CHAPTER 7

Likelihood functions

7.1 Introduction

Let Y be a random variable whose density or distribution function $f_Y(y;\theta)$ depends on the p-dimensional parameter vector θ. Usually, we think of Y as vector valued with n independent components, but this consideration is important only in large sample approximations. For the most part, since it is unnecessary to refer to the individual components, we write Y without indices. A realization of Y is called an observation and, when we wish to make a clear distinction between the observation and the random variable, we write y for the observation. Of course, y is just a number or ordered set of numbers but, implicit in Y is the sample space or set of possible observations, one of which is y. This distinction is made at the outset because of its importance in the remainder of this chapter.

The parameter vector θ with components $\theta^1, \ldots, \theta^p$ is assumed to lie in some subset, Θ, of R^p. Often, in fact, $\Theta = R^p$, but this assumption is not necessary in the discussion that follows. For technical reasons, it helps to assume that Θ is an open set in R^p: this condition ensures, for example, that there are no equality constraints among the components and that the parameter space is genuinely p-dimensional.

Associated with any observed value y on Y, there is a particular parameter value $\theta_T \in \Theta$, usually unknown, such that Y has density function $f_Y(y;\theta_T)$. We refer to θ_T as the 'true' value. Often, however, when we wish to test a null hypothesis value, we write θ_0 and in subsequent calculations, θ_0 is treated as if it were the true value. Ideally, we would like to know the value of θ_T, but apart from exceptional cases, the observed data do not determine θ_T uniquely or precisely. Inference, then, is concerned with probability statements concerning those values in Θ that are consistent with the observed y. Usually, it is both unreasonable and undesirable to quote a single 'most consistent' parameter value: interval estimates, either in the form of confidence sets or Bayes intervals, are preferred.

Since the eventual goal is to make probabilistic statements concerning those parameter values that are consistent in some sense with the observed data, the conclusions must be unaffected by two kinds of transformation:

(i) invertible transformation of Y
(ii) invertible transformation of θ.

Invariance under the first of these groups is guaranteed if we work with the log likelihood function, defined up to an arbitrary additive function of y. Invariance under re-parameterization is a main concern of this chapter. For that reason, we are interested in quantities that transform as tensors under change of coordinates on Θ. Thus Θ, and not the sample space, is here regarded as the space of primary interest. By convention, therefore, we use superscripts to represent the coordinates of an arbitrary point in Θ. In this respect, the notation differs from that used in the previous chapter, where transformations of the sample space were considered.

In almost all schools of statistical inference, the log likelihood function for the observed data

$$l(\theta; y) = \log f_Y(y;\theta)$$

plays a key role. One extreme view, (Edwards, 1972), is that nothing else matters. In the Bayesian framework on the other hand, it is necessary to acquire a prior distribution, $\pi(\theta)$ that describes

'degree of belief' or personal conviction prior to making the observation. Bayes's theorem then gives

$$\pi(\theta|y) = \pi(\theta) f_Y(y;\theta)/c(y)$$

as the posterior distribution for θ given y, where $c(y)$ is the normalization factor, $\int \pi(\theta) f_Y(y;\theta) d\theta$. All probability statements are then based on the posterior distribution of θ given y. Other schools of inference hold that, in order to conceive of probabilities as relative frequencies rather than as degrees of belief, it is necessary to take account of the sample space of possible observations. In other words, $l(\theta; y)$ must be regarded as the observed value of the random variable $l(\theta; Y)$. One difficulty with this viewpoint is that there is often some leeway in the choice of sample space.

A thorough discussion of the various schools of statistical inference is beyond the scope of this book, but can be found, for example, in Cox & Hinkley (1974) or Berger & Wolpert (1984). In the discussion that follows, our choice is to regard $l(\theta; Y)$ and its derivatives with respect to θ as random variables. The sample space is rarely mentioned explicitly, but it is implicit when we talk of moments or cumulants, which involve integration over the sample space.

Tensor methods are particularly appropriate and powerful in this context because of the requirement that any inferential statement should be materially unaffected by the parameterization chosen. The parameterization is simply a convenient but arbitrary way of specifying the various probability models under consideration. An inferential statement identifies a subset of these distributions and that subset should, in principle at least, be unaffected by the particular parameterization chosen. Unless otherwise stated, therefore, when we talk of tensors in this chapter, we refer implicitly to arbitrary invertible transformations of the parameter vector. Particular emphasis is placed on invariants, which may be used to make inferential statements independent of the coordinate system. The most important invariant is the log likelihood function itself. Other invariants are connected with the likelihood ratio statistic and its distribution.

7.2 Log likelihood derivatives

7.2.1 *Null cumulants*

In what follows, it is assumed that the log likelihood function has continuous partial derivatives up to the required order and that these derivatives have finite moments, again up to the required order, which is obvious from the context. These derivatives at an arbitrary point θ, are written as

$$\begin{aligned} U_r &= u_r(\theta; Y) = \partial l(\theta; Y)/\partial\theta^r \\ U_{rs} &= u_{rs}(\theta; Y) = \partial^2 l(\theta; Y)/\partial\theta^r\partial\theta^s \\ U_{rst} &= u_{rst}(\theta; Y) = \partial^3 l(\theta; Y)/\partial\theta^r\partial\theta^s\partial\theta^t \end{aligned}$$

and so on. Our use of subscripts here is not intended to imply that the log likelihood derivatives are tensors. In fact, the derivatives with respect to an alternative parameterization $\phi = \phi^1, \ldots, \phi^p$, are given by

$$\begin{aligned} \bar{U}_r &= \theta^i_r U_i \\ \bar{U}_{rs} &= \theta^i_r \theta^j_s U_{ij} + \theta^i_{rs} U_i \\ \bar{U}_{rst} &= \theta^i_r \theta^j_s \theta^k_t U_{ijk} + \theta^i_r \theta^j_{st} U_{ij}[3] + \theta^i_{rst} U_i \end{aligned} \tag{7.1}$$

and so on, where $\theta^i_r = \partial\theta^i/\partial\phi^r$ is assumed to have full rank, $\theta^i_{rs} = \partial^2\theta^i/\partial\phi^r\partial\phi^s$ and so on. Thus U_r is a tensor but subsequent higher-order derivatives are not, on account of the higher derivatives that appear in the transformation formulae. The log likelihood derivatives are tensors under the smaller group of linear or affine transformations, but this is of no substantial importance in the present context.

For reasons that will become clear shortly, it is desirable to depart to some extent from the notation used in Chapters 2 and 3 for moments and cumulants. The null moments of U_r, U_{rs}, $U_{rst}, \ldots$ are written as

$$\begin{aligned} \mu_r &= E(U_r;\theta), & \mu_{r,s} &= E(U_r U_s;\theta), \\ \mu_{rs} &= E(U_{rs};\theta), & \mu_{r,st} &= E(U_r U_{st};\theta), \\ \mu_{rst} &= E(U_{rst};\theta), & \mu_{r,st,uvw} &= E(U_r U_{st} U_{uvw};\theta) \end{aligned}$$

and so on. The word 'null' here refers to the fact that the twin processes of differentiation and averaging both take place at the same value of θ. The null cumulants are defined by

$$\kappa_r = \mu_r, \qquad \kappa_{r,s} = \mu_{r,s} - \mu_r\mu_s \qquad \kappa_{rs,tu} = \mu_{rs,tu} - \mu_{rs}\mu_{tu}$$

and so on.

Neither the set of moments nor the set of cumulants is linearly independent. To see how the linear dependencies arise, we note that for all θ, integration over the sample space gives

$$\int f_Y(y;\theta)dy = 1.$$

Differentiation with respect to θ and reversing the order of differentiation and integration gives

$$\mu_r = \kappa_r = \int u_r(\theta;y) f_Y(y;\theta)dy = 0.$$

Further differentiation gives

$$\begin{aligned} \mu_{[rs]} &= \mu_{rs} + \mu_{r,s} = 0 \\ \mu_{[rst]} &= \mu_{rst} + \mu_{r,st}[3] + \mu_{r,s,t} = 0 \\ \mu_{[rstu]} &= \mu_{rstu} + \mu_{r,stu}[4] + \mu_{rs,tu}[3] + \mu_{r,s,tu}[6] + \mu_{r,s,t,u} = 0. \end{aligned}$$

In terms of the null cumulants, we have

$$\begin{aligned} \kappa_{[rs]} &= \kappa_{rs} + \kappa_{r,s} = 0 \\ \kappa_{[rst]} &= \kappa_{rst} + \kappa_{r,st}[3] + \kappa_{r,s,t} = 0 \\ \kappa_{[rstu]} &= \kappa_{rstu} + \kappa_{r,stu}[4] + \kappa_{rs,tu}[3] + \kappa_{r,s,tu}[6] + \kappa_{r,s,t,u} = 0, \end{aligned} \tag{7.2}$$

and so on, with summation over all partitions of the indices. In the remainder of this chapter, the enclosure within square brackets of a set of indices implies summation over all partitions of that set, as in the expressions listed above. In addition, $\kappa_{r,[st]}$ is synonomous with the combination $\kappa_{r,s,t} + \kappa_{r,st}$, the rule in this case applying to a subset of the indices. Details of the argument leading to (7.2) are given in Exercise 7.1. In particular, to reverse the order of differentiation and integration, it is necessary to assume that the sample space does not depend on θ.

In the univariate case, power notation is often employed in the form

$$i_{rst} = E\left\{\left(\frac{\partial l}{\partial\theta}\right)^r \left(\frac{\partial^2 l}{\partial\theta^2}\right)^s \left(\frac{\partial^3 l}{\partial\theta^3}\right)^t ;\theta\right\}.$$

The moment identities then become $i_{10} = 0$,

$$\begin{aligned} &i_{01} + i_{20} = 0, \\ &i_{001} + 3i_{11} + i_{30} = 0, \\ &i_{0001} + 4i_{101} + 3i_{02} + 6i_{21} + i_{40} = 0. \end{aligned}$$

Similar identities apply to the cumulants, but we refrain from writing these down, in order to avoid further conflict of notation.

7.2.2 *Non-null cumulants*

Given a test statistic based on the log likelihood derivatives at the hypothesized value, only the null distribution, or the null cumulants, are required in order to compute the significance level. However, in order to assess the suitability of a proposed test statistic, it is necessary to examine the sensitivity of the statistic to changes in the parameter value. Suppose then that U_r, $U_{rs}, \ldots$ are the log likelihood derivatives at an arbitrary point θ and that the 'true' parameter point is θ_T. We may then examine how the cumulants of U_r, $U_{rs}, \ldots$ depend on the value of $\delta = \theta_T - \theta$. Thus, we write in an obvious notation,

$$\mu_r(\theta;\theta_T) = E\{U_r;\theta_T\} = \int \frac{\partial \log f_Y(y;\theta)}{\partial\theta} f_Y(y;\theta_T)dy$$

and similarly for $\mu_{r,s}(\theta;\theta_T)$, $\mu_{rs}(\theta;\theta_T)$ and so on. The null values are $\mu_r(\theta;\theta) = \mu_r$, $\mu_{r,s}(\theta;\theta) = \mu_{r,s}$ and so on. The non-null cumulants are written $\kappa_r(\theta;\theta_T)$, $\kappa_{rs}(\theta;\theta_T)$, $\kappa_{r,s}(\theta;\theta_T)$ and so on, where, for example,

$$\begin{aligned}\kappa_r(\theta;\theta_T) &= \mu_r(\theta;\theta_T)\\ \kappa_{r,s}(\theta;\theta_T) &= \mu_{r,s}(\theta;\theta_T) - \mu_r(\theta;\theta_T)\mu_s(\theta;\theta_T).\end{aligned}$$

For small values of δ, it is possible to express the non-null cumulants as a power series in δ, with coefficients that involve the null cumulants alone. In fact, from the Taylor expansion

$$\begin{aligned}\frac{f_Y(Y;\theta_T)}{f_Y(Y;\theta)} &= 1 + U_r\delta^r + (U_{rs} + U_rU_s)\delta^r\delta^s/2!\\ &\quad + (U_{rst} + U_rU_{st}[3] + U_rU_sU_t)\delta^r\delta^s\delta^t/3! + \cdots,\end{aligned}$$

we find the following expansions.

$$\begin{aligned}\mu_r(\theta;\theta_T) &= \mu_r + \mu_{r,s}\delta^s + \mu_{r,[st]}\delta^s\delta^t/2! + \mu_{r,[stu]}\delta^s\delta^t\delta^u/3! + \cdots\\ \mu_{r,s}(\theta;\theta_T) &= \mu_{r,s} + \mu_{r,s,t}\delta^t + \mu_{r,s,[tu]}\delta^t\delta^u/2!\\ &\quad + \mu_{r,s,[tuv]}\delta^t\delta^u\delta^v/3! + \cdots\\ \mu_{r,st}(\theta;\theta_T) &= \mu_{r,st} + \mu_{r,st,u}\delta^u + \mu_{r,st,[uv]}\delta^u\delta^v/2! + \cdots\end{aligned} \tag{7.3}$$

where, for example,

$$\mu_{r,[stu]} = \mu_{r,stu} + \mu_{r,s,tu}[3] + \mu_{r,s,t,u}$$

involves summation over all partitions of the bracketed indices.

Identical expansions hold for the cumulants. For example,

$$\begin{aligned}\kappa_{r,s}(\theta;\theta_T) &= \kappa_{r,s} + \kappa_{r,s,t}\delta^t + \kappa_{r,s,[tu]}\delta^t\delta^u/2!\\ &\quad + \kappa_{r,s,[tuv]}\delta^t\delta^u\delta^v/3! + \cdots\end{aligned} \tag{7.4}$$

where

$$\kappa_{r,s,[tuv]} = \kappa_{r,s,tuv} + \kappa_{r,s,t,uv}[3] + \kappa_{r,s,t,u,v}.$$

Note that $\kappa_{r,s,[t]}$, $\kappa_{r,s,[tu]}, \ldots$ are the vector of first derivatives and the array of second derivatives with respect to the second argument only, of the null cumulant, $\kappa_{r,s}(\theta;\theta)$. There is therefore, a close formal similarity between these expansions and those derived by Skovgaard (1986a) for the derivatives of $\kappa_{r,s}(\theta) = \kappa_{r,s}(\theta;\theta)$ and similar null cumulants. Skovgaard's derivatives involve additional terms that arise from considering variations in the two arguments simultaneously. The derivation of (7.4) can be accomplished along the lines of Section 3 of Skovgaard's paper.

7.2.3 *Tensor derivatives*

One peculiar aspect of the identities (7.2) and also of expansions (7.3) and (7.4) is that, although the individual terms, in general, are not tensors, nevertheless the identities and expansions are valid in all coordinate systems. Consider, for example, the identity $\kappa_{rs} + \kappa_{r,s} = 0$ in (7.2). Now, U_r is a tensor and hence all its cumulants are tensors. Thus $\kappa_{r,s}$, $\kappa_{r,s,t}$ and so on are tensors and hence $\kappa_{rs} = -\kappa_{r,s}$ must also be a tensor even though U_{rs} is not a tensor. This claim is easily verified directly: see Exercise 7.2. Similarly, from the identity $\kappa_{rst} + \kappa_{r,st}[3] + \kappa_{r,s,t} = 0$, it follows that $\kappa_{rst} + \kappa_{r,st}[3]$ must be a tensor. However, neither κ_{rst} nor $\kappa_{r,st}$ are tensors. In fact, the transformation laws are

$$\begin{aligned}\bar\kappa_{r,st} &= \theta^i_r\theta^j_s\theta^k_t\kappa_{i,jk} + \theta^i_r\theta^j_{st}\kappa_{i,j}\\ \bar\kappa_{rst} &= \theta^i_r\theta^j_s\theta^k_t\kappa_{ijk} + \theta^i_r\theta^j_{st}\kappa_{ij}[3].\end{aligned}$$

From these, it can be seen that $\kappa_{rst} + \kappa_{r,st}[3]$ is indeed a tensor.

The principal objection to working with arrays that are not tensors is that it is difficult to recognize and to construct invariants. For example, the log likelihood ratio statistic is invariant and, when we expand it in terms of log likelihood derivatives, it is helpful if the individual terms in the expansion are themselves invariants. For that reason, we seek to construct arrays V_r, V_{rs}, $V_{rst}, \ldots$ related to the log likelihood derivatives, such that the Vs are tensors. In addition, in order to make use of Taylor expansions such as (7.3) and (7.4), we require that the Vs be ordinary log likelihood derivatives in *some* coordinate system. This criterion excludes covariant derivatives as normally defined in differential geometry: it also excludes least squares residual derivatives, whose cumulants do not obey (7.2). See Exercise 7.3.

Let θ_0 be an arbitrary parameter value and define

$$\beta^r_{st} = \kappa^{r,i}\kappa_{i,st}, \quad \beta^r_{stu} = \kappa^{r,i}\kappa_{i,stu}, \quad \beta^r_{stuv} = \kappa^{r,i}\kappa_{i,stuv},$$

where $\kappa^{r,s}$ is the matrix inverse of $\kappa_{r,s}$. In these definitions, only null cumulants at θ_0 are involved. This means that the β-arrays can be computed at each point in Θ without knowing the 'true' value, θ_T. The βs are the regression coefficients of U_{st}, $U_{stu}, \ldots$ on U_r where the process of averaging is carried out under θ_0, the same point at which derivatives were computed. Now consider the parameter transformation defined in a neighbourhood of θ_0 by

$$\begin{aligned}\phi^r - \phi^r_0 = \theta^r - \theta^r_0 &+ \beta^r_{st}(\theta^s - \theta^s_0)(\theta^t - \theta^t_0)/2!\\ &+ \beta^r_{stu}(\theta^s - \theta^s_0)(\theta^t - \theta^t_0)(\theta^u - \theta^u_0)/3! + \cdots\end{aligned} \tag{7.5}$$

Evidently, θ_0 transforms to ϕ_0. From (7.1), the derivatives at ϕ_0 with respect to ϕ, here denoted by V_i, V_{ij}, $V_{ijk}, \ldots$, satisfy

$$\begin{aligned}U_r &= V_r\\ U_{rs} &= V_{rs} + \beta^i_{rs}V_i\\ U_{rst} &= V_{rst} + \beta^i_{rs}V_{it}[3] + \beta^i_{rst}V_i\\ U_{rstu} &= V_{rstu} + \beta^i_{rs}V_{itu}[6] + \beta^i_{rs}\beta^j_{tu}V_{ij}[3] + \beta^i_{rst}V_{iu}[4] + \beta^i_{rstu}V_i\end{aligned} \tag{7.6}$$

with summation over all partitions of the free indices. Evidently, the Vs are genuine log likelihood derivatives having the property that the null covariances

$$\nu_{r,st} = \mathrm{cov}(V_r, V_{st}), \quad \nu_{r,stu} = \mathrm{cov}(V_r, V_{stu}), \ldots$$

of V_r with the higher-order derivatives, are all zero. Note, however, that

$$\begin{aligned}\nu_{rs,tu} &= \mathrm{cov}(V_{rs}, V_{tu}) = \kappa_{rs,tu} - \kappa_{rs,i}\kappa_{tu,j}\kappa^{i,j}\\ \nu_{r,s,tu} &= \mathrm{cum}(V_r, V_s, V_{tu}) = \kappa_{r,s,tu} - \kappa_{r,s,i}\kappa_{tu,j}\kappa^{i,j}\end{aligned}$$

are non-zero in general. It follows that identities (7.2) apply to the Vs in the form

$$\begin{aligned}&\nu_{rs}+\nu_{r,s}=0,\\&\nu_{rst}+\nu_{r,s,t}=0,\\&\nu_{rstu}+\nu_{rs,tu}[3]+\nu_{r,s,tu}[6]+\nu_{r,s,t,u}=0.\end{aligned}$$

It remains to show that the Vs are tensors. To do so, we need to examine the effect on the Vs of applying a non-linear transformation to θ. One method of proof proceeds as follows. Take $\theta_0=\phi_0=0$ for simplicity and consider the transformation

$$\bar{\theta}^r=a^r_i\theta^i+a^r_{ij}\theta^i\theta^j/2!+a^r_{ijk}\theta^i\theta^j\theta^k/3!+\cdots$$

so that the log likelihood derivatives transform to $\bar{U}_r$, $\bar{U}_{rs}$, $\bar{U}_{rst}$ satisfying

$$\begin{aligned}U_r&=a^i_r\bar{U}_i\\U_{rs}&=a^i_ra^j_s\bar{U}_{ij}+a^i_{rs}\bar{U}_i\\U_{rst}&=a^i_ra^j_sa^k_t\bar{U}_{ijk}+a^i_ra^j_{st}\bar{U}_{ij}[3]+a^i_{rst}\bar{U}_i\end{aligned}\tag{7.7}$$

and so on. These are the inverse equations to (7.1). Proceeding quite formally now, equations (7.6) may be written using matrix notation in the form

$$\mathbf{U}=\mathbf{BV}.\tag{7.8}$$

Similarly, the relation between U and $\bar{U}$ in (7.8) may be written

$$\mathbf{U}=\mathbf{A}\bar{\mathbf{U}}.$$

These matrix equations encompass all of the derivatives up to whatever order is specified. In the particular case where $a^r_i=\delta^r_i$, but not otherwise, the array of coefficients $\mathbf{B}$ transforms to $\bar{\mathbf{B}}$, where

$$\mathbf{B}=\mathbf{A}\bar{\mathbf{B}},$$

as can be seen by examining equations (7.7) above. More generally, if $a^i_r\neq\delta^i_r$, we may write $\mathbf{B}=\mathbf{A}\bar{\mathbf{B}}\mathbf{A}^{*-1}$, where $\mathbf{A}^*$ is a direct product matrix involving a^i_r alone (Exercise 7.8). Premultiplication of (7.8) by $\mathbf{A}^{-1}$ gives

$$\bar{\mathbf{U}}=\mathbf{A}^{-1}\mathbf{U}=\mathbf{A}^{-1}\mathbf{BV}=\bar{\mathbf{B}}\mathbf{V}.$$

Since, by definition, the transformed Vs satisfy $\bar{\mathbf{U}}=\bar{\mathbf{B}}\bar{\mathbf{V}}$, it follows that if $a^r_i=\delta^r_i$, then $\bar{\mathbf{V}}=\mathbf{V}$. Thus, the Vs are unaffected by non-linear transformation in which the leading coefficient is δ^r_i. Since all quantities involved are tensors under the general linear group, it follows that the Vs must be tensors under arbitrary smooth invertible parameter transformation.

The Vs defined by (7.6) are in no sense unique. In fact any sequence of coefficients τ^i_{rs}, $\tau^i_{rst},\ldots$ that transforms like the βs will generate a sequence of symmetric tensors when inserted into (7.6). One possibility is to define

$$\tau^r_{st}=\kappa^{r,i}\kappa_{i,[st]},\qquad\tau^r_{stu}=\kappa^{r,i}\kappa_{i,[stu]}$$

and so on. These arrays transform in the same way as the βs: any linear combination of the two has the same property. In fact, the Vs defined by (7.6) can be thought of as derivatives in the 'canonical' coordinate system: the corresponding tensors obtained by replacing β by τ are derivatives in the 'mean-value' coordinate system. This terminology is taken from the theory of exponential family models.

Barndorff-Nielsen (1986) refers to the process of obtaining tensors via (7.6) as the *intertwining* of *strings*, although *unravelling of strings* might be a better description of the process. In this terminology, the sequence of coefficients β^r_{st}, $\beta^r_{stu},\ldots$ or the alternative sequence τ^r_{st}, $\tau^r_{stu},\ldots$ is known as a *connection* string. The log likelihood derivatives themselves, and any sequence that transforms like (7.1) under re-parameterization, forms an infinite *co-string*. *Contra-strings* are defined by analogy. The main advantage of this perspective is that it forces one to think of the *sequence* of derivatives as an indivisible object: the higher-order derivatives have no invariant interpretation in the absence of the lower-order derivatives. See Foster (1986) for a light-hearted but enlightened discussion on this point.

7.3 Large sample approximation

7.3.1 *Log likelihood derivatives*

Suppose now that Y has n independent and identically distributed components so that the log likelihood for the full data may be written as the sum

$$l(\theta;Y)=\sum_i l(\theta;Y_i).$$

The derivatives U_r, $U_{rs},\ldots$ are then expressible as sums of n independent and identically distributed random variables. Under mild regularity conditions, therefore, the joint distribution of U_r, $U_{rs},\ldots$ may be approximated for large n by the normal distribution, augmented, if necessary by Edgeworth corrections.

It is convenient in the calculations that follow to make the dependence on n explicit by writing

$$\begin{aligned} U_r &= n^{1/2}Z_r \\ U_{rs} &= n\kappa_{rs}+n^{1/2}Z_{rs} \\ U_{rst} &= n\kappa_{rst}+n^{1/2}Z_{rst} \end{aligned} \tag{7.9}$$

and so on for the higher-order derivatives. Thus,

$$\kappa_{r,s}=-\kappa_{rs}=-E\{\partial^2 l(\theta;Y_i)/\partial\theta^r\partial\theta^s;\theta\}$$

is the Fisher information per observation and κ_{rst}, $\kappa_{rstu},\ldots$ are higher-order information measures per observation. Moreover, assuming θ to be the true value, it follows that Z_r, Z_{rs}, $Z_{rst},\ldots$ are $O_p(1)$ for large n.

More generally, if, as would normally be the case, the components of Y are not identically distributed but still independent, $\kappa_{r,s}$ is the average Fisher information per observation and κ_{rst}, $\kappa_{rstu},\ldots$ are higher-order average information measures per observation. Additional fairly mild assumptions, in the spirit of the Lindeberg-Feller condition, are required to ensure that $\kappa_{r,s}=O(1)$, $Z_r=O_p(1)$, $Z_{rs}=O_p(1)$ and so on. Such conditions are taken for granted in the expansions that follow.

7.3.2 *Maximum likelihood estimation*

The likelihood equations $u_r(\hat\theta;Y)=0$ may be expanded in a Taylor series in $\hat\delta=n^{1/2}(\hat\theta-\theta)$ to give

$$\begin{aligned} 0=n^{1/2}Z_r&+(n\kappa_{rs}+n^{1/2}Z_{rs})\hat\delta^s/n^{1/2}+(n\kappa_{rst}+n^{1/2}Z_{rst})\hat\delta^s\hat\delta^t/(2n)\\ &+n\kappa_{rstu}\hat\delta^s\hat\delta^t\hat\delta^u/(6n^{3/2})+O_p(n^{-1}). \end{aligned}$$

For future convenience, we write

$$\kappa^{rs}=\kappa^{r,i}\kappa^{s,j}\kappa_{ij},\quad \kappa^{rst}=\kappa^{r,i}\kappa^{s,j}\kappa^{t,k}\kappa_{ijk},$$

and so on, using the tensor $\kappa_{i,j}$ and its matrix inverse $\kappa^{i,j}$ to lower and raise indices. We may now solve for $\hat\delta$ in terms of the Zs, whose joint cumulants are known, giving

$$\begin{aligned} \hat\delta^r=&\kappa^{r,s}Z_s+n^{-1/2}(\kappa^{r,s}\kappa^{t,u}Z_{st}Z_u+\kappa^{rst}Z_sZ_t/2)\\ &+n^{-1}\big(\kappa^{r,s}\kappa^{t,u}\kappa^{v,w}Z_{st}Z_{uv}Z_w+\kappa^{rst}\kappa^{u,v}Z_sZ_{tu}Z_v\\ &+\kappa^{r,s}\kappa^{tuv}Z_{st}Z_uZ_v/2+\kappa^{rst}\kappa^{uvw}\kappa_{t,w}Z_sZ_uZ_w/2\\ &+\kappa^{r,s}\kappa^{t,u}\kappa^{v,w}Z_{suw}Z_tZ_v/2+\kappa^{rstu}Z_sZ_tZ_u/6\big)+O_p(n^{-3/2}). \end{aligned} \tag{7.10}$$

Terms have been grouped here in powers of $n^{1/2}$. It is worth pointing out at this stage that $\hat{\delta}^r$ is not a tensor. Hence the expression on the right of the above equation is also not a tensor. However, the first-order approximation, namely $\kappa^{r,s}Z_s$, is a tensor.

From the above equation, or at least from the first two terms of the equation, some useful properties of maximum likelihood estimates may be derived. For example, we find

$$\begin{aligned} E(\hat{\delta}^r) &= n^{-1/2}(\kappa^{r,s}\kappa^{t,u}\kappa_{st,u} + \kappa^{rst}\kappa_{s,t}/2) + O(n^{-3/2}) \\ &= -n^{-1/2}\kappa^{r,s}\kappa^{t,u}(\kappa_{s,t,u} + \kappa_{s,tu})/2 + O(n^{-3/2}). \end{aligned}$$

This is $n^{1/2}$ times the bias of $\hat{\theta}^r$. In addition, straightforward calculations give

$$\begin{aligned} \operatorname{cov}(\hat{\delta}^r, \hat{\delta}^s) &= \kappa^{r,s} + O(n^{-1}) \\ \operatorname{cum}(\hat{\delta}^r, \hat{\delta}^s, \hat{\delta}^t) &= n^{-1/2}\kappa^{r,i}\kappa^{s,j}\kappa^{t,k}(\kappa_{ijk} - \kappa_{i,j,k}) + O(n^{-3/2}). \end{aligned}$$

Higher-order cumulants are $O(n^{-1})$ or smaller.

In the univariate case we may write

$$\begin{aligned} E(\hat{\theta} - \theta) &= -n^{-1}(i_{30} + i_{11})/(2i_{20}^2) + O(n^{-2}) \\ \operatorname{var}(\hat{\theta}) &= i_{20}^{-1}/n + O(n^{-2}) \\ \kappa_3(\hat{\theta}) &= n^{-2}(i_{001} - i_{30})/i_{20}^3 + O(n^{-3}), \end{aligned}$$

where i_{rst} is the generalized information measure per observation.

More extensive formulae of this type are given by Shenton & Bowman (1977, Chapter 3): their notation, particularly their version of the summation convention, differs from that used here. See also Peers & Iqbal (1985), who give the cumulants of the maximum likelihood estimate up to and including terms of order $O(n^{-1})$. In making comparisons, note that $\kappa^{ij} = -\kappa^{i,j}$ is used by Peers & Iqbal to raise indices.

7.4 Maximum likelihood ratio statistic

7.4.1 *Invariance properties*

In the previous section, the approximate distribution of the maximum likelihood estimate was derived through an asymptotic expansion in the log likelihood derivatives at the true parameter point. Neither $\hat{\theta}$ nor $\hat{\delta}$ are tensors and consequently the asymptotic expansion is not a tensor expansion. For that reason, the algebra tends to be a little complicated: it is not evident how the arrays involved should transform under a change of coordinates. In this section, we work with the maximized log likelihood ratio statistic defined by

$$W(\theta) = 2l(\hat{\theta}; Y) - 2l(\theta; Y)$$

where $l(\theta; Y)$ is the log likelihood for the full data comprising n independent observations. Since W is invariant under re-parameterization, it may be expressed in terms of other simpler invariants. The distributional calculations can be rendered tolerably simple if we express W as an asymptotic expansion involving invariants derived from the tensors V_r, V_{rs}, $V_{rst}, \ldots$ and their joint cumulants. The known joint cumulants of the Vs can then be used to determine the approximate distribution of W to any required order of approximation.

First, however, we derive the required expansion in arbitrary coordinates.

7.4.2 *Expansion in arbitrary coordinates*

Taylor expansion of $l(\hat{\theta};Y) - l(\theta;Y)$ about θ gives

$$\begin{aligned}\tfrac{1}{2}W(\theta) &= l(\hat{\theta};Y) - l(\theta;Y)\\ &= U_r\hat{\delta}^r/n^{1/2} + U_{rs}\hat{\delta}^r\hat{\delta}^s/(2n) + U_{rst}\hat{\delta}^r\hat{\delta}^s\hat{\delta}^t/(6n^{3/2}) + \cdots.\end{aligned}$$

Note that U_r, $U_{rs}, \ldots$ are the log likelihood derivatives at θ, here assumed to be the true parameter point. If we now write

$$\hat{\delta}^r = Z^r + c^r/n^{1/2} + d^r/n + O_p(n^{-3/2}),$$

where $Z^r = \kappa^{r,s}Z_s$, c^r and d^r are given by (7.10), we find that $\frac{1}{2}W(\theta)$ has the following expansion:

$$\begin{aligned}&n^{1/2}Z_r(Z^r + c^r/n^{1/2} + d^r/n + \cdots)/n^{1/2}\\ &+ (n\kappa_{rs} + n^{1/2}Z_{rs})\{Z^rZ^s + 2Z^rc^s/n^{1/2} + (c^rc^s + 2Z^rd^s)/n + \cdots\}/(2n)\\ &+ (n\kappa_{rst} + n^{1/2}Z_{rst})(Z^rZ^sZ^t + 3Z^rZ^sc^t/n^{1/2} + \cdots)/(6n^{3/2})\\ &+ (n\kappa_{rstu} + n^{1/2}Z_{rstu})(Z^rZ^sZ^tZ^u + \cdots)/(24n^2) + O_p(n^{-3/2}).\end{aligned}$$

This expansion includes all terms up to order $O_p(n^{-1})$ in the null case and involves quartic terms in the expansion of $l(\hat{\theta};Y)$. On collecting together terms that are of equal order in n, much cancellation occurs. For example, in the $O_p(n^{-1/2})$ term, the two expressions involving c^r cancel, and likewise for the two expressions involving d^r in the $O_p(n^{-1})$ term. For this reason, the expansion to order $O_p(n^{-1})$ of $W(\theta)$ does not involve d^r, and c^r occurs only in the $O_p(n^{-1})$ term.

Further simplification using (7.10) gives

$$\begin{aligned}\tfrac{1}{2}W(\theta) &= \tfrac{1}{2}Z_rZ_s\kappa^{r,s} + n^{-1/2}\{\kappa_{rst}Z^rZ^sZ^t/3! + Z_{rs}Z^rZ^s/2!\}\\ &\quad + n^{-1}\{(Z_{ri}Z^i + \tfrac{1}{2}\kappa_{rij}Z^iZ^j)\kappa^{r,s}(Z_{si}Z^i + \tfrac{1}{2}\kappa_{sij}Z^iZ^j)/2\\ &\quad + \kappa_{rstu}Z^rZ^sZ^tZ^u/4! + Z_{rst}Z^rZ^sZ^t/3!\}\end{aligned} \tag{7.11}$$

when terms that are $O_p(n^{-3/2})$ are ignored. Note that the $O_p(1)$, $O_p(n^{-1/2})$ and $O_p(n^{-1})$ terms are each invariant. This is not immediately obvious because Z_{rs}, Z_{rst}, κ_{rst} and κ_{rstu} are not tensors. In particular, the individual terms in the above expansion are not invariant.

7.4.3 *Invariant expansion*

The simplest way to obtain an invariant expansion is to use (7.11) in the coordinate system defined by (7.5). We simply replace all κs by νs and re-define the Zs to be

$$\begin{aligned}Z_r &= n^{-1/2}V_r\\ Z_{rs} &= n^{-1/2}(V_{rs} - n\nu_{rs})\\ Z_{rst} &= n^{-1/2}(V_{rst} - n\nu_{rst}).\end{aligned} \tag{7.12}$$

The leading term in the expansion for $W(\theta)$ is $Z_rZ_s\nu^{r,s}$, also known as the score statistic or the quadratic score statistic. The $O_p(n^{-1/2})$ term is

$$n^{-1/2}(Z_{rs}Z^rZ^s - \nu_{r,s,t}Z^rZ^sZ^t/3),$$

which involves the skewness tensor of the first derivatives as well as a 'curvature' correction involving the residual second derivative, Z_{rs}. Note that Z_{rs} is zero for full exponential family models in which the dimension of the sufficient statistic is the same as the dimension of the parameter. See Section 6.2.2, especially Equation (6.7).

7.4.4 *Bartlett factor*

From (7.11) and (7.12), the mean of $W(\theta)$ can be obtained in the form

$$\begin{aligned} p + n^{-1}\{&\nu_{rst}\nu^{r,s,t}/3 + \nu_{r,s,tu}\nu^{r,t}\nu^{s,u} + \nu_{rij}\nu_{skl}\nu^{r,s}\nu^{i,j}\nu^{k,l}/4 \\ &+ \nu_{rij}\nu_{skl}\nu^{r,s}\nu^{i,k}\nu^{j,l}/2 + \nu_{rs,tu}\nu^{r,t}\nu^{s,u} + \nu_{rstu}\nu^{r,s}\nu^{t,u}/4\} \\ &+ O(n^{-3/2}). \end{aligned}$$

Often the mean is written in the form

$$E(W(\theta);\theta) = p\{1 + b(\theta)/n + O(n^{-3/2})\}$$

where $b(\theta)$, known as the Bartlett correction factor, is given by

$$\begin{aligned} pb(\theta) = \rho_{13}^2/4 + \rho_{23}^2/6 - (\nu_{r,s,t,u} - \nu_{rs,tu}[3])\nu^{r,s}\nu^{t,u}/4 \\ - (\nu_{r,s,tu} + \nu_{rs,tu})\nu^{r,s}\nu^{t,u}/2. \end{aligned} \tag{7.13}$$

In deriving the above, we have made use of the identities

$$\begin{aligned} \nu_{rst} &= -\nu_{r,s,t} \\ \nu_{rstu} &= -\nu_{r,s,t,u} - \nu_{r,s,tu}[6] - \nu_{rs,tu}[3] \end{aligned}$$

derived in Sections 7.2.1 and 7.2.3, and also,

$$\begin{aligned} \rho_{13}^2 &= \nu_{i,j,k}\nu_{l,m,n}\nu^{i,j}\nu^{k,l}\nu^{m,n}, \\ \rho_{23}^2 &= \nu_{i,j,k}\nu_{l,m,n}\nu^{i,l}\nu^{j,m}\nu^{k,n}, \\ \rho_4 &= \nu_{i,j,k,l}\nu^{i,j}\nu^{k,l}, \end{aligned}$$

which are the invariant standardized cumulants of V_r.

The reason for the unusual grouping of terms in (7.13) is that, not only are the individual terms invariant under re-parameterization, but with this particular grouping they are nearly invariant under the operation of conditioning on ancillary statistics. For example, ρ_4 defined above is not invariant under conditioning. This point is examined further in the following chapter: it is an important point because the expression for $b(\theta)$ demonstrates that the conditional mean of $W(\theta)$ is independent of all ancillary statistics, at least to the present order of approximation. In fact, subsequent calculations in the following chapter show that $W(\theta)$ is statistically independent of *all* ancillary statistics to a high order of approximation.

Since Z_r is asymptotically normal with covariance matrix $\kappa_{r,s}$, it follows from (7.11) that the likelihood ratio statistic is asymptotically χ_p^2. This is a first-order approximation based on the leading term in (7.11). The error term in the distributional approximation appears to be $O(n^{-1/2})$, but as we shall see, it is actually $O(n^{-1})$. In fact, it will be shown that the distribution of

$$W' = W/\{1 + b(\theta)/n\}, \tag{7.14}$$

the Bartlett corrected statistic, is $\chi_p^2 + O(n^{-3/2})$. Thus, not only does the Bartlett factor correct the mean of W to this order of approximation, but it also corrects all of the higher-order cumulants of W to the same order of approximation. This is an unusual and surprising result. There is no similar correction for the quadratic score statistic, $U_r U_s \kappa^{r,s}$, which is also asymptotically χ_p^2.

7.4.5 *Tensor decomposition of W*

In order to decompose the likelihood ratio statistic into single degree of freedom contrasts, we define the vector with components

$$\begin{aligned} W_r = Z_r &+ n^{-1/2}\{Z_{rs}Z^s/2 + \nu_{rst}Z^sZ^t/3!\} \\ &+ n^{-1}\{Z_{rst}Z^sZ^t/3! + \nu_{rstu}Z^sZ^tZ^u/4! + 3Z_{rs}Z^{st}Z_t/8 \\ &+ 5Z_{rs}Z_tZ_u\nu^{stu}/12 + \nu_{rst}\nu_{uvw}\nu^{t,u}Z^sZ^vZ^w/9\}. \end{aligned} \tag{7.15}$$

It is then easily shown that

$$W = W_rW_s\nu^{r,s} + O_p(n^{-3/2}).$$

Further, taking the Zs as defined in (7.12), W_r is a tensor.

The idea in transforming from W to W_r is that the components of W_r may be interpreted as single degree of freedom contrasts, though they are not independent. In addition, as we now show, the joint distribution of W_r is very nearly normal, a fact that enables us to derive the distribution of W.

A straightforward but rather lengthy calculation shows that the joint cumulants of the W_r are

$$\begin{aligned} E(W_r;\theta) &= n^{-1/2}\nu_{rst}\nu^{s,t}/3! + O(n^{-3/2}) \\ &= -n^{-1/2}\nu_{r,s,t}\nu^{s,t}/3! + O(n^{-3/2}) \\ \operatorname{cov}(W_r,W_s;\theta) &= \nu_{r,s} + n^{-1}\left(\nu_{rstu}\nu^{t,u}/4 + \nu_{rt,su}\nu^{t,u} + \nu_{r,t,su}\nu^{t,u}\right. \\ &\quad \left. +\nu_{r,i,j}\nu_{s,k,l}\nu^{i,k}\nu^{j,l}/6 + 2\nu_{r,s,i}\nu_{j,k,l}\nu^{i,j}\nu^{k,l}/9\right) + O(n^{-2}) \\ \operatorname{cum}(W_r,W_s,W_t;\theta) &= O(n^{-3/2}) \\ \operatorname{cum}(W_r,W_s,W_t,W_u;\theta) &= O(n^{-2}). \end{aligned}$$

Higher-order joint cumulants are of order $O(n^{-3/2})$ or smaller. In other words, ignoring terms that are of order $O(n^{-3/2})$, the components of W_r are jointly normally distributed with the above mean and covariance matrix. To the same order of approximation, it follows that W has a scaled non-central χ^2_p distribution with non-centrality parameter

$$n^{-1}\nu_{r,s,t}\nu_{u,v,w}\nu^{r,s}\nu^{t,u}\nu^{v,w} = n^{-1}p\bar\rho^3_{12} = a/n,$$

which is a quadratic form in $E(W_r;\theta)$. The scale factor in this distribution is a scalar formed from the covariance matrix of W_r, namely

$$\begin{aligned} &1 + \{\nu_{rstu}\nu^{r,s}\nu^{t,u}/4! + \nu_{rt,su}\nu^{r,s}\nu^{t,u} + \rho^2_{23}/6 + 2\rho^2_{13}/9\}/(np) \\ &= 1 + c/n. \end{aligned}$$

The rth cumulant of W (Johnson & Kotz, 1970, p. 134) is

$$\begin{aligned} &2^{r-1}(r-1)!(1+c/n)^r\{1 + ar/(np)\} + O(n^{-3/2}) \\ &= 2^{r-1}(r-1)!p\{1+b/n\}^r + O(n^{-3/2}) \end{aligned}$$

where $b = b(\theta)$ is given by (7.13). Thus the rth cumulant of $W' = W/(1+b/n)$ is $2^{r-1}(r-1)!p + O(n^{-3/2})$, to this order of approximation, the same as the rth cumulant of a χ^2_p random variable. We conclude from this that the corrected statistic (7.14) has the χ^2_p distribution to an unusually high order of approximation.

The argument just given is based entirely on formal calculations involving moments and cumulants. While it is true quite generally, for discrete as well as continuous random variables, that the cumulants of W' differ from those of χ^2_p by $O(n^{-3/2})$, additional regularity conditions are required in order to justify the 'obvious' conclusion that $W' \sim \chi^2_p + O(n^{-3/2})$. Discreteness has an effect that is of order $O(n^{-1/2})$, although the error term can often be reduced to $O(n^{-1})$ if a continuity correction is made. Despite these caveats, the correction is often beneficial even for discrete random variables for which the 'obvious' step cannot readily be justified. The argument is formally correct provided only that the joint distribution of W_r has a valid Edgeworth expansion up to and including the $O(n^{-1})$ term.

7.5 Some examples

7.5.1 *Exponential regression model*

Suppose, independently for each i, that Y_i has the exponential distribution with mean μ_i satisfying the log-linear model

$$\eta^i = \log(\mu_i) = x_r^i \beta^r. \tag{7.16}$$

The notation used here is close to that used in the literature on generalized linear models where η is known as the *linear predictor*, $\mathbf{X} = \{x_r^i\}$ is called the model matrix and β is the vector of unknown parameters. If we let $Z_i = (Y_i - \mu_i)/\mu_i$, then the first two derivatives of the log likelihood may be written in the form

$$U_r = x_r^i Z_i \quad \text{and} \quad U_{rs} = -\sum_i x_r^i x_s^i (Y_i/\mu_i).$$

The joint cumulants are as follows

$$\begin{aligned} n\kappa_{r,s} &= x_r^i x_s^j \delta_{ij} = \mathbf{X}^T\mathbf{X}, & \kappa^{r,s} &= n(\mathbf{X}^T\mathbf{X})^{-1} \\ n\kappa_{r,s,t} &= 2x_r^i x_s^j x_t^k \delta_{ijk} & n\kappa_{r,st} &= -x_r^i x_s^j x_t^k \delta_{ijk} \\ n\kappa_{r,s,t,u} &= 6x_r^i x_s^j x_t^k x_u^l \delta_{ijkl} & n\kappa_{r,s,tu} &= -2x_r^i x_s^j x_t^k x_u^l \delta_{ijkl} \\ n\kappa_{rs,tu} &= x_r^i x_s^j x_t^k x_u^l \delta_{ijkl}. \end{aligned}$$

In addition, we have the following tensorial cumulants

$$\begin{aligned} \nu_{rs,tu} &= \kappa_{rs,tu} - \kappa_{rs,i}\kappa_{tu,j}\kappa^{i,j} \\ \nu_{r,s,tu} &= \kappa_{r,s,tu} - \kappa_{r,s,i}\kappa_{tu,j}\kappa^{i,j}. \end{aligned}$$

In order to express the Bartlett adjustment factor using matrix notation, it is helpful to define the following matrix and vector, both of order n.

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \quad \mathbf{V} = \mathrm{diag}(\mathbf{P}).$$

Note that $\mathbf{P}$ is the usual projection matrix that projects on to the column space of $\mathbf{X}$: it is also the asymptotic covariance matrix of $\hat{\eta}$, the maximum likelihood estimate of η. Thus, the components of $\mathbf{P}$ and $\mathbf{V}$ are $O(n^{-1})$. Straightforward substitution now reveals that the invariant constants in the Bartlett correction term may be written as follows

$$n^{-1}\rho_{13}^2 = 4\mathbf{V}^T\mathbf{P}\mathbf{V}, \quad n^{-1}\rho_{23}^2 = 4\sum_{ij} P_{ij}^3, \quad n^{-1}\rho_4 = 6\mathbf{V}^T\mathbf{V},$$

$$\begin{aligned} n^{-1}\nu^{r,s}\nu^{t,u}\nu_{rt,su} &= \mathbf{V}^T\mathbf{V} - \sum_{ij} P_{ij}^3, \\ n^{-1}\nu^{r,s}\nu^{t,u}\nu_{rs,tu} &= \mathbf{V}^T\mathbf{V} - \mathbf{V}^T\mathbf{P}\mathbf{V}, \\ n^{-1}\nu^{r,s}\nu^{t,u}\nu_{r,s,tu} &= -2\mathbf{V}^T\mathbf{V} + 2\mathbf{V}^T\mathbf{P}\mathbf{V}. \end{aligned}$$

After collecting terms, we find that

$$\epsilon_p = n^{-1}pb(\theta) = \sum_{ij} P_{ij}^3/6 - \mathbf{V}^T(\mathbf{I} - \mathbf{P})\mathbf{V}/4 \tag{7.17}$$

which is independent of the value of the parameter, as is to be expected from considerations of invariance.

In the particular case where $p = 1$ and $\mathbf{X} = \mathbf{1}$, a vector of 1s, we have $\mathbf{V}^T(\mathbf{I} - \mathbf{P})\mathbf{V} = 0$ and the adjustment reduces to $b = 1/6$. This is the adjustment required in testing the hypothesis $H_0 : \mu_i = 1$ (or any other specified value) against the alternative $H_1 : \mu_i = \mu$, where the value μ is left unspecified. More generally, if the observations are divided into k sets each of size m, so that $n = km$, we may wish to test for homogeneity of the means over the k sets. For this purpose, it is convenient to introduce two indices, i indicating the set and j identifying the observation within a set. In other words, we wish to test H_1 against the alternative $H_2 : \mu_{ij} = \mu_i$, where the k means, $\mu_1, \ldots, \mu_k$, are left unspecified. In this case, $\mathbf{X}$ is the incidence matrix for a balanced one-way or completely randomized design. Again, we have $\mathbf{V}^T(\mathbf{I} - \mathbf{P})\mathbf{V} = 0$ and the value of the adjustment is given by

$$\epsilon_k = \sum_{ij} P_{ij}^3/6 = k/(6m).$$

This is the adjustment appropriate for testing H_0 against H_2. To find the adjustment appropriate for the test of H_1 against H_2, we subtract, giving

$$\epsilon_k - \epsilon_1 = k/(6m) - 1/(6n).$$

More generally, if the k sets are of unequal sizes, $m_1, \ldots, m_k$, it is a straightforward exercise to show that

$$\epsilon_k - \epsilon_1 = \sum m_i^{-1}/6 - n^{-1}/6.$$

The test statistic in this case,

$$T = -2 \sum m_i \log(\bar{y}_i/\bar{y})$$

in an obvious notation, is formally identical to Bartlett's (1937) test for homogeneity of variances. It is not difficult to verify directly that the first few cumulants of $(k-1)T/\{k - 1 + \epsilon_k - \epsilon_1\}$ are the same as those of χ^2_{k-1} when terms of order $O(m_i^{-2})$ are ignored (Bartlett, 1937).

The claim just made does not follow from the results derived in Section 7.4.5, which is concerned only with simple null hypotheses. In the case just described, H_1 is composite because the hypothesis does not determine the distribution of the data. Nevertheless, the adjustment still corrects all the cumulants.

7.5.2 *Poisson regression model*

Following closely the notation of the previous section, we assume that Y_i has the Poisson distribution with mean value μ_i satisfying the log-linear model (7.16). The first two derivatives of the log likelihood are

$$U_r = x_r^i(Y_i - \mu_i) \quad \text{and} \quad U_{rs} = -\sum_i x_r^i x_s^i \mu_i = -\mathbf{X}^T\mathbf{W}\mathbf{X},$$

where $\mathbf{W} = \text{diag}(\mu_i)$. In this case, U_{rs} is a constant and all cumulants involving U_{rs} vanish. Since all cumulants of Y_i are equal to μ_i, the cumulants of U_r are

$$n\kappa_{r,s} = \sum_i x_r^i x_s^i \mu_i = \mathbf{X}^T\mathbf{W}\mathbf{X}, \quad \kappa^{r,s} = n(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$

$$n\kappa_{r,s,t} = \sum_i x_r^i x_s^i x_t^i \mu_i, \qquad n\kappa_{r,s,t,u} = \sum_i x_r^i x_s^i x_t^i x_u^i \mu_i.$$

Now define the matrix $\mathbf{P}$ and the vector $\mathbf{V}$ by

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T \quad \text{and} \quad \mathbf{V} = \text{diag}(\mathbf{P})$$

so that $\mathbf{P}$ is the asymptotic covariance matrix of $\hat{\eta}$ and $\mathbf{PW}$ projects on to the column space of $\mathbf{X}$. It follows then that the invariant standardized cumulants of U_r are

$$n^{-1}\rho_{13}^2 = \mathbf{V}^T\mathbf{WPWV}$$
$$n^{-1}\rho_{23}^2 = \sum_{ij}\mu_i\mu_j P_{ij}^3$$
$$n^{-1}\rho_4 = \mathbf{V}^T\mathbf{WV}.$$

Thus, the Bartlett adjustment is given by

$$\epsilon_p = n^{-1}pb(\theta) = \sum_{ij}\mu_i\mu_j P_{ij}^3/6 - \mathbf{V}^T\mathbf{W}(\mathbf{I}-\mathbf{PW})\mathbf{V}/4. \tag{7.18}$$

This expression can be simplified to some extent in those cases where $\mathbf{X}$ is the incidence matrix for a decomposable log-linear model (Williams, 1976). See also Cordeiro (1983) who points out that, for decomposable models, $\mathbf{V}$ lies in the column space of $\mathbf{X}$ implying that the second term on the right of (7.18) vanishes for such models. In general, however, the second term in (7.18) is not identically zero: see Exercise 7.15.

7.5.3 *Inverse Gaussian regression model*

The inverse Gaussian density function, which arises as the density of the first passage time of Brownian motion with positive drift, may be written

$$f_Y(y;\mu,\nu) = \left(\frac{\nu}{2\pi y^3}\right)^{1/2}\exp\left(\frac{-\nu(y-\mu)^2}{2\mu^2 y}\right) \qquad y,\mu>0. \tag{7.19}$$

For a derivation, see Moran (1968, Section 7.23). This density is a member of the two-parameter exponential family. The first four cumulants of Y are

$$\kappa_1 = \mu, \quad \kappa_3 = 3\mu^5/\nu^2, \quad \kappa_2 = \mu^3/\nu, \quad \kappa_4 = 15\mu^7/\nu^3.$$

These can be obtained directly from the generating function

$$K_Y(\xi) = \nu\{b(\theta+\xi/\nu) - b(\theta)\}$$

where $\theta = -(2\mu^2)^{-1}$, $b(\theta) = -(-2\theta)^{1/2}$ and θ is called the canonical parameter if ν is known. See, for example, Tweedie (1957a,b).

Evidently, ν is a precision parameter or 'effective sample size' and plays much the same role as σ^{-2} in normal-theory models. To construct a linear regression model, we suppose for simplicity that ν is given and constant over all observations. The means of the independent random variables $Y_1,\ldots,Y_n$ are assumed to satisfy the inverse linear regression model

$$\eta^i = x_r^i\beta^r, \qquad \eta^i = 1/\mu_i,$$

where $\beta^1,\ldots,\beta^p$ are unknown parameters. This is a particular instance of a generalized linear model in which the variance function is cubic and the link function is the reciprocal. The 'canonical' link function in this instance is $\theta = \mu^{-2}$. In applications, other link functions, particularly the log, might well be found to give a better fit: it is essential, therefore, to check for model adequacy but this aspect will not be explored here.

Using matrix notation, the first two derivatives of the log likelihood may be written

$$U_r = -\nu\{\mathbf{X}^T\mathbf{YX}\beta - \mathbf{X}^T\mathbf{1}\},$$
$$U_{rs} = -\nu\mathbf{X}^T\mathbf{YX},$$

where $\mathbf{Y}$ =diag$\{y_1, \ldots, y_n\}$ is a diagonal matrix of observed random variables.

The above derivatives are formally identical to the derivatives of the usual normal-theory log likelihood with $\mathbf{Y}$ and $\mathbf{1}$ taking the place of the weight matrix and response vector respectively. This analogy is obvious from the interpretation of Brownian motion and from the fact that the likelihood does not depend on the choice of stopping rule. For further discussion of this and related points, see Folks & Chhikara (1978) and the ensuing discussion of that paper.

It follows that the maximum likelihood estimate of β is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{Y}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{1}.$$

This is one of the very rare instances of a generalized linear model for which closed form estimates of the regression parameters exist whatever the model matrix.

The joint cumulants of the log likelihood derivatives are

$$n\kappa_{r,s} = \nu\sum_i x_r^i x_s^i \mu_i = \nu\mathbf{X}^T\mathbf{W}\mathbf{X}, \quad \kappa^{r,s} = n(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}/\nu$$

$$n\kappa_{r,s,t} = -3\nu\sum_i x_r^i x_s^i x_t^i \mu_i^2, \qquad n\kappa_{r,s,t,u} = 15\nu\sum_i x_r^i x_s^i x_t^i x_u^i \mu_i^3,$$
$$n\kappa_{r,st} = \nu\sum_i x_r^i x_s^i x_t^i \mu_i^2, \qquad n\kappa_{rs,tu} = \nu\sum_i x_r^i x_s^i x_t^i x_u^i \mu_i^3,$$
$$n\kappa_{r,s,tu} = -3\nu\sum_i x_r^i x_s^i x_t^i x_u^i \mu_i^3.$$

With $\mathbf{P}$, $\mathbf{V}$ and $\mathbf{W}$ as defined in the previous section, it follows after the usual routine calculations that the $O(n^{-1})$ bias of $\hat{\beta}$ is

$$\text{bias}(\hat{\beta}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{V}/\nu.$$

In addition, the Bartlett adjustment factor may be simplified to

$$\epsilon_p = \Big(\sum_{ij} P_{ij}^3 \mu_i^2 \mu_j^2 + \mathbf{V}^T\mathbf{W}^2\mathbf{P}\mathbf{W}^2\mathbf{V} - 2\mathbf{V}^T\mathbf{W}^3\mathbf{V}\Big)\Big/\nu. \tag{7.20}$$

This expression is suitable for routine computation using simple matrix manipulations. It may be verified that if $\mathbf{X}$ is the incidence matrix for a one-way layout, then $\epsilon_p = 0$, in agreement with the known result that the likelihood ratio statistic in this case has an exact χ_p^2 distribution. More generally, for designs that have sufficient structure to discriminate between different link functions, the Bartlett adjustment is non-zero, showing that the exact results do not extend beyond the one-way layout.

7.6 Bibliographic notes

The emphasis in this chapter has been on likelihood ratio statistics and other invariants derived from the likelihood function. In this respect, the development follows closely to that of McCullagh & Cox (1986). Univariate invariants derived from the likelihood are discussed by Peers (1978).

The validity of adjusting the likelihood ratio statistic by means of a straightforward multiplicative correction was first demonstrated by Lawley (1956). His derivation is more comprehensive than ours in that it also covers the important case where nuisance parameters are present.

Bartlett factors and their role as normalization factors are discussed by Barndorff-Nielsen & Cox (1984).

Geometrical aspects of normal-theory non-linear models have been discussed by a number of authors from a slightly different perspective: see, for example, Beale (1960), Bates & Watts (1980) or Johansen (1983).

Amari (1985) deals with arbitrary likelihoods for regular problems, and emphasises the geometrical interpretation of various invariants. These 'interpretations' involve various kinds of *curvatures* and *connection coefficients*. It is not clear that such notions, borrowed from the differential geometry of esoteric spaces, necessarily shed much light on the interpretation of statistical invariants. The subject, however, is still young!

For a slightly more positive appraisal of the role of differential geometry in statistical theory, see the review paper by Barndorff-Nielsen, Cox & Reid (1986).

Further aspects of differential geometry in statistical theory are discussed in a forthcoming IMS monograph by Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao.

7.7 Further results and exercises 7

7.1 Let X_r, X_{rs}, $X_{rst}, \ldots$ be a sequence of arrays of arbitrary random variables. Such a sequence will be called *triangular*. Let the joint moments and cumulants be denoted as in Section 7.2.1 by

$$\begin{aligned}\mu_{r,st,uvw} &= E\{X_r X_{st} X_{uvw}\} \\ \kappa_{r,st,uvw} &= \operatorname{cum}\{X_r, X_{st}, X_{uvw}\}\end{aligned}$$

and so on. Now write $\mu_{[\ldots]}$ and $\kappa_{[\ldots]}$ for the sum over all partitions of the subscripts as follows

$$\begin{aligned}\mu_{[rs]} &= \mu_{rs} + \mu_{r,s} \\ \mu_{[rst]} &= \mu_{rst} + \mu_{r,st}[3] + \mu_{r,s,t}\end{aligned}$$

and so on, with identical definitions for $\kappa_{[rs]}$, $\kappa_{[rst]}$ and so on. Show that $\kappa_{[r]} = \mu_{[r]}$,

$$\begin{aligned}\kappa_{[rs]} &= \mu_{[rs]} - \mu_{[r]}\mu_{[s]} \\ \kappa_{[rst]} &= \mu_{[rst]} - \mu_{[r]}\mu_{[st]}[3] + 2\mu_{[r]}\mu_{[s]}\mu_{[t]} \\ \kappa_{[rstu]} &= \mu_{[rstu]} - \mu_{[r]}\mu_{[stu]}[4] - \mu_{[rs]}\mu_{[tu]}[3] + 2\mu_{[r]}\mu_{[s]}\mu_{[tu]}[6] \\ &\quad - 6\mu_{[r]}\mu_{[s]}\mu_{[t]}\mu_{[u]}.\end{aligned}$$

Hence show that the cumulants of the log likelihood derivatives satisfy $\kappa_{[\ldots]} = 0$, whatever the indices.

7.2 Give a probabilistic interpretation of $\mu_{[\ldots]}$ and $\kappa_{[\ldots]}$ as defined in the previous exercise.

7.3 Give the inverse formulae for $\mu_{[\ldots]}$ in terms of $\kappa_{[\ldots]}$.

7.4 By first examining the derivatives of the null moments of log likelihood derivatives, show that the derivatives of the null cumulants satisfy

$$\begin{aligned}\frac{\partial \kappa_{r,s}(\theta)}{\partial \theta^t} &= \kappa_{rt,s} + \kappa_{r,st} + \kappa_{r,s,t} \\ \frac{\partial \kappa_{rs}(\theta)}{\partial \theta^t} &= \kappa_{rst} + \kappa_{rs,t} \\ \frac{\partial \kappa_{r,st}(\theta)}{\partial \theta^u} &= \kappa_{ru,st} + \kappa_{r,stu} + \kappa_{r,st,u} \\ \frac{\partial \kappa_{r,s,t}(\theta)}{\partial \theta^u} &= \kappa_{ru,s,t} + \kappa_{r,su,t} + \kappa_{r,s,tu} + \kappa_{r,s,t,u}.\end{aligned}$$

State the generalization of this result that applies to

(i) cumulants of arbitrary order (Skovgaard, 1986a)
(ii) derivatives of arbitrary order.

Hence derive identities (7.2) by repeated differentiation of $\kappa_r(\theta)$.

7.5 Using expansions (7.3) for the non-null moments, derive expansions (7.4) for the non-null cumulants.

7.6 Express the four equations (7.6) simultaneously using matrix notation in the form

$$\mathbf{U} = \mathbf{B}\mathbf{V}$$

and give a description of the matrix $\mathbf{B}$. It may be helpful to define $\beta_r^i = \delta_r^i$.

7.7 Show that equations (7.7) may be written in the form

$$\mathbf{U} = \mathbf{A}\bar{\mathbf{U}}$$

where $\mathbf{A}$ has the same structure as $\mathbf{B}$ above.

7.8 Show that under transformation of coordinates on Θ, the coefficient matrix $\mathbf{B}$ transforms to $\bar{\mathbf{B}}$, where

$$\mathbf{B} = \mathbf{A}\bar{\mathbf{B}}\mathbf{A}^{*-1}$$

and give a description of the matrix $\mathbf{A}^*$.

7.9 Using the results of the previous three exercises, show that the arrays V_r, $V_{rs}, \ldots$, defined at (7.6), behave as tensors under change of coordinates on Θ.

7.10 Let s_i^2, $i = 1, \ldots, k$ be k independent mean squares calculated from independent normal random variables. Suppose

$$E(s_i^2) = \sigma_i^2, \quad \text{var}(s_i^2) = 2\sigma_i^2/m_i$$

where m_i is the number of degrees of freedom for s_i^2. Derive the likelihood ratio statistic, W, for testing the hypothesis $H_0 : \sigma_i^2 = \sigma^2$ against the alternative that leaves the variances unspecified. Using the results of Section 7.5.1, show that under H_0,

$$E(W) = k - 1 + \tfrac{1}{3}\textstyle\sum m_i^{-1} - \tfrac{1}{3}m_\bullet^{-1}$$

where $m_\bullet$ is the total degrees of freedom. Hence derive Bartlett's test for homogeneity of variances (Bartlett, 1937).

7.11 Suppose that $Y_1, \ldots, Y_n$ are independent Poisson random variables with mean μ. Show that the likelihood ratio statistic for testing $H_0 : \mu = \mu_0$ against an unspecified alternative is

$$W = 2n\{\bar{Y}\log(\bar{Y}/\mu_0) - (\bar{Y} - \mu_0)\}.$$

By expanding in a Taylor series about $\bar{Y} = \mu_0$ as far as the quartic term, show that

$$E(W; \mu_0) = 1 + \frac{1}{6n\mu_0} + O(n^{-2}).$$

7.12 Derive the result stated in the previous exercise directly from (7.18). Check the result numerically for $\mu_0 = 1$ and for $n = 1, 5, 10$. Also, check the variance of W numerically. You will need either a computer or a programmable calculator.

7.13 Using the notation of the previous two exercises, let $\pm W^{1/2}$ be the signed square root of W, where the sign is that of $\bar{Y} - \mu_0$. Using the results given in Section 7.4.5, or otherwise, show that

$$\begin{aligned} E(\pm W^{1/2}) &= -\frac{1}{6(n\mu_0)^{1/2}} + O(n^{-3/2}), \\ \text{var}(\pm W^{1/2}) &= 1 + \frac{1}{8n\mu_0} + O(n^{-2}), \\ \kappa_3(\pm W^{1/2}) &= O(n^{-3/2}), \qquad \kappa_4(\pm W^{1/2}) = O(n^{-2}). \end{aligned}$$

Hence show that under $H_0 : \mu = \mu_0$,

$$S = \frac{\pm W^{1/2} + (n\mu_0)^{-1/2}/6}{1 + (16n\mu_0)^{-1}}$$

has the same moments as those of $N(0,1)$ when terms of order $O(n^{-3/2})$ are ignored. Why is it technically wrong in this case to say that

$$S \sim N(0,1) + O(n^{-3/2})?$$

7.14 Repeat the calculations of the previous exercise, this time for the exponential distribution in place of the Poisson. Compare numerically the transformation $\pm W^{1/2}$ with the Wilson-Hilferty cube root transformation

$$3n^{1/2}\left\{\left(\frac{\bar{Y}}{\mu_0}\right)^{1/3} + \frac{1}{9n} - 1\right\},$$

which is also normally distributed to a high order of approximation. Show that the cube root transformation has kurtosis of order $O(n^{-1})$, whereas $\pm W^{1/2}$ has kurtosis of order $O(n^{-2})$. For a derivation of the above result, see Kendall & Stuart (1977, Section 16.7).

7.15 Show that the second term on the right in (7.17) is zero if $\mathbf{X}$ is the incidence matrix for

(i) an unbalanced one-way layout
(ii) a randomized blocks design (two-way design) with equal numbers of replications per cell.
(iii) a Latin square design.

Show that the second term is not zero if $\mathbf{X}$ is the model matrix for an ordinary linear regression model in which more than two x-values are observed.

7.16 Find expressions for the first term in (7.17) for the four designs mentioned in the previous exercise.

7.17 Simplify expression (7.18) in the case of a two-way contingency table and a model that includes no interaction term. Show that the second term is zero.

7.18 Using expression (7.13) for $b(\theta)$ together with the expressions given for the cumulants in Section 7.5.1, derive (7.17) as the Bartlett correction applicable to the exponential regression model (7.16).

7.19 Comment on the similarity between the correction terms, (7.17) and (7.18).

7.20 Using expression (7.15) for W_r, show that the joint third and fourth cumulants are $O(n^{-3/2})$ and $O(n^{-2})$ respectively. Derive the mean vector and covariance matrix. Hence justify the use of the Bartlett adjustment as a multiplicative factor.

7.21 *Normal circle model*: Suppose Y is bivariate normal with mean $(\rho\cos(\theta), \rho\sin(\theta))$ and covariance matrix $n^{-1}I_2$. Let $\rho = \rho_0$ be given.

(i) Find the maximum likelihood estimate of θ.
(ii) Derive the likelihood ratio statistic for testing the hypothesis $H_0 : \theta = 0$.
(iii) Interpret the first two log likelihood derivatives and the likelihood ratio statistic geometrically.
(iv) Show that the Bartlett correction for testing the hypothesis in (ii) is $b(\theta) = 1/(4\rho_0^2)$.
(v) Show that the first derivative of the log likelihood function is normally distributed: find its variance.

7.22 Using the results derived in the previous exercise for $n = 4$ and $\rho_0 = 1$, construct 95% confidence intervals for θ based on (a) the score statistic and (b) the likelihood ratio statistic. For numerical purposes, consider the two data values $(y_1, y_2) = (0.5, 0.0)$ and $(1.5, 0.0)$. Is the value $\theta = \pi$ consistent with either observation? Plot the log likelihood function. Plot the first derivative against θ. Comment on the differences between the confidence intervals.

7.23 *Normal spherical model*: Suppose Y is a trivariate normal random vector with mean $(\rho\cos(\theta)\cos(\phi), \rho\cos(\theta)\sin(\phi), \rho\sin(\theta))$ and covariance matrix $n^{-1}I_3$. Let $\rho = \rho_0$ be given.

(i) Find the maximum likelihood estimate of (θ, ϕ).
(ii) Derive the likelihood ratio statistic for testing the hypothesis $H_0 : \theta = 0, \phi = 0$.
(iii) Show that the Bartlett correction for testing the hypothesis in (ii) is identically zero regardless of the value of ρ_0.

7.24 *Normal hypersphere model*: Repeat the calculations of the previous exercise, replacing the spherical surface in 3-space by a p-dimensional spherical surface in R^{p+1}. Show that the Bartlett adjustment reduces to

$$b(\theta) = \frac{-(p-2)}{4\rho_0^2},$$

which is negative for $p \geq 3$. (McCullagh & Cox, 1986).

7.25 By considering the sum of two independent inverse Gaussian random variables, justify the interpretation of ν in (7.19) as an 'effective sample size'.

7.26 Show that (7.20) vanishes if $\mathbf{X}$ is the incidence matrix for an unbalanced one-way layout.

7.27 Derive the expression analogous to (7.20) for the log link function replacing the reciprocal function. Simplify in the case $p = 1$.

7.28 For the exponential regression model of Section 7.5.1, show that the $O(n^{-1})$ bias of $\hat{\beta}$ is

$$\text{bias}(\hat{\beta}) = -(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}/2.$$

Show that, for a simple random sample of size 1, the bias is exactly $-\gamma$, where $\gamma = 0.57721$ is Euler's constant. Find the exact bias for a simple random sample of size n and compare with the approximate formula.

7.29 Repeat the calculations of the previous exercise for the Poisson log-linear model of Section 7.5.2.