may do this first for the case of a known \( \lambda = 26.962 \) (the average number of goals per team scored in a match), then for the case of both parameters \( \lambda, a \) unknown. (The maximum likelihood estimate for \( a \) turns out to be 0.209, with a 90% interval from 0.008 to 0.380.)

(c) The preceding may be generalised to the case of Poisson processes, say \( X(t) = X_0(t) + Z(t) \) and \( Y(t) = Y_0(t) + Z(t) \), where \( X_0, Y_0, Z \) now are independent Poisson processes with rate parameters \( (1 - a)\lambda, (1 - a)\lambda, a\lambda \). Show that the correlation between \( X(t) \) and \( Y(t) \) is \( a \), regardless of the time evolved \( t \).
Figure 8.18 The probabilities of Norway winning (solid line), of Spain winning (dashed line), and the match ending in a draw (lower dotted line), as a function of real match time, for the Women’s Handball European Championship finale December 2014, based on the independent Poisson processes assumption. See Exercise 8.11 for a more elaborate model.

(d) Figure 8.18 is based on the finale for the Women’s Handball European Championships tournament December 2014, where Norway won 28-25 against Spain. It depicts the real time real excitement plots, giving the probabilities \( p_N(t), p_S(t), p_D(t) \) of Norway winning, of Spain winning, and of the match ending in a draw by 60 minutes, as a function of the match time. Specifically, \( p_N(t) = P\{N(t) + N' > S(t) + S'\} \), and so forth, where \( N(t) \) and \( S(t) \) count the goals scored by the two teams up to time \( t \), and the final results are \( N(t) + N' \) vs. \( S(t) + S' \), where \( N' \) and \( S' \) are taken there as independent Poisson with equal parameters \( \lambda(60 - t) \). Thus Spain’s best moment in the match occurs after 18 minutes, leading 10–5, where the chance of Spain winning is 76.9% to Norway’s 18.5% (and with chance 4.6% of having a draw); when Norway leads 26–22 with 8 minutes to go the three probabilities are 90.2%, 4.9%, 4.9%. Carry out these computations and produce a version of Figure 8.18.

(e) Now construct a parallel figure, where \( N' \) and \( S' \) have the more relevant but complicated bivariate Poisson distribution given earlier, with mean parameters \( \lambda(60 - t) \) but with correlation \( a = 0.209 \). Compare the two figures. Apply these algorithms to display real time real excitement plots for other handball and football matches, in particular predicting who wins, as a function of real match time.

(f) Finally construct more elaborate methods that do not merely give point estimates (f) for these \( p_N(t), p_S(t), p_D(t) \) probabilities, but full confidence curves for each of these, taking on board uncertainties in the parameter estimates for \( \lambda \) and (in particular) \( a \). A cautious better could, for example, decide to put money on Norway only when say the full 80% confidence interval for \( p_N(t) \) lies to the right of a threshold value.

8.12 Leukaemia: These are some details and amendments relating to Example 8.6.

(a) For the model with \( Y \) exponential with mean \( \lambda_i = \exp(\beta_0 + \beta_1(x_i - \bar{x})) \), estimate parameters and compute the AIC score.
Exercises

8.13 Log-polynomial densities and exponential tilting: Apply the (8.13) model to the 799 MHAQ data of Example 8.2 for different model orders. Show that the best fit, according to the AIC, is for $p = 4$. For that order, compute and display optimal confidence distributions for the parameters $\theta_1, \ldots, \theta_4$. Compare this model to that of the tilting expansion method (8.14) of Example 8.7. For the latter, which turns out to have a clearly better AIC score, compute the optimal confidence distribution for the tilting parameter $\gamma$. We note that Mogstad (2013) works with such tilting schemes, creating a semiparametric bridge from the usual fitted normal density to the nonparametric via exponential tilting, by slowly increasing the model order with larger sample sizes, via both AIC and FIC methods.

8.14 Tilting the Poisson: In the spirit of Example 8.7, consider tilting the Poisson model via the (8.14) idea, using

$$f(y, \theta, \gamma) = \exp(-\theta)(\theta^{y^2}/y!) \exp(\gamma y^2 - k(\theta, \gamma))$$

for $y = 0, 1, 2, \ldots$, with $k(\theta, \gamma) = \log(\sum_{y=0}^{\infty} \exp(-\theta)(\theta^{y^2}/y!) \exp(\gamma y^2))$. This is an exponential family in $(y, y^2)$, and $\gamma$ reflects the degree of potential departure from the Poisson distribution. Consider the dataset

3, 1, 1, 4, 0, 4, 0, 4, 8, 1, 5, 3, 4, 3, 7.

(a) Compute the empirical variance to mean ratio $W_n$ and find the associated $p$-value as a test for Poisson-ness, that is, $\gamma = 0$. For this use first the simple normal approximation based on $\sqrt{n}(W_n - 1) \rightarrow_d N(0, 2)$ under Poisson conditions (cf. Exercise A.1), and then the $p$-value $p = P_{0}(W_n \geq 1.75 | \sum_{i=1}^{n} Y_i = 48)$, associated with the conditional test given the observed $\sum_{i=1}^{n} Y_i$.

(b) Compute and display the approximate confidence curve $cc(\gamma)$ using the deviance, and also an approximate 95% interval for $\gamma$. Supplement this analysis with a simulation-based calculation of the optimal confidence distribution for $\gamma$.

8.15 Strauss: Consider the Strauss model and the little dataset worked with in Example 8.9, with the parametrisation $f_0(x, a) = k_0(a) \exp(-aM(x))$, and $M(x)$ the number of pairs among $n = 25$ points $x_i$ closer to each other than $r = 0.15$. The parametrisation used in (8.16) corresponds to $\rho = \exp(-a)$. For the dataset in question, $M(\text{obs}) = 4$.

(a) Use simulations to compute the maximum likelihood $\hat{a}$ and hence $\hat{\rho} = \exp(-\hat{a})$, and also the variance of $M(X)$ at that position in the parameter space. You may attempt to construct the required simulation schemes from scratch, or perhaps consult the spatstat package of Baddeley and Turner (2005) and Baddeley et al. (2015).

(b) Now assume that four further point patterns are observed too, supposed to stem from precisely the same Strauss model and with the same parameter, and with the same number $n = 25$ of points inside each square, with $M(x)$ equal to 5, 3, 5, 6. Compute the maximum likelihood estimate $\hat{a}$ and the variance of $\hat{M} = (1/5) \sum_{i=1}^{5} M(x_i)$ at that position in the parameter space. Finally compute and display the optimal confidence distribution for $\rho$. 

8.16 **Heteroscedasticity:** Consider the model used to analyse the speedskating data in Example 8.11, with independent observations $Y_i = x_i^\gamma + \sigma \exp(\psi z_i) + \epsilon_i$ with standard normal $\epsilon_i$, where $x_i$ and $z_i$ are of dimensions $p$ and $q$, respectively.

(a) Find the maximum likelihood estimators $\hat{\beta}(\phi)$ and $\hat{\gamma}(\phi)$ for fixed heteroscedasticity parameter $\phi$. Use this to find the profiled log-likelihood function $\ell_{\text{prof}}(\phi)$, which may then be maximised separately, yielding in its turn $\hat{\beta}(\hat{\phi})$ and $\hat{\gamma}(\hat{\phi})$. This optimisation, in lower dimension $q$, is often easier numerically than attempting to optimise directly in dimension $p + q + 1$.

(b) For the speedskating data, fit the model, with parameters $a,b,\sigma,\phi_1,\phi_2$, and find confidence distributions for the heteroscedasticity parameters $\phi_1$ and $\phi_2$, as well as for the $b$.

(c) For three skaters with personal best 6:05, 6:10, 6:20 on the 5k, find confidence distributions for the expected value of their 10k abilities, that is, for the appropriate $a + bx_0$.

8.17 **Doubly linear regression models:** Consider the survival analysis worked with in Example 8.12 using the linear-linear gamma model, with $a_i = \exp(a_0 + a_1 x_{i,1} + a_2 x_{i,2} + a_3 x_{i,3})$ and $b_i = \exp(b_0 + b_1 x_{i,1})$. Find confidence distributions for $a_1, a_2, a_3$ and for $b_3$. Find also the confidence distribution for the median time to event (i.e., death of liver transplantation) for a patient with given values of covariates $x_{1,1}, x_{3,2}, x_{3,3}$. Illustrate this for a few types of patients.

8.18 **Stretching Poisson regression models:** Consider the probability distribution

$$ f(y; \mu, \kappa) = \exp(-\mu) \frac{\mu^y}{y^y} \frac{1}{k(\mu, \kappa)} $$

for $y = 0, 1, 2, \ldots$.

(a) Show that the variance to mean ratio for this distribution is greater than 1 for $\gamma < 1$ and smaller than 1 for $\gamma > 1$.

(b) Consider the bird species data encountered in Table 4.5 and Exercise 4.18, with number of species $y$ along with covariates $x_{1}$ and $x_{2}$. Compare ordinary Poisson regression, which uses $\mu_i = \exp(x_i^\gamma)$ and $\gamma = 1$, to the extended model that uses $\gamma$ as an extra free parameter. Demonstrate that this extended overdispersion model does better than ordinary Poisson, in terms of the AIC score. Also compute and display an approximate confidence distribution for $\gamma$, using the deviance method of Section 3.4. (The point estimate is 0.568, and a 90% interval is from 0.275 to 1.010.)

(c) In the spirit of generalised linear-linear models, as in Section 8.8, investigate the model with $\mu_i = \exp(x_i^\gamma)$ and $\gamma_i = \exp(x_{i,1} + a_1 w_i)$, where $w_i = (x_{i,2} - \bar{x}_2)/sd(x_2)$. Show that this yields better AIC score than the version with a fixed $\gamma$.

(d) For various positions $(x_{1,0}, x_{2,0})$ in the covariate space, compute and display the fitted probability point mass functions $f_1(y), f_2(y), f_3(y)$ from the three models, and comment on what might be lost when applying the standard Poisson regression tools.

(e) Compare your results from this stretched Poisson model to those reached in Exercise 4.18 using negative binomial regression to account for overdispersion. Which of the approaches leads to the best AIC score?

8.19 **Cross-country ski-sprint:** This exercise pertains to some of the details of Example 8.8. Access the rank data for the 114 top events (57 for men, 57 for women), from the CLP book website www.mn.uio.no/math/english/services/knowledge/CLP/.

(a) Suppose positions 1, $\ldots$, $N$ are to be labelled A or B, with $n_A$ of type A and $n_B$ of type B. Assume all $\binom{N}{n_A}$ positions $X_1, \ldots, X_{n_A}$ are equally likely and consider $Z_A = \sum X_i$. Show that $\mathbb{E} Z_A = \frac{1}{2} n_A (N + 1)$ and that $\text{Var} Z_A = \frac{1}{12} n_A n_B (N + 1)$. 


8.10 Exercises

(b) For a ski sprint final with six skiers, let, as in Example 8.8, $Z_A$ be the sum of the ranks of the A skiers. Show that $Z_A$ has mean 7, 10.5, 14 for finals of type (2, 4), (3, 3), (4, 2), and that its variance is 4.667, 5.250, 4.667 for these three types.

c) Consider the null hypothesis $H_0$ of fairness, that the 15 combinations for $(X_1, X_2)$ ranks for the two A skiers are equally likely for the (2, 4) type finals, that the 20 combinations for $(X_1, X_2, X_3)$ ranks for the three A skiers are equally likely for the (3, 3) type finals, and finally that the 15 combinations for $(X_1, X_2, X_3, X_4)$ ranks for the four A skiers are equally likely for the (4, 2) type finals. With $Z_A = \sum X_i$ we know the mean and variance of $Z_A$ from (b), for the three types (2, 4), (3, 3), (4, 2). Give the normal approximation to the null hypothesis distribution of $Z_A^*$, the sum of all $Z_A$ over the 57 events. For the men, there were 34, 9, 14 finals of these three types, and for the women there were 19, 15, 23. Test the null hypothesis using $Z_A^*$, via the normal approximation, or by simulation of its null distribution.

d) For the men and the women separately, compute the log-likelihood function $\ell(\theta)$ using the balance tilting model of Example 8.8, and reproduce a version of Figure 8.10.

e) Compute and display the confidence distribution for the parameter $\alpha$, the probability that the winner of the final is an A skier. This may be analysed using $\alpha = \frac{1}{4}\alpha_{24} + \frac{1}{5}\alpha_{33} + \frac{1}{4}\alpha_{32}$, where $\alpha_{24}$ is the probability that the winner is an A skier in a final of type (2, 4), and so on. Make a version of Figure 8.19, also supplemented with a pointwise 90% confidence band.

Figure 8.19 The probability that the winner in an Olympic final is one of the skiers from the first of the two semifinals, as a function of the imbalance parameter $\theta$ of Example 8.8. Olympic fairness corresponds to $\theta = 0$ and the probability being $\frac{1}{2}$. Analysis shows $\theta$ is negative, and more so for the men than for the women.
9

Confidence distributions in higher dimensions

This chapter is concerned with confidence distributions for vector parameters, defined over nested families of regions in parameter space. For a given degree of confidence the region with that confidence is a confidence region in the sense of Neyman. Such regions are often called simultaneous confidence regions. They form level sets for the confidence curve representing the confidence distribution. Analytic confidence curves for vector parameters, and also for parameters of infinite dimensions, for example, all linear functions of the mean parameter vector in the normal distribution, are available in some cases. Some of these are reviewed and certain generalisations are discussed.

9.1 Introduction

Fisher’s general concept of fiducial distribution is difficult in higher dimensions. Joint fiducial distributions are not subject to ordinary probability calculus. Marginals and other derived distributions need in fact not be fiducial; cf. also Chapter 6. For vector parameters one must therefore settle for a less ambitious construct to capture the inferential uncertainty. Pitman (1939, 1957) noted that the fiducial probabilities in higher dimensions must be restricted to sets of specific forms, to avoid inconsistencies. One consequence of this insight is that only certain types of dimension reduction by integrating the higher dimensional fiducial density are valid. See Chapter 6 for examples of inconsistencies when not observing these restrictions, which usually are hard to specify.

Neyman (1941) was more restrictive. He looked only for a confidence region of specific degree. We lean towards Neyman and define confidence distributions in higher dimensions as confidence distributed over specified nested families of regions.

We first look at confidence distributions for the mean vector \( \mu \) in a multinormal distribution obtained from a sample, of dimension \( p \), say. When the covariance matrix is known, the confidence distribution is simply, as will be seen, a multivariate normal distribution about the sample mean. This is obtained from a natural multivariate pivot, yielding the cumulative distribution function \( C(\mu) \) assigning confidence to intervals \( \{ m : m \leq \mu \} \) in \( \mathbb{R}^p \). Thus confidence is also assigned to intervals and rectangles in \( \mathbb{R}^p \). The fiducial debate clarified that fiducial probability and hence confidence cannot be extended to the Borel sets. The confidence density for \( \mu \) is thus not particularly useful, although directly obtained. Though integration over intervals yields valid confidence statements, integration over other sets might not do so. We might also limit the support of the multinormal confidence distribution to the Mahalanobis ellipsoids around the sample mean.
The confidence distribution is then conveniently represented by the confidence curve, which turns out to be a probability transform of the deviance function.

When the covariance matrix is unknown and must be estimated, the confidence distribution is found from Hotelling’s $t^2$ distribution. The confidence curve will then have elliptic contours, and is also a probability transform of the deviance function.

In parametric models the deviance function evaluated at the parameter value will often have a distribution close to a chi-squared distribution. When the distribution of the deviance is close to invariant, approximate confidence curves are available by appropriate probability transforms. In complex models the distribution of the deviance might be different from chi-squared when the data are weak. Simulation, usually in the form of parametric bootstrapping, will then be a good tool for estimating the deviance distribution and to investigate its invariance. The fine-tuning methods of Chapter 7 are also very useful in this regard.

The model parameter $\theta$ is often of high dimension, while only a few components, or otherwise derived parameters, that is, functions of $\theta$, might be of primary interest. There is then a need to reduce the full confidence distribution to marginal confidence distributions for the parameters of interest. In general this cannot be done by integrating out the confidence density, as was learned from the fiducial controversy. The distribution obtained by integrating the confidence density might, however, approximately be a confidence distribution. Whether the approximation is satisfactory must be checked, usually by simulation. Another method of marginalisation is to first compute the profile likelihood for the interest parameter and then to probability transform the related deviance function. Bootstrapping is another method. From the bootstrap sample a bootstrap distribution for the interest parameter is used to form a confidence distribution, say by the abc method of Section 7.6.

There are often confidence intervals available for the individual coordinates of the model parameter $\theta$ or of a parameter vector of interest. These can be transformed to a box-shaped confidence curve by adjusting their levels to becoming a collection of simultaneous confidence intervals. The box-shaped confidence curves coming out of this adjustment will be called product confidence curves.

In Section 14.6 an application involving Norwegian income is considered, and a product confidence curve is obtained for income growth from year 2000 to 2010, defined as the ratio of quantiles $\psi(p) = q_{2010}(p)/q_{2000}(p)$, over a range of $p$. This confidence curve is graphed as a graded simultaneous confidence band around the observed quantiles.

The adjustment necessary to make a collection of individual confidence curves into a product confidence curve is of some interest. We suggest comparing this adjustment to the adjustment function used for stochastically independent component confidence curves. A measure of collective dependency in the collection of individual confidence curves is suggested, in Section 9.7.

### 9.2 Normally distributed data

Consider a sample $Y_1, \ldots, Y_n$ from the $p$-dimensional multivariate normal distribution $N_p(\mu, \Sigma)$ with known covariance matrix $\Sigma$. The Mahalanobis ellipsoids around the observed
Confidence distributions in higher dimensions

mean vector $\tilde{Y}$ are based on $n(\mu - \tilde{Y})'\Sigma^{-1}(\mu - \tilde{Y})$ being chi-squared distributed with $p$ degrees of freedom. Let $\Gamma_p$ be the distribution function of this chi-squared distribution. Then

$$\{\mu : n(\mu - \tilde{Y})'\Sigma^{-1}(\mu - \tilde{Y}) \leq \Gamma_p^{-1}(1 - \varepsilon)\}$$

is Neyman’s confidence region of degree $1 - \varepsilon$ for the mean parameter $\mu$. These shall also be the level sets for our confidence curve for $\mu$. This confidence curve is then

$$cc_\varepsilon(\mu) = \Gamma_p(n(\mu - \tilde{Y}_{\text{obs}})'\Sigma^{-1}(\mu - \tilde{Y}_{\text{obs}})).$$

A $p$-dimensional pivot is available in this case, namely $\sqrt{n}/\Sigma^{-1/2}(\mu - \tilde{Y})$. The pivot is $N_p(0, I)$ distributed, with say $N_p$ as its cumulative distribution function and $np$ its density function. As in the univariate case the cumulative confidence distribution is found by applying the pivotal distribution function to the pivot,

$$C(\mu) = N_p(\sqrt{n}/\Sigma^{-1/2}(\mu - \tilde{Y})).$$

The confidence density is

$$c(\mu) = \sqrt{n}/|\Sigma|^{-1/2}np(\sqrt{n}/\Sigma^{-1/2}(\mu - \tilde{Y})).$$

If say $p = 2$ the confidence of an interval in $\mathbb{R}^2$, say the product set $[a_1, b_1] \times [a_2, b_2]$, is

$$C(a_1, b_1) - C(a_1, b_2) - C(a_2, b_1) + C(b_1, b_2).$$

In Section 9.5, confidence for product sets based on continuous intervals for each coordinate at $\mu$ is discussed.

Singh et al. (2007) noted that linear parameters $\theta = A\mu$, with $A$ a $q \times p$ matrix, have confidence distributions obtained by integrating the joint confidence density. This is also the case in some other situations, and they suggested the term confidence distributions in the linear sense ($\ell$-CD).

To avoid inconsistencies we restrict confidence to be distributed over Neyman’s confidence ellipsoids, that is, the level sets of the preceding confidence curve. In this case dimension reduction by integrating the joint confidence density is, however, possible for many derived parameters. For a subset of $q$ coordinates, for example, integrating out the other yields the right $q$-dimensional normal distribution. Another example is for $\psi = \mu_1/\mu_2$. By the method used in Section 4.6 for the Fieller problem, integrating out yields the cumulative confidence distribution

$$C(\psi) = \Phi\left(\sqrt{n}\frac{\psi \tilde{Y}_2 - \tilde{Y}_1}{\sigma_1^2 + 2\psi \rho_{1,2} \sigma_1 \sigma_2 + \psi^2 \sigma_2^2}\right).$$

The level sets of the confidence curve, that is, the Mahalanobis ellipsoids, that we suggest should be given confidence, are also the level sets of the confidence density for $\mu$. Because in this case the likelihood and the confidence density are equal, $L(\mu) = c(\mu)$, the likelihood function and the confidence curve have identical level curves. The confidence curve is simply the chi-squared transform of the deviance function, $cc(\mu) = \Gamma_p(\log L(\mu))$.

For an example where integration goes wrong for finding the confidence distributions for most derived parameters, consider the following.
9.2 Normally distributed data

Example 9.1 When integration goes wrong
The confidence distribution for the parameter vector \( \theta = (\mu, \sigma) \) of the normal distribution is found from the two standard pivots. The t-pivot for \( \mu \) is independent of the chi-squared pivot for \( \sigma \). The two marginal confidence densities lead to a confidence density for \( \theta \) by their product, yielding confidence to intervals in the half-plane. Bartlett (1936) claimed that integration to have the confidence density for \( \psi = a\mu + b\sigma \) goes wrong when \( a \) and \( b \) are constants. Fisher found Bartlett to be wrong, the marginal density is indeed a confidence density in the sense that it distributes confidence correctly over intervals for \( \psi \). It has been found that for any continuous function \( \psi \) of \( \theta \) that is varying in both coordinates, and is not linear, as Bartlett’s parameter, integration goes wrong. See also Section 6.2.

Confidence regions for the mean vector \( \mu \) of a \( p \)-dimensional normal distribution with unknown covariance matrix were looked for after the discovery of the confidence intervals based on the t-distribution when \( p = 1 \). This problem was solved by Working and Hotelling (1929), followed up by Hotelling (1931). Their elliptical confidence regions are actually the same as those found when the covariance matrix is known, when \( \Sigma \) is substituted by its estimate. The confidence of an ellipsoid is of course affected by the covariance matrix being unknown and is estimated. These ellipsoids are determined by

\[
r^2 = n(\mu - \bar{Y})' \Sigma^{-1}(\mu - \bar{Y}),
\]

with the covariance matrix estimated by the sample covariance matrix

\[
\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'.
\]

Then \((n-p)r^2/(p(n-1))\) is F-distributed with \( p \) degrees of freedom in the numerator and and \( n-p \) degrees of freedom in the denominator; cf. Hotelling (1931). Let \( F_{p,n-p} \) be the distribution function of this F-distribution. The confidence curve for \( \mu \) is consequently

\[
cc(\mu) = F_{p,n-p} \left( \frac{n-p}{p(n-1)} n(\mu - \bar{Y})' \hat{\Sigma}^{-1}(\mu - \bar{Y}) \right).
\]

In some cases a confidence curve is needed for more complex parameters than vectors. In the linear normal model, for example, it might be desirable to have a confidence curve for the parameter \( \theta \) representing all linear functions of the mean parameter of the model. With \( w \) a coefficient vector, \( w'\mu \) is a coordinate of \( \theta \). Even when the coefficient vectors are restricted to have unit length, there are more than countably many coordinates. The method of multiple comparison of Scheffé (1959) yields in fact a confidence curve for \( \theta \) when the coefficient vectors are restricted to a linear subspace of a given dimension. This confidence curve or distribution is difficult to grasp or depict in its total complexity. The simultaneous confidence intervals or regions in low dimensions could, however, be graphed. For \( q \) the dimension of the coefficient vectors \( w \), and for \( \hat{\sigma}_w \) being the estimated standard deviation of \( w'\mu \), these confidence intervals at simultaneous level \( \alpha \) are \( w'\mu \pm \hat{\sigma}_w \sqrt{qF_{q,n-q-1}(\alpha)}^{1/2} \). For two linear parameters the simultaneous confidence regions are the corresponding ellipsoids.
Confidence distributions in higher dimensions

The problem of dimension reduction of this complex confidence distribution of Scheffé is for linear transforms of the infinite-dimensional parameter $\theta$ solved by applying his method to appropriate linear subspaces.

Note also that the simultaneous confidence intervals are equal to t-intervals, but with confidence level adjusted. The Scheffé confidence curve might therefore be regarded as a product confidence curve, to be discussed later, of the one-dimensional t-confidence curves.

Tukey’s method of simultaneous inference for pairwise differences (Scheffé, 1959) might be regarded as a problem of developing a product confidence curve for the vector of pairwise differences, and could be solved by simulation as outlined below, also when the studentised range distribution is of questionable validity, say because of deviations from the normal distribution.

9.3 Confidence curves from deviance functions

The deviance function $D(\theta) = 2 \log \left( \frac{L(\hat{\theta})}{L(\theta)} \right) = 2(\ell(\hat{\theta}) - \ell(\theta))$, evaluated at the true value of the parameter, is asymptotically chi-squared distributed with $p$ degrees of freedom as the sample size increases to infinity, when the dimension of the model parameter $\theta$ is $p$. A corresponding result holds also for the profile deviance. Conditions for this to hold are discussed in Chapter 2. This important result implies that $cc(\theta) = \Gamma_p(D(\theta))$ is an approximate (and asymptotically correct) confidence curve for the model parameter.

Accuracy is improved by the methods of Chapter 7, for example, median-bias correction. By results discussed in that chapter, the rate of convergence for confidence curves based on bias corrected deviance functions are third-order correct in dimension 1. When the median of the maximum likelihood estimator is available, it is good practice to bias-correct also in the $p$-dimensional case. What to understand as the median of a stochastic vector is open to choice. One possibility is the geometric median minimising the expected Euclidean distance in the distribution. Another possibility is to use the coordinate-wise medians. When the parameter is natural, say because it is of primary interest, the coordinate-wise median is usually a good choice. This median is, however, not invariant to transformations that is not a product of monotone coordinate-wise transformations.

Example 9.2 Ratio of two standard deviations

Consider a situation with two variance parameters $\sigma_1^2$ and $\sigma_2^2$, with independent estimators distributed proportional to chi-squared distributions with 9 and 4 degrees of freedom respectively. We are primarily interested in $\psi = \sigma_2^2 / \sigma_1^2$. The maximum likelihood estimator $\hat{\psi}$ has median 0.95 while $\hat{\sigma}_1$ has median 0.96 $\sigma_1$. Applying median-bias correction to each component separately, the deviance function yields the confidence curve that is contoured in Figure 9.1. The maximum likelihood estimates behind this confidence curve are $\hat{\sigma}_1 = 1$ and $\hat{\psi}_2 = 2$. The distribution of the deviance at the true value is independent of the parameter, and is found by simulation. It is nearly a $\chi^2_2$. The confidence curve based on the bias-corrected profile deviance for $\psi$, with null distribution nearly that of $1.08 \chi^2_2$, is shown in the lower panel, together with the exact tail-symmetric confidence curve based on the F-distribution. The two confidence curves are indistinguishable.
9.4 Potential bias and the marginalisation paradox

A joint fiducial distribution does in general not lead to valid marginal fiducial distributions for derived parameters. This is illustrated by the length problem (Example 6.3). Does dimension reduction of a confidence curve obtained by probability transforming the deviance function derived by profiling have the same difficulty? Let \( \theta = (\psi, \chi) \) where \( \psi \) is scalar, and let \( D_{\text{prof}}(\psi) = D((\psi, \hat{\chi}(\psi)) \) be the profile deviance, where \( D(\psi, \chi) \) is the deviance of the full parameter. Here \( \hat{\chi}(\psi) \) is the maximum likelihood estimate of the nuisance parameter when \( \psi \) is fixed. Assume that \( D_{\text{prof}}(\psi) \) has distribution function \( H_\psi \), free of \( \chi \), making \( cc(\psi) = H_\psi(D_{\text{prof}}(\psi)) \) an exact confidence curve. From the Wilks theorem of Chapter 2, \( H_\psi = \Gamma_1 \), the \( \chi^2_1 \) distribution function, in smooth models. Despite being exact in the sense that \( cc(\psi) \) is uniformly distributed when \( \psi \) is the true parameter, the confidence curve might be biased in the sense that its anchoring point, the maximum likelihood estimate \( \hat{\psi} \) is not median unbiased. In that case the confidence intervals obtained as the level sets \( \{ \psi : cc(\psi) \leq a \} \) will typically not be tail symmetric. That is, they will be of the form \( (C^{-1}(a), C^{-1}(1-b)) \) with \( a + b = a \) but with \( a \neq \frac{1}{2} a \).

In the length problem, and in other models, the joint Bayesian posterior based on a flat prior might yield posterior distributions for derived parameters that are biased in the frequentist sense. In the discussion of Dawid et al. (1973) this was pointed out. The paper emphasised, however, what they called ‘the marginalisation paradox’, that is, that the Bayesian might conclude differently from the same prior when the posterior for \( \psi \) is
Confidence distributions in higher dimensions

based on the statistic $T$, and when it is based on the marginal posterior for $\psi$, even when the marginal posterior depends only on $T$. When the prior is flat the phenomenon is due to the integrated likelihood $\int L(\psi, x, X) \, d\chi = L_{int}(\psi, T)$ being different from the marginal likelihood $L_{marg}(\psi, T)$.

The same ‘paradox’ is present for confidence distributions. When the profile deviance $D_{prof}(\psi, T)$ depending only on $T$ is different from the marginal deviance based on $T$ the confidence curve based on the profile deviance will differ from that based on the marginal deviance. When $T$ yields a pivot, the related confidence distribution is clearly the one to choose. In the more complex situation, the procedure of choice might be to work with the modified likelihood; see Chapter 7.

9.5 Product confidence curves

When no analytic confidence curve for the parameter vector of interest, $\theta$, is available the product method of Beran (1988a) can be used when confidence curves are available for each component of the vector. Beran (1988a) developed further the simultaneous confidence bands for survival functions of Nair (1984) and Hjort (1985). Beran dealt with confidence intervals, but we will as in Schweder (2007) present this idea in the context of confidence curves. From a collection of confidence curves, called the root, the product confidence curve for $\theta$ is obtained by forming its level sets as the product of level sets for each component of the root, at various levels of nominal confidence. The product confidence curve will have rectangular level sets. These rectangular level sets will typically have smaller coverage probabilities than what the individual intervals have. The product confidence curve is obtained by probability transforming the product formed by the root.

In more precise terms, let $cc_t$ be the confidence curve for $\theta_t$, varying only in the $t$th coordinate. Each $cc_t(\theta)$ is uniformly distributed at the true value of the parameter. The product confidence curve has nested rectangular level sets determined by $\max_t cc_t(\theta) \leq K^{-1}(\alpha)$ with confidence level $K(\alpha)$, for a suitable transformation $K: (0,1) \rightarrow (0,1)$. The level sets of the resulting confidence curve are the product sets of level sets of the individual confidence curves $cc_t$. This is the reason for the term product confidence curve.

**Definition 9.1 (Product confidence curve)** If $\max_{t \leq T} cc_t(\theta)$ has a distribution with distribution function $K_\theta$, then

$$cc(\theta) = K(\max_{t \leq T} cc_t(\theta))$$

is a product confidence curve. If $K_\theta = K$ is independent of $\theta$, $cc(\theta)$ is a balanced product confidence curve, and $K$ is the adjustment function.

We shall primarily be concerned with product confidence curves, that is, in cases where $K_\theta = K$ does not depend on $\theta$. That product confidence curves indeed represent collections of simultaneous confidence intervals is seen from

$$P_\theta(\text{cc}(\theta) \leq a) = P_\theta \left\{ \max_{t \leq T} cc_t(\theta) \leq K^{-1}(\alpha) \right\} = \alpha$$

(9.1)

for all $\theta$ and $0 < \alpha < 1$, which is equivalent to $\cap_{t=1,...,T} \{ cc_t(\theta) \leq K^{-1}(\alpha) \}$. 
9.5 Product confidence curves

Figure 9.2 Left: The balanced product confidence curve for \((\psi = \sigma_2/\sigma_1, \sigma_1)\), based on the individual confidence curves in Example 9.3. Right: The adjustment function (solid line) and the Bonferroni adjustment function (dashed line).

Example 9.3 Ratio of standard deviations via product confidence

In Example 9.2 a confidence curve for \(\theta = (\psi, \sigma_1)\) was obtained, with \(\psi = \sigma_2/\sigma_1\), based on the deviance function. A product confidence curve would have been an alternative. The root has two elements. The first is the tail-symmetric confidence curve \(cc_\psi\) based on the F-pivot for \(\psi\). The second is the tail-symmetric confidence curve \(cc_{\sigma_1}\) for \(\sigma_1\) based on chi-squared pivot. The product confidence curve is \(cc(\theta) = K(\max\{cc_\psi(\psi), cc_{\sigma_1}(\sigma_1)\})\). The adjustment function \(K\) is the distribution function of

\[ V = \max\{|1 - 2 F_{\nu_1}(X)|, |1 - 2 F_{\nu_2}(Y/\nu_2)/(X/\nu_1)|\} \]

where \(X\) and \(Y\) are independent chi-squared distributed with \(\nu_1\) and \(\nu_2\) degrees of freedom respectively. The adjustment function is clearly independent of \(\theta\). Figure 9.2 shows the adjustment function \(K\) together with the adjustment function determined by the simple Bonferroni method. For the latter, see Exercise 9.5.

The product confidence curve is called balanced because all elements in the root are treated equally. Balance is not always desirable for product confidence curves. Some components of the parameter might be of more interest than others. For the more important components of \(\theta\), narrower component level sets might be desirable for the rectangular product level sets, on the expense of wider level sets for less interesting components. A practical weighting scheme is provided by component-specific transformations of the form

\[ cc_w(\theta) = K_w\left(\max_{t \leq T} cc_t(\theta)^{1/w_t}\right). \]
Confidence distributions in higher dimensions

The adjustment function $K_{\theta}$ depends on the weighting scheme, and is now the cumulative distribution function of $V_w = \max_{t \leq T} cc_t(\theta)^{1/w_t}$.

Leaving unbalanced product confidence curves aside, when are the adjustment functions independent of the parameter? Note first that all the component confidence curves in the root are uniformly distributed at the true value of the parameter. Variability in $K_{\theta}$ must therefore stem from dependencies between the root confidence curves varying with $\theta$. In many cases such varying dependency can be ruled out, but not always.

Example 9.4 Parameters of the binormal distribution

Consider a sample of size $n = 10$ from the bivariate normal distribution with known unit variances, but parameter $\theta = (\mu_1, \mu_2, \rho)$ where the first two elements are the mean parameters and the third is the correlation coefficient. A product confidence curve is desired for $\theta$. The two first members of the root are $cc_t(\theta) = |1 - 2\Phi((\mu_t - \hat{\mu}_t)/\sqrt{n})|$ for $t = 1, 2$. For the $\rho$ parameter, Fisher’s transform $F(r) = \frac{1}{2} \log \left\{ \frac{1+r}{1-r} \right\}$ is used. It turns out that there is a nonnegligible bias $b = \text{median} \{ F(r) \} - F(\rho)$ for small samples. The empirical correlation coefficient $r$ is consequently bias corrected on the transformed scale. There is some departure from standard normality in Fisher’s bias corrected pivot $F(\rho) - \{ F(r) - b \} \sqrt{n - 3}$. It is turned into a confidence curve $cc_3(\theta) = G(1 - 2\Phi(F(\rho) - \{ F(r) - b \} \sqrt{n - 3})$ by a distribution function $G$ estimated by simulation. The adjustment curve $K_{\theta}$ is clearly independent of the mean parameters, but not of the correlation coefficient (see Figure 9.3).

![Figure 9.3](image)

Figure 9.3 Adjustment functions for various values of the correlation function $\rho$, in Example 9.4. Each displayed function is based on 1000 simulations. There are two replicate functions for $\rho = 0.5$ and $\rho = 0.95$. 
Confidence curves are often found through simulation. From a bootstrap sample of size $B$ of $\tilde{\theta}$, confidence curves are made separately for each component. These make up the root. Each member of the root, $cc^*_i$ is constructed to make $V^*_i = cc^*_i(\theta^0)$ uniformly distributed, where $\theta^0$ is the reference value used in the simulation. These uniform variables are typically stochastically dependent. From the bootstrap experiment there are $B$ replicates of vectors with coordinates $V^*_i$. The adjustment function is now simply the distribution function of $\max V^*_i$, which is estimated from the bootstrap data.

In Chapter 7 we discuss a variant of the acceleration and bias corrected percentile intervals for one-dimensional parameters introduced by Efron (1987). The idea of the abc method is that on some transformed but monotonous scale $h$, $\tilde{\gamma} = h(\tilde{\psi})$ is normally distributed with mean $\gamma - b(1 + a^2)$ and variance $(1 + a^2)^2$, and with $\gamma = \theta(\psi)$. With a value for the acceleration constant $a$, which might take some effort to find, and for the bias constant $b = \Phi^{-1}(H(\psi))$, the tail-symmetric confidence curve found by the abc method is

$$cc_{abc}(\psi, \hat{\psi}) = |1 - 2C_{abc}(\psi)|,$$

$$C_{abc}(\psi, \hat{\psi}) = \Phi \left( \frac{\Phi^{-1}(H(\psi)) - b}{1 + a(\Phi^{-1}(H(\psi)) - b)} - b \right),$$

where $H$ is the distribution function of the bootstrap distribution for the estimator $\hat{\psi}$ assumed here to be based on $B = \infty$ replicates. The scale transformation is related to $H$ as

$$\theta(s) = (1 + a\tilde{\gamma}) \{\Phi^{-1}(H(s)) - b\} + \tilde{\gamma}.$$

Consider a root of confidence curves for the components of $\theta$, constructed by bootstrapping and applying the abc method for each member of the root, producing a confidence distribution and bootstrap replicates $C_{abc}(\theta, \theta^*_t)$ for each $t$. The root is turned into a confidence curve for $\theta$ when the adjustment function is available. When the bootstrapping is nonparametric the bootstrap data yields the adjustment function in an easy way. Then $K$ is actually the distribution function of

$$V = \max_{t \in T} |1 - 2H_t(\theta^*_t)|,$$

where $H_t$ is the distribution function of the bootstrapped component estimates $\theta^*_t$.

The basis for this result is that in nonparametric bootstrapping, the bootstrapped parameter estimate $\theta^*$ has nearly distribution $P_{\tilde{\theta}}$ when the estimate obtained from the observed data $\tilde{\theta}$ have distribution $P_{\theta}$. The distribution of $V$, with $K$ as its distribution function, is thus found from the joint distribution of $C_{abc}(\tilde{\theta}, \theta^*_t)$. Consider a given component $t$. Using (9.3) for $h_t$ and utilising the invariance property of confidence distributions,

$$C_{abc}(\tilde{\theta}, \theta^*_t) = C_{abc}(\tilde{\theta}, \gamma^*_t) = \Phi \left( \frac{\tilde{\gamma}_t - h_t(\theta^*_t)}{1 + a_t\tilde{\gamma}_t} - b_t \right) = 1 - H_t(\theta^*_t).$$

This shows that the adjustment function $K$ indeed is given by (9.4). At least when it can be regarded as independent of $\theta$, which often is reasonable.

The number of bootstrap replicates $B$ is in practice large but finite. The bootstrap data are used to calculate the confidence curves in the root, and also the adjustment function $K$. 
Confidence distributions in higher dimensions

For a sufficiently large number of replicates these results are correct to first order in the sample size of the observed data when the assumptions behind the abc method holds for each member of the root.

In some cases the goal is to estimate a parameter of infinite dimension, say a Beran (1988a) showed that sampling from the infinite index set solves the problem, however, at least asymptotically. For regression curves over compact support, Claeskens and Van Keilegom (2003) found asymptotically correct simultaneous confidence sets based on bootstrapping.

9.6 Confidence bands for curves

Confidence bands are used to display the estimation uncertainty for curve estimates. A curve \( f(t) \) where \( t \) runs over an index set \( I \), usually an interval of the real line or a finite set, is estimated by \( \hat{f}(t) \). An upper function \( \hat{f}_u(t) \) and a lower function \( \hat{f}_l(t) \) computed from the data make up a confidence band. It is a pointwise confidence band of level \( \alpha \) if

\[
P(\hat{f}_l(t) \leq f(t) \leq \hat{f}_u(t) \text{ for all } t \in I) = \alpha
\]

(9.5)

It is a simultaneous confidence band of level \( \alpha \) if

\[
P(\hat{f}_l(t) \leq f(t) \leq \hat{f}_u(t) \text{ for all } t \in I) = \alpha;
\]

(9.6)

that is, the band is a confidence interval in \( \mathbb{R}^I \) and is thus a level set for a confidence curve on this set. Constructing the first type is often easy, whereas the second type of construction is harder, as it involves the joint distribution of all \( [\hat{f}_l, \hat{f}_u] \). Also, there would typically be different types of bands, say with different shapes, all managing the requirement of joint coverage probability \( \alpha \).

Example 9.5 Evolution of Olympic body mass index

In Example 3.8 we studied body mass index (BMI) for the speedskaters of the Vancouver Olympics 2010. We actually have similar data from previous Olympic games too, at least for many of the skaters, back to 10 male skaters in 1952 and 19 female skaters in 1960 (when ladies were first allowed to take part). Interestingly, the mean BMI index has been through some dynamic evolution over the years. Figure 9.4 gives a detailed close-up for the four Olympics 1998, 2002, 2006, 2010, displaying confidence curves for the four parameters, say \( \psi_{98}, \psi_{02}, \psi_{06}, \psi_{10} \). The BMI is decreasing, indicating a phasing out of the biggest muscular bodies in favour of the slimmer and technically superior types.

Whereas Figure 9.4 deals with four parameters in a pointwise confidence manner, that is, as with (9.5), Figure 9.5 displays simultaneous bands, that is, as with (9.6), concerning Olympic parameters \( \psi_{98}, \ldots, \psi_{10} \). The coverage probability \( \alpha \) in the latter figure equals 0.10, 0.50, 0.90. Constructing these bands, which are of the fixed-width type \( \psi_j \pm z_{\alpha} \), amounts to finding critical values \( z_{\alpha} \) such that

\[
G(z) = P(|\psi_j - \hat{\psi}_j| \leq z \text{ for all } 11 \, j) = \alpha.
\]

This is done via \( G(z) = P(\max\sigma_j N_j \leq z) = \prod \Gamma_1(z^2/\sigma_j^2) \), where \( \sigma_j \) is the standard deviation for the mean BMI estimator \( \hat{\psi}_j \), from which it is not difficult to find the critical
9.6 Confidence bands for curves

Figure 9.4  Confidence curves for the mean parameter of the body mass index distribution for Olympic male speedskaters, over four recent Olympics. The four curves are marked by their associated maximum likelihood tags corresponding to estimates 23.707, 23.911, 24.093, 24.268 at A (Vancouver, 2010), B (Torino, 2006), C (Salt Lake City, 2002), D (Nagano, 1998).

Figure 9.5  Simultaneous confidence bands for the mean parameter of the BMI distribution for Olympic male speedskaters, over the 11 games from 1972 to 2010, of degree 0.10 (inner band), 0.50 (middle band), 0.90 (outer band). Point estimates $\hat{\mu}_j$ are indicated with stars, accompanied by pointwise 0.90 confidence intervals.

values $z$ numerically. An alternative is to make the band have width proportional to the standard deviations, say $\psi_j \pm w_\alpha \sigma_j$, with $w_\alpha$ found from

$$H(w) = P(\frac{|\psi_j - \hat{\psi}_j|}{\sigma_j} \leq w \text{ for all } j) = \Gamma_1(w^2)^{11} = \alpha.$$
Confidence distributions in higher dimensions

Note that the individual pointwise 90% intervals become rather shorter than when using the simultaneous method.

The methods used in the preceding example are related to the product confidence curve methods of Section 9.5, and the calculations are relatively easy owing to the independence of the ensemble of estimators. Similar methods may be developed for other ensembles of parameters, perhaps via transformations. An instructive example is that of the probability parameters \( p_1, \ldots, p_k \) in a multinomial model, where one observes \( N_1, \ldots, N_k \), with \( N_j \) counting the number of independent trials ending up as being of type \( j \). The pointwise intervals are of the type 
\[
\hat{p}_j \pm z_{\alpha} \frac{\hat{p}_j (1 - \hat{p}_j)}{\sqrt{n}},
\]
with \( \hat{p}_j = N_j / n \) and \( z_{\alpha} \) the appropriate normal quantile. A simultaneous band, of the same type, but wider, that is, 
\[
\hat{p}_j \pm w_{\alpha} \frac{\hat{p}_j (1 - \hat{p}_j)}{\sqrt{n}},
\]
can be constructed, with \( w_{\alpha} \) the required quantile from the distribution of 
\[
D_n = \max \sqrt{n} |\hat{p}_j - p_j| / (\hat{p}_j (1 - \hat{p}_j)),
\]
which may be simulated. This method does not take the special multinomial structure into account, however. Good alternatives for such simultaneously valid confidence regions are
\[
R_A(\alpha) = \{ p : \Gamma_{k-1}(A_n(p)) \leq \alpha \} \quad \text{and} \quad R_B(\alpha) = \{ p : \Gamma_{k-1}(B_n(p)) \leq \alpha \},
\]
where
\[
A_n(p) = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \quad \text{and} \quad B_n(p) = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{N_j}
\]
are the two versions of the classical Pearson goodness-of-fit statistic. Both are converging to the \( \chi^2_{k-1} \) with sample size, and the approximation tends to work well also for moderate and small \( n \). In particular, \( A_n(p) \) and \( B_n(p) \) have distributions close to being constant across the parameter space, that is, they are pivots.

**Example 9.6 Carlsen versus Anand**

Magnus Carlsen and Viswanath Anand have since 2005 and up to and including the world championship match of 2013 played each other a total of 39 times (where we focus on classical matches and hence disregard rapid games and exhibition games). Of these, Carlsen and Anand have won 6 times each, with 27 games ending in remis. We take these 39 games to have been statistically independent and also assume, for the present purposes of illustration, that the winning and draw chances have remained the same over this period. Hence \( (6,6,27) \) result from a trinomial distribution, and Figure 9.6 indicates the confidence regions for the two winning probabilities. Note that the \( R_B \) regions are circles here, as Carlsen and Anand had won the same number of games (up to the end of 2013). Both sets of regions \( R_A \) and \( R_B \) define valid sets of confidence; their somewhat different shapes merely reflect that they distribute their confidences in different ways. See also Exercise 9.4.

We learn from these examples and calculations that simultaneous bands for a high number of parameters might become very broad, however. In fact the width of the band increases proportionally to \( \Gamma_1^{-1}(1 - d/n)^{1/2} \) for \( n \) such parameters, with \( d \) reflecting the confidence level, provided the estimators associated with the parameters are independent. Exercise 9.5 gives some details concerning this, in particular revealing that the widths of such bands are of size proportional to \( (\log n)^{1/2} \). In certain situations, involving parameters for which
9.6 Confidence bands for curves

Figure 9.6 Simultaneous confidence regions for the winning probabilities for Carlsen and Anand, based on two versions of the Pearson statistic, associated with \( A_n \) to the left and \( B_n \) to the right. The coverage levels are 0.50 and 0.90 for the indicated regions.

The background estimators are serially correlated, however, the bands may become stable. In particular this happens when it comes to simultaneous bands for cumulative distribution functions and cumulative hazard rates, as we shall briefly explain now.

**Bands for a distribution function**

Let \( Y_1, \ldots, Y_n \) be i.i.d. from some unknown continuous distribution function \( F \). With the usual empirical distribution function \( F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq t) \), consider the random functions

\[
A_n(t) = F_n(t) - F(t) \quad \text{and} \quad B_n(u) = A_n(F^{-1}(u)) - u = G_n(u) - u.
\]

The point of carrying out this time transformation is that \( G_n(u) = \frac{1}{n} \sum_{i=1}^{n} I(U_i \leq u) \), that is, the empirical distribution function in terms of \( U_i = F(Y_i) \), and these are independent and standard uniform. Functionals expressed in terms of the \( B_n \) functions are in particular pivots, with distributions not depending on \( F \) at all.

The functional associated with the Glivenko–Cantelli theorem and the Kolmogorov–Smirnov test statistic is the maximum absolute value,

\[
D_n = \|A_n\| = \max_t |F_n(t) - F(t)| = \max_u |G_n(u) - u| = \|B_n\|.
\]

Its distribution may easily be simulated and tabulated, for each given sample size \( n \); in fact,

\[
D_n = \max_{1 \leq i \leq n} \{|U_i - i/n|, |U_i - (i-1)/n|\}
\]

in terms of the order statistics of an i.i.d. uniform sample. Hence the fixed-width band

\[
R_n(\alpha) = \{t : F_n(t) - z_{n,\alpha} \leq t \leq F_n(t) + z_{n,\alpha} \text{ for all } t\}
\]

has exact confidence \( \alpha \), where \( z_{n,\alpha} \) is the corresponding quantile for \( D_n \).
Confidence distributions in higher dimensions

It is fruitful to understand these issues via approximations for larger \( n \), involving convergence of the empirical process \( Z_n = \sqrt{n}(F_n - F) \). A classical result in the literature on convergence of such empirical processes is that \( Z_n \rightarrow_d Z \), where \( Z(t) = W^0(F(t)) \) is a time-transformed Brownian bridge; specifically, \( Z \) is Gaussian, has zero mean, and covariance function

\[
\text{cov}[Z(t_1), Z(t_2)] = F(t_1)[1 - F(t_2)] \quad \text{for } t_1 \leq t_2.
\]

See, for example, Billingsley (1968) for the necessary theory and for proofs and applications of this result. One of the corollaries is that

\[
\sqrt{n}D_n = \max_t |Z_n(t)| \rightarrow_d D = \max_t |W^0(F(t))| = \max_{0 \leq s \leq 1} |W^0(s)|,
\]

and the distribution of \( D \) is tabulated (or may be simulated for any given occasion without much trouble). Hence the band

\[
F(t) \in F_n(t) \pm z_n / \sqrt{n} \tag{9.7}
\]

has accurate simultaneous coverage \( \alpha \), when \( z_n \) is the \( \alpha \) quantile of the distribution of \( D \).

Here (9.7) is the classical fixed-width Kolmogorov–Smirnov band, which, for example, may be used to test whether a given hypothesised distribution function fits the data – if all of it is inside such a 95% band, say, then one is tempted to accept the hypothesis in question. Various related bands may also be constructed, based on the same set of ideas and the empirical process convergence \( Z_n \rightarrow_d Z \) given earlier. One may, for example, prefer a band with width proportional to the pointwise one, and this is accomplished via the result

\[
M_n = \max_{a \leq s \leq b} \left[ F_n(0.5)[1 - F_n(0.5)] \right]^{1/2} \rightarrow_d M = \max_{a \leq s \leq b} \left[ |W^0(F(t))| \right]^{1/2},
\]

valid for a finite interval \([a, b]\) for which \( F(a) > 0 \) and \( F(b) < 1 \). But \( M \) is also the same as \( \max_{F(a) \leq s \leq F(b)} |W^0(s)| / \sqrt{3} \), and this distribution may be simulated, for the \([a, b]\) interval required. An illustration for a similar problem is given in the text that follows.

Bands for a cumulative hazard and a survival function

Consider the setup of Section 4.7, involving survival data \((t_i, \delta_i)\) for \( i = 1, \ldots, n \), where \( t_i \) is the perhaps censored lifelength for individual \( i \) and \( \delta_i \) is an indicator for noncensoring. Somewhat more formally, \( t_i = \min(t_i^\delta, c_i) \) and \( \delta_i = I(t_i^\delta < c_i) \) are what one may observe, where \( t_i^\delta \) is the real lifetime and \( c_i \) a censoring time. This would be the situation in a case where a statistical analysis needs to be carried out before a certain date, where some of the individuals under study are still alive. The \( t_i^\delta \) are assumed to stem from a lifetime distribution with hazard rate function \( \alpha(s) \), with cumulative hazard rate \( A(t) = \int_0^t \alpha(s) \, ds \) and associated survival function

\[
S(t) = P(t_i^\delta \geq t) = \exp(-A(t)).
\]

We shall explain how to construct both pointwise and simultaneous confidence bands for \( A \) and for \( S \).
9.6 Confidence bands for curves

As with (4.6), we shall work with the number at risk process and the counting process

\[ Y(s) = \sum_{i=1}^{\alpha} I[t_i \geq s] \quad \text{and} \quad N(t) = \sum_{i=1}^{\alpha} I[t_i \leq t, \delta_i = 1]. \]

The Nelson–Aalen and Kaplan–Meier estimators, for respectively \( A \) and \( S \), are then defined as

\[
\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{t_i \leq t} \frac{\Delta N(t_i)}{Y(t_i)},
\]

\[
\hat{S}(t) = \prod_{0 \leq s \leq t} \left(1 - \frac{dN(s)}{Y(s)}\right) = \prod_{0 \leq s \leq t} \left(1 - \frac{\Delta N(t_i)}{Y(t_i)}\right).
\]

Here \( \Delta N(t_i) \) is the number of observed transitions at time point \( t_i \); with continuous data, this would be one or zero, but there could be ties. Results in the literature concerning the empirical process \( \sqrt{n}(A - A) \) include the fact that there is convergence in distribution,

\[
\sqrt{n}[\hat{A}(t) - A(t)] \to_d W(\sigma(t)^2), \quad \text{where} \quad \sigma(t)^2 = \int_0^t \frac{\alpha(s)\,ds}{y(s)}.
\]

Here \( W(\cdot) \) is a Brownian motion process, with \( \text{Var} W(u) = u \), so that the limit process \( W(\sigma(t)^2) \) has variance \( \sigma(t)^2 \). Also involved here is \( y(s) \), the limit in probability of \( Y(s)/n \). The limiting variance here is consistently estimated using

\[
\hat{\sigma}(t)^2 = n \int_0^t \frac{dN(s)}{Y(s)^2} = \sum_{t_i \leq t} \frac{n\delta_i}{Y(t_i)^2}.
\]

Hence a pointwise confidence band can be constructed, as \( \hat{A}(t) \pm z_{\alpha}\hat{\sigma}(t)/\sqrt{n} \), with \( z_{\alpha} \) the quantile point of \( |N(0,1)| \). One may also construct a pointwise band via large-sample normality for the log-transform, that is,

\[
\hat{A}(t) \exp\{\pm z_{\alpha}\hat{\sigma}(t)/\sqrt{n}\}.
\]

This option has the advantage that the band does not slip below zero.

Our interest here lies with simultaneous bands, however, and as with the cumulative distribution function \( F \) above there are several venues. A direct solution, utilising the Brownian motion limit, starts from

\[
D_y = \max_{u \leq s \leq b} \sqrt{n}|\hat{A}(t) - A(t)| \to_d D = \max_{\sigma(t)^2 \leq u \leq \delta(t)^2} |W(u)|.
\]

This distribution of \( D \) may actually be found exactly, but in a somewhat cumbersome form, by combining results of Billingsley (1968, chapter 3) with the representation \( W(u) = W(c) + W(u - c) \) for \( u \in [c, d] \), where \( W(c) \) is normal and independent of \( W(u - c) \), a new Brownian motion. It may, however, be easiest simply to simulate the \( D \) distribution, for the \([a,b]\) interval of interest, which then yields the band \( \hat{A}(t) \pm w_{\alpha}/\sqrt{n} \). A fruitful alternative to this fixed-width band is to construct one having widths proportional to the pointwise one, as
Confidence distributions in higher dimensions

independently proposed and worked with in Nair (1984) and Hjort (1985). The starting point there is that

\[ M_n = \max_{a \leq t \leq b} \sqrt{n} |\hat{A}(t) - A(t)|/\hat{\sigma}(t) \]

tends in distribution to

\[ M = \max_{a \leq u \leq d} |W(\sigma(u)^2)|/\sigma(u) = \max_{c \leq u \leq d} |W(u)|/\sqrt{u}, \]

with \([a, b]\) mapped to \([c, d] = [\sigma(a)^2, \sigma(b)^2]\). This distribution may again be simulated, for the \([a, b]\) in question, yielding a simultaneous confidence band for \(A\) on that interval. For further details and some exploration of modifications for smaller sample sizes, see Bie et al. (1987).

Example 9.7 Surviving in Roman era Egypt

Consider again the lifelengths in Roman era Egypt, where we now focus on the age interval 15 to 40 years. We give the estimated survival distribution \(\exp\{−\hat{A}(t)\}\), along with bands of coverage degree 0.50 and 0.90, produced via critical points found by simulating the limiting distribution of

\[ \max_{15 \leq t \leq 40} |\sqrt{n}(\hat{A}(t) - A(t))|/\hat{\sigma}(t) \]

as per the recipe above. See Figure 9.7. We also include the estimated survival curves using respectively the Weibull and the Gamma distributions. For details here, see Exercise 9.6.

\[ \begin{array}{c}
\text{age} \\
\text{survival probability}
\end{array} \]

Figure 9.7 Survival curve estimates for lives in Roman era ancient Egypt, for age interval 15 to 40; nonparametric (step function, solid line), based on the Gamma distribution (continuous curve, dashed line), and based on the Weibull distribution (continuous curve, dotted line, lying slightly above the Gamma one). The inner and outer step functions give simultaneous nonparametric confidence bands of degree 0.50 and 0.95.
9.7 Dependencies between confidence curves

The adjustment functions for turning roots of individual confidence curves into product confidence curves are compared to the Bonferroni adjustment functions in Example 9.3. They could also be compared to the adjustment function under independence between the root members as was done in Figure 9.2. This latter curve is $K_{ind}(p) = p^T$ when there are $T$ members in the root.

Adjustment curves turn often out to follow a power law, $K(p) \approx p^{\hat{T}}$. Figure 9.8 shows the adjustment function for the root of size $T = 29$ for a quantile regression of capital income on other income on the log-log scale. The linearity is striking. The slope is $\hat{T} = 7.05$.

For adjustment functions of approximate power law form, the ratio $\hat{T}/T$ reflects how far from independence the confidence curves in the root are. Under complete dependence, $\hat{T} = 1$. The number of stochastically independent factors behind the root is estimated by $\hat{T}$.

A measure of dependencies between confidence curves in the root, that ranges from zero (complete independence) to one (complete dependence), is given by

$$r = \frac{T - \hat{T}}{T - 1}$$

(9.8)

For the quantile regression curve for income in Norway the dependencies in confidence curves in the root is estimated to be $r = 0.78$. Some of this dependency results from the smoothing applied to the quantile regression curve. The bulk of the dependency must, however, be ascribed to the stability in the distribution of capital income on other income in Norway.

![Figure 9.8](image-url) Simultaneous level versus nominal level on log-log scale: 95% quantile regression of capital income on other income, Norway 2002, at $T = 29$ levels of other income.
Confidence distributions in higher dimensions

9.8 Notes on the literature

Fiducial probability is not ordinary probability, as Fisher claimed; see Chapter 6. Joint confidence distributions such as the normal distribution for $\mu$ based on a $p$-dimensional pivot yields confidence to rectangular regions by their distribution functions. They do possess joint confidence densities, but these cannot be integrated to provide confidence to arbitrary Borel measurable sets. For linear parameters integration works, however. Singh et al. (2007) discuss what they call multivariate confidence distributions of a linear sense (termed $\ell$-CD), where integration of the confidence density over linearly restricted sets yields valid confidence of the sets.

Elliptical confidence curves are available for the mean parameter $\mu$ of a normal distribution, and also more generally for an elliptical distribution (Härdle and Simar, 2012).

Survival analysis is used primarily in medical research (Andersen et al., 1993, Collett, 2003) and engineering (Rausand and Høyland, 2004). Andersen et al. (1993), Aalen et al. (2008) and Fleming and Harrington (1991) present theory of counting processes and survival analysis, along with detailed discussion of applications. There is a long tradition in statistics for constructing simultaneous confidence bands around functions (e.g., cumulative distribution or hazard functions), typically associated with limits of empirical processes; see, for example, Billingsley (1968) and Andersen et al. (1993) for appropriate theory. Bands associated with certain nonparametric likelihood tests are developed in Owen (1995).

Product confidence curves were introduced by Schweder (2007). They extend the theory of product confidence sets of Beran (1988a), and simplified his theory somewhat. Schweder illustrated the method by simultaneous confidence bands for the growth in personal income in Norway by percentile. A new version of this study is presented in Chapter 14.

There is a huge literature on simultaneous testing; see, for example, Romano and Wolf (2007). Concepts and methods in simultaneous testing are useful for confidence regions, and consequently for confidence curves in higher dimensions. In this chapter we have discussed various confidence curves with level sets covering the full parameter vector with probability equal to the level. In hypothesis speak, this corresponds to tests with controlled family-wise error rate (FWER). Romano and Wolf (2007) discuss tests with a controlled probability of $k$ or more of the individual hypotheses being falsely rejected. Corresponding $k$-FWER confidence curves are discussed in Exercise 9.7.

Constructing confidence regions by contouring the deviance function by the chi-square distribution has been done since Wilks (1938). To go the other way, from confidence distributions to likelihood, was suggested by Efron (1993) and by Schweder and Hjort (1996, 2002); see Chapter 10.

Exercises

9.1 Multidimensional fiducial integration: Consider a sample of size $n$ from the $p$-dimensional normal distribution $N_p(\mu, \Sigma)$, and let $\bar{Y}$ be the sample mean vector. When the covariance matrix $\Sigma$ is known, $N_p(\bar{Y}, \Sigma/n)$ is a $p$-variate confidence distribution for the mean vector $\mu$. Find a confidence distribution for $\psi = A\mu$ where $A$ is a $q \times p$ matrix. Show that this confidence distribution is also found by integration in the full confidence distribution. Does integration also work for linear parameters when $\Sigma$ is unknown and must be estimated?
9.2 The value of knowing $\Sigma$: Consider the ellipsoidal confidence curve for the $p$-dimensional mean vector $\mu$ from a normal sample of size $n$ from the normal distribution, both when the covariance matrix is known and when it must be estimated. Find the adjustment function that turns the first confidence curve into the second when the estimated covariance matrix equals the true. Plot the adjustment function and suggest an overall measure of information loss of having to estimate the covariance matrix. Calculate the information loss for a few values of $p$ and $n$.

9.3 Glivenko-Cantelli confidence curves: The confidence band for a distribution function based on the Glivenko–Cantelli lemma might be seen as a product confidence curve. What is its root? Does its adjustment function follow a power law, say for $p = 5, 10, 50, 500$?

9.4 Carlsen versus Anand: Assume $(X, Y, Z)$ is trinomial with $n$ trials and probability parameters $(p, q, r)$. Set up a scheme for computing and displaying a confidence region for $(p, q)$, of coverage degree $\alpha$, based on the two versions of the Pearson statistic (9.7).

(a) Do this, for $n = 39$ and observed counts $(6, 6, 27)$, hence duplicating a version of Figure 9.6. Find the corresponding product confidence curve and compare.

(b) Update this construction, and the relevant figures, now taking into account also the matches throughout 2014, that is, the Zürich draw in February and the eleven world champion games of Sochi (three wins for Carlsen, one win for Anand, seven draws). The score as the end of 2014 is hence 9 and 7 wins for Carlsen and Anand and 35 draws.

9.5 Big bands: Suppose one wishes to construct a simultaneous confidence band for an ensemble of parameters $\theta_1, \ldots, \theta_m$, and that estimators $\hat{\theta}_j$ are available for which $\hat{\theta}_j \sim N(\theta_j, \sigma_j^2)$, with known standard deviation parameters. At the outset we do not impose any further conditions here regarding independence or dependence. Consider the simultaneous band

$$R(z) = \prod_{j=1}^m |\hat{\theta}_j - z\sigma_j| = \{\theta : |\hat{\theta}_j - \theta_j|/\sigma_j \leq z \text{ for all } j \leq m\}.$$  

(a) For any events $A_1, \ldots, A_m$, show that $P(\bigcup_{j=1}^m A_j) \leq \sum_{j=1}^m P(A_j)$. Deduce the complementary version, that for any events $B_1, \ldots, B_m$,

$$P(\bigcap_{j=1}^m B_j) \geq 1 - \sum_{j=1}^m P(B_j).$$

In particular, if each $P(B_j) \geq 1 - \varepsilon/m$, then $P(\bigcap_{j=1}^m B_j) \geq 1 - \varepsilon$. This leads to the Bonferroni method solution, mentioned in Example 9.3. With $z_1$ chosen such that $P(|N| \geq z_1) = \varepsilon/m$, or $z_1 = \Phi^{-1}(1 - \frac{1}{2}\varepsilon/m)$, then the band $R(z_1)$ has coverage level at least $1 - \varepsilon$. Note that this method works whether the individual $\hat{\theta}_j$ are independent or not.

(b) Assume for this point that the estimators are statistically independent. Show that the confidence level of $R(z)$ is $P(|N| = z) = \Gamma_1(z^2)^m$, with $N \sim N(0, 1)$ and $\Gamma_1(\cdot)$ the distribution function for the $\chi^2_1$. Hence show that to achieve coverage $1 - \varepsilon$, say 0.95, one needs $\Gamma(z^2) = (1 - \varepsilon)^{1/m}$, and that this leads to selecting $z_2 = \Phi^{-1}(1 - \frac{1}{2}d/m)$, where $d = -\log(1 - \varepsilon)$.

(c) Show that the bands produced via method (b) are slimmer, but only very slightly so, than the ostensibly cruder ones produced via method (a). For more on situations where Bonferroni intervals become slimmer than those produced by more laborious methods, see Hjort (1988b).

(d) Investigate briefly how broad bands of this type become for increasing ensemble size $m$. For what $m$ are the individual intervals as broad as $\hat{\theta}_j \pm 4\sigma_j$, and for how big $m$ do they have the size $\hat{\theta}_j \pm 5\sigma_j$?
Confidence distributions in higher dimensions

9.6 Confidence bands for survival in ancient Egypt: For the survival data of Pearson (1902), available at the CLP book website www.mn.uio.no/math/english/services/knowledge/CLP/, compute the Nelson–Aalen estimator and its survival function cousin \( \exp(-\hat{A}) \). Simulate the distribution of \( \max_{(c,d)} |W(u)|/\sqrt{n} \) to construct the simultaneous confidence band at levels 0.50, 0.75, 0.90, 0.95, say over the interval that corresponds to age window \([a,b] = [10,50]\); see Example 9.7. Display such estimates and bands, for the men and the women.

9.7 \( k \)-FWER confidence curves: The level curve at \( \alpha \) of a confidence curve for a vector parameter \( \theta \) covers its true value with probability \( \alpha \). In the context of hypotheses testing Romano and Wolf (2007) relax this criterion. In our context their criterion is that with confidence \( \alpha \) at most \( k \) of the components of \( \theta \) are outside of what is called \( k \)-FWER confidence curve, for family-wise error rate. The product curve of Section 9.5 might easily be modified to be of \( k \)-FWER type. The trick is simply to replace the adjustment function with the distribution of the order statistic \( c_{(\tau-k)}(\theta, X) \). Let \( \theta \) be the vector of off-diagonal elements in a correlation matrix of dimension \( p \). Simulate a normal sample of size and parameters of your choice and calculate a \( k \)-FWER confidence curve for \( \theta \). Does its adjustment function follow a power law?
10

Likelihoods and confidence likelihoods

The likelihood function is the bridge between the data and the statistical inference when the model is chosen. The dimension of the likelihood might be reduced by profiling, conditioning or otherwise to allow a confidence distribution for a parameter of interest to emerge. We shall now turn this around and develop methods for converting confidence distributions for focus parameters into likelihoods, which we call confidence likelihoods. This is particularly useful for summarising the most crucial part of the inference in a manner convenient for further use, for example, in meta-analyses, as discussed further in Chapter 13. We shall in particular discuss normal conversion methods along with acceleration and bias corrected modifications, and illustrate their uses. Connections to bootstrapping of likelihoods are also examined.

10.1 Introduction

The likelihood function is the link between data as probabilistically modelled and statistical inference in terms of confidence distributions or otherwise. This chapter prepares the ground for Chapter 13 on meta-analysis. In a meta-analysis the results of several studies that address a set of related research hypotheses are combined. In its simplest form, the problem is to combine estimates of a common parameter from independent studies, which is traditionally solved by a weighted average with weights obtained from the respective standard errors. This would be optimal if the estimators of the individual studies were normally distributed.

From the confidence distributions or confidence intervals of the studies, it might be seen that the normal assumption is far-fetched. Rather than just averaging, it would be better if a pseudo-likelihood could be constructed from the information provided in each of the reports, and combine these by multiplication to a collected pseudo-likelihood. We will call a pseudo-likelihood obtained from a confidence distribution or a confidence interval a confidence likelihood (c-likelihood), and methods for such constructions are developed in the text that follows.

Fisher (1922) introduced the likelihood function as the primary tool for statistical analysis of data. This and the wealth of new concepts, new points of views and new theory constitute the third revolution in parametric statistical inference (Hald, 1998, p. 1), after the first when Laplace (1774) introduced the method of inverse probability and that of Gauss and Laplace, who introduced the least squares method and linear models for normally distributed data in 1809–1812. The likelihood function is a minimal but sufficient reduction of the data. The locus of its maximum, the maximum likelihood estimator, is in many cases the optimal point estimator, and the log-likelihood ratio provides confidence inference. Furthermore,
Likelihoods and confidence likelihoods

Independent data are combined in an optimal way by multiplying their likelihood functions together. This last property is of special interest in the current chapter. In the context of simple meta-analysis the independent pieces of data are the individual estimates, say in the form of confidence distributions. Assuming that these studies are representative of the phenomenon of interest, i.e. that the parameter of interest, that is, is understood in the same way across studies, and that there is no file-drawer problem so that studies are published independent of the results regarding the focus parameter, the Fisherian solution would be to construct an integrated likelihood function by multiplication, and to proceed with this summary to obtain a collective estimate, or preferably a confidence distribution.

There are two snags to this. First, published studies seldom provide the full likelihood of the data considered, but rather gross summaries of the data in terms of p-values and confidence intervals or confidence distributions, and there is a need to convert this information to a pseudo-likelihood. The other snag is that much of the work behind the individual studies consists of developing an appropriate model, checking its fit to the data, and also of squeezing out information about the main parameters accounting for the presence of nuisance parameters and other peculiarities. In the integrated analysis one would not want to redo this research work, partly because it should be based on a thorough understanding of the methods of measurements and other contributing factors behind the data, an understanding that perhaps seldom will be in place outside the original research team. The second best option would thus be to establish a pseudo-likelihood function from the published results from each study. This likelihood should be as close as possible to the profile likelihoods (or sometimes the conditional likelihood) of the individual studies. We shall assume that each publication provides a confidence distribution for the parameter $\psi$ in focus, or sufficient information for a confidence distribution to be established.

Confidence curves for scalar parameters were in Chapter 1 introduced as probability transforms of deviance functions, or profile deviance functions in the presence of nuisance parameters. For interest parameters of dimension $p \geq 2$ the (profile) deviance function gives rise to a nested set of confidence regions by probability transformation. These confidence regions are understood as contours of a confidence curve obtained from a joint confidence distribution, but such confidence distributions are not measures over the ordinary Borel sigma-algebra but rather measures over the much smaller family of nested confidence regions. The link from likelihood to confidence is thus clear when the model dictates how to probability transform.

If this link could be reversed, we would obtain a likelihood function from a confidence distribution. The link is the reverse of the probability transform, from probability, or rather confidence, to quantile. Consider a scalar parameter $\theta$ and a continuous probability distribution $P_\theta$ for the data $Y$. In the absence of nuisance parameters the deviance function is $D(\theta, Y)$ and the directed likelihood root is

$$r(\theta, Y) = \text{sgn}(\theta - \hat{\theta})D(\theta, Y)^{1/2},$$

where $\hat{\theta}$ is the maximum likelihood estimator.

The cumulative confidence distribution for observed data $y$ is then

$$C(\theta) = P_\theta \{ r(\theta, Y) \leq r(\theta, y_{\text{obs}}) \},$$
10.1 Introduction

when \( r \) is monotone in \( \theta \). With \( H_\theta(r) = P_\theta\{r(\theta, Y) \leq r\} \) being invertible with quantiles \( H_{\theta}^{-1} \), the directed likelihood root is recovered from the confidence distribution as \( r(\theta, y) = H_{\theta}^{-1}(C(\theta)) \). The bridge between the likelihood and the confidence distribution is clearly founded on the sampling distribution of the data. It is thus not sufficient to know the confidence distribution to obtain the likelihood function, just as the confidence distribution cannot be obtained from the likelihood function without further information.

Profile deviances are converted to confidence curves by probability transformations, and confidence curves are converted into confidence likelihoods by the same probability transformations, but in inverse form. If \( F_\psi \) is the cumulative distribution of the profile deviance \( D(\psi) \) evaluated at the true parameter value, then \( cc(\psi) = F_\psi(D(\psi)) \) is the confidence curve. The assumption here is that the profile deviance is distributed independently of the nuisance parameter \( \chi \). The deviance is thus \( D_c(\psi) = F_{\psi}^{-1}(cc(\psi)) \) and the confidence log-likelihood obtained from a given confidence distribution represented by its confidence curve is

\[
\ell_c(\psi) = -\frac{1}{2}F_{\psi}^{-1}(cc(\psi)).
\]

Under the assumption of the probability transform being independent of the nuisance parameter, the confidence likelihood recovers the original profile likelihood on which the confidence curve is based. When this assumption does not hold the chosen probability transform might be the one determined by a particular value of the nuisance parameter, say the maximum likelihood estimate, and the confidence distribution obtained from the profile log-likelihood will typically only be approximate for other values of the nuisance parameter. In this case the c-likelihood will only approximate the profile likelihood, and should be regarded as a pseudo-likelihood.

In Examples 1.2 and 2.2 we saw there is no one-to-one correspondence between likelihoods and confidence distributions. The relation depends on the protocol for the observations. In the first example one protocol is to observe the number of points in a Poisson process over \( t \) units of time. Another one is to observe the waiting time until \( x \) points occur in the process. The confidence distributions are not equal in the two cases. For the same reasons, the confidence likelihood based on a confidence distribution depends on the observational protocol. See Example 10.6 below.

In Bayesian analysis, prior information is represented as prior distributions, and amalgamated with the likelihood function of the new data to posterior distributions. If prior information is in the form of a distribution, the frequentist has been at loss. Traditionally the only way the non-Bayesian statistician has been able to incorporate prior information into the analysis is through the choice of a statistical model for the new data, and by imposing functional restrictions on the parameter. But if the available prior distribution could be turned into a likelihood function an additional tool is available. Then a collected likelihood is obtained by combining this likelihood with the other likelihood components, and inference should be based on this combined likelihood. Often such prior distributions are based on data, sometimes of the soft kind. From Fisher we have learned that inference should be based on the available data through the likelihood function. It makes for good order in the inferential process to represent each independent piece of data with its likelihood function, and to combine data by multiplying together their component likelihoods. This goes also for data behind prior distributions.


Likelihoods and confidence likelihoods

10.2 The normal conversion

By the Wilks Theorem 2.4 the profile deviance for a scalar focus parameter $\psi$ evaluated at the true value $\psi_0$ has asymptotically a chi-squared distribution with df = 1, as the data grow, provided basically that the nuisance parameters stay fixed. The contours of the profile deviance function $D(\psi)$ are thus probability transformed to a confidence curve by the cumulative chi-squared distribution. An approximate confidence log-likelihood is obtained from a given confidence curve $cc(\psi)$ based on large samples through the chi-squared quantile function $/Gamma_1^{-1}$, that is, $D(\psi) = /Gamma_1^{-1}(cc(\psi))$. The approximated profile likelihood is thus

$$L_c(\psi) = \exp\{-\frac{1}{2} /Gamma_1^{-1}(cc(\psi))\} = \exp\{-\frac{1}{2}\{\Phi^{-1}(C(\psi))\}^2\};$$

see Exercise 10.7.

Efron (1993) found this profile likelihood by considering a scalar parameter $\psi$ in an exponential family model. From an exact confidence distribution $C_n$ based on a sample of size $n$ with confidence density $c_n$ and a corresponding confidence density $c_{2n}$ based on the doubled data, Efron argues that $L_c(\psi) = c_{2n}(\psi)/c_n(\psi)$, which would be the case if the confidence densities were Bayesian posterior densities, based on a prior $\pi(\psi)$. In that case $c_n(\psi) \propto \pi(\psi)L_c(\psi)$ and $c_{2n}(\psi) \propto \pi(\psi)L_c(\psi)^2$. In exponential families he then finds that confidence intervals obtained by his abc method (cf. Section 7.6), yield confidence densities that lead to $L_c$ in exponential families. Efron’s argument is quite different from our simple argument above based on the asymptotic chi-squared distribution of the deviance.

Normal conversion works just as well for the interest parameter $\psi$ having dimension $p \geq 2$. Then the Wilks theorem is that the profile deviance evaluated at the true value of $\psi$ is asymptotically distributed as a $\chi^2_p$ with cumulative distribution function $/Gamma_p$. A confidence curve $cc(\psi)$ is then converted to a likelihood via

$$D_c(\psi) = /Gamma_p^{-1}(cc(\psi))$$

and

$$L_c(\psi) = \exp\{-\frac{1}{2} /Gamma_p^{-1}(cc(\psi))\}$$

is obtained.

Example 10.1 Male lifelength in ancient Egypt (Example 1.4 continued)

Continuing Example 1.4, Figure 10.1 displays the confidence curve and log-likelihood for $\lambda = ET$ where $T$ is the stochastic lifelength of an ancient Egyptian male. The solid lines represent exact curves while the dashed curves are approximate. The approximate confidence curve is the chi-squared probability transform of the exact deviance function, while the dashed log-likelihood curve is the chi-squared quantile transform of the exact confidence curve. In both cases, df = 1. The approximations are good because the sample size is as large as $n = 82$, and despite the skewness and long tail in the assumed exponential distribution.

For mean parameters in linear normal models such as regression models, exact confidence curves are available, and the c-log-likelihood obtained by converting an exact confidence curve by the chi-squared distribution is almost the modified profile likelihood $L_{mprof}$ for the parameter (Barndorff-Nielsen and Cox, 1994, p. 265); see the discussion of Section 7.7. The approximation is very good. If say $Y_1, \ldots, Y_n$ are independent and $Y_j \sim N(\mu + z_j/\beta, \sigma^2)$ with $\mu$ the parameter of interest, then the exact confidence distribution is
10.2 The normal conversion

Figure 10.1 Confidence curve and confidence log-likelihood function for $\lambda$, the male life expectancy in ancient Egypt, cf. Example 10.1. Solid lines are exact under the exponential model, while dashed curves are approximations.

obtained from the usual Student pivot. With $q$ regression parameters, and the design matrix with rows $(1, z^\top_j)$ having rank $q + 1$, the confidence curve is

$$cc(\mu) = G_v(t(\mu)),$$

with $t(\mu) = (\mu - \hat{\mu})/se(\hat{\mu})$,

where $G_v$ is the distribution function of the t distribution with $v = n - q - 1$ degrees of freedom. The profile likelihood is

$$L_{\text{prof}}(\mu) = \left\{1 + \frac{t(\mu)^2}{n - 1 - q}\right\}^{-n/2}$$

and its modification, via methods of Section 7.7, is

$$L_{\text{mprof}}(\mu) = \left\{1 + \frac{t(\mu)^2}{n - 1 - q}\right\}^{-(n-q)/2+1}.$$

The expressions for a regression parameter, or a linear contrast, are identical. To convert the confidence curve for a single regression parameter $\beta$ based on a t-statistic to a confidence likelihood, first convert the confidence curve to a confidence deviance,

$$\hat{D}_{\text{mprof}}(\beta) = \Gamma_1^{-1}(cc(\beta)).$$

Then convert further to an improved confidence likelihood,

$$\hat{L}_{\text{prof}}(\beta) = \exp\left\{-\frac{1}{2} \frac{n}{n - q - 2} \hat{D}_{\text{mprof}}(\beta)\right\}.$$

This should be an excellent approximation.
Likelihoods and confidence likelihoods

Example 10.2 Broken limits to life expectancy

Oeppen and Vaupel (2002, p. 1029) point to an astonishing fact: “Female life expectancy in the record-holding country has risen for 160 years at a steady pace of almost three months per year.” Their ‘broken limits’ refer to the fact that prediction intervals for future years nearly always were broken by the realised life expectancy, which turn out higher than predicted. A record-holding country in a year is the country that has the highest life expectancy that year among countries with acceptable mortality data. In 1840 the record was held by Swedish women, who lived on average a little more than 45 years. In the year 2000 Japanese woman held the record with a life expectancy of almost 85 years. As did Oeppen and Vaupel, we will use simple regression analysis to measure the rate of rise in female life expectancy. Let \( Y \) be record female life expectancy with \( x \) the calendar year. The regression model is simply \( Y = a + bx + \epsilon \), with residuals \( \epsilon \) assumed i.i.d. \( N(0, \sigma^2) \). The residual variance was larger in the early part of the series because fewer countries kept acceptable demographic statistics; cf. the left panel of Figure 10.2, but we disregard this complication. The regression result is \( \hat{b} = 0.2430 \) at a standard error of \( se = 0.0026 \). The right panel of Figure 10.2 shows the exact confidence curve for the slope \( b \) (solid line) along with its approximation based on normal conversion (dashed line). These are virtually indistinguishable, because a \( t_{n-2} \) is so close to a standard normal. We return to these data in Example 12.8 to check what they may say about our future.

![Figure 10.2](image-url) The ‘broken records’ data (left), displaying the expected lifelength for women of the record-holding country, along with two (nearly indistinguishable) confidence curves for the slope parameter \( b \) (right), one based on the exact distribution for the pivot (solid line) and the other based on normal conversion (dashed line).
Example 10.3 Bowhead whales off Alaska (Example 8.3 continued)

In Section 8.2 we found the conditional likelihood for total abundance given the ancillary statistics (the annual numbers of near captures) from a capture–recapture experiment (assuming a closed population of bowhead whales). We also found the confidence distribution. Numerical work, not displayed here, shows that the conditional deviance function and the confidence deviance obtained by normal conversion are very similar.

10.3 Exact conversion

Exact confidence distributions are found from pivots. Without making much out of it, Fisher (1930) also gave the formula for the exact likelihood function based on a confidence distribution. Let $C(\psi, T)$ be the cumulative confidence distribution based on a one-dimensional statistic $T$ with a continuous distribution. Being a function of the statistic $T$, the probability density of $T$ and thus the likelihood follows by the chain rule:

$$f(T, \psi) = g(C(\psi, T)) \left| \frac{\partial C(\psi, T)}{\partial t} \right|.$$ 

Since $C$ is a confidence distribution it has a uniform distribution with probability density $g(C(\psi, T)) = 1$. The confidence likelihood is thus

$$L_c(\psi, t) = \left| \frac{\partial C(\psi, t)}{\partial t} \right|.$$ 

This construction holds for any pivot $\text{piv} = \text{piv}(T, \psi) = F^{-1}(C(\psi, T))$ with pivotal cumulative distribution $F$ and density $f$. Thus

$$L_c(\psi, t) = f(\text{piv}(t, \psi)) \left| \frac{\partial \text{piv}(t, \psi)}{\partial t} \right|.$$ 

In the case of discrete distribution on the integers, such as the Poisson distribution, the difference in the direction of $t$ yields with half-correction $C(\psi, t+1) - C(\psi, t) = \frac{1}{2} [L(\psi, t) + L(\psi, t+1)]$. Without half-correction $C(\psi, t) = P_\psi(T \geq t) or C(\psi, t) = P_\psi(T \leq t)$ and the confidence difference is precisely the likelihood based on $T$, $|C(\psi, t+1) - C(\psi, t)| = L(\psi, t)$.

Example 10.4 Normal confidence likelihood

When $C(\psi, T) = \Phi\left(\frac{\mu(\psi) - T}{\sigma(\psi)}\right)$ for some location and scale functions $\mu$ and $\sigma$ the confidence log-likelihood is

$$\ell(\psi) = -\frac{1}{2} (\Phi^{-1}(C(\psi)))^2 - \log \sigma(\psi).$$

This is correct since $T \sim N(\mu(\psi), \sigma(\psi)^2)$.

The confidence likelihood is the probability density of the statistic. The confidence density is also a density, but it is not generally a likelihood. It is actually $dC(\psi, T)/d\psi$. It is important to keep the confidence likelihood apart from the confidence density.

Example 10.5 Variance estimation, normal data

The likelihood for $\sigma$ is based on the pivot $SSR/\sigma^2$ or directly on the sum of squares $SSR = \nu S^2$ part of the data. It is the density of $SSR = \sigma^2 \chi^2_{\nu}$, which is proportional to

$$L(\sigma) = \sigma^{-\nu} \exp\left(-\frac{1}{2} \nu S^2 / \sigma^2\right).$$
Likelihoods and confidence likelihoods

This is the marginal likelihood based on the sufficient statistic $SSR$. The cumulative confidence for $\sigma$ is $C(\sigma) = \Gamma(\nu, \nu S^2/\sigma^2)$, and the likelihood based on $C$ comes out the same. The confidence density is, on the other hand,

$$c(\sigma) = \frac{dC(\sigma)}{d\sigma} \propto \sigma^{-\nu-1} \exp(-\frac{1}{2} \nu S^2/\sigma^2) = \frac{L(\sigma)}{\sigma}.$$

Likelihoods are invariant to parameter transformations, while densities must be adjusted by the Jacobian. In the parameter $\tau = \log \sigma$ the confidence density is proportional to the likelihood. The log-likelihood has a nicer shape in $\tau$ than in $\sigma$, where it is less neatly peaked. It is of interest to note that the improper prior $\pi(\sigma) = \sigma^{-1}$, regarded as the canonical 'noninformative' prior for scale parameters such as the present $\sigma$, yields when combined with the likelihood $L$ the confidence distribution as the Bayes posterior distribution. See also the more general comment in Section 10.7.

An important fact is that additional information is needed to convert a confidence distribution to a confidence likelihood and vice versa. In addition to the confidence distribution $C(\psi)$ as a function of $\psi$ we must also know how this confidence distribution has been constructed from the data, $C(\psi, T)$, where $T$ is the statistic on which the confidence distribution is based. Exact confidence likelihoods might also be obtained from a confidence curve, and an exact confidence curve is obtained from a deviance function if it is known what the sampling variation is. If the deviance function $D(\psi, Y)$ has cumulative distribution $F(d, \psi)$ then $cc(\psi) = F(D(\psi, Y), \psi)$, and the likelihood follows. The inverse exact conversion is of course $D(\psi) = F^{-1}(cc(\psi), \psi)$.

10.4 Likelihoods from prior distributions

In Bayesian analysis prior distributions are included in the analysis through Bayes’ formula. If the prior distribution is based on data, however soft, these 'data' should in a Fisherian analysis be included by way of a likelihood function. One way of converting a distribution to a likelihood function is by normal conversion, as discussed previously. This is not the only way, and the following example illustrates that a distribution can lead to different likelihoods.

A problem that might occur in a Bayesian analysis is that there sometimes are several independent sources of information on a given parameter. There might thus be several independent prior distributions for the parameter. The Bayesian method is based on one and only one joint prior distribution for the vector of parameters, and the set of independent priors would thus need to be amalgamated into one distribution before Bayes’ lemma can be applied. With two prior densities $p_1(\theta)$ and $p_2(\theta)$ Poole and Raftery (2000) put them together by $p(\theta) = k(\alpha)p_1(\theta)^{\alpha}p_2(\theta)^{1-\alpha}$ where $k(\alpha)$ is a normalising constant and $0 < \alpha < 1$ gives weight to the priors. They call this Bayesian melding. If the two priors are identical, the melded version is equal to both. In Bayesian melding there is thus no information gain from having two independent prior distributions for a parameter. If, however, the independent prior distributions were represented by likelihoods $L_1$ and $L_2$ the pooled information is represented by $L_1(\theta)L_2(\theta)$. Then the pooled likelihood is more informative than each of its parts, not the least when they are identical.
10.4 Likelihoods from prior distributions

Example 10.6 Different likelihoods with the same confidence distribution

In a Bayesian analysis of assessing the status of bowhead whales off of Alaska, Brandon and Wade (2006) assume a uniform prior distribution over $[0.4, 0.8]$ for one of the parameters. This prior distribution is based on data from other fin whale species and other populations, and should be regarded as data, although soft. We shall take a look at three different generating mechanisms that yield quite different likelihoods. The cumulative confidence distribution function is

$$C(\psi) = (\psi - 0.4)/0.4 \quad \text{for } 0.4 \leq \psi \leq 0.8. \quad (10.1)$$

**Shift model:** In this model, the confidence distribution is based on the uniform pivot

$$U = (\psi - T - 0.2)/0.5 \quad \text{for } 0.4 < \psi < 0.8,$$

where $U$ is uniform $(0, 1)$. The observed value is $T_{\text{obs}} = 0.6$ does indeed result in the confidence distribution (10.1). The resulting likelihood is flat,

$$L_1(\psi, T) = \partial C/\partial T = \partial U/\partial T = 1/0.4 \text{ over the interval } (0.4, 0.8) \text{ and is zero outside.}$$

**Scale model:** The confidence distribution is now based on

$$U = (\psi - 0.4)/(2T - 0.8).$$

The observed value $T_{\text{obs}} = 0.6$ leads to the uniform confidence distribution over $[0.4, 0.8]$. In this scale model the likelihood is

$$L_2(\psi, T) = \partial U/\partial T \propto \psi - 0.4 \text{ over the permissible interval.}$$

**Transformed normal model:** The normally converted likelihood

$$L_3(\psi) \propto \exp \left\{ -\frac{1}{2} \Phi^{-1}\left( \frac{\psi - 0.4}{0.4} \right)^2 \right\}$$

is obtained from the confidence distribution

$$C(\psi, T) = \Phi\left( \Phi^{-1}\left( \frac{T - 0.4}{0.4} \right) - \Phi^{-1}\left( \frac{\psi - 0.4}{0.4} \right) \right).$$

Again, with $T_{\text{obs}} = 0.6$, the confidence distribution is the uniform over the interval $[0.4, 0.8]$. The likelihood is once more obtained as $L_3(\psi, T) = \partial C/\partial T$. Normal conversion is thus seen as generated from the normal score of the confidence distribution. Figure 10.3 displays the three likelihood functions based on the same uniform confidence distribution. They are indeed different. \[\Box\]

Example 10.7 Normal data and a prior

Let $\tilde{\psi} \sim N(\psi, \sigma^2)$ be sufficient with $\sigma$ known. The prior distribution is $\psi \sim N(\psi_p, \tau^2)$. The Bayesian posterior is then $N(\tilde{\psi}/\tau^2)$ where

$$\tilde{\psi} = \frac{\psi/\sigma^2 + \psi_p/\tau^2}{1/\sigma^2 + 1/\tau^2} \quad \text{and} \quad \frac{1}{\tau^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}.$$
Figure 10.3 Three likelihoods consistent with a uniform confidence distribution over $[0.4,0.8]$; cf. Example 10.6. “Many likelihoods informed me of this before, which hung so tottering in the balance that I could neither believe nor misdoubt.” – SHAKESPEARE.

The confidence curve of the prior is simply $cc(\psi) = |2\Phi((\psi - \psi_p)/\tau) - 1|$ and the confidence likelihood is $L_c(\psi) = \exp\{-\frac{1}{2}\Gamma^{-1}_1(cc(\psi))\}$. Since $\Gamma_1(x) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x})$ the combined log-likelihood comes out as

$$-\frac{1}{2}\left\{\left(\frac{\psi - \tilde{\psi}}{\sigma}\right)^2 + \left(\frac{\psi - \psi_p}{\tau}\right)^2\right\} = -\frac{1}{2}\left(\frac{\psi - \tilde{\psi}}{\tilde{\tau}}\right)^2 + \text{const.},$$

which represents the same normal distribution as found as the Bayesian posterior. A more direct argument is to note that the confidence likelihood of $N(\psi_p, \tau^2)$ is simply proportional to the normal density $L_c(\psi) = \phi((\psi - \psi_p)/\tau)$. The combined likelihood is then for the Fisherian proportional to $\phi((\psi - \psi_p)/\tau)\phi((\psi - \tilde{\psi})/\sigma)$, while the Bayesian posterior density is proportional to that likelihood.

Noninformativity is a difficult notion in the Bayesian paradigm. The method of inverse probability was the Bayesian method with flat priors to reflect absence of prior information. Flatness of a probability function is, however, not an invariant property. If you know nothing of $\psi$ you know nothing of $\theta = t(\psi)$ when $t$ is a monotone transformation, and the prior density for $\theta$ will not be flat if the transformation is nonlinear and the prior for $\psi$ is flat.

This problem of noninformativity was Fisher’s basic motivation for inventing the fiducial distribution in 1930. In a Fisherian analysis, noninformativity is represented by a flat prior.
likelihood, or equivalently, with no prior likelihood. Because likelihoods are invariant to parameter transformations, noninformativity is an invariant property of prior likelihoods.

In the case of prior information this might apply only to one component of the parameter vector. If, say, \( \theta = (\psi, \chi) \) and there is a prior distribution for \( \psi \), but no prior information on \( \chi \), the Bayesian is in trouble, while the Fisherian would work out a confidence curve for \( \psi \) from its prior distribution, and from this a prior likelihood \( L_c(\psi) \) for this part of the parameter vector. The prior likelihood for \( \theta \) is then \( L_c(\theta) = L_c(\psi) \). If, however, there is independent prior information on \( \chi \) represented by \( L_c(\chi) \), then \( L_c(\theta) = L_c(\psi) L_c(\chi) \).

Bootstrapping and simulation have emerged as indispensable tools in statistical inference. Bootstrapping might be used to establish a confidence likelihood, and it might further be used to draw confidence inference from confidence likelihoods, perhaps in a meta-analysis context. When dealing with original data confidence distributions are often obtained by bootstrapping, say by the t-bootstrap of Section 4.5. When reporting, the numerical confidence curve might then be plotted, and perhaps the 95% confidence interval and the point estimate are numerically given. To allow the results to be used in a meta-analysis later on the corresponding likelihood function should also be given. A parametric approximation to the likelihood based on the bootstrap distribution would be helpful in a later meta-analysis. In case only a confidence interval is available, an approximate likelihood might be constructed; see the following section.

A confidence likelihood to be bootstrapped is usually obtained from a pivot based on an estimator. One typically observes aspects of the bootstrap distribution of the estimator, and uses these to reach accurate inference. In our context it is also useful to focus on bootstrap distributions of the full log-likelihood curves or surfaces. Thus in Example 7.5, the bootstrapped log-likelihood curve takes the form \( \ell^*(\sigma) = -\nu \log \sigma - \frac{1}{2} \nu (S^*)^2 / \sigma^2 \), where \( S^* \) is drawn from the \( S(\chi^2/\nu)^1/2 \) distribution. When the pivot is linear and normal the cumulative confidence is \( C(\psi, T) = \Phi(\mu(\psi) - T) \) and the confidence likelihood is \( L(\psi, T) = \varphi(\Phi^{-1}(C(\psi, T))) \), as obtained from normal conversion. The bootstrapped likelihood is then

\[
L^*(\psi) = \varphi(\Phi^{-1}(C(\psi, T_{\text{obs}}) + Z))
\]

where \( Z \) is a standard normal draw.

10.5 Likelihoods from confidence intervals

Scientific reporting is often briefed to the 95% confidence interval \( (\psi_1, \psi_2) \) for \( \psi \) and the point estimate \( \hat{\psi} \). To merge this information with other information, a likelihood based on the information would indeed be helpful. This is elaborated further in Chapter 13.

The more information there is about how the confidence interval is established, the better a confidence likelihood based on the confidence interval is grounded. We will discuss some possible models summarising the available background information, or perhaps being just assumed models if no background information is available. The idea is to extrapolate from the point estimate and the confidence interval to a confidence distribution, in view of the model. The confidence distribution is then turned into a likelihood by exact or normal conversion.
Likelihoods and confidence likelihoods

Assume the point estimate to be median unbiased, thus $\hat{\psi}$ is the confidence median. With degree of confidence $1 - \alpha$, the confidence limits are respectively the $\alpha/2$ and $1 - \alpha/2$ confidence quantiles. If more than these three confidence quantiles are available, say in the form of confidence intervals of different degrees but based on the same data, extrapolation to the whole confidence distribution and thus the confidence likelihood is strengthened.

A statistical model behind the three confidence quantiles should be argued from the information surrounding the confidence interval. After having had a look at symmetric confidence intervals we will consider how transformation models might be used to obtain a likelihood through exact conversion of a confidence distribution obtained from the interval and how the interval might directly yield a likelihood by normal conversion.

The linear normal model for symmetric confidence intervals

A confidence interval is symmetric when $\hat{\psi} - \psi_1 = \psi_2 - \hat{\psi}$, and is otherwise skewed. Most confidence intervals obtained by statistical software are symmetric. They are often based on large-sample theory ensuring that $\hat{\psi}$ is approximately normally distributed with a standard error $s$ estimated from the Hessian of the log-likelihood function. Thus, $(\psi - \hat{\psi})/s$ is an approximate pivot with a standard normal pivotal distribution, and

$$C(\psi, \hat{\psi}) = \Phi((\psi - \hat{\psi})/s)$$

is the cumulative confidence distribution. With $z = \Phi^{-1}(1 - \alpha/2)$ the standard error is recovered from the symmetric interval as $s = (\psi_2 - \psi_1)/(2z)$. The confidence likelihood is then obtained by exact conversion, and the log-likelihood is simply

$$\ell_c(\psi) = \log \frac{d}{d\psi} C(\psi, \hat{\psi}) + \log s = -\frac{1}{2} \left( \frac{\psi - \hat{\psi}}{s} \right)^2.$$  

Normal conversion yields the same result, $\ell_n(\psi) = -\frac{1}{2} \Phi^{-1}(C(\psi, \hat{\psi}))^2$.

Normal models in transformed parameters for skewed confidence intervals

Some intervals are based on a linear normal pivot in an increasing function $h$ of $\psi$ and $\hat{\psi}$, with cumulative confidence

$$C(\psi, \hat{\psi}) = \Phi \left( \frac{h(\psi) - h(\hat{\psi})}{s} \right).$$

Both exact and normal convergence turn this confidence distribution into the confidence log-likelihood

$$\ell_c(\psi) = -\frac{1}{2} \left( \frac{h(\psi) - h(\hat{\psi})}{s} \right)^2.$$  

This model is similar to the pivot Efron uses in his abc method on bootstrapped data; cf. Section 7.6. The transformation $h$ is then estimated from the bootstrap distribution. In our case, with only three confidence quantiles at hand, strong assumptions are needed to estimate $h$. 
10.5 Likelihoods from confidence intervals

There are some cases of special interest. When \( \psi \) is population size or some other positive parameter, the log-normal pivot may apply. The cumulative confidence is then

\[
C(\psi, \hat{\psi}) = \Phi\left(\frac{\log \psi - \log \hat{\psi}}{s}\right) = \Phi\left(\frac{\log \psi - \log \hat{\psi}}{s}\right).
\]

Positive parameters might also be power transformed, say \( h(\psi, \alpha) = \text{sgn}(\alpha) \psi^{\alpha} \). The power \( \alpha \) and the scale \( s \) must then be estimated by solving the two equations

\[
\psi^\alpha - \hat{\psi}^\alpha = -sz \quad \text{and} \quad \psi^{\alpha-1} - \hat{\psi}^{\alpha-1} = sz
\]

where \( z \) is the upper \( \frac{1}{2}\alpha \) quantile in the normal distribution.

For the correlation coefficient Fisher found the zeta transform

\[
h(\psi) = \frac{1}{2} \log \left\{ \frac{(1 + \psi)}{(1 - \psi)} \right\}
\]

to provide a nearly normal pivot; cf. Section 7.8. When \( \psi \) is a probability the logit function \( h(\psi) = \log(\psi) - \log(1 - \psi) \) might work. There are also other transformations available, such as the probit \( h = \Phi^{-1} \).

**Example 10.8 Exponential model**

To see how well a reconstructed log-likelihood approximates its aim, consider an observation \( y \) from an exponential distribution with rate parameter \( \psi \). The information upon which to base the reconstruction are

\[
(\psi_1, \hat{\psi}, \psi_2) = (-\log(1 - \frac{1}{2}\alpha), \log 2, -\log(\frac{1}{2}\alpha)) / y.
\]

For \( \alpha = 0.05 \) the coefficients come out \( a = 0.32 \) and \( s = 1.32 \) independent of \( y \). Figure 10.4 shows the reconstructed log-likelihood and the log-likelihood \( \ell = \log(\psi) - \psi y \), both normalised to have maximum value zero.

**Example 10.9 Effective population size for cod**

The effective population size \( N_e \) of a given stable population is the size of a hypothetical stable population where each individual has a binomially distributed number of reproducing offsprings. The hypothetical population maintains the same genetic variability as the actual population over the generations. In a study of a cod population a point estimate and a 95% confidence interval for \( N_e \) was found from genetic data by a relatively complex method involving jackknifing; see Knutsen et al. (2011). An estimate and a 95% confidence interval were also found for the actual population size \( N \). This confidence interval was found by the standard linear normal pivot. The results for \( N_e \) and \( N \) are stochastically independent. The confidence quantiles came out as follows.

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_e )</td>
<td>106</td>
<td>198</td>
<td>1423</td>
</tr>
<tr>
<td>( N )</td>
<td>800</td>
<td>1847</td>
<td>2893</td>
</tr>
</tbody>
</table>

From these data a confidence interval is sought for the ratio \( \psi = N_e / N \). This ratio can be used to estimate the variance in number of reproductive offsprings, which is hard to estimate by other more direct methods. The confidence log-likelihood for \( N \) is simply

\[
\ell_c(N) = -\frac{1}{2} \left( N - 1847 \right)^2 / 534^2
\]

since we find \( s = 534 \). For \( N_e \) we chose the linear normal...
Figure 10.4 The exponential log-likelihood (solid line) and the reconstructed log-likelihood from a 95% confidence interval by the power transformation method (dashed line). The single observation is $y = 1$.

The confidence log-likelihood is

$$\ell_c(N_e) = -\frac{1}{2}(N_e^a - 198^a)^2/s^2.$$  

The confidence curve shown in Figure 10.5 is obtained by the $\chi^2_1$ probability transform of the profile of the confidence log-likelihood. The 95% confidence interval for $\psi$ is [0.017, 0.792]. The right confidence limit has rather misleading precision due to the flatness of the confidence curve to the right. From biological theory, $\psi \leq 1$. Since $cc(1) = 0.957$, balanced confidence intervals at levels above 95.7% have the upper limit $\psi = 1$ as their right end points. For more details, see Schweder and Hjort (2013b).

Hypothetical data are simulated to investigate the performance of the confidence curve for $\psi = N_e/N$. Balanced confidence intervals of degree 95% and point estimate, $\hat{N} \pm 534\Phi^{-1}(0.975)$ and $\hat{N}$, with $\hat{N} \sim N(1847, 534^2)$, are simulated. Skewed confidence intervals were also simulated for $N_e$. Samples of size 5 were taken from the gamma distribution with shape parameter 0.55. With $l$ being the smallest and $h$ being the largest observation, the interval $[l, h]$ has degree of confidence $\alpha = 0.0625$. The median of the five observations was used as the point estimate. The scale of the gamma distribution was chosen to yield a ‘true value’ for $N_e$ of 198. The corresponding value for the ratio is $\psi_0 = 0.1072$. 
10.5 Likelihoods from confidence intervals

Confidence intervals and point estimates for $N$ and $N_e$ were generated by this scheme in 1000 replicates, and confidence curves were calculated by the method used for the observed data in each replicate. In 433 of the replicates the simulated point estimate (at the minimum of the simulated confidence curve) were below $\psi_0$. There is thus a slight positive bias in the confidence curve when applied to data from the simulation model. This makes $cc(\psi)$ slightly unbalanced. The coverage probabilities of the level sets of $cc(\psi)$ are, however, not that far from their nominal values; see Figure 10.6. At low levels the coverage probabilities are a bit too high, whereas they are slightly too low at intermediate levels. At high levels the coverage probabilities are close to their nominal targets.

A nonlinear transformation model for skewed confidence intervals

Consider a parametric family of transformations functions $h(x, a)$. For each value of the parameter $a$ the transformation is increasing in $x$. Let $T$ be a given function of $\psi$ and $\hat{\psi}$, increasing in $\psi$. For a scale parameter $s > 0$ the model is that $(1/s)h(T(\psi, \hat{\psi}), a)$ is a pivot with given pivotal distribution. The pivotal distribution will often be the normal, making

$$C(\psi, \hat{\psi}) = \Phi \left( \frac{1/s h(T(\psi, \hat{\psi}), a)}{s} \right)$$

the cumulative confidence distribution behind the confidence interval. With degree of confidence $1 - \alpha$, the three equations

$$C(\hat{\psi}, \psi) = \frac{1}{2}, \quad C(\psi_1, \hat{\psi}) = \frac{1}{2}\alpha, \quad C(\psi_2, \hat{\psi}) = 1 - \frac{1}{2}\alpha$$
determine the two parameters \( a \) and \( s \) of the model. The confidence likelihood is found by
exact conversion (Section 7.3), and its logarithm is

\[
\ell_c(\psi) = -\frac{1}{2} \Phi^{-1}(C(\psi, \widehat{\psi}))^2 + \log\left( \frac{\partial}{\partial T} h(T, a) \right) + \log\left( \frac{\partial}{\partial \psi} T(\psi, \widehat{\psi}) \right).
\]

The power transformation model could be useful. For positive \( x \) the transforming
function is

\[
h(x, a) = \left( \frac{1}{a} \right)(x^a - 1), \quad \text{with } h(x, 0) = \log x.
\]

This transformation goes well together with \( T = \psi / \widehat{\psi} \), and yields

\[
\ell_c(\psi) = -\frac{1}{2} \Phi^{-1}(C(\psi))^2 + a \log |\psi|.
\] (10.2)

The distributional assumption behind the normal power transform model is that the statistic
\( \widehat{\psi} \) has a distribution with scale parameter proportional to \( \psi \). For \( a \geq 0 \) the distribution is
skewed with the longer tail to the right. For moderate negative power parameters \((-1 < a < -0.05)\) the distribution is more nearly symmetric, quite similar to the normal distribution,
but a bit skewed to the right for \( a \) close to zero.

When \( \psi \) is known to be positive the power transformation might be a bit unsatisfactory
for \( a > 0 \), as then \( C(0, \widehat{\psi}) = \Phi(-1/(sa)) > 0 \). For \( a \leq 0 \) this is not a problem. For \( a = 0 \) the
model is that \( \widehat{\psi} \) is log-normally distributed with median \( \psi \).

If \( \psi \) is a probability, or is otherwise restricted to an interval \( I \), a function \( g(\psi, \widehat{\psi}) \) from
\( I \) to the positive real line such that \( g(\widehat{\psi}, \widehat{\psi}) = 1 \) might be used together with the power
transformation. The model is then that

\[ C(\psi, \hat{\psi}) = \Phi\left( \frac{1}{s a} \{ g(\psi, \hat{\psi}) - 1 \} \right) \]

is the cumulative confidence.

When the observed confidence quantiles appear to be obtained through a pivot, or perhaps through an approximate pivot by way of bootstrapping, the confidence distribution obtained by extrapolation should indeed be turned into a confidence likelihood by exact conversion. When, however, the confidence interval appear to have been found by contouring a deviance function by the appropriate quantile of the chi-squared distribution, this deviance function is found by reversing the normal conversion (Section 7.2). A structural form must be assumed for the confidence distribution also in this case. Consider the normal power model. The parameters \( a \) and \( s \) are estimated from the confidence quantiles as previously. The confidence log-likelihood is then simply \( \ell_c(\psi) = -\frac{1}{2} \Phi^{-1}(C(\psi, \hat{\psi}))^2 \) and the term \( \log |\psi| \) has been lost relative to the result by exact conversion.

Sometimes more than one confidence interval is given, for example, for degrees of confidence 0.90, 0.95 and 0.99. It is then possible to evaluate how well the normal-power model fits these intervals. If the fit is unsatisfactory, the normal distribution might be replaced by a t-distribution, and the degree of freedom estimated from the confidence quantiles.

10.6 Discussion

The likelihood function is the ideal vehicle for carrying the information in a set of data into the inference. It gives a sufficient summary of the data, and it allows the future user to give the past data the appropriate weight in the combined analysis. In situations with one interest parameter and several nuisance parameters, there is a need to be freed of the nuisance parameters. This could be done in the final analysis, with past data represented by the full likelihood, or it could be done by representing the past data with a confidence likelihood for the parameter of interest. Getting rid of nuisance parameters is often a taxing operation. A thorough understanding of the data and the underlying probability mechanism is required. For this reason, and also to avoid duplication of work, it will often be natural to carry out this reduction in the initial analysis. Because the confidence likelihood usually comes for free when the pivot behind the confidence distribution is identified, it certainly seems rational to present the confidence likelihood for parameters for which the estimate is suspected to be subject to later updating.

In meta-analysis, one often experiences more variability in estimates across studies than stipulated by the individual likelihoods or standard errors. In such cases, a latent variable representing unaccounted variability in the individual confidence likelihoods might be called for; see Chapter 13. At any rate, the confidence likelihoods are better suited for meta-analysis than only estimates accompanied by standard errors.

Bootstrapping or simulation is a widely used tool in confidence estimation. The future user would thus like to simulate the past reduced data along with the new data. This is easily done when the pivot underlying the confidence likelihood representing the past data is known.
Likelihoods and confidence likelihoods

A pivot in several parameters and in the same number of statistics determines a confidence likelihood exactly as in Section 10.3. It does also dictate how to simulate the data reduced to these statistics.

When data are weak, it is sometimes tempting to include a component in the likelihood that represents prior beliefs rather than prior data. This subjective component of the likelihood should then, perhaps, be regarded as a penalising term rather than a likelihood term. Schweder and Ianelli (2001) used this approach to assess the status of the stock of bowhead whales subject to inuit whaling off of Alaska; see Section 14.3, where the bowhead whales are revisited. Instead of using the prior density as a penalising term, they used the corresponding normal-based likelihood. In one corresponding Bayesian analysis, the prior density was, of course, combined multiplicatively with the likelihood of the observed data (Raftery et al., 1995).

10.7 Notes on the literature

There is naturally a vast literature on likelihoods, their interpretation, constructions, combinations and multiple uses, since they were introduced by Fisher (1922). A part of this relates to likelihood-based methodology for construction of confidence intervals, containing therefore also a strand of work related to setting up distributions of confidence, from Bayesian posterior distributions to Fisher’s fiducial and our frequentist confidence distributions. There has been relatively little attention on aspects of the inverse problem, that of going the other way, from confidence to likelihood. A natural reason for this is the basic point made a couple of times in this chapter (see, e.g., Example 10.6), that this inverse transformation cannot be unique. That different likelihoods may lead to the same system of confidence intervals has been noted and discussed in Schweder and Hjort (1996, 2002, 2003), along with recipes for approximations, for example, based on further information regarding the data protocol.

Efron (1993) was concerned with such problems, and derived what he called the implied likelihood via an argument involving a ‘double sample’ and a parallel to Bayesian posterior calculations. Efron later expressed a bit of surprise that his method did not catch on more (personal communication to T. S., 1998). The implied likelihood, however, turns out to be the same as what we reach with our normal conversion; cf. Section 10.2 and its more general machinery. The potential uses of such implied or confidence likelihoods are manifold, but they ought to have particular promise in meta-analysis type problems, where the task is to reach inference statements when combining information across different sources; cf. Chapter 13. Such ideas are followed up in Efron (1996, 1998).

The likelihood synthesis approach of Schweder and Hjort (1996), developed further in this chapter, grew out of certain concerns with the approach called Bayesian synthesis proposed by Raftery et al. (1995), where in particular crucial issues related to the so-called Borel paradox create problems. Schweder and Ianelli (1998, 2001) use likelihood synthesis for bowhead whale population assessment; see also Schweder et al. (2010). A later refinement of the Bayesian synthesis method is the so-called Bayesian melding approach of Poole and Raftery (2000).

The nonlinear transformation models we consider in Section 10.5 have some of the flavour of, for example, Box–Cox transformations; see Box and Cox (1964). Their uses
in the present connection, that of constructing approximate likelihoods from confidence intervals, are novel.

**Exercises**

10.1 **Two Poisson experiments**: In Example 3.4 we saw that two experiments could have identical likelihood functions, but different confidence distributions. The first experiment is to observe the waiting time $T_i$ until $x$ points are observed in a Poisson point process of intensity $\lambda$, and in the other the number of points $X(t)$ observed over the period $(0, t]$ is observed. If $T_i = t$ and $X(t) = x$ the likelihoods are identical. In the waiting time experiment the confidence distribution $C(\lambda, t) = \Gamma_x(\lambda t)$ is exact. When using the half-correction in the Poisson distribution, the confidence distribution based on $X(t)$ is $C(\lambda, x) = \Gamma_x(\lambda t) - \frac{1}{2} \exp(-\lambda t)(\lambda t)^{4/3}$. The difference between the two confidence curves is small for large $x$. The normal conversion of this confidence distribution is the likelihood $L_x(\lambda) = \Gamma^{-1}_x(C(\lambda, x))$.

The square root transformation is known to bring the Poisson distribution close to the normal. By results of Exercise 8.9, $V = 2((\lambda t)^{1/2} - X(t)^{1/2})$ is approximately $\mathcal{N}(0, 1)$ and is accordingly approximately a pivot for moderate to large $\lambda$. The log-likelihood is approximately $\ell(\lambda) = \log\psi(2((\lambda t)^{1/2} - X(t)^{1/2}))$. For $x = 5$ and $t = 0.5$, construct a figure showing the exact Poisson deviance function, that obtained by normal conversion of the half-corrected Poisson confidence distribution $C(\lambda, x)$, and that based on the $V$. Averaging over $\lambda$, say over $(2, 20)$, the Poisson maximum likelihood estimator has an average median bias of $m = \text{average}[	ext{median}(X(1) - \lambda))] \approx -0.17$. A bias-corrected version of the normally converted deviance function is obtained by moving the deviance function this amount, $D_{\text{mod}}(\lambda) = 2(\ell(\lambda)) - \ell(\lambda))$. Our point here is to illustrate the approximation techniques; when the exact likelihood is available it is of course preferable, but normal conversion of the half-corrected confidence distribution does also work well. See the following exercise for more on this topic.

10.2 **Improved Poisson confidence distribution**: In Example 7.3 confidence distributions for the mean parameter $\mu$ of a Poisson distribution were discussed.

(a) Compare the likelihood to confidence likelihood obtained from the half-corrected confidence distribution; and also from the approximate confidence distributions, say for an observed value $T_{\text{obs}} = 5$.

(b) For $X$ Poisson distributed with mean $\lambda$ and $V = 2(\sqrt{X} - \sqrt{\lambda})$ show by simulation that $\exp(V/1000)$ is nearly $t_{50}$ distributed except for a scaling constant. Use this to develop a confidence distribution, and compare to the half-corrected confidence distribution with and without median bias correction.

(c) Develop the likelihood function based on the $t$-distribution of $\exp(V/1000)$. Does it provide a better approximation than that obtained by normal conversion and bias correction?

10.3 **Oxygen to the brain**: Newborn babies are tested for problems with oxygen supply to the brain (Elstad et al., 2011). Of the infants with normal temperature who tested negatively, 76% were correctly predicted to not have problems with the oxygen supply, with 95% confidence interval [69%, 82%]. Among hypothermic infants testing negatively, 78% were correctly predicted to have a normal supply of oxygen, with 95% confidence interval [67%, 86%]. Estimate likelihood functions for the probability of correct prediction for each group, first using that the confidence intervals were based on binomial variation, and then without this assumption. Test whether the probabilities are equal, and find a combined confidence distribution assuming homogeneity.
Likelihoods and confidence likelihoods

10.4 A single exponential: Consider the simplest of all cases, that there is \( n = 1 \) exponentially distributed observation \( Y \) with distribution function \( 1 - \exp(-y/\psi) \) for \( y \) positive. Find the confidence curve for \( \psi \) by exact conversion of the deviance. What is the tail asymmetry? Compare this confidence curve to the confidence curve based on the bias-corrected deviance. The confidence curve based on the directed square root deviance might also be derived. Compare these confidence curves to the tail-symmetric confidence curve resulting from pivoting.

10.5 Uniforms: Work out the median bias-corrected confidence curve obtained from the deviance for a sample of size \( n \) from the uniform distribution over the interval \((0, \psi)\). At what rate does this confidence curve approach the tail-symmetric confidence curve found by pivoting the largest observation?

10.6 Normal variance: The problem is to estimate \( \sigma \) from a sample of size \( n = 5 \) from \( N(\mu, \sigma^2) \). In Example 10.5 the confidence curve was obtained by exact conversion of the bias-corrected profile deviance. What is the confidence curve found by exact conversion of the marginal deviance, after bias correction? The tail-symmetric confidence curve is also available from the signed square root deviance, whether based on profiling or marginalisation. Compare this to the other confidence curves.

10.7 Squared normals: Here we work through simple but important formulae for the inverse \( \chi^2_1 \) transform, used also in connection with the normal conversion. Use first \( N(0, 1)^2 \sim \chi^2_1 \) to show that \( \Gamma_1(x) = 2\Phi(\sqrt{x}) - 1 \). Then show that \( \Gamma^{-1}_1(\psi) = \Phi^{-1}(\frac{1}{2} + \frac{1}{2}p)^2 \). Via \( cc(\psi) = |1 - 2C(\psi)| \), show that \( \Gamma^{-1}_1(cc(\psi)) = \Phi^{-1}(C(\psi))^2 \).

10.8 Exponential data: Consider the model of Example 10.8. Check Figure 10.4, say for an observation \( y = 2 \). Why are the coefficients of the Box–Cox transformation obtained from a confidence interval with midpoint independent of \( y \), but slightly dependent on \( \alpha \)? Compare your reconstructed confidence distribution to the true one, \( C(\psi) = 1 - \exp(-\psi y) \). Does the method associated with (10.2) work well in this case? Check also that the confidence distribution reconstructed by this method matches its target well.

10.9 Odds ratio: Elstad et al. (2011) discuss whether lowering the body temperature of a newborn by 3.5 degrees Celsius for 72 hours (therapeutic hypothermia) might prevent fatal problems with blood supply. They compared results under this treatment with published results for non-treated newborns in the same risk group (normothermia) with respect to a dichotomous index for bad development. Point estimates and confidence intervals for treated and nontreated are as follows.

<table>
<thead>
<tr>
<th></th>
<th>0.025</th>
<th>0.500</th>
<th>0.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>normothermia</td>
<td>0.73</td>
<td>0.84</td>
<td>0.91</td>
</tr>
<tr>
<td>hypothermia</td>
<td>0.45</td>
<td>0.60</td>
<td>0.74</td>
</tr>
</tbody>
</table>

The results for normothermia are collected from various published sources and reflect some overdispersion.

(a) When \( \hat{p} \) is an estimate of a binomial probability, \( \arcsin(\hat{p}^{1/2}) \) is close to normally distributed with a variance \( \sigma^2 \) nearly independent of \( p \). Assume this applies also for the results for normothermia. Estimate \( \sigma \) for treated and untreated newborns from the two sets of summary data. Do you see any sign of \( \arcsin(\hat{p}^{1/2}) \) not behaving as expected for the untreated – or perhaps that a method different from the arcsin method had been used for the summary?
Figure 10.7 The profile deviance for the odds ratio (OR), with the 95% quantile of $\chi^2_1$.

(b) Find a confidence likelihood for $p$ from the summary data. For the untreated a transformation different from Fisher’s arcsin might be used. Find an alternative transformation.

(c) The problem is to draw inference on the odds ratio the formula is given as

$$\text{OR} = \frac{p_{\text{nom}}/(1-p_{\text{nom}})}{p_{\text{hyp}}/(1-p_{\text{hyp}})},$$

see also Example 8.1. Calculate and display the profile deviance for OR, as in Figure 10.7. (Answers: $\sigma_{\text{nom}} = 0.062$, $\sigma_{\text{hyp}} = 0.077$). The point estimate and 95% confidence interval for OR by contouring the profile deviance by the $\chi^2_1$ distribution is respectively 3.53, and (1.45, 9.02).

10.10 *Humpback whales:* Paxton et al. (2006) estimated the abundance of a certain North Atlantic population of humpback whales in the years 1995 and 2001. Their estimates are obtained by aggregating line transect data from the various parts of the area by spatial smoothing. Confidence intervals were obtained by bootstrapping. Data are independent between years. This gives little help in choosing the model behind the confidence intervals, other than
abundance necessarily being nonnegative. Paxton et al. (2006) report their results by the following table:

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>3439</td>
<td>9810</td>
<td>21457</td>
</tr>
<tr>
<td>2001</td>
<td>6651</td>
<td>11319</td>
<td>21214</td>
</tr>
</tbody>
</table>

Find a confidence likelihood and a confidence curve for the yearly growth rate for this population of humpbacks by first finding a joint confidence likelihood for abundance in the two years.
11

Confidence in non- and semiparametric models

The previous chapters have been concerned primarily with confidence inference for quantities related to parametric models. Conceptually there is nothing inherently difficult with lifting the definitions to focus parameters associated with semi- and nonparametric models, however, even though one typically must rely on approximations rather than exact solutions when constructing confidence distributions. An exception is the case of nonparametric inference for cumulative distribution functions and their quantiles, where we develop both exact and approximate methods. We also examine the empirical likelihood as a means of generating nonparametric confidence curves.

11.1 Introduction

In earlier chapters we worked with so-called parametric models, where the mechanism generating data depends on some finite-dimensional parameter $\theta$. Here we have developed motivation and machinery for confidence distributions and their links with various likelihood functions when it comes to inference for focus parameters, say $\psi = a(\theta)$.

But even nonparametric models, where there is no finite-dimensional parameter describing the class of probability distributions, have focus parameters. Examples include the mean, standard deviation and skewness of an unknown probability distribution; the probability of surviving five years after an operation; the probability that $X$ is larger than $Y$ in a situation with random $(X, Y)$; the correlation coefficient; and so on. Such focus parameters may also be estimated, say with $\hat{\psi}$ estimating an appropriate $\psi = \psi(F)$, with $F$ denoting the unknown and not parametrically modelled distribution. Hence it certainly makes sense to define and construct confidence distributions and associated confidence curves and densities for such parameters. In semi- and nonparametric setups we must, however, expect to rely more on distributional approximations than in various parametric models where exact calculations are feasible. Notable exceptions to this are nonparametric analysis of distribution functions and their quantiles, treated in respectively Sections 11.2 and 11.3.

We have also seen in earlier chapters how confidence distributions are linked to various likelihood functions (likelihoods are used to construct confidence distributions, and confidence distributions in their turn may be used to construct implied likelihoods). This is different and more difficult in semi- and nonparametric settings. A useful tool here is the so-called nonparametric empirical likelihood, however, which we work with in Section 11.5.
11.2 Confidence distributions for distribution functions

Suppose $Y_1, \ldots, Y_n$ are an i.i.d. sample from an unknown distribution $F$. Such a distribution function may be estimated nonparametrically, that is, without any further assumptions, using

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I\{Y_i \leq y\}.$$ 

This is simply the binomial estimate of $F(y)$, counting the number of data points falling in the $(-\infty, y]$ window. As such we know that $F_n(y)$ is approximately a normal, with mean $F(y)$ and variance $F(y)[1 - F(y)]/n$, and that the interval

$$F_n(y) \pm z_\alpha \hat{\sigma}(y)/\sqrt{n},$$

with $\hat{\sigma}(y) = [F_n(y)(1 - F_n(y))]^{1/2}$, has level approximately equal to $\alpha$ when $z_\alpha$ is the appropriate $\Phi^{-1}(\frac{1}{2} + \frac{1}{2}\alpha)$ quantile of the standard normal. An approximation to the full confidence distribution is

$$C_{n,0}(p) = \Phi\left(\sqrt{n}(p - F_n(y))/\hat{\sigma}(y)\right),$$

with the probability $p = F(y)$ ranging from zero to one.

A more precise version is, however, possible, using that $A_n(y) = nF_n(y)$ is exactly a binomial $(n, F(y))$ variable. Using the theory of Section 3.2, see specifically Example 3.5, we have

$$C_n(p) = P_p\{A_n(y) > A_{n,obs}(y)\} + \frac{1}{2}P_p\{A_n(y) = A_{n,obs}(y)\},$$

with $B_n(a, p)$ and $b_n(a, p)$ denoting the cumulative and probability functions for the binomial $(n, p)$. This function may be used in the usual fashion to produce the confidence curve $cc_n(p) = |1 - 2C_n(p)|$ and the confidence density $cn(p) = C_n'(p)$.

Example 11.1 Confidence distributions for ancient Egyptian survival

Consider again the lifelengths of 59 women and 82 men from Roman Egypt, examined earlier in Examples 1.4 and 3.7. Figure 11.1 shows the empirical distribution functions $F_n$ for women (solid curve) and men. It is apparent that women tended to die earlier than men. To avoid making the figure too crowded we show a pointwise 95% confidence band only for the women’s age distribution, not for the men. Age limits $y_0$ equal to 10, 20, 40 years are also marked in the figure, as the associated probabilities $F(y_0)$ of dying before these age limits are handled next, in Figure 11.2. That figure gives exact confidence densities for these $F(y_0)$ probabilities, for women (upper panel) and for men (lower panel). We note in particular that the survival chances for women beyond 40 years are rather worse than for men.

11.3 Confidence distributions for quantiles

Suppose $Y_1, \ldots, Y_n$ form an i.i.d. sample from some continuous distribution $F$ on some interval on the real line. The topic discussed here is how to construct confidence intervals
11.3 Confidence distributions for quantiles

Figure 11.1  The figure displays the empirical distribution function $F_n$ for the Roman era Egyptian lifetimes, for 59 women (upper line) and for 82 men (lower line). For the women we also supply pointwise upper and lower pointwise 95% curves. Also indicated are the positions $y_0$ equal to 10, 20, 40 years along the age axis, as Figure 11.2 shows confidence densities for these $F(y_0)$.

Figure 11.2  Complementing Figure 11.1, this figure displays confidence densities for the probabilities $F(y_0)$ of having died before reaching age $y_0$ equal to 10, 20, 40 years, for women (top) and men (lower).
Confidence in non- and semiparametric models

and confidence distributions for the population median \( \mu = F^{-1}(\frac{1}{2}) \) and for other quantiles. For simplicity of presentation we take \( F \) to be increasing (i.e., without ‘flat regions’) so that these population quantiles are unique and well defined.

The order statistics \( Y_{(1)} < \cdots < Y_{(n)} \) can be represented via \( Y_{(i)} = F^{-1}(U_{(i)}) \), where \( U_{(1)} < \cdots < U_{(n)} \) are the order statistics from an i.i.d. sample from the uniform distribution over the unit interval. Focusing first on the median \( \mu \), it follows that

\[
\begin{align*}
    r_n(a, b) &= P_F(Y_{(n)} \leq \mu \leq Y_{(1)}) = P(U_{(n)} \leq \frac{1}{2} \leq U_{(1)}),
\end{align*}
\]

where \( 1 \leq a < b \leq n \). This probability can be computed exactly, as shown later, with the consequence that \((a, b)\) can be selected to provide, for example, a 90% confidence interval for \( \mu \) – modulo the mild complication that there are only finitely many such candidate intervals, so we cannot expect to find an interval with coverage exactly equal to 0.90.

The joint density of \((U, V) = (U_{(a)}, U_{(b)})\) is proportional to

\[
    u^{a-1}(v-u)^{b-a-1}(1-v)^{n-b} \quad \text{for} \quad 0 < u < v < 1,
\]

and the density of \( U \) proportional to \( u^{a-1}(1-u)^{n-a} \), that is, a Beta\((a, n-a+1)\); see, for example, David and Nagaraja (2003). Dividing the former with the latter we find that \( V|U = u \) has a density proportional to \((v-u)^{b-a-1}(1-v)^{n-b}\) on \((u, 1)\), which we may more conveniently represent in terms of \( W = (V-u)/(1-u) \) having a Beta\((b-a, n-b+1)\) distribution. Hence

\[
    r_n(a, b) = \int_{0}^{1/2} P\left[ W \geq \frac{1}{2} - \frac{u}{1-u} \bigg| U \right] \text{Be}(u, a, n-a+1) \, du
    = \int_{0}^{1/2} \left\{ 1 - \text{Be}\left( \frac{1}{2} - \frac{u}{1-u}, b-a, n-b+1 \right) \right\} \text{Be}(u, a, n-a+1) \, du,
\]

in terms of the density function \( \text{Be}(u, c, d) \) and cumulative distribution function \( \text{Be}(u, c, d) \) for the Beta\((c, d)\) distribution. The point is that these coverage probabilities can be computed via numerical integration, and so to speak kept and accessed as a table for given sample size \( n \).

**Example 11.2 Nonparametric confidence for the median**

For an illustration we generated a sample of size \( n = 100 \) from the unit exponential distribution, where the true median is \( \log 2 = 0.693 \). The machinery is applied to all \( n/2 = 50 \) candidate confidence intervals of the type \([Y_{(a)}, Y_{(n-a)}]\), with the preceding integral representation used to compute the associated exact coverages \( r_n(a, n-a) \). Thus the shortest nontrivial interval \([Y_{(50)}, Y_{(51)}]\) has coverage 0.079, intervals \([Y_{(42)}, Y_{(59)}]\) and \([Y_{(40)}, Y_{(61)}]\) have coverage degrees 0.911 and 0.965, and so forth. Figure 11.3 displays these intervals.

We note that other combinations of pairs are possible too, in addition to these fairly natural \((a, n-a)\) pairs. Degrees of nonsymmetry in the distributions are, however, automatically reflected in the intervals, that is, the \([Y_{(a)}, Y_{(n-a)}]\) are not symmetric around the median; cf. again Figure 11.3.

The preceding methods may be extended without difficulties to general population quantiles, say \( \mu = F^{-1}(p) \) for any \( p \in (0, 1) \). Here (11.2) and its ensuing representation
11.3 Confidence distributions for quantiles

Figure 11.3  Exact confidence intervals with a smoothed confidence curve for the median $\mu$ based on a sample of size $n = 100$ from the unit exponential distribution. The true value $\mu = \log 2$ is also indicated.

generalise to

$$s_n(a, b) = P_F\{Y(a) \leq \mu \leq Y(b)\}$$
$$= P\{U(a) \leq \mu \leq U(b)\}$$

$$= \int_0^q \left[ 1 - \text{Be}\left(\frac{\mu - u}{1 - u}, b - a, n - b + 1\right) \right] \text{Be}(u, a, n - a + 1) \, du,$$

and these may be computed and utilised for, for example, pairs of order statistics of the type $a = \lfloor np \rfloor - j$ and $b = \lfloor np \rfloor + j$.

Example 11.3 Quantiles for birthweight distributions

There is a large literature on factors that influence chances of having babies with low or too low birthweight. Recently more attention has been given to analyses of factors related to having big or too big babies, however, as such events may indicate various other types of health risks, for both the baby and the mother. Voldner et al. (2008) consider a certain cohort of 1028 mothers, all giving birth at the Oslo University Hospital Rikshospitalet inside the 2001–2008 period, generally aiming to extend insights into maternal metabolic syndrome and determinants of foetal macrosomia. The birthweights of the 548 boys range from 1275 to 5140 grams, and those of the 480 girls from 600 to 5420 grams.

We may now apply the method and formulae above to construct confidence intervals and indeed confidence distributions for any quantiles of interest, say, $q_1, q_2, q_3$ corresponding to levels 0.05, 0.50, 0.95. The intervals we work with are of the type

$$q_1 \in [Y(j_1 - a_1), Y(j_1 + a_1)], \quad q_2 \in [Y(j_2 - a_2), Y(j_2 + a_2)], \quad q_3 \in [Y(j_3 - a_3), Y(j_3 + a_3)],$$
Confidence in non- and semiparametric models

Figure 11.4  Confidence intervals and curves for the 0.05, 0.50, 0.95 quantiles for the birthweight distribution of 1028 babies of Example 11.3, in grams, for boys (upper) and girls (lower). The estimated quantiles are respectively 2801, 3690, 4584 for boys and 2569, 3536, 4429 for girls.

where for the boys $j_1 = 27, j_2 = 274, j_3 = 521$ correspond to the estimated quantiles, which are 2801, 3690, 4584; similarly, for the girls, $j_1 = 24, j_2 = 240, j_3 = 456$, corresponding to estimated quantiles 2569, 3536, 4429. Figure 11.4 displays all these confidence intervals along with their degrees of confidence and enveloped by confidence curves. Note that the confidence curves for the medians are rather tighter than for the extreme quantiles. The reason for this is roughly speaking simply that there are more data in the middle of these distributions; the median is easier to estimate with good precision than outer quantiles.

For a further illustration, Figure 11.5 displays the exact cumulative confidence distribution for the 0.95 quantile of this birthweight distribution, along with its normal approximation.

As we saw the collection of confidence intervals generated by the preceding method gives rise via (11.3) to a confidence curve, say $cc_n(\mu)$, for the quantile $\mu = F^{-1}(p)$. This may be translated to a confidence distribution curve, say

$$C_n(\mu) = \begin{cases} \frac{1}{2}(1 - cc_n(\mu)) & \text{if } cc_n(\mu) \leq \frac{1}{2}, \\ \frac{1}{2}(1 + cc_n(\mu)) & \text{if } cc_n(\mu) \geq \frac{1}{2}. \end{cases}$$

The machinery is exact, but is associated with the modest practical inconvenience that not all confidence levels are attained and that the $cc_n$ and $C_n$ curves are not smooth.
11.3 Confidence distributions for quantiles

An alternative is to turn to large-sample theory and the associated approximate normality of sample quantiles. It is known that

\[ \sqrt{n} \{ \hat{\mu}(p) - \mu(p) \} \rightarrow_d q(p) W^0(p) \sim N(0, q(p)^2 p(1 - p)), \]  

where \( \hat{\mu}(p) = Q_n(p) = F_n^{-1}(p) \) is the sample quantile estimator of the population quantile \( \mu(p) \). Here \( W^0(\cdot) \) is a Brownian bridge, a zero-mean normal process with variance \( p(1 - p) \) at position \( p \), and \( q(p) = (F^{-1})'(p) = 1/f(\mu(p)) \) is the so-called quantile density. It follows that

\[ \sqrt{n} \{ \mu(p) - \hat{\mu}(p) \} \rightarrow_d N(0, 1), \]

as long as \( \hat{q}(p) \) is constructed to be a consistent estimator of \( q(p) \), and that the ensuing

\[ C_n(\mu(p)) = \Phi \left( \frac{\sqrt{n} \{ \mu(p) - \hat{\mu}(p) \}}{\hat{q}(p) |p(1 - p)|^{1/2}} \right) \]

is a first-order correct confidence distribution for the \( p \)-quantile. For \( \hat{q}(p) \) various alternatives are available; one simple method is

\[ \hat{q}(p) = \{Q_n(p + \varepsilon) - Q_n(p - \varepsilon)\}/(2\varepsilon), \]

for a suitably small \( \varepsilon \). Technically speaking it is necessary that \( \varepsilon \rightarrow 0 \) (the bias tending to zero) and \( n\varepsilon \rightarrow \infty \) (the variance tending to zero) as \( n \) increases, to ensure the required consistency. A practical alternative is to use \texttt{density} in \texttt{R}. For more details and discussion of smoother alternatives, see Sheather and Marron (1990), Jones (1992) and Hjort and Petrone (2007).

Figure 11.5  Exact confidence distribution (solid line) for the 0.95 quantile of the birthweight distribution for the \( n = 548 \) baby boys of Example 11.3, in grams, along with the normal approximation (dashed line).
Confidence in non- and semiparametric models

Figure 11.6  Exact confidence curves for the five deciles (quantiles of level 0.1, 0.3, 0.5, 0.7, 0.9), for the body mass index datasets pertaining to all Olympic speed skaters from 1952 to 2010, amounting to respectively 1080 men and 741 women. As many as 17% of these male Olympians are overweight, according to the WHO criteria.

Figure 11.5 illustrates this normal approximation based on (11.5), along with the exact confidence distribution based on (11.3), for the case of the 0.95 quantile of the birthweight distribution for boys considered in Example 11.3. The approximation is satisfactory, in this case even for the 0.95 quantile, where the quantile density \( q(p) \) is hard to estimate accurately; the approximation is typically more accurate for less extreme quantiles.

Example 11.4 Deciles of BMI distribution
Consider the datasets on body mass index (weight/height\(^2\), in kg/m\(^2\)) gathered on all participants in Olympic speedskating events from 1952 to 2010, comprising a total of 1080 men and 741 women. We use the method of exact confidence curves for quantiles to provide such plots for the five deciles (associated with levels 0.1, 0.3, 0.5, 0.7, 0.9), for men and for women; see Figure 11.6. Identical scales are used for men (on top) and women (below), to highlight the difference in their BMI distributions. Interestingly, the World Health Organization considers a total of 187 of these 1080 male Olympians (and 19 of the 741 women) overweight (or ‘pre-obese’, to use WHO’s technical term); also, 6 of the women would be classified as underweight. See also Example 3.8.

11.4 Wilcoxon for location
The location \( \mu \) of a continuous distribution is often of particular interest. When the distribution \( F \) is symmetric, that is, \( F(\mu - y) = 1 - F(\mu + y) \) for all \( y \), \( \mu \) is the median of the distribution. Previously we discussed confidence inference for quantiles, including the median, without assuming symmetry. This method is related to the sign test which rejects the \( H_0: \mu \leq \mu_0 \) if sufficiently many of the differences \( Y_i - \mu_0 \) are positive in the sample of
11.5 Empirical likelihood

The critical number is found from the binomial distribution with parameters \( n \) and \( p = \frac{1}{2} \).

Under symmetry, the Wilcoxon test (Lehmann, 1975, Oja, 2010) is known to be more powerful than the sign test. The two-sided Wilcoxon test for \( H_0: \mu = \mu_0 \) is based on the sum of ranks of \(|Y_i - \mu_0|\) belonging to observations above \( \mu_0 \). If this sum is either critically small or big, \( H_0 \) is rejected. The p-value is easily computed in R and also in other software. Let \( p(m) \) be the p-value when \( m = \mu_0 \), and let \( \text{CI}_\alpha = \{m: p(m) \geq 1 - \alpha\} \) which is an \( \alpha \) level confidence interval. We call the confidence curve having these intervals as level curves Wilcoxon confidence curves.

The median unbiased point estimate obtained from the Wilcoxon confidence curve is called the Hodges–Lehmann estimator. It might be computed as the median of the pairwise means \((Y_i + Y_j)/2\) obtained from the sample, with \( i \leq j \) (i.e., the diagonal elements where \( i = j \) are included). The Hodges–Lehmann estimator is asymptotically normally distributed about \( \mu \) and with a standard error that might be computed by a window type estimator; see Schweder (1975) or Oja (2010). For a large sample this yields a quick and good approximate confidence distribution. In a little detail, \( \sqrt{n}\hat{\mu}/\kappa \) is asymptotically normal with \( \kappa^2 = (1/12)/\lambda^2 \), where \( \lambda = \int f(y)^2 \, dy \), and various density estimation methods lead to good estimates of \( \lambda \) and hence \( \kappa \). The approximation to the Wilcoxon confidence distribution is \( \Phi(\sqrt{n}\hat{\mu}/\kappa) \).

Rank methods might be used for a variety of purposes. Linear regression might be estimated by rank methods, for example, see Oja (2010). Such analyses are robust against long tails in the residual distribution, but might fail if there is strong asymmetry in the distribution. Nonparametric methods tackling this problem are discussed in the next section.

Example 11.5 Median of BMI distribution (continued)

The BMI for female Olympic speedskaters is roughly symmetric; see Figure 11.6. Plotting the empirical distribution functions of \( y_i - \hat{\mu} \) and \( \hat{\mu} - y_i \), where \( \hat{\mu} = 22.110 \) is the Hodges–Lehmann estimate, yields two curves very close to each other; see Exercise 11.6. Figure 11.7 shows the Wilcoxon confidence curve together with the confidence curve of the sign test. The approximate confidence curve based on the Hodges–Lehmann estimator is also shown.

11.5 Empirical likelihood

Likelihoods are naturally associated with parametric models, but certain constructions in the literature have aimed at lifting this concept to semi- and nonparametric settings. One such is the empirical likelihood (EL) of Owen (1990, 2001). We shall explain here how the EL may be used to form confidence distributions for population parameters also outside the realm of parametric models.

**Empirical likelihood for i.i.d. data**

Discussing the i.i.d. case first, suppose \( Y_1, \ldots, Y_n \) are independent observations from some unknown distribution \( F \) on some sample space, and suppose a \( p \)-dimensional parameter \( \theta_0 = \theta_0(F) \) is identified via a suitable estimating equation, say \( E_F m(Y_i, \theta) = 0 \), where \( m(y, \theta) \)
Figure 11.7  Confidence distributions for the symmetry point of location for the BMI of 741 female Olympic speedskaters, based on the sign test, the Wilcoxon test and the normal approximation to the Hodges–Lehmann estimator. The latter two are almost identical.

contains components \( m_1(y, \theta), \ldots, m_p(y, \theta) \). An example is \( m_j(y, \theta) = g_j(y) - \theta_j \) for chosen \( g_1, \ldots, g_p \), in which case \( \theta_{0,j} \) is the mean value \( E_g(Y) \). For another illustration, assume data are one-dimensional, and consider \( m_j(y, \theta) = I\{y \leq \theta_j\} - j/10 \) for \( j = 1, \ldots, 9 \); then \( \theta_{0,1}, \ldots, \theta_{0,9} \) are the nine deciles of \( F \).

The EL is now defined as

\[
EL_n(\theta) = \max \left\{ \prod_{i=1}^{n} (nw_i) : \text{each } w_i > 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i m(Y_i, \theta) = 0 \right\}. \tag{11.6}
\]

The background motivating this definition is as follows. Consider the probability distribution with weights \( w_1, \ldots, w_n \) attached to data points \( y_1, \ldots, y_n \). Then the likelihood expression \( L_n(w) = \prod_{i=1}^{n} w_i \) reaches its unconstrained maximum when these probability weights are equal, that is, for \( w_0^* = 1/n \), see Exercise 11.3. Thus in this particular sense the empirical distribution \( F_n = \sum_{i=1}^{n} (1/n) \delta(y_i) \), associating weight \( 1/n \) to each data point, may be interpreted as the unconstrained maximum likelihood estimator of the unknown distribution \( F \). It also follows that \( EL_n(\theta) \) is the maximum of the natural multinomial-type likelihood-ratio quantity \( L_n(w)/L_n(w^0) \), under the \( p \) constraints \( \sum_{i=1}^{n} w_i m_j(y_i, \theta) = 0 \) for \( j = 1, \ldots, p \). See Owen (2001) and Hjort et al. (2009) for further discussion pertaining to this.

To compute and assess performance of the EL, let us consider the Lagrange multiplier problem associated with the preceding maximum operation;

\[
\sum_{i=1}^{n} \log w_i - n\lambda_0 \left( \sum_{i=1}^{n} w_i - 1 \right) - n\lambda^* \sum_{i=1}^{n} w_i M_i,
\]
with \( n\lambda_0 \) and \( n\lambda = (n\lambda_1, \ldots, n\lambda_p)^t \) Lagrange multipliers and with \( M_i \) shorthand for \( m(Y_i, \theta)/\sqrt{n} \). The particular scaling used, with Lagrange multipliers proportional to \( n \) and \( m(Y_i, \theta) \) scaled down with \( \sqrt{n} \), is not strictly necessary, but facilitates our mathematical presentation that follows. Taking derivatives yields equations

\[
1/w_i = n\lambda_0 + n\lambda_i'M_i, \quad \sum_{i=1}^n w_i = 1, \quad \sum_{i=1}^n w_i M_i = 0.
\]

Multiplying the first equation with \( w_i \) leads to \( \lambda_0 = 1 \), so \( nw_i = 1/(1 + \lambda_i'M_i) \) and

\[
\log \text{EL}_n(\theta) = -\sum_{i=1}^n \log(1 + \lambda_i'M_i),
\]

where the random \( \lambda \) in question is determined by

\[
n^{-1} \sum_{i=1}^n M_i/(1 + \lambda_i'M_i) = 0.
\]

With a bit of further analysis it becomes clear that this \( \lambda \) is the maximiser of

\[
G_n(\lambda) = 2 \sum_{i=1}^n \log(1 + \lambda_i'M_i)
\]

and that

\[
Z_n(\theta) = -2 \log \text{EL}_n(\theta) = \max G_n(\cdot) = G_n(\hat{\lambda}).
\]

Here the \( G_n \) function also depends on \( \theta \), but we choose not to overburden the notation to indicate this. This provides a practical way to actually compute the EL statistic of (11.6), or its logarithm, via EL

\[
\text{EL}_n(\theta) = \exp(-1/2 Z_n(\theta)),
\]

and with the \( Z_n(\theta) \) computed for each candidate value \( \theta \) via numerical optimisation of the \( G_n \) function. A practically useful start value for such an iterative optimisation algorithm is given in the text that follows, in fact as \( \hat{\lambda} = S_n^{-1}U_n \); see (11.9) and (11.10).

To study certain useful approximations to \( Z_n(\theta) \) and hence \( \text{EL}_n(\theta) \), and to learn about their behaviour and performance, we need to study the random maximiser and random maximum of functions of the type \( G_n \). Clearly this behaviour depends on the properties of the \( M_1, \ldots, M_n \). To formulate and prove the fundamental lemma driving these large-sample approximations and limit distributions, assume at the moment merely that \( M_1, \ldots, M_n \) are certain random variables, that is, not necessarily of the precise form \( m(Y_i, \theta)/\sqrt{n} \) used earlier.

We shall find it useful to work with the two-term Taylor approximation to \( G_n \), and the associated maximiser and maximum of this function,

\[
G^*_n(\lambda) = 2\lambda'U_n - \lambda'S_n\lambda \quad \text{and} \quad Z^*_n(\theta) = \max G^*_n(\cdot) = G^*_n(\lambda^*),
\]

in terms of

\[
U_n = \sum_{i=1}^n M_i \quad \text{and} \quad S_n = \sum_{i=1}^n M_iM_i'.
\]

In fact \( \lambda^* = S_n^{-1}U_n \), leading to the explicit formula \( Z^*_n(\theta) = U_n'S_n^{-1}U_n \) for this approximation to the more complicated \( Z_n(\theta) \). The fundamental lemma is now as follows, stated here with fairly canonical conditions.

**Lemma 11.1 (Basic lemma for empirical likelihood)** Let \( M_1, \ldots, M_n \) be random \( p \)-dimensional variables and consider the ensuing random functions \( G_n(\lambda) \) and \( G^*_n(\lambda) \)
as in (11.7) and (11.9). As \( n \) goes to infinity, assume the following conditions hold, with \( U \) a random vector and the limit \( p \times p \) matrix \( \Sigma \) of full rank:

\[
(A0) \quad P(\max G_n = 0) \to 0; \\
(A1) \quad U_n = \sum_{i=1}^n M_i \to_d U; \\
(A2) \quad S_n = \sum_{i=1}^n M_i M_i^T \to_{pr} \Sigma; \\
(A3) \quad D_n = \max_{i \leq n} \|M_i\| \to_{pr} 0.
\]

Then \( Z_n = \max G_n = G_n(\hat{\lambda}) \) and \( Z^*_n = \max G^*_n = G^*_n(\lambda^*) \) are asymptotically equivalent, and \( Z_n \to_d U^T \Sigma^{-1} U \).

Here (A0) is the basic existence condition needed for EL to be useful, and is equivalent to \( P(0 \in C_n) \to 1 \), where \( C_n \) is the interior of the convex hull spanned by \( M_1, \ldots, M_n \). When this indeed holds, the function \( G_n \) has a unique nonzero maximiser, which also translates to there being a unique maximiser of \( \prod_{i=1}^n (n w_i) \) of the EL definition (11.6), with the constraints given there, inside the interior of the simplex. The (A0) condition holds with very few restrictions in the case of finite \( p \) and growing \( n \), as here; see Owen (2001, p. 219). The situation is more complicated with growing \( p \); see Hjort et al. (2009).

**Proof.** The essence of the proof is as follows, with more details and further discussion available in Hjort et al. (2009, section 2). Condition (A3) sees to it that the difference \( G_n(\hat{\lambda}) - G^*_n(\lambda^*) \) is zero in probability, uniformly over all compact regions of \( \lambda \). Both \( G_n \) and \( G^*_n \) are concave, and the maximiser of the latter is \( \lambda^* = S_n^{-1} U_n \). Via arguments in the paper mentioned, we have both \( \|\lambda\| \) and \( \|\lambda^*\| \) of order \( O_p(1) \), under conditions (A1) and (A2), which then lead to \( Z_n - Z^*_n \to_{pr} 0 \). But it is easy enough to work with \( Z^*_n = U^*_n S_n^{-1} U_n \) and to prove that it indeed converges in distribution to \( U^* \Sigma^{-1} U \) under conditions (A1) and (A2), and the conclusion follows.

In most applications the limit \( U \) of (A1) is multinormal, say \( N_p(\zeta, \Omega) \), in which case the quadratic form limit \( \Lambda = U^T \Sigma^{-1} U \) may be expressed as a linear combination of independent noncentral \( \chi^2 \) variables. If in particular \( U \sim N_p(0, \Omega) \), then some analysis reveals that \( \sum_{i=1}^n d_i N_i^2 \), where the \( N_i \) are i.i.d. standard normal and the \( d_i \) are the eigenvalues of \( \Omega \Sigma^{-1} \). For the case of \( \Omega = \Sigma \), the quadratic form is simply a \( \chi^2 \). See Exercise 11.4.

With Lemma 11.1 sorted out we may return to the study of the EL. Under the true distribution \( F \), with the implied true value of \( \theta_0 = \theta_0(F) \), the case of EL\(_n\)\((\theta_0)\) corresponds to the situation of Lemma 11.1 with \( M_i = m(Y_i, \theta_0) / \sqrt{n} \). As long as \( M_i \) has a finite variance matrix \( \Sigma \), it is clear from the law of large numbers that (A2) holds with this matrix, and from the central limit theorem that (A1) holds with \( U \sim N_p(0, \Sigma) \), featuring the same matrix. Since also (A3) holds, under the finite second moment condition (see again Hjort et al., 2009), we have shown that

\[
Z_n(\theta_0) = -2 \log \text{EL}_n(\theta_0) \to_{d} \chi^2_p, \tag{11.11}
\]

the so-called Wilks theorem for the EL. In particular this leads to nonparametrically formed confidence curves, say

\[
\text{cc}(\theta) = \Gamma_p(\sqrt{Z_n(\theta)}).
\]
Example 11.6  EL confidence curve for a birthweight quantile

Consider again the $n = 548$ birthweights of Example 11.3, where we now use the EL to construct confidence intervals and indeed a full confidence curve for the 0.95 quantile $\mu$. For this we use the EL machinery of (11.6) with estimating equation $m(y, \mu) = I(y \leq \mu) - 0.95$, and first compute $Z_n(\mu)$ of (11.8) via numerical optimisation, for each of a fine grid of $\mu$ values. In view of (11.11), the confidence curve $cc(\mu) = \Gamma_1(Z_n(\mu))$ is first-order correct, that is, confidence intervals can be read off from $cc(\mu) \leq \alpha$. The same is true for its approximation $\Gamma_1(Z_n^*(\mu))$, which is seen in Figure 11.8 to be very close indeed. The third curve given in the figure is that associated with the exact order statistics calculations of (11.3). \[ \blacksquare \]

The estimating equation $m(y, \mu)$ used for the preceding example is rather non-smooth (changing value from $-0.95$ to $0.05$ as $\mu$ passes from the left to the right of $y$), reflected in the stepwise flat nature of the resulting $Z_n(\mu)$ and $Z_n^*(\mu)$ curves and their further chi-squared transformed versions depicted in Figure 11.8. For various other applications, for example, when using $m(y, \theta) = g(y) - \theta$ for estimating the mean of $g(Y)$, there is full smoothness in $\theta$, also for the $Z_n$ and $Z_n^*$ curves.

Importantly, the EL can also be profiled, in cases where a focus parameter $\psi = a(\theta)$ is of interest. We then let

\[ \text{PEL}_n(\psi) = \max \{ \text{EL}_n(\theta): a(\theta) = \psi \}, \]

Figure 11.8  For the 0.95 quantile $\mu$ of the birthweight distribution of the $n = 548$ baby boys considered in Example 11.3, the figure displays the two confidence curves $\Gamma_1(Z_n(\mu))$ (solid line) and $\Gamma_1(Z_n^*(\mu))$ (dashed line) associated with the empirical likelihood and its approximation. Also given is the exact confidence curve based on (11.3) (smoother curve).
along with
\[
\Pi_n(\psi) = -2\log \text{PEL}_n(\psi) = \min \{ Z_n(\theta): \ a(\theta) = \psi \}. \tag{11.12}
\]

One may now show that there is a Wilks theorem also for this profile function, that is, that
\[
\Pi_n(\psi_0) = -2\log \text{PEL}_n(\psi_0) \rightarrow_d \chi^2_1, \tag{11.13}
\]
under the conditions of the true distribution \(F\), with \(\psi_0 = a(\theta_0)\) the implied true value of the parameter \(\psi\). Proving this may essentially be accomplished appealing to the proximity of \(Z_n(\theta)\) to the simpler quadratic criterion function \(Z^*_n(\theta)\), via the closeness of \(G_n\) and \(G^*_n\), and then argue as in Section 2.4; see in particular Theorem 2.4 and Remark 2.5. For further details and discussion, see Qin and Lawless (1994) and Hjort et al. (2009). The consequence is that confidence intervals may be read off from \(\Pi_n(\psi) \leq \Gamma_1^{-1}(\alpha)\), or directly from \(cc(\psi) = \Gamma_1(\Pi_n(\psi)) \leq \alpha\).

**Remark 11.2 Using EL to robustify maximum likelihood analysis**

Assume i.i.d. data \(Y_1, \ldots, Y_n\) are to be fitted to some smooth parametric model \(f(y, \theta)\) with say \(\theta = (\theta_1, \ldots, \theta_p)^t\). The maximum likelihood estimator \(\hat{\theta}\) can be defined as the solution to the equations \(\sum_{i=1}^n u_i(Y_i, \theta) = 0\) for \(j = 1, \ldots, p\), where the \(u_j(y, \theta) = \partial \log f(y, \theta) / \partial \theta\) are the score functions; cf. the theory of Sections 2.2–2.4. This fits the EL machinery, with the \(u_j(y, \theta)\) as estimating equations; hence \(\log \text{EL}_n(\theta)\) can be computed and compared to the traditional \(\ell_n(\theta)\). The theory developed above, applied to the case \(M_i = u(Y_i, \theta)/\sqrt{n}\), implies that
\[
\log \text{EL}_n(\theta) = -\frac{1}{2}Z_n(\theta) \approx -\frac{1}{2}Z^*_n(\theta) = -\frac{1}{2}n\hat{U}_n(\theta)^t S_n(\theta)^{-1} \hat{U}_n(\theta),
\]
in terms of \(\hat{U}_n(\theta) = n^{-1} \sum_{i=1}^n u(Y_i, \theta)\) and
\[
S_n(\theta) = n^{-1} \sum_{i=1}^n [u(Y_i, \theta) - \hat{U}_n(\theta)] [u(Y_i, \theta) - \hat{U}_n(\theta)]^t.
\]
The first point now is that \(\log \text{EL}_n(\cdot)\) and \(\ell_n(\cdot)\) are both maximised at the same position, that is, for the maximum likelihood estimator \(\hat{\theta}\), and that they are both aiming for the same least false parameter value \(\theta_0\) which minimises the Kullback–Leibler distance from truth to model; see Section A.5. The second point is, however, that they have different degrees of tightness around this maximum point. The usual log-likelihood function trusts the model but may give the wrong impression of the uncertainty of the maximum likelihood estimator. The EL method succeeds nonparametrically in detecting the right amount of uncertainty, however, and the \(\chi^2_p\) results of (11.11) and (11.13) provide correct confidence regions and intervals. For an illustration of this, see Figure 11.9.

**Empirical likelihood for regression models**

The EL machinery can also be extended to regression models and there typically provides a model robust alternative to traditional methods based on parametric likelihoods. Suppose such a regression model postulates that \(Y_i | \alpha_i\) stems from \(f(y | \alpha_i, \theta)\), for \(i = 1, \ldots, n\), and that the \(Y_i\) are independent. The \(\theta\) parameter would typically comprise regression coefficients
11.5 Empirical likelihood

Figure 11.9 Here \( n = 100 \) data points are generated from the \( N(\theta, \sigma^2) \) distribution, with \( \sigma_0 = 1.5 \), whereas the model used to compute likelihood function and confidence curves is that of the \( N(\theta, 1) \). The left panel displays the model-based \( \ell_n(\theta) \) curve (dotted line), which here is tighter than the two other curves, respectively \( \log EL_n(\theta) \) and its approximation \( -\frac{1}{2} Z_n^*(\theta) \); see Remark 11.2, with \( m(y, \theta) = y - \theta \) as the basic estimating equation. The right panel displays the consequent confidence curves, again with the model deviance based \( cc(\theta) \) being too optimistic but with the two others, based on the nonparametric EL, providing adequate and model-robust intervals for the mean.

along with other parameters related to the error distribution, like the spread. The ordinary log-likelihood function is \( \ell_n = \sum_{i=1}^{n} \log f(y_i | x_i, \theta) \), and the maximum likelihood estimator is the solution to

\[
\sum_{i=1}^{n} u(y_i | x_i, \theta) = 0,
\]

where \( u(y | x, \theta) = \frac{\partial \log f(y | x, \theta)}{\partial \theta} \) is the score function. This invites using the EL apparatus with \( u(y | x, \theta) \) as estimating equation; that is, \( EL_n(\theta) \) is as in (11.6), maximising \( \prod_{i=1}^{n} w_i \) over all \( w_1, \ldots, w_n \) in the probability simplex under the constraints \( \sum_{i=1}^{n} w_i u(Y_i | x_i, \theta) = 0 \). Lemma 11.1 applies, with \( M_i = u(Y_i | x_i, \theta) / \sqrt{n} \), and again \( Z_n(\theta) = -2 \log EL_n(\theta) \) is close to

\[
Z_n^*(\theta) = U_n(\theta)^T S_n(\theta)^{-1} U_n(\theta),
\]

where

\[
U_n(\theta) = n^{-1/2} \sum_{i=1}^{n} u(Y_i | x_i, \theta) \quad \text{and} \quad S_n(\theta) = n^{-1} \sum_{i=1}^{n} u(Y_i | x_i, \theta) u(Y_i | x_i, \theta)^T.
\]
Confidence in non- and semiparametric models

Under the true state of affairs, corresponding also to the least false parameter value \( \theta_0 \), see the discussion of Section 2.6, we have

\[
Z_n(\theta_0) = -2 \log \text{EL}_n(\theta_0) \to_d U^1 \Sigma^{-1} U \sim \chi^2_p,
\]

where \( U \sim N_p(0, \Sigma) \) is the limit distribution of \( U_n(\theta_0) \) and \( \Sigma \) indeed is the same as the probability limit of \( S_n(\theta_0) \). These results, related to securing conditions (A0)–(A3) of Lemma 11.1, hold under mild assumptions via the Lindeberg theorem. Similarly, for a profiled EL, with a focus parameter \( \psi = a(\theta) \) of interest, we may form \( \text{PEL}_n(\psi) \) and \( \Pi_n(\psi) = -2 \log \text{PEL}_n(\psi) \) as with (11.12), and the Wilks type theorem (11.13) continues to hold.

To illustrate the use of the EL for regression models, consider count data \( Y_i \) associated with \( p \)-dimensional covariate vectors \( x_i \), and which we model as \( Y_i | x_i \sim \text{Pois}(\mu_i) \) with \( \mu_i = \exp(x_i^\top \beta) \). The log-likelihood function is

\[
\ell_n(\beta) = \sum_{i=1}^n \{y_i x_i^\top \beta - \exp(x_i^\top \beta) - \log(y_i!}\},
\]

and the estimating equations governing the EL are \( \sum_{i=1}^n (y_i - \mu_i) x_i = 0 \), i.e. \( m(y|x,\beta) = (y - \exp(x^\top \beta))x \). With \( M_i = \{Y_i - \exp(x_i^\top \beta_0)/\sqrt{n} \) at the true or least false parameter value \( \beta_0 \) we have

\[
U_n(\beta_0) = \sum_{i=1}^n M_i = n^{-1/2} \sum_{i=1}^n (Y_i - \exp(x_i^\top \beta_0))/\sqrt{n} \to_d U \sim N_p(0,K),
\]

where \( K \) is the limit in probability matrix of \( K_n = n^{-1} \sum_{i=1}^n \text{Var}(Y_i|x_i)x_i x_i^\top \); similarly, \( S_n(\beta_0) = n^{-1} \sum_{i=1}^n (Y_i - \exp(x_i^\top \beta_0))^2 x_i x_i^\top \), from the general recipe of Lemma 11.1 tends to the same \( \Sigma \), under mild conditions on the sequence of \( x_i \). Thus the Wilks theorems hold for the EL, for the full \( \beta \) vector as well as for any profiled version to deal with focus parameters \( \psi = a(\beta) \).

The point here is as with Remark 11.2, that the nonparametric EL succeeds in automatically finding the correct amount of variability inherent in the maximum likelihood estimator \( \hat{\beta} \); in the preceding calculations the Poisson structure of the model is not used beyond giving rise to the estimating equations \( \sum_{i=1}^n (y_i - \exp(x_i^\top \beta))x_i = 0 \). The ordinary likelihood method, which, however, does use the Poisson structure to its fuller intent, involves the appropriate Hessian matrix of the log-likelihood function, that is, \( J_n = n^{-1} \sum_{i=1}^n \exp(x_i^\top \beta_0)x_i x_i^\top \) and its limit \( J \), and from the theory of Chapter 2 we know that \( \sqrt{n}(\hat{\beta} - \beta_0) \to_d N_p(0,J^{-1}KJ^{-1}) \). Only under Poisson conditions can we expect \( J = K \). When there is overdispersion, for example, with elements of \( K \) bigger than those of \( J \), ordinary likelihood analysis yields too small confidence intervals and indeed too optimistically tight confidence distributions. The EL automatically repairs for this.

### 11.6 Notes on the literature

Inference for a cumulative distribution function \( F \) at a given location \( y_0 \) is an easy task via the empirical distribution function \( F_n \), in that \( F_n(y_0) \) is simply a binomial proportion. When
joint inference is required at several positions one needs to exploit the structure of $F_n - F$ more fully, associated, for example, with weak convergence of the $\sqrt{n}(F_n - F)$ process to a time-transformed Brownian bridge, say $W^\theta(F(t))$; see Billingsley (1968) and our treatment in Section 9.6.

Nonparametric inference for quantiles is less well explored, though the basic large-sample results related to (11.4) are essentially well known; cf. David and Nagaraja (2003), also for various relevant consequences when it comes to linear or smooth functions of order statistics, and so forth; cf. Stigler (1974). For nonparametric quantile inference from the Bayesian perspective, see Hjort and Petrone (2007) and Hjort and Walker (2009).

Empirical likelihood has gathered a large literature, starting with Owen (1990, 1991), with Owen (2001) being an important source also for other material. Hjort et al. (2009) extend the scope of the empirical likelihood in several directions, including that of allowing a growing number of interest parameters with sample size.

We noted in Section 11.1 that there are no particular conceptual difficulties lifting the apparatus of confidence distributions to non- and semiparametric models, as long as the focus parameter $\psi$ is such that there is an estimator $\hat{\psi}$ with an appropriately relevant large-sample property, say $n^{1/2}(\psi - \hat{\psi})/\hat{\kappa} \to_d Z$ for a suitable rate $\alpha$. The prototypical case remains that of $\alpha = \frac{1}{12}$ and a normal limit, but in big models and with complex information there are other rates. Nevertheless, confidence distributions along the lines of $R(n^{1/2}(\psi - \hat{\psi})/\hat{\kappa})$ can then be constructed, with $R(z)$ the cumulative distribution function for $Z$, perhaps with further amendments as for the parametric cases surveyed in Chapter 7. Various problems involving kernel type estimators for smooth functions are associated with the rate $n^{2/5}$. Instances in which the cube root rate $n^{1/3}$ applies include those treated in Kim and Pollard (1990).

Traditional regression models for hazard rates, for example, Cox’s proportional hazards construction, are semiparametric in nature, with separate estimation strategies for the parametric and the nonparametric parts. There are $\sqrt{n}$ rate results about these; see, for example, Andersen et al. (1993) and Aalen et al. (2008), leading as per the recipe given earlier also to confidence distributions for regression parameters and so forth. For the $\beta$ regression coefficient vector of the Cox model inference typically uses the so-called partial likelihood, for which also log-profile results are available and where confidence distributions may be put up following the deviance type recipes of Chapters 3, 4 and 8.

Exercises

11.1 Quantile confidence: In Section 11.3 we gave nonparametric confidence distribution methods for quantiles, via the formula (11.3) that makes it possible to assess the confidence level for any $[Y_{i0}, Y_{j0}]$. In some of our illustrations we then used that method to display all confidence intervals of the type $[Y_{i-a}, Y_{j+a}]$, for given $j$ and varying $a$. Now construct an alternative version of Figures 11.4 and 11.6, which instead uses the cc$(\mu) = |1 - 2C(\mu)|$ construction for $\mu = F^{-1}(p)$, with

$$C(Y_{i0}) = P(\mu \leq Y_{i0}) = P(p \leq U_{i0}) = 1 - Be(p,i,n-i+1).$$
Confidence in non- and semiparametric models

11.2 Confidence for skewness: For a distribution with finite third moment, its skewness is \( \gamma = E((Y-\xi)/\sigma)^3 \), where \( Y \) denotes a variable having the distribution in question, with mean \( \xi \) and standard deviation \( \sigma \). The natural empirical estimate, based on a sample \( Y_1, \ldots, Y_n \), is \( \hat{\gamma} = n^{-1} \sum_{i=1}^{n} w_i^3 \), with \( w_i = (Y_i - \xi)/\hat{\sigma} \), featuring the usual data mean and standard deviation.

(a) Explain why \( \sqrt{n}(\hat{\gamma} - \gamma) \) has a limit distribution of the type \( N(0, \kappa^2) \), provided also higher moments up to order six are finite, and attempt to find a formula for \( \kappa \). Deduce that \( C_n(\gamma) = \Phi(\gamma - \hat{\gamma}/\hat{\kappa}) \) is an asymptotically correct confidence distribution for \( \gamma \), where \( \hat{\kappa} \) is any consistent estimator for \( \kappa \).

(b) Finding an explicit estimator for \( \kappa \) is somewhat cumbersome, however, and easier practical alternatives include those of bootstrapping and jackknifing. The latter uses

\[
\hat{\kappa}^2 = \frac{n+1}{n} \sum_{i=1}^{n} (\hat{\gamma}(i)-\hat{\gamma})^2,
\]

where \( \hat{\gamma}(i) \) is the empirical skewness computed with data point \( y_i \) pushed out of the dataset, and \( \hat{\gamma} = n^{-1} \sum_{i=1}^{n} \hat{\gamma}(i) \). For the data with ancient Egyptian lifelengths, for 82 mean and 59 women (cf. Example 3.7), compute the skewnesses for the men and the women (these are found to be respectively 0.323 and 1.445). Use jackknifing to find their standard errors and then display their associated normal-approximated confidence distributions.

(c) Use t-bootstrap, as in Section 7.6, with scale estimate for \( \hat{\gamma} \) taken to be that afforded by the jackknife method given earlier, to provide more accurate nonparametric confidence distributions for the two skewnesses. Comment on the degree to which the simpler normal approximations succeed in coming close to the more accurate t-bootstrap results.

11.3 Empirical likelihood: Show that \( \prod_{i=1}^{n} w_i \), with \( w_i > 0 \) and \( \sum_{i=1}^{n} w_i = 1 \), is maximised for \( w_i = 1/n \). Set up equations and a computational scheme for maximising \( \prod_{i=1}^{n} w_i \) under the side conditions that \( \sum_{i=1}^{n} w_i = 1 \) and \( \sum_{i=1}^{n} w_i (Y_i - \xi) = 0 \), for given dataset \( y_1, \ldots, y_n \). Use this to compute the log-empirical-likelihood function for the mean of the distribution of ancient Egyptian lifelengths, for men and for women. Invert these to display empirical likelihood based confidence distributions.

11.4 Quadratic forms: Suppose \( U \sim N_p(0, \Sigma) \) and consider the quadratic form \( \Lambda = U^t \Sigma^{-1} U \), with \( \Sigma \) a full rank variance matrix. This is the limit distribution of the \( -2\log EL_n(\theta_0) \) statistic, under appropriate conditions. In given situations its distribution may be easily simulated, with consistent estimates for \( \Omega \) and \( \Sigma \), but it is also of interest to learn more about its precise distribution.

(a) Write \( U = \Omega^{1/2} Z \) with \( Z \sim N_p(0, I_p) \) to show that \( \Lambda = Z^t K Z \) with \( K = \Omega^{1/2} \Sigma^{-1} \Omega^{1/2} \). Then use the spectral theorem of linear algebra to pick an orthonormal matrix \( P \) with the property that \( PKP^t = D \), with \( D \) a diagonal matrix with diagonal elements \( d_1, \ldots, d_p \), being the eigenvalues of \( \Omega \Sigma^{-1} \).

(b) Use this to demonstrate that \( \Lambda = \sum_{j=1}^{p} d_i N_j^2 \), with the \( N_j \) being independent and standard normal.

(c) Show that the characteristic function of \( \Lambda \) can be expressed as \( \phi(t) = \prod_{j=1}^{p} (1 - 2it^j)^{-1/2} \). Devise an algorithm that suitably inverts this characteristic function to compute the density \( g(\lambda) \) of \( \Lambda \).

(d) Generalise the preceding to noncentral case, where \( U \sim N_p(\xi, \Omega) \).

11.5 EL versus ML: Choose a reasonable dataset on which to try out the traditional normal model \( N(\xi, \sigma^2) \), and estimate the two parameters using maximum likelihood.
Exercises 335

(a) Put up confidence distributions for $\xi$ and $\sigma$, based on normal-model theory.

(b) Then complement the aforementioned curves with confidence distributions based on the empirical likelihood. Comment on any differences, in view of Remark 11.2.

11.6 Wilcoxon and Hodges–Lehmann: These are some details regarding the Hodges–Lehmann and Wilcoxon methods applied to the women’s BMI; cf. Example 11.4. To see whether the asymmetry in the distribution makes the Wilcoxon confidence curve invalid we have conducted a small bootstrap exercise. For say 1000 bootstrap replicates of the data ($n = 741$ observations drawn with replacement from the sample), compute the p-value when testing $H_0: \mu = \hat{\mu}$ by the two-sided Wilcoxon test. Display these bootstrapped p-values, and discuss whether they can be said to follow the uniform distribution. Explain why departure from uniformity indicates invalidity of the Wilcoxon confidence curve.