CHAPTER 0

Tensors

0.1 What is a tensor?

0.1.1 *Difficult questions*

In one form or another, the title of this section is a question that I have frequently been asked, often by professional colleagues who should know better, but sometimes by unsuspecting strangers at social gatherings. This is not an ideal conversation starter, but politeness demands an answer. The question is usually asked out of innocent curiosity or out of simple naivety, but I invariably assume that my interlocutor has a burning desire to know the truth, at least as I see it. The question could hardly be more straightforward, and yet I have been woefully unsuccessful at providing anything resembling a satisfactory answer. Even a partial answer that elicited a follow-up question would rate as a success, but this is rare, especially so in mixed company at cocktail parties. Why is it that the simplest questions are so often the hardest to answer? This note explains my current thoughts on the topic, and will serve as a guide should someone be foolish enough to raise the question at a future occasion.

As every successful politician knows, the essential first step in answering any difficult question is to 're-phrase the question.' This can mean anything from 'the question I think was intended' to 'the question I know how to answer' to 'the question I want to be asked.' Even a straightforward, apparently unambiguous, question such as 'what is a tensor?' is susceptible to this ploy. Here are some re-phrased versions of the question, roughly in increasing order of difficulty.

1. What is the mathematical definition of a tensor?
2. Is a tensor a kind of vector?
3. Is a vector a kind of tensor?
4. Is a matrix a special kind of tensor?
5. When you write λ^i, do you mean a row vector or a column vector?
6. What are tensors used for?
7. I know that a vector has magnitude and direction, but what does a tensor look like?
8. How would I recognize a tensor if I met one in the dark?

The first of these questions is easily disposed of by the stock answer, courtesy of vector spaces, 'a tensor is an element of a tensor space.' The second and third questions are answered in a single word, 'yes.' Question 4 is only slightly more difficult, the answer being 'yes and no.' Question 5 deals with typographical strategy: it all depends on whether the elements are written across the page, down the page, or diagonally.

Audiences are invariably nonplussed by the directness and simplicity of these answers. The obvious correctness of the answer to question 1 is eloquent testimony to the persuasive power of modern mathematics. What more is there to say?

Regarding question 7, one could preface a reply by stating that vectors need have neither magnitude nor direction, and the same is true for tensors. But that, I suspect, is more likely to confuse than to enlighten.

Question 8 is the really tough nut. How, for example, can one tell from observation that an electric field is a vector and a magnetic field is a tensor? After all, they do not look very different. In my experience, for these and other reasons, the latter questions are best evaded. Distraction tactics often work well. Offer to extemporize on Aristotelian philosophy. Even volunteer to explain confidence intervals. On no account be drawn into geometrical descriptions or physical analogies in response to questions 7–8. That way lies certain failure. These, of course, are precisely the questions that every inquisitive student wants to ask, but it is a futile and anti-social exercise to give anything other than a flippant answer, particularly at social gatherings. The more noble the attempt at answering honestly, the more demoralizing the failure, so much so that even an expert begins to have doubts.

0.1.2 *Definitions and/or examples*

With a little training it is fairly easy even for a child to identify a chimpanzee, a tiger or a human being by sight. But to define what is meant by the term 'chimpanzee' in language that would satisfy a logician or mathematician is a much more difficult task. One would not normally cite this inability to produce a definition as evidence that the child does not know what a chimpanzee is. For taxonomic purposes, examples and descriptive characteristics are more useful than definitions.

The purpose of definitions in mathematics is to crystallize essential properties, while eliminating irrelevant subject-matter details. When divorced from examples this is a mindless antiseptic exercise, but when viewed in its proper context a good definition can be unifying and enlightening. A vector, for example, can be described as a quantity with magnitude and direction. This is not satisfactory as a mathematical definition because no mention is made of addition or the triangle rule, but it is nonetheless a useful description in many physical contexts. The standard mathematical definition as an element of a certain kind of set, is cleaner, in many ways more powerful, but is totally lacking in descriptive power and geometrical insight.

In the historical development of science and mathematics, concepts are invariably encountered before adequate definitions are in place. On the whole, the opposite is true in textbooks. However, the lack of a satisfactory definition need not be a barrier to progress. Indeed substantial progress and good understanding are essential in order to formulate good definitions. It can be a useful didactic tool to explore a number of examples before the unifying theme is revealed. The notes in this chapter are neither complete nor self-contained. The treatment of examples is of necessity somewhat condensed. Concepts arise in examples before they are defined. This reversal of the familiar order will confuse some readers, but it may help others to appreciate better the relation between examples and definitions. Even if a satisfactory definition is not achieved, the examples will have served their purpose, and we will have a good working idea of what a tensor is.

0.2 Straightforward answers?

0.2.1 *What is a vector?*

It is assumed that the reader has some familiarity with vectors, yet it is felt necessary to address this question head-on. A vector space is a set $\mathcal{V}$ that is closed under a commutative binary operation called addition, and closed under multiplication by scalars. In symbols,

$$\alpha_1 v_1 + \alpha_2 v_2 = \alpha_2 v_2 + \alpha_1 v_1 \in \mathcal{V}$$

for all $v_1, v_2 \in \mathcal{V}$ and for all scalars α_1, α_2. Unless otherwise specified, scalars are taken to be real numbers, although in many cases complex numbers would serve equally well. A vector is then defined to be an element of a vector space.

The complete definition of a vector space requires a number of additional conditions that are trivially satisfied by all examples in this book, but which need to be checked in more exotic cases. In particular, it is required that there exist a point $0 \in \mathcal{V}$ such that $v + 0 = v$ and $v + (-v) = 0$ for every v. In addition, associative and distributive conditions are needed to ensure that $1v = v$ and $v + v = 2v$ for every v. Full details can be found in Halmos, (1958, section 2).

The vector property is thus a property of a set, not a property of individual elements taken in isolation. According to the definition, therefore, the ordered pair $v = (0.25, 0.75)$ is not a vector unless and until we specify a suitable set $\mathcal{V}$ to which v belongs. Consider in particular the following sets:

$$\begin{aligned} \mathcal{V}_1 &= \{(\xi_1, \xi_2) : \xi_1, \xi_2 \in \mathcal{R}\}; \\ \mathcal{V}_2 &= \{(\xi_1, \xi_2) : 0 \le \xi_1, \xi_2 \le 1\}; \\ \mathcal{V}_3 &= \{(\xi, 3\xi) : \xi \in \mathcal{R}\}; \end{aligned}$$

where $\mathcal{R}$ denotes the real line. Evidently v is a point in each of these sets, but only $\mathcal{V}_1$ and $\mathcal{V}_3$ are vector spaces with the usual definitions of addition and multiplication by scalars.

The preceding definitions of vector and vector space are entirely abstract. No attempt is made to convey a picture of what it is that constitutes a vector or of what a vector 'looks like' in a physical or operational sense. That failure is understandable, inevitable, and ultimately desirable in view of the enormous variety of objects, even within statistics, that can be considered as vectors. In order to strike a balance, therefore, we make liberal use of examples to supplement the definitions throughout these notes.

0.2.2 *What is a tensor?*

In section 1.5 we define the tensor product of vector spaces, $\mathcal{V}_1$ and $\mathcal{V}_2$. It suffices for the moment to say that the tensor product $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ is a new vector space constructed from the component spaces $\mathcal{V}_1$ and $\mathcal{V}_2$. A tensor of order one is a vector, i.e. an element of a vector space. A tensor of order two is an element of a tensor product space $\mathcal{V}_1 \otimes \mathcal{V}_2$. Tensors of higher order are defined by repeated tensor products of component spaces. Thus all vectors are tensors, and, since the tensor product space is a vector space, all tensors are vectors. This may sound a little strange, even paradoxical, perhaps. In practice, when we say that some quantity is a tensor both the order and the component spaces are usually implied by notation and context. In contexts where tensors of various orders occur, it is common to use the word 'vector' to mean 'tensor of order one,' implying that the space in question is not a tensor product of component spaces.

Even at this stage, although we have not yet defined the tensor product, it is possible to give a simple example to convey the flavour of what is meant by tensor product.

Let $\mathcal{V}$ be the space of real-valued quadratic functions on the interval $[0, 1]$. An element in $\mathcal{V}$ is thus a function v whose numerical value at the point $x \in [0, 1]$ is given by the formula

$$v(x) = \xi_0 + \xi_1 x + \xi_2 x^2,$$

for some real-valued coefficients ξ_0, ξ_1, ξ_2. Although the elements in $\mathcal{V}$ are real-valued functions, they are in 1–1 correspondence with the set of vectors $(\xi_0, \xi_1, \xi_2) \in \mathcal{R}^3$. Evidently $\mathcal{V}$ is a three-dimensional vector space with respect to the usual definitions of addition of polynomials and multiplication of polynomials by scalars. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be independent copies of $\mathcal{V}$. In other words, elements of $\mathcal{V}_1$ are quadratic functions of a variable x_1: elements of $\mathcal{V}_2$ are quadratic functions of a variable x_2. The tensor product space $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ is the space of functions on the unit square, that are quadratic in x_1 for each fixed value of x_2, and quadratic in x_2 for each fixed value of x_1. In symbols,

$$\mathcal{V}_1 = \mathrm{span}\{1,\, x_1,\, x_1^2\}; \qquad \mathcal{V}_2 = \mathrm{span}\{1,\, x_2,\, x_2^2\};$$
$$\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2 = \mathrm{span}\{1,\, x_1,\, x_1^2,\, x_2,\, x_1x_2,\, x_1^2x_2,\, x_2^2,\, x_1x_2^2,\, x_1^2x_2^2\}.$$

Elements in $\mathcal{W}$ are thus linear combinations of formal products of elements in the component spaces. In this example where $\mathcal{V}_1$ and $\mathcal{V}_2$ are essentially (apart from notation) the same space, the tensor product space is often denoted by the symbol $\mathcal{V}^{\otimes 2}$.

Two examples of functions in $\mathcal{W}$ are the following:

$$\begin{aligned} w_1(x) &= 4 - x_1^2 - x_2^2 + 2x_1x_2(3 + x_1 + x_2) - x_1^2x_2^2; \\ w_2(x) &= x_1 - x_2 + 2x_1x_2(x_2 - x_1). \end{aligned}$$

Such polynomials are uniquely determined by their coefficient vectors, which in this case are conveniently portrayed as the 3×3 matrices

$$\begin{array}{c|ccc} & 1 & x_2 & x_2^2 \\ \hline 1 & 4 & 0 & -1 \\ x_1 & 0 & 6 & 2 \\ x_1^2 & -1 & 2 & -1 \end{array} \quad \text{and} \quad \begin{array}{c|ccc} & 1 & x_2 & x_2^2 \\ \hline 1 & 0 & -1 & 0 \\ x_1 & 1 & 0 & 2 \\ x_1^2 & 0 & -2 & 0 \end{array}.$$

Evidently, in this example $\mathcal{W}$ contains two non-overlapping subspaces, one consisting of symmetric functions satisfying $w(x_1, x_2) = w(x_2, x_1)$, the other consisting of skew-symmetric functions satisfying $w(x_1, x_2) = -w(x_2, x_1)$.

0.2.3 *Components of a vector*

Let $\mathcal{V}$ be a vector space of dimension n, and let $\{e_1, \ldots, e_n\}$ be a basis in $\mathcal{V}$. Then each element $v \in \mathcal{V}$ has a unique representation as a linear combination of the basis vectors:

$$v = v^1 e_1 + \cdots + v^n e_n. \tag{1.1}$$

We say that $(v^1, \ldots, v^n)$, or simply v^i, are the components of v with respect to the basis $\{e_1, \ldots, e_n\}$. An essential aspect of tensor notation in this context is that superscripts denote components of a vector in $\mathcal{V}$, and are not to be confused with powers.

It is important at the outset to understand that $(v^1, \ldots, v^n)$, as an ordered set of real numbers in $\mathcal{R}^n$, serves to identify $v \in \mathcal{V}$, but is not in any sense the same point as v because the objects in $\mathcal{V}$ need not be ordered sets of real numbers. The quadratic and bi-quadratic functions at the end of the preceding section illustrate the distinction. The following example is similar but simpler.

Let $\mathcal{V}$ be the space of functions on $[0, 2\pi)$ spanned by the vectors $e_1 = \cos\theta$ and $e_2 = \sin\theta$. Then each element in $\mathcal{V}$ is a linear combination of the form $v(\theta) = 2\cos\theta + 5\sin\theta$, i.e. a function $v(\theta)$ defined on the interval $[0, 2\pi)$. The components of v, in this example $(2, 5)$, may be regarded as a point in $\mathcal{R}^2$, a very different kind of object from a function on the real line. In the bi-quadratic example at the end of the preceding section, the component vector was portrayed as a matrix, quite a different kind of object from a polynomial.

There is another major distinction between the abstract notion of a point $v \in \mathcal{V}$ and its description by means of components $(v^1, \ldots, v^n)$. While it is possible to conceive of a point $v \in \mathcal{V}$ in the abstract, its description via components necessarily involves a basis. The choice of basis is usually arbitrary up to linear transformation. If a new basis is chosen, the description of v by its components must be altered from the original $(v^1, \ldots, v^n)$ to new components $(\bar{v}^1, \ldots, \bar{v}^n)$, say. This transformation is necessary in order that the description $(\bar{v}^1, \ldots, \bar{v}^n)$ with respect to the new basis should identify the same point in $\mathcal{V}$ as $(v^1, \ldots, v^n)$ with respect to the original basis. It is precisely this transformation property that is an automatic and integral part of the tensor calculus.

Despite the conceptual distinction between the vector of components $(v^1, \ldots, v^n)$ and the point $v \in \mathcal{V}$, we shall occasionally refer to 'the point $v = (v^1, \ldots, v^n)$' when in fact we mean 'the point $v \in \mathcal{V}$ whose components with respect to the basis $\{e_1, \ldots, e_n\}$ are $(v^1, \ldots, v^n)$.' Usually the correct meaning is obvious from the context.

0.2.4 *Subspaces and cosets*

We now examine a number or ways in which an existing vector space can be partitioned or extended. Two notions are particularly important in this respect, namely the notion of a vector subspace, and the notion of the span of a collection of vectors.

DEFINITION: A non-empty subset $\mathcal{U}$ of a vector spave $\mathcal{V}$ is a *subspace* if it is closed under linear combinations. In other words, $\mathcal{U}$ is a subspace if, for each pair of points x, y in $\mathcal{U}$, every linear combination $\alpha x + \beta y$ is also in $\mathcal{U}$.

A subspace of a vector space is itself a vector space with the same operations of addition and scalar multiplication derived from the parent space. In particular, every subspace includes the origin. Two extreme subspaces are the set $\mathcal{O}$ consisting of the origin only, and the subspace comprising the whole space $\mathcal{V}$.

The intersection of two or more subspaces is again a subspace, as can be seen from the definition. The union of two subspaces is not ordinarily a subspace because the union does not contain linear combinations. Venn diagrams are thus inappropriate for the representation of subspace operations.

Two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ are said to be *non-overlapping* if the origin is the only point of intersection, i.e. $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{O}$. Note that $\mathcal{O}$ is not the empty set but rather a vector space of dimension zero, so this definition is at odds with the definition of disjoint sets.

DEFINITION: Let U be any subset of $\mathcal{V}$, not necessarily a subspace. The *span* of U, denoted by $\operatorname{span}(U)$, is that subset of $\mathcal{V}$ consisting of linear combinations of the elements of U. If U is a finite set, the span is often written in the form $\operatorname{span}\{u_1, \ldots, u_p\}$.

The span of any collection of vectors in $\mathcal{V}$ is a subspace.

The span of two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ is the set of vectors $x + y$ with $x \in \mathcal{U}_1$ and $y \in \mathcal{U}_2$. This set, denoted by $\mathcal{U}_1 + \mathcal{U}_2$, is a subspace of $\mathcal{V}$. Note that $\mathcal{U} + \mathcal{V} = \mathcal{V}$, and $\mathcal{U} + \mathcal{U} = \mathcal{U}$.

DEFINITION: Let $\mathcal{U}$ be a subspace of $\mathcal{V}$. The set of vectors $x + y$ with $y \in \mathcal{U}$, is called the *translate of $\mathcal{U}$ by x*, or a *coset* of $\mathcal{U}$, and is denoted by $x + \mathcal{U}$. If x happens to be an element of $\mathcal{U}$, then $x + \mathcal{U} = \mathcal{U}$, in which case the coset is a vector space. In general, however, if x is not in $\mathcal{U}$, the coset is not a vector space because it does not include the origin. Note that if $x + \mathcal{U} = y + \mathcal{U}$ it does not follow that x and y are the same point in $\mathcal{V}$. It does follow that $x - y$ is a point in $\mathcal{U}$.

For mathematical purposes, the following statistical examples are somewhat simplified to avoid the kinds of complications that invariably arise in applied work.

Example 1: Linear regression. Multiple linear regression refers to a class of statistical models used for studying the dependence of a response variable Y on a (small) number of explanatory variables $x_1, \ldots, x_p$. For present purposes, all of these variables are conceived of as points, vectors or functions in a finite-dimensional vector space $\mathcal{V}$, often but not necessarily equal to R^n. Usually the systematic part of the model is written in the form

$$E(Y \mid x) = \beta_1 x_1 + \cdots + \beta_p x_p$$

where $\beta_1, \ldots, \beta_p$ are unknown scalar coefficients to be estimated. The ultimate aims of multiple regression are many and varied. From our present perspective, if $\beta_1, \ldots, \beta_p$ are unrestricted coefficients, the model states simply that the conditional expected value $E(Y \mid x)$ lies in the subspace defined by the span of the vectors $x_1, \ldots, x_p$. In most cases, x_1 is the constant vector, and β_1 is then called the intercept.

Example 2: Polynomial regression. Let $Y(x)$ be the response observed at the point $x = (x_1, x_2)$ in the plane, or in some subset thereof. For example, Y might denote chemical yield (possibly on a logarithmic scale as a percent of the theoretical maximum), x_1 reaction temperature, and x_2 reaction pressure. Let $\mathcal{U}$ be the space spanned by the functions $1, x_1, x_2, x_1^2, x_2^2, x_1 x_2$. In this case, the response vector, or rather, the expected response vector, is conceived of as a point in the space of functions on 'reaction space', and the subspace $\mathcal{U}$ is the space of polynomials of total degree at most two. In practice, observations are taken at a finite grid of points in the reaction space, in which case $\mathcal{V}$ is the space of functions on this grid, and $\mathcal{U}$ is the subspace of quadratic functions on the same grid.

Example 3: Additive effects model. Let S denote the grid of points $\{1, \ldots, m\} \times \{1, \ldots, n\}$ in the plane, and let $\mathcal{V}$ be the vector space of real-valued functions of two variables on S. Let $\mathcal{U}_1$ be the subspace of functions of one variable x on the points $\{1, \ldots, m\}$, and let $\mathcal{U}_2$ be the subspace of functions of the second variable t on $\{1, \ldots, n\}$. Although we write $f_1(x)$ and $f_2(t)$ for typical elements of $\mathcal{U}_1$ and $\mathcal{U}_2$, it should be borne in mind that f_1 is conceived here as a function on S, i.e. a function of two variables that happens not to depend on t. Consequently, $\mathcal{U}_0$, the space of functions that are constant on S, is a subspace of both $\mathcal{U}_1$ and $\mathcal{U}_2$. In fact $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{U}_0$. Finally, $\mathcal{U}_1 + \mathcal{U}_2$, the span of the two subspaces, is the space of functions of the form $f_1(x) + f_2(t)$ on S.

In certain areas of applied statistical work, the response at (i, j) is denoted by Y_{ij}, and the additive effects model is written algebraically as $E(Y_{ij}) = \alpha_i + \beta_j$. An alternative, so-called symbolic or model-formula, notation is also employed in the form $A + B$, where A and B are called factors. From the present viewpoint, these factors are vector spaces coinciding with $\mathcal{U}_1$ and $\mathcal{U}_2$, and the operator '+' is the vector span.

0.2.5 *Dimension of a subspace*

We need only cite two facts about subspaces. First, a subspace $\mathcal{U}$ in a vector space $\mathcal{V}$ of dimension n is a vector space of dimension not more than n. Second, given a p-dimensional subspace $\mathcal{U}$ in $\mathcal{V}$, there exists a basis $\{x_1, \ldots, x_p, x_{p+1}, \ldots, x_n\}$ in $\mathcal{V}$ such that $x_1, \ldots, x_p$ are in $\mathcal{U}$ and form a basis for $\mathcal{U}$. (Halmos, 1958, section 12).

0.2.6 *Complementary spaces*

Two non-overlapping vector subspaces having the property that their span coincides with the entire space are said to be *complementary*. In the symbols introduced above, two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ are complementary in $\mathcal{V}$ if $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{O}$ and $\mathcal{U}_1 + \mathcal{U}_2 = \mathcal{V}$. If $\mathcal{V}$ is a vector space of dimension n, and $\mathcal{U}_1$ is a subspace of dimension p then the complementary space has dimension $n - p$. Conversely, if $\mathcal{U}_1$ and $\mathcal{U}_2$ are non-overlapping subspaces of dimensions p and $n - p$ in a vector space of dimension n, then $\mathcal{U}_1$ and $\mathcal{U}_2$ are complementary spaces. (Halmos 1958, sections 11, 12).

The importance of complementary subspaces lies in the fact that any vector v in $\mathcal{V}$ can be written as the sum of two components $v = x + y$ with $x \in \mathcal{U}_1$ and $y \in \mathcal{U}_2$. In this respect, both parts of the definition are important. The second part, $\mathcal{U}_1 + \mathcal{U}_2 = \mathcal{V}$ ensures that points $x \in \mathcal{U}_1$ and $y \in U_2$ exist satisfying $x + y = v$ for any $v \in \mathcal{V}$. The non-overlapping condition on the subspaces ensures that the decomposition is unique.

Example 4: Additive effects model continued. In the notation previously established, let $\mathcal{U}_1'$ be the set of functions on $\{1, \ldots, m\}$ satisfying $\sum f_1(x) = 0$. This is a subspace of $\mathcal{U}_1$ satisfying $\mathcal{U}_0 \cap \mathcal{U}_1' = \mathcal{O}$, and $\mathcal{U}_0 + \mathcal{U}_1' = \mathcal{U}_1$, so that $\mathcal{U}_0$ and $\mathcal{U}_1'$ are complementary subspaces of $\mathcal{U}_1$. Likewise, if $\mathcal{U}_2'$ is the set of functions on $\{1, \ldots, n\}$ satisfying $\sum f_2(y) = 0$, $\mathcal{U}_0$ and $\mathcal{U}_2'$ are complementary subspaces of $\mathcal{U}_2$. Also, $\mathcal{U}_0$, $\mathcal{U}_1'$ and $\mathcal{U}_2'$ are complementary subspaces of $\mathcal{U}_1 + \mathcal{U}_2$: they are non-overlapping and their span is $\mathcal{U}_1 + \mathcal{U}_2$. Finally, if we let $\mathcal{U}_{12}'$ be the set of functions on S satisfying

$$\begin{aligned} \sum_x f(x, y) = 0 \quad &\text{for} \quad y = 1, \ldots, n, \\ \sum_y f(x, y) = 0 \quad &\text{for} \quad x = 1, \ldots, m, \end{aligned}$$

it can be seen that $\mathcal{U}_{12}'$ is a subspace complementary to $\mathcal{U}_1 + \mathcal{U}_2$ in $\mathcal{V}$.

As a numerical example with $m = 3$ and $n = 4$ we offer the decomposition

$$\begin{aligned} \begin{pmatrix} 4 & 2 & 3 \\ 5 & 0 & 1 \\ 3 & 0 & 0 \\ 0 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix} + \begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \end{pmatrix} &+ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \\ &+ \begin{pmatrix} 0 & 0 & 0 \\ 2 & -1 & -1 \\ 1 & 0 & -1 \\ -3 & 1 & 2 \end{pmatrix} \end{aligned}$$

The elements on the right, whether they be called functions, vectors or matrices, are in the subspaces $\mathcal{U}_0$, $\mathcal{U}_1'$, $\mathcal{U}_2'$ and $\mathcal{U}_{12}'$ respectively.

Back now to the general case, if $\mathcal{U} = \mathcal{V}$ there is no complementary subspace other than the trivial subspace $\mathcal{O}$ consisting of the origin only. Likewise, if $\mathcal{U} = \mathcal{O}$, the only complementary subspace is $\mathcal{V}$ itself. In general, however, for a given subspace $\mathcal{U}$ of a vector space $\mathcal{V}$, there exist infinitely many subspaces that are complementary to $\mathcal{U}$ in $\mathcal{V}$. For example, if $\mathcal{V}$ is the real plane and $\mathcal{U}$ is the real axis, any line through the origin other than the horizontal axis can serve as a complement of $\mathcal{U}$. If we take the y-axis as a complement to $\mathcal{U}$, a point (x, y) in $\mathcal{V}$ has the familiar decomposition $(x, 0) + (0, y)$ as the sum of a vector in $\mathcal{U}$ and a vector in the complement. If, on the other hand, we choose the equi-angular line (x, x) as the complement to $\mathcal{U}$, the decomposition becomes

$$(x, y) = (x - y, 0) + (y, y).$$

In the additive effects example described above, the spaces $\mathcal{U}_1, \mathcal{U}_2$ and their span are prescribed by subject-matter coinsiderations. But the complementary space $\mathcal{U}_{12}'$ depends on a simple, but ultimately arbitrary, choice. A different choice of complementary space would yield a different decomposition.

0.2.7 *Quotient space*

Let $\mathcal{V}$ be a vector space. For a given subspace $\mathcal{U}$ there are, usually, infinitely many complementary subspaces. Without some additional structure, there is no natural way to select one of these complements for preferential treatment. There is, however, a natural way of patching together all these complementary subspaces to form a new vector space that, for all practical purposes, gives a unique complement of $\mathcal{U}$ in $\mathcal{V}$.

The idea is to manufacture a new vector space whose elements are not *points* in $\mathcal{V}$, but rather *subsets* of $\mathcal{V}$, in fact the cosets of $\mathcal{U}$. To do so, we need to define the vector space operations on cosets of $\mathcal{U}$. Addition of two cosets is defined by addition of the elements: this operation yields a new coset. Scalar multiplicaton is defined likewise, again yielding a new coset. Thus

$$\begin{aligned}(v_1 + \mathcal{U}) + (v_2 + \mathcal{U}) &= v_1 + v_2 + \mathcal{U} + \mathcal{U} = v_2 + v_2 + \mathcal{U},\\ \alpha(v_1 + \mathcal{U}) &= \alpha v_1 + \alpha\mathcal{U} = \alpha v_1 + \mathcal{U}\end{aligned}$$

for any scalar α. The cosets thus form a vector space, $\mathcal{U}$ itself being the zero element. This vector space is called the *quotient space*, and is denoted by the symbol $\mathcal{V}/\mathcal{U}$. The dimension of the quotient space is $\dim(\mathcal{V}) - \dim(\mathcal{U})$, the same as the dimension of any complementary space. (Halmos, 1958, sections 21, 22).

One way to visualize the quotient space, illustrated graphically in Fig 1.2, is as follows. Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be two vector spaces, both complementary to $\mathcal{U}$ in $\mathcal{V}$. Since $\mathcal{V} = \mathcal{U} + \mathcal{W}_1 = \mathcal{U} + \mathcal{W}_2$, any point v in $\mathcal{V}$ can be expressed as the sum of a pair of points in two ways:

$$v = u_1 + w_1 = u_2 + w_2,$$

where u_1, u_2 are in $\mathcal{U}$, $w_1 \in \mathcal{W}_1$, and $w_2 \in \mathcal{W}_2$. It follows then that $w_1 - w_2 = u_1 - u_2$ is a point in $\mathcal{U}$. Consequently $w_1 + \mathcal{U}$ and $w_2 + \mathcal{U}$ are two representations of the same coset, and $v + \mathcal{U}$ is a third representation of the same coset. The coset associated with a point v is thus the set of projections of $\mathcal{V}$ along $\mathcal{U}$ on to the various complementary spaces. The cosets of $\mathcal{U}$ partition the space into equivalence classes: two points in $\mathcal{V}$ are equivalent in this sense if their difference is in $\mathcal{U}$.

Example 5: Linear regression with offset. Consider the linear regression model

$$E(Y \mid x) = x_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where $x_0, \ldots, x_p$ are vectors in $\mathcal{V}$. If $\mathcal{U} = \text{span}\{x_1, \ldots, x_p\}$, this model asserts that the mean of Y lies on a particular coset of $\mathcal{U}$. Since x_0 has unit coeficient, the model specifies the coset $x_0 + \mathcal{U}$, but it does not otherwise identify the point on the coset.

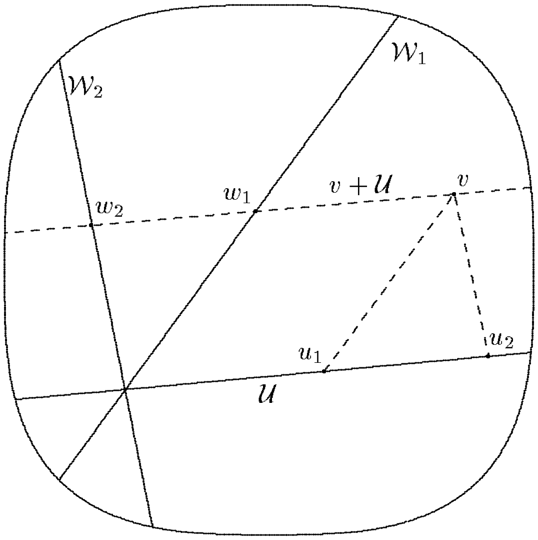

Figure 1.1. *Stone diagram illustrating a subspace $\mathcal{U}$ and two complementary spaces $\mathcal{W}_1$ and $\mathcal{W}_2$. The coset $v+\mathcal{U}$ is illustrated by a dashed line parallel to $\mathcal{U}$. The two decompositions $v = u_1 + w_1$ and $v = u_2 + w_2$ are also illustrated geometrically by the parallelogram rule for vector addition.*

Example 6: Residual vector in linear models. Consider the linear model, written in conventional matrix notation as

$$Y = X\beta + \epsilon,$$

where X is a given matrix whose columns span a subspace $\mathcal{X}$ of $\mathcal{V} = \mathcal{R}^n$, and ϵ is a random variable in $\mathcal{R}^n$, whose distribution depends on unknown parameters, θ. Let P be any projection matrix having range $\mathcal{X}$, and let $Q = I - P$ be the complementary, or residual, projection having null space $\mathcal{X}$. For example, we could take $P = X(X^TX)^{-1}X^T$ or $P' = X(X^TWX)X^TW$ for any symmetric positive definite matrix W. Whatever the choice of P, $R = QY$ is a residual vector. Since $R = Q\epsilon$, the residual vector has a distribution independent of β, but dependent on θ. Unlike the vector of fitted values whose definition depends on the choice of inner product or weight matrix W, all residual vectors are equivalent in the sense that they contain the same information. For example, the residual $R' = Q'Y$ can be obtained from $R = QY$ by the linear transformation $R' = Q'R$, and conversely $R = QR'$. Note that, although R' and R are vectors in different spaces, the difference $R' - R$ is necessarily a linear combination of the columns of X, i.e. an element of $\mathcal{X}$. For this reason, it is generally preferable to regard the residual vector as an element of the quotient space $\mathcal{V}/\mathcal{X}$ rather than an element of any particular subspace complementary to $\mathcal{X}$ in $\mathcal{V}$. By this device, we make precise the notion that all residual vectors are equivalent.

From a slightly different point of view, a residual projection on $\mathcal{V}$ having null space $\mathcal{X}$ transforms all points in a given coset to the single point where the range intersects the coset. Since all points in a coset are transformed to the same point in the range, the transformation on $\mathcal{V}$ can alternatively be considered as a transformation from the quotient space to itself. From this point of view, each residual projections on $\mathcal{V}$ acts as the identity on $\mathcal{V}/\mathcal{X}$, so all residual projections on $\mathcal{V}$ are identical on $\mathcal{V}/\mathcal{X}$.

In fact, it is technically not necessary to construct a projection matrix or form a residual vector at all: we may talk simply of the distribution of the random variable $Y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$, and use the resulting 'residual likelihood' for purposes of estimating θ. So far as its dependence on the data is concerned, the residual likelihood is a function on the space $\mathcal{V}/\mathcal{X}$, i.e. constant on the cosets of $\mathcal{X}$. We return to this example in section ?.?.

0.2.8 *Direct sum and tensor sum of vector spaces*

Given two vector spaces, we may form a new, and larger, vector space, in at least two different ways. The first of these sums is called the direct sum or Cartesian product, and is defined as follows.

DEFINITION: *Direct sum of vector spaces.* Let $\mathcal{U}$ and $\mathcal{V}$ be two vector spaces over the same field of scalars. The direct sum $\mathcal{W} = \mathcal{U} \oplus \mathcal{V}$ is a new vector space whose elements are all the ordered pairs (u, v) with $u \in \mathcal{U}$ and $v \in \mathcal{V}$. The linear operations are defined component-wise

$$\alpha(u_1, v_1) + \beta(u_2, v_2) = (\alpha u_1 + \beta u_2,\ \alpha v_1 + \beta v_2)$$

in the obvious way.

The subspaces consisting of the elements $(u, 0)$ and $(0, v)$ may be identified with $\mathcal{U}$ and $\mathcal{V}$ respectively. In this sense, although there is a slight technical innacuracy, we may speak of $\mathcal{U}$ and $\mathcal{V}$ as non-overlapping subspaces of $\mathcal{W}$, and $\mathcal{W} = \mathcal{U} + \mathcal{V}$, the span of the two complementary subspaces. The dimension of $\mathcal{U} \oplus \mathcal{V}$ is $\dim(\mathcal{U}) + \dim(\mathcal{V})$.

The Cartesian product of two sets U and V, in general not vector spaces, is the set of ordered pairs (u, v) with $u \in U$ and $v \in V$. So far as it goes, this coincides with the definition of direct sum. Unlike the direct sum of vector spaces, however, the term Cartesian product implies no linear operations.

It is worth pausing for a moment to examine the nature of the elements of the direct sum space just defined. The definition of direct sum is exactly what is required for constructing the real plane from the real line, or, by extension, $\mathcal{R}^n$ from n copies of the real line. Suppose, however, that the elements of $\mathcal{U}$ and $\mathcal{V}$ are real-valued functions on domains D_1 and D_2. According to the definition, the elements of $\mathcal{U} \oplus \mathcal{V}$ are ordered pairs of functions, in other words, functions from $D_1 \times D_2$ taking values in the plane. In a real sense, the elements of $\mathcal{U} \oplus \mathcal{V}$ are quite unlike the elements of the component spaces. In addition, even though the elements of $\mathcal{U}$ and $\mathcal{V}$ both take values in the reals, elements of the complementary spaces $(\mathcal{U}, 0)$ and $(0, \mathcal{V})$ do not take values in the same set. These considerations lead to a different kind of sum, what we shall call the tensor sum of two vector spaces.

Let 1 be the one-dimensional vector space of functions that are constant on D_2. Then $\mathcal{U} \otimes 1$ is the vector space of functions on $D_1 \times D_2$ of the form $f(x_1, x_2) = g(x_1)$ for $g \in \mathcal{U}$. Likewise, $1 \otimes \mathcal{V}$ is the vector space of functions on $D_1 \times D_2$ of the form $f(x_1, x_2) = h(x_2)$ for $h \in \mathcal{V}$. The span $\mathcal{U} \otimes 1 + 1 \otimes \mathcal{V}$, more commonly denoted in statistical work by $\mathcal{U} + \mathcal{V}$, is the vector space of functions on $D_1 \times D_2$ of the form $g(x_1) + h(x_2)$.

DEFINITION: *Tensor sum of two function spaces.* Let $\mathcal{U}$ and $\mathcal{V}$ be vector spaces of real-valued functions on domains D_1 and D_2 respectively. The tensor sum $\mathcal{U} + \mathcal{V}$ is the vector space of functions of the form $u + v$ with $u \in \mathcal{U}$ and $v \in \mathcal{V}$, acting on the domain $D_1 \times D_2$.

With the domain extended in this way, the tensor sum is simply the span of the two subspaces, so there is no need for a new symbol for this operation. The important condition is that the vector spaces $\mathcal{U}$ and $\mathcal{V}$ should be such that addition of elements from the two spaces is defined in a way that is compatible with addition in each of the component spaces. In other words, $\mathcal{U}$ and $\mathcal{V}$ must be vector spaces of functions. Otherwise the tensor sum is not defined.

The example that follows shows that, in general, $\mathcal{U}$ and $\mathcal{V}$ have a non-trivial intersection, so the dimension of the tensor sum is less that the sum of the dimensions of the component spaces. The dimension of the tensor sum is

$$\dim(\mathcal{U} + \mathcal{V}) = \dim(\mathcal{U}) + \dim(\mathcal{V}) - \dim(\mathcal{U} \cap \mathcal{V}).$$

Example 7: Direct sum versus tensor sum. Let $\mathcal{U}$ be the span of the polynomials $\{1, x, x^2\}$ on $[0, 1]$, and let $\mathcal{V}$ be the span of $\{1, y\}$ on $\mathcal{R}$. By the definition given above, the elements of $\mathcal{U} \oplus \mathcal{V}$

are ordered pairs of functions $(p_2(x), p_1(y))$ from the set $[0,1] \times \mathcal{R}$ into $\mathcal{R}^2$. Here, p_r denotes a polynomial of degree not exceeding r. A typical element of the direct sum $\mathcal{U} \oplus \mathcal{V}$ has the form

$$(a_0 + a_1 x + a_2 x^2,\, b_0 + b_1 y)$$

with arbitrary coefficients a_0, a_1, a_2, b_0, b_1. The space clearly has dimension 5, the sum of the dimensions of the component spaces.

The tensor sum, on the other hand, is the set of functions $p_2(x) + p_1(y)$ from $[0,1] \times \mathcal{R}$ into $\mathcal{R}$. This space is spanned by the functions $\{1, x, x^2, y\}$, and is thus of dimension 4. Note that functions in $\mathcal{U}$ are constant in y, and functions in $\mathcal{V}$ are constant in x. The intersection, $\mathcal{U} \cap \mathcal{V}$, the space of functions that are constant in both variables, has dimension 1.

Example 8: Randomized blocks model:. In a randomized blocks design, the expected response when treatment j is applied to block i is often modelled as an additive function of block and treatment effects. To put this in the language of vector spaces, let $\mathcal{U}$ be the space of functions on the block labels, and let $\mathcal{V}$ be the space of functions on the treatments. Then $\mathcal{U} + \mathcal{V}$ is the space of functions $u(i) + v(j)$ on the Cartesian product of the domains, $\{1, \ldots, m\} \times \{1, \ldots, n\}$. In the notation commonly used in statistics for specifying linear models, this space is denoted by $B + T$ where B is the factor for blocks and T is the factor for treatment. A factor in this context is the vector space of functions on a finite set of labels, and the operator '+' denotes the span or tensor sum.

0.2.9 *Dual basis*

The following theorem is taken from Halmos (1958, section 15).

THEOREM 1. Let $\mathcal{V}$ be an n-dimensional vector space and let $\{e_1, \ldots, e_n\}$ be a basis in $\mathcal{V}$. Then there is a uniquely determined basis $\{y^1, \ldots, y^n\}$ in $\mathcal{V}^*$ with the property that $y^i(e_j) = \delta^i_j$. ($\delta^i_j = 1$ for $i = j$ and zero otherwise is an extremely useful symbol known as Kronecker's delta.) Consequently the dual space is n-dimensional.

For a proof of this assertion, see Halmos (1958, section 15).

The basis $\{y^1, \ldots, y^n\}$, which is the set of coordinate projections on to $\{e_1, \ldots, e_n\}$, is called the *dual basis* in $\mathcal{V}^*$ of $\{e_1, \ldots, e_n\}$. If $\mathcal{V}$ is a real inner product space with inner product $\langle \cdot, \cdot \rangle$, and if G is the matrix with elements $G_{ij} = \langle e_i, e_j \rangle$, then the coordinate projections are given by

$$y^i(v) = \sum_j G^{ij} \langle e_j, v \rangle, \tag{1.2}$$

where G^{ij} is the (i,j) element of the matrix G^{-1}.

0.2.10 *Examples*

The following examples are of a non-statistical nature and are intended to clarify the distinction between $\mathcal{V}$ and $\mathcal{V}^*$.

Example 9: Harmonics on the circle. Let $\mathcal{V}$ be the space of zero and first-order harmonics on the unit circle. In symbols,

$$\mathcal{V} = \operatorname{span}\{e_0 = 1, e_1 = \cos\theta, e_2 = \sin\theta\}$$

for $0 \le \theta < 2\pi$. We can think of the elements of $\mathcal{V}$ either as periodic functions defined on an interval of the real line, or preferably, as continuous functions on the unit circle. Typical elements of $\mathcal{V}$ include functions such as $\sin(\theta + \pi/4)$ and $\cos^2(\theta/2)$. Every vector in $v \in \mathcal{V}$ can be expressed in a unique way as a linear combination of the basis vectors

$$v(\theta) = v^0 1 + v^1 \cos\theta + v^2 \sin\theta.$$

The following are examples of linear functionals on $\mathcal{V}$.

$$\begin{aligned} y^0(v) &= \frac{1}{2\pi}\int_0^{2\pi} v(\theta)\,d\theta, \\ y^1(v) &= \frac{1}{\pi}\int_0^{2\pi} v(\theta)\cos\theta\,d\theta, \\ y^2(v) &= \frac{1}{\pi}\int_0^{2\pi} v(\theta)\sin\theta\,d\theta, \\ \int_0^{2\pi} v(\theta)\cos^2(\theta/2)\,d\theta &= \pi y^0(v) + \pi y^1(v)/2. \end{aligned}$$

In terms of the basis vectors $\{e_0, e_1, e_2\}$, in $\mathcal{V}$, the value of the ith linear functional on the jth basis vector is $y^i(e_j) = \delta^i_j$. Thus $\{y^0, y^1, y^2\}$ are the dual basis vectors in $\mathcal{V}^*$. Consequently, $y^j(v) = v^j$, the jth component of v with respect to the given basis vectors. For instance, the fourth linear functional in the preceding list has components $(\pi, \pi/2, 0)$ with respect to the dual basis. In other words, an element v^* in $\mathcal{V}^*$ with coefficient vector (v^*_0, v^*_1, v^*_2) is a linear functional

$$\begin{aligned} v^*(v) &= v^*_0 y^0(v) + v^*_1 y^1(v) + v^*_2 y^2(v) = v^*_0 v^0 + v^*_1 v^1 + v^*_2 v^2 \\ &= \frac{1}{2\pi}\int_0^{2\pi} (v^*_0 + 2v^*_1\cos\theta + 2v^*_2\sin\theta)v(\theta)\,d\theta. \end{aligned}$$

It is common practice to use the word 'coefficient' for the components of a vector in the dual space. Each coefficient vector represents a different linear combination or linear functional.

Example 10: Contrast vectors. Let $\mathcal{V}$ be the set of contrast vectors in $\mathcal{R}^3$. An element of $\mathcal{V}$ is an ordered set of real numbers $(\bar{v}^1, \bar{v}^2, \bar{v}^3)$ whose sum is zero. We take as basis vectors the two linearly independent contrast vectors

$$e_1 = (1, -1, 0) \quad \text{and} \quad e_2 = (-1, 0, 1).$$

Any contrast vector $v = (\bar{v}^1, \bar{v}^2, \bar{v}^3)$ can be written as a linear combination of e_1 and e_2:

$$v = v^1 e_1 + v^2 e_2.$$

Using (1.2), the coefficients, which are linear functionals on $\mathcal{V}$, are given by

$$\begin{aligned} v^1(v) &= (\bar{v}^1 - 2\bar{v}^2 + \bar{v}^3)/3, \\ v^2(v) &= (-\bar{v}^1 - \bar{v}^2 + 2\bar{v}^3)/3. \end{aligned}$$

Since $v^i(e_j) = \delta^i_j$, $\{v^1, v^2\}$ is the dual basis in $\mathcal{V}^*$ of $\{e_1, e_2\}$. In other words, the point $\alpha = (\alpha_1, \alpha_2)$ in $\mathcal{V}^*$ denotes the linear functional

$$\alpha(v) = \alpha_1 v^1(v) + \alpha_2 v^2(v).$$

0.2.11 *Notation and terminology*

In statistical contexts, variables have names such as 'degrees Celsius,' 'kilometres east,' or 'log weight in kg.' These names form a basis in $\mathcal{V}$ called the incidence basis, denoted here by $\{e_1, \ldots, e_p\}$, where p is the number of variables. Thus, e_i is synonymous with the term 'name of variable i', and v^i is synonymous with 'value of variable i,' or 'value of v on variable i.' The complete description of $v \in \mathcal{V}$ is thus $v = v^i e_i$, or equivalently, $(v^1, \ldots, v^p)$, the basis vectors being understood from the context. Stone (1987) uses the notation v_i in place of e_i for the name of variable i. By a slight modification of notation in Halmos (1958), $[e, v_i]$ is used by Stone in place of our v^i for the value on variable i.

According to theorem 2, every linear functional on $\mathcal{V}$ can be written as a linear combination of the dual basis vectors or coordinate projections

$$\alpha(v) = \alpha_1 v^1 + \cdots + \alpha_n v^n = \alpha_i v^i,$$

with implicit summation over the repeated index. Each element in the dual space may therefore be identified with a coefficient vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ in $\mathcal{R}^n$. These are the components of the linear functional $\alpha \in \mathcal{V}^*$ with respect to the dual basis $\{v^1, \ldots, v^n\}$. Here, we have identified the dual space with the set of coefficient vectors α. Stone (1987) identifies $\mathcal{V}^*$ with the space of 'evaluators,' so that our linear functional $\alpha_i v^i$ coincides with his $[\alpha, v]$.

In tensor notation, the components of $v \in \mathcal{V}$ are identified by means of superscripts. The summation convention then permits (1.1) to be abbreviated to $v = v^i e_i$. The components of a linear functional $\alpha \in \mathcal{V}^*$ are denoted by α_i, a coefficient vector with subscripts. The linear functional $\alpha(v)$ may therefore be written in the form $\alpha(v) = \alpha_i v^i$. To achieve uniformity of notation and to blend in with the summation convention, components of vectors in $\mathcal{V}$ are indexed by superscripts: components of vectors in $\mathcal{V}^*$, also called coefficients, are indexed by subscripts. The distinction made in matrix notation between row vectors and column vectors is related to, but does not coincide uniformly with the distinction between component vectors in $\mathcal{V}$ and $\mathcal{V}^*$. For further discussion on this point, see section 1.4.4.

It may be helpful at this point to digress briefly on the subject of matrix notation. Consider for example, the identity matrix of order n, written in matrix notation as I or I_n. Such a matrix arising as a linear transformation from $\mathcal{R}^n$ to $\mathcal{R}^n$ is written in index notation using Kronecker's delta in the form δ^i_j. However, I_n could also arise as a covariance matrix of the random variables $(X^1, \ldots, X^n)$, in which case we would write $\delta^{ij} = \mathrm{cov}(X^i, X^j)$ with superscripts since all tensor equations must be balanced with respect to the positions of indices. Although δ^i_j is numerically equal to δ^{ij}, the two symbols are not equal as tensors because it is implicit in the notation that they are designed to operate in quite different ways. To say the same thing in a different way, tensor notation implies or suggests that these are the components of vectors in different vector spaces. The fact that the components are numerically equal is then irrelevant when the vectors in question do not lie in the same space.

0.3 Transformation rules

0.3.1 *Duality and dual spaces*

Let $\mathcal{V}$ be a vector space with basis $\{e_1, \ldots, e_n\}$. In the usual expression for $v \in \mathcal{V}$ as a linear combination of the basis vectors, we write $v = v^i e_i$. The notation exhibits a certain formal symmetry between the components v^i and the basis vectors e_i, but the interpretation, as usually given, is quite asymmetric: v^i is a 'scalar' whereas e_i is a 'vector' in $\mathcal{V}$. However, $\{v^1, \ldots, v^n\}$, as linear functionals, form a basis in $\mathcal{V}^*$, in fact the dual basis of $\{e_1, \ldots, e_n\}$. Thus, the expression $v^i e_i$ could equally well be interpreted as a point in $\mathcal{V}^*$ with components $(e_1, \ldots, e_n)$ relative to the dual basis. Tensor notation permits, and even encourages, this dual interpretation. In this dual sense, $v^i e_i$ is formally an invariant. It identifies a point in $\mathcal{V}$ independent of the choice of basis: it also identifies a point in $\mathcal{V}^*$ independent of the basis.

The following arguments, although they do not attempt to exploit the symmetry of the duality described above, explain why expressions such as $v^i e_i$, having no free index, are invariant.

0.3.2 *Change of basis vectors in $\mathcal{V}$*

Let $\mathcal{V}$ be a vector space of dimension n. The general idea is that we want to talk of a point in $\mathcal{V}$ in a way that is automatically independent of the choice of basis vectors. At the same time the methods employed must be sufficiently concrete to permit computation in the usual ways. Obviously the *representation* of $v \in \mathcal{V}$ in terms of its components $(v^1, \ldots, v^n)$ with respect to $\{e_1, \ldots, e_n\}$ does depend on the choice of basis vectors: if the basis vectors are altered, the *representation* will also be altered, but the point itself remains unchanged.

Introduce a new set of basis vectors $\{\bar{e}_1, \ldots, \bar{e}_n\}$. Each member of the new basis is a linear combination of the old basis with coefficients K as follows:

$$\bar{e}_r = K_r^j e_j. \tag{1.3}$$

The inverse transform giving the old basis vectors in terms of the new is also linear, but with coefficients L:

$$e_j = L_j^r \bar{e}_r.$$

On substituting one of these into the other we obtain

$$\bar{e}_r = K_r^j L_j^s \bar{e}_s, \quad \text{and} \quad e_j = K_r^i L_j^r e_i.$$

In other words K and L are matrix inverses. We write $K_r^j L_j^s = \delta_r^s$, $K_r^i L_j^r = \delta_j^i$, where δ_j^i is the identity matrix, or Kronecker's delta. In contrast to matrix notation, the order of symbols in a tensor formula is irrelevant because scalar multiplication is commutative.

Let $(v^1, \ldots, v^n)$ be the components of v with respect to the basis $\{e_1, \ldots, e_n\}$, and let $(\bar{v}^1, \ldots, \bar{v}^n)$ be the components of the same point with respect to $\{\bar{e}_1, \ldots, \bar{e}_n\}$. Then we have

$$v = v^i e_i = v^i L_i^r \bar{e}_r = \bar{v}^r \bar{e}_r.$$

In other words, under the change of basis (1.3) with matrix K, the components of v are transformed linearly from v^i to $\bar{v}^i$ by

$$\bar{v}^r = L_i^r v^i. \tag{1.4}$$

This linear transformation involves the matrix *inverse* of K in (1.3). It is this transformation that is implied when we say that $(v^1, \ldots, v^n)$ is a *contravariant vector* or *contravariant tensor* of order one.

The dual space is associated with coefficient vectors whose components are $(\alpha_1, \ldots, \alpha_n)$ with respect to the dual basis in $\mathcal{V}^*$ of $\{e_1, \ldots, e_n\}$. When the basis vectors in $\mathcal{V}$ are transformed by (1.3) there is a corresponding change in the dual basis vectors from v^i to $\bar{v}^i$ as in (1.4) or (1.2). By the argument given above, the components $(\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ with respect to the new basis vectors are given by

$$\bar{\alpha}_r = K_r^j \alpha_j, \tag{1.5}$$

involving the matrix inverse of L in (1.4). The net effect of these two changes is that the value of any linear functional such as $\alpha_i v^i$ is unchanged:

$$\alpha_i v^i = \alpha_i K_r^i \bar{v}^r = \bar{\alpha}_i \bar{v}^i.$$

It is transformation (1.5) that is implied when we say that the coefficient vector $(\alpha_1, \ldots, \alpha_n)$, or simply α_i, is a *covariant vector* or *covariant tensor* of order one.

A tensor is thus a component vector used to identify points in a vector space or in the dual space. Implicit in the tensorial recipe is the notion that the basis vectors are arbitrary up to linear transformation. In order that the description v^i should represent the same point in $\mathcal{V}$ when a new basis is selected according to (1.3), it is necessary for the description to change linearly by rule (1.4). Likewise, in the dual space of coefficient vectors, in order that the vector α should represent a specific linear functional independent of the choice of basis, it is necessary that the components α_i undergo transformation by rule (1.5) when a new basis is introduced in $\mathcal{V}$. These transformation rules are implicit in tensor notation. One advantage of the notation is that it permits us to use the basis vectors for computational purposes. At the same time, the explicit recognition that the basis vectors are arbitrary up to linear transformation effectively means that all calculations are independent of the choice of basis. In particular, all scalars are invariant, i.e. independent of the choice of basis.

0.3.3 *Group action*

An alternative description, leading essentially to the same result, can be phrased in terms of group action. Let $\mathcal{G}$ be a group acting on a real n-dimensional vector space $\mathcal{V}$. That is to say that for each $g \in \mathcal{G}$ and $v \in \mathcal{V}$, gv is a point in $\mathcal{V}$. Further, $g_2(g_1 v)$ is the same point as $(g_2 g_1)v$. The elements of the group often have a physical description or interpretation such as 'transposition,' 'cyclic permutation,' 'rotation,' 'reflection,' 'dilation,' 'isotropic expansion' and so on. Such a group often, but not always, has a representation in terms of $n \times n$ matrices acting on $\mathcal{R}^n$ in the sense of matrix multiplication. It is an essential part of any representation that the matrix representation of the group product $g = g_2 g_1$ be the same as the matrix product of the representations of g_2 and g_1. Assuming that $\mathcal{G}$ has such a matrix representation, we may think of g as a matrix, v as a column vector in $\mathcal{V} = \mathcal{R}^n$, and gv representing matrix multiplication. The action of g is denoted by the symbol $v \mapsto gv$ for each $v \in \mathcal{V}$.

Any action of the group on $\mathcal{V}$ is accompanied by an 'equal and opposite reaction' on $\mathcal{V}^*$. In the space of linear functionals on $\mathcal{V}$, the components of the linear functional α with respect to the dual basis are denoted by $(\alpha_1, \ldots, \alpha_n)$, i.e. a vector in $\mathcal{R}^n$. It is convenient in our matrix representation to take α to be a row vector. This helps to distinguish $\mathcal{V}$ from $\mathcal{V}^*$, both being isomorphic to $\mathcal{R}^n$. The actions of g on $\mathcal{V}$ and $\mathcal{V}^*$ are given by

$$g : v \mapsto gv; \quad \alpha \mapsto \alpha g^{-1}.$$

In tensor notation with superscripts indexing rows and subscripts indexing columns,

$$g : v^i \mapsto g^i_r v^r; \quad \alpha_i \mapsto h^r_i \alpha_r, \tag{1.6}$$

where g^i_r are the elements of the matrix g and h^r_i are the elements of g^{-1}.

In this representation, a linear functional on $\mathcal{V}$ is given by the product αv, or in tensor notation by $\alpha_i v^i$. It is immediately apparent from the definition of group action that linear functionals are invariant under the group. With respect to the action of the group $\mathcal{G}$, component vectors in $\mathcal{V}$ are said to be contravariant; component vectors in $\mathcal{V}^*$ are said to be covariant.

In this description there is no suggestion of a change of basis, so v and gv represent different points in $\mathcal{V}$. The emphasis is thus quite different from that in the previous section where v^i and $\bar{v}^i$ were taken to be alternative representations of the same point in $\mathcal{V}$, but with respect to different bases. The primary emphasis in both cases is on scalars, i.e. invariants either under the action of $\mathcal{G}$, or under change of basis. If $\mathcal{G}$ is the group of all linear transformations on $\mathcal{R}^n$, the matrices g and h in (1.6) may be identified with L and K in (1.3) and (1.5) respectively. In this context at least, group action and change of basis are alternative ways of motivating the tensor transformation rules (1.6), or (1.4), (1.5). Although the motivations are very different, the aims are identical and the net result, the tensor transformation rules, is the same in both cases.

The present description in terms of group action is a little more flexible, however, because we may tailor the group to suit the circumstances. In some physical contexts the orthogonal group may be appropriate, particularly where 'real space' $\mathcal{R}^3$ is involved. The matrices appearing in (1.6) are then required to be orthogonal, so that scalars are invariant under rotation and reflection, but not necessarily under dilation or linear transformation. Where moments and cumulants of random variables are involved, the affine group of transformations is usually the appropriate choice for $\mathcal{G}$. In other contexts, the group of lower triangular matrices may be the appropriate choice.

0.3.4 *Adjoint transformation*

Let A be a linear transformation on a vector space $\mathcal{V}$, i.e. $A\colon \mathcal{V} \mapsto \mathcal{V}$. For any linear functional α, there exists an associated linear functional $A'\alpha$ such that the numerical value of α at Ax is equal to the value of $A'\alpha$ at x. Further, A' is a linear transformation on $\mathcal{V}^*$, called the adjoint linear transformation of A (Halmos, 1958, section 44).

Different notational conventions express this notion in different ways. Halmos writes $[x, \alpha]$ to denote the value of the linear functional α at x where others might write $\alpha(x)$. Then the definition is $[Ax, \alpha] = [x, A'\alpha]$, and the assertion is that A' is a linear transformation on $\mathcal{V}^*$.

In one form of matrix notation, elements of $\mathcal{V}$ are column vectors and elements of $\mathcal{V}^*$ are row vectors. The linear transformation A on $\mathcal{V}$ is represented by matrix multiplication Ax, and $[x, \alpha]$ is the scalar αx. Then the value of α at Ax is αAx, which is the same as αA evaluated at x. The adjoint linear transformation $A'\alpha$ on $\mathcal{V}^*$ is thus αA, or matrix post-multiplication.

Using index notation, the components of Ax are $a^i_r x^r$ and the value of α at x is $\alpha_i x^i$. Then $[Ax, \alpha] = \alpha_i a^i_r x^r = [x, A'\alpha]$. The adjoint operation $A'\alpha$ on $\mathcal{V}^*$ is thus $a^i_r \alpha_i$. So the adjoint operation is an immediate and automatic consequence of this notational convention. No transposition of indices is involved.

0.3.5 *Examples*

The examples that follow are intended to illustrate the variety of objects that can be considered to be tensors.

Example 11: Vector product in $\mathcal{R}^3$. Let $\mathcal{V} = \mathcal{R}^3$ represent real space in the Newtonian sense. For any two elements $a = (a^1, a^2, a^3)$ and $b = (b^1, b^2, b^3)$ in $\mathcal{V}$, define the vector product or cross product $c = a \times b$ by its components

$$\begin{aligned} c^1 &= a^2b^3 - a^3b^2, \\ c^2 &= a^3b^1 - a^1b^3, \\ c^3 &= a^1b^2 - a^2b^1. \end{aligned}$$

Certainly c is an ordered triple of real numbers. By some definitions, that property alone would qualify c for membership of $\mathcal{V}$, but our requirements are more stringent than this. Use of superscripts suggests that c is a contravariant vector in $\mathcal{V}$, but this is far from obvious from the definition. The question at issue, then, is whether or not c identifies the same point in $\mathcal{V}$ when we select a new basis for $\mathcal{V}$.

Consider changing basis vectors from e_r to $\bar{e}_r$ as in (1.3). Components with respect to this new basis are denoted by $\bar{a}^i$, $\bar{b}^i$ and $\bar{c}^i$. Since a and b are the components of vectors in $\mathcal{V}$, they transform by rule (1.4) for contravariant vectors. Hence, from the definition of the cross product, the components of c with respect to the new basis are

$$\begin{aligned} \bar{c}^1 &= \bar{a}^2\bar{b}^3 - \bar{a}^3\bar{b}^2 \\ \bar{c}^2 &= \bar{a}^3\bar{b}^1 - \bar{a}^1\bar{b}^3, \\ \bar{c}^3 &= \bar{a}^1\bar{b}^2 - \bar{a}^2\bar{b}^1. \end{aligned}$$

Simplification of the expression for $\bar{c}^1$ gives

$$\bar{c}^1 = L_i^2 a^i L_j^3 b^j - L_j^3 a^j L_i^2 b^i = L_i^2 L_j^3 (a^i b^j - a^j b^i),$$

showing that $\bar{c}^1$ is a linear function of $\{c^1, c^2, c^3\}$. In order for c^i to be a tensor or vector in $\mathcal{V}$ its components must transform to $\bar{c}^i = L_r^i c^r$. The above expression, which is quadratic in L, shows that this is not the case for general L, so the cross product is not a contravariant vector in the sense of section 1.3.

A question of interest related to the discussion of group action is the following. For what group of transformations, or set of matrices L, does c obey the transformation law of a contravariant vector? We now investigate this point in detail.

Further simplification of the expression for $\bar{c}^1$ gives

$$\bar{c}^1 = (L_1^2 L_2^3 - L_2^2 L_1^3)c^3 + (L_3^2 L_1^3 - L_1^2 L_3^3)c^2 + (L_2^2 L_3^3 - L_3^2 L_2^3)c^1.$$

There are similar expressions for $\bar{c}^2$ and $\bar{c}^3$. Now, if L is 3×3 and orthogonal, it can be shown that L_j^i is equal to its co-factor multiplied by $|L|$. Thus, if L is orthogonal, we have

$$\bar{c}^i = |L| \, L_r^i c^r.$$

In other words, $c = a \times b$ is a contravariant vector in $\mathcal{V}$, but only with respect to O^+, the group of rotations in $\mathcal{R}^3$. Some authors, for example Jeffreys (1952), and Synge and Schild (1978, p. 128), use the term 'Cartesian tensor' to mean a tensor in this restricted sense.

As will be shown later, the skew-symmetric array $c^{ij} = a^i b^j - a^j b^i$ forms a contravariant tensor of order two quite generally. This array has $n(n-1)/2$ independent non-zero elements. It so happens that if $n = 3$, the three independent elements can be dressed up to masquerade as a vector, at least to the limited extent described above. For many purposes, rotations, angular momentum and similar physical concepts are best thought of as skew-symmetric second-order tensors rather than as vectors. The term 'axial vector' (as opposed to 'polar vector') is sometimes used in physics to distinguish 'vectors' such as $a \times b$ from ordinary vectors in $\mathcal{R}^3$. For example, an electric field is an ordinary (polar) vector in $\mathcal{R}^3$, but a magnetic field is an axial vector (skew-symmetric tensor).

Example 12: Semi-invariants and the affine group. Let $\mathcal{G}$ be the group of $(n+1) \times (n+1)$ matrices of the form

$$g = \begin{pmatrix} 1 & 0 \\ g_0 & g_1 \end{pmatrix},$$

in which $g_0 \in \mathcal{R}^n$, and g_1 is $n \times n$ and non-singular. Let $\mathcal{F}$ be the set of vectors in $\mathcal{R}^{n+1}$ of the form $(1, v)$ with $v \in \mathcal{R}^n$. The action of $\mathcal{G}$ on $\mathcal{F}$

$$\begin{pmatrix} 1 & 0 \\ g_0 & g_1 \end{pmatrix} \begin{pmatrix} 1 \\ v \end{pmatrix} = \begin{pmatrix} 1 \\ g_0 + g_1 v \end{pmatrix} = \begin{pmatrix} 1 \\ \bar{v} \end{pmatrix}.$$

is isomorphic to the action of the general affine group, $GA(\mathcal{R}^n)$ acting on points v in $\mathcal{V} = \mathcal{R}^n$:

$$\bar{v} = g_0 + g_1 v. \tag{1.7}$$

The action of $\mathcal{G}$ on $\mathcal{F}$ is a representation of $GA(\mathcal{R}^n)$ as a sub-group of $GL(\mathcal{R}^{n+1})$, the general linear group acting on $\mathcal{R}^{n+1}$. The set $\mathcal{F}$ is called an affine space in $\mathcal{R}^{n+1}$: it is a translation of its tangent space. The same terminology is used for $\mathcal{V}$, a space subject to affine re-parameterization, even though $\mathcal{V}$ is not obviously a translation of its tangent space in the geometrical sense. Many physical variables such as time, position and velocity, which have no well-defined origin, are subject to affine re-scaling in this sense. The same is true of everyday temperatures recorded in °C or °F.

Now, F is not itself a vector space because it does not include the origin, but it does have a tangent space given by

$$T = \{x_1 - x_2 : x_1, x_2 \in F\} = \{(0, v_1) - (0, v_2) : v_1, v_2 \in \mathcal{V}\}.$$

Evidently, $\mathcal{T}$ is a vector space isomorphic with $\mathcal{R}^n$. In fact, $\mathcal{T}$ may be identified with the set of differences $v_1 - v_2$ with $v_1, v_2 \in \mathcal{V}$. It might seem that $\mathcal{T}$ is the same as $\mathcal{V}$, but it is critically important to distinguish the two for the following reason. The action of the group on a vector $dv \in \mathcal{T}$ is given by

$$d\bar{v} = (\bar{v}_1 - \bar{v}_2) = g_1(v_1 - v_2) = g_1 dv,$$

which, in contrast to (1.7), is linear. If we re-write (1.7) using index notation in the form

$$\bar{v}^i = g^i + g^i_r v^r,$$

we obtain the tensor transformation rule for vectors dv^i in $\mathcal{T}$:

$$d\bar{v}^i = g^i_r dv^r.$$

In other words, although vectors in $\mathcal{V}$ are affected by the change of origin, vectors in the tangent space of $\mathcal{V}$ are unaffected. For that reason, vectors in $\mathcal{T}$ or tensors in $\mathcal{T}^{\otimes k}$ or in $\mathcal{T}^{*\otimes k}$ are sometimes said to be semi-invariant. Cumulants of order two or more are the principal examples of such tensors in statistics.

In the old days when temperature scales were taught in primary schools, the more pedantic teachers, no doubt appreciating the difference between an affine space and its tangent space, insisted on drawing a distinction between degrees centigrade (°C), indicating a temperature, and centigrade degrees (C°), indicating a rise or fall in temperature. While most students eventually came to recognize the distinction, not all of them at the time considered it to be a distinction worth making.

Example 13: Tangent spaces and log likelihood derivatives. Let $\theta = (\theta^1, \ldots, \theta^n)$ be a point in an n-dimensional set or manifold, Θ, and let $l(\theta)$ be a differentiable function defined on Θ. In statistical contexts, θ is a parameter and $l(\theta)$ is a log likelihood function. There is no suggestion that Θ is a vector space, even though the term 'parameter space' is commonly used. By a re-parameterization $\theta \mapsto \phi$ is meant a new coordinate system on Θ, i.e. a set of functions $(\phi^1(\theta), \ldots, \phi^n(\theta))$ whose Jacobian is non-zero. It is usually best to think of the parameter space as a static entity transcending its description by coordinates: a coordinate system then functions like a basis in the sense that it provides a systematic method for identifying points in Θ. The set of such re-parameterizations evidently constitutes a group. The derivative matrix of the transformation is

$$\phi^i_r = \partial\phi^i/\partial\theta^r.$$

The inverse matrix is $\theta^i_r = \partial\theta^i/\partial\phi^r$. Since the transformation is not linear, these matrices are not constant over Θ.

Let $\mathcal{T}_\theta$ be the tangent space at θ, i.e. formally the space spanned by the differential operators $\{\partial/\partial\theta^1, \ldots, \partial/\partial\theta^n\}$ at a specific point θ. To form a mental image of this, it may be helpful to imagine Θ as a two-dimensional surface $x(\theta)$ in $\mathcal{R}^3$ parameterized by (θ^1, θ^2). Then $\partial x/\partial\theta^1, \partial x/\partial\theta^2$ are vectors in $\mathcal{R}^3$, parallel to the tangent plane at θ. The components of such a vector in $\mathcal{T}_\theta$ are denoted by $(d\theta^1, \ldots, d\theta^n)$. Consider by way of example the spherical polar representation of the unit sphere in $\mathcal{R}^3$,

$$x(\theta) = (\cos\theta^1, \sin\theta^1\cos\theta^2, \sin\theta^1\sin\theta^2)$$

for $0 < \theta^1 < \pi$, $-\pi < \theta^2 \le \pi$. The derivative vectors at $\theta^1 = \theta^2 = \pi/4$ are $e_1 = (-1/\sqrt{2}, 1/2, 1/2)$ and $e_2 = (0, -1/2, 1/2)$. These are the basis vectors in $\mathcal{T}_{\pi/4}$, the tangent space at $\theta = (\pi/4, \pi/4)$. A point in the tangent space is identified by its components relative to this basis, i.e. the components $(d\theta^1, d\theta^2)$ signify the point $e_1 d\theta^1 + e_2 d\theta^2$ in the tangent space. Note that by definition of a space, the tangent space must include the origin, or a point called zero. Thus, $x(\theta)$ is not ordinarily a point in the tangent space. However, in order to enhance the image it is helpful mentally to transport the origin to $x(\theta)$, so that the algebraic definition of tangent space is made to coincide with the geometric notion of tangent plane, which is the set of points $x(\theta) + \lambda^1 e_1 + \lambda^2 e_2$ with $(\lambda^1, \lambda^2) \in \mathcal{R}^2$.

In the dual space of linear functionals on $\mathcal{T}_\theta$, $\alpha = (\alpha_1, \ldots, \alpha_n)$ is identified with the linear functional $\alpha_i d\theta^i$ on $\mathcal{T}_\theta$. Under the action of the group on Θ, basis vectors in $\mathcal{T}_\theta$ are transformed linearly by

$$\frac{\partial}{\partial\theta^i} \mapsto \frac{\partial}{\partial\phi^i} = \frac{\partial}{\partial\theta^r}\theta_i^r.$$

Hence the components transform inversely by the rule

$$d\theta^i \mapsto d\phi^i = \phi_r^i d\theta^r.$$

To draw a parallel with section 1.4.2, $\theta_i^r = K_i^r$ is the matrix of the linear transformation on the space $\mathcal{T}_\theta$, so log-likelihood derivatives transform linearly as covariant tensors as in (1.5). Component vectors $d\theta^i$ are transformed linearly by the inverse matrix $\phi_r^i = L_r^i$, in the manner of contravariant tensors. The only major departure here is that the matrix of the transformation varies from point to point in the parameter space.

0.4 Tensor products

0.4.1 *Examples*

The formation of tensor products is a method for putting together two or more vector spaces to form a new vector space called the tensor product space. The process is much like formal multiplication in the sense that the new space is linear in each of the component spaces. Some familiar examples of tensor product spaces will serve as a guide to motivate the definition.

Example 14: Tensor product of polynomials. Let $\mathcal{V}_1$ be the vector space of affine functions of one variable, x_1. This is a two-dimensional space spanned by the functions $\{1, x_1\}$ on $\mathcal{R}$. Likewise, let $\mathcal{V}_2$ be another vector space of quadratic functions of the variable x_2. Thus $\mathcal{V}_2 = \text{span}\{1, x_2, x_2^2\}$. Then $\mathcal{V}_1 \otimes \mathcal{V}_2$ is the space of functions on the plane spanned by all the formal products

$$\mathcal{V}_1 \otimes \mathcal{V}_2 = \text{span}\{1, x_1, x_2, x_1x_2, x_2^2, x_1x_2^2\}.$$

There is a minor technical distinction to be drawn here between the function $x_2 \in \mathcal{V}_2$, a function of one variable, and $x_2 \in \mathcal{V}_1 \otimes \mathcal{V}_2$, a function on the plane which happens to depend only on one of the two components. On balance it seems unnecessary and unhelpful to make such distinctions a part of the notation.

Note that the dimension of the tensor product space is the product of the dimensions of $\mathcal{V}_1$ and $\mathcal{V}_2$.

Example 15: Tensor product of Heaviside functions. Let $H(x)$, the unit Heaviside function, be the indicator function for $x \geq 0$. Define

$$\mathcal{V}_1 = \text{span}\{1, H(x_1)\}, \quad \mathcal{V}_2 = \text{span}\{1, H(x_2)\}.$$

Then $\mathcal{V}_1 \otimes \mathcal{V}_2$, the space of formal products as functions on the plane, is the space of functions that are constant on quadrants. In symbols,

$$\mathcal{V}_1 \otimes \mathcal{V}_2 = \text{span}\{1, H(x_1), H(x_2), H(x_1)H(x_2)\}.$$

Contrast the direct sum space $\mathcal{V}_1 \oplus \mathcal{V}_2$, which is

$$\mathcal{V}_1 \oplus \mathcal{V}_2 = \mathrm{span}\{1, H(x_1), H(x_2)\}.$$

In statistical terminology, if binary factor A is associated with x_1, and factor B with x_2, positive values denoting high levels, then $\mathcal{V}_1 \oplus \mathcal{V}_2$ is the main-effects model $A+B$, whereas $\mathcal{V}_1 \otimes \mathcal{V}_2$ includes the interaction and is written as $A{*}B$ or $A.B$. By one convention, $A{*}B$ means $\mathcal{V}_1 \otimes \mathcal{V}_2$ with the basis given above: $A.B$ denotes the same space with quadrant indicator functions as basis vectors.

Example 16: Tensor product of contrast vectors. Let $\mathcal{V}$ be the space of contrast vectors in $\mathcal{R}^3$ spanned by

$$e_1 = (1, 0, -1), \quad \text{and} \quad e_2 = (0, 1, -1).$$

Then the basis vectors of $\mathcal{V} \otimes \mathcal{V} = \mathcal{V}^{\otimes 2}$ are the formal products $e_{ij} = e_i \otimes e_j$. If we regard e_i as a column vector it is simplest to portray $e_{ij} = e_i e_j^T$ as a rank-one matrix.

$$e_{11} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad e_{12} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

$$e_{21} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}, \qquad e_{22} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Note that e_{11}, e_{22} and $e_{12} + e_{21}$ are symmetric matrices whereas $e_{12} - e_{21}$ is skew-symmetric. The symmetric matrices span a subspace of $\mathcal{V}^2$ called $\mathrm{sym}_2(\mathcal{V})$: the skew-symmetric matrices span a subspace called $\mathrm{alt}_2(\mathcal{V})$. For most statistical purposes, only the symmetric part is relevant.

The same space and subspaces, though with different basis vectors, would be obtained were we to use a different basis for $\mathcal{V}$.

0.4.2 *Definition of tensor product*

Probably the easiest way to envisage a tensor product of two vectors $v_1 \in \mathcal{V}_1$ and $v_2 \in \mathcal{V}_2$ is to think of vectors as functions of one variable. A vector $(\xi_1, \ldots, \xi_n) \in \mathcal{R}^n$ can be expressed as a real-valued function $\xi(\cdot)$ defined on the first n positive integers, whose value at $x = i$ is $\xi(i) = \xi_i$. The tensor product $v_1 \otimes v_2$ is then a function on the plane whose value at (x_1, x_2) is equal to the product $v_1(x_1)v_2(x_2)$.

Definition: Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. For each $v_1, v_2 \in \mathcal{V}$ and $w_1, w_2 \in \mathcal{W}$, the symbol $v \otimes w$ is defined to be the bi-linear formal product satisfying

$$\begin{aligned}(\alpha_1 v_1 + \alpha_2 v_2) \otimes w &= \alpha_1 v_1 \otimes w + \alpha_2 v_2 \otimes w, \\ v \otimes (\alpha_1 w_1 + \alpha_2 w_2) &= \alpha_1 v \otimes w_1 + \alpha_2 v \otimes w_2.\end{aligned}$$

for all scalars α_1, α_2. Then $\mathcal{V} \otimes \mathcal{W}$ is the vector space of all linear combinations $\sum \alpha_i v_i \otimes w_i$ for $v_i \in \mathcal{V}$ and $w_i \in \mathcal{W}$. If $\{e_1, \ldots, e_m\}$ is a basis in $\mathcal{V}$ and $\{e'_1, \ldots, e'_n\}$ is a basis in $\mathcal{W}$ then the set of mn symbols $\{e_{ij} = e_i \otimes e'_j\}$ is a basis, called a product basis, in $\mathcal{V} \otimes \mathcal{W}$.

0.4.3 *Bi-linear functionals*

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. A bi-linear functional $y(v, w)$ on $\mathcal{V} \times \mathcal{W}$ is a scalar-valued function, defined for each $v \in \mathcal{V}$ and $w \in \mathcal{W}$, that is linear in v for fixed w, and linear in w for fixed v. In symbols

$$\begin{aligned} y(\alpha_1 v_1 + \alpha_2 v_2, w) &= \alpha_1 y(v_1, w) + \alpha_2 y(v_2, w), \\ y(v, \alpha_1 w_1 + \alpha_2 w_2) &= \alpha_1 y(v, w_1) + \alpha_2 y(v, w_2) \end{aligned}$$

for each $v, v_1, v_2 \in \mathcal{V}$, for $w, w_1, w_2 \in \mathcal{W}$, and for all scalars α_1, α_2. It is clear from the definition that if $y_1(\cdot, \cdot)$ and $y_2(\cdot, \cdot)$ are bi-linear functionals, so also is the linear combination $\alpha_1 y_1 + \alpha_2 y_2$. In other words, the set of bi-linear functionals on $\mathcal{V} \times \mathcal{W}$ is a vector space. It is possible now to give an alternative definition of the tensor product space $\mathcal{V} \otimes \mathcal{W}$ as follows:

Definition: The tensor product of two vector spaces $\mathcal{V}$ and $\mathcal{W}$ is the dual of the space of bi-linear functionals on $\mathcal{V} \times \mathcal{W}$.

If $(v^1, \ldots, v^m)$ are the components of $v \in \mathcal{V}$ and $(w^1, \ldots, w^n)$ are the components of $w \in \mathcal{W}$, then v^i is an example of a linear functional on $\mathcal{V}$ and w^i is an example of a linear functional on $\mathcal{W}$. In fact, these are the dual bases vectors in $\mathcal{V}^*$ and $\mathcal{W}^*$ respectively. Consequently, the product $(v, w) \mapsto v^i w^j$ is an example of a bi-linear functional on $\mathcal{V} \times \mathcal{W}$. In fact these mn bi-linear functionals constitute the dual basis in $\mathcal{V}^* \otimes \mathcal{W}^*$.

0.4.4 *Components of a tensor*

Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be vector spaces with bases $\{e_1, \ldots, e_m\}$ and $\{e'_1, \ldots, e'_n\}$ respectively. Then $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ is a vector space of dimension mn with basis vectors

$$\{e_{ij} = e_i \otimes e'_j; \qquad i = 1, \ldots, m; j = 1, \ldots, n\}.$$

Each point $w \in \mathcal{W}$ has a unique representation as a linear combination of the basis vectors. We write

$$w = w^{ij} e_i \otimes e'_j = w^{ij} e_{ij},$$

making use of the summation convention in which the range of summation for each index is to be understood from the context. Then w^{ij} are the components of the tensor w. It is an obvious abuse of terminology to say that w^{ij}, meaning the array of real numbers, is a (contravariant) tensor, but such abuses are not uncommon.

0.4.5 *Tensor transformation rules*

Let $\{e_1, \ldots, e_m\}$ be a basis in $\mathcal{V}_1$, $\{e'_1, \ldots, e'_n\}$ be a basis in $\mathcal{V}_2$, and let w^{ij} be the components of a tensor or vector in $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ with respect to the product basis $\{e_{ij} = e_1 \otimes e'_j\}$. In symbols,

$$w = w^{ij} e_{ij}.$$

Consider now the effect of introducing a new basis $\{\bar{e}_i\}$ in $\mathcal{V}_1$ and $\{\bar{e}'\}$ in $\mathcal{V}_2$. We write

$$\bar{e}_r = K^i_r e_i, \qquad \bar{e}'_r = K'^i_r e'_i,$$

defining the new basis vectors in terms of the old. The product basis in $\mathcal{W}$ is thereby transformed from $\{e_{ij}\}$ to $\{\bar{e}_{ij}\}$ given by

$$\bar{e}_{rs} = K^i_r K'^j_s e_i \otimes e'_j = K^i_r K'^j_s e_{ij}. \tag{1.8}$$

Accordingly, if we denote by $\bar{w}^{ij}$ the components of w with respect to the new basis, we have

$$w = \bar{w}^{rs} \bar{e}_{rs} = \bar{w}^{rs} K^i_r K'^j_s e_{ij} = w^{ij} e_{ij}.$$

Thus the components of w with respect to the new basis are

$$\bar{w}^{rs} = L^r_i L'^s_j w^{ij} \tag{1.9}$$

in which L is the matrix inverse of K, and L' is the inverse of K'. This is the tensor transformation rule for the components of a contravariant tensor in $\mathcal{V}_1 \otimes \mathcal{V}_2$.

Components of vectors in the dual space $\mathcal{W}^*$ of linear functionals on $\mathcal{W}$ are denoted by means of subscripts. Their tensor transformation rule is given by (1.8).

In most of the applications that we have in mind only one space $\mathcal{V}$, and its tensor products $\mathcal{V}^{\otimes k}$, is involved. Thus $\mathcal{V}_1 = \mathcal{V}_2$, $m = n$, $K = K'$ and $L = L'$.

The extension of the tensor transformation rules (1.8)and (1.9) for tensors of higher order should be immediately apparent.

0.4.6 *Tensor multiplication and contraction*

Operationally, a tensor is an array of numbers or functionals that obeys a particular multi-linear transformation rule under a change of basis in the relevant space or spaces. Suppose for simplicity that a single space $\mathcal{V}$ is involved, and that the change of basis vectors is given by

$$\bar{e}_i = a^r_i e_r \quad \text{or} \quad e_r = b^i_r \bar{e}_i,$$

where b^i_r is the matrix inverse of a^r_i. These matrices belong to some group relevant to the problem, usually the full linear group, but in appropriate circumstances, the orthogonal group or the symmetric group. An array $\omega = \omega^{ijk}$ whose values in the new coordinate system are given by

$$\bar{\omega}^{ijk} = b^i_r b^j_s b^k_t \omega^{rst}$$

is said to be a contravariant tensor of order three. Technically, ω^{ijk} are the components of a point or vector in $\mathcal{V}^{\otimes 3}$. Likewise, an array γ_{ij} that transforms by the rule

$$\bar{\gamma}_{ij} = a^r_i a^s_j \gamma_{rs}$$

is said to be covariant of order two. The direct product of two tensors is also a tensor. For example,

$$\psi^{ijk}_{rs} = \omega^{ijk} \gamma_{rs}$$

is a tensor of covariant order two and contravariant order three. It obeys the transformation rule for a mixed tensor.

Direct multiplication results in a tensor of increased order. In general, the indices may have different ranges and may refer to quite unrelated spaces. However, when a superscript and a subscript refer to components in a space and its dual, they necessarily have the same range. In such cases it is legitimate to sum over the index, a process known as contraction, resulting in a tensor of lower order. In the preceding example,

$$\phi^i = \psi^{irs}_{rs} = \omega^{irs} \gamma_{rs}$$

is a contravariant vector, or the component vector of a point in $\mathcal{V}$. It is an elementary exercise to show that ϕ^i does obey the transformation rule for component vectors in $\mathcal{V}$.

Tensor calculus is simply the application of these multiplication and contraction rules: it is ordinarily not legitimate, for example, to sum over a single index, or to make two superscripts equal and to sum over the 'diagonal.'

One of the most important cases of tensor contraction occurs when it results in a tensor of order zero, otherwise known as a scalar or an invariant. Such quantities are particularly important because their value is the same in every coordinate system within the group.

0.5 Examples

The following is a brief selection of spaces and tensors that arise in various statistical operations.

Example 17: Moments of random variables. A random variable $X(\omega)$ is conceived of as a function from the outcome space Ω with values in a vector space $\mathcal{V}$. The average $\mu = E(X) = \int X(\omega)\, dP(\omega)$ is a linear combination of vectors in $\mathcal{V}$. Hence $\mu \in \mathcal{V}$. To say the same thing in an equivalent way, the vector with components

$$\mu^r = E(X^r) = \int X^r(\omega)\, dP(\omega)$$

are the components of a vector in $\mathcal{V}$. Likewise, since $X^r(\cdot)$ is a linear functional on $\mathcal{V}$ and $X^r X^s$ is a bi-linear functional, it follows that

$$\mu^{rs} = E(X^r X^s) = \int X^r(\omega) X^s(\omega)\, dP(\omega)$$

are the components of a vector in $\mathcal{V}^{\otimes 2}$.

Usually, we omit mention of the outcome space Ω and write the first three moments of X as follows:

$$\begin{aligned} \kappa^i &= E(X^i), \\ \kappa^{ij} &= E(X^i X^j), \\ \kappa^{ijk} &= E(X^i X^j X^k). \end{aligned}$$

these being the components of vectors in $\mathcal{V}$, $\mathcal{V}^{\otimes 2}$, and $\mathcal{V}^{\otimes 3}$ respectively.

To say the same thing in another way, consider the linear transformation $\bar{X}^i = a^i_r X^r$ as in (1.6). Then the moments of the transformed random variable are

$$\begin{aligned} \bar{\kappa}^i &= E(a^i_r X^r) = a^i_r E(X^r) = a^i_r \kappa^r, \\ \bar{\kappa}^{ij} &= E(a^i_r a^j_s X^r X^s) = a^i_r a^j_s E(X^r X^s) = a^i_r a^j_s \kappa^{rs}, \end{aligned}$$

and so on. These are the transformation rules for the components of contravariant tensors in $\mathcal{V}$, $\mathcal{V}^{\otimes 2}$ and so on.

Example 18: Covariance matrix and the affine group. Let $X = (X^1, \ldots, X^n)$ be the components of a random vector in an affine space $\mathcal{V} = \mathcal{R}^n$. We consider $\mathcal{V}$ to be an affine space in the sense that it has no definite origin, or in the sense that it is subject to affine re-parameterization. One way to define the covariance matrix of X is as as follows. Let X_1 and X_2 be independent copies of X, and let $dX = X_1 - X_2$. Then the components of the covariance matrix of X are given by

$$\kappa^{i,j} = \tfrac{1}{2} E(dX^i\, dX^j).$$

In the terminology of Example 4 in section 1.4.3, dX^i are the components of a point in $\mathcal{T}$, the tangent space of $\mathcal{V}$. Since the expectation operator is linear, and $dX^i\, dX^j$ is bi-linear, it follows that $\kappa^{i,j}$ are the coordinates of a point (contravariant tensor) in $\mathcal{T}^{\otimes 2}$. By contrast, the moment matrix with components $\kappa^{ij} = E(X^i X^j)$, is not a tensor in $\mathcal{T}^{\otimes 2}$. The previous example shows that κ^{ij} are the components of a tensor in $\mathcal{V}^{\otimes 2}$, but only if $\mathcal{V}$ is regarded as a vector space rather than an affine space.

We now check the preceding argument in the standard way. The covariance matrix of X has components $\kappa^{i,j}$ defined by

$$\kappa^{i,j} = E(X^i X^j) - E(X^i)E(X^j) = \kappa^{ij} - \kappa^i \kappa^j.$$

Consider now the general affine group

$$\bar{X}^i = a^i + a^i_r X^r$$

acting on $\mathcal{V}$. In order to show that $\kappa^{i,j}$ are the components of a vector in $\mathcal{T}^{\otimes 2}$ we need to show that the covariance matrix of the transformed variable satisfies $\bar\kappa^{i,j} = a^i_r a^j_s \kappa^{r,s}$. From the linearity of the expectation operator, we have

$$\begin{aligned}\bar\kappa^{i} &= E(\bar X^i) = a^i + a^i_r\kappa^r\,,\\ \bar\kappa^{ij} &= E(\bar X^i\bar X^j) = a^i a^j + a^i a^j_r\kappa^r + a^j a^i_r\kappa^r + a^i_r a^j_s\kappa^{rs}.\end{aligned}$$

Then the product of the mean vectors is given by

$$\bar\kappa^i\bar\kappa^j = a^i a^j + a^i a^j_r\kappa^r + a^i a^j_r\kappa^r + a^i_r a^j_s\kappa^r\kappa^s.$$

Finally, the difference gives

$$\bar\kappa^{i,j} = \bar\kappa^{ij} - \bar\kappa^i\bar\kappa^j = a^i_r a^j_s(\kappa^{rs} - \kappa^r\kappa^s) = a^i_r a^j_s\kappa^{r,s}\,,$$

which is the required transformation rule for tensors in $\mathcal{T}^{\otimes 2}$.

Example 19: Metric tensor in Euclidean space. Let $x = (x^1, \ldots, x^n)$ be the coordinates of a point in an n-dimensional affine space $\mathcal{E}$. To say that the space is Euclidean is to say that the squared distance between points x and y in $\mathcal{E}$ is given by the quadratic form

$$ds^2 = d^i d^j g_{ij}\,, \tag{1.10}$$

where $d^i = x^i - y^i$. The coefficients g_{ij} are the components of a symmetric positive definite matrix, called the metric or metric tensor in $\mathcal{E}$.

Now, d^i are the components of a vector in $\mathcal{T}$, the tangent space of $\mathcal{E}$. In order for ds^2 to be invariant under affine transformation it is necessary for g_{ij} to be the components of a tensor in $\mathcal{T}^* \otimes \mathcal{T}^*$. In other words, g_{ij} are the components of a covariant tensor of order two.

The most natural example of a metric tensor is the inner product matrix $G_{ij} = \langle e_i,\, e_j\rangle$, where $\{e_i\}$ are the basis vectors in $\mathcal{T}$. Since $\langle\cdot,\,\cdot\rangle$ is a symmetric bi-linear functional on $\mathcal{T}$, it follows automatically that G_{ij} are the components of a symmetric covariant tensor, called the fundamental metric tensor in $\mathcal{T}$, or in $\mathcal{E}$.

Example 20: Metric tensor in Riemannian space. In a Riemannian space, the metric tensor g_{ij} is not constant over the space, but varies continuously from point to point in the manifold E. Then (1.10) gives the infinitesimal squared distance between neighbouring points x and $x + dx$ in terms of the metric g_{ij} at x. Under arbitrary differentiable re-parameterizations of the manifold, dx^i are the components of a vector in the tangent space, $\mathcal{T}_x$, of E at x. By the argument given above for Euclidean spaces, since ds^2 is invariant, g_{ij} must be the components of a tensor in $\mathcal{T}^*_x \otimes \mathcal{T}^*_x$.

Example 21: Fisher information matrix. Let $l(\theta; y)$ denote the log likelihood function defined on $\Theta \times \mathcal{Y}$, where Θ denotes the parameter space, and $\mathcal{Y}$ denotes the sample space. The components of the observed and expected Fisher information matrices are as follows:

$$I_{ij} = -\frac{\partial^2 l(\hat\theta; y)}{\partial\theta^i\partial\theta^j}; \qquad \mathcal{I}_{ij} = -E_\theta\Big(\frac{\partial^2 l(\theta; Y)}{\partial\theta^i\partial\theta^j}\Big).$$

Under differentiable re-parameterizations $\theta \mapsto \phi$, component vectors in the tangent space of Θ undergo linear transformation via

$$d\theta^i \mapsto d\phi^i = \phi^i_r d\theta^r\,,$$

where $\phi^i_r = \partial\phi^i/\partial\theta^r$ is the Jacobian matrix at θ. It is a straightforward exercise to check that the components of I and $\mathcal{I}$ undergo linear transformation to

$$\bar I_{ij} = \hat\theta^r_i\hat\theta^s_j I_{rs} \quad \text{and} \quad \bar{\mathcal{I}}_{ij} = \theta^r_i\theta^s_j\mathcal{I}_{rs}\,, \tag{1.11}$$

where $\theta^r_i = \partial\theta^r/\partial\phi^i$, and $\hat\theta^r_i$ is the same matrix evaluated at $\hat\theta$. To show that I is a tensor, it is necessary to assume that $\partial l(\theta; y)/\partial\theta^i = 0$ at $\hat\theta$. To show that $\mathcal{I}$ is a tensor, it is necessary to assume that $E_\theta(\partial l(\theta; Y)/\partial\theta^i) = 0$. As a consequence of (1.11), we say that $\mathcal{I}_{ij}$ are the components of a covariant tensor in $\mathcal{T}^*_\theta \otimes \mathcal{T}^*_\theta$, the tensor product of the dual of the tangent space at θ: I_{ij} are the components of a covariant tensor in $\mathcal{T}^*_{\hat\theta} \otimes \mathcal{T}^*_{\hat\theta}$, the tensor product of the dual of the tangent space at $\hat\theta$. So, although both information matrices are symmetric covariant tensors, they are not covariant tensors in the same tangent space.

Note that $\mathcal{I}$ is a function on Θ, in fact a covariant tensor field on Θ. As such, Θ can be regarded as a Riemannian space with $\mathcal{I}$ as metric tensor.

Example 22: Kronecker delta. Kronecker's delta is the symbol δ^j_i, the identity transformation from a vector space to itself. Consider, however, the symbols δ^{ij}, δ_{ij} taking the value 1 if $i = j$ and zero otherwise. What does it mean for these to be the components of a tensor, presumably in the tensor product spaces $\mathcal{V} \otimes \mathcal{V}$ and $\mathcal{V}^* \otimes \mathcal{V}^*$? Unless these are the components relative to a very particular coordinate system, for example the orthogonal coordinate system in $\mathcal{R}^3$ with equal units of distance in each coordinate direction, it is to be understood that δ^{ij} are the components in all coordinate systems. Under linear transformation in $\mathcal{V}$ the components must change to $\bar\delta^{ij}$, also taking the value 1 if $i = j$ and zero otherwise. In symbols, we must have

$$\bar\delta^{ij} = g^i_r g^j_s \delta^{rs} = \delta^{ij},$$

for all matrices g in the group $\mathcal{G}$. This identity holds only for the orthogonal group, and in that restricted sense only, δ^{ij} are the components of a tensor in $\mathcal{V}^{\otimes 2}$. In the same sense δ_{ij} are the components of a tensor in the dual space. Such tensors, whose components are the same under all relevant coordinate transformations, are said to be isotropic (Jeffreys, 19??).

In real space $g_{ij} = \langle e_i, e_j\rangle$ is the inner product or metric tensor on $\mathcal{V}$. There does exist a basis for which $g_{ij} = \delta_{ij}$, namely any standard orthogonal basis with equal units in each of the coordinate directions. From the viewpoint of tensor calculus, it is usually best not to place any restrictions on the basis vectors. In other words, if the calculations involve a metric space, it is best to use a symbol g_{ij}, recognizing that these are the components of a tensor in $\mathcal{T}^* \otimes \mathcal{T}^*$, the tensor product of the dual of the tangent space of $\mathcal{V}$ with itself. Explicit numerical computations may subsequently exploit the orthogonality of the basis vectors, but it is usually a tactical mistake to impose such restrictions at an early stage in the algebra.

Example 23: Extended Kronecker delta. Let δ^{ijk}, δ^{ijkl} take the value 1 if all indices are equal, and zero otherwise. These are isotropic tensors, but only under the symmetric group (permutation of coordinates or indices), which is a discrete sub-group of the orthogonal group on $\mathcal{R}^n$.

There are occasions when this group is of statistical interest, for example when we deal with symmetric functions of independent and identically distributed random variables. From an operational or notational point of view, these extended Kronecker symbols should not appear unless the symmetric group is the relevant group of symmetries on $\mathcal{V}$.

0.6 Inner products and conjugate functionals

Notions of magnitude, direction and orthogonality, familiar from our everyday experience in 'real space,' are absent from the standard definition of a vector space. To include this additional structure, we define an inner product. An inner product on a real vector space $\mathcal{V}$ is a real-valued symmetric bi-linear functional satisfying the following conditions.

$$\begin{aligned}
\langle x, y\rangle &= \langle y, x\rangle;\\
\langle \alpha_1 x_1 + \alpha_2 x_2, y\rangle &= \alpha_1\langle x_1, y\rangle + \alpha_2\langle x_2, y\rangle;\\
\langle x, x\rangle &\ge 0;\\
\text{if}\quad \langle x, x\rangle &= 0, \quad\text{then}\quad x = 0.
\end{aligned}$$

As usual, x_1, x_2 and y are elements of $\mathcal{V}$, and α_1, α_2 are real scalars.

These definitions are the same as those used by Halmos (1958). Stone (1987) eschews the final positive definiteness condition.

In terms of the component vectors, an inner product has the form

$$\langle x,\, y\rangle = x^i y^j g_{ij},$$

where $g_{ij} = \langle e_i,\, e_j\rangle$, the fundamental metric tensor on the space, is positive definite.

For any linear functional A on $\mathcal{V}$, there exists an associated linear functional A^* such that, for all x, y,

$$\langle Ax,\, y\rangle = \langle x,\, A^* y\rangle.$$

We say that A^* is the *conjugate linear functional* of A. By symmetry,

$$\langle x,\, A^*y\rangle = \langle A^*y,\, x\rangle = \langle y,\, A^{**}x\rangle = \langle A^{**}x,\, y\rangle,$$

so that $A^{**} = A$. In matrix notation, if $\langle x,\, y\rangle = x^T Gy$, then

$$\langle Ax,\, y\rangle = (Ax)^T Gy = x^T A^T Gy = x^T G(G^{-1}A^T Gy) = \langle x,\, A^*y\rangle$$

so the conjugate matrix of A is $A^* = G^{-1}A^T G$. In particular, if G is proportional to the identity, the conjugate matrix is the transposed matrix.

Using index notation, and denoting the components of the linear transformation Ax by $a_i^r x^i$, the components of the conjugate transformation are

$$a^{*r}_{\ i} = g^{rs} g_{ij} a_s^j$$

Unlike the adjoint matrix of section 1.4.4, the conjugate requires a non-trivial transformation, a tensor-preserving transposition of subscripts and superscripts.

A matrix that is self-conjugate satisfies the condition $A^T G = GA$, so that GA is symmetric.

0.7 Generalized inverse matrices

A square matrix A of full rank has a unique inverse usually denoted by A^{-1}. If A does not have full rank, any matrix A^- satisfying

$$AA^-A = A \tag{1.12}$$

is said to be a generalized inverse of A. Apart from the full-rank case, there are infinitely many matrices A^- satisfying the defining property of a generalized inverse. However, certain derived properties or derived vectors in statistical applications are independent of the choice of generalized inverse. In particular, since $(AA^-)(Ax) = Ax$, it follows that AA^- is the identity on the range of A. Consequently, the rank of AA^- is at least equal to the dimension of the range of A, i.e. $\mathrm{rank}(AA^-) \geq \mathrm{rank}(A)$. But the rank of a matrix product cannot exceed the rank of either matrix. It follows that $\mathrm{rank}(AA^-) = \mathrm{rank}(A) \leq \mathrm{rank}(A^-)$. Moreover, from the defining property, we have $(AA^-)(AA^-) = AA^-$. In other words AA^- is idempotent, a projection matrix whose eigenvalues are either zero or one. The same is true of A^-A. Consequently,

$$\mathrm{tr}(AA^-) = \mathrm{rank}(AA^-) = \mathrm{rank}(A) \leq \mathrm{rank}(A^-),$$

independently of the choice of generalized inverse.

If B is a full-rank matrix, then $B^{T^{-1}}A^-B^{-1}$ is a generalized inverse of BAB^T because

$$BAB^T(B^{T^{-1}}A^-B^{-1})BAB^T = BAA^-AB = BAB^T,$$

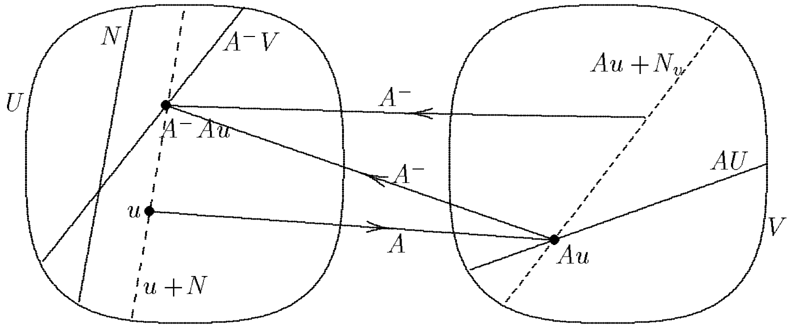

Figure 1.2. *Stone diagram illustrating the definition of an annihilating generalized inverse of a linear transformation from U to V. The range of A is the solid line marked AU in V: the range of A^- is the solid line marked A^-V in U. The solid line N is the null space of A in U. The dashed line $u+N$, the translate of the null space by u, is the pre-image of Au. In the inverse transformation, all points in $Au+N_v$ (dashed line in V), the translate of the null space of A^-, are transformed by A^- to the same point in the pre-image. In the Moore-Penrose inverse, A^-V is orthogonal to N in U, and AU is orthogonal to N_v in V.*

and the defining property is satisfied.

Reverting now to index notation, and supposing A to be symmetric, we write a^{ij} for the components of A and a_{ij} for the components of A^-. The rank of A is then given by $a^{ij}a_{ij}$, the trace of the matrix product. Furthermore, if a^{ij} are the components of a contravariant tensor, the equation above shows that a_{ij} are the components of a covariant tensor. Consequently, $a^{ij}a_{ij}$ is the contraction of a tensor product, or a tensor of order zero, and is invariant.

These results are entirely independent of the choice of generalized inverse. Occasionally, however, it is convenient or even necessary to choose a generalized inverse with the egalitarian property that A is a generalized inverse of A^-. In symbols,

$$A^-AA^- = A^-. \tag{1.13}$$

This property ensures that the two matrices have equal rank. A matrix satisfying the first property is called a *minimal inverse*: a matrix satisfying both properties is called an *annihilating inverse* (Stone, 1987), or a *reflexive inverse* (Rao, 1973).

It may be helpful to explain in coordinate-free geometrical language the rationale for conditions (1.12) and (1.13). Let A be a linear transformation between two vector spaces $\mathcal{U}$ and $\mathcal{V}$ (Fig. 1.1). The null space, $\mathcal{N}$, is the set of vectors in $\mathcal{U}$ such that $Au = 0$. Consequently, all vectors in the set $u+\mathcal{N}$, and only those vectors, are mapped to the point $y = Au$ in the range. The set of points that are mapped to y by A is called the *pre-image of y in $\mathcal{U}$*. Any matrix A^- that transforms y to a point in the pre-image of y is a generalized inverse of A In symbols, the minimal requirement for a generalized inverse is that, for each u, $A^-y = A^-Au$ should lie in $N+u$. Consequently $AA^-Au = Au$, and (1.12) follows.

The minimal property is a condition on the action of A^- on the range of A in $\mathcal{V}$. It says nothing about the action of A^- on any v that is not in the range. The annihilating condition requires A^- to have a null space $N_v \subset \mathcal{V}$ that is complementary to the range. In other words, each $v \in \mathcal{V}$ has a unique decomposition $v = v_0 + y$ with $A^-v_0 = 0$, and $y = Au$ in the range of A. Consequently, the set of points $y + N_v$ is transformed by A^- to a single point A^-y in the pre-image of y. In symbols, for any v_0 in the null space, we have $A^-AA^-v_0 = A^-v_0 = 0$, while for any $y = Au$ in the range, the minimal property gives

$$A^-AA^-y = A^-AA^-Au = A^-Au = A^-y.$$

Consequently, $A^-AA^-v = A^-v$ for all $v \in \mathcal{V}$, and (1.13) follows.

Figure 1.1 illustrates these relationships by means of a Stone diagram in which vector spaces are represented by cartouches, subspaces by solid lines, translated subspaces by dashed lines, and transformations by directed lines between two cartouches. Stone diagrams have a certain runic quality that makes them appear impenetrable at first glance. Unlike Venn diagrams, they require careful study to understand, and considerable practice to construct. But, once mastered, Stone diagrams provide the kind of information that is not readily extracted from matrix formulae.

For statistical work, minimal and annihilating inverses are ordinarily the only types of generalized inverse required. The geometrical description makes it clear that it is always possible to choose a generalized inverse satisfying both (1.12) and (1.13), but these conditions do not define a unique inverse. It is of some interest to investigate what further conditions are required to ensure uniqueness. Provided that $\mathcal{U}$ and $\mathcal{V}$ are inner product spaces, two further conditions are enough. Let the inner products be given by $\langle u_1,\, u_2\rangle = u_1^T W_u u_2$ and $\langle v_1,\, v_2\rangle = v_1^T W_v v_2$, where W_u and W_v are real positive definite symmetric matrices. The Moore-Penrose inverse requires that the null space of A be orthogonal to the range of A^-, and, symmetrically, that the null space of A^- be orthogonal to the range of A. For any u in $\mathcal{U}$, A^-Au lies in the range of A^-, and $u - A^-Au$ lies in the null space of A. The first Moore-Penrose orthogonality condition is thus $\langle A^-Au,\, u - A^-Au\rangle = 0$ for all $u \in \mathcal{U}$. In matrix notation, we have

$$u^T(A^-A)^T W_u(I - A^-A)u = 0.$$

Consequently, the matrix of this quadratic form must be skew-symmetric.

$$(A^-A)^T W_u - (A^-A)^T W_u A^-A = -W_u A^-A + (A^-A)^T W_u A^-A.$$

Pre-multiplication by $(A^-A)^T$ and post-multiplication by A^-A give

$$(A^-A)^T W_u = (A^-A)^T W_u A^-A = W_u A^-A.$$

In other words, A^-A is required to be a self-conjugate projection matrix on $\mathcal{U}$. The same argument with A and A^- reversed gives the second orthogonality condition in $\mathcal{V}$:

$$(AA^-)^T W_v = (AA^-)^T W_v AA^- = W_v AA^-.$$

In particular, the familiar least-squares projection matrices $P_X = X(X^TWX)^{-1}X^TW$ and $I - P_X$, are self-conjugate and self-inverse in the Moore-Penrose sense. These matrices are considered as transformations from the sample space $\mathcal{U}$ to itself, the inner product matrix being W.

Different choices of inner product matrices on the two spaces yield different Moore-Penrose inverses, a point that is rarely emphasized. In common with most authorities, Rao (1973, p. 26) and Kruskal (1975) take W_u and W_v to be the identity matrices. Stone's definition coincides with ours.

Example 24: Multinomial covariance matrix. Let $\mathcal{V}$ be an inner product space with metric tensor $W = \{w_{ij}\}$, symmetric and positive definite, and let $\mathcal{U} = \mathcal{V}^*$ be the dual of $\mathcal{V}$. The dual inner product matrix in $\mathcal{V}^*$ is W^{-1}. In statistical applications, $W^{-1} = \Sigma$ is usually a covariance matrix. A contravariant tensor with components a^{ij} may be regarded as the components of a linear transform from $\mathcal{U} = \mathcal{V}^*$ to $\mathcal{V}$. Thus $W_u = \Sigma$ and $W_v = \Sigma^{-1}$. Suppose that A is given in matrix notation by

$$A = W_u(I - X(X^TW_uX)^{-1}X^TW_u) = W_u(I - P),$$

where X is a given matrix whose columns are linearly independent vectors in $\mathcal{U}$. Then the Moore-Penrose inverse is

$$A^- = (I - P)W_v$$

In particular, if $W_u = \mathrm{diag}\{\pi_1, \ldots, \pi_n\}$ with $\sum \pi_j = 1$, and $X = 1$, the constant vector, then A is the multinomial covariance matrix whose Moore-Penrose inverse is

$$\mathrm{diag}\{1/\pi_1, \ldots, 1/\pi_n\} - 11^T.$$

In this instance, the null space of A is the span of 1, while the null space of A^- is the span of the vector π.

0.8 Orthogonal projection

Let $\mathcal{V}$ be a real vector space with inner product $\langle\cdot,\,\cdot\rangle$, symmetric and linear in each argument. By the orthogonal projection of $y \in \mathcal{V}$ on to the subspace $\mathcal{V}_1 \subset \mathcal{V}$, we mean a vector $\hat{y} \in \mathcal{V}_1$ satisfying $\langle y - \hat{y},\, \hat{y}\rangle = 0$. In order to construct this projection, or the components of the projection with respect to a basis $\{e_r\}$ in $\mathcal{V}_1$, we proceed as follows.

Denote by $g_{rs} = \langle e_r,\, e_s\rangle$ the fundamental metric tensor on $\mathcal{V}_1$. As usual, the inverse matrix is denoted by g^{rs}. Define $\hat{y}_r = \langle e_r,\, v\rangle$. These are the components of a vector in $\mathcal{V}_1^*$. Further, $\hat{y}^r = g^{rs}\hat{y}_s$ are the components of $\hat{y} \in \mathcal{V}_1$ with respect to the basis $\{e_r\}$. In symbols,

$$\hat{y} = \hat{y}^r e_r = g^{rs} e_r \langle e_s,\, y\rangle$$

is the projection of y on to $\mathcal{V}_1$.

To write the same thing using matrix notation, suppose that $\langle x,\, y\rangle = x^T W y$, where W is $n \times n$ and positive definite. Suppose, in addition, that e_r is identified with the rth column of the $n \times p$ matrix X, so that $g_{rs} = (X^T W X)_{rs}$. Then $\hat{y}_r = (X^T W y)_r$, $\hat{y}^r = ((X^T W X)^{-1} X^T W Y)_r$, and $\hat{y} = X(X^T W X)^{-1} X^T W Y$, a familiar expression from linear models.

Note that $X^- = (X^T W X)^{-1} X^T W$ satisfies the minimal condition for a generalized inverse of X, so we may write $\hat{y} = X X^- y$ with $X X^-$ as an explicit projection matrix. This particular choice of generalized inverse yields the orthogonal, or self-conjugate, projection: other choices of generalized inverse yield non-orthogonal projections on to the same space.

The preceding analysis assumes that the columns of X are linearly independent so that $X^T W X$ is non-singular. For a variety of reasons, this condition is frequently not satisfied in statistical work. In order to make progress, it is natural to construct the matrix

$$P = X(X^T W X)^- X^T W$$

for some choice of generalized inverse, with the intention that PY should be the required projection on to the column space of X. It is straightforward to check that P is a projection matrix ($P^2 = P$), if and only if the generalized inverse is annihilating. Furthermore, if $X^T W X$ has the same rank as X, all annihilating inverses yield the same projection matrix, and hence the same fitted value or least-squares projection.

Example 25: Estimating functions. Let Y be a random vector in an n-dimensional affine space $\mathcal{V}$. Define $\mu = E(Y)$, $\Sigma = \mathrm{cov}(Y)$, and $W = \Sigma^{-1}$, where Σ is taken to be positive definite. A model specifies that μ lies in a p-dimensional manifold $\mu(\theta) \subset \mathcal{V}$, with θ varying in a p-dimensional parameter set Θ. In general, Σ depends on μ, and hence on θ. Now, $Y - \mu$ is a vector in $\mathcal{T}$, the tangent space of $\mathcal{V}$, and W is a metric tensor on $\mathcal{T}$. Since W depends on μ, we denote by $\langle\cdot,\,\cdot\rangle_\theta$, the inner product using the metric tensor W evaluated at $\mu(\theta)$.

Let $e_r = \partial\mu/\partial\theta^r$ for $r = 1, \ldots, p$. In matrix notation, $D = (e_1, \ldots, e_p)$ is the derivative matrix of the mean-value vector with respect to the parameters at θ. These are the natural basis vectors in $\mathcal{T}_\theta$, the tangent space at θ, which may be regarded as a subspace of $\mathcal{T}$. The vector $U \in \mathcal{T}_\theta^*$ with components

$$U_r = \langle e_r,\, Y - \mu\rangle_\theta = (D^T W(Y - \mu))_r$$

is known as the quasi-likelihood estimating function, or quasi-likelihood score vector. The root of the equation $U_r(\hat{\theta}) = 0$ is the quasi-likelihood estimator: if it exists, it may not be unique. Note that $g^{rs}U_s$ are the components of the projection of $Y - \mu$ on to $\mathcal{T}_\theta$ with respect to the basis $\{e_r\}$. This is the first Newton step towards $\hat{\theta}$.

0.9 Mixed tensors

0.9.1 *Linear transformation of vector spaces*

The space of linear transformations from $\mathcal{U}$ of dimension n to $\mathcal{V}$ of dimension m is the set of functions g on $\mathcal{U}$ with range in $\mathcal{V}$, satisfying

$$g(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 g(x_1) + \alpha_2 g(x_2).$$

Evidently, if g_1 and g_2 are in this set, any linear combination $\alpha_1 g_1 + \alpha_2 g_2$ is also in the set. Consequently the set of such transformations constitutes a vector space, in fact a vector space of dimension mn. The components of g with respect to the basis vectors in $\mathcal{U}$ and $\mathcal{V}$ form a $m \times n$ matrix $\{g_r^i\}$: the components of $y = g(x)$ are given by

$$y^i = g_r^i x^r.$$

Thus g_r^i are the components of the mixed tensor g in the space $\mathcal{U} \otimes \mathcal{V}^*$.

0.9.2 *Factorial design and Yates's algorithm*

Consider a laboratory experiment in which yield is observed at each of four temperatures, (factor A), each of three concentrations of solute, (factor B), and for each of two catalysts, (factor C). The incidence bases for factor A, B and C are

$$\begin{aligned} A &= \mathrm{span}\{e_1 = \delta_{i1},\, e_2 = \delta_{i2},\, e_3 = \delta_{i3},\, e_4 = \delta_{i4}\}, \\ B &= \mathrm{span}\{e'_1 = \delta_{j1},\, e'_2 = \delta_{j2},\, e'_3 = \delta_{j3}\}, \\ C &= \mathrm{span}\{e''_1 = \delta_{k1},\, e''_2 = \delta_{k2}\}. \end{aligned}$$

To say that Y^{ijk} is the observation with temperature at level i, concentration at level j, and catalyst at level j, is to say that Y^{ijk} are the components of a vector Y in $A \otimes B \otimes C$ relative to the product basis $e_{ijk} = e_i \otimes e'_j \otimes e''_k$. In symbols, $Y = Y^{ijk} e_{ijk}$.

In comparative experiments, which are concerned with contrasts rather than levels, it is frequently required to transform to a new basis of orthogonal contrasts for A, such as the polynomial contrasts

$$\begin{pmatrix} \bar{e}_1 \\ \bar{e}_2 \\ \bar{e}_3 \\ \bar{e}_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}.$$

We write this using index notation as $\bar{e}_r = K_r^j e_j$, with inverse transform $e_j = L_j^r \bar{e}_r$. Note that $KK^T = \Lambda = \mathrm{diag}\{4, 20, 4, 20\}$, so that $L = K^T \Lambda^{-1}$, or

$$L^T = \begin{pmatrix} 1/4 & & & \\ & 1/20 & & \\ & & 1/4 & \\ & & & 1/20 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{pmatrix}.$$

The usual polynomial contrast matrix for factor B, giving $\bar{e}'$ in terms of e' is

$$K' = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}.$$

Since the rows are orthogonal, the inverse matrix, $L' = K'^T \Lambda'^{-1}$, can be obtained in the same way as for factor A.

The components of Y with respect to the contrast basis $\bar{e}_{rst} = \bar{e}_r \otimes \bar{e}'_s \otimes \bar{e}''_t$ are $Y = \bar{Y}^{rst}\bar{e}_{rst}$, where

$$\bar{Y}^{rst} = L_i^r L_j'^s L_k''^t Y^{ijk}. \tag{1.14}$$

These are the standardized factorial contrasts and interactions that emerge at the third and final step in the standard implementation of Yates's algorithm. The unstandardized or raw contrasts are those obtained by omitting the 'divisors' Λ, Λ' and Λ'' from the definitions of L, L' and L'', i.e. by replacing L^T by K rather than $\Lambda^{-1}K$. Each matrix multiplication in (1.14) represents one step of Yates's algorithm.

In the present notation, i, j, k refer to factor levels, whereas r, s, t refer to factor contrasts. Thus $\bar{Y}^{111}$ is the mean response, often denoted by I, $\bar{Y}^{121}$ is the linear effect of solute, usually denoted by B_L, and $\bar{Y}^{312}$ is the quadratic by linear interaction of factors A and C, usually denoted by $A_Q C_L$.

The number of arithmetic operations required to compute the contrast vector in (1.14) is $4n+3n+2n$ multiplications and a similar number of additions. Here $n = 4\times3\times2$ is the total number of observations or factor combinations. In general, if there are p factors with $n_1, \ldots, n_p$ levels, the number of operations required by Yates's algorithm is $n(n_1+\cdots+n_p)$. This number is considerably fewer than n^2, the number of operations required to multiply a vector by an unstructured $n \times n$ matrix. In particular, if $n = 2^p$, the number of operations required is $2n\log_2(n)$.

For further details, see Good (1958) or Takemura (198?).

0.9.3 *Fast Fourier transform*

The discrete Fourier transform F_j of a sequence of numbers x_t is a linear transformation given by

$$nF_j = \sum_{t=0}^{n-1} x_t W^{jt},$$

where $W = \exp(2\pi i/n)$ is a complex nth root of unity, and the superscript on W is a power. The fast Fourier transform is a computational device modelled on Yates's algorithm. In cases where n is a composite number, $n = n_1 n_2 \cdots n_p$, the time index is represented as a multi-index, so x_t becomes $x_{t_1 t_2 \ldots t_p}$. A similar indexing system is used for the sequence of Fourier coefficients. Although the multi-index representation of the discrete Fourier transform is not quite so simple as (1.14), so-called twiddle factors can be introduced to achieve similar gains in computational efficiency. For details of the algorithm, see Bloomfield (1976, ch. 4).

0.10 Cosets, quotient spaces and annihilators

0.10.1 *Quotient space*

Let $\mathcal{V}$ be a vector space, and let $\mathcal{X}$ be a subspace of $\mathcal{V}$. A subspace $\mathcal{Y}$ is said to be *complementary* to $\mathcal{X}$ in $\mathcal{V}$ if $\mathcal{X}$ and $\mathcal{Y}$ are disjoint, $(\mathcal{X} \cap \mathcal{Y} = \{0\})$, and $\mathcal{Y} + \mathcal{X} = \mathcal{V}$. For a given subspace $\mathcal{X}$ there are, usually, infinitely many complementary subspaces. Without some additional structure, there is no natural way to select one of these complements for preferential treatment. There is, however, a natural way of patching together all these complementary subspaces to form a new vector space that, for all practical purposes, gives a unique complement of $\mathcal{X}$ in $\mathcal{V}$.

For any $v \in \mathcal{V}$, $v + \mathcal{X}$ denotes the set of vectors $v + x$ with $x \in \mathcal{X}$, called a *coset* of $\mathcal{X}$, or the translate of $\mathcal{X}$ by v. Unless $v \in \mathcal{X}$, this is not a subspace because it does not include zero. Note that if $v_1 + \mathcal{X} = v_2 + \mathcal{X}$, it does not follow that v_1 and v_2 are equal. It does, however, follow that $v_1 - v_2 \in \mathcal{X}$, so we may say that '$v_1 = v_2$ modulo $\mathcal{X}$.' The cosets thus partition $\mathcal{V}$ into non-overlapping subsets or equivalence classes.

Consider now the set whose elements are the cosets of $\mathcal{X}$ in $\mathcal{V}$. Addition of two cosets is defined by addition of the elements, and scalar multiplicaton is defined likewise. Thus

$$(v_1 + \mathcal{X}) + (v_2 + \mathcal{X}) = v_1 + v_2 + \mathcal{X} + \mathcal{X} = v_2 + v_2 + \mathcal{X},$$
$$\alpha(v_1 + \mathcal{X}) = \alpha v_1 + \alpha\mathcal{X} = \alpha v_1 + \mathcal{X}$$

for any scalar α. The cosets thus form a vector space, $\mathcal{X}$ itself being the zero element. This vector space is called the *quotient space*, and is denoted by the symbol $\mathcal{V}/\mathcal{X}$. The dimension of the quotient space is $\dim(\mathcal{V}) - \dim(\mathcal{X})$.

Example 26: Residual vector in linear models. Consider the linear model, written in conventional matrix notation as

$$Y = X\beta + \epsilon,$$

where X is a given matrix whose columns span a subspace $\mathcal{X}$ of $\mathcal{V} = \mathcal{R}^n$, and ϵ is a random variable in $\mathcal{R}^n$, whose distribution depends on unknown parameters, θ. Let P be any projection matrix having range $\mathcal{X}$, and let $Q = I - P$ be the complementary projection having null space $\mathcal{X}$. For example, we could take $P = X(X^TX)^{-1}X^T$ or $P' = X(X^TWX)X^TW$ for any positive definite matrix W. Then $R = QY$ is a residual vector having a distribution independent of β, but dependent on θ. All residuals are equivalent in the sense that they contain the same information. For example, the residual $R' = Q'Y$ can be obtained from $R = QY$ by the linear transformation $R' = Q'R$, and conversely $R = QR'$. Note that, although R' and R are vectors in different spaces, the difference $R' - R$ is necessarily a linear combination of the columns of X, i.e. an element of $\mathcal{X}$. For this reason, it is generally preferable to regard the residual vector as an element of the quotient space $\mathcal{V}/\mathcal{X}$ rather than an element of any particular subspace complementary to $\mathcal{X}$ in $\mathcal{V}$.

In fact, it is technically not necessary to construct a projection matrix or form a residual vector at all: we may talk simply of the distribution of the random variable $Y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$, and use the resulting 'residual likelihood' for purposes of estimating θ. So far as its dependence on the data is concerned, the residual likelihood is a function on the space $\mathcal{V}/\mathcal{X}$, i.e. constant on the cosets of $\mathcal{X}$.

0.10.2 *Annihilator of a subspace*

Let $\mathcal{V}$ be a vector space of dimension n, $\mathcal{X}$ a subspace of $\mathcal{V}$ of dimension p, and let $\mathcal{V}^*$ be the dual space of linear functionals on $\mathcal{V}$. Let $\mathcal{X}^0$ be a that subset of linear functionals taking the value zero on $\mathcal{X}$. Then $\mathcal{X}^0$ is a subspace of $\mathcal{V}^*$ of dimension $n - p$, called the annihilator of $\mathcal{X}$ (Halmos, 1958, section 17).

The elements of $\mathcal{X}^0$ are linear functionals taking the value zero on $\mathcal{X}$, and therefore constant on the cosets of $\mathcal{X}$. To say the same thing in another way, $\mathcal{X}^0$ is the dual space of linear functionals on the quotient space $\mathcal{V}/\mathcal{X}$, and we may write formally $\mathcal{X}^0 = (\mathcal{V}/\mathcal{X})^*$. A basis in $\mathcal{X}^0$ thus serves as the component vector of a 'point' in the quotient space relative to the dual basis.

0.11 Determinants and measures

0.11.1 *Measures and densities*

A measure μ in $\mathcal{V}$ is a real-valued function defined on a collection of subsets of $\mathcal{V}$. The main property of a measure is that it is additive for disjoint sets: $\mu(A \cup B) = \mu(A) + \mu(B)$ if A and B are disjoint. A measure thus encompasses a variety of notions such as length, area, volume, electric charge, mass and probability, all of which are additive for disjoint sets. Some measures, such as counting measure on integers, are purely atomic in the sense that μ is concentrated on a countable number of points. Other measures are continuous or a mixture of atomic and continuous parts. In the majority of applications, measures are non-negative. However, signed measures do occur, for example in electrostatics where $\mu(A)$, the net charge in A, may be positive or negative.

While it is possible, and conceptually helpful, to conceive of a measure, or any other function on $\mathcal{V}$, in the abstract, most computations require functions to be expressed in terms of the component vector x^i relative to a given basis or coordinate system. In what follows, we consider f to be a function on $\mathcal{V}$, not directly defined in a coordinate-free manner, but indirectly as a function of the component-vector x^i in $\mathcal{R}^n$. The physical appearance of such a function thus depends on the coordinate system used. Typically, therefore, f_x and f_y represent the same function on $\mathcal{V}$, but with respect to different coordinate systems.

Continuous measures have the property that μ can be expressed as an integral with respect to Lebesgue measure in the form

$$\mu(S) = \int_S f_x(x)\, dx,$$

where f_x is called the density function of the measure relative to the given basis vectors. This is an integral with respect to Lebesgue measure in the space of component vectors. However, the subscript on the integral is a subset of $\mathcal{V}$, not a subset of the component vectors in $\mathcal{R}^n$. To make sense of this, we agree that the symbol $\int_S$ denotes an integral over those component vectors x^i such that $x^i e_i$ lies in S.

One awkwardness of this construction in the present context is the dependence on the units of measurement and on the coordinate system. The Lebesgue measure of the set $[0,1]^n$ in $\mathcal{R}^n$, or the set of component vectors x^i with $0 \le x^i \le 1$ is the dimensionless number one. The meaning of this number clearly depends on whether the basis vectors e_i are in units of centimetres or kilometres, and indeed, on whether they are mutually perpendicular. The interpretation given to a unit interval in the first coordinate direction is the magnitude of e_1, the first basis vector. This unit of measure is denoted by $\|e_1\| = \langle e_1,\, e_1\rangle^{1/2}$. Assuming that the basis vectors have the physical dimension of length, the interpretation of the unit square $0 \le x^1, x^2 \le 1$ in the plane of the first two components is that of area. The units of area are given by the square root of the determinant of the 2×2 matrix with components $g_{ij} = \langle e_i,\, e_j\rangle$. For example, if e_1 denotes 'metres east' and e_2 denotes 'centimetres vertical', then $g_{11} = 10^6$, $g_{12} = g_{21} = 0$, and $g_{22} = 1$, the fundamental unit of measure being cm^2. Then the unit interval $0 \le x^1, x^2 \le 1$ in the space of component vectors represents an area of $1 \times |G|^{1/2}$ cm^2. This notion extends without modification to spaces of higher dimension.

In general, therefore, since a measure has physical units inherited from the magnitudes and orientations of the basis vectors, it is preferable to represent the density in the form

$$\mu(S) = \int_S p(x)\, |G|^{1/2}\, dx. \qquad (*)$$

The introduction of the determinantal factor may seem inconsequential because G is a constant matrix and the determinant is a constant factor in the integral. However, because G is a covariant tensor of order two, the latter representation is invariant with respect to coordinate transformation, as we shall show presently. In fact, we could replace G in $(*)$ by any symmetric positive definite covariant tensor because this implies a simple multiplicative re-definition of the units of measure.

Consider now the effect of a change of basis vectors from the original e_j to $\bar e_i = a_i^j e_j$ with matrix A. The component vectors tranform contravariantly from x^i to $y^i = b_j^i x^j$ with matrix $B = A^{-1}$. The transformed metric tensor is $\bar g_{rs} = a_r^i a_s^j g_{ij}$, so the transformed units of measure are $|\bar G|^{1/2} = |G|^{1/2} \times J$, where J is the absolute value of the determinant of A. Making the required change of variables ($y = Bx$) in the integral $(*)$, we find

$$\mu(S) = \int_S p(Ay)\, |\bar G|^{1/2}\, dy = \int_S \bar p(y)\, |\bar G|^{1/2}\, dy.$$

The subscript on the integral is unchanged because of our convention that this is an integral over those component vectors, y, such that $y^i \bar e_i$ lies in $S \subset \mathcal{V}$.

From the point of view of tensor transformation properties, any function f_x that undergoes transformation by the rule

$$f_y(y) = f_x(Ay)\, J$$

when the component vectors are transformed linearly by $y = Bx$, is called a density on $\mathcal{V}$. The term *relative invariant of weight one* is also used for the same purpose. By this definition, the square root of the determinant of a second-order covariant tensor is a density. Also, the inverse square root of the determinant of a second-order contravariant tensor is a density.

By way of contrast, a function that transforms by the rule $f_y(y) = f_x(Ay)$ is called an invariant, or an absolute invariant on $\mathcal{V}$. The function $p(x)$ in $(*)$ is an absolute invariant. A likelihood function, unlike a probability density function, is an absolute invariant in this sense, with respect to re-parameterization. Potential functions associated with gravitational or electrostatic fields are examples of absolute invariants in physics.

0.11.2 *Factors of measures*

Let $\mathcal{V}_1, \mathcal{V}_2$ be complementary subspaces of $\mathcal{V}$ such that $e_1, \ldots, e_m$ is a basis in $\mathcal{V}_1$, and $e_{m+1}, \ldots, e_n$ is a basis in $\mathcal{V}_2$. The fundamental metric tensor on $\mathcal{V}$, with components $g_{ij} = \langle e_i,\, e_j \rangle$, can be partitioned as follows.

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}; \qquad G^{-1} = \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix}.$$

Thus, G_{11} is the metric tensor on $\mathcal{V}_1$, and G_{22} is the metric tensor on $\mathcal{V}_2$. Further, $|G|^{1/2}$ is a density on $\mathcal{V}$, $|G_{11}|^{1/2}$ is a density on $\mathcal{V}_1$, and $|G_{22}|^{1/2}$ is a density on $\mathcal{V}_2$.

Now, G^{-1} is the fundamental metric tensor on $\mathcal{V}^*$, so G^{22} is the metric tensor on $\mathcal{V}_2^*$, the dual space of linear functionals on $\mathcal{V}_2$. The dual basis is the set of linear functionals v^j such that $v^j(e_i) = \delta_i^j$. Consequently the basis linear functionals in $\mathcal{V}_2^*$ all take the value zero on $\mathcal{V}_1$, so $\mathcal{V}_2^*$ is the annihilator of $\mathcal{V}_1$. Also, $\mathcal{V}_2^*$ is the space of linear functionals on the quotient space $\mathcal{V}/\mathcal{V}_1$. Conversely, $\mathcal{V}_1^*$ is the annihilator of $\mathcal{V}_2$, and the space of linear functionals on $\mathcal{V}/\mathcal{V}_2$.

Using formulae for the inverse of a partitioned matrix, G^{22} satisfies

$$G^{22} = \left(G_{22} - G_{21}G_{11}^{-1}G_{12}\right)^{-1}.$$

The determinant of G is given by

$$\begin{aligned} |G|^{1/2} &= |G_{11}|^{1/2}\; |(G_{22} - G_{21}G_{11}^{-1}G_{12})|^{1/2} \\ &= |G_{11}|^{1/2}\; |G^{22}|^{-1/2}. \end{aligned} \qquad (**)$$

In words, the density $|G|^{1/2}$ on $\mathcal{V}$ factors into a product of two densities, $|G_{11}|^{1/2}$ on $\mathcal{V}_1$, and $|G^{22}|^{-1/2}$ on $\mathcal{V}/\mathcal{V}_1$. The alternative factorization $|G^{11}|^{-1/2}|G_{22}|^{1/2}$ is also a product of two densities, on $\mathcal{V}/\mathcal{V}_2$ and $\mathcal{V}_2$, showing that the factorization is not unique.

To understand better the nature of these two factorizations, consider a change of basis vectors in $\mathcal{V}$ in which the subspace $\mathcal{V}_1$ is preserved. In other words, consider a re-parameterization $\bar{e}_r = a_r^i e_i$ in which a_r^i is zero for $r \le m$ and $i > m$. In matrix notation, if r denotes the row and i the column, the components a_r^i and b_r^i of A and B are

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}; \qquad B = A^{-1} = \begin{pmatrix} A^{11} & 0 \\ A^{21} & A^{22} \end{pmatrix}.$$

Note that A^{11} is the inverse matrix of A_{11}, and likewise for A^{22}. The transformed metric tensor and its inverse are AGA^T and $B^TG^{-1}B$, or, in tensor notation, $a_r^i a_s^j g_{ij}$ and $b_r^i b_s^j g^{rs}$. The first factor in (**),

$$\bar{G}_{11} = A_{11}^T G_{11} A_{11}$$

is the transformed metric tensor on $\mathcal{V}_1$. The complementary factor

$$(\bar{G}^{22})^{-1} = A_{22}^T(G_{22} - G_{21}G_{11}^{-1}G_{12})A_{22}$$

functions as a metric tensor on any space complementary to $\mathcal{V}_1$, i.e. on the space $\mathcal{V}/\mathcal{V}_1$. The components of this tensor are unaffected by the sub-matrix A_{21} in the transformation.

0.11.3 *Singular normal likelihood*

Let Y be a normally distributed random variable in $\mathcal{V} = \mathcal{R}^n$ with mean μ, and covariance matrix Σ depending on unknown parameters to be estimated. It is convenient to let $\mathcal{V}$ be an inner product space with inner product matrix $G = \Sigma^{-1}$. The joint density of Y at y is given by

$$f(y) = (2\pi)^{-n/2}\,|G|^{1/2}\,\exp\bigl(-\tfrac{1}{2}||y-\mu||^2\bigr),$$

where $||y||^2 = \langle y,\, y\rangle = y^TGy$ is the invariant squared length of y.

Suppose that $\mu = E(Y)$ is known to lie in a given p-dimensional subspace $\mathcal{X}$, specified, for example, as the span of the columns of a matrix X. Let $\bar{e}_1, \ldots, \bar{e}_p$ be a basis in $\mathcal{X}$, and let $\bar{e}_{p+1}, \ldots, \bar{e}_n$ be a basis in any complementary space such that the transformation from the old to the new basis has unit determinant. Let P be the orthogonal projection from $\mathcal{V}$ to $\mathcal{X}$, and let $Q = I - P$ be the complementary projection. Note in particular that QY is a residual vector in $\mathcal{V}/\mathcal{X}$. Pythagoras's theorem gives $||y||^2 = ||Py||^2 + ||Qy||^2$. Further, the transformed inner product matrix with respect to the new basis is

$$\bar{G} = \begin{pmatrix} \bar{G}_{11} & \bar{G}_{12} \\ \bar{G}_{21} & \bar{G}_{22} \end{pmatrix}.$$

The determinant of $\bar{G}$ factors as the product

$$|G| = |\bar{G}| = |\bar{G}_{11}|\,|\bar{G}^{22}|^{-1}, \qquad (*)$$

such that $|\bar{G}_{11}|^{1/2}$ is a density on $\mathcal{X}$, and $|\bar{G}^{22}|^{-1/2}$ is a density on $\mathcal{V}/\mathcal{X}$. Then the joint density of Y factors into a product as follows:

$$(2\pi)^{-n/2}|\bar{G}_{11}|^{1/2}\exp\bigl(-\tfrac{1}{2}||P(y-\mu)||^2\bigr) \times |\bar{G}^{22}|^{-1/2}\exp\bigl(-\tfrac{1}{2}||Qy||^2\bigr).$$

The first factor is the density of the fitted values in $\mathcal{X}$. The second is the residual density on $\mathcal{V}/\mathcal{X}$, constant on cosets of $\mathcal{X}$.

The matrix $\bar{G}_{11}$ is the induced metric tensor on $\mathcal{X}$. In matrix notation, if the columns of X are used as the basis in $\mathcal{X}$, this is X^TGX, or $X^T\Sigma^{-1}X$. From equation (*), we have

$$|\bar{G}^{22}|^{-1} = |G|/|\bar{G}_{11}| = |\Sigma|^{-1}\,|X^T\Sigma^{-1}X|^{-1}.$$

Consequently, the likelihood function based on the singular normal distribution of the residuals in $\mathcal{V}/\mathcal{X}$ is

$$|\Sigma|^{-1/2}\,|X^T\Sigma^{-1}X|^{-1/2}\exp\bigl(-\tfrac{1}{2}||Qy||^2\bigr) \qquad (**)$$

In this expression, Q is the orthogonal complement of P, not an arbitrary projection having null space $\mathcal{X}$. As a matrix Q is given by $I - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$.

This derivation asumes that the transformation to the new basis vectors spanning $\mathcal{X}$ and its complement has unit determinant. If this assumption is not satisfied, a determinantal factor independent of Σ enters into the density calculations. But this determinant is a constant and has no effect on the residual likelihood, so (**) is unaffected.

0.12 Vector Fields

0.12.1 *Background*

A vector field on a set S is a vector-valued function on S. To each element $x \in S$ is assigned a vector $f(x)$, read as 'f at x.' In fluid dynamics, for example, x denotes position in real space, and f at x is the velocity vector of the fluid at x. In gravitation, mechanics or electrostatics, $f(x)$ may denote the force field at x. In applications connected with statistical estimation, x is a point in the parameter space, and $f(x)$ is the value of the estimating function at x. In most physical applications, it is appropriate to imagine the collection of vectors at different points in S as vectors in the same space. Such vectors can therefore be added or subtracted. For example, the average velocity of a fluid is obtained by adding the velocity vectors of all the particles and dividing by the number of particles. But, in general, although $f(x_1)$ may be numerically equal to $f(x_2)$, the physical effects (on the motion of a rigid body, for example), may be quite different. In statistical applications, the metric tensor or covariance matrix frequently depends on the parameter. Even if $f(x_1)$ is numerically equal to $f(x_2)$, an assessment of statistical significance may depend on the metric tensor, leading to a difference in conclusions. In general, therefore, $f(x)$ is best regarded as a vector in a space $\mathcal{T}_x$, usually the tangent space at x or its dual. Each point has its own tangent space, possibly with its own metric tensor. The collection of these spaces is called the tangent bundle.

Physical conservation laws require fields to have certain characteristic properties. For example, if f is an electrostatic or gravitational force field, conservation of energy implies that the net work required or energy gained in traversing a closed loop γ must be zero. Otherwise, energy could be created by traversing the loop indefinitely. Now, f is a vector in the same space as x, or in the tangent space, so it is appropriate to denote the components by x^i and f^i, in the manner of contravariant tensors. The conservation of energy condition is then

$$\int_\gamma \langle f,\, dx\rangle = \int_\gamma f^i\, g_{ij}\, dx^j = 0,$$

where g_{ij} is the metric tensor associated with the inner product. Fields satisfying this condition are called conservative. Likewise, conservation of mass considerations in the flow of incompressible fluids require that, for any element of volume, the net inflow should equal the net outflow, a condition known as zero divergence.

0.12.2 *Divergence of a vector field*

Let $\mathcal{V}$ be an affine space, and let $f(\cdot)$ be a vector field defined on $\mathcal{V}$ with values in $\mathcal{T}$, the tangent space of $\mathcal{V}$. The components of f at v are denoted by $f^i(v)$. Many familiar examples of vector fields in physics are of this type. Real space has no natural origin, so it is best envisaged as an affine space (ignoring relativistic curvature due to gravitation). Although x is a point in $\mathcal{V}$, f itself is a vector in the tangent space of $\mathcal{V}$. Now, $\partial/\partial x^i$ are the components of a vector in $\mathcal{T}^*$, the dual of $\mathcal{T}$. Consequently, assuming that $f(\cdot)$ is differentiable,

$$f^i_j(x) = \partial f^i(x)/\partial x^j$$

are the components at $x \in \mathcal{V}$ of a tensor in $\mathcal{T} \otimes \mathcal{T}^*$. Further, since δ^j_i are the components of a tensor in the same space, the scalar $d(x) = \delta^j_i f^i_j(x)$ is an invariant, known as the divergence of the field $f(\cdot)$ at x.

Invariance of the divergence of a contravariant vector field is an automatic consequence of tensor notation, but tensor notation does not offer an explanation of what divergence measures. In fact, divergence of f at x_0 is the average value of $\langle f,\, x - x_0\rangle / \langle x - x_0,\, x - x_0\rangle$. This average must be interpreted in an invariant sense as an integral with respect to invariant measure over the surface or interior of the infinitesimal sphere $\langle x - x_0,\, x - x_0\rangle \le \rho^2$ centered at x_0. In this sense, $\rho d(x_0)$ is the surface-average radial component of the field on the surface of this sphere: $d(x_0)$ is the volume-average of $\langle f,\, x - x_0\rangle/\langle x - x_0,\, x - x_0\rangle$ over the interior of the sphere. In fluid mechanics, the divergence of the velocity vector at x is proportional to net the outward flow of fluid from x. Incompressible fluids are characterized by zero divergence at every point. Fields that have zero divergence are said to be solenoidal, Aris (1989, section 3.43).

0.12.3 *Rotational fields*

A rotational field is a mathematical model of vortex motion. The simplest rotational field is one in which the field is linear in x and orthogonal to x at all points. Now, $\langle x,\, Ax\rangle = x^T GAx$ is identically zero if and only if GA is skew-symmetric. Thus, a necessary and sufficient condition that the field $y^r = a^r_i x^i$ be purely rotational is that $a_{ik} = g_{ij}a^j_k$ be a skew-symmetric covariant tensor. More generally, a non-linear contravariant vector field f is said to be irrotational at x if its partial derivatives $f^i_r = \partial f^i/\partial x^r$ satisfy the symmetry condition $g_{ik}f^k_j(x) = g_{jk}f^k_i(x)$. A field that is irrotational at all points is said to be conservative.

In general, rotational motion and turning forces, such as angular velocity, angular momentum, magnetic fields and torque, are represented in physics by skew-symmetric matrices. For example, the positive, or anticlockwise, moment of the force (f^1, f^2) applied at the point (x^1, x^2) in the plane is equal to $x^1 f^2 - x^2 f^1$. More generally, the moment of a force is represented by the skew-symmetric contravariant tensor $x^i f^j - x^j f^i$. This is simply an alternative, and in many ways preferable, representation of the cross product or vector product $x \times f$ in $\mathcal{R}^3$.

0.12.4 *Curl of a covariant vector field*

Let f be a covariant vector field on an affine space $\mathcal{V}$. In statistics, a score function or vector estimating function is of this type. The components of f in $\mathcal{T}^*$, the dual of the tangent space, are denoted by $f_r(x)$. The partial derivatives of f form a covariant tensor of order two with components $f_{rs} = \partial f_r/\partial x^s$. Any square matrix can be decomposed into a symmetric and a skew-symmetric part:

$$2f_{rs} = (f_{rs} + f_{sr}) + (f_{rs} - f_{sr}).$$

The first term represents the components of a tensor in $\mathrm{sym}(\mathcal{T}^{*\otimes 2})$: the second term gives the components of a tensor in the complementary space $\mathrm{alt}(\mathcal{T}^{*\otimes 2})$. The latter component is called the curl of the field f, generalizing the familiar definition for vectors in $\mathcal{R}^3$. For further details, see Synge and Schild (1978, p. 246).

The term curl originates in Maxwell's work in electromagnetism, but this particular combination of derivatives had arisen earlier in work on the refraction of light (MacCullagh, 1846).

0.12.5 *Newton's second law*

Let $x^i \equiv x^i(t)$ be the position at time t of a particle of mass m subject to a force with components f^i. Newton's second law states that the force is equal to the rate of change of momentum. In symbols,

$$f^i = \frac{d}{dt}(mv^i) = \frac{d}{dt}(m\dot{x}^i).$$

This is a contravariant vector equation. If there are many particles in the system, we may sum these vector contributions, one vector for each particle. Then the total vector force is equal to the rate of change of the total momentum vector.

The motion of a rigid body is constrained to translation and rotation. The following equations, derived from Newton's second law, relate directly to rotary motion. The moment about the origin of a force f applied at x is given by the skew-symmetric tensor $M^{rs} = x^r f^s - x^s f^r$. The angular momentum, or moment of momentum, is given by

$$H^{rs} = m(x^r v^s - x^s v^r),$$

where v denotes the velocity vector. Differentiation of H, followed by application of Newton's second law, gives $dH^{rs}/dt = M^{rs}$. In words, the moment of the force on a particle is equal to the rate of change of angular momentum.

Further details of the application of tensor calculus to classical dynamics can be found in Synge and Schild (1978, chapter 5).

0.13 Invariants

An invariant is a function whose value is unaffected by a relevant class of transformations. The class of transformations is usually implied by the context. The principal idea is that fundamentally important aspects of a problem are those that are unaffected by unimportant or cosmetic details such as the choice of basis vectors or the choice of coordinate system in parametric statistics. Tensor notation, although paradoxically very much tied to coordinate systems, does generate invariants in an automatic fashion as tensors of order zero. It does not supply a geometric interpretation of such invariants.

The invariance argument is mathematically very powerful, and quite persuasive in many circumstances. But it can be carried to excess, particularly in problems where the statistical model has unintended or accidental symmetries that the physical process does not possess. For example, the multivariate normal model has rotational symmetries that physical processes usually do not possess. Likewise, the univariate Cauchy model has symmetries beyond the usual location and scale (Exercise ?.?).

To give a simple example of an invariant, let ω be a linear transformation from a vector space to itself. In terms of component vectors, the transformation may be written in the form

$$y^r = \omega_i^r x^i,$$

and ω_i^r are the components of a mixed tensor in $\mathcal{V} \otimes \mathcal{V}^*$. Now, δ_i^r are the components of the identity function, a mixed tensor in the same space. Since the tensor property is preserved under multiplication and contraction, it follows that

$$\lambda = \omega_i^r \delta_r^i$$

is a tensor of order zero, or an invariant. In matrix notation, a change of basis in $\mathcal{V}$ induces a tensorial transformation from the original Ω to a new matrix of the form $\bar{\Omega} = A^{-1}\Omega A$. This is known as a similarity transform or unitary transform (Dirac, 1958, ch. 26). In Cartan's terminology (Cartan, 1981, section 41), the matrices Ω and $\bar{\Omega}$ are said to be equivalent. It is shown in Exercise 1.? that the eigenvalues of a (1,1) tensor are invariant under this transformation. In other words, equivalent matrices have the same eigenvalues. In particular, $\lambda = \mathrm{tr}(\Omega)$ is equal to the sum of the eigenvalues and is invariant.

The second example is closer in spirit to the material in the following chapters in that it involves random variables in an explicit way. Let $\sigma^{i,j}$ be the covariance matrix of X, and consider the effect of making an *orthogonal* transformation from X to new variables $\bar{X} = AX$. In matrix notation Σ is transformed to $\bar{\Sigma} = A\Sigma A^T$. Since the characteristic polynomial of $\bar{\Sigma}$ is the same as the characteristic polynomial of Σ, it follows that the eigenvalues are invariant. This invariance holds only for orthogonal transformation of X. Sylvester's law of inertia states that only the rank and signature are invariant under linear transformation. Reverting to index notation, $\sigma^{i,j}\delta_{ij}$, the trace of Σ, is invariant under orthogonal transformation because both $\sigma^{i,j}$ and δ_{ij} transform as tensors under this group. δ_{ij} is not a tensor under the general linear group, so the indicated trace is not invariant under the larger group.

Log likelihood functions are automatically invariant under invertible transformation of the sample space. When we talk of invariance in this context, we usually mean invariance under a group of transformations on the parameter space. The log likelihood itself is invariant under re-parameterization, but its derivatives are not. The log likelihood ratio statistic $l(\hat{\theta}; y) - l(\theta; y)$ is invariant, and consequently its distribution is invariant. Chapter ? discusses ways of approximating the distribution in terms of invariants.

Example 27: Mahalanobis distance. Let $X = (X^1, \ldots, X^n)$ be the components of a random vector in an affine space $\mathcal{V}$ with tangent space $\mathcal{T}$. Example 10 shows that the covariance matrix $\kappa^{i,j}$ are the components of a tensor in $\mathcal{T}^{\otimes 2}$, and that the inverse covariance matrix $\kappa_{i,j}$ are the components of a tensor in $\mathcal{T}^* \otimes \mathcal{T}^*$. In other words, $\kappa_{i,j}$ is a metric tensor on $\mathcal{V}$. If X_1 and X_2 are two points in $\mathcal{V}$, then $d = X_1 - X_2$ lies in $\mathcal{T}$, and $d^i d^j$ are the components of a tensor in $\mathcal{T}^{\otimes 2}$. Consequently, the Mahalanobis squared distance $d^i d^j \kappa_{i,j}$ is invariant under affine transformation in $\mathcal{V}$.

0.14 Densities and invariant measures

A measure μ in a set $\mathcal{V}$ is said to be invariant under the group $\mathcal{G}$ acting on $\mathcal{V}$ if, for each $A \subset \mathcal{V}$ and $g \in \mathcal{G}$,

$$\mu(gA) = \mu(A)$$

where gA denotes the translation of A by g.

An invariant measure need not exist: if it exists it need not be unique. The set of translates of $x \in \mathcal{V}$ by $\mathcal{G}$ is called an *orbit*, and x is an *orbit representative*. If $\mathcal{V}$ comprises a single orbit in the sense that $\mathcal{V} = \{gx : g \in \mathcal{G}\}$, then the action is said to be *transitive*. If the action is transitive and if an invariant measure exists, then it is unique up to a scalar multiplier.

Three examples will illustrate the ideas. Let $\mathcal{V}$ be the finite set $\{1, \ldots, n\}$, and let $\mathcal{G}$ be the symmetric group of permutations. The action is transitive, and cardinality measure or counting measure on subsets is invariant. However, if $\mathcal{G}$ is the group of cyclic permutations, $\mathcal{V}$ comprises $(n-1)!$ orbits, and the action is not transitive. In this case, any measure constant on orbits is invariant. Second, Lebesgue measure in $\mathcal{R}$ is invariant under translation: there is no measure that is invariant under location-scale transformation on $\mathcal{R}$. A less obvious example is the following. Let $\mathcal{V}$ be the complex plane (omitting the real axis and with conjugate pairs identified). Let $\mathcal{G}$ be the group of fractional linear transformations on $\mathcal{V}$ of the form

$$gz = \frac{az + b}{cz + d},$$

with a, b, c, d real and $ad - bc \neq 0$. It should be checked that this is indeed a group, and that the 2×2 matrix with elements a, b, c, d is a representation of g, group composition corresponding to matrix multiplication. Invariant measure on subsets of the upper half plane, ($z = x + iy$ with $y > 0$), is given by

$$\mu(A) = \int_A dx\, dy / y^2$$

provided that this integral exists. It is not obvious, but it is nonetheless true, that $\mu(gA) = \mu(A)$. This measure is also invariant under all subgroups of $\mathcal{G}$, in particular the location-scale subgroup $(a, b) \circ (x, y) \mapsto (x + a, by)$, with $b > 0$, acting on the upper half plane.

0.15 Bibliographic notes

For an excellent orthodox and comprehensive introduction to finite-dimensional vector spaces, see Halmos (1958). For a more iconoclastic description in line with sections 1 and 2 of this chapter, see Hoffmann (1966).

We have chosen in this chapter to include all units of measurement in the definition of the basis vectors. As a consequence, component vectors are ordered sets of real numbers. This practice is not universal: for a forcefully argued contrary view in which symbols denote physical magnitudes, see Jeffreys and Jeffreys (1956, p. 3–4). In most treatments, particularly in undergraduate physics texts, it is unclear where the units of measurement reside. Kibble (1985, p. 4) gives an abbreviated discussion but his convention is not entirely clear on this point. Most mainstream mathematics texts such as Halmos (1958) or Birkoff and MacLane (1948) seem to regard units of measurement as a non-issue not worth discussing. Other than in examples, the basis vectors are never *defined* in physical terms, so this hands-off approach is mathematically quite defensible, perhaps even necessary. Hoffmann (1966, p. 57), however, is explicit in using the same convention as this book. Among statistics texts, the conventions are often not spelled out; an exception is Stone (1986) whose convention coincides with ours.

Tensor notation is widely used in applied mathematics, to areas such as fluid dynamics, elasticity, electromagnetism, relativity, and differential geometry. Definitions and notation vary to some extent from subject to subject and from author to author. For example, Jeffreys (1952) and Jeffreys and Jeffreys (1956) are concerned with the effect on equations of motion of rotating the axes or frame of reference. Consequently, their definition of what they call a Cartesian tensor refers only to the group of rigid motions (rotation and translation), and not, in general, to reflections or linear transformations.

Other useful references, again with a bias towards applications in physics, include, in order of first publication, McConnell (1931), Synge and Schild (1978), Schouten (1989), Aris (1989), Wrede (1972), Thomas (1965), Lawden (1968), and Richtmyer (1981). Most of these authors define a tensor as a set of quantities that obey the tensor transformation rules under a change of coordinates. With the exception of Richtmyer, there is little discussion of vector spaces and dual spaces.

In connection with differential geometry and calculus on manifolds, see the books by Eisenhart (1926), Weatherburn (1950), Sokolnikoff (1951), Stoker (1969), Lovelock and Rund (1989), Bishop and Goldberg (1980), Willmore (1982) and Spivak (1970).

The summation convention was first introduced by Einstein (1916) in his paper on the foundation of general relativity.

For a discussion of the geometry of generalized inverses, see Kruskal (1975), Rao (1973), or Stone (1987).

0.16 Further results and exercises 1

0.1 Vectors in $\mathcal{R}^2$ are sometimes specified in polar coordinates by magnitude and direction. All vectors have a definite magnitude. All except one have a definite direction. Which one?

0.2 In a non-metric space the notion of length or distance is absent. Bearing this in mind, answer question 7 of section 0.1.1.

0.3 Show that the 'parallelogram law of addition' for component vectors is a consequence of the definition of linear functionals in section 0.3.1.

0.4 Let $\{1, x, x^2\}$ be a basis for quadratic functions on $[0, 1]$. What are the components of $v(x) = 2x^2 - 3x + 1$ with respect to this basis? What are the components of v with respect to the basis $\{1, 2x - 1, 6x^2 - 6x + 1\}$?

0.5 For square-integrable functions on $[0, 1]$, define the inner product

$$\langle v_1, v_2 \rangle = \int_0^1 v_1(x)\, v_2(x)\, dx.$$

Show that for the first choice of basis in the preceding exercise,

$$G = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}, \qquad G^{-1} = \begin{pmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{pmatrix}.$$

Use equation (1.2) to find the components of $v(x) = x^2 - 3x - 1$ with respect to the two bases given in the preceding exercise.

0.6 Let $\mathcal{V}$ be the extended complex plane, and let the modulus and argument of v_j be denoted by $|v_j|$ and ϕ_j respectively. The point zero is defined to have argument zero. Multiplication by real scalars is defined in the standard way, but addition is defined as follows.

$$v_1 + v_2 = \frac{|v_1| + |v_2|}{1 - |v_1|\,|v_2|} e^{i\phi_1 + i\phi_2}.$$

Show that the associativity condition $v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$ is satisfied. Examine the conditions in section 2 of Halmos (1958) and explain why $\mathcal{V}$ is or is not a vector space.

0.7 Show that the eigenvalues of a real skew-symmetric matrix are purely imaginary ($\bar{\lambda} = -\lambda$). Hence deduce that a skew-symmetric matrix of odd order must have at least one zero eigenvalue.

0.8 Let $A^{ij} = a^i b^j - a^j b^i$ be a skew-symmetric matrix of order three. Show that the vector product $c = a \times b$, as defined in section 1.4.5, lies in the null space of A.

0.9 Let X be a $n \times p$ matrix of rank $q < p \le n$, and let W be symmetric positive definite of order n. Define two matrices P_1 and P_2, both of the form

$$P = X(X^T W X)^- X^T W,$$

but with different choices of generalized inverse. Show that P is a self-conjugate projection matrix if the generalized inverse is annihilating. On the assumption that both generalized inverses are annihilating, and that the rank of $X^T W X$ is the same as the rank of X, show that $P_1 = P_2$.

0.10 Show that the identity matrix is a generalized inverse of any projection matrix. Show also that a projection matrix, not necessarily symmetric, is its own annihilating generalized inverse.

0.11 Let ω^{ij}, with inverse ω_{ij} be the components of a $p \times p$ matrix of rank p. Show that

$$\gamma^{rs,ij} = \omega^{ri}\omega^{sj} + \omega^{rj}\omega^{si},$$

regarded as a $p^2 \times p^2$ matrix with rows indexed by (r, s) and columns by (i, j), is symmetric with rank $p(p+1)/2$. Show also that $\omega_{ri}\omega_{sj}/2$ is a generalized inverse.

0.12 *Singular values of a matrix*: Let the $p \times q$ matrix A, with components a_i^r, be considered as defining a linear transformation from the domain, $\mathcal{R}^q$, to the range in $\mathcal{R}^p$. Interpret the *singular values* of A as invariants under independent orthogonal transformation of the domain and range spaces. For the definition of singular values and their application in numerical linear algebra, see Chambers (1977, Section 5.e).

0.13 Let A, A^{-1} and X be symmetric matrices with components a_{ij}, a^{ij} and x_{ij} respectively. Show that the Taylor expansion for the log determinant of $A+X$ about the origin may be written

$$\begin{aligned}\log\det(A+X) = \log\det(A) + x_{ij}a^{ij} - x_{ij}x_{kl}a^{ik}a^{jl}/2 \\ + x_{ij}x_{kl}x_{mn}a^{ik}a^{jm}a^{ln}/3 + \ldots\end{aligned}$$

Describe the form of the general term in this expansion. Generalize this expansion to asymmetric matrices.

0.14 *Hooke's Law*: In the mechanics of deformable solids, the components of the *stress tensor*, p_{ij}, measure force per unit area in the following sense. Let e_i be the unit vector in the ith coordinate direction and let $\bar{e}_i$ be the orthogonal plane. Then p_{ii} is the force per unit area normal to the plane, also called the *normal stress*, and p_{ij}, $j \neq i$ are the *shear stresses* acting in the plane. The components of the *strain tensor*, q_{ij}, which are dimensionless, measure percentage deformation or percentage change in length. Both arrays are symmetric tensors under the orthogonal group.

In the case of *elastic* deformation of an *isotropic* material, Hooke's law in its most general form states that the relationship between stress and strain is linear. Thus,

$$p_{rs} = b^{ij}_{rs}q_{ij},$$

where b^{ij}_{rs} is an isotropic fourth-order tensor given by

$$b^{ij}_{rs} = \lambda\delta^{ij}\delta_{rs} + 2\mu\delta^i_r\delta^j_s,$$

for constants λ, μ that are characteristic of the material.

Show that the inverse relationship giving the strains in terms of the stresses may be written in the form

$$q_{ij} = \left(\lambda'\delta^{rs}\delta_{ij} + 2\mu'\delta^r_i\delta^s_j\right)p_{rs},$$

where the new constants are given by

$$\mu' = \frac{1}{4\mu}, \qquad \lambda' + 2\mu' = \frac{\lambda+\mu}{\mu(3\lambda+2\mu)} = E^{-1}.$$

In the terminology used in Mechanics, E is known as *Young's modulus* or *modulus of elasticity*, μ is called the *rigidity* or *shear modulus* and $\sigma = \lambda/\{2(\lambda+\mu)\}$ is *Poisson's ratio*. Note that $E = 2(1+\sigma)\mu$, implying that two independent constants entirely determine the three-dimensional elastic properties of the material. (Murnaghan, 1951, Chapters 3,4; Jeffreys & Jeffreys 1956, Section 3.10; Drucker, 1967, Chapter 12).

0.15 *Alternating tensor*: The permutation symbol or alternating tensor of order n is defined as follows:

$$\epsilon^{i_1,\ldots,i_n}_{r_1,\ldots,r_n} = \begin{cases} 1 & \text{if } r_1,\ldots,r_n \text{ is an even permutation of } i_1,\ldots,i_n; \\ -1 & \text{if } r_1,\ldots,r_n \text{ is an odd permutation of } i_1,\ldots,i_n; \\ 0 & \text{otherwise.}\end{cases}$$

The value zero is understood to apply if any subscript or any superscript is duplicated. Show that, for any $n \times n$ matrix A

$$\epsilon^{i_1,\ldots,i_n}_{1,\ldots,n}a^1_{i_1}a^2_{i_2}\cdots a^n_{i_n} = \det(A).$$

Let $\mathcal{V}$ be a vector space of dimension n. Let ϵ be the component matrix of a transformation from $\mathcal{V}^{\otimes n}$ to itself as follows:

$$y^{i_1,\ldots,i_n} = \epsilon^{i_1,\ldots,i_n}_{r_1,\ldots,r_n}x^{r_1,\ldots,r_n}.$$

where x and y are tensors in $\mathcal{V}^{\otimes n}$. Now introduce a change of basis vectors in $\mathcal{V}$ as follows: $\bar{e}_r = a^i_r e_i$. Prove that the component vectors in the new coordinate system satisfy

$$\bar{y}^{i_1,\ldots,i_n} = \epsilon^{i_1,\ldots,i_n}_{r_1,\ldots,r_n}\bar{x}^{r_1,\ldots,r_n},$$

showing that the components of ϵ are the same in all coordinate systems, (Synge and Schild, Ch. 7).

0.16 Define an alternative permutation symbol by

$$\epsilon_{r_1,\ldots,r_n} = \epsilon^{1,\ldots,n}_{r_1,\ldots,r_n}.$$

Show that this symbol transforms as an isotropic tensor under O^+, the orthogonal group with positive determinant (Ames and Murnaghan 1929, p. 440), (Wrede 1972, p. 85), (Jeffreys and Jeffreys, 1956, Sections 2.07, 3.03).

0.17 Explain why the multinomial generalized inverse matrix given by Watson (1996) is not the same as that given at the end of section 1.8.

0.18 *Rigid body dynamics*: Let $y(x;t)$ be the position at time t of the point originally at x at time $t=0$. Thus, $y(x;0)=x$. For each t, $y(\cdot;t)$ is a transformation of an affine space $\mathcal{V}$ to itself satisfying the rigid-body condition $|y(x_1;t)-y(x_2;t)|^2 = |x_1-x_2|^2$. Deduce that $y(x;t)=b(t)+A(t)x$, where $b(t)\in\mathcal{V}$, and $A(t)$ is an orthogonal matrix.

Let $v=v(x;t)=\dot{y}(x;t)$ be the velocity vector of x. Show that, for any pair of points, x_1, x_2, the velocity vectors satisfy the orthogonality condition

$$\langle v_1 - v_2,\, y_1 - y_2\rangle = 0.$$

One solution to this differential equation is $v_1(t)=v_2(t)$, implying that all points move in parallel and in unison. Suppose, however, that the motion is such that at time t_0, one point, $y(x_0;t_0)$, is stationary, i.e. $\dot{y}(x_0;t_0)=0$. Show then that for each $y=y(x;t_0)$,

$$\langle \dot{A}A^{-1}(y-y_0),\, y-y_0\rangle = 0,$$

and hence that $G\dot{A}A^{-1}$ is skew-symmetric. Deduce that the covariant components of the velocity vector at time t_0 satisfy

$$v_i = g_{ij}v^j = \omega_{ij}(y^j - y_0^j),$$

where ω_{ir} is a skew-symmetric covariant tensor.

CHAPTER 1

Index notation

1.1 Introduction

It is a fact not widely acknowledged that, with appropriate choice of notation, many multivariate statistical calculations can be made simpler and more transparent than the corresponding univariate calculations. This simplicity is achieved through the systematic use of index notation and special arrays called tensors. For reasons that are given in the following sections, matrix notation, a reliable workhorse for many second-order calculations, is totally unsuitable for more complicated calculations involving either non-linear functions or higher-order moments. The aim of this book is to explain how index notation simplifies many statistical calculations, particularly those involving moments or cumulants of non-linear functions. Other applications where index notation greatly simplifies matters include k-statistics, Edgeworth and conditional Edgeworth approximations, saddlepoint and Laplace approximations, calculations involving conditional cumulants, moments of maximum likelihood estimators, likelihood ratio statistics and the construction of ancillary statistics. These topics are the subject matter of later chapters.

In some ways, the most obvious and, at least initially, most disconcerting aspect of index notation is that the components of the vector of primary interest, usually a parameter, θ, or a random variable, X, are indexed using superscripts. Thus, θ^2, the second component of the vector θ, is not to be confused with the square of any component. For this reason, powers are best avoided unless the context leaves no room for ambiguity, and the square of θ^2 is written simply as $\theta^2\theta^2$. In view of the considerable advantages achieved, this is a very modest premium to pay.

1.2 The summation convention

Index notation is a convention for the manipulation of multi-dimensional arrays. The elements of these arrays are called either *components* or *coefficients* depending on the context. In the context of parametric inference and in manipulations associated with likelihood functions, it is appropriate to take the unknown parameter as the vector of interest: see the first example in Section 2.4. Here, however, we take as our vector of interest the p-dimensional random variable X with components $X^1, \ldots, X^p$. In this context, arrays of constants used in the formation of linear combinations are called *coefficients*. This terminology is merely a matter of convention but it appears to be useful and the notation does emphasize it. Thus, for example, $\kappa^i = E(X^i)$ is a one-dimensional array whose components are the means of the components of X and $\kappa^{ij} = E(X^i X^j)$ is a two-dimensional array whose components are functions of the joint distributions of pairs of variables.

Probably the most convenient aspect of index notation is the implied summation over any index repeated once as a superscript and once as a subscript. The range of summation is not stated explicitly but is implied by the positions of the repeated index and by conventions regarding the range of the index. Thus,

$$a_i X^i \equiv \sum_{i=1}^{p} a_i X^i \tag{1.1}$$

specifies a linear combination of the Xs with coefficients $a_1, \ldots, a_p$. Quadratic and cubic forms in

X with coefficients a_{ij} and a_{ijk} are written in the form

$$a_{ij}X^iX^j \quad \text{and} \quad a_{ijk}X^iX^jX^k \tag{1.2}$$

and the extension to homogeneous polynomials of arbitrary degree is immediate.

For the sake of simplicity, and with no loss of generality, we take all multiply-indexed arrays to be symmetric under index permutation but, of course, subscripts may not be interchanged with superscripts. The value of this convention is clearly apparent when we deal with scalars such as $a_{ij}a_{kl}\omega^{ijkl}$, which, by convention only, is the same as $a_{ik}a_{jl}\omega^{ijkl}$ and $a_{il}a_{jk}\omega^{ijkl}$. For instance, if $p = 2$ and $a_{ij} = \delta_{ij} = 1$ if $i = j$ and 0 otherwise, then, without the convention,

$$a_{ij}a_{kl}\omega^{ijkl} - a_{ik}a_{jl}\omega^{ijkl} = \omega^{1122} + \omega^{2211} - \omega^{1212} - \omega^{2121}$$

and this is not zero unless ω^{ijkl} is symmetric under index permutation.

Expressions (2.1) and (2.2) produce one-dimensional or scalar quantities, in this case scalar random variables. Suppose instead, we wish to construct a vector random variable Y with components $Y^1, \ldots, Y^q$, each of which is linear in X, we may write

$$Y^r = a^r_i X^i \tag{1.3}$$

and $r = 1, \ldots, q$ is known as a *free* index. Similarly, if the components of Y are homogeneous quadratic forms in X, we may write

$$Y^r = a^r_{ij}X^iX^j. \tag{1.4}$$

Non-homogeneous quadratic polynomials in X may be written in the form

$$Y^r = a^r + a^r_i X^i + a^r_{ij}X^iX^j.$$

Where two sets of indices are required, as in (2.3) and (2.4), one referring to the components of X and the other to the components of Y, we use the sets of indices $i, j, k, \ldots$ and $r, s, t, \ldots$. Occasionally it will be necessary to introduce a third set, $\alpha, \beta, \gamma, \ldots$ but this usage will be kept to a minimum.

All of the above expressions could, with varying degrees of difficulty, be written using matrix notation. For example, (2.1) is typically written as $\mathbf{a}^T\mathbf{X}$ where $\mathbf{a}$ and $\mathbf{X}$ are column vectors; the quadratic expression in (2.2) is written $\mathbf{X}^T\mathbf{A}\mathbf{X}$ where $\mathbf{A}$ is symmetric, and (2.3) becomes $\mathbf{Y} = \mathbf{A}^*\mathbf{X}$ where $\mathbf{A}^*$ is of order $q \times p$. From these examples, it is evident that there is a relationship of sorts between column vectors and the use of superscripts, but the notation $\mathbf{X}^T\mathbf{A}\mathbf{X}$ for $a_{ij}X^iX^j$ violates the relationship. The most useful distinction is not in fact between rows and columns but between coefficients and components and it is for this reason that index notation is preferred here.

1.3 Tensors

The term *tensor* is used in this book in a well-defined sense, similar in spirit to its meaning in differential geometry but with minor differences in detail. It is not used as a synonym for array, index notation or the summation convention. A cumulant tensor, for example, is a symmetric array whose elements are functions of the joint distribution of components of the random variable of interest, X say. The values of these elements in any one coordinate system are real numbers but, when we describe the array as a tensor, we mean that the values in one coordinate system, Y say, can be obtained from those in any other system, X say, by the application of a particular transformation formula. The nature of this transformation is the subject of Sections 3.4 and 4.5, and in fact, we consider not just changes of basis, but also non-invertible transformations.

When we use the adjectives *covariant* and *contravariant* in reference to tensors, we refer to the way in which the arrays transform under a change of variables from the original x to new variables y. In statistical calculations connected with likelihood functions, x and y are typically parameter vectors but in Chapters 2 and 3, x and y refer to random variables. To define the adjectives covariant and contravariant more precisely, we suppose that ω is a d-dimensional array whose elements are functions of the components of x, taken d at a time. We write $\omega = \omega^{i_1 i_2 \cdots i_d}$ where the d components need not be distinct. Consider the transformation $y = g(x)$ from x to new variables $y = y^1, \ldots, y^p$ and let $a_i^r \equiv a_i^r(x) = \partial y^r / \partial x^i$ have full rank for all x. If $\bar{\omega}$, the value of ω for the transformed variables, satisfies

$$\bar{\omega}^{r_1 r_2 \cdots r_d} = a_{i_1}^{r_1} a_{i_2}^{r_2} ... a_{i_d}^{r_d} \omega^{i_1 i_2 \cdots i_d} \tag{1.5}$$

then ω is said to be a contravariant tensor. On the other hand, if ω is a covariant tensor, we write $\omega = \omega_{i_1 i_2 \cdots i_d}$ and the transformation law for covariant tensors is

$$\bar{\omega}_{r_1 r_2 \cdots r_d} = b_{r_1}^{i_1} b_{r_2}^{i_2} ... b_{r_d}^{i_d} \omega_{i_1 i_2 \cdots i_d} \tag{1.6}$$

where $b_r^i = \partial x^i / \partial y^r$, the matrix inverse of a_i^r, satisfies $a_i^r b_r^j = \delta_i^j = a_r^j b_i^r$.

The function $g(.)$ is assumed to be an element of some group, either specified explicitly or, more commonly, to be inferred from the statistical context. For example, when dealing with transformations of random variables or their cumulants, we usually work with the general linear group (2.3) or the general affine group (2.8). Occasionally, we also work with the smaller orthogonal group, but when we do so, the group will be stated explicitly so that the conclusions can be contrasted with those for the general linear or affine groups. On the other hand, when dealing with possible transformations of a vector of parameters, it is natural to consider non-linear but invertible transformations and $g(.)$ is then assumed to be a member of this much larger group. In other words, when we say that an array of functions is a tensor, the statement has a well defined meaning only when the group of transformations is specified or understood.

It is possible to define hybrid tensors having both subscripts and superscripts that transform in the covariant and contravariant manner respectively. For example, if ω^{ij} and ω_{ijk} are both tensors, then the product $\gamma_{klm}^{ij} = \omega^{ij}\omega_{klm}$ is a tensor of covariant order 3 and contravariant order 2. Furthermore, we may sum over pairs of indices, a process known as *contraction*, giving

$$\gamma_{kl}^{i} = \gamma_{klj}^{ij} = \omega^{ij}\omega_{klj}.$$

A straightforward calculation shows that γ_{kl}^i is a tensor because, under transformation of variables, the transformed value is

$$\bar{\gamma}_{klm}^{ij} = \gamma_{tuv}^{rs} a_r^i a_s^j b_k^t b_l^u b_m^v$$

and hence, summation over $m = j$ gives

$$\bar{\gamma}_{kl}^{i} = \gamma_{tu}^{r} a_r^i b_k^t b_l^u.$$

Thus, the tensor transformation property is preserved under multiplication and under contraction. An important consequence of this property is that scalars formed by contraction of tensors must be invariants. In effect, they must satisfy the transformation law of zero order tensors. See Section 2.6.

One of the problems associated with tensor notation is that it is difficult to find a satisfactory notation for tensors of arbitrary order. The usual device is to use subscripted indices as in (2.5) and (2.6), but this notation is aesthetically unpleasant and is not particularly easy to read. For these reasons, subscripted indices will be avoided in the remainder of this book. Usually we give explicit expressions involving up to three or four indices. The reader is then expected to infer the necessary generalization, which is of the type (2.5), (2.6) if we work with tensors but is usually more complicated if we work with arbitrary arrays.

1.4 Examples

In this and in the following chapter, X and Y are random variables but when we work with log likelihood derivatives, it is more appropriate to contemplate transformation of the parameter vector and the terms covariant and contravariant then refer to parameter transformations and not to data transformations. To take a simple example, relevant to statistical theory, let $l(\theta; Z) = \log f_Z(Z; \theta)$ be the log likelihood function for $\theta = \theta^1, \ldots, \theta^p$ based on observations Z. The partial derivatives of l with respect to the components of θ may be written

$$\begin{aligned} U_r(\theta) &= \partial l(\theta; Z)/\partial\theta^r \\ U_{rs}(\theta) &= \partial^2 l(\theta; Z)/\partial\theta^r\partial\theta^s \end{aligned}$$

and so on. The maximum likelihood estimate of θ satisfies $U_r(\hat{\theta}) = 0$ and the observed information for θ is $I_{rs} = -U_{rs}(\hat{\theta})$, with matrix inverse I^{rs}. Suppose now that we were to re-parameterize in terms of $\phi = \phi^1, \ldots, \phi^p$. If we denote by an asterisk derivatives with respect to ϕ, we have

$$\begin{aligned} U^*_r &= \theta^i_r U_i, \qquad & U^*_{rs} &= \theta^i_r \theta^j_s U_{ij} + \theta^i_{rs} U_i, \\ I^*_{rs} &= \theta^i_r \theta^j_s I_{ij}, \qquad & I^{*rs} &= \phi^r_i \phi^s_j I^{ij} \end{aligned} \tag{1.7}$$

where

$$\theta^i_r = \partial\theta^i/\partial\phi^r, \qquad \theta^i_{rs} = \partial^2\theta^i/\partial\phi^r\partial\phi^s$$

and θ^i_r is assumed to have full rank with matrix inverse $\phi^r_i = \partial\phi^r/\partial\theta^i$. Arrays that transform like U_r, I_{rs} and I^{rs} are tensors, the first two being covariant of orders 1 and 2 respectively and the third being contravariant of order 2. The second derivative, $U_{rs}(\theta)$, is not a tensor on account of the presence of second derivatives with respect to θ in the above transformation law. Note also that the array U^*_{rs} cannot be obtained by transforming the array U_{rs} alone: it is necessary also to know the value of the array U_r. However $E\{U_{rs}(\theta); \theta\}$, the Fisher information for θ, is a tensor because the second term in U^*_{rs} has mean zero at the true θ.

To take a second example, closer in spirit to the material in the following two chapters, let $X = X^1, \ldots, X^p$ have mean vector $\kappa^i = E(X^i)$ and covariance matrix

$$\kappa^{i,j} = \text{cov}(X^i, X^j) = E(X^i X^j) - E(X^i)E(X^j).$$

Suppose we make an affine transformation to new variables $Y = Y^1, \ldots, Y^q$, where

$$Y^r = a^r + a^r_i X^i. \tag{1.8}$$

The mean vector and covariance matrix of Y are easily seen to be

$$a^r + a^r_i \kappa^i \quad \text{and} \quad a^r_i a^s_j \kappa^{i,j}$$

where $a^r_i = \partial Y^r/\partial X^i$. Thus, even though the transformation may not be invertible, the covariance array transforms like a contravariant tensor. Arrays that transform in this manner, but only under linear or affine transformation of X, are sometimes called *Cartesian tensors* (Jeffreys, 1952). Such transformations are of special interest because $a^r_i = \partial Y^r/\partial X^i$ does not depend on X. It will be shown that cumulants of order two or more are not tensors in the sense usually understood in differential geometry, but they do behave as tensors under the general affine group (2.8). Under non-linear transformation of X, the cumulants transform in a more complicated way as discussed in Section 4.4.

Tensors whose components are unaffected by coordinate transformation are called *isotropic.* This terminology is most commonly used in Mechanics and in the physics of fluids, where all three coordinate axes are measured in the same units. In these contexts, the two groups of most relevance are the orthogonal group, O, and the orthogonal group with positive determinant, O^+. In either case, δ^i_j, δ_{ij} and δ^{ij} are isotropic tensors. There is exactly one isotropic third-order tensor under O^+ (Exercise 1.22). However, this tensor, called the *alternating tensor*, is anti-symmetrical and does not occur in the remainder of this book. All fourth-order isotropic tensors are functions of the three second-order isotropic tensors (Jeffreys, 1952, Chapter 7). The only symmetrical isotropic fourth-order tensors are

$$\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}, \quad \delta^{ij}\delta^{kl} + \delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk},$$
$$\delta^{ij}\delta_{kl} \quad \text{and} \quad \delta^i_k\delta^j_l + \delta^i_l\delta^j_k,$$

(Thomas, 1965, Section 7). Isotropic tensors play an important role in physics (see Exercise 1.21) but only a minor role in statistics.

1.5 Elementary matrix theory

For later use, we state here without detailed proof, some elementary matrix-theory results using tensor notation and terminology. Our main interest lies in matrix inverses and spectral decompositions or eigenvalue decompositions of real symmetric matrices. We consider first the tensorial properties of generalized inverse matrices.

1.5.1 *Generalized inverse matrices*

Let ω_{ij} be a symmetric covariant tensor written as Ω using matrix notation. A generalized inverse of Ω is any matrix Ω^- satisfying

$$\Omega\Omega^-\Omega = \Omega, \tag{1.9}$$

implying that $\Omega\Omega^-$ acts as the identity on the range of Ω. It follows that $\text{rank}(\Omega\Omega^-) \geq \text{rank}(\Omega)$, and hence that $\text{rank}(\Omega^-) \geq \text{rank}(\Omega)$. Post-multiplication of (2.9) by Ω^- gives $(\Omega\Omega^-)(\Omega\Omega^-) = \Omega\Omega^-$. In other words, $\Omega\Omega^-$ is idempotent and hence

$$\text{tr}(\Omega\Omega^-) = \text{rank}(\Omega\Omega^-) = \text{rank}(\Omega). \tag{1.10}$$

This conclusion is independent of the choice of generalized inverse. In addition, if Ω^- is a generalized inverse of Ω and if $\mathbf{A}$ is any full-rank matrix, then $\mathbf{A}\Omega^-\mathbf{A}$ is a generalized inverse of $\mathbf{A}^{-1}\Omega\mathbf{A}^{-1}$. The last result follows directly from the definition (2.9).

Reverting now to tensor notation, we write the generalized matrix inverse of ω_{ij} as ω^{ij}. The inverse can in fact always be chosen to be symmetric, so that this notation does not conflict with our convention regarding symmetry. It follows from (2.10) that, whatever the choice of inverse,

$$\text{rank}(\omega_{ij}) = \omega_{ij}\omega^{ij}.$$

In addition, if ω^{ij} is any generalized inverse of ω_{ij}, then $a^r_i a^s_j \omega^{ij}$ is a generalized inverse of $b^i_r b^j_s \omega_{ij}$, where a^i_r is a full rank matrix with inverse b^r_i. In other words, ω^{ij} is a contravariant tensor.

These arguments are entirely independent of the choice of generalized inverse. Occasionally, however, it is convenient to choose a generalized inverse with the additional property that

$$\Omega^-\Omega\Omega^- = \Omega^- \tag{1.11}$$

implying that $\text{rank}(\Omega^-) = \text{rank}(\Omega)$. In other words, Ω is a generalized inverse of Ω^-. The symmetry of the tensor formulae is greatly enhanced if the generalized inverse matrix is chosen to have the property (2.11) for then we need not distinguish between ω^{ij} as a generalized inverse matrix of ω_{ij} and ω_{ij} as a generalized inverse of ω^{ij}.

In fact, it is always possible to choose a generalized inverse with the properties (2.9) and (2.11) and having the additional symmetry property that

$$\Omega^- \Omega = (\Omega^- \Omega)^T, \quad \Omega \Omega^- = (\Omega \Omega^-)^T.$$

Such a generalized inverse is unique and is known as the Moore-Penrose inverse (Rao, 1973, Section 1b). See also Exercise 1.10.

Conditions (2.9), (2.11) are entirely natural whether Ω is the matrix representation of a symmetric covariant tensor, a symmetric contravariant tensor or an asymmetric (1,1) tensor. On the other hand, the symmetry conditions, as stated above, appear to be quite unnatural if Ω is the matrix representation of an asymmetric (1,1) tensor, but otherwise, the conditions seem sensible. The symmetry condition arises naturally if the usual Euclidean inner product with equal weights is used to determine orthogonality: see Kruskal (1975, Section 6). On the other hand, if a weighted inner product is appropriate, as it often is in statistical applications, then the symmetry condition would seem to be inappropriate. See, for example, Exercise (2.11).

1.5.2 *Spectral decomposition*

Any real symmetric covariant tensor ω_{ij} may be written in the form

$$\omega_{ij} = \sigma_i^r \sigma_j^s \Lambda_{rs} \tag{1.12}$$

where $\Lambda_{rr} = \lambda_r$, a real number, $\Lambda_{rs} = 0$ for $r \neq s$ and σ_i^r is a real orthogonal matrix satisfying

$$\sigma_i^r \sigma_j^s \delta_{rs} = \delta_{ij}.$$

The values $\lambda_1, \ldots, \lambda_p$ are known as the eigenvalues of ω_{ij} and (2.12) is known as the eigenvalue decomposition or spectral decomposition of ω. This decomposition implies that the quadratic form $Q = \omega_{ij} x^i x^j$ may be written as $Q = \sum_1^p \lambda_r (y^r)^2$ where $y^r = \sigma_i^r x^i$ is an orthogonal transformation of x. The set of eigenvalues is unique but evidently the representation (2.12) is not unique because we may at least permute the components of y. Further, if some of the eigenvalues are equal, say $\lambda_1 = \lambda_2$, any orthogonal transformation of the components (y^1, y^2) satisfies (2.12).

Under orthogonal transformation of x, ω_{ij} transforms to $\bar{\omega}_{ij} = a_i^k a_j^l \omega_{kl}$ where a_i^k is orthogonal. The spectral decomposition (2.12) then becomes

$$\bar{\omega}_{ij} = (a_i^k \sigma_k^r)(a_j^k \sigma_l^s) \Lambda_{rs} \tag{1.13}$$

where $a_i^k \sigma_k^r$ is an orthogonal matrix. On comparing (2.12) with (2.13) we see that the set of eigenvalues of ω_{ij} is invariant under orthogonal transformation of coordinates. The eigenvalues are not invariant under arbitrary nonsingular transformation because $a_i^k \sigma_k^r$ is not, in general, orthogonal unless a_i^k is orthogonal.

Consider now the alternative decomposition

$$\omega_{ij} = \tau_i^r \tau_j^s \epsilon_{rs} \tag{1.14}$$

where $\epsilon_{rr} = \pm 1$ or zero, $\epsilon_{rs} = 0$ for $r \neq s$ and no constraints are imposed on τ_i^r other than that it should be real and have full rank. The existence of such a decomposition follows from (2.12). Again, the representation (2.14) is not unique because, if we write $y^r = \tau_i^r x^i$, then $Q = \omega_{ij} x^i x^j$ becomes

$$Q = \sum{}^{+} (y^r)^2 - \sum{}^{-} (y^r)^2 \tag{1.15}$$

where the first sum is over those y^r for which $\epsilon_{rr} = +1$ and the second sum is over those components for which $\epsilon_{rr} = -1$. Two orthogonal transformations, one for the components y^r

for which $\epsilon_{rr} = +1$ and one for the components for which $\epsilon_{rr} = -1$, leave (2.15) unaffected. Furthermore, the components for which $\epsilon_{rr} = 0$ may be transformed linearly and all components may be permuted without affecting the values of ϵ_{rs} in (2.14). At most, the order of the diagonal elements, ϵ_{rr} can be changed by the transformations listed above.

Under linear transformation of x, (2.14) becomes

$$\bar{\omega}_{ij} = (a_i^k \tau_k^r)(a_j^l \tau_l^s)\epsilon_{rs},$$

so that the matrix rank

$$\omega_{ij}\omega^{ij} = \epsilon_{rs}\epsilon^{rs}$$

and signature,

$$\delta^{rs}\epsilon_{rs} = \sum \epsilon_{rr}$$

are invariant functions of the covariant tensor ω_{ij}. This result is known as Sylvester's law of inertia (Gantmacher, 1960, Chapter X; Cartan, 1981, Section 4).

The geometrical interpretation of Sylvester's law is that the equation

$$x^0 = Q = \omega_{ij}x^i x^j$$

describes a hypersurface of dimension p in R^{p+1} and the qualitative aspects of the shape of this surface that are invariant under linear transformation of $x^1, \ldots, x^p$ are the numbers of positive and negative principal curvatures. This makes good sense because the effect of such a transformation is to rotate the coordinates and to re-define distance on the surface. The surface, in a sense, remains intact. If we were to change the sign of x^0, the positive and negative curvatures would be reversed.

In the particular case where ω_{ij} is positive definite of full rank, the matrix τ_i^r in (2.14) is known as a matrix square root of ω. For this case, if σ_r^s is an orthogonal matrix, then $\tau_i^r\sigma_r^s$ is also a matrix square root of ω_{ij}. Subject to this choice of orthogonal transformation, the matrix square root is unique.

1.6 Invariants

An invariant is a function whose value is unaffected by transformations within a specified class or group. To take a simple example, let ω_i^r be a mixed tensor whose value under linear transformation becomes

$$\bar{\omega}_i^r = a_s^r \omega_j^s b_i^j,$$

where $a_s^r b_i^s = \delta_i^r$. In matrix notation, $\bar{\Omega} = \mathbf{A}\Omega\mathbf{A}^{-1}$ is known as a *similarity* transformation or *unitary* transformation, (Dirac, 1958, Chapter 26). In Cartan's terminology (Cartan, 1981, Section 41), the matrices Ω and $\bar{\Omega}$ are said to be *equivalent.* It is an elementary exercise to show that $\mathrm{tr}(\bar{\Omega}) = \mathrm{tr}(\Omega)$, so that the sum of the eigenvalues of a real (1,1) tensor is (a) real and (b) invariant. The same is true of any symmetric polynomial function of the eigenvalues. In particular, the determinant is invariant. For an interpretation, see Exercise 1.13.

The second example is closer in spirit to the material in the following chapters in the sense that it involves random variables in an explicit way. Let $\kappa^{i,j}$ be the covariance matrix of the components of X and consider the effect on $\kappa^{i,j}$ of making an orthogonal transformation of X. By (2.13), the set of eigenvalues is unaffected. Thus the set of eigenvalues is an invariant of a symmetric contravariant tensor, but only within the orthogonal group. Only the rank and signature are invariant under nonsingular linear or affine transformation. Other examples of invariant functions of the cumulants are given in Section 3.8.

These examples pinpoint one serious weakness of matrix notation, namely that no notational distinction is made between a (1,1) tensor whose eigenvalues are invariant under the full linear group, and a (2,0) tensor, invariably symmetric, wwhose eigenvalues are invariant only under the smaller orthogonal group.

The log likelihood function itself is invariant under arbitrary smooth parameter transformation, not necessarily linear. Under nonsingular transformation of the data, the log likelihood is not invariant but transforms to $l(\theta; z) + c(z)$ where $c(z)$ is the log determinant of the Jacobian of the transformation. However $l(\hat{\theta}; z) - l(\theta; z)$, the maximized log likelihood ratio statistic, is invariant under transformation of the data. Some authors define the log likelihood as the equivalence class of all functions that differ from $l(\theta; z)$ by a function of z and, in this sense, the log likelihood function is an invariant.

To test the simple hypothesis that θ takes on some specified value, say θ_0, it is desirable in principle to use an invariant test statistic because consistency of the observed data with the hypothesized value θ_0 is independent of the coordinate system used to describe the null hypothesis value. Examples of invariant test statistics include the likelihood ratio statistic, $l(\hat{\theta}; z) - l(\theta_0; z)$, and the quadratic score statistic, $U_r U_s i^{rs}$, where i^{rs} is the matrix inverse of $-E\{U_{rs}; \theta\}$.

One of the main reasons for working with tensors, as opposed to arbitrary arrays of functions, is that it is easy to recognize and construct invariants. For example, any scalar derived from tensors by the process of contraction is automatically an invariant. This is a consequence of the tensor transformation properties (2.5) and (2.6). If the arrays are tensors only in some restricted sense, say under linear transformation only, then any derived scalars are invariants in the same restricted sense.

By way of illustration, consider the group of orthogonal transformations and suppose that the array ω^{ij} satisfies the transformation laws of a tensor under orthogonal transformation of X. Since we are dealing with orthogonal transformations only, and not arbitrary linear transformations, it follows that δ^{ij} and δ_{ij} are tensors. This follows from

$$a_i^r a_j^s \delta^{ij} = \delta^{rs}$$

where a_i^r is any orthogonal matrix. Note, however, that δ^{ijk} and δ_{ijk} are not tensors (Exercise 1.19). Hence the scalars

$$\omega^{ij}\delta_{ij}, \quad \omega^{ij}\omega^{kl}\delta_{ik}\delta_{jl}, \ldots$$

are invariants. In terms of the eigenvalues of ω^{ij}, these scalars may be written as power sums,

$$\sum_j \lambda_j, \quad \sum_j \lambda_j^2, \ldots.$$

Another function invariant under nonsingular linear transformation is the matrix rank, which may be written $\omega^{ij}\omega_{ij}$, where ω_{ij} is any generalized inverse of ω^{ij}. However not every invariant function can easily be derived by means of tensor-like manipulations of the type described here. For example, Sylvester's law of inertia states that the numbers of positive and negative eigenvalues of a real symmetric matrix are invariant under nonsingular linear transformation. In other words, if a_i^r is a nonsingular matrix then the sign pattern of the eigenvalues of ω^{ij} is the same as the sign pattern of the eigenvalues of $a_i^r a_j^s \omega^{ij}$. There appears to be no simple way to deduce this result from tensor-like manipulations alone. See, however, Exercise 1.9 where the signature is derived by introducing an auxiliary tensor ω_{ij}^+ and forming the invariant $\omega^{ij}\omega_{ij}^+$.

1.7 Direct product spaces

1.7.1 *Kronecker product*

Suppose that $Y^1, \ldots, Y^n$ are independent and identically distributed vector-valued random variables, each having p components. Should we so wish, we can regard $\mathbf{Y} = (Y^1, \ldots, Y^n)$ as a point in R^{np}, but usually it is preferable to consider $\mathbf{Y}$ explicitly as an element of the direct product space $R^n \times R^p$. One advantage of this construction is that we can then require derived statistics to be invariant under one group of transformations acting on R^n and to be tensors under a different group acting on R^p. For example, if $\kappa^{r,s}$ is the $p \times p$ covariance matrix of Y^1, then $\kappa^{r,s}$ is a tensor under the action of the affine group on R^p. Any estimate $k^{r,s}$, say, ought to have the same property. Furthermore, since the joint distribution of $Y^1, \ldots, Y^n$ is unaffected by permuting the n vectors, $k^{r,s}$ ought to be invariant under the action of the symmetric group (of permutations) on R^n.

Using tensor notation, the covariance matrix of $\mathbf{Y}$ may be written as $\delta^{ij}\kappa^{r,s}$, where indices i and j run from 1 to n while indices r and s run from 1 to p. Both δ^{ij} and $\kappa^{r,s}$ are tensors, but under different groups acting on different spaces. The Kronecker product, $\delta^{ij}\kappa^{r,s}$, is a tensor under the direct product group acting on $R^n \times R^p$.

More generally, if $\omega^{ij}\kappa^{r,s}$ is a tensor under the direct product of two groups acting on R^n and R^p, it may be necessary to compute the matrix inverse or generalized inverse in terms of ω_{ij} and $\kappa_{r,s}$, the generalized inverses on R^n and R^p respectively. It is immediately apparent that $\omega_{ij}\kappa_{r,s}$ is the required generalized inverse and that the inverse is covariant in the sense implied by the positions of the indices.

In matrix notation, the symbol $\otimes$ is usually employed to denote the Kronecker product. However, $\mathbf{A} \otimes \mathbf{B}$ is not the same matrix as $\mathbf{B} \otimes \mathbf{A}$ on account of the different arrangement of terms. No such difficulties arise with index notation because multiplication of real or complex numbers is a commutative operation.

Other kinds of direct products such as the Hadamard product (Rao, 1973, p. 30) do not arise in tensor calculations.

1.7.2 *Factorial design and Yates's algorithm*

Suppose that one observation is taken at each combination of the levels of factors A, B and C. Denote by Y^{ijk}, the yield or response recorded with A at level i, B at level j and C at level k. We make no claim that Y is, in any useful sense, a tensor. In fact, we occasionally write Y_{ijk} in place of Y^{ijk} where convenient to do so. This is the conventional notation for factorial models, where the indices are ordered according to the factors, and the array is not symmetric under index permutation. Further, unless the factors have equal numbers of levels, the indices have unequal ranges. This application is included here more to illustrate the value of index notation and the summation convention than as an example of a tensor.

Corresponding to each factor, we introduce *contrast matrices*, a^i_r, b^j_s, c^k_t, where the letters i,j,k refer to factor levels and the letters r,s,t refer to factor contrasts. The term 'contrast' is misused here because usually the first column of the contrast matrix is not a contrast at all, but a column of unit values. The remaining columns typically sum to zero and are therefore contrasts in the usual sense. The contrast matrices have full rank and are usually chosen to be orthogonal in the sense that

$$\begin{aligned} a^i_r a^{i'}_{r'} \delta_{ii'} &= a_{rr'} = 0 \quad \text{if } r \neq r' \\ b^j_s b^{j'}_{s'} \delta_{jj'} &= b_{ss'} = 0 \quad \text{if } s \neq s' \\ c^k_t c^{k'}_{t'} \delta_{kk'} &= c_{tt'} = 0 \quad \text{if } t \neq t' \end{aligned}$$

For example, if A has two levels, B has three ordered levels and C has four ordered levels, it is customary to make use of the orthogonal polynomial contrasts

$$a_r^i = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \qquad b_s^j = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} \qquad c_t^k = \begin{pmatrix} 1 & -3 & 1 & -1 \\ 1 & -1 & -1 & 3 \\ 1 & 1 & -1 & -3 \\ 1 & 3 & 1 & 1 \end{pmatrix}$$

Hence, the inner products give

$$a_{rr'} = \mathrm{diag}\{2, 2\}, \quad b_{ss'} = \mathrm{diag}\{3, 2, 6\}, \quad c_{tt'} = \mathrm{diag}\{4, 20, 4, 20\}.$$

By convention, contrast matrices are arranged so that rows refer to factor levels and columns to factor contrasts.

Whatever the contrasts chosen, the design matrix $\mathbf{X}$ corresponding to the factorial model $A{*}B{*}C$ is just the Kronecker product

$$x_{rst}^{ijk} = a_r^i b_s^j c_t^k .$$

In other symbols, the saturated factorial model is just

$$E(Y^{ijk}) = x_{rst}^{ijk}\beta^{rst} = a_r^i b_s^j c_t^k \beta^{rst},$$

where β^{rst} is the 'interaction' of contrast r of factor A with contrast s of factor B and contrast t of factor C. For instance, with the contrast matrices given above, β^{111} is just the mean and β^{132} is written conventionally as $B_Q C_L$. In other words, β^{132} is a measure of the change in the quadratic effect of B per unit increase in the level of C.

The so-called 'raw' or unstandardized contrasts that are produced by the three steps of Yates's algorithm, are given by

$$b_{rst} = a_r^i b_s^j c_t^k Y_{ijk}. \tag{1.16}$$

The linear combinations that are implied by the above expression are exactly those that arise when Yates's algorithm is performed in the conventional way.

To derive the least squares estimates of the parameters, we raise the indices of b, using the expression

$$\hat{\beta}^{rst} = a^{rr'} b^{ss'} c^{tt'} b_{r's't'} .$$

If the contrast matrices are each orthogonal, this expression reduces to

$$\hat{\beta}^{rst} = b_{rst}/(a_{rr} b_{ss} c_{tt}),$$

with variance $\sigma^2/(a_{rr} b_{ss} c_{tt})$, no summation intended.

The extension to an arbitrary number of factors, each having an arbitrary number of levels, is immediate. For further discussion, see Good (1958, 1960) or Takemura (1983).

From the numerical analytic point of view, the number of computations involved in (2.16) is considerably fewer than what would be required to solve n linear equations if the factorial structure of $\mathbf{X}$ were not utilized. For example, with k factors each at two levels, giving $n = 2^k$ observations, (2.16) requires nk additions and subtractions as opposed to $O(n^2)$ operations if the factorial structure were ignored. Herein lies the appeal of Yates's algorithm and also the fast Fourier transform, which uses the same device.

1.8 Bibliographic notes

Tensor notation is used widely in applied mathematics, mathematical physics and differential geometry. Definitions and notation vary to some extent with the context. For example, Jeffreys (1952) and Jeffreys & Jeffreys (1956) are concerned only with the effect on equations of motion of rotating the frame of reference or axes. Consequently, their definition of what they call a *Cartesian tensor* refers only to the orthogonal group and not, in general to arbitrary linear or non-linear transformation of coordinates. Their notation differs from that used here, most noticeably through the absence of superscripts.

Other useful references, again with a bias towards applications to physics, include McConnell (1931) and Lawden (1968). Thomas (1965, Section 6) emphasises the importance of transformation groups in the definition of a tensor.

For more recent work, again connected mainly with mathematical physics, see Richtmyer (1981).

In the theory of differential geometry, which is concerned with describing the local behaviour of curves and surfaces in space, notions of curvature and torsion are required that are independent of the choice of coordinate system on the surface. This requirement leads naturally to the notion of a tensor under the group of arbitrary invertible parameterizations of the surface. Gaussian and Riemannian curvature as well as mean curvature are invariants derived from such tensors. This work has a long history going back to Levi-Cività, Ricci, Riemann and Gauss's celebrated *theorem egregium*. Details can be found in the books by Eisenhart (1926), Weatherburn (1950), Sokolnikoff (1951) and Stoker (1969). For a more recent treatment of Riemannian geometry, see Willmore (1982) or Spivak (1970).

For a discussion of the geometry of generalized inverses, see Kruskal (1975) and the references therein.

The notion of a *spinor* is connected with rotations in Euclidean space and has applications in the theory of special relativity. Inevitably, there are strong similarities with quaternions, which are also useful for studying rotations. Tensors arise naturally in the study of such objects. See, for example, Cartan (1981).

Despite the widespread use in statistics of multiply-indexed arrays, for example, in the study of factorial, fractional factorial and other designs, explicit use of tensor methods is rare, at least up until the past few years. For an exception, see Takemura (1983). The reasons for this neglect are unclear: matrix notation abounds and is extraordinarily convenient provided that we do not venture far beyond linear models and second-moment assumptions. In any case, the defects and shortcomings of matrix notation become clearly apparent as soon as we depart from linear models or need to study moments beyond the second. For example, many of the quantities that arise in later chapters of this book cannot be expressed using matrix notation. This is the realm of the tensor, and it is our aim in the remainder of this book to demonstrate to the reader that great simplification can result from judicious choice of notation.

1.9 Further results and exercises 1

1.1 Derive the transformation laws (2.7) for log likelihood derivatives.

1.2 Show that if ω^{ijk} is a contravariant tensor and ω_{ijk} is a covariant tensor, then $\omega^{ijk}\omega_{ijk}$ is an invariant.

1.3 Show directly, using the notation in (2.7), that $U_r U_s I^{rs}$ is invariant under the group of invertible transformations acting on the parameter space.

1.4 Let $i_{rs} = -E\{U_{rs};\theta\}$ and let i^{rs} be the matrix inverse. Under which group of transformations is $U_r U_s i^{rs}$ an invariant?

1.5 Show, using the notation in (2.7), that

$$V_{ij} = U_{ij} - \kappa_{ij,k} i^{kl} U_l$$

is a covariant tensor, where $\kappa_{rs,t}$ is the covariance of U_{rs} and U_t.

1.6 Let a^i_j be the elements of a square matrix, not necessarily symmetrical, and let its inverse, b^j_i satisfy $a^i_j b^j_k = \delta^i_k = a^j_k b^i_j$. Show that the derivatives satisfy

$$\frac{\partial b^j_i}{\partial a^r_s} = -b^j_r b^s_i$$

$$\frac{\partial a^j_i}{\partial b^r_s} = -a^j_r a^s_i$$

1.7 Show that the spectral decomposition of the the symmetric matrix $\mathbf{A}$

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T, \quad \mathbf{Q}\mathbf{Q}^T = \mathbf{I}, \quad \Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$$

is unique up to permutations of the columns of $\mathbf{Q}$ and the elements of Λ if the eigenvalues of $\mathbf{A}$ are distinct.

1.8 Show that there exists a linear transformation $Y^r = a^r_i X^i$ from X to Y such that the quadratic form $\omega_{ij}X^iX^j$ may be written as $\epsilon_{rs}Y^rY^s$ where $\epsilon_{ii} = \pm 1$ and $\epsilon_{ij} = 0$ if $i \neq j$.

1.9 Consider the decomposition $\omega_{ij} = \tau^r_i \tau^s_j \epsilon_{rs}$ of the symmetric covariant tensor ω_{ij}, where the notation is that used in (2.14). Define

$$\omega^+_{ij} = \tau^r_i \tau^s_j |\epsilon_{rs}|.$$

Show that ω^+_{ij} is a covariant tensor and that the scalar

$$s = \omega^{ij}\omega^+_{ij}$$

is independent of the choice of τ^r_i and also of the choice of generalized inverse ω^{ij}. Show that s is the signature of ω_{ij}.

1.10 In the notation of the previous exercise, let

$$\omega^{ij} = \gamma^i_r \gamma^j_s \epsilon^{rs},$$

where $\epsilon^{rs} = \epsilon_{rs}$ and $\gamma^i_r \tau^r_j = \delta^i_j$. Show that ω^{ij} is the Moore-Penrose inverse of ω_{ij}.

1.11 Show that the identity matrix is a generalized inverse of any projection matrix. Show also that a projection matrix, not necessarily symmetric, is its own generalized inverse satisfying (2.9) and (2.11). Under what conditions is a projection matrix self-inverse in the Moore-Penrose sense?

1.12 Let ω^{ij}, with inverse ω_{ij}, be the components of a $p \times p$ symmetric matrix of rank p. Show that

$$\gamma^{rs,ij} = \omega^{ri}\omega^{sj} + \omega^{rj}\omega^{si},$$

regarded as a $p^2 \times p^2$ matrix with rows indexed by (r, s) and columns by (i, j), is symmetric with rank $p(p+1)/2$. Show also that $\omega_{ri}\omega_{sj}/2$ is a generalized inverse. Find the Moore-Penrose generalized inverse.

1.13 Consider the linear mapping from R^p to itself given by

$$\bar{X}^r = \omega^r_i X^i$$

where ω^r_i is nonsingular. Show that, under simultaneous change of coordinates

$$Y^r = a^r_i X^i, \quad \bar{Y}^r = a^r_i \bar{X}^i,$$

ω^r_i transforms as a mixed tensor. By comparing the volume of a set, B say, in the X coordinate system with the volume of the transformed set, $\bar{B}$, interpret the determinant of ω^r_i as an invariant. Give similar interpretations of the remaining $p-1$ invariants, e.g. in terms of surface area and so on.

1.14 Let $\pi_1, \ldots, \pi_k$ be positive numbers adding to unity and define the multinomial covariance matrix

$$\omega_{ij} = \begin{cases} \pi_i(1-\pi_i) & i = j \\ -\pi_i\pi_j & i \neq j \end{cases}.$$

Show that ω_{ij} has rank $k-1$ and that

$$\omega^{ij} = \begin{cases} 1/\pi_i & i = j \\ 0 & \text{otherwise} \end{cases}$$

is a generalized inverse of rank k. Find the Moore-Penrose generalized inverse.

1.15 Let the $p\times q$ matrix $\mathbf{A}$, with components a^r_i, be considered as defining a linear transformation from the domain, R^q, to the range in R^p. Interpret the *singular values* of $\mathbf{A}$ as invariants under independent orthogonal transformation of the domain and range spaces. For the definition of singular values and their application in numerical linear algebra, see Chambers (1977, Section 5.e).

1.16 Let $\mathbf{A}$, $\mathbf{A}^{-1}$ and $\mathbf{X}$ be symmetric matrices with components a_{ij}, a^{ij} and x_{ij} respectively. Show that the Taylor expansion for the log determinant of $\mathbf{A}+\mathbf{X}$ about the origin may be written

$$\begin{aligned}\log\det(\mathbf{A}+\mathbf{X}) = \log\det(\mathbf{A}) + x_{ij}a^{ij} - x_{ij}x_{kl}a^{ik}a^{jl}/2 \\ + x_{ij}x_{kl}x_{mn}a^{ik}a^{jm}a^{ln}/3 + \cdots\end{aligned}$$

Describe the form of the general term in this expansion. Compare with Exercise 1.6 and generalize to asymmetric matrices.

1.17 Justify the claim that Kronecker's delta, δ^j_i, is a tensor.

1.18 Show that δ_{ij} is a tensor under the orthogonal group but not under any larger group.

1.19 Show that δ_{ijk}, $\delta_{ijkl}, \ldots$ are tensors under the *symmetric* group but not under any larger group.
[The symmetric group, which is most conveniently represented by the set of permutation matrices, arises naturally in the study of sample moments based on simple random samples, where the order in which the observations are recorded is assumed to be irrelevant. See Chapter 4.]

1.20 Show that the symmetric group is a subgroup of the orthogonal group, which, in turn, is a subgroup of the general linear group.

1.21 *Hooke's Law*: In the mechanics of deformable solids, the components of the *stress tensor*, p_{ij}, measure force per unit area in the following sense. Let $\mathbf{e}_i$ be the unit vector in the ith coordinate direction and let $\bar{\mathbf{e}}_i$ be the orthogonal plane. Then p_{ii} is the force per unit area normal to the plane, also called the *normal stress*, and p_{ij}, $j \neq i$ are the *shear stresses* acting in the plane. The components of the *strain tensor*, q_{ij}, which are dimensionless, measure percentage deformation or percentage change in length. Both arrays are symmetric tensors under the orthogonal group.

In the case of *elastic* deformation of an *isotropic* material, Hooke's law in its most general form states that the relationship between stress and strain is linear. Thus,

$$p_{rs} = b_{rs}^{ij} q_{ij},$$

where b_{rs}^{ij} is an *isotropic* fourth-order tensor given by

$$b_{rs}^{ij} = \lambda \delta^{ij} \delta_{rs} + 2\mu \delta_r^i \delta_s^j,$$

for constants λ, μ that are characteristic of the material.

Show that the inverse relationship giving the strains in terms of the stresses may be written in the form

$$q_{ij} = \left(\lambda' \delta^{rs} \delta_{ij} + 2\mu' \delta_i^r \delta_j^s\right) p_{rs},$$

where the new constants are given by

$$\mu' = \frac{1}{4\mu}, \qquad \lambda' + 2\mu' = \frac{\lambda + \mu}{\mu(3\lambda + 2\mu)} = E^{-1}.$$

In the terminology used in Mechanics, E is known as *Young's modulus* or *modulus of elasticity*, μ is called the *rigidity* or *shear modulus* and $\sigma = \lambda/\{2(\lambda + \mu)\}$ is *Poisson's ratio.* Note that $E = 2(1 + \sigma)\mu$, implying that two independent constants entirely determine the three-dimensional elastic properties of the material. (Murnaghan, 1951, Chapters 3,4; Jeffreys & Jeffreys 1956, Section 3.10; Drucker, 1967, Chapter 12).

1.22 The array ϵ_{ijk} of order $3 \times 3 \times 3$ defined by

$$\begin{aligned} \epsilon_{123} &= \epsilon_{231} = \epsilon_{312} = 1 \\ \epsilon_{213} &= \epsilon_{132} = \epsilon_{321} = -1 \\ \epsilon_{ijk} &= 0 \quad \text{otherwise,} \end{aligned}$$

is known as the *alternating tensor* (Ames & Murnaghan, 1929, p. 440). For any 3×3 matrix a_r^i, show that

$$\epsilon_{ijk} a_r^i a_s^j a_t^k = \epsilon_{rst} \det(\mathbf{A}).$$

Hence show that ϵ_{ijk} is an isotropic tensor under O^+, the orthogonal group with positive determinant (Jeffreys & Jeffreys, 1956, Sections 2.07, 3.03).

Write down the generalization of the alternating tensor appropriate for a $p \times p \times p$ array.

CHAPTER 2

Elementary theory of cumulants

2.1 Introduction

This chapter deals with the elementary theory of cumulants in the multivariate case as well as the univariate case. Little prior knowledge is assumed other than some familiarity with moments and the notion of mathematical expectation, at least in the univariate case. In what follows, all integrals and infinite sums are assumed to be convergent unless otherwise stated. If the random variable X has a density function $f_X(x)$ defined over $-\infty < x < \infty$, then the expectation of the function $g(X)$ is

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \tag{2.1}$$

If the distribution of X is discrete, the integral is replaced by a sum over the discrete values. More generally, to take care of both the discrete case and the continuous case simultaneously, we may replace $f_X(x)dx$ in the above integral by $dF_X(x)$, where F_X is a probability measure. Such differences, however, need not concern us here. All that is required is a knowledge of some elementary properties of the integral and the expectation operator. In the particular case where $g(X)$ is the rth power of X, (2.1) is called the rth *moment* of X.

In the multivariate case, essentially the same definitions may be used except that the integral over R^1 is replaced by an integral over R^p. Not only do we have to consider the moments of each component of X but also the cross moments such as $E(X^1 X^2)$ and $E(X^1 X^1 X^2)$. As always, in the multivariate case, superscripts denote components and not powers. The moments of a given component of X give information regarding the marginal distribution of that component. The cross moments are required to give information concerning the joint distribution of the components.

Cumulants are normally introduced as functions of the moments. It is entirely natural to inquire at the outset why it is preferable to work with cumulants rather than moments since the two are entirely equivalent. A single totally convincing answer to this query is difficult to find and, in a sense, Chapters 2 to 6 provide several answers. Simplicity seems to be the main criterion as the following brief list shows.

(i) Most statistical calculations using cumulants are simpler than the corresponding calculations using moments.

(ii) For independent random variables, the cumulants of a sum are the sums of the cumulants.

(iii) For independent random variables, the cross cumulants or mixed cumulants are zero.

(iv) Edgeworth series used for approximations to distributions are most conveniently expressed using cumulants.

(v) Where approximate normality is involved, higher-order cumulants can usually be neglected but not higher-order moments.

2.2 Generating functions

2.2.1 *Definitions*

As always, we begin with the random variable X whose components are $X^1, \ldots, X^p$. All arrays bearing superscripts refer to these components. Unless otherwise specified, the moments of X about the origin are assumed finite and are denoted by

$$\begin{aligned} \kappa^i &= E(X^i), \quad \kappa^{ij} = E(X^i X^j), \\ \kappa^{ijk} &= E(X^i X^j X^k) \end{aligned} \tag{2.2}$$

and so on. Moments about the mean, also called central moments, are rarely considered explicitly here, except as a special case of the above with $\kappa^i = 0$. The indices need not take on distinct values and there may, of course, be more than p indices, implying repetitions. Thus, for example, κ^{11} is the mean square of X^1 about the origin and κ^{222} is the mean cube of X^2, the second component of X. In this context, superscripts must not be confused with powers and powers should, where possible, be avoided. In this book, powers are avoided for the most part, the principal exceptions arising in Sections 2.5 and 2.6 where the connection with other notations is described and interpretations are given.

Consider now the infinite series

$$\begin{aligned} M_X(\xi) = 1 + \xi_i\kappa^i + \xi_i\xi_j\kappa^{ij}/2! + \xi_i\xi_j\xi_k\kappa^{ijk}/3! \\ + \xi_i\xi_j\xi_k\xi_l\kappa^{ijkl}/4! + \ldots, \end{aligned} \tag{2.3}$$

which we assume to be convergent for all $|\xi|$ sufficiently small. The sum may be written in the form

$$M_X(\xi) = E\{\exp(\xi_i X^i)\},$$

and the moments are just the partial derivatives of $M_X(\xi)$ evaluated at $\xi = 0$.

The cumulants are most easily defined via their generating function,

$$K_X(\xi) = \log M_X(\xi),$$

which has an expansion

$$\begin{aligned} K_X(\xi) = \xi_i\kappa^i + \xi_i\xi_j\kappa^{i,j}/2! + \xi_i\xi_j\xi_k\kappa^{i,j,k}/3! \\ + \xi_i\xi_j\xi_k\xi_l\kappa^{i,j,k,l}/4! + \ldots. \end{aligned} \tag{2.4}$$

This expansion implicitly defines all the cumulants, here denoted by κ^i, $\kappa^{i,j}$, $\kappa^{i,j,k}$ and so on, in terms of the corresponding moments. The major departure from standard statistical notation is that we have chosen to use the same letter for both moments and cumulants. Both are indexed by a set rather than by a vector, the only distinction arising from the commas, which are considered as separators for the cumulant indices. Thus the cumulants are indexed by a set of indices fully partitioned and the moments by the same set unpartitioned. One curious aspect of this convention is that the notation does not distinguish between moments with one index and cumulants with one index. This is convenient because the first moments and first cumulants are identical.

The infinite series expansion (2.3) for $M_X(\xi)$ may be divergent for all real $|\xi| > 0$ either because some of the higher-order moments are infinite or because the moments, though finite, increase sufficiently rapidly to force divergence (Exercise 2.2). In such cases, it is entirely legitimate to work with the finite series expansions up to any specified number of terms. This device can be justified by taking ξ to be purely imaginary, in which case, the integral $E\{\exp(\xi_i X^i)\}$ is convergent and its Taylor approximation for small imaginary ξ is just the truncated expansion. Similarly for

$K_X(\xi)$. Thus cumulants of any order are well defined if the corresponding moment and all lower-order marginal moments are finite: see (2.9) and Exercise 2.1.

One difficulty that arises in using moment or cumulant calculations to prove results of a general nature is that the infinite set of moments is, in general, not sufficient to determine the joint distribution uniquely. Feller (1971, Section VII.3) gives a pair of non-identical univariate density functions having identical moments of all orders. Non-uniqueness occurs only when the function $M_X(\xi)$ is not analytic at the origin. Thus, for a large class of problems, non-uniqueness can be avoided by including the condition that the series expansion for $M_X(\xi)$ be convergent for $|\xi| < \delta$ where $\delta > 0$ (Moran, 1968, Section 6.4; Billingsley, 1985, Exercise 30.5). In the univariate case, other conditions limiting the rate of increase of the even moments are given by Feller (1971, Sections VII.3, VII.6 and XV.4).

2.2.2 *Some examples*

Multivariate normal distribution: The multivariate normal density with mean vector κ^i and covariance matrix $\kappa^{i,j}$ may be written in the form

$$(2\pi)^{-p/2}|\kappa^{i,j}|^{-1/2}\exp\{-\tfrac{1}{2}(x^i-\kappa^i)(x^j-\kappa^j)\kappa_{i,j}\},$$

where $\kappa_{i,j}$ is the matrix inverse of $\kappa^{i,j}$ and x ranges over R^p. The moment generating function may be found by completing the square in the exponent and using the fact that the density integrates to 1. This gives

$$M_X(\xi) = \exp\{\xi_i\kappa^i + \tfrac{1}{2}\xi_i\xi_j\kappa^{i,j}\}$$

and

$$K_X(\xi) = \xi_i\kappa^i + \tfrac{1}{2}\xi_i\xi_j\kappa^{i,j}.$$

In other words, for the normal distribution, all cumulants of order three or more are zero and, as we might expect, the second cumulant is just the covariance array.

Multinomial distribution: For a second multivariate example, we take the multinomial distribution on k categories with index m and parameter $\pi = \pi_1, \ldots, \pi_k$. The joint distribution or probability function may be written

$$\mathrm{pr}(X^1 = x^1, \ldots, X^k = x^k) = \binom{m}{x^1, \ldots, x^k}\prod_{j=1}^{k}\pi_j^{x^j}$$

where $0 \le x^j \le m$ and $\sum_j x^j = m$. The moment generating function may be found directly using the multinomial theorem, giving

$$M_X(\xi) = \sum_x \binom{m}{x^1, \ldots, x^k}\prod_{j=1}^{k}\exp(\xi_j x^j)\pi_j^{x^j} = \left(\sum \pi_j \exp(\xi_j)\right)^m.$$

The cumulant generating function is

$$K_X(\xi) = m\log\left(\sum \pi_j\exp(\xi_j)\right).$$

Thus all cumulants are finite and have the form $m \times$ (function of π). The first four are given in Exercise 2.16.

Student's distribution: Our third example involves a univariate distribution whose moments of order three and higher are infinite. The t distribution on three degrees of freedom has density function

$$f_X(x) = \frac{2}{\pi\sqrt{3}(1+x^2/3)^2} \qquad -\infty < x < \infty.$$

The moment generating function, $M_X(\xi) = E\{\exp(\xi X)\}$, diverges for all real $|\xi| > 0$ so that the function $M_X(\xi)$ is not defined or does not exist for real ξ. However, if we write $\xi = i\zeta$ where ζ is real, we find

$$M_X(i\zeta) = \int_{-\infty}^{\infty} \frac{2\exp(i\zeta x)}{\pi\sqrt{3}(1 + x^2/3)^2}\, dx.$$

The integrand has poles of order 2 at $x = \pm i\sqrt{3}$ but is analytic elsewhere in the complex plane. If $\zeta > 0$, the integral may be evaluated by deforming the contour into the positive complex half-plane, leaving the residue at $x = +i\sqrt{3}$. If $\zeta < 0$, it is necessary to deform in the other direction, leaving the residue at $x = -i\sqrt{3}$. This procedure gives

$$\begin{aligned}
M_X(i\zeta) &= \exp(-\sqrt{3}|\zeta|)\{1 + \sqrt{3}|\zeta|\} \\
K_X(i\zeta) &= -\sqrt{3}|\zeta| + \log(1 + \sqrt{3}|\zeta|) \\
&= \begin{cases} -3\zeta^2/2 + \sqrt{3}\zeta^3 - \ldots & \text{if } \zeta > 0 \\ -3\zeta^2/2 - \sqrt{3}\zeta^3 - \ldots & \text{if } \zeta < 0 \end{cases}
\end{aligned}$$

Thus $K_X(i\zeta)$ has a unique Taylor expansion only as far as the quadratic term. It follows that $E(X) = 0$, $\operatorname{var}(X) = 3$ and that the higher-order cumulants are not defined.

2.3 Cumulants and moments

To establish the relationships connecting moments with cumulants, we write $M_X(\xi) = \exp\{K_X(\xi)\}$ and expand to find

$$\begin{aligned}
1 + \xi_i\kappa^i &+ \xi_i\xi_j(\kappa^{i,j}/2! + \kappa^i\kappa^j/2!) \\
&+ \xi_i\xi_j\xi_k\kappa^{i,j,k}/3! \quad + \xi_i\xi_j\xi_k\xi_l\kappa^{i,j,k,l}/4! + \ldots \\
&+ \xi_i\xi_j\xi_k\kappa^i\kappa^{j,k}/2! \; + \xi_i\xi_j\xi_k\xi_l\{\kappa^i\kappa^{j,k,l}/6 + \kappa^{i,j}\kappa^{k,l}/8\} + \ldots \\
&+ \xi_i\xi_j\xi_k\kappa^i\kappa^j\kappa^k/3! + \xi_i\xi_j\xi_k\xi_l\kappa^i\kappa^j\kappa^{k,l}/4 + \ldots \\
&\qquad\qquad\qquad\qquad + \xi_i\xi_j\xi_k\xi_l\kappa^i\kappa^j\kappa^k\kappa^l/4! + \ldots \\
&\qquad\qquad\qquad\qquad\qquad\qquad + \ldots\,.
\end{aligned} \tag{2.5}$$

After combining terms and using symmetry, we find the following expressions for moments in terms of cumulants:

$$\begin{aligned}
\kappa^{ij} &= \kappa^{i,j} + \kappa^i\kappa^j \\
\kappa^{ijk} &= \kappa^{i,j,k} + (\kappa^i\kappa^{j,k} + \kappa^j\kappa^{i,k} + \kappa^k\kappa^{i,j}) + \kappa^i\kappa^j\kappa^k \\
&= \kappa^{i,j,k} + \kappa^i\kappa^{j,k}[3] + \kappa^i\kappa^j\kappa^k \\
\kappa^{ijkl} &= \kappa^{i,j,k,l} + \kappa^i\kappa^{j,k,l}[4] + \kappa^{i,j}\kappa^{k,l}[3] + \kappa^i\kappa^j\kappa^{k,l}[6] \\
&\quad + \kappa^i\kappa^j\kappa^k\kappa^l,
\end{aligned} \tag{2.6}$$

where, for example,

$$\kappa^{i,j}\kappa^{k,l}[3] = \kappa^{i,j}\kappa^{k,l} + \kappa^{i,k}\kappa^{j,l} + \kappa^{i,l}\kappa^{j,k}$$

is the sum over the three 2^2 partitions of four indices. The bracket notation is simply a convenience to avoid listing explicitly all 15 partitions of four indices in the last equation (2.6). Only the five distinct types, each corresponding to a partition of the *number* 4, together with the number of partitions of each type, need be listed. The following is a complete list of the 15 partitions of four items, one column for each of the five types.

$ijkl$	$i\|jkl$	$ij\|kl$	$i\|j\|kl$	$i\|j\|k\|l$
	$j\|ikl$	$ik\|jl$	$i\|k\|jl$	
	$k\|ijl$	$il\|jk$	$i\|l\|jk$	
	$l\|ijk$		$j\|k\|il$	
			$j\|l\|ik$	
			$k\|l\|ij$	

In the case of fifth-order cumulants, there are 52 partitions of seven different types and our notation makes the listing of such partitions feasible for sets containing not more than eight or nine items. Such lists are given in Tables 1 and 2 of the Appendix, and these may be used to find the expressions for moments in terms of cumulants. We find, for example, from the partitions of five items, that

$$\begin{aligned}\kappa^{ijklm} &= \kappa^{i,j,k,l,m} + \kappa^i\kappa^{j,k,l,m}[5] + \kappa^{i,j}\kappa^{k,l,m}[10]\\ &\quad + \kappa^i\kappa^j\kappa^{k,l,m}[10] + \kappa^i\kappa^{j,k}\kappa^{l,m}[15]\\ &\quad + \kappa^i\kappa^j\kappa^k\kappa^{l,m}[10] + \kappa^i\kappa^j\kappa^k\kappa^l\kappa^m.\end{aligned}$$

If $\kappa^i = 0$, all partitions having a unit part (a block containing only one element) can be ignored. The formulae then simplify to

$$\begin{aligned}\kappa^{ij} &= \kappa^{i,j}, \quad \kappa^{ijk} = \kappa^{i,j,k}\\ \kappa^{ijkl} &= \kappa^{i,j,k,l} + \kappa^{i,j}\kappa^{k,l}[3]\\ \kappa^{ijklm} &= \kappa^{i,j,k,l,m} + \kappa^{i,j}\kappa^{k,l,m}[10]\\ \kappa^{ijklmn} &= \kappa^{i,j,k,l,m,n} + \kappa^{i,j}\kappa^{k,l,m,n}[15] + \kappa^{i,j,k}\kappa^{l,m,n}[10]\\ &\quad + \kappa^{i,j}\kappa^{k,l}\kappa^{m,n}[15].\end{aligned}$$

These are the formulae for the central moments in terms of cumulants.

The reverse formulae giving cumulants in terms of moments may be found either by formal inversion of (2.6) or by expansion of $\log M_X(\xi)$ and combining terms. The expressions obtained for the first four cumulants are

$$\begin{aligned}\kappa^{i,j} &= \kappa^{ij} - \kappa^i\kappa^j\\ \kappa^{i,j,k} &= \kappa^{ijk} - \kappa^i\kappa^{jk}[3] + 2\kappa^i\kappa^j\kappa^k\\ \kappa^{i,j,k,l} &= \kappa^{ijkl} - \kappa^i\kappa^{jkl}[4] - \kappa^{ij}\kappa^{kl}[3] + 2\kappa^i\kappa^j\kappa^{kl}[6] - 6\kappa^i\kappa^j\kappa^k\kappa^l.\end{aligned} \tag{2.7}$$

Again, the sum is over all partitions of the indices but this time, the coefficient $(-1)^{\nu-1}(\nu-1)!$ appears, where ν is the number of blocks of the partition. The higher-order formulae follow the same pattern and the list of partitions in Tables 1,2 of the Appendix may be used for cumulants up to order eight.

More generally, the relationships between moments and cumulants may be written as follows. Let $\Upsilon = \{\upsilon_1, \ldots, \upsilon_\nu\}$ be a partition of a set of p indices into ν non-empty blocks. (Υ and υ are the upper and lower cases of the Greek letter 'upsilon'). For example, if $p = 4$ and $\nu = 2$ we might have $\Upsilon = \{(i,j),(k,l)\}$, $\{(i,k),(j,l)\}$ or $\{(i),(j,k,l)\}$ or any one of the four other possible partitions into 2 blocks. The partition comprising just a single block is denoted by Υ_1 and the partition comprising p unit blocks is denoted by Υ_p. There is no standard statistical notation for these partitions and it is sometimes convenient to use the alternatives, 1 and 0, borrowed from lattice theory: see Section 3.6. The moment involving all p indices is written $\kappa(\Upsilon_1)$ and the corresponding cumulant $\kappa(\Upsilon_p)$. In the above example, $\Upsilon_1 = \{(i,j,k,l)\}$ is the partition into one block, $\kappa(\Upsilon_1) = \kappa^{ijkl}$ is the corresponding moment, $\Upsilon_4 = \{(i),(j),(k),(l)\}$ is the partition into 4 blocks and $\kappa(\Upsilon_4) = \kappa^{i,j,k,l}$ is the corresponding cumulant. Cumulants involving only those indices in block υ_j are written $\kappa(\upsilon_j)$ and the corresponding moment is written $\mu(\upsilon_j)$.

By examining the general term in expansion (2.5) for $M_X(\xi)$, it is not difficult to see that the expression for the moment $\kappa(\Upsilon_1)$ in terms of cumulants may be written

$$\kappa(\Upsilon_1) = \sum_{\Upsilon} \kappa(\upsilon_1)\ldots\kappa(\upsilon_\nu) \tag{2.8}$$

where the sum extends over all partitions of the indices. Equivalently, the above may be expressed as a double sum, first over ν and then over all partitions of the indices into ν blocks.

The corresponding expression for the cumulant $\kappa(\Upsilon_p)$ in terms of the moments is

$$\kappa(\Upsilon_p) = \sum_{\Upsilon}(-1)^{\nu-1}(\nu-1)!\mu(\upsilon_1)\ldots\mu(\upsilon_\nu). \tag{2.9}$$

Note that the block sizes do not enter into either of the above expressions.

In fact, we may take (2.9) as an alternative definition of cumulant, more directly applicable than the definition relying on generating functions. The advantage of (2.9) as a definition is that it makes explicit the claim made at the end of Section 2.2, that cumulants of order r are well defined when the corresponding rth-order moment and all lower-order marginal moments are finite. Note that, apart from the univariate case, $\kappa^{ijk} < \infty$ does not imply that $\kappa^{i,j,k} < \infty$: see Exercise 2.1.

2.4 Linear and affine transformations

The objective in this section is to examine how the moment arrays κ^{ij}, $\kappa^{ijk}, \ldots$ and the cumulant arrays $\kappa^{i,j}$, $\kappa^{i,j,k}, \ldots$ change when we make a simple transformation from the original variables $X^1, \ldots, X^p$ to new variables $Y = Y^1, \ldots, Y^q$. If Y is a linear function of X, we may write

$$Y^r = a_i^r X^i,$$

where a_i^r is an array of constants. It is not difficult to see that the moments of Y are

$$a_i^r\kappa^i, \quad a_i^r a_j^s \kappa^{ij}, \quad a_i^r a_j^s a_k^t \kappa^{ijk}, \ldots$$

while the cumulants are

$$a_i^r\kappa^i, \quad a_i^r a_j^s \kappa^{i,j}, \quad a_i^r a_j^s a_k^t \kappa^{i,j,k}, \ldots. \tag{2.10}$$

In other words, under linear transformation, both moments and cumulants transform like contravariant tensors. Note however, that the matrix a_i^r need not have full rank.

Affine transformations involve a change of origin according to the equation

$$Y^r = a^r + a_i^r X^i.$$

The cumulants of Y, derived at the end of this section, are

$$a^r + a_i^r\kappa^i, \quad a_i^r a_j^s \kappa^{i,j}, \quad a_i^r a_j^s a_k^t \kappa^{i,j,k}, \quad a_i^r a_j^s a_k^t a_l^u \kappa^{i,j,k,l} \tag{2.11}$$

and so on. The change of origin affects only the mean vector or first cumulant. For this reason, cumulants are sometimes called semi-invariants. On the other hand, the moments of Y are

$$\begin{aligned}
&a^r + a_i^r\kappa^i, \\
&a^r a^s + a^r a_i^s \kappa^i[2] + a_i^r a_j^s \kappa^{ij}, \\
&a^r a^s a^t + a^r a^s a_i^t \kappa^i[3] + a^r a_i^s a_j^t \kappa^{ij}[3] + a_i^r a_j^s a_k^t \kappa^{ijk}
\end{aligned}$$

and so on, where $a^r a_i^s \kappa^i[2] = a^r a_i^s \kappa^i + a^s a_i^r \kappa^i$. Thus, unlike the cumulants, the moments do not transform in a pleasant way under affine transformation of coordinates.

The transformation law for cumulants is similar to the transformation law of Cartesian tensors, (Jeffreys, 1952), the only difference being the dependence of the first cumulant on the choice of origin. To avoid any ambiguity of terminology, we use the term *cumulant tensor* to describe any array of quantities that transforms according to (2.11) under affine transformation of X. In Section

3.3, we develop rules for the transformation of cumulant tensors under non-linear transformation of X. These rules are quite different from tensor transformation laws that arise in differential geometry or in theoretical physics.

To prove (2.11) we use the method of generating functions, giving

$$\begin{aligned} M_Y(\xi) &= E[\exp\{\xi_r(a^r + a_i^r X^i)\}] \\ &= \exp(\xi_r a^r) M_X(\xi_r a_i^r). \end{aligned}$$

In other words,

$$K_Y(\xi) = \xi_r a^r + K_X(\xi_r a_i^r)$$

from which expressions (2.10) and (2.11) follow directly.

2.5 Univariate cumulants and power notation

Much of the literature on cumulants concentrates on the univariate case, $p = 1$ and uses the condensed *power* notation κ_r for the rth cumulant of X^1, written in this section as X without indices. In this and in the following section, we move quite freely from power notation to index notation and back: this should cause no confusion on account of the different positions of indices in the two notations. Following Kendall & Stuart (1977, Chapter 3) we write

$$\mu_r' = E(X^r) \quad \text{and} \quad \mu_r = E(X - \mu_1')^r$$

where the superscript here denotes a power. Expressions (2.6) giving moments in terms of cumulants become

$$\begin{aligned} \mu_2' &= \kappa_2 + \kappa_1^2 \\ \mu_3' &= \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3 \\ \mu_4' &= \kappa_4 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + 6\kappa_1^2\kappa_2 + \kappa_1^4 \end{aligned}$$

where superscripts again denote powers. The reverse formulae (2.7) become

$$\begin{aligned} \kappa_2 &= \mu_2' - (\mu_1')^2 \\ \kappa_3 &= \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 \\ \kappa_4 &= \mu_4' - 4\mu_1'\mu_3' - 3(\mu_2')^2 + 12(\mu_1')^2\mu_2' - 6(\mu_1')^4. \end{aligned}$$

In this notation, the permutation factors, previously kept in [.], and the arithmetic factors $(-1)^{\nu-1}(\nu - 1)!$ become combined, with the result that the essential simplicity of the formulae (2.7) disappears. For this reason, unless we are dealing with a single random variable or a set of independent and identically distributed random variables, it is usually best to use index notation, possibly reverting to power notation at the last step of the calculations.

An extended version of power notation is sometimes used for bivariate cumulants corresponding to the random variables X and Y. For example, if $\mu_{rs}' = E(X^rY^s)$, the corresponding cumulant may be written κ_{rs}. Note that in this notation, $\kappa_{rs} \neq \kappa_{sr}$. In other words, with power notation, the cumulants are indexed by a vector whereas, with index notation, the cumulants are indexed by a set or, more generally, by the partitions of a set.

To establish the relationships between bivariate moments and bivariate cumulants in this notation, we simply convert the terms in (2.7) into power notation. For $r = 2$, $s = 1$ this gives

$$\kappa_{21} = \mu_{21}' - 2\mu_{10}'\mu_{11}' - \mu_{01}'\mu_{20}' + 2(\mu_{10}')^2\mu_{01}'$$

and, for $r = 2$, $s = 2$,

$$\begin{aligned} \kappa_{22} = \mu_{22}' &- 2\mu_{10}'\mu_{12}' - 2\mu_{01}'\mu_{21}' - \mu_{20}'\mu_{02}' - 2(\mu_{11}')^2 \\ &+ 2(\mu_{10}')^2\mu_{02}' + 2(\mu_{01}')^2\mu_{20}' + 8\mu_{01}'\mu_{10}'\mu_{11}' - 6(\mu_{10}')^2(\mu_{01}')^2. \end{aligned}$$

Additional, more impressive formulae of this type may be found in David, Kendall & Barton (1966, Tables 2.1.1 and 2.1.2). The formula for κ_{44}, for example, occupies 24 lines.

Simpler formulae are available in terms of central moments, but it is clear that the above notation conceals the simplicity of the formulae. For example, if μ'_{01} and μ'_{10} are both equal to zero, the formulae for κ_{44}, κ_{35}, κ_{26} and so on, can be found from the list of partitions of $1, 2, \ldots, 8$ in Table 2 of the Appendix. For this reason, power notation is best avoided at least in formal manipulations. It is, however, very useful and convenient in the univariate case and is also useful more generally for interpretation.

2.6 Interpretation of cumulants

Although the definition of cumulants given in Sections 2.2 and 2.3 covered both univariate and mixed cumulants, their interpretations are best considered separately. Roughly speaking, mixed cumulants have an interpretation in terms of dependence or independence: univariate cumulants have a simple interpretation in terms of the shape of the marginal distribution. Of course, mixed cumulants could be interpreted also in terms of the shape of the joint distribution but such interpretations are not given here. We deal first with univariate cumulants using the notation of Section 2.5 and work our way up to mixed cumulants involving four distinct variables.

The first cumulant of X, denoted by κ_1, is the mean value and the second cumulant, κ_2, is the variance. In rigid body mechanics, κ_1 is the x-coordinate of the centre of gravity and κ_2 is the moment of inertia about the axis $x = \kappa_1$ of a uniform laminar body of unit mass in the (x, y) plane, bounded by $0 \le y \le f(x)$ where $f(x)$ is the density of X.

The third cumulant of X is a measure of asymmetry in the sense that $\kappa_3 = E(X - \kappa_1)^3$ is zero if X is symmetrically distributed. Of course $\kappa_3 = 0$ does not, on its own, imply symmetry: to guarantee symmetry, we require all odd cumulants to vanish and the distribution to be determined by its moments. For an example of an asymmetrical distribution whose odd cumulants are zero, see Kendall & Stuart (1977, Exercise 3.26), which is based on the note by Churchill (1946). The usual measure of skewness is the standardized third cumulant, $\rho_3 = \kappa_3/\kappa_2^{3/2}$, which is unaffected by affine transformations $X \to a + bX$ with $b > 0$. If $b < 0$, $\rho_3 \to -\rho_3$. Third and higher-order standardized cumulants given by $\rho_r = \kappa_r/\kappa_2^{r/2}$, can be interpreted as summary measures of departure from normality in the sense that if X is normal, all cumulants of order three or more are zero. This aspect is developed in greater detail in Chapter 5 where Edgeworth expansions are introduced.

Suppose now we have two random variables X^1 and X^2. With index notation, the mixed cumulants are denoted by $\kappa^{1,2}$, $\kappa^{1,1,2}$, $\kappa^{1,2,2}$, $\kappa^{1,1,1,2}, \ldots$. The corresponding quantities in power notation are κ_{11}, κ_{21}, κ_{12}, κ_{31} and so on. To the extent that third and higher-order cumulants can be neglected, we find from the bivariate normal approximation that

$$\begin{aligned} E(X^2|X^1 = x^1) &\simeq \kappa_{01} + (\kappa_{11}/\kappa_{20})(x^1 - \kappa_{10}) \\ E(X^1|X^2 = x^2) &\simeq \kappa_{10} + (\kappa_{11}/\kappa_{02})(x^2 - \kappa_{01}) \end{aligned} \tag{2.12}$$

so that $\kappa_{11} > 0$ implies positive dependence in the sense of increasing conditional expectations. Refined versions of the above, taking third- and fourth-order cumulants into account, are given in Chapter 5.

The simplest interpretations of bivariate cumulants are given in terms of independence. If X^1 and X^2 are independent, then all mixed cumulants involving X^1 and X^2 alone are zero. Thus, $\kappa^{1,2} = \kappa^{1,1,2} = \kappa^{1,2,2} = \ldots = 0$ or, more concisely using power notation, $\kappa_{rs} = 0$ for all $r, s \ge 1$. Provided that the moments determine the joint distribution, the converse is also true, namely that if $\kappa_{rs} = 0$ for all $r, s \ge 1$, then X^1 and X^2 are independent. The suggestion here is that if $\kappa_{rs} = 0$ for $r, s = 1, \ldots, t$, say, then X^1 and X^2 are approximately independent in some sense. However it is difficult to make this claim rigorous except in the asymptotic sense of Chapter 5 where all higher-order cumulants are negligible.

Consider now the case of three random variables whose mixed cumulants may be denoted by κ_{111}, κ_{211}, κ_{121}, κ_{112}, κ_{311}, κ_{221}, and so on by an obvious extension of the notation of Section 2.5. It is not difficult to see that if X^1 is independent of (X^2, X^3), or if X^2 is independent of (X^1, X^3), or if X^3 is independent of (X^1, X^2) then $\kappa_{111} = 0$ and, in fact, more generally, $\kappa_{rst} = 0$ for all $r, s, t \geq 1$. These independence relationships are most succinctly expressed using generalized power notation and it is for this reason that we switch here from one notation to the other.

More generally, if we have any number of random variables that can be partitioned into two independent blocks, then all mixed cumulants involving indices from both blocks are zero. Note that if X^1 and X^2 are independent, it does *not* follow from the above that, say, $\kappa^{1,2,3} = 0$. For example, if $X^3 = X^1X^2$, then it follows from (3.2) that $\kappa^{1,2,3} = \kappa_{20}\kappa_{02}$ and this is strictly positive unless X^1 or X^2 is degenerate. For a more interesting example, see the bivariate exponential recurrence process $\{X_j, Y_j\}$, $j = \ldots, -1, 0, 1, \ldots$ described in Exercise 2.35, in which each X_j is independent of the marginal process $\{Y_j\}$ and conversely, Y_j is independent of the marginal process $\{X_j\}$, but the two processes are dependent.

To see how the converse works, we note that $\kappa_{rs0} = 0$ for all $r, s \geq 1$ implies independence of X^1 and X^2, but only if the joint distribution is determined by its moments. Similarly, $\kappa_{r0s} = 0$ implies independence of X^1 and X^3. This alone does not imply that X^1 is independent of the pair (X^2, X^3): for this we require, in addition to the above, that $\kappa_{rst} = 0$ for all $r, s, t \geq 1$.

2.7 The central limit theorem

2.7.1 *Sums of independent random variables*

Suppose that $X_1, \ldots, X_n$ are n independent vector-valued random variables where X_r has components $X_r^1, \ldots, X_r^p$. We do not assume that the observations are identically distributed and so the cumulants of X_r are denoted by $\kappa_r^i, \kappa_r^{i,j}, \kappa_r^{i,j,k}$ and so on. One of the most important properties of cumulants is that the joint cumulants of $X_{\bullet} = X_1 + \cdots + X_n$ are just the sums of the corresponding cumulants of the individual variables. Thus we may write $\kappa_{\bullet}^i$, $\kappa_{\bullet}^{i,j}$, $\kappa_{\bullet}^{i,j,k}$ for the joint cumulants of $X_{\bullet}$ where, for example, $\kappa_{\bullet}^{i,j} = \sum_{r=1}^n \kappa_r^{i,j}$. This property is not shared by moments and there is, therefore, some risk of confusion if we were to write $\kappa_{\bullet}^{ij} = E(X_{\bullet}^i X_{\bullet}^j)$ because

$$\kappa_{\bullet}^{ij} = \kappa_{\bullet}^{i,j} + \kappa_{\bullet}^i\kappa_{\bullet}^j \neq \sum_r \kappa_r^{ij}.$$

For this reason, it is best to avoid the notation $\kappa_{\bullet}^{ij}$.

To demonstrate the additive property of cumulants we use the method of generating functions and write

$$M_{X_{\bullet}}(\xi) = E[\exp\{\xi_i(X_1^i + \cdots + X_n^i)\}].$$

By independence we have,

$$M_{X_{\bullet}}(\xi) = M_{X_1}(\xi)\cdots M_{X_n}(\xi)$$

and thus

$$K_{X_{\bullet}}(\xi) = K_{X_1}(\xi) + \cdots + K_{X_n}(\xi).$$

The required result follows on extraction of the appropriate coefficients of ξ.

In the particular case where the observations are identically distributed, we have that $\kappa_{\bullet}^i = n\kappa^i$, $\kappa_{\bullet}^{i,j} = n\kappa^{i,j}$ and so on. If the observations are not identically distributed, it is sometimes convenient to define average cumulants by writing

$$\kappa_{\bullet}^i = n\bar\kappa^i, \quad \kappa_{\bullet}^{i,j} = n\bar\kappa^{i,j}, \quad \kappa_{\bullet}^{i,j,k} = n\bar\kappa^{i,j,k}$$

and so on where, for example, $\bar\kappa^{i,j}$ is the average covariance of components i and j.

2.7.2 *Standardized sums*

With the notation of the previous section, we write

$$Y^i = n^{-1/2}\{X^i_{\bullet} - n\bar{\kappa}^i\}.$$

The cumulants of $X_{\bullet}$ are $n\bar{\kappa}^i$, $n\bar{\kappa}^{i,j}$, $n\bar{\kappa}^{i,j,k}$ and so on. It follows from Section 2.4 that the cumulants of Y are

$$0, \quad \bar{\kappa}^{i,j}, \quad n^{-1/2}\bar{\kappa}^{i,j,k}, \quad n^{-1}\bar{\kappa}^{i,j,k,l}, \ldots$$

the successive factors decreasing in powers of $n^{-1/2}$. Of course, the average cumulants are themselves implicitly functions of n and without further, admittedly mild, assumptions, there is no guarantee that $n^{-1/2}\bar{\kappa}^{i,j,k}$ or $n^{-1}\bar{\kappa}^{i,j,k,l}$ will be negligible for large n. We avoid such difficulties in the most direct way, simply by assuming that $\bar{\kappa}^{i,j}, \bar{\kappa}^{i,j,k}, \ldots$ have finite limits as $n \to \infty$ and that the limiting covariance matrix, $\bar{\kappa}^{i,j}$ is positive definite. In many problems, these assumptions are entirely reasonable but they do require checking. See Exercise 2.10 for a simple instance of failure of these assumptions.

The cumulant generating function of Y is

$$K_Y(\xi) = \xi_i\xi_j\bar{\kappa}^{i,j}/2! + n^{-1/2}\xi_i\xi_j\xi_k\bar{\kappa}^{i,j,k}/3! + \ldots. \tag{2.13}$$

Under the assumptions just given, and for complex ξ, the remainder in this series after r terms is $O(n^{-r/2})$. Now, $\xi_i\xi_j\bar{\kappa}^{i,j}/2$ is the cumulant generating function of a normal random variable with zero mean and covariance matrix $\bar{\kappa}^{i,j}$. Since convergence of the cumulant generating function implies convergence in distribution, subject to continuity of the limiting function at the origin (Moran, 1968, Section 6.2), we have just proved a simple version of the central limit theorem for independent but non-identically distributed random variables. In fact, it is not necessary here to use the generating function directly. Convergence of the moments implies convergence in distribution provided that the limiting moments uniquely determine the distribution, as they do in this case. For a more accurate approximation to the density of Y, we may invert (2.13) formally, leading to an asymptotic expansion in powers of $n^{-1/2}$. This expansion is known as the Edgeworth series after F.Y. Edgeworth (1845–1926). Note however, that although the error in (2.13) after two terms is $O(n^{-1})$, the error in probability calculations based on integrating the formal inverse of (2.13) need not be $O(n^{-1})$. In discrete problems, the error is typically $O(n^{-1/2})$.

Stronger forms of the central limit theorem that apply under conditions substantially weaker than those assumed here, are available in the literature. In particular, versions are available in which finiteness of the higher-order cumulants is not a requirement. Such theorems, on occasion, have statistical applications but they sometimes suffer from the disadvantage that the error term for finite n may be large and difficult to quantify, even in an asymptotic sense. Often the error is $o(1)$ as opposed to $O(n^{-1/2})$ under the kind of assumptions made here. Other forms of the central limit theorem are available in which certain specific types of dependence are permitted. For example, in applications related to time series, it is often reasonable to assume that observations sufficiently separated in time must be nearly independent. With additional mild assumptions, this ensures that the asymptotic cumulants of derived statistics are of the required order of magnitude in n and the central limit result follows.

2.8 Derived scalars

Suppose we are interested in the distribution of the statistic

$$T^2 = (X^i - \kappa_0^i)(X^j - \kappa_0^i)\kappa_{i,j}$$

where $\kappa_{i,j}$ is the matrix inverse or generalized inverse of $\kappa^{i,j}$. The Mahalanobis statistic, T^2, is a natural choice that arises if we are testing the hypothesis $H_0 : \kappa^i = \kappa_0^i$, where the higher-order cumulants are assumed known. One reason for considering this particular statistic is that it is invariant under affine transformation of X. Its distribution must therefore depend on scalars derived from the cumulants of X that are invariant under affine nonsingular transformations

$$X^r \to a^r + a_i^r X^i$$

where a_i^r is a $p \times p$ matrix of rank p. These scalars are the multivariate generalizations of ρ_3^2, ρ_4, ρ_6, ρ_5^2 and so on, in the univariate case: see Section 2.6.

To obtain the multivariate generalization of $\rho_3^2 = \kappa_3^2/\kappa_2^3$, we first write $\kappa^{i,j,k}\kappa^{l,m,n}$ as the generalization of κ_3^2. Division by κ_2 generalizes to multiplication by $\kappa_{r,s}$. Thus we require a scalar derived from

$$\kappa^{i,j,k}\kappa^{l,m,n}\kappa_{r,s}\kappa_{t,u}\kappa_{v,w}$$

by contraction, i.e. by equating pairs of indices and summing. This operation can be done in just two distinct ways giving two non-negative scalars

$$\bar\rho_{13}^2 = \rho_{13}^2/p = \kappa^{i,j,k}\kappa^{l,m,n}\kappa_{i,j}\kappa_{k,l}\kappa_{m,n}/p, \tag{2.14}$$

$$\bar\rho_{23}^2 = \rho_{23}^2/p = \kappa^{i,j,k}\kappa^{l,m,n}\kappa_{i,l}\kappa_{j,m}\kappa_{k,n}/p. \tag{2.15}$$

Similarly, to generalize $\rho_4 = \kappa_4/\kappa_2^2$, we obtain just the single expression

$$\bar\rho_4 = \rho_4/p = \kappa^{i,j,k,l}\kappa_{i,j}\kappa_{k,l}/p. \tag{2.16}$$

In the univariate case, (2.14) and (2.15) reduce to the same quantity and $\bar\rho_4$ satisfies the familiar inequality $\bar\rho_4 \geq \bar\rho_3^2 - 2$. The multivariate generalization of this inequality applies most directly to $\bar\rho_{13}^2$, giving

$$\bar\rho_4 \geq \bar\rho_{13}^2 - 2.$$

Equality is achieved if and only if the joint distribution of the Xs is concentrated on some conic in p-space in which the coefficients of the quadratic term are $\kappa_{i,j}$. The support of the distribution may be degenerate at a finite number of points but it is assumed here that it is not contained in any lower dimensional subspace. Otherwise, the covariance matrix would be rank deficient and p would be replaced by the rank of the subspace. In the univariate case, equality is achieved if and only if the distribution is concentrated on two points: see Exercise 2.12.

The corresponding inequality for $\bar\rho_{23}^2$ is

$$\bar\rho_4 \geq \bar\rho_{23}^2 - p - 1.$$

See Exercise 2.14. This limit is attained if and only if the joint distribution is concentrated on $p+1$ points not contained in any linear subspace of R^p. The inequality for $\bar\rho_{23}^2$ is obtained by taking the trace of the residual covariance matrix of the products after linear regression on the linear terms. The trace vanishes only if this matrix is identically zero. Thus, achievement of the bound for $\bar\rho_{23}^2$ implies achievement of the bound for $\bar\rho_{13}^2$ and also that the higher-order cumulants are determined by those up to order four. See Section 3.8.

The simplest example that illustrates the difference between the two skewness scalars is the multinomial distribution with index m and parameter vector $\pi_1, \ldots, \pi_k$. The joint cumulants are given in Exercise 2.16 and the covariance matrix $\kappa^{i,j} = m\{\pi_i\delta_{ij} - \pi_i\pi_j\}$, has rank $p = k - 1$. The simplest generalized inverse is $\kappa_{i,j} = \{m\pi_i\}^{-1}$ for $i = j$ and zero otherwise. Substitution of the expressions in Exercise 2.16 for the third and fourth cumulants into (2.14) – (2.16) gives

$$\begin{aligned} m(k-1)\bar{\rho}_{13}^2 &= \sum\nolimits_j \pi_j^{-1} - k^2 \\ m(k-1)\bar{\rho}_{23}^2 &= \sum\nolimits_j (1-\pi_j)(1-2\pi_j)/\pi_j \\ m(k-1)\bar{\rho}_4 &= \sum\nolimits_j \pi_j^{-1} - k^2 - 2(k-1). \end{aligned}$$

Thus $\bar{\rho}_{13}^2$ is zero for the uniform multinomial distribution even though $\kappa^{i,j,k}$ is not identically zero. On the other hand, $\bar{\rho}_{23}^2$ is zero only if $\kappa^{i,j,k}$ is identically zero and, in the case of the multinomial distribution, this cannot occur unless $k = 2$ and $\pi = \frac{1}{2}$. For additional interpretations of the differences between these two scalars: see Exercise 2.15.

The above invariants are the three scalars most commonly encountered in theoretical work such as the expansion of the log likelihood ratio statistic or computing the variance of T^2. However, they are not the only invariant functions of the first four cumulants. A trivial example is $\kappa^{i,j}\kappa_{i,j} = p$, or more generally, the rank of $\kappa^{i,j}$. Also, if we were to generalize $\rho_4^2 = \kappa_4^2/\kappa_2^4$ by considering quadratic expressions in $\kappa^{i,j,k,l}$, there are two possibilities in addition to $(\bar{\rho}_4)^2$. These are

$$\begin{aligned} \bar{\rho}_{14}^2 &= \kappa^{i,j,k,l}\kappa^{r,s,t,u}\kappa_{i,j}\kappa_{k,r}\kappa_{l,s}\kappa_{t,u}/p, \\ \bar{\rho}_{24}^2 &= \kappa^{i,j,k,l}\kappa^{r,s,t,u}\kappa_{i,r}\kappa_{j,s}\kappa_{k,t}\kappa_{l,u}/p. \end{aligned}$$

In addition to these, there are integer invariants of a qualitatively different kind, obtained by extending the notions of rank and signature to multi-way arrays. Kruskal (1977) defines the rank of a three-way asymmetrical array in a way that is consistent with the standard definition for matrices. In addition, the four-way array, $\kappa^{i,j,k,l}$ can be thought of as a symmetric $p^2 \times p^2$ matrix whose rank and signature are invariants. See Section 1.5.2 and Exercise 1.9. There may also be other invariants unconnected with the notions of rank or signature but none have appeared in the literature. However, it seems unnatural to consider $\kappa^{i,j,k,l}$ as a two-way array and, not surprisingly, the integer invariants just mentioned do not arise in the usual statistical calculations. For completeness, however, it would be good to know the complete list of all invariants that can be formed from, say, the first four cumulants.

To see that (2.14)–(2.16) are indeed invariant under affine transformation, we note that

$$\kappa^{i,j,k} \to a_i^r a_j^s a_k^t \kappa^{i,j,k} \quad \text{and} \quad \kappa_{i,j} \to b_r^i b_s^j \kappa_{i,j},$$

where b_r^i is the matrix inverse of a_i^r. Direct substitution followed by cancellation reveals the invariance. This invariance property of scalars derived by contraction of tensors is an elementary consequence of the tensor transformation property. Provided that we work exclusively with tensors, it is not necessary to check that scalars derived in this way are invariant.

2.9 Conditional cumulants

Suppose we are given the conditional joint cumulants of the random variables $X^1, \ldots, X^p$ conditional on some event A. How do we combine the conditional cumulants to obtain the unconditional joint cumulants? In the case of moments, the answer is easy because

$$E(X^1X^2\cdots) = E_A E(X^1X^2\cdots|A). \qquad (2.17)$$

In other words, the unconditional moments are just the average of the conditional moments. However, it is not difficult to show, for example, that the covariance of X^i and X^j satisfies

$$\begin{aligned}\kappa^{i,j} =& E_A\{\mathrm{cov}(X^i, X^j|A)\}\\ &+ \mathrm{cov}_A\{E(X^i|A), E(X^j|A)\}.\end{aligned} \tag{2.18}$$

To see how these expressions generalize to cumulants of arbitrary order, we denote the conditional cumulants by λ^i, $\lambda^{i,j}$, $\lambda^{i,j,k}$ and we use the identity connecting the moment generating functions

$$M_X(\xi) = E_A M_{X|A}(\xi).$$

Expansion of this identity and comparison of coefficients gives

$$\begin{aligned}\kappa^i &= E_A\{\lambda^i\}\\ \kappa^{i,j} + \kappa^i\kappa^j &= E_A\{\lambda^{i,j} + \lambda^i\lambda^j\}\\ \kappa^{i,j,k} + \kappa^i\kappa^{j,k}[3] + \kappa^i\kappa^j\kappa^k &= E_A\{\lambda^{i,j,k} + \lambda^i\lambda^{j,k}[3] + \lambda^i\lambda^j\lambda^k\}.\end{aligned}$$

Expression (2.18) for the unconditional covariance follows from the second expression above. On using this result in the third expression, we find

$$\kappa^{i,j,k} = E(\lambda^{i,j,k}) + \kappa_2(\lambda^i, \lambda^{j,k})[3] + \kappa_3(\lambda^i, \lambda^j, \lambda^k).$$

The generalization is easy to see though, for notational reasons, a little awkward to prove, namely

$$\kappa(\Upsilon_p) = \sum_{\Upsilon} \kappa_\nu\{\lambda(\upsilon_1), \ldots, \lambda(\upsilon_\nu)\}$$

with summation over all partitions of the p indices. In this expression, $\lambda(\upsilon_j)$ is the conditional mixed cumulant of the random variables whose indices are in υ_j and $\kappa_\nu\{\lambda(\upsilon_1), \ldots, \lambda(\upsilon_\nu)\}$ is the νth order cumulant of the ν random variables listed as arguments. For details of a proof see the papers by Brillinger (1969) or Speed (1983).

In many circumstances, it is required to compute conditional cumulants from joint unconditional cumulants, the converse of the result just described. A little reflection soon shows that the converse problem is considerably more difficult and the best that can be expected are approximate conditional cumulants. Expansions of this type are given in Chapter 5.

However, if $M_{X,Y}(\xi,\zeta) = E\{\exp(\xi_i X^i + \zeta_r Y^r)\}$ is the joint moment generating function of X, Y, we can at least write down an expression for the conditional moment generating function $M_{X|Y}(\xi)$. Since

$$M_{X,Y}(\xi,\zeta) = \int \exp(\zeta_r Y^r) M_{X|Y}(\xi) f_Y(y)dy,$$

we may invert the integral transform to find

$$k\, M_{X|Y}(\xi) f_Y(y) = \int_{c-i\infty}^{c+i\infty} M_{X,Y}(\xi,\zeta)\exp(-\zeta_r y^r)d\zeta.$$

Division by $f_Y(y)$ gives the conditional moment generating function in the form

$$M_{X|Y}(\xi) = \frac{\int M_{X,Y}(\xi,\zeta)\exp(-\zeta_r y^r)d\zeta}{\int M_{X,Y}(0,\zeta)\exp(-\zeta_r y^r)d\zeta}. \tag{2.19}$$

This expression, due to Bartlett (1938), can be used for generating expansions or approximations for conditional moments: it is rarely used directly for exact calculations. See, however, Moran (1968, Section 6.14).

2.10 Bibliographic notes

Cumulants were first defined by Thiele in about 1889. Thiele's work is mostly in Danish and the most accessible English translation is of his book *Theory of Observations* reprinted in *The Annals of Mathematical Statistics* (1931), pp.165-308. In Chapter 6 of that book, Thiele defines the univariate cumulants and calls them *half-invariants* on account of their simple transformation properties. He also derives a version of the central limit theorem by showing that the higher-order cumulants of a standardized linear combination of random variables converge rapidly to zero provided that 'the coefficient of any single term is not so great ... that it throws all the other terms into the shade', a delightful statement closely approximating the spirit of the Lindeberg-Feller condition.

For an excellent readable account of the central limit theorem and the Lindeberg condition, see LeCam (1986).

Fisher (1929), in an astonishing tour de force, rediscovered cumulants, recognized their superiority over moments, developed the corresponding sample cumulants and cross-cumulants, gave the formulae for the cumulants and cross-cumulants of the sample cumulants and formulated combinatorial rules for computing the cumulants of such statistics. Both Thiele and Fisher used what we call 'power notation' for cumulants and cross-cumulants and their achievements are the more remarkable for that reason.

Formulae giving univariate moments in terms of cumulants and vice versa are listed in Kendall & Stuart (1977, Chapter 3): multivariate versions of these formulae are given in Chapter 13. See also David, Kendall & Barton (1966, Tables 2.1.1 and 2.1.2) and David & Barton (1962, Chapter 9). Similar formulae are given by Brillinger (1975, Chapter 2).

The derived scalars $\bar\rho_4$ and $\bar\rho_{23}^2$ were given by Mardia (1970) as summary measures of multivariate kurtosis and skewness. The additional skewness scalar, $\bar\rho_{13}^2$ is given by McCullagh & Cox (1986) who show how it arises in calculations involving likelihood ratio tests. See also Davis (1980) who shows how the scalar $\psi_p = p(\bar\rho_{13}^2 - \bar\rho_{23}^2)$ arises in calculations concerning the effect of non-normality on the distribution of Wilks's Λ.

2.11 Further results and exercises 2

2.1 Let X have density function given by

$$f_X(x) = \begin{cases} 2/x^3 & x \ge 1 \\ 0 & \text{otherwise.} \end{cases}$$

Show that the mean of X is finite but that all higher-order moments are infinite. Find an expression for the density of $Y = 1/X$ and show that all moments and cumulants are finite. Let $\mu'_{rs} = E(X^r Y^s)$ be the joint moment of order $r+s$ (using the power notation of Section 2.5). Show that $\mu'_{21} = 2$ but that the corresponding cumulant, κ_{21}, is infinite.

2.2 Let X be a standard normal random variable and set $Y = \exp(X)$. Show that the rth moment of Y about the origin is $\mu'_r = \exp(r^2/2)$. Hence find expressions for the first four cumulants of Y. Show that the series expansions for $M_Y(\xi)$ and $K_Y(\xi)$ about $\xi = 0$ are divergent for all real $\xi > 0$ even though all cumulants are finite (Heyde, 1963).

2.3 Let X be a scalar random variable. Prove by induction on r that the derivatives of $M_X(\xi)$ and $K_X(\xi)$, if they exist, satisfy

$$M_X^{(r)}(\xi) = \sum_{j=1}^{r} \binom{r-1}{j-1} M_X^{(r-j)}(\xi) K_X^{(j)}(\xi) \qquad r \ge 1.$$

Hence show that

$$\mu'_r = \kappa_r + \sum_{j=1}^{r-1} \binom{r-1}{j-1} \kappa_j \mu'_{r-j}$$

and, for $r \geq 4$, that

$$\mu_r = \kappa_r + \sum_{j=2}^{r-2} \binom{r-1}{j-1} \kappa_j \mu_{r-j}$$

(Thiele, 1897, eqn. 22; Morris, 1982).

2.4 If $\mu(\Upsilon)$ and $\kappa(\Upsilon)$ denote the ordinary moment and the ordinary cumulant corresponding to the indices in $\Upsilon = \{i_1, \ldots, i_p\}$, show that

$$\mu(\Upsilon) = \kappa(\Upsilon) + \sum_{\{\upsilon_1, \upsilon_2\}} \kappa(\upsilon_1)\mu(\upsilon_2)$$

where $\{\upsilon_1, \upsilon_2\}$ is a partition of Υ into two non-empty blocks and the sum extends over all partitions such that $i_1 \in \upsilon_1$. What purpose does the condition $i_1 \in \upsilon_1$ serve? Show that this result generalizes the univariate identity in Exercise 2.3.

2.5 Show that the central and non-central moments satisfy

$$\mu'_r = \sum_{j=0}^{r} \binom{r}{j} \mu_{r-j} (\mu'_1)^j$$

$$\mu_r = \sum_{j=0}^{r} \binom{r}{j} \mu'_{r-j} (-\mu'_1)^j .$$

(Kendall and Stuart, 1977, p. 58).

2.6 The density function of Student's distribution on ν degrees of freedom is

$$\frac{(1+t^2/\nu)^{-(\nu+1)/2}}{\nu^{1/2} B(\frac{1}{2}, \nu/2)} \qquad -\infty < t < \infty,$$

where $B(.,.)$ is the beta function. Show that the odd moments that exist are zero and that the even moments are

$$\mu_{2r} = \frac{1.3.\cdots(2r-1)\nu^r}{(\nu-2)\cdots(\nu-2r)} \qquad 2r < \nu.$$

Hence show that

$$\rho_4 = 6/(\nu-4) \quad \text{for} \quad \nu > 4$$

$$\rho_6 = 240/\{(\nu-4)(\nu-6)\} \quad \text{for} \quad \nu > 6.$$

2.7 Prove directly using (2.7) that if X^i is independent of the pair (X^j, X^k), then $\kappa^{i,j,k} = 0$.

2.8 Show that if X^1 is independent of (X^2, X^3), then the cumulant κ_{rst} of order $r+s+t$ with $r, s, t \geq 1$ is equal to zero.

2.9 Derive expressions (2.7) for cumulants in terms of moments (i) directly from (2.6) and (ii) by expansion of $\log M_X(\xi)$.

2.10 Let $h_r(x)$ be the standardized Hermite polynomial of degree r satisfying the condition $\int h_r(x)h_s(x)\phi(x)dx = \delta_{rs}$ where $\phi(x)$ is the standard normal density. If $X_i = h_i(Z)$ where Z is a standard normal variable, show that $X_1, \ldots$ are uncorrelated but not independent. Show also that the second cumulant of $n^{1/2}\bar{X}$ is exactly one but that the third cumulant does not converge to zero. Construct a similar example in which the first three cumulants converge to those of the standard normal density, but where the central limit theorem does not apply.

2.11 Verify that $\bar{\rho}_4 = p^{-1}\kappa^{i,j,k,l}\kappa_{i,j}\kappa_{k,l}$ is invariant under affine non-singular transformation of X.

2.12 By considering the expression

$$\operatorname{var}\{\kappa_{i,j}X^i X^j - c_i X^i\}$$

with $c_i = \kappa_{i,r}\kappa_{s,t}\kappa^{r,s,t}$, show that $\bar{\rho}_4 \geq \bar{\rho}^2_{13} - 2$, where $\bar{\rho}_4$ and $\bar{\rho}^2_{13}$ are defined by (2.14) and (2.16). In addition, by examining the expression

$$\int (a + a_i x^i + a_{ij}x^i x^j)^2 f_X(x)\, dx,$$

show that $\bar{\rho}_4 = \bar{\rho}^2_{13} - 2$ if and only if the joint distribution of X is concentrated on a particular class of conic.

2.13 Show that $\bar{\rho}^2_{23} \geq 0$ with equality only if $\kappa^{i,j,k} = 0$ identically.

2.14 Show that $\kappa^{ij,kl} - \kappa^{i,j,r}\kappa^{k,l,s}\kappa_{r,s}$, regarded as a $p^2 \times p^2$ symmetric matrix, is non-negative definite. By examining the trace of this matrix in the case where $\kappa^i = 0$, $\kappa^{i,j} = \delta^{ij}$, show that

$$\bar{\rho}_4 \geq \bar{\rho}^2_{23} - p - 1,$$

with equality if and only if the joint distribution is concentrated on $p + 1$ points not contained in any linear subspace of R^p. Deduce that $\bar{\rho}_4 = \bar{\rho}^2_{23} - p - 1$ implies $\bar{\rho}_4 = \bar{\rho}^2_{13} - 2$ and that $\bar{\rho}_4 > \bar{\rho}^2_{13} - 2$ implies $\bar{\rho}_4 > \bar{\rho}^2_{23} - p - 1$.

2.15 Show that $\bar{\rho}^2_{13} = 0$ if and only if every linear combination $a_i X^i$ is uncorrelated with the quadratic form $\kappa_{i,j}X^i X^j$. (Take $\kappa^i = 0$.) Show also that $\bar{\rho}^2_{23} = 0$ if and only if every linear combination $a_i X^i$ is uncorrelated with every quadratic form $a_{ij}X^i X^j$.

2.16 Show that the multinomial distribution with index m and probability vector $\pi_1, \ldots, \pi_k$ has cumulant generating function

$$m\{k(\theta + \xi) - k(\theta)\},$$

where $k(\theta) = \log[\Sigma \exp(\theta_j)]$ and

$$\pi_i = \exp(\theta_i) / \sum \exp(\theta_j).$$

Hence show that the first four cumulants are

$$\begin{aligned}
\kappa^i &= m\pi_i \\
\kappa^{i,j} &= m\{\pi_i\delta_{ij} - \pi_i\pi_j\} \\
\kappa^{i,j,k} &= m\{\pi_i\delta_{ijk} - \pi_i\pi_j\delta_{ik}[3] + 2\pi_i\pi_j\pi_k\} \\
\kappa^{i,j,k,l} &= m\{\pi_i\delta_{ijkl} - \pi_i\pi_j(\delta_{ik}\delta_{jl}[3] + \delta_{jkl}[4]) + 2\pi_i\pi_j\pi_k\delta_{il}[6] \\
&\qquad - 6\pi_i\pi_j\pi_k\pi_l\},
\end{aligned}$$

where, for example, $\delta_{ijk} = 1$ if $i = j = k$ and zero otherwise, and no summation is implied where indices are repeated at the same level.

2.17 Evaluate explicitly the fourth cumulants of the multinomial distribution for the five distinct index patterns.

2.18 Show, for the multinomial distribution with index $m = 1$, that the moments are $\kappa^i = \pi_i$, $\kappa^{ij} = \pi_i\delta_{ij}$, $\kappa^{ijk} = \pi_i\delta_{ijk}$ and so on, where no summation is implied. Hence give an alternative derivation of the first four cumulants in Exercise 2.16.

2.19 For the multinomial distribution with $p = \operatorname{rank}(\kappa^{i,j}) = k-1$, show that

$$\begin{aligned}(k-1)\bar\rho_{13}^2 &= m^{-1}\{\textstyle\sum \pi_j^{-1} - k^2\},\\ (k-1)\bar\rho_{23}^2 &= m^{-1}\{\textstyle\sum \pi_j^{-1} - 3k + 2\}\\ (k-1)\bar\rho_4 &= m^{-1}\{\textstyle\sum \pi_j^{-1} - k^2 - 2(k-1)\},\end{aligned}$$

and hence that

$$\bar\rho_4 = \bar\rho_{13}^2 - 2/m = \bar\rho_{23}^2 - k/m,$$

showing that the inequalities in Exercises 2.12 and 2.14 are sharp for $m = 1$. Show also that the minimum value of $\bar\rho_{23}^2$ for the multinomial distribution is $(k-2)/m$.

2.20 Hölder's inequality for a pair of random variables X and Y is

$$E|XY| \le \{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q}$$

where $p^{-1} + q^{-1} = 1$. Deduce from the above that

$$\{E|X_1X_2\cdots X_r|\}^r \le E|X_1|^r \cdots E|X_r|^r$$

for random variables $X_1, \ldots, X_r$. Hence prove that if the diagonal elements of cumulant tensors are finite then all other elements are finite.

2.21 Using (2.6) and (2.7), express $\kappa^{i,jkl} = \operatorname{cov}(X^i, X^jX^kX^l)$ in terms of ordinary moments and hence, in terms of ordinary cumulants.

2.22 Let $a = a^1, \ldots, a^p$ and $b = b^1, \ldots, b^p$, where $a^j \le b^j$, be the coordinates of two points in R^p and denote by (a, b) the Cartesian product in R^p of the intervals (a^j, b^j). Let $f(x) = f(x^1, \ldots, x^p)$ be a p-dimensional joint density function and define

$$F(a,b) = \int_{x\in(a,b)} f(x)\,dx$$

where $dx = dx^1\cdots dx^p$. Express $F(a,b)$ in terms of the cumulative distribution function $F(x) \equiv F(-\infty, x)$ evaluated at points with coordinates $x^j = a^j$ or b^j. Comment briefly on the similarities with and differences from (2.9).

2.23 Let a^{ij} be the elements of a square matrix, not necessarily symmetrical, and let its inverse, a_{ij}, satisfy $a^{ij}a_{kj} = \delta^i_k = a^{ji}a_{jk}$, being careful regarding the positions of the indices. Show that the derivatives satisfy

$$\begin{aligned}\partial a_{ij}/\partial a^{rs} &= -a_{is}a_{rj}\\ \partial a^{ij}/\partial a_{rs} &= -a^{is}a^{rj}.\end{aligned}$$

2.24 Show that if a_i are the components of a vector of coefficients satisfying $a_ia_j\kappa^{i,j} = 0$, then

$$a_i\kappa^{i,j} = 0, \quad a_i\kappa^{i,j,k} = 0, \quad a_i\kappa^{i,j,k,l} = 0$$

and so on. Hence deduce that the choice of generalized inverse has no effect on the scalars derived in Section 2.8.

2.25 Let $X_1, \ldots, X_n$ be independent and identically distributed scalar random variables having cumulants κ_1, κ_2, $\kappa_3, \ldots$. Show that

$$\begin{aligned} 2\kappa_2 &= E(X_1 - X_2)^2 \\ 3\kappa_3 &= E(X_1 + \omega X_2 + \omega^2 X_3)^3 \quad (\omega = e^{2\pi i/3}) \\ 4\kappa_4 &= E(X_1 + \omega X_2 + \omega^2 X_3 + \omega^3 X_4)^4 \quad (\omega = e^{2\pi i/4}) \end{aligned}$$

where $i^2 = -1$. Hence, by writing $\omega = e^{2\pi i/r}$ and

$$r\kappa_r = \lim_{n\to\infty} n^{-1}[X_1 + \omega X_2 + \omega^2 X_3 + \cdots + \omega^{nr-1} X_{nr}]^r,$$

give an interpretation of cumulants as coefficients in the Fourier transform of the randomly ordered sequence $X_1, X_2, \ldots$. Express κ_r as a *symmetric* function of $X_1, X_2, \ldots$, (Good, 1975, 1977).

2.26 Show that if $\Upsilon = \{\upsilon_1, \ldots, \upsilon_\nu\}$ is a partition of the indices $j_1, \ldots, j_n$ and $\omega = e^{2\pi i/n}$ is a primitive nth root of unity, then

$$\sum_j \omega^{j_1 + \cdots + j_n}\, \delta(\upsilon_1) \cdots \delta(\upsilon_\nu) = \begin{cases} 0 \text{ if } \Upsilon < 1 \\ n \text{ if } \Upsilon = 1 \end{cases}$$

where the sum extends over all positive integer vectors having components in the range $(1, n)$.

2.27 Let $X_1, \ldots, X_n$ be independent and identically distributed p-dimensional random vectors having cumulants κ^r, $\kappa^{r,s}$, $\kappa^{r,s,t}, \ldots$. Define the random vector $Z_{(n)}$by

$$Z^r_{(n)} = \sum_{j=1}^n X^r_j \exp(2\pi i j/n)$$

where $i^2 = -1$. Using the result in the previous exercise or otherwise, show that the nth-order moments of $Z_{(n)}$ are the same as the nth-order cumulants of X, i.e.

$$E(Z^{r_1}_{(n)} \cdots Z^{r_n}_{(n)}) = \kappa^{r_1, \ldots, r_n},$$

(Good, 1975, 1977). Hence give an interpretation of mixed cumulants as Fourier coefficients along the lines of the interpretation in Exercise 2.25.

2.28 Let $X_1, \ldots, X_n$ be independent χ^2_1 random variables. Show that the joint cumulant generating function is $-\frac{1}{2}\sum_i \log(1 - 2\xi_i)$. Show also that the joint cumulant generating function of $Y_1 = \sum X_j$ and $Y_2 = \sum \lambda_j X_j$ is $-\frac{1}{2}\sum_i \log(1 - 2\xi_1 - 2\lambda_i \xi_2)$. Hence show that the joint cumulants of (Y_1, Y_2) are given by

$$\kappa_{rs} = 2^{r+s-1}(r + s - 1)! \sum \lambda_i^s$$

using the notation of Section 2.5.

2.29 Show that if the ratio $R = Y/X$ is independent of X, then the moments of the ratio are the ratio of the moments, i.e.

$$\mu'_r(R) = \mu'_r(Y)/\mu'_r(X).$$

2.30 In the notation of Exercise 2.28, let $R = Y_2/Y_1$. Show that the first four cumulants of this ratio are

$$\begin{aligned} \kappa_1(R) &= \kappa_1(\lambda) \\ \kappa_2(R) &= 2\kappa_2(\lambda)/(n+2) \\ \kappa_3(R) &= 8\kappa_3(\lambda)/\{(n+2)(n+4)\} \\ \kappa_4(R) &= 48\kappa_4(\lambda)/\{(n+2)(n+4)(n+6)\} \\ &\quad + 48n\kappa_2^2(\lambda)/\{(n+2)^2(n+4)(n+6)\} \end{aligned}$$

and explain what is meant by the notation $\kappa_r(\lambda)$.

2.31 Using the notation of Exercises 2.28 and 2.30, show that if the eigenvalues decrease exponentially fast, say $\lambda_j = \lambda^j$, with $|\lambda| < 1$, then nR has a non-degenerate limiting distribution for large n, with cumulants

$$\kappa_r(nR) \simeq 2^{r-1}(r-1)!\lambda^r/(1-\lambda^r).$$

Show that this result is false if λ is allowed to depend on n, say $\lambda_j = 1 - 1/j$.

2.32 Using (2.18), show that if X and Y are independent real-valued random variables, then the variance of the product is

$$\mathrm{var}(XY) = \kappa_{10}^2\kappa_{02} + \kappa_{01}^2\kappa_{20} + \kappa_{20}\kappa_{02}.$$

2.33 Let X_1 and X_2 be independent and identically distributed p-dimensional random variables with zero mean and identity covariance matrix. The spherical polar representation of X is written (R, θ) where θ has $p-1$ components and $R = |X|$ is the Euclidean norm of X. Show that

$$p\bar{\rho}_{13}^2 = E(R_1^3R_2^3\cos\theta_{12})$$
$$p\bar{\rho}_{23}^2 = E(R_1^3R_2^3\cos^3\theta_{12})$$

where $R_1R_2\cos\theta_{12} = X_1^iX_2^j\delta_{ij}$, so that θ_{12} is the random angle between X_1 and X_2. Hence give a geometrical interpretation of these two scalars in the special case where $p = 2$. Show also that

$$4p\bar{\rho}_{23}^2 - 3p\bar{\rho}_{13}^2 = E\{R_1^3R_2^3\cos(3\theta_{12})\}$$

and that this quantity is non-negative if $p \le 2$.

2.34 Let X be a scalar random variable and write

$$M_X^{(n)}(\xi) = 1 + \mu_1'\xi + \mu_2'\xi^2/2! + \cdots + \mu_n'\xi^n/n!$$

for the truncated moment generating function. The zeros of this function, $a_1^{-1}, \ldots, a_n^{-1}$, not necessarily real, are defined by

$$M_X^{(n)}(\xi) = (1-a_1\xi)(1-a_2\xi)\cdots(1-a_n\xi).$$

Show that the symmetric functions of the as

$$\langle rs\cdots u\rangle = \sum_{i<j<\cdots<l} a_i^r a_j^s \cdots a_k^u$$

are semi-invariants of X (unaffected by the transformation $X \to X + c$) if and only if the powers $r, s, \ldots, u$ that appear in the symmetric function are at least 2. Show also that the cumulants are given by the particular symmetric functions

$$\kappa_r = -(r-1)!\sum_i a_i^r = -(r-1)! < r > \qquad (r < n).$$

Express the semi-invariant $\langle 22\rangle$ in terms of κ_2 and κ_4. (MacMahon, 1884, 1886; Cayley, 1885).

2.35 Let $\{\epsilon_j\}$, $\{\epsilon'_j\}$ and $\{Z_j\}$, $j = \ldots, -1, 0, 1, \ldots$ be three doubly infinite, mutually independent sequences of independent unit exponential random variables. The bivariate sequence $\{X_j, Y_j\}$, $j = \ldots, -1, 0, 1, \ldots$, defined by

$$X_{j+1} = \begin{cases} X_j - Z_j & \text{if } X_j > Z_j \\ \epsilon_{j+1} & \text{otherwise} \end{cases}$$

$$Y_{j+1} = \begin{cases} Y_j - Z_j & \text{if } Y_j > Z_j \\ \epsilon'_{j+1} & \text{otherwise} \end{cases}$$

is known as a bivariate *exponential recurrence* process. Show that

(i) X_j and Y_j are unit exponential random variables.
(ii) $\mathrm{cov}(X_j, X_{j+h)} = 2^{-h}$ where $h \geq 0$.
(iii) X_i is independent of Y_j.
(iv) X_i is independent of the sequence $\{Y_j\}$.

Hence deduce that all third-order mixed cumulants involving both Xs and Ys are zero. Show also that

$$\mathrm{cum}(X_j, X_{j+1}, Y_j, Y_{j+1}) = 1/12$$

(McCullagh, 1984c).

2.36 In the two-dimensional case, show that the homogeneous cubic form $\kappa^{i,j,k} w_i w_j w_k$ can be written, using power notation, in the form

$$Q_3(w) = \kappa_{30} w_1^3 + \kappa_{03} w_2^3 + 3\kappa_{21} w_1^2 w_2 + 3\kappa_{12} w_1 w_2^2.$$

By transforming to polar coordinates, show that

$$Q_3(w) = r^3\{\tau_1 \cos(\theta - \epsilon_1) + \tau_3 \cos(3\theta - 3\epsilon_3)\},$$

where

$$16\tau_1^2 = 9(\kappa_{30} + \kappa_{12})^2 + 9(\kappa_{03} + \kappa_{21})^2$$
$$16\tau_3^2 = (\kappa_{30} - 3\kappa_{12})^2 + (\kappa_{03} - 3\kappa_{21})^2.$$

Find similar expressions for ϵ_1 and ϵ_3 in terms of the κs.

2.37 By taking X to be a two-dimensional standardized random variable with zero mean and identity covariance matrix, interpret $Q_3(w)$, defined in the previous exercise, as a directional standardized skewness. [Take w to be a unit vector.] Show that, in the polar representation, $\epsilon_3 - \epsilon_1$ is invariant under rotation of X, but changes sign under reflection. Find an expression for this semi-invariant in terms of κ_{30}, κ_{03}, κ_{21} and κ_{12}. Discuss the statistical implications of the following conditions:

(i) $4\rho_{23}^2 - 3\rho_{13}^2 = 0$;
(ii) $\rho_{13}^2 = 0$;
(iii) $\epsilon_3 - \epsilon_1 = 0$.

2.38 In the two-dimensional case, show that the homogeneous quartic form $\kappa^{i,j,k,l} w_i w_j w_k w_l$ can be written, using power notation, in the form

$$Q_4(w) = \kappa_{40} w_1^4 + \kappa_{04} w_2^4 + 4\kappa_{31} w_1^3 w_2 + 4\kappa_{13} w_1 w_2^3 + 6\kappa_{22} w_1^2 w_2^2.$$

By transforming to polar coordinates, show that

$$Q_4(w) = r^4\{\tau_0 + \tau_2 \cos(2\theta - 2\epsilon_2) + \tau_4 \cos(4\theta - 4\epsilon_4)\}.$$

Show that $8\tau_0 = 3\kappa_{40} + 3\kappa_{04} + 6\kappa_{22}$ is invariant under orthogonal transformation of X. Find similar expressions for τ_2, τ_4, ϵ_2 and ϵ_4 in terms of the κs.

2.39 By taking X to be a two-dimensional standardized random variable with zero mean and identity covariance matrix, interpret $Q_4(w)$, defined in the previous exercise, as a directional standardized kurtosis. Taking ϵ_1 as defined in Exercises 2.36 and 2.37, show, using the polar representation, that that $\epsilon_2 - \epsilon_1$ and $\epsilon_4 - \epsilon_1$ are both invariant under rotation of X, but change sign under reflection. Find expressions for these semi-invariants in terms of the κs. Interpret τ_0 as the mean directional kurtosis and express this as a function of $\bar{\rho}_4$. Discuss the statistical implications of the following conditions:

(i) $\tau_0 = 0$;
(ii) $\tau_2 = 0$;
(iii) $\tau_4 = 0$;
(iv) $\epsilon_4 - \epsilon_2 = 0$.

2.40 In the one-dimensional case, the functions of the cumulants of X that are invariant under affine transformation are

$$\frac{\kappa_{2r}}{\kappa_2^r}, \quad \frac{\kappa_{2r+1}^2}{\kappa_2^{2r+1}}, \quad r = 1, 2, \ldots$$

using power notation. All other invariants can be expressed as functions of this sequence. By extending the results described in the previous four exercises, describe the corresponding complete list of invariants and semi-invariants in the bivariate case.

2.41 Discuss the difficulties encountered in extending the above argument beyond the bivariate case.

2.42 *Spherically symmetric random variables*: A random variable X is said to be spherically symmetric if its distribution is unaffected by orthogonal transformation. Equivalently, in spherical polar coordinates, the radial vector is distributed independently of the angular displacement, which is uniformly distributed over the unit sphere. Show that the odd cumulants of such a random variable are zero and that the even cumulants must have the form

$$\kappa^{i,j} = \tau_2\delta^{ij}, \quad \kappa^{i,j,k,l} = \tau_4\delta^{ij}\delta^{kl}[3], \quad \kappa^{i,j,k,l,m,n} = \tau_6\delta^{ij}\delta^{kl}\delta^{mn}[15]$$

and so on, for some set of coefficients τ_2, $\tau_4, \ldots$. Show that the standardized cumulants are

$$\bar{\rho}_4 = \tau_4(p+2)/\tau_2^2, \quad \bar{\rho}_6 = \tau_6(p+2)(p+4)/\tau_2^3$$

and hence that $\tau_4 \geq -2\tau_2^2/(p+2)$.

2.43 For the previous exercise, show that

$$\tau_6 \geq -\tau_2\tau_4 - \tau_2^3 + \frac{3(\tau_4 + \tau_2^2)^2}{(p+4)\tau_2}.$$

CHAPTER 3

Generalized cumulants

3.1 Introduction and definitions

In Chapter 2 we examined in some detail how cumulant tensors transform under affine transformation of X. Cumulants of order two and higher transform like Cartesian tensors, but the first-order cumulant does not. In this chapter, we show how cumulant tensors transform under non-linear or non-affine transformation of X. The algorithm that we describe relies heavily on the use of index notation and is easy to implement with the assistance of suitable tables. Applications are numerous. For example, the maximum likelihood estimator and the maximized likelihood ratio statistic can be expressed as functions of the log likelihood derivatives at the true but unknown parameter point, θ. The distribution of these derivatives at the true θ is known as a function of θ. With the methods developed here, we may compute moments or cumulants of any derived statistic, typically as an asymptotic approximation, to any required order of approximation.

In general, it is a good deal more convenient to work with polynomial functions rather than, say, exponential or logarithmic functions of X. The first step in most calculations is therefore to expand the function of interest as a polynomial in X and to truncate at an appropriate point. The essential ingredient when working with polynomial functions is to develop a notation capable of coping with generalized cumulants of the type

$$\kappa^{i,jk} = \mathrm{cov}(X^i, X^j X^k).$$

It seems obvious and entirely natural to denote this quantity by $\kappa^{i,jk}$, thereby indexing the set of generalized cumulants by partitions of the indices. In other words, to each partition there corresponds a unique cumulant and to each cumulant there corresponds a unique partition. For example,

$$\begin{aligned}\kappa^{i,jkl} &= \mathrm{cov}(X^i, X^j X^k X^l)\\ \kappa^{ij,kl} &= \mathrm{cov}(X^i X^j, X^k X^l)\\ \kappa^{i,j,kl} &= \mathrm{cum}(X^i, X^j, X^k X^l).\end{aligned} \tag{3.1}$$

Thus $\kappa^{i,j,kl}$, the third-order cumulant of the three variables X^i, X^j and the product $X^k X^l$, is said to be of order $b = 3$ and degree $p = 4$. The order is the number of blocks of the partition and the degree is the number of indices. *Ordinary* cumulants with $b = p$ and *ordinary* moments with $b = 1$ are special cases of generalized cumulants. The order of the blocks and of the indices within blocks in (3.1) is immaterial provided only that the partition is preserved. In this way, the notion of symmetry under index permutation is carried over to generalized cumulants.

3.2 The fundamental identity for generalized cumulants

Just as moments can be expressed as combinations of ordinary cumulants according to (2.6), so too generalized cumulants can be expressed in a similar way. First, we give the expressions for the four generalized cumulants listed above and then the general formula is described. The following four formulae may be derived from first principles using (2.6) and (2.7).

$$\begin{aligned}
\kappa^{i,jk} &= \kappa^{ijk} - \kappa^{i}\kappa^{jk} \\
&= \kappa^{i,j,k} + \kappa^{j}\kappa^{i,k} + \kappa^{k}\kappa^{i,j} \\
\kappa^{i,jkl} &= \kappa^{ijkl} - \kappa^{i}\kappa^{jkl} \\
&= \kappa^{i,j,k,l} + \kappa^{j}\kappa^{i,k,l}[3] + \kappa^{i,j}\kappa^{k,l}[3] + \kappa^{i,j}\kappa^{k}\kappa^{l}[3] \\
\kappa^{ij,kl} &= \kappa^{ijkl} - \kappa^{ij}\kappa^{kl} \\
&= \kappa^{i,j,k,l} + \kappa^{i}\kappa^{j,k,l}[2] + \kappa^{k}\kappa^{i,j,l}[2] + \kappa^{i,k}\kappa^{j,l}[2] \\
&\qquad + \kappa^{i}\kappa^{k}\kappa^{j,l}[4] \\
\kappa^{i,j,kl} &= \kappa^{ijkl} - \kappa^{i}\kappa^{jkl} - \kappa^{j}\kappa^{ikl} - \kappa^{ij}\kappa^{kl} + 2\kappa^{i}\kappa^{j}\kappa^{kl} \\
&= \kappa^{i,j,k,l} + \kappa^{k}\kappa^{i,j,l}[2] + \kappa^{i,k}\kappa^{j,l}[2].
\end{aligned} \tag{3.2}$$

Again, the bracket notation has been employed, but now the interpretation of groups of terms depends on the context. Thus, for example, in the final expression above, $\kappa^{i,k}\kappa^{j,l}[2] = \kappa^{i,k}\kappa^{j,l} + \kappa^{i,l}\kappa^{j,k}$ must be interpreted in the context of the partition on the left, namely $i|j|kl$. The omitted partition of the same type is $ij|kl$, corresponding to the cumulant product $\kappa^{i,j}\kappa^{k,l}$. Occasionally, this shorthand notation may lead to ambiguity and if so, it becomes necessary to list the individual partitions explicitly. However, the reader quickly becomes overwhelmed by the sheer number of terms that complete lists involve. For this reason we make every effort to avoid explicit complete lists.

An alternative notation, useful in order to avoid the kinds of ambiguity alluded to above, is to write $\kappa^{i,k}\kappa^{j,l}[2]_{ij}$ for the final term in (3.2). However, this notation conflicts with the summation convention and, less seriously, $\kappa^{i,k}\kappa^{j,l}[2]_{ij}$ is the same as $\kappa^{i,k}\kappa^{j,l}[2]_{kl}$. For these reasons, the unadorned bracket notation will be employed where there is no risk of ambiguity.

From the above examples, it is possible to discern, at least qualitatively, the rule that expresses generalized cumulants in terms of ordinary cumulants. An arbitrary cumulant of order b involving p random variables may be written as $\kappa(\Upsilon^*)$ where $\Upsilon^* = \{\upsilon_1^*|\ldots|\upsilon_b^*\}$ is a partition of p indices into b non-empty blocks. Rather conveniently, every partition that appears on the right in (3.2) has coefficient $+1$ and, in fact, the general expression may be written

$$\kappa(\Upsilon^*) = \sum_{\Upsilon\vee\Upsilon^*=1} \kappa(\upsilon_1)\cdots\kappa(\upsilon_\nu), \tag{3.3}$$

where the sum is over all $\Upsilon = \{\upsilon_1|\ldots|\upsilon_\nu\}$ such that Υ and Υ^* are not both sub-partitions of any partition other than the full set, $\Upsilon_1 = \{(1, 2, \ldots, p)\}$ containing one block. Partitions satisfying this condition are said to be *complementary* to Υ^*. The notation and terminology used here are borrowed from lattice theory where $\Upsilon\vee\Upsilon^*$, also equal to $\Upsilon^*\vee\Upsilon$, is the *least upper bound* of Υ and Υ^*. A proof of this result is given in Section 3.6, using properties of the lattice of set partitions. In practice, the following graph-theoretical description of the condition $\Upsilon\vee\Upsilon^* = 1$ seems preferable because it is easier to visualize.

Any partition of p items or indices, say $\Upsilon^* = \{\upsilon_1^*|\ldots|\upsilon_b^*\}$, can be represented as a graph on p vertices. The edges of the graph consist of all pairs (i, j) that are in the same block of Υ^*. Thus the graph of Υ^* is the union of b disconnected complete graphs and we use the notation Υ^* interchangeably for the graph and for the partition. Since Υ and Υ^* are two graphs sharing the same vertices, we may define the edge sum graph $\Upsilon\oplus\Upsilon^*$, whose edges are the union of the edges of Υ and Υ^*. The condition that $\Upsilon\oplus\Upsilon^*$ be connected is identical to the condition $\Upsilon\vee\Upsilon^* = 1$

in (3.3). For this reason, we use the terms *connecting partition* and *complementary partition* interchangeably. In fact, this graph-theoretical device provides a simple way of determining the least upper bound of two or more partitions: the blocks of $\Upsilon \vee \Upsilon^*$ are just the connected components of the graph $\Upsilon \oplus \Upsilon^*$. The connections in this case need not be direct. In other words, the blocks of $\Upsilon \vee \Upsilon^*$ do not, in general, correspond to cliques of $\Upsilon \oplus \Upsilon^*$.

Consider, for example, the (4,6) cumulant $\kappa^{12,34,5,6}$ with $\Upsilon^* = 12|34|5|6$. Each block of the partition $\Upsilon = 13|24|56$ joins two blocks of Υ^*, but Υ and Υ^* are both sub-partitions of $1234|56$, so condition (3.3) is not satisfied. In the graph-theoretical representation, the first four vertices of $\Upsilon \oplus \Upsilon^*$ are connected and also the last two, but the vertices fall into two disconnected sets. By contrast, $\Upsilon = 1|23|456$ satisfies the required condition $\Upsilon \vee \Upsilon^* = 1$, giving rise to the contribution $\kappa^1 \kappa^{2,3} \kappa^{4,5,6}$.

Expression (3.3) gives generalized cumulants in terms of ordinary cumulants. The analogous expression for generalized cumulants in terms of ordinary moments is

$$\kappa(\Upsilon^*) = \sum_{\Upsilon \geq \Upsilon^*} (-1)^{\nu-1}(\nu-1)!\, \mu(\upsilon_1) \cdots \mu(\upsilon_\nu) \tag{3.4}$$

where $\mu(\upsilon_j)$ is an ordinary moment, and ν is the number of blocks of Υ. The sum extends over all partitions Υ such that Υ^* is a sub-partition of Υ. This expression follows from the development in Section 2.4 and can be regarded as effectively equivalent to (2.9). See also Section 3.6.

It is not difficult to see that (2.8) and (2.9) are special cases of (3.3) and (3.4). If we take $\Upsilon^* = \Upsilon_1$, the unpartitioned set, then every partition Υ satisfies the condition required in (3.3), giving (2.8). On the other hand, if we take $\Upsilon^* = \Upsilon_p$, the fully partitioned set, then Υ^* is a sub-partition of every partition. Thus every partition contributes to (3.4) in this case, giving (2.9).

3.3 Cumulants of homogeneous polynomials

For definiteness, consider two homogeneous polynomials

$$P_2 = a_{ij} X^i X^j \quad \text{and} \quad P_3 = a_{ijk} X^i X^j X^k$$

of degree 2 and 3 respectively. In many ways, it is best to think of P_2 and P_3, not as quadratic and cubic forms, but as linear forms in pairs of variables and triples of variables respectively. From this point of view, we can see immediately that

$$\begin{aligned}
E(P_2) &= a_{ij}\kappa^{ij} = a_{ij}\{\kappa^{i,j} + \kappa^i\kappa^j\} \\
E(P_3) &= a_{ijk}\kappa^{ijk} = a_{ijk}\{\kappa^{i,j,k} + \kappa^i\kappa^{j,k}[3] + \kappa^i\kappa^j\kappa^k\} \\
\operatorname{var}(P_2) &= a_{ij}a_{kl}\kappa^{ij,kl} \\
&= a_{ij}a_{kl}\{\kappa^{i,j,k,l} + \kappa^i\kappa^{j,k,l}[2] + \kappa^k\kappa^{i,j,l}[2] \\
&\qquad + \kappa^{i,k}\kappa^{j,l}[2] + \kappa^i\kappa^k\kappa^{j,l}[4]\} \\
&= a_{ij}a_{kl}\{\kappa^{i,j,k,l} + 4\kappa^i\kappa^{j,k,l} + 2\kappa^{i,k}\kappa^{j,l} + 4\kappa^i\kappa^k\kappa^{j,l}\}
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(P_2, P_3) &= a_{ij}a_{klm}\kappa^{ij,klm} \\
= a_{ij}a_{klm}\{&\kappa^{i,j,k,l,m} + \kappa^i\kappa^{j,k,l,m}[2] + \kappa^k\kappa^{i,j,l,m}[3] \\
&+ \kappa^{i,k}\kappa^{j,l,m}[6] + \kappa^{k,l}\kappa^{i,j,m}[3] + \kappa^i\kappa^k\kappa^{j,l,m}[6] + \kappa^k\kappa^l\kappa^{i,j,m}[3] \\
&+ \kappa^{i,k}\kappa^{j,l}\kappa^m[6] + \kappa^{i,k}\kappa^{l,m}\kappa^j[6] + \kappa^{i,k}\kappa^j\kappa^l\kappa^m[6]\}
\end{aligned}$$

For the final expression above where $\Upsilon^* = ij|klm$, the list of 42 complementary partitions can be found in Table 1 of the Appendix, where Υ^* is coded numerically as $123|45$. Since the arrays of coefficients a_{ij} and a_{ijk} are symmetrical, the permutation factors in [.] can be changed into ordinary arithmetic factors. Thus the 42 complementary partitions contribute only 10 distinct terms. These cannot be condensed further except in special cases. In the expression for $\mathrm{var}(P_2)$, on the other hand, the two classes of partitions $i|jkl[2]$ and $k|ijl[2]$ make equal contributions and further condensation is then possible as shown above.

In specific applications, it is often the case that the arrays a_{ij} and a_{ijk} have some special structure that can be exploited in order to simplify expressions such as those listed above. Alternatively, it may be that the cumulants have simple structure characteristic of independence, exchangeability, or identical distributions. In such cases, the joint cumulants listed above can be condensed further using power notation.

By way of illustration, we suppose that a_{ij} is a residual projection matrix of rank r, most commonly written as $\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ in the notation of linear models where $\mathbf{X}$ is a model matrix of known constants. We suppose in addition that the random variables $Y^i - \kappa^i$ are independent and identically distributed so that $\kappa^{i,j} = \kappa_2\delta^{ij}$, and the fourth-order cumulants are $\kappa_4\delta^{ijkl}$. On the assumption that $E(Y)$ does indeed lie in the column space of $\mathbf{X}$, it follows that $a_{ij}\kappa^j = 0$, and hence

$$\begin{aligned} E(P_2) &= a_{ij}\kappa^{i,j} = \kappa_2 \textstyle\sum a_{ii} = r\kappa_2 \\ \mathrm{var}(P_2) &= a_{ij}a_{kl}\{\kappa^{i,j,k,l} + 2\kappa^{i,k}\kappa^{j,l}\} \\ &= \kappa_4 \textstyle\sum a_{ii}^2 + 2r\kappa_2^2 \end{aligned}$$

In this example, P_2 is just the residual sum of squares on r degrees of freedom after linear regression on $\mathbf{X}$, where the theoretical errors are independent and identically distributed but not necessarily normal.

3.4 Polynomial transformations

It is not difficult now to develop the transformation law for cumulants under arbitrary non-linear transformation to new variables Y. The formulae developed in Section 3.2 refer to the particular polynomial transformation

$$Y^1 = \prod_{j\in v_1^*} X^j, \quad Y^2 = \prod_{j\in v_2^*} X^j, \quad \ldots, \quad Y^b = \prod_{j\in v_b^*} X^j.$$

To the extent that any continuous function can be approximated with arbitrary accuracy by means of a polynomial, there is little loss of generality in considering polynomial transformations

$$Y^r = a^r + a_i^r X^i + a_{ij}^r X^i X^j + a_{ijk}^r X^i X^j X^k + \cdots. \tag{3.5}$$

It is necessary at this stage to insist that the infinite expansion (3.5) be convergent, in principle for all X. In practice, in asymptotic calculations, the expansion must be convergent for all X for which the probability is appreciable.

To state the transformation law of cumulants in a concise way, it is helpful to abbreviate (3.5) by using 'matrix' notation as follows:

$$Y^r = (A_0^r + A_1^r + A_2^r + A_3^r + \cdots)X \tag{3.6}$$

where, for example, $A_2^r X$ is understood to represent a vector whose components are quadratic or bilinear forms in X. We abbreviate (3.6) further by introducing the operators $P^r = A_0^r + A_1^r + A_2^r + \cdots$ and writing

$$Y^r = P^r X. \tag{3.7}$$

The cumulant generating function of $Y = Y^1, \ldots, Y^q$ may now be written purely formally as

$$K_Y(\xi) = \exp(\xi_r P^r)\kappa_X \tag{3.8}$$

where $\exp(\xi_r P^r)$ is an operator acting on the cumulants of X as follows.

$$\begin{aligned} K_Y(\xi) = \{1 + \xi_r P^r + \xi_r\xi_s P^r P^s/2! \\ + \xi_r\xi_s\xi_t P^r P^s P^t/3! + \cdots\}\kappa_X. \end{aligned} \tag{3.9}$$

We define $1\kappa_X = 0$ and

$$P^r\kappa_X = a^r + a^r_i\kappa^i + a^r_{ij}\kappa^{ij} + a^r_{ijk}\kappa^{ijk} + \cdots .$$

Compound operators $P^r P^s$ and $P^r P^s P^t$ acting on κ_X produce generalized cumulants of order $b = 2$ and $b = 3$ respectively. For example, $A^r_1 A^s_2\kappa_X = a^r_i a^s_{jk}\kappa^{i,jk}$, which is not to be confused with $A^r_2 A^s_1\kappa_X = a^r_{ij} a^s_k\kappa^{ij,k}$. In like manner, third-order compound operators produce terms such as

$$\begin{aligned} A^r_1 A^s_1 A^t_1\kappa_X &= a^r_i a^s_j a^t_k\kappa^{i,j,k} \\ A^r_1 A^s_2 A^t_1\kappa_X &= a^r_i a^s_{jk} a^t_l\kappa^{i,jk,l}. \end{aligned}$$

Compound operators involving A_0 produce terms such as

$$\begin{aligned} A^r_0 A^s_1\kappa_X &= a^r a^s_i\kappa^{,i} = 0 \\ A^r_1 A^s_0 A^t_2\kappa_X &= a^r_i a^s a^t_{kl}\kappa^{i,,kl} = 0 \end{aligned}$$

and these are zero because they are mixed cumulants involving, in effect, one variable that is a constant.

Expansion (3.9) for the first four cumulants gives

$$\begin{aligned} K_Y(\xi) &= \xi_r\{a^r + a^r_i\kappa^i + a^r_{ij}\kappa^{ij} + a^r_{ijk}\kappa^{ijk} + \cdots\} \\ &+ \xi_r\xi_s\{a^r_i a^s_j\kappa^{i,j} + a^r_i a^s_{jk}\kappa^{i,jk}[2] + a^r_{ij} a^s_{kl}\kappa^{ij,kl} + \cdots\}/2! \\ &+ \xi_r\xi_s\xi_t\{a^r_i a^s_j a^t_k\kappa^{i,j,k} + a^r_i a^s_j a^t_{kl}\kappa^{i,j,kl}[3] + a^r_i a^s_{jk} a^t_{lm}\kappa^{i,jk,lm}[3] + \cdots\}/3! \\ &+ \xi_r\xi_s\xi_t\xi_u\{a^r_i a^s_j a^t_k a^u_l\kappa^{i,j,k,l} + a^r_i a^s_j a^t_k a^u_{lm}\kappa^{i,j,k,lm}[4] + \cdots\}/4! \\ &+ \cdots \end{aligned} \tag{3.10}$$

The leading terms in the above expansions are the same as (2.11), the law governing affine transformation.

The proof of (3.8) is entirely elementary and follows from the definition of generalized cumulants, together with the results of Section 2.4. A similar, though less useful formal expression can be developed for the moment generating function $M_X(\xi)$, which may be written

$$M_Y(\xi) = \exp(\xi_r P^r) * \kappa_X \tag{3.11}$$

where $1 * \kappa_X = 1$, $P^r * \kappa_X = P^r\kappa_X$ as before, and commas are omitted in the application of compound operators giving

$$\begin{aligned} A^r_0 A^s_1 * \kappa_X &= a^r a^s_i\kappa^i \\ A^r_1 A^s_1 * \kappa_X &= a^r_i a^s_j\kappa^{ij} \\ A^r_1 A^s_0 A^t_2 * \kappa_X &= a^r_i a^s a^t_{jk}\kappa^{ijk} \end{aligned}$$

and so on. Once again, the proof follows directly from the definition of moments.

In practice, in order to make use of (3.10), it is usually necessary to re-express all generalized cumulants in terms of ordinary cumulants. This exercise involves numerous applications of (3.3) and the formulae become considerably longer as a result.

3.5 Classifying complementary partitions

In order to use the fundamental identity (3.3) for the generalized cumulant $\kappa(\Upsilon^*)$ we need to list all partitions complementary to Υ^*. If the number of elements of Υ^* is more than, say six to eight, the number of complementary partitions can be very large indeed. If the listing is done by hand, it is difficult to be sure that no complementary partitions have been omitted. On the other hand, programming a computer to produce the required list is not an easy task. Further, it is helpful to group the partitions into a small number of equivalence classes in such a way that all members of a given class make the same contribution to the terms that occur in (3.10).

Suppose, for example, that we require the covariance of the two scalars

$$a_{ij}X^iX^j \quad \text{and} \quad b_{ij}X^iX^j$$

both of which are quadratic in X. We find

$$\mathrm{cov}\{a_{ij}X^iX^j\,,\; b_{ij}X^iX^j\} = a_{ij}b_{kl}\kappa^{ij,kl}.$$

To simplify the following expressions, we take $\kappa^i = 0$, giving

$$a_{ij}b_{kl}\{\kappa^{i,j,k,l} + \kappa^{i,k}\kappa^{j,l} + \kappa^{i,l}\kappa^{j,k}\}.$$

Since, by assumption, $a_{ij} = a_{ji}$ and $b_{ij} = b_{ji}$, this expression reduces to

$$a_{ij}b_{kl}\{\kappa^{i,j,k,l} + 2\kappa^{i,k}\kappa^{j,l}\}.$$

Thus the two partitions $ik|jl$ and $il|jk$ make identical contributions to the covariance and therefore they belong to the same equivalence class. This classification explains the grouping of terms in (3.2).

The classification of complementary partitions is best described in terms of the intersection matrix $M = \Upsilon^* \cap \Upsilon$, where m_{ij} is the number of elements in $\upsilon_i^* \cap \upsilon_j$. Since the order of the blocks is immaterial, this matrix is defined only up to independent row and column permutations. Two partitions, Υ_1 and Υ_2, whose intersection matrices are M_1 and M_2, are regarded as equivalent if $M_1 = M_2$ after suitably permuting the blocks of Υ_1 and Υ_2 or the columns of M_1 and M_2. It is essential in this comparison that the ith rows of M_1 and M_2 refer to the same block of Υ^*.

To take a simple example, suppose that $\Upsilon^* = 12|34|5$, $\Upsilon_1 = 135|24$, $\Upsilon_2 = 123|45$ and $\Upsilon_3 = 134|25$ with intersection matrices

$$M_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}; \qquad M_2 = \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}; \qquad M_3 = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

These matrices are all distinct. However, the partitions 145|23, 235|14 and 245|13 have M_1 as intersection matrix, the partition 124|35 has M_2 as intersection matrix and 234|15 has M_3 as intersection matrix. Thus these eight complementary partitions are written as

$$135|24[4] \;\cup\; 123|45[2] \;\cup\; 134|25[2].$$

If Υ_1 and Υ_2 are equivalent partitions in the sense just described, they must have the same number of blocks and also identical block sizes. Further, when permuting columns, we need only consider blocks of equal size: by convention, the blocks are usually arranged in decreasing size.

Tables 1 and 2 in the Appendix give lists of complementary partitions classified according to the above scheme. A typical element of each equivalence class is given and the number of elements in that class follows in [.].

3.6 Elementary lattice theory

3.6.1 *Generalities*

A lattice is a finite partially ordered set $\mathcal{L}$ having the additional property that for every pair of elements $a, b \in \mathcal{L}$ there is defined a unique greatest lower bound $c = a \wedge b$ and a unique least upper bound $d = a \vee b$ where $c, d \in \mathcal{L}$. These additional properties should not be taken lightly and, in fact, some commonly occurring partially ordered sets are not lattices because not every pair has a unique least upper bound or greatest lower bound. One such example is described in Exercise 3.34.

Lattices of various types arise rather frequently in statistics and probability theory. For example, in statistics, the class of factorial models, which can be described using the operators $+$ and $*$ on factors A, B, $C, \ldots$, forms a lattice known as the *free distributive lattice*. Typical elements of this lattice are $a = A + B * C$ and $b = A{*}B + B{*}C + C{*}A$, each corresponding to a factorial model. In the literature on discrete data, these models are also called *hierarchical*, but this usage conflicts with standard terminology in the analysis of variance where hierarchical refers to the presence of several variance components. In this particular example, $a < b$ because a is a sub-model of b, and the partial order has a useful statistical interpretation.

In probability theory or in set theory where $A_1, A_2, \ldots$ are subsets of Ω, it is sometimes useful to consider the lattice with elements Ω, A_i, $A_i \cap A_j$, $A_i \cap A_j \cap A_k$ and so on. We say that $a < b$ if $a \subset b$; the lattice so formed is known as the *binary lattice*. The celebrated inclusion-exclusion principle for calculating $\mathrm{pr}(A_1 \cup A_2 \cup \cdots)$ is a particular instance of (3.12) below (Rota, 1964). Exercise 2.22 provides a third example relevant both to statistics and to probability theory.

In this book, however, we are concerned principally with the lattice of set partitions where $a < b$ if a is a sub-partition of b. The least upper bound, $a \vee b$ was described in Section 3.2 as the partition whose blocks are the connected vertices of the graph $a \oplus b$: the greatest lower bound $a \wedge b$ is the partition whose blocks are the non-empty intersections of the blocks of a and b. Figure 3.2 gives the Hasse diagrams of the partition lattices of sets of up to four items. Each partition in the ith row of one of these diagrams contains i blocks. Partitions in the same row are unrelated in the partial order: partitions in different rows may or may not be related in the partial order. Notice that 123|4 has three immediate descendants whereas 13|24 has only two.

Let $\mathcal{L}$ be an arbitrary lattice, and let $f(\cdot)$ be a real-valued function defined on $\mathcal{L}$. We define the new function, $F(.)$ on $\mathcal{L}$ by

$$F(a) = \sum_{b \le a} f(b), \tag{3.11a}$$

analogous to integrating over the interval $[0, a]$. The formal inverse operation analogous to differentiation may be written

$$f(b) = \sum_{c \le b} m(c, b) F(c) \tag{3.12}$$

where $m(a, b)$ is known as the Möbius function for the lattice. See Exercise 2.?? for a justification of this analogy with differential and integral calculus.

It may be helpful by way of comparison to consider the matrix equation $F = Lf$, where f and F are vectors, and L_{ab} is the indicator function for $b \le a$. The inverse relation is $f = MF$ where the Möbius matrix $M = L^{-1}$ is also a lower triangular matrix. Note also that M^T is the matrix inverse of L^T. Consequently, if we define $\bar{F}(a) = \sum_{b \ge a} f(b)$, then the inverse relation involves the transposed Möbius matrix $f(b) = \sum_{c \ge b} m(b, c) \bar{F}(c)$.

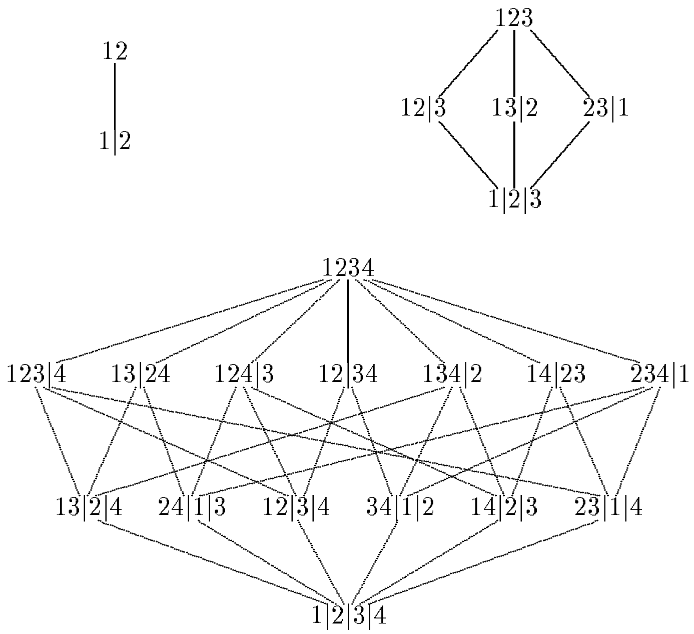

Figure 3.1: *Hasse diagrams for smaller partition lattices.*

Substitution of (3.12) into (3.11a) or (3.11a) into (3.12) gives

$$\begin{aligned} F(a) &= \sum_{b\le a}\sum_{c\le b} m(c,b)F(c) & \qquad f(b) &= \sum_{a\le b}\sum_{c\le a} m(a,b)f(c) \\ &= \sum_{c}\sum_{c\le b\le a} m(c,b)F(c) & &= \sum_{c}\sum_{c\le a\le b} m(a,b)f(c) \\ &= \sum_{c} \delta_{ac}F(c). & &= \sum_{c} \delta_{bc}f(c). \end{aligned}$$

The relation is required to hold for all functions F or f, which justifies the final line above. It follows, for fixed $a \le c$, that the sums over the lattice interval $[a, c]$,

$$\sum_{a\le b\le c} m(a,b) = \sum_{a\le b\le c} m(b,c) = \delta(a,c)$$

are zero unless $a = c$. In this sense, the Möbius function can be thought of as a series of contrasts or difference operators on all lattice intervals. The Möbius function may therefore be obtained recursively by $m(a, a) = 1$ followed by either

$$m(a,c) = -\sum_{a\le b<c} m(a,b), \quad \text{or} \quad m(a,c) = -\sum_{a<b\le c} m(b,c),$$

for $a < c$, and $m(a, c) = 0$ otherwise.

3.6.2 *Möbius function for the partition lattice*

We now show by induction that the Möbius function satisfies

$$m(a,1) = (-1)^{|a|-1}(|a|-1)! \tag{3.13}$$

where $|a|$ is the number of blocks of the partition. We use the property established above, that

$$m(a,1) = -\sum_{a<b\leq 1} m(b,1).$$

The Stirling number of the second kind, S_n^m, is the number of ways of partitioning a set of n elements into m non-empty subsets. A new element can be placed into one of the existing blocks or into a new block. Consequently, S_n^m satisfies the recurrence relation

$$S_{n+1}^m = mS_n^m + S_n^{m-1}.$$

Let a be a partition having $n+1$ blocks, and suppose that (3.13) is satisfied for all partitions having up to n blocks. Application of the formula for the Möbius function gives

$$-m(a,1) = \sum_{m=1}^{n} S_{n+1}^m (-1)^{m-1}(m-1)!.$$

The recurrence relation for Stirling numbers gives

$$\begin{aligned}-m(a,1) &= \sum_{m=1}^{n} S_n^m (-1)^{m-1} m! + \sum_{m=1}^{n} S_n^{m-1}(-1)^{m-1}(m-1)! \\ &= S_n^n(-1)^{n-1}n!.\end{aligned}$$

Since $S_n^n = 1$, the result (3.13) follows.

More generally, $m(\Upsilon, \Upsilon^*)$ is the product of terms like (3.13). Suppose that $\Upsilon < \Upsilon^*$ and that $\Upsilon^* = \{v_1^*|\ldots|v_\nu^*\}$ is a partition into ν blocks. In the partition Υ, each block of Υ^* is partitioned into smaller subsets, v_j^* being partitioned into b_j blocks in the finer partition. Then

$$m(\Upsilon, \Upsilon^*) = \prod_j (-1)^{b_j - 1}(b_j - 1)!.$$

For example, $m(1|2|3|4, 12|34)$ is equal to the product of $m(1|2, 12)$ and m(3—4, 34), $whichis(-1)^2$ while $m(1|2|34|56, 1234|56)$ is equal to $m(1|2|34, 1234)m(56, 56)$. The pairs 34 and 56 may be regarded as single indexes, so the Möbius function is equal to 2×1.

3.6.3 *Inclusion-exclusion and the binary lattice*

Let $A_1, \ldots, A_k$ be subsets of the sample space Ω. The binary lattice, $\mathcal{B}$, on these k generators, is the set whose 2^k elements are Ω, the sets A_i, the pairwise intersections $A_i \cap A_j$ with $(i < j)$, the triple intersections, and so on up to $A_1 \cap \cdots \cap A_k$. The partial order is defined by set inclusion, $a < b$ if a is a proper subset of b. The Hasse diagram of this lattice is illustrated for $k = 3$ in the first diagram in Fig. 3.2.

It is convenient to define, in association with $\mathcal{B}$, a new isomorphic lattice $\mathcal{B}'$ whose elements are the 2^k non-overlapping sets

$$\{A_1, \bar{A}_1\} \cap \{A_2, \bar{A}_2\} \cap \cdots \cap \{A_k, \bar{A}_k\}.$$

Elements in $\mathcal{B}'$ are associated with elements of $\mathcal{B}$ by deleting all the $\bar{A}$s, or equivalently by replacing all the $\bar{A}$s by Ω. The new lattice inherits its partial order from $\mathcal{B}$, so it is isomorphic with $\mathcal{B}$. Figure 3.1 illustrates the two lattices for $k = 3$.

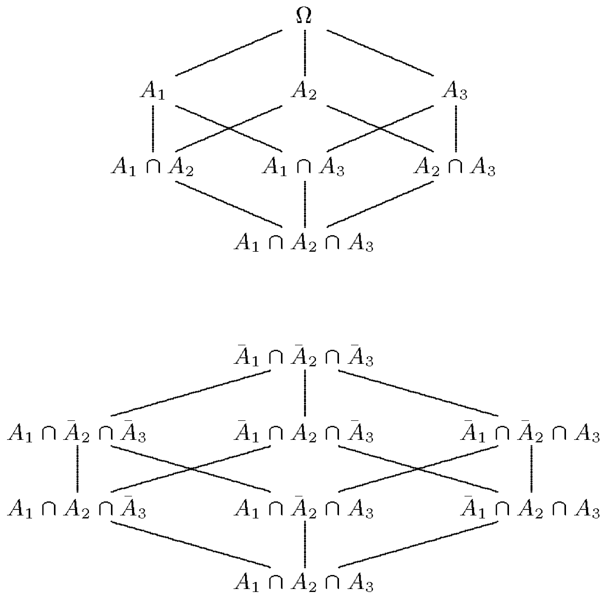
Ω
A_1
A_2
A_3
$A_1 \cap A_2$
$A_1 \cap A_3$
$A_2 \cap A_3$
$A_1 \cap A_2 \cap A_3$
$\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3$
$A_1 \cap \bar{A}_2 \cap \bar{A}_3$
$\bar{A}_1 \cap A_2 \cap \bar{A}_3$
$\bar{A}_1 \cap \bar{A}_2 \cap A_3$
$A_1 \cap A_2 \cap \bar{A}_3$
$A_1 \cap \bar{A}_2 \cap A_3$
$\bar{A}_1 \cap A_2 \cap A_3$
$A_1 \cap A_2 \cap A_3$

Figure 3.2: *Hasse diagram for the binary lattice on three generators (top), with the associated lattice of non-overlapping sets (bottom).*

The important relation between $\mathcal{B}$ and $\mathcal{B}'$ is the following. Let b and b' be corresponding elements in the two lattices. Then

$$b = \bigcup_{a' \leq b'} a',$$

primes being used to denote elements of $\mathcal{B}'$. In other words, b is the union of b' and its descendants in the associated lattice. Since the elements of $\mathcal{B}'$ are disjoint, the probability of the union is the sum of the probabilities,

$$\mathrm{pr}(b) = \sum_{a' \leq b'} \mathrm{pr}(a'),$$

analogous to integration over the lattice. The inverse relation is

$$\mathrm{pr}(b') = \sum_{a \leq b} m(a, b)\, \mathrm{pr}(b).$$

As is shown in Exercise 3.??, the Möbius function for the binary lattice satisfies $m(a, \Omega) = (-1)^{|a|}$, where $|a|$ is the number of sets intersecting in a. In particular, taking $b' = \bar{A}_1 \cap \cdots \cap \bar{A}_k$ and $b = \Omega$, we obtain the celebrated inclusion-exclusion rule

$$\begin{aligned}
&\mathrm{pr}(A_1 \cup \cdots \cup A_k) = 1 - \mathrm{pr}(\bar{A}_1 \cap \cdots \cap \bar{A}_k) \\
&\quad = 1 - \textstyle\sum m(b, \Omega)\ \mathrm{pr}(b) \\
&\quad = 1 - \mathrm{pr}(\Omega) + \textstyle\sum \mathrm{pr}(A_j) - \sum \mathrm{pr}(A_i \cap A_j) + \cdots \\
&\quad = \textstyle\sum \mathrm{pr}(A_j) - \sum \mathrm{pr}(A_i \cap A_j) + \cdots + (-1)^{k-1} \mathrm{pr}(A_1 \cap \cdots \cap A_k).
\end{aligned}$$

3.6.4 *Cumulants and the partition lattice*

Let $\mathcal{L}$ be the partition lattice on p labels or indices. Each element of $\mathcal{L}$ is a partition $\Upsilon = \{\upsilon_1|\ldots|\upsilon_b\}$ into a number of non-empty sets called blocks. We now define the following functions on $\mathcal{L}$.

$$\begin{aligned} f(\Upsilon) &= \kappa(\upsilon_1)\cdots\kappa(\upsilon_b) && \text{cumulant product}\\ F(\Upsilon) &= \mu(\upsilon_1)\cdots\mu(\upsilon_b) && \text{moment product}\\ g(\Upsilon) &= \kappa(\Upsilon) && \text{generalized cumulant}\end{aligned}$$

It is helpful notationally to denote by 0 and 1 the minimal and maximal elements of $\mathcal{L}$. In particular, the element 0 has p blocks of size one each; the element 1 has a single block of size p.

We begin with the established result that any moment is expressible as a sum of cumulant products, summed over the relevant partition lattice. In the current notation,

$$g(1) = F(1) = \sum_{0\leq a\leq 1} f(a) = \sum_{0\leq\Upsilon\leq 1} \kappa(\upsilon_1)\cdots\kappa(\upsilon_b).$$

Let Υ^* be an arbitrary partition with blocks $\upsilon_1^*|\ldots|\upsilon_b^*$. Now the moment $\mu(\upsilon_1^*)$ is a sum of cumulant products, summed over the partitions of υ_1^*. On multiplying together b such expressions, one for each block of Υ^*, we obtain the following:

$$\begin{aligned} F(\Upsilon^*) &= \mu(\upsilon_1^*)\cdots\mu(\upsilon_b^*)\\ &= \sum_{\Upsilon\leq\Upsilon^*} \kappa(\upsilon_1)\cdots\kappa(\upsilon_\nu)\\ &= \sum_{\Upsilon\leq\Upsilon^*} f(\Upsilon) \end{aligned} \tag{3.14}$$

The lattice interval $[0,\Upsilon^*]$ is the direct product of complete sub-lattices $[0,\upsilon_1^*]\times\cdots\times[0,\upsilon_b^*]$, which explains the second line above. The relation between F and f justifies our choice of notation.

Application of the Möbius inversion formula gives

$$f(\Upsilon^*) = \sum_{\Upsilon\leq\Upsilon^*} m(\Upsilon,\Upsilon^*)\,F(\Upsilon). \tag{3.15}$$

In particular, if $\Upsilon^* = 1$, we may use the fact that $m(\Upsilon,1) = (-1)^{b-1}(b-1)!$, to obtain the expression for ordinary cumulants in terms of moment products:

$$f(1) = \sum_{\Upsilon}(-1)^{b-1}(b-1)!\,\mu(\upsilon_1)\cdots\mu(\upsilon_b). \tag{3.16}$$

Now suppose that each index in the preceding expression is in fact a compound index representing a product of variables, so that the expanded partition becomes Υ^*. Then $f(1)$ becomes the generalized cumulant $\kappa(\Upsilon^*)$ given by

$$\begin{aligned} \kappa(\Upsilon^*) &= \sum_{\Upsilon\geq\Upsilon^*} m(\Upsilon,1)\,\mu(\upsilon_1)\cdots\mu(\upsilon_b)\\ &= \sum_{\Upsilon\geq\Upsilon^*} m(\Upsilon,1)\,F(\Upsilon) \end{aligned}$$

Substitution of expression (3.14) for $F(\Upsilon)$ gives

$$\kappa(\Upsilon^*) = \sum_{\Upsilon\geq\Upsilon^*} m(\Upsilon,1)\sum_{\Pi\leq\Upsilon} f(\Pi).$$

Finally, reversal of the order of summation gives

$$\begin{aligned}\kappa(\Upsilon^*) = g(\Upsilon^*) &= \sum_{\Pi} \sum_{\Upsilon \geq \Upsilon^* \vee \Pi} m(\Upsilon, 1)\, f(\Pi) \\ &= \sum_{\Pi \vee \Upsilon^* = 1} \kappa(\pi_1) \cdots \kappa(\pi_\sigma) \qquad (3.17)\end{aligned}$$

where $\pi_1 | \ldots | \pi_\sigma$ are the blocks of Π. This completes the proof of the fundamental identity (3.3).

In matrix notation, for the special case of four variables, we may write this equation as shown in Table 3.1.

Table 3.1 *Generalized cumulants expressed in terms of ordinary cumulants*

κ^{rstu}		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$\kappa^{r,s,t,u}$
$\kappa^{rsu,t}$		1	1		1	1	1	1	1		1		1		1		$\kappa^{r,s,u}\kappa^{t}$
$\kappa^{rtu,s}$		1	1	1		1	1	1	1	1			1	1			$\kappa^{r,t,u}\kappa^{s}$
$\kappa^{stu,r}$		1	1	1	1		1	1	1	1	1	1					$\kappa^{s,t,u}\kappa^{r}$
$\kappa^{rs,tu}$		1	1	1	1	1		1	1		1	1	1	1			$\kappa^{r,s}\kappa^{t,u}$
$\kappa^{rt,su}$		1	1	1	1	1	1		1	1		1	1		1		$\kappa^{r,t}\kappa^{s,u}$
$\kappa^{ru,st}$	=	1	1	1	1	1	1	1		1	1			1	1		$\kappa^{r,u}\kappa^{s,t}$
$\kappa^{rs,t,u}$		1			1	1		1	1								$\kappa^{r,s}\kappa^{t}\kappa^{u}$
$\kappa^{rt,s,u}$		1		1		1	1		1								$\kappa^{r,t}\kappa^{s}\kappa^{u}$
$\kappa^{ru,s,t}$		1	1			1	1	1									$\kappa^{r,u}\kappa^{s}\kappa^{t}$
$\kappa^{st,r,u}$		1		1	1		1	1									$\kappa^{s,t}\kappa^{r}\kappa^{u}$
$\kappa^{su,r,t}$		1	1		1		1		1								$\kappa^{s,u}\kappa^{r}\kappa^{t}$
$\kappa^{tu,r,s}$		1	1	1				1	1								$\kappa^{t,u}\kappa^{r}\kappa^{s}$
$\kappa^{r,s,t,u}$		1															$\kappa^{r}\kappa^{s}\kappa^{t}\kappa^{u}$

3.6.5 *Further relationships among cumulants*

The three functions described at the beginning of the previous section are defined in general as follows:

$$F(b) = \sum_{a \leq b} f(a) \qquad\qquad f(a) = \sum_{b \leq a} m(b, a) F(b)$$

$$K(a) = \sum_{b \geq a} m(b, 1) F(b) \qquad\qquad K(a) = \sum_{a \vee b = 1} f(b)$$

Ignoring, for the moment, the statistical interpretation of these functions, the first three expressions are simply a matter of definition, and could be applied to any lattice. The final expression for K in terms of f is a consequence of these definitions, and also applies to any lattice.

In order to obtain an expression for f in terms of K, it is helpful to use matrix notation since the relationships are linear and invertible. In this notation, the various expressions may be written

$$F = Lf, \quad f = L^{-1}F = MF, \quad K = L^T W F = L^T W L f,$$

where $W = \operatorname{diag}\{m(a, 1)\}$. Matrix notation makes it obvious that the transformation from f to K involves a symmetric matrix. What is less obvious is that the elements of $L^T W L$ are all zero or one, and that $(L^T W L)_{ab} = I(a \vee b = 1)$. In any event, the inverse relation is $f = M W^{-1} M^T K = \Sigma K$, where where Σ is a symmetric matrix with elements

$$\sigma(b, c) = \sum_a \frac{m(a, b)\, m(a, c)}{m(a, 1)},$$

and summation may be restricted to $a \leq b \wedge c$.

Explicitly, if $p = 3$, we may write the five partitions in order as $a_1 = 123$, $a_2 = 1|23$, $a_3 = 2|13$, $a_4 = 3|12$ and $a_5 = 1|2|3$. The matrix giving generalized cumulants in terms of cumulant products in (3.3) is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where a unit entry in position (i, j) corresponds to the criterion $a_i \vee a_j = 1$. The matrix inverse corresponding to (3.18) above is

$$\frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 0 & 2 \\ 0 & -1 & 1 & 1 & -1 \\ 0 & 1 & -1 & 1 & -1 \\ 0 & 1 & 1 & -1 & -1 \\ 2 & -1 & -1 & -1 & 1 \end{pmatrix}$$

Application of (3.18), using the second row of the above matrix, gives

$$\kappa^i \kappa^{j,k} = \{\kappa^{j,ik} + \kappa^{k,ij} - \kappa^{i,jk} - \kappa^{i,j,k}\}/2$$

and this particular expression can be verified directly.

For $p = 4$, the matrix $\sigma(.,.)$ in (3.18) is shown in Table 3.2.

Table 3.2 *Matrix giving cumulant products in terms of generalized cumulants*

	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
	0	−1	−1	−1	−1	1	1	1	−1	−1	2	−1	2	2	−2
	0	−1	−1	−1	−1	1	1	1	−1	2	−1	2	−1	2	−2
	0	−1	−1	−1	−1	1	1	1	2	−1	−1	2	2	−1	−2
	0	−1	−1	−1	−1	1	1	1	2	2	2	−1	−1	−1	−2
	0	1	1	1	1	−1	−1	−1	−2	1	1	1	1	−2	−1
1	0	1	1	1	1	−1	−1	−1	1	−2	1	1	−2	1	−1
–	0	1	1	1	1	−1	−1	−1	1	1	−2	−2	1	1	−1
6	0	−1	−1	2	2	−2	1	1	2	−1	−1	−1	−1	−1	1
	0	−1	2	−1	2	1	−2	1	−1	2	−1	−1	−1	−1	1
	0	2	−1	−1	2	1	1	−2	−1	−1	2	−1	−1	−1	1
	0	−1	2	2	−1	1	1	−2	−1	−1	−1	2	−1	−1	1
	0	2	−1	2	−1	1	−2	1	−1	−1	−1	−1	2	−1	1
	0	2	2	−1	−1	−2	1	1	−1	−1	−1	−1	−1	2	1
	6	−2	−2	−2	−2	−1	−1	−1	1	1	1	1	1	1	−1

The most important of these identities, or at least the ones that occur most frequently in the remainder of this book, are (3.15) and its inverse and the expression involving connecting partitions for generalized cumulants in terms of cumulant products. The inverse of the latter expression seems not to arise often. It is given here in order to emphasize that the relationship is invertible.

Finally, we note from (3.15) that any polynomial in the moments, homogeneous in the sense that every term is of degree 1 in each of p variables, can be expressed as a similarly homogeneous polynomial in the cumulants, and vice-versa for cumulants in terms of moments. In addition, from (3.16) we see that every generalized cumulant of degree p is uniquely expressible as a homogeneous polynomial in the moments. Inversion of (3.16) shows that every homogeneous polynomial in the moments is expressible as a *linear* function of generalized cumulants, each of degree p. Similarly, every homogeneous polynomial in the cumulants is expressible as a linear function of generalized cumulants. It follows that any polynomial in the cumulants or in the moments, homogeneous or not, is expressible as a linear function of generalized cumulants. Furthermore, this linear representation is unique because the generalized cumulants are linearly independent (Section 3.8). In fact, we have previously made extensive use of this important property of generalized cumulants. Expansion (3.10) for the cumulant generating function of the polynomial (3.5) is a linear function of generalized cumulants.

3.7 Some examples involving linear models

The notation used in this section is adapted to conform to the conventions of linear models, where $y = y^1, \ldots, y^n$ is the observed value of the random vector $Y = Y^1, \ldots, Y^n$ whose cumulants are κ^i, $\kappa^{i,j}$, $\kappa^{i,j,k}$ and so on. The common case where the observations are independent is especially important but, for the moment, the cumulant arrays are taken as arbitrary with no special structure. Now, Y and y lie in R^n, but the usual linear model specifies that the mean vector with components κ^i lies in the p-dimensional subspace, S_p of R^n spanned by the vectors $x_1, \ldots, x_p$ with components x_r^i. Thus we may write $E(Y) = x_r\beta^r$, or using components,

$$E(Y^i) = \kappa^i = x_r^i\beta^r$$

where β is a p-dimensional vector of parameters to be estimated.

It is important at this stage, to ensure that the notation distinguish between, on the one hand, vectors such as y^i, κ^i, $\kappa^{i,j}$ in R^n and its tensor products, and on the other hand, vectors such as β^r, $\beta^{r,s}$ in S_p and its tensor products. To do so, we use the letters $i, j, k, \ldots$ to indicate components of vectors in R^n, and $r, s, t, \ldots$ to indicate components of vectors in S_p.

Assume now that $\kappa^{i,j}$ is a known matrix of full rank, and that its matrix inverse is $\kappa_{i,j}$. The cumulants of $Y_i = \kappa_{i,j}Y^j$ may be written with subscripts in the form κ_i, $\kappa_{i,j}$, $\kappa_{i,j,k}$, and so on, and y_i is a quantity that can be computed. The least squares estimate of β is a vector b with components b^r satisfying the orthogonality condition $\langle y - x_r b^r,\, x_s\rangle = 0$ for each s. In other words, the residual vector $y - x_r b^r$ is required to be orthogonal to each of the basis vectors of S_p. This condition evidently requires a metric tensor to determine orthogonality in R^n, the natural one in this case being the inverse covariance matrix $\kappa_{i,j}$. Thus, we arrive at the condition $x_s^i\kappa_{i,j}(y^j - x_r^j b^r) = 0$, or

$$(x_r^i x_s^j \kappa_{i,j})b^s = x_r^i\kappa_{i,j}Y^j = x_r^i Y_i. \tag{3.19}$$

Since the information matrix $x_r^i x_s^j\kappa_{i,j}$ arises rather frequently, we denote it by $\beta_{r,s}$ and its inverse by $\beta^{r,s}$. The reason for this unusual choice of notation is that $\beta^r, \beta^{r,s}$ and $\beta_{r,s}$ play exactly the same roles in S_p as κ^i, $\kappa^{i,j}$ and $\kappa_{i,j}$ in R^n. In the terminology used in differential geometry and metric spaces, $\beta_{r,s}$ are the components of the induced metric tensor on S_p. Thus (3.19) becomes

$$\beta_{r,s}b^s = b_r = x_r^i Y_i$$

so that b^r and b_r play exactly the same role in S_p as Y^i and Y_i in R^n.

The cumulants of b_r are

$$\begin{aligned} x_r^i\kappa_{i,j}\kappa^j &= x_r^i\kappa_{i,j}x_s^j\beta^s = \beta_{r,s}\beta^s = \beta_r, \\ x_r^i x_s^j\kappa_{i,j} &= \beta_{r,s}, \qquad x_r^i x_s^j x_t^k\kappa_{i,j,k} = \beta_{r,s,t}, \\ x_r^i x_s^j x_t^k x_u^l\kappa_{i,j,k,l} &= \beta_{r,s,t,u} \end{aligned}$$

and so on. The cumulants of the least squares estimate b^r are obtained by raising indices giving β^r, $\beta^{r,s}$,

$$\beta^{r,s,t} = \beta^{r,u}\beta^{s,v}\beta^{t,w}\beta_{u,v,w}$$

and so on.

When we use the term tensor in connection with y^i, κ^i, $\kappa^{i,j}$, $\kappa_{i,j}$ and so on, we allude to the possibility of a change of basis vectors in R^n. In most applications, the transformations that would normally be contemplated in this context are rather limited, but the general theory permits arbitrary linear transformation. In connection with vectors b^r, β^r, $\beta^{r,s}$ and so on, the term tensor refers to the possibility of reparameterization by choosing an alternative set of vectors spanning S_p. In this sense, x^i_r are the components of a tensor in $S^*_p \otimes R^n$, the space of linear transformations from $S_p to\ R^n$.

The residual sum of squares may be written as the difference between the total sum of squares and the regression sum of squares as follows.

$$S^2 = y^i y^j \kappa_{i,j} - b^r b^s \beta_{r,s} = y^i y^j (\kappa_{i,j} - \lambda_{i,j}),$$

where $\lambda^{i,j} = x^i_r x^j_s \beta^{r,s}$ is the covariance matrix of the fitted values, and $\lambda_{i,j} = \kappa_{i,k}\kappa_{j,l}\lambda^{k,l}$ is one choice of annihilating generalized inverse. Note also that

$$\lambda_{i,j} x^j_r = \kappa_{i,k}\lambda^{k,l}\kappa_{l,j} x^j_r = \kappa_{i,j} x^j_r$$

and

$$\kappa^{i,j}\lambda_{i,j} = p.$$

These identities are more easily demonstrated using matrix notation. It follows that the cumulants of S and b are given by

$$\begin{aligned} E(S^2) &= \kappa^{i,j}(\kappa_{i,j} - \lambda_{i,j}) = n - p \\ \operatorname{var}(S^2) &= (\kappa_{i,j} - \lambda_{i,j})(\kappa_{k,l} - \lambda_{k,l})\kappa^{ij,kl} \\ &= (\kappa_{i,j} - \lambda_{i,j})(\kappa_{k,l} - \lambda_{k,l})\{\kappa^{i,j,k,l} + \kappa^{i,k}\kappa^{j,l}[2]\} \\ &= 2(n-p) + \kappa^{i,j,k,l}(\kappa_{i,j} - \lambda_{i,j})(\kappa_{k,l} - \lambda_{k,l}) \\ \operatorname{cov}(b^r, S^2) &= \beta^{r,s} x^i_s \kappa_{i,j}(\kappa_{k,l} - \lambda_{k,l})\kappa^{j,kl}. \end{aligned}$$

In the important special case where the observations are independent, we may write, in an obvious notation,

$$\operatorname{var}(S^2) = 2(n-p) + \sum_i \rho^i_4 (1 - h_{ii})^2$$

where ρ^i_4 is the standardized fourth cumulant of Y^i, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$ is the projection matrix producing fitted values. Also,

$$\operatorname{cov}(\mathbf{b},\, S^2) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{C}_3$$

where $c^i_3 = \kappa^i_3(1 - h_{ii})/\kappa^i_2$. Alternatively, we may write

$$\operatorname{cov}(\hat\mu,\, S^2) = \mathbf{H}\mathbf{C}_3$$

where $\hat\mu = \mathbf{X}\mathbf{b}$ is the vector of fitted values.

3.8 Cumulant spaces

We examine here the possibility of constructing distributions in R^p having specified cumulants or moments up to a given finite order, n. Such constructions are simpler in the univariate case and the discussion is easier in terms of moments than in terms of cumulants. Thus, we consider the question of whether or not there exists a distribution on the real line having moments $\mu'_1, \mu'_2, \ldots, \mu'_n$, the higher-order moments being left unspecified. If such a distribution exists, it follows that

$$\int (a_0 + a_1 x + a_2 x^2 + \cdots)^2 \, dF(x) dx \geq 0 \tag{3.20}$$

for any polynomial, with equality only if the density is concentrated at the roots of the polynomial. The implication is that for each $r = 1, 2, \ldots, [n/2]$, the matrix of order $r + 1 \times r + 1$

$$M'_r \;=\; \begin{pmatrix} 1 & \mu'_1 & \mu'_2 & \cdots & \mu'_r \\ \mu'_1 & \mu'_2 & \mu'_3 & \cdots & \mu'_{r+1} \\ \mu'_2 & \mu'_3 & \mu'_4 & \cdots & \mu'_{r+2} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \mu'_r & \mu'_{r+1} & \mu'_{r+2} & \cdots & \mu'_{2r} \end{pmatrix}$$

must be non-negative definite. Equivalently, we may eliminate the first row and column and work with the reduced matrix whose (r, s) element is $\mu_{r,s} = \mathrm{cov}(X^r, X^s)$, where indices here denote powers. The first two inequalities in this sequence are $\kappa_2 \geq 0$ and $\rho_4 \geq \rho_3^2 - 2$.

The above sequence of conditions is necessary but not sufficient to establish that a distribution having the required moments exists. By way of illustration, if we take $\mu'_1 = \mu'_2 = 1$, it follows that $\kappa_2 = 0$ giving $f_X(x) = \delta(x - 1)$, where $\delta(\cdot)$ is Dirac's delta function giving unit probability mass to the origin. In other words, all higher-order moments are determined by these particular values of μ'_1 and μ'_2. Similarly, if $\mu'_1 = 0$, $\mu'_2 = 1$, $\mu'_3 = 0$, $\mu'_4 = 1$, giving $\rho_4 = -2$, we must have

$$f_X(x) = \tfrac{1}{2}\delta(x - 1) + \tfrac{1}{2}\delta(x + 1)$$

because M'_2 has a zero eigenvalue with eigenvector $(-1, 0, 1)$, implying that F must be concentrated at the roots of the polynomial $x^2 - 1$. Since $\mu'_1 = 0$, the weights must be $\frac{1}{2}$ at $x = \pm 1$. Again, all higher-order moments and cumulants are determined by this particular sequence of four moments. It follows that there is no distribution whose moments are $0, 1, 0, 1, 0, 2, \ldots$ even though the corresponding M'_3 is positive semi-definite.

More generally, if for some $k \geq 1$, the matrices $M'_1, \ldots, M'_{k-1}$ are positive definite and $|M'_k| = 0$, then F is concentrated on exactly k points. The k points are the roots of the polynomial whose coefficients are given by the eigenvector of M'_k whose eigenvalue is zero. The probability mass at each point is determined by the first $k - 1$ moments. Since the value of any function at k distinct points can be expressed as a linear combination of the first $k - 1$ polynomials at those points, it follows that

$$\mathrm{rank}(M'_r) = \min(k, r). \tag{3.21}$$

Equivalently, the above rank condition may be deduced from the fact that the integral (3.20) is a sum over k points. Thus M'_r is a convex combination of k matrices each of rank one. The positive definiteness condition together with (3.21) is sufficient to ensure the existence of $f(\cdot)$.

In the case of multivariate distributions, similar criteria may be used to determine whether any particular sequence of arrays is a sequence of moments from some distribution. The elements of the multivariate version of M'_r are arrays suitably arranged. For example, the $(2, 2)$ component of M'_r is a square array of second moments whereas the $(1, 3)$ component is the same array arranged as a row vector. The existence of a distribution is guaranteed if, for each $r = 1, 2, \ldots,$ M'_r is non-negative definite and the rank is maximal, namely

$$\operatorname{rank}(M'_r) = \binom{p+r}{r}.$$

See Exercise 3.31. In the case of rank degeneracy, additional conditions along the lines of (3.21) are required to ensure consistency.

The moment space, $\mathcal{M}_n$, is the set of vectors $(\mu'_1, \ldots, \mu'_n)$ having n components that can be realized as moments of some distribution. Since the moments of the degenerate distribution $\delta(x-t)$ are t^r, any vector of the form $(t, t^2, \ldots, t^n)$ must lie in $\mathcal{M}_n$. In addition, since the moments of a mixture of distributions are mixtures of the moments, $\mathcal{M}_n$ is convex. Consequently, all convex combinations of polynomial vectors, $(t, t^2, \ldots, t^n)$, lie in $\mathcal{M}_n$. Finally, $\mathcal{M}_n$ has dimension n because there are n linearly independent vectors of the above polynomial type. Hence, the moments are functionally independent in the sense that no non-trivial function of any finite set of moments is identically zero for all distributions. Discontinuous functions such as $1 - H(\kappa_2)$ that are identically zero on $\mathcal{M}_n$ are regarded as trivial. The argument just given applies also to the multivariate case where the basis vectors of $\mathcal{M}_n^{(p)}$ are $(t^i, t^j t^k, \ldots)$, superscripts now denoting components. It follows immediately that if F has a finite number of support points, say k, then $\operatorname{rank}(M'_r) \le k$ and that this limit is attained for sufficiently large r.

The cumulant space, $\mathcal{K}_n$, is defined in the obvious way as the set of all vectors that can be realized as cumulants of some distribution. The transformation from $\mathcal{M}_n$ to $\mathcal{K}_n$ is nonlinear and the convexity property of $\mathcal{M}_n$ is lost in the transformation. To see that $\mathcal{K}_n$ is not convex for $n \ge 4$, it is sufficient to observe that the set $\kappa_4 \ge -2\kappa_2^2$ is not convex in the (κ_2, κ_4) plane. However, it is true that if t_1 and t_2 are vectors in $\mathcal{K}_n$, then $t_1 + t_2$ is also in $\mathcal{K}_n$ (Exercise 3.29). As a consequence, if t is in $\mathcal{K}_n$ then λt is also in $\mathcal{K}_n$ for any positive integer $\lambda \ge 1$. This property invites one to suppose that λt lies in $\mathcal{K}_n$ for all $\lambda \ge 1$, not necessarily an integer, but Exercise 3.30 demonstrates that this is not so for $n \ge 6$. The claim is true by definition for infinitely divisible distributions and the counterexample is related to the fact that not all distributions are infinitely divisible. Finally, the transformation from $\mathcal{M}_n$ to $\mathcal{K}_n$ is one to one and continuous, implying that $\mathcal{K}_n$ has dimension n. As a consequence, there are no non-trivial functions of the cumulants that are zero for all distributions. A similar result with obvious modifications applies in the multivariate case.

Identity (3.3) shows that the generalized cumulants are functionally dependent. However, they are *linearly* independent as the following argument shows. Without loss of generality, we may consider an arbitrary linear combination of generalized cumulants, each of the same degree, with coefficients $c(\Upsilon)$. The linear combination, $\sum_{\Upsilon^*} c(\Upsilon^*)\kappa(\Upsilon^*)$ may be written as

$$\begin{aligned}
&\sum_{\Upsilon^*} c(\Upsilon^*) \sum_{\Upsilon \ge \Upsilon^*} (-1)^{\nu-1}(\nu-1)!\, \mu(\upsilon_1)\cdots\mu(\upsilon_\nu) \\
&= \sum_{\Upsilon} (-1)^{\nu-1}(\nu-1)!\, \mu(\upsilon_1)\cdots\mu(\upsilon_\nu) C(\Upsilon)
\end{aligned} \tag{3.22}$$

where $C(\Upsilon) = \sum_{\Upsilon^* \le \Upsilon} c(\Upsilon^*)$ is an invertible linear function of the original coefficients. The implication is that $\sum c(\Upsilon)\kappa(\Upsilon) = 0$ with $c \ne 0$ implies a syzygy in the moments. From the previous discussion, this is known to be impossible because the moments are functionally independent. A simple extension of this argument covers the case where the linear combination involves generalized cumulants of unequal degree.

3.9 Bibliographic notes

There is some difficulty in tracing the origins of the fundamental identity (3.3). Certainly, it is not stated in Fisher's (1929) paper on k-statistics but Fisher must have known the result in some form in order to derive his rules for determining the joint cumulants of k-statistics. In fact, Fisher's procedure was based on the manipulation of differential operators (Exercise 3.11) and involved an expression for $M_X(\xi)$ essentially the same as (3.11) above. His subsequent calculations for joint cumulants were specific to the k-statistics for which many of the partitions satisfying $\Upsilon \vee \Upsilon^* = 1$ vanish on account of the orthogonality of the k-statistics. Rather surprisingly, despite the large number of papers on k-statistics that appeared during the following decades, the first explicit references to the identity (3.3) did not appear until the papers by James (1958), Leonov & Shiryaev (1959) and James & Mayne (1962). The statement of the result in these papers is not in terms of lattices or graphs. James (1958) uses the term *dissectable* for intersection matrices that do not satisfy the condition in (3.3). He gives a series of rules for determining the moments or cumulants of any homogeneous polynomial symmetric function although his primary interest is in k-statistics. He notes that for k-statistics, only the pattern of non-zero elements of the intersection matrix is relevant, but that in general, the numerical values are required: see Chapter 4. Leonov & Shiryaev (1959) use the term *indecomposability* defined essentially as a connectivity condition on the matrix $\Upsilon^* \cap \Upsilon$: see Exercises 3.4 and 3.5.

Rota's (1964) paper is the source of the lattice-theory notation and terminology. The notation and the derivation of (3.3) are taken from McCullagh (1984b). Alternative derivations and alternative statements of the result can be found in Speed (1983). The earliest statement of the result in a form equivalent to (3.3), though in a different notation, appears to be in Malyshev (1980) under the colourful title 'vacuum cluster expansions'.

There is also a more specialized literature concerned with variances of products: see, for example, Barnett (1955), or Goodman (1960, 1962).

For a thorough discussion of moment spaces, the reader is referred to Karlin & Studden (1966).

3.10 Further results and exercises 3

3.1 Let $X^1, \ldots, X^n$ be independent and identically distributed. By expressing $\bar{X}$ as a linear form and the sample variance, s^2 as a quadratic form, show that $\mathrm{cov}(\bar{X}, s^2) = \kappa_3/n$. Hence show that $\mathrm{corr}(\bar{X}, s^2) \to \rho_3/(2+\rho_4)^{1/2}$ as $n \to \infty$. Show also that $\mathrm{cov}(\bar{X}^2, s^2) = \kappa_4/n^2 + 2\kappa_1\kappa_3/n$ and show that the limiting correlation is, in most cases, non-zero for non-normal variables.

3.2 Let $Y^1, \ldots, Y^n$ be independent and identically distributed random variables with zero mean, variance κ_2, and higher-order cumulants κ_3, κ_4 and so on. Consider the following statistics.

$$k_2 = \sum(Y^i - \bar{Y})^2/(n-1) \qquad l_2 = \sum(Y^i)^2/n$$

$$k_3 = n\sum(Y^i - \bar{Y})^3/((n-1)(n-2)) \qquad l_3 = \sum(Y^i)^3/n$$

Express k_2 and k_3 as homogeneous polynomials in the form $\phi_{ij}Y^iY^j$ and $\phi_{ijk}y^iY^jY^k$. Show that the coefficients ϕ take the values $1/n$, $-1/(n(n-1))$ or $2/(n(n-1)(n-2))$ depending on the number of distinct indices. Show that k_2 and l_2 are unbiased for κ_2, and that k_3 and l_3 are unbiased for κ_3. Find the variances of all four statistics. Discuss briefly the efficiency of each statistic under the assumption that κ_3 and all higher-order cumulants can be neglected.

3.3 Let M be the intersection matrix $\Upsilon^* \cap \Upsilon$. Columns j_1 and j_2 are said to *hook* if, for some i, $m_{ij_1} > 0$ and $m_{ij_2} > 0$. The set of columns is said to *communicate* if there exists a sequence $j_1, j_2, \ldots, j_\nu$ such that columns j_l and j_{l+1} hook. The matrix M is said to be *indecomposable* if its columns communicate. Show that M is indecomposable if and only if M^T is indecomposable.

3.4 Show, in the terminology of Exercise 3.4, that M is indecomposable if and only if $\Upsilon \vee \Upsilon^* = 1$. (Leonov & Shiryaev, 1959; Brillinger, 1975, Section 2.3.)

3.5 If Υ^* is a 2^k partition (a partition of $2k$ elements into k blocks of 2 elements each), show that the number of 2^k partitions Υ satisfying $\Upsilon \vee \Upsilon^* = 1$, is $2^{k-1}(k-1)!$.

3.6 If $X = X^1, \ldots, X^p$ are jointly normal with zero mean and covariance matrix $\kappa^{i,j}$ of full rank, show that the rth order cumulant of $Y = \kappa_{i,j} X^i X^j$ is $2^{r-1}p(r-1)!$. Hence show that the cumulant generating function of Y is $-\frac{1}{2}p\log(1-2\xi)$ and therefore that Y has the χ^2 distribution on p degrees of freedom.

3.7 Show that $\kappa^{r,s}\kappa^{t,u}\xi_{rt}\xi_{su} = \mathrm{tr}\{(\kappa\xi)^2\}$ where $\kappa\xi$ is the usual matrix product.

3.8 For any positive definite matrix $\mathbf{A}$, define the matrix $\mathbf{B} = \log(\mathbf{A})$ by

$$\mathbf{A} = \exp(\mathbf{B}) = \mathbf{I} + \mathbf{B} + \mathbf{B}^2/2! + \cdots + \mathbf{B}^r/r! + \cdots.$$

By inverting this series, or otherwise, show that

$$\log|\mathbf{A}| = \mathrm{tr}\log(\mathbf{A}),$$

where $|\mathbf{A}|$ is the determinant of $\mathbf{A}$.

3.9 If $X = X^1, \ldots, X^p$ are jointly normal with zero mean, show that the joint cumulant generating function, $\log E\exp(\xi_{ij}Y^{ij})$ of $Y^{ij} = X^i X^j$ is $K_Y(\xi) = -\frac{1}{2}\log|\mathbf{I} - 2\xi\kappa|$. Hence derive the cumulant generating function for the Wishart distribution.

3.10 Let X be a scalar random variable with moments μ_r and cumulants κ_r in the notation of Section 2.5. Show that the rth moment of the polynomial

$$Y = P(X) = a_0 + a_1 X + a_2 X^2 + \cdots$$

is given formally by $P(d)M_X(\xi)|_{\xi=0}$ where $d = d/d\xi$. Hence show that the moment generating function of Y is

$$M_Y(\zeta) = \exp\{\zeta P(d)\}M_X(\xi)|_{\xi=0}$$

in which the operator is supposed to be expanded in powers before attacking the operand (Fisher, 1929, Section 10).

3.11 Show that any polynomial expression in $X^1, \ldots, X^p$, say

$$\begin{aligned} Q_4 = a_0 + a_i X^i + a_{ij}X^i X^j/2! + a_{ijk}X^i X^j X^k/3! \\ + a_{ijkl}X^i X^j X^k X^l/4!, \end{aligned}$$

can be expressed as a homogeneous polynomial in $X^0, X^1, \ldots, X^p$

$$Q_4 = \sum_{ijkl=0}^{p} b_{ijkl}X^i X^j X^k X^l/4!\,,$$

where $X^0 = 1$. Give expressions for the coefficients b in terms of the as. Find $E(Q_4)$ and $\mathrm{var}(Q_4)$ and express the results in terms of the as.

3.12 Consider the first-order autoregressive process $Y_0 = \epsilon_0 = 0$, $Y_j = \beta Y_{j-1} + \epsilon_j$, $j = 1, \ldots, n$, where $|\beta| < 1$ and ϵ_j are independent $N(0,1)$ random variables. Show that the log likelihood function has first derivative $U = \partial l/\partial\beta = T_1 - \beta T_2$ where $T_1 = \Sigma Y_j Y_{j-1}$ and $T_2 = \Sigma Y_{j-1}^2$ with summation from 1 to n. By expressing U as a quadratic form in ϵ, show that the first three cumulants of U are $E(U) = 0$,

$$\begin{aligned}\kappa_2(U) &= \sum_{1\leq i<j\leq n} \beta^{2j-2i-2}\\ &= (1-\beta^2)^{-1}\{n - (1-\beta^{2n})/(1-\beta^2)\} = E(T_2)\\ \kappa_3(U) &= 6 \sum_{1\leq i<j<k\leq n} \beta^{2k-2i-3}\\ &= \frac{6\{n\beta(1-\beta^2) - 2\beta + n\beta^{2n-1}(1-\beta^2) + 2\beta^{2n+1}\}}{(1-\beta^2)^3}.\end{aligned}$$

3.13 Let $Y = Y^1, \ldots, Y^p$ have zero mean and covariance matrix $\kappa^{i,j}$. Show that the 'total variance', $\sigma^2 = E(Y^i Y^j \delta_{ij})$, is invariant under orthonormal transformation of Y. For any given direction, ϵ, define

$$\begin{aligned}\sigma_\epsilon^2 &= \mathrm{var}(\epsilon_i Y^i) = \epsilon_i \epsilon_j \kappa^{i,j}\\ \tau_\epsilon^2 &= \epsilon_i\epsilon_j\{\sigma^2\delta^{ij} - \kappa^{i,j}\} = \epsilon_i\epsilon_j I^{ij}.\end{aligned}$$

Give an interpretation of σ_ϵ^2 and τ_ϵ^2 as regression and residual variances respectively. Show also that, in mechanics, τ_ϵ^2 is the moment of inertia of a rigid body of unit mass about the axis ϵ. Hence, interpret I^{ij} as the *inertia tensor* (Synge & Griffith, 1949, Section 11.3; Jeffreys & Jeffreys, 1956, Section 3.08).

3.14 *MacCullagh's formula*: Using the notation of the previous exercise, let $X^i = Y^i + \kappa^i$, where κ^i are the components of a vector of length $\rho > 0$ in the direction ϵ. Show that

$$\begin{aligned}E\left(\frac{1}{|X|}\right) &= \frac{1}{\rho} + \frac{1}{2\rho^3}\left(-\sigma^2 + 3\sigma_\epsilon^2\right) + O(\rho^{-4})\\ &= \frac{1}{\rho} + \frac{1}{2\rho^3}\left(2\sigma^2 - 3\tau_\epsilon^2\right) + O(\rho^{-4}),\end{aligned}$$

(MacCullagh, 1855; Jeffreys & Jeffreys, 1956, Section 18.09). [The above expression gives the potential experienced at an external point (the origin) due to a unit mass or charge distributed as $f_X(x)$. The correction term is sometimes called the *gravitational quadrupole* (Kibble, 1985, Chapter 6).]

3.15 Show that the sum

$$\sum_{\Upsilon\geq\Upsilon^*} (-1)^{\nu-1}(\nu-1)! = \delta(\Upsilon^*, 1)$$

is zero unless $\Upsilon^* = 1$.

3.16 By reversing the order of summation in (3.17), show that the generalized cumulant may be written

$$\kappa(\Upsilon^*) = \sum_{\Pi} \kappa(\pi_1)\cdots\kappa(\pi_\sigma) \sum_{\substack{\Upsilon\geq\Upsilon^*\\ \Upsilon\geq\Pi}} (-1)^{\nu-1}(\nu-1)!.$$

Hence, using the result of the previous exercise, prove the fundamental identity (3.3).

3.17 Let X be a scalar random variable whose distribution is Poisson with mean 1. Show that the cumulants of X of all orders are equal to 1. Hence show that the rth moment is

$$\mu'_r = E(X^r) = B_r$$

where B_r, the rth Bell number, is the number of partitions of a set of r elements. Hence derive a generating function for the Bell numbers.

3.18 Let $\Upsilon^* = \{v_1^*, \ldots, v_b^*\}$ be a partition of a set of p elements into b non-empty blocks of sizes $|v_1^*|, \ldots, |v_b^*|$. By using (3.16) or otherwise, show that the number of partitions complementary to Υ^* is

$$\sum_{\substack{\Upsilon \geq \Upsilon^* \\ \Upsilon = \{v_1, \ldots, v_\nu\}}} (-1)^{\nu-1}(\nu-1)!\, B_{|v_1|} \cdots B_{|v_\nu|}.$$

3.19 Interpret the expression in Exercise 3.19 as the cumulant of order b

$$\mathrm{cum}\left(X^{|v_1^*|}, \ldots, X^{|v_b^*|}\right)$$

where X is defined in Exercise 3.18 and the superscripts denote powers. Simplify this result in the special case where Υ^* is a 2^k partition and compare with Exercise 3.6.

3.20 Show that the number of sub-partitions of Υ^* is given by

$$\sum_{\Upsilon \leq \Upsilon^*} 1 = B_{|v_1^*|} \cdots B_{|v_b^*|}.$$

3.21 Using the result given in Exercise 3.19, show that the total number of ordered pairs of partitions (Υ_1, Υ_2) satisfying $\Upsilon_1 \vee \Upsilon_2 = 1$ is

$$C_p^{(2)} = \sum_{\Upsilon} (-1)^{\nu-1}(\nu-1)!\, B^2_{|v_1|} \cdots B^2_{|v_\nu|}$$

where the partitions contain p elements and B_r^2 is the square of the rth Bell number. Deduce also that $C_p^{(2)}$ is the pth cumulant of $Y = X_1X_2$ where the Xs are independent Poisson random variables with unit mean.

3.22 Show that the number of ordered pairs (Υ_1, Υ_2) satisfying $\Upsilon_1 \vee \Upsilon_2 = \Upsilon^*$ for some fixed partition Υ^*, is

$$\sum_{\Upsilon \leq \Upsilon^*} m(\Upsilon, \Upsilon^*) B^2_{|v_1|} \cdots B^2_{|v_\nu|},$$

where $m(\Upsilon, \Upsilon^*)$ is the Möbius function for the partition lattice, defined below (3.12). Hence prove that the total number of ordered triplets $(\Upsilon_1, \Upsilon_2, \Upsilon_3)$ satisfying $\Upsilon_1 \vee \Upsilon_2 \vee \Upsilon_3 = 1$ is

$$C_p^{(3)} = \sum_{\Upsilon} (-1)^{\nu-1}(\nu-1)!\, B^3_{|v_1|} \cdots B^3_{|v_\nu|}.$$

Show also, in the notation of Exercise 3.22, that $C_p^{(3)}$ is the pth cumulant of the triple product $Y = X_1X_2X_3$.

3.23 Generalize the result of the previous exercise to ordered k-tuplets of partitions. Give a simple explanation for this result.

3.24 An alternative way of representing a partition by means of a graph is to use p labelled edges instead of labelled vertices. In this form, the graph of $\Upsilon^* = \{\upsilon_1, \ldots, \upsilon_\nu\}$ comprises p labelled edges emanating from b unlabelled vertices, giving a total of $b + p$ vertices of which p are 'free'. If $\Upsilon = \{\upsilon_1, \ldots, \upsilon_\nu\}$ is another partition of the same indices represented as a graph in the same way, we define the graph $\Upsilon \otimes \Upsilon^*$ by connecting corresponding free vertices of the two graphs. This gives a graph with $b + \nu$ unlabelled vertices and p labelled edges, parallel edges being permitted. Show that the graph $\Upsilon \otimes \Upsilon^*$ is connected if and only if $\Upsilon \vee \Upsilon^* = 1$.

3.25 In the notation of the previous exercise, prove that all cycles in the graph $\Upsilon \otimes \Upsilon^*$ have even length. (A cycle is a path beginning and ending at the same vertex.) Such a graph is said to be even. Show that all even connected graphs have a unique representation as $\Upsilon \otimes \Upsilon^*$. Hence prove that the number of connected even graphs having p labelled edges is $(C_p^{(2)} + 1)/2$ where $C_p^{(2)}$ is defined in Exercise 3.22 (Gilbert, 1956).

3.26 In the terminology of the previous two exercises, what does $C_p^{(3)}$ in Exercise 3.23 correspond to?

3.27 By considering the mixture density, $pf_1(x) + (1 - p)f_2(x)$, show that the moment space, $\mathcal{M}_n$, is convex.

3.28 By considering the distribution of the sum of two independent random variables, show that the cumulant space, $\mathcal{K}_n$ is closed under vector addition.

3.29 Show that there exists a unique distribution whose odd cumulants are zero and whose even cumulants are $\kappa_2 = 1$, $\kappa_4 = -2$, $\kappa_6 = 16$, $\kappa_8 = -272$, $\kappa_{10} = 7936, \ldots$. Let $M_r'(\lambda)$ be the moment matrix described in Section 3.8, corresponding to the cumulant sequence, $\lambda\kappa_1, \ . \ . \ . \ , \lambda\kappa_{2r}$. Show that for the particular cumulant sequence above, the determinant of $M_r'(\lambda)$ is

$$|M_r'| = 1!\,2! \ldots r!\,\lambda^r(\lambda - 1)^{r-1} \cdots (\lambda - r + 1),$$

for $r = 1, 2, 3, 4$. Hence prove that there is no distribution whose cumulants are $\{\lambda\kappa_r\}$ for non-integer $\lambda < 3$. Find the unique distribution whose cumulants are $\{\lambda\kappa_r\}$ for $\lambda = 1, 2, 3$.

3.30 By counting the number of distinguishable ways of placing r identical objects in $p + 1$ labelled boxes, show that, in p dimensions,

$$\operatorname{rank}(M_r') \le \binom{p + r}{r},$$

where M_r' is defined in Section 3.8.

3.31 Show that achievement of the limit $\bar{\rho}_4 = \bar{\rho}_{23}^2 - p - 1$ implies the following constraint on the third cumulants

$$\bar{\rho}_{23}^2 - \bar{\rho}_{13}^2 - p + 1 = 0.$$

3.32 Show that, if complex-valued random variables are permitted, there are no restrictions on the moment spaces or on the cumulant spaces such as those discussed in Section 3.8.

3.33 Describe the usual partial order on the set of partitions of the *number* k. Explain why this set, together with the usual partial order, does *not* form a lattice for $k \ge 5$. Verify by direct inspection that the structure is a lattice for each $k \le 4$.

CHAPTER 4

Sample cumulants

4.1 Introduction

We now address problems of a more inferential flavour, namely problems concerning point estimation or, better, interval estimation of population cumulants based on observed data. Moment and cumulant estimators have a lengthy history in the statistical literature going back to the work of K. Pearson, Chuprov, Student, Fisher and others in the early part of this century. In recent years, the volume of work on such topics has declined, partly because of concerns about robustness, sensitivity to outliers, misrecorded or miscoded values and so on. Estimates of the higher-order cumulants are sensitive to such errors in the data and, for some purposes, this sensitivity may be considered undesirable. In addition, the variance of such an estimate depends on cumulants up to twice the order of the estimand and these are even more difficult to estimate accurately. If there is insufficient data it may not be possible to estimate such cumulants at all. We are confronted immediately and forcibly with what Mosteller & Tukey (1977, Chapter 1) call 'the misty staircase'. That is to say that, to assess the variability of a primary statistic, we compute a secondary statistic, typically more variable than the primary statistic. It is then possible to estimate the variability of the secondary statistic by computing a third statistic and so ad infinitum. Fortunately, we usually stop long before this extreme stage.

Section 4.2 is concerned with simple random samples from an infinite population. The emphasis is on the various symmetric functions, i.e. functions of the data that are unaffected by re-ordering the data values. The functions having the most pleasing statistical properties are the k-statistics and, to a lesser extent, the generalized k-statistics. The k-statistics are unbiased estimates of ordinary cumulants and generalized k-statistics are estimates of generalized cumulants, including moments as a special case.

Section 4.3 is concerned with simple random samples from a finite population. The idea here is to construct statistics whose expectation under simple random sampling is just the value of the statistic computed in the whole population. The generalized k-statistics have this property but it turns out to be more convenient to work with linear combinations called 'polykays'. These are unbiased estimates of population polykays or, in the infinite population case, unbiased estimates of cumulant products. Again, they are most conveniently indexed by set partitions.

The remaining sections are concerned with estimates of cumulants and cumulant products in the presence of identifiable systematic structure, typically in the mean value. A common example is to estimate the second- and higher-order cumulants when the data are divided into k groups differing only in mean value. A second example is to estimate the cumulants of the error distribution based on residuals after linear regression. The difficulty here is that neither the raw data nor the observed residuals are identically distributed, so there is no compelling reason for restricting attention to symmetric functions in the usual sense. Indeed, better estimates can be obtained by using functions that are not symmetric: see Exercise 4.22.

4.2 k-statistics

4.2.1 *Definitions and notation*

Let $Y_1, \ldots, Y_n$ be independent and identically distributed p-dimensional random variables where Y_i has components $Y_i^1, \ldots, Y_i^p$. The cumulants and generalized cumulants of Y_i are written κ^r, $\kappa^{r,s}$, $\kappa^{r,s,t}$, $\kappa^{r,st}$ and so on. No subscripts are necessary because of the assumption that the observations are identically distributed. For each generalized cumulant, κ, with appropriate superscripts, there is a unique polynomial symmetric function, denoted by k with matching superscripts, such that k is an unbiased estimate of κ. The lower-order k-statistics are very familiar, though perhaps by different names. Thus, for example, the simplest k-statistic

$$k^r = n^{-1}\sum\nolimits_i Y_i^r = \bar{Y}^r \tag{4.1}$$

is just the sample mean, an unbiased estimate of κ^r. Also,

$$\begin{aligned} k^{r,s} &= \sum\nolimits_i (Y_i^r - \bar{Y}^r)(Y_i^s - \bar{Y}^s)/(n-1) \\ &= n^{-1}\phi^{ij} Y_i^r Y_j^s, \end{aligned} \tag{4.2}$$

where $\phi^{ii} = 1$ and $\phi^{ij} = -1/(n-1)$ for $i \neq j$, is just the usual sample covariance matrix. It is well known that $k^{r,s}$ is an unbiased estimate of $\kappa^{r,s}$.

We need not restrict attention to ordinary cumulants alone. A straightforward calculation shows that

$$k^{rs} = n^{-1}\sum\nolimits_i Y_i^r Y_i^s \tag{4.3}$$

is an unbiased estimate of κ^{rs} and that

$$k^{r,st} = n^{-1}\sum\nolimits_{ij} \phi^{ij} Y_i^r Y_j^s Y_j^t \tag{4.4}$$

is an unbiased estimate of $\kappa^{r,st}$.

The four statistics (4.1) to (4.4) are all examples of k-statistics. Following the terminology of Chapter 3, we refer to (4.1) and (4.2) as ordinary k-statistics and to (4.3) and (4.4) as generalized k-statistics. It is important at the outset to emphasize that while

$$\kappa^{rs} \equiv \kappa^{r,s} + \kappa^r \kappa^s,$$

the corresponding expression with κ replaced by k is false. In fact, we may deduce from (4.1) to (4.3) that

$$nk^{rs} = (n-1)k^{r,s} + nk^r k^s.$$

Equivalently, we may write

$$\begin{aligned} k^{rs} - k^{r,s} &= n^{-1}\sum\nolimits_{ij} (\delta^{ij} - \phi^{ij}) Y_i^r Y_j^s \\ &= \sum\nolimits^{\#} Y_i^r Y_j^s / n^{(2)}, \end{aligned}$$

where $n^{(2)} = n(n-1)$, which is an unbiased estimate of the product $\kappa^r\kappa^s$, and is not the same as $k^r k^s$. The symbol, $\sum^{\#}$, which occurs frequently in the calculations that follow, denotes summation over unequal values of the indices, $i, j, \ldots$.

4.2.2 *Some general formulae for k-statistics*

It follows from the definition of moments that

$$\begin{aligned} k^r &= n^{-1}\delta^i Y_i^r \\ k^{rs} &= n^{-1}\delta^{ij} Y_i^r Y_j^s \\ k^{rst} &= n^{-1}\delta^{ijk} Y_i^r Y_j^s Y_k^t \end{aligned}$$

and so on, where $\delta^{ijk} = 1$ if $i = j = k$ and zero otherwise, are unbiased estimates of the moments κ^r, κ^{rs}, κ^{rst} and so on. To construct unbiased estimates of the ordinary cumulants, we write

$$\begin{aligned} k^{r,s} &= n^{-1}\phi^{ij} Y_i^r Y_j^s \\ k^{r,s,t} &= n^{-1}\phi^{ijk} Y_i^r Y_j^s Y_k^t \\ k^{r,s,t,u} &= n^{-1}\phi^{ijkl} Y_i^r Y_j^s Y_k^t Y_l^u \end{aligned} \tag{4.5}$$

and so on, and aim to choose the coefficients ϕ to satisfy the criterion of unbiasedness. The k-statistics are required to be symmetric in two different senses. First, they are required to be symmetric functions in the sense that they are unaffected by permuting the n observations $Y_1, \ldots, Y_n$. As a consequence, for any permutation $\pi_1, \ldots, \pi_n$ of the first n integers, it follows that

$$\phi^{ijkl} = \phi^{\pi_i \pi_j \pi_k \pi_l}.$$

This means, for example, that $\phi^{iijj} = \phi^{1122}$ but it does not follow from the above criterion that ϕ^{1221} is the same as ϕ^{1122}. In fact, however, it turns out that the coefficients ϕ are symmetric under index permutation. This follows not from the requirement that the ks be symmetric functions, but from the requirement that the k-statistics, like the corresponding cumulants, be symmetric under index permutation. In Section 4.3.2, symmetric functions will be introduced for which the coefficients are not symmetric in this second sense. It follows that ϕ^{ijk} can take on at most three distinct values depending on whether $i = j = k$, $i = j \neq k$ or all three indices are distinct. Similarly, ϕ^{ijkl} can take on at most five distinct values, namely ϕ^{1111}, ϕ^{1112}, ϕ^{1122}, ϕ^{1123} and ϕ^{1234}, corresponding to the five partitions of the *number* 4.

On taking expectations in (4.5), we find that the following identities must be satisfied by the coefficients, ϕ.

$$\begin{aligned} \kappa^{r,s} &= n^{-1}\phi^{ij}\left(\kappa^{r,s}\delta_{ij} + \kappa^r\kappa^s\delta_i\delta_j\right) \\ \kappa^{r,s,t} &= n^{-1}\phi^{ijk}\left(\kappa^{r,s,t}\delta_{ijk} + \kappa^r\kappa^{s,t}\delta_i\delta_{jk}[3] + \kappa^r\kappa^s\kappa^t\delta_i\delta_j\delta_k\right) \\ \kappa^{r,s,t,u} &= n^{-1}\phi^{ijkl}\left(\kappa^{r,s,t,u}\delta_{ijkl} + \kappa^r\kappa^{s,t,u}\delta_i\delta_{jkl}[4] + \kappa^{r,s}\kappa^{t,u}\delta_{ij}\delta_{kl}[3]\right. \\ &\qquad \left. + \kappa^r\kappa^s\kappa^{t,u}\delta_i\delta_j\delta_{kl}[6] + \kappa^r\kappa^s\kappa^t\kappa^u\delta_i\delta_j\delta_k\delta_l\right). \end{aligned}$$

Thus we must have

$$\begin{array}{llll} \phi^{ij}\delta_{ij} = n, & \phi^{ij}\delta_i = 0, & & \\ \phi^{ijk}\delta_{ijk} = n, & \phi^{ijk}\delta_{ij} = 0, & \phi^{ijk}\delta_i = 0, & \\ \phi^{ijkl}\delta_{ijkl} = n, & \phi^{ijkl}\delta_{ijk} = 0, & \phi^{ijkl}\delta_{ij} = 0, & \phi^{ijkl}\delta_i = 0, \end{array} \tag{4.6}$$

and so on. From these formulae, we find that, for i, j, k, l all distinct,

$$\begin{aligned} \phi^{ii} &= \phi^{iii} = \phi^{iiii} = 1 \\ \phi^{ij} &= \phi^{iij} = \phi^{iiij} = \phi^{iijj} = -1/(n-1) \\ \phi^{ijk} &= \phi^{iijk} = 2/\{(n-1)(n-2)\} \\ \phi^{ijkl} &= -6/\{(n-1)(n-2)(n-3)\}. \end{aligned}$$

One can then show by induction that the general expression for the coefficients ϕ is

$$(-1)^{\nu-1}\Big/\binom{n-1}{\nu-1} = (-1)^{\nu-1}(\nu-1)!/(n-1)^{(\nu-1)} \tag{4.7}$$

where $\nu \le n$ is the number of distinct indices, and

$$(n-1)^{(\nu-1)} = (n-1)(n-2)\cdots(n-\nu+1).$$

There are no unbiased estimates for cumulants of order greater than n.

Many of the pleasant statistical properties of k-statistics stem from the orthogonality of the ϕ-arrays and the δ-arrays as shown in (4.6). More generally, if υ_1, υ_2 are sets of indices, we may write $\langle \phi(\upsilon_1), \delta(\upsilon_2)\rangle$ for the sum over those indices in $\upsilon_1 \cap \upsilon_2$. This notation gives

$$\langle \phi(\upsilon_1),\, \delta(\upsilon_2)\rangle = \begin{cases} 0, & \text{if } \upsilon_2 \subset \upsilon_1; \\ n, & \text{if } \upsilon_2 = \upsilon_1; \\ \delta(\upsilon_2 - \upsilon_1), & \text{if } \upsilon_1 \subset \upsilon_2, \end{cases} \tag{4.8}$$

no simplification being possible otherwise. In the above expression, the symbol $\subset$ is to be interpreted as meaning 'proper subset of'.

For an alternative derivation of expression (4.7) for ϕ, we may proceed as follows. Unbiased estimates of the moments are given by

$$k^{rs} = n^{-1}\sum Y_i^r Y_i^s, \quad k^{rst} = n^{-1}\sum Y_i^r Y_i^s Y_i^t$$

and so on. Unbiased estimates of products of moments are given by the so-called symmetric means

$$\begin{aligned} k^{(r)(s)} &= \textstyle\sum^{\#} Y_i^r Y_j^s/n^{(2)}, & k^{(r)(st)} &= \textstyle\sum^{\#} Y_i^r Y_j^s Y_j^t/n^{(2)}, \\ k^{(rs)(tu)} &= \textstyle\sum^{\#} Y_i^r Y_i^s Y_j^t Y_j^u/n^{(2)} & k^{(r)(stu)} &= \textstyle\sum^{\#} Y_i^r Y_j^s Y_j^t Y_j^u/n^{(2)} \\ k^{(r)(s)(t)} &= \textstyle\sum^{\#} Y_i^r Y_j^s Y_k^t/n^{(3)}, & k^{(r)(st)(u)} &= \textstyle\sum^{\#} Y_i^r Y_j^s Y_j^t Y_k^u/n^{(3)} \end{aligned}$$

and so on, with summation extending over unequal subscripts. It is a straightforward exercise to verify that $E(k^{(r)(st)(u)}) = \kappa^r\kappa^{st}\kappa^u$ and similarly for the remaining statistics listed above. All linear combinations of these statistics are unbiased for the corresponding parameter. Thus

$$k^{rst} - k^{(r)(st)}[3] + 2k^{(r)(s)(t)} \tag{4.9}$$

is an unbiased estimate of $\kappa^{rst} - \kappa^r\kappa^{st}[3] + 2\kappa^r\kappa^s\kappa^t = \kappa^{r,s,t}$. By expressing (4.9) as a symmetric cubic polynomial with coefficients ϕ^{ijk}, it can be seen that the coefficients must satisfy (4.7), the numerator coming from the Möbius function and the denominator from the above sums over unequal subscripts.

A similar argument applies to higher-order k-statistics.

4.2.3 *Joint cumulants of ordinary k-statistics*

Before giving a general expression for the joint cumulants of k-statistics, it is best to examine a few simple cases. Consider first the covariance of k^r and k^s, which may be written

$$\operatorname{cov}(k^r, k^s) = n^{-2}\phi^i\phi^j\kappa_{i,j}^{r,s}$$

where $\kappa_{i,j}^{r,s}$ is the covariance of Y_i^r and Y_j^s. This covariance is zero unless $i = j$, in which case we may write $\kappa_{i,j}^{r,s} = \kappa^{r,s}\delta_{ij}$ because the random variables are assumed to be identically distributed. Thus

$$\operatorname{cov}(k^r, k^s) = n^{-2}\phi^i\phi^j\delta_{ij}\kappa^{r,s} = \kappa^{r,s}/n.$$

Similarly, for the covariance of $k^{r,s}$ and k^t, we may write

$$\mathrm{cov}(k^{r,s}, k^t) = n^{-2}\phi^{ij}\phi^k\kappa^{rs,t}_{ij,k}.$$

On expansion of the generalized cumulant using (3.3), and after taking independence into account, we find

$$\mathrm{cov}(k^{r,s}, k^t) = n^{-2}\phi^{ij}\phi^k\left(\kappa^{r,s,t}\delta_{ijk} + \kappa^s\kappa^{r,t}\delta_j\delta_{ik}[2]\right).$$

Application of (4.8) gives $\mathrm{cov}(k^{r,s}, k^t) = n^{-1}\kappa^{r,s,t}$. Similarly, the covariance of two sample covariances is

$$\begin{aligned}\mathrm{cov}(k^{r,s}, k^{t,u}) &= n^{-2}\phi^{ij}\phi^{kl}\kappa^{rs,tu}_{ij,kl}\\ &= n^{-2}\phi^{ij}\phi^{kl}\{\kappa^{r,s,t,u}\delta_{ijkl} + \kappa^r\kappa^{s,t,u}\delta_i\delta_{jkl}[4]\\ &\qquad + \kappa^{r,t}\kappa^{s,u}\delta_{ik}\delta_{jl}[2] + \kappa^r\kappa^t\kappa^{s,u}\delta_i\delta_l\delta_{jk}[4]\}\\ &= n^{-1}\kappa^{r,s,t,u} + \kappa^{r,t}\kappa^{s,u}[2]\textstyle\sum_{ij}\phi^{ij}\phi^{ij}/n^2\\ &= n^{-1}\kappa^{r,s,t,u} + \kappa^{r,t}\kappa^{s,u}[2]/(n-1).\end{aligned}$$

In the case of third and higher joint cumulants of the ks, it is convenient to introduce a new but obvious notation. For the joint cumulant of $k^{r,s}$, $k^{t,u}$, and $k^{v,w}$ we write

$$\kappa_k(r,s|t,u|v,w) = n^{-3}\phi^{ij}\phi^{kl}\phi^{mn}\kappa^{rs,tu,vw}_{ij,kl,mn}.$$

In the expansion for the 3,6 cumulant, all partitions having a unit part can be dropped because of (4.8). The third-order joint cumulant then reduces to

$$\begin{aligned}\kappa_k(r,s|t,u|v{,}w) = n^{-3}\phi^{ij}\phi^{kl}\phi^{mn}\big(&\kappa^{r,s,t,u,v,w}\delta_{ijklmn}\\ &+ \kappa^{r,s,t,v}\kappa^{u,w}\delta_{ijkm}\delta_{ln}[12] + \kappa^{r,s,t}\kappa^{u,v,w}\delta_{ijk}\delta_{lmn}[6]\\ &+ \kappa^{r,t,v}\kappa^{s,u,w}\delta_{ikm}\delta_{jln}[4] + \kappa^{r,t}\kappa^{s,v}\kappa^{u,w}\delta_{ik}\delta_{jm}\delta_{ln}[8]\big),\end{aligned}$$

where the third term has zero contribution on account of orthogonality. To simplify this expression further, we need to evaluate the coefficients of the cumulant products. For example, the coefficient of the final term above may be written as

$$n^{-3}\sum_{ijk}\phi^{ij}\phi^{ik}\phi^{jk}$$

and this sum, known as pattern function, has the value $(n-1)^{-2}$. See Table 4.1 under the pattern coded 12/13/23. On evaluating the remaining coefficients, we find

$$\begin{aligned}\kappa_k(r,s|t,u|v,w) &= \kappa^{r,s,t,u,v,w}/n^2 + \kappa^{r,s,t,v}\kappa^{u,w}[12]/\{n(n-1)\}\\ &+ \kappa^{r,t,v}\kappa^{s,u,w}[4](n-2)/\{n(n-1)^2\} + \kappa^{r,t}\kappa^{s,v}\kappa^{u,w}[8]/(n-1)^2.\end{aligned}$$

In the univariate case where k^r and $k^{r,s}$ are commonly written as $\bar{Y}$ and s^2, we may deduce from the cumulants listed above that

$$\begin{aligned}&\mathrm{var}(\bar{Y}) = \kappa_2/n, \qquad \mathrm{cov}(\bar{Y}, s^2) = \kappa_3/n,\\ &\mathrm{var}(s^2) = \kappa_4/n + 2\kappa_2^2/(n-1)\end{aligned}$$

and the third cumulant of s^2 is

$$\begin{aligned}\kappa_3(s^2) = \kappa_6/n^2 + 12\kappa_4\kappa_2/\{n(n-1)\} + 4(n-2)\kappa_3^2/\{n(n-1)^2\}&\\ + 8\kappa_2^3/(n-1)^2.&\end{aligned}$$

More generally, the joint cumulant of several k-statistics can be represented by a partition, say Υ^*. On the right of the expression for $\kappa_k(\Upsilon^*)$ appear cumulant products corresponding to the partitions Υ complementary to Υ^*, multiplied by a coefficient that depends on n and on the intersection matrix $\Upsilon^* \cap \Upsilon$. This coefficient is zero for all partitions Υ having a unit block and also for certain other partitions that satisfy the first condition in (4.8), possibly after simplification by the third condition in (4.8). For example, if $\Upsilon^* = \{ij|kl|mn\}$ and $\Upsilon = \{ijk|lmn\}$, then the coefficient

$$\phi^{ij}\phi^{kl}\phi^{mn}\delta_{ijk}\delta_{klm} = \phi^{kl}\phi^{mn}\delta_k\delta_{lmn} = 0$$

is zero even though no block of Υ is a subset of a block of Υ^*. In general, the coefficient of the complementary partition $\Upsilon = \{v_1, \ldots, v_\nu\}$ may be written as

$$n^{-\alpha}\langle\phi(\Upsilon^*), \delta(\Upsilon)\rangle = \sum \phi(v_1^*)\cdots\phi(v_\alpha^*)\,\delta(v_1)\cdots\delta(v_\nu)$$

with summation over all indices. With this notation, the joint cumulant of several k-statistics may be written as

$$\kappa_k(\Upsilon^*) = \sum_{\Upsilon \vee \Upsilon^* = 1} n^{-\alpha}\langle\phi(\Upsilon^*), \delta(\Upsilon)\rangle\,\kappa(v_1)\cdots\kappa(v_\nu). \qquad (4.10)$$

The main difficulty in using this formula lies in computing the coefficients $\langle\phi(\Upsilon^*), \delta(\Upsilon)\rangle$.

4.2.4 *Pattern matrices and pattern functions*

We now examine various ways of expressing and computing the coefficients $\langle\phi(\Upsilon^*), \delta(\Upsilon)\rangle$, also called *pattern functions*, that arise in (4.10). Evidently the coefficient is a function of n that depends on the intersection matrix $\Upsilon^* \cap \Upsilon$. In fact, since the value of ϕ depends only on the number of distinct indices and not on the number of repetitions of any index, it follows that $\langle\phi(\Upsilon^*), \delta(\Upsilon)\rangle$ must depend only on the pattern of non-zero values in $\Upsilon^* \cap \Upsilon$ and not on the actual intersection numbers. The so-called *pattern matrix* is determined only up to separate independent permutations of the rows and columns.

To take a few simple examples, suppose that $\Upsilon^* = ijk|lmn$, $\Upsilon_1 = ijl|kmn$ and $\Upsilon_2 = il|jkmn$. The intersection matrices are

$$\Upsilon^* \cap \Upsilon_1 = \begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array} \quad \text{and} \quad \Upsilon^* \cap \Upsilon_2 = \begin{array}{cc} 1 & 2 \\ 1 & 2 \end{array},$$

which are different, but the pattern matrices are identical. The pattern functions $\langle\phi(\Upsilon^*), \delta(\Upsilon)\rangle$, written explicitly as sums of coefficients, are

$$\sum \phi^{iij}\phi^{ijj} \quad \text{and} \quad \sum \phi^{ijj}\phi^{ijj},$$

both of which reduce to $\sum(\phi^{ij})^2 = n^2/(n-1)$. Similarly, if we take $\Upsilon^* = ij|kl|mn$ and $\Upsilon = ik|jm|ln$, the pattern matrix is

$$\Upsilon^* \cap \Upsilon = \begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{array} \qquad (4.11)$$

corresponding to the algebraic expression $\sum \phi^{ij}\phi^{ik}\phi^{jk}$, as if the columns of the pattern matrix were labelled i, j and k. This pattern function appears in the final term of the expression for $\kappa_k(r, s|t, u|v, w)$ above and takes the value $n^3/(n-1)^2$: see Exercise 4.6.

Because of the orthogonality of the ϕ-arrays and the δ-arrays of coefficients, many pattern functions are identically zero and it is helpful to identify these at the outset. Evidently, from (4.10), if $\Upsilon^* \vee \Upsilon < 1$, the pattern function is zero. Moreover, additional vanishing pattern functions

Table 4.1 *Some useful non-zero pattern functions*

Pattern	*Pattern function*
$1/1/\cdots/1$	n
	Two rows
$12/12$	$n^2/(n-1)$
$123/123$	$n^3/(n-1)^{(2)}$
$1234/1234$	$n^3(n+1)/(n-1)^{(3)}$
$12345/12345$	$n^4(n+5)/(n-1)^{(4)}$
$12356/123456$	$n^3(n+1)(n^2+15n-4)/(n-1)^{(5)}$
	Three rows
$12/12/12$	$n^2(n-2)/(n-1)^2$
$12/13/23$, $123/12/13$	$n^3/(n-1)^2$
$123/123/12$	$n^3(n-3)/\{(n-1)^2(n-2)\}$
$123/123/123$	$\dfrac{n^3(n^2-6n+10)}{(n-1)^2(n-2)^2}$
$123/124/34$, $1234/123/34$	$\dfrac{n^4}{(n-1)^2(n-2)}$
$123/124/134$	$\dfrac{n^4(n-3)}{(n-1)^2(n-2)^2}$
$123/24/1234$	$\dfrac{n^4(n-4)}{(n-1)^2(n-2)^2}$
$1234/1234/12$	$\dfrac{n^3(n^2-4n-1)}{(n-1)^2(n-2)(n-3)}$
$1234/1234/123$	$\dfrac{n^3(n^3-18n^2+17n+2)}{(n-1)^2(n-2)^2(n-3)}$
$1234/1234/1234$	$\dfrac{n^3(n^4-12n^3+51n^2-74n-18)}{(n-1)^2(n-2)^2(n-3)^2}$

can be identified using (4.8), but this test must, in general, be applied iteratively. For example, if $\Upsilon^* = ij|kl|m$ and $\Upsilon = ijk|lm$, we find that

$$\phi^{ij}\phi^{kl}\phi^{m}\delta_{ijk}\delta_{lm} = \phi^{kl}\phi^{m}\delta_{k}\delta_{lm} = 0$$

on applying (4.8) twice. The pattern matrix is

$$\begin{array}{cc} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{array}$$

and the corresponding pattern function is zero because the columns of this matrix can be partitioned into two blocks that connect only through one row. The pattern matrix for $\Upsilon = ikm|jl$ cannot be partitioned in this way and the corresponding pattern function is $n^2/(n-1)$. Rows containing a single entry may be deleted since $\phi^i = 1$. Table 4.1 gives a list of some of the more useful pattern functions.

Table 4.1 *Non-zero pattern functions (continued)*

Pattern	*Pattern function*
	Three rows (contd.)
1234/125/345, 12345/215/435	$n^5/(n-1)^2(n-2)^2$
1234/1235/45, 12345/1235/34	$\dfrac{n^4(n+1)}{(n-1)^2(n-2)(n-3)}$
1234/1235/145	$\dfrac{n^4(n^2-4n-1)}{(n-1)^2(n-2)^2(n-3)}$
1234/1235/1245	$\dfrac{n^4(n^3-9n^2+19n+5)}{(n-1)^2(n-2)^2(n-3)^2}$
12345/1234/125	$\dfrac{n^4(n^2-5n-2)}{(n-1)^2(n-2)^2(n-3)}$
12345/1245/12	$\dfrac{n^4(n^2-4n-9)}{(n-1)(n-1)^{(4)}}$
1234/1256/3456	$\dfrac{n^4(n+1)(n^2-5n+2)}{(n-1)^2(n-1)^2(n-3)^2}$
12345/1236/456	$\dfrac{n^4(n^3-9n^2+19n+5)}{(n-1)^2(n-2)^2(n-3)^2}$
	Four rows
12/12/12/12	$n^2(n^2-3n+3)/(n-1)^3$
13/13/12/12	$n^3/(n-1)^2$
13/23/12/12, 123/23/12/12	$n^3(n-2)/(n-1)^3$
123/13/23/12	$n^3(n-3)/(n-1)^3$
123/123/12/12	$\dfrac{n^3(n^2-4n+5)}{(n-1)^3(n-2)}$
123/123/12/13	$\dfrac{n^3(n^2-5n+7)}{(n-1)^3(n-2)}$
123/123/123/12	$\dfrac{n^3(n-3)(n^2-4n+6)}{(n-1)^3(n-2)^2}$
123/123/123/123	$\dfrac{n^3(n^4-9n^3+33n^2-60n+48)}{(n-1)^3(n-2)^3}$

To conserve space, the patterns in Table 4.1 are coded numerically. For example, the pattern (4.11) is coded under the section marked *Three rows* as 12/13/23 and takes the value $n^3/(n-1)^2$. Patterns that can be derived from (4.11) by permuting rows or columns are not listed explicitly.

4.3 Related symmetric functions

4.3.1 *Generalized k-statistics*

Generalized k-statistics are the sample versions of the generalized cumulants, so that, for example, $k^{r,st}$ is a symmetric function and is an unbiased estimate of $\kappa^{r,st}$. A fairly simple extension of the argument used in Section 4.2.2 shows that

$$k^{r,st} = n^{-1}\sum \phi^{ij} Y_i^r Y_j^s Y_j^t \tag{4.12}$$

is the required symmetric function. Similarly,

$$k^{rs,tu} = n^{-1} \sum \phi^{ij} Y_i^r Y_i^s Y_j^t Y_j^u \tag{4.13}$$

$$k^{r,s,tu} = n^{-1} \sum \phi^{ijk} Y_i^r Y_j^s Y_k^t Y_k^u \tag{4.14}$$

and so on. Note that the coefficients ϕ^{ij} in (4.12) and (4.13) could have been replaced by ϕ^{ijj} and ϕ^{iijj}, while that in (4.14) could have been written ϕ^{ijkk}, matching the partition corresponding to the required k-statistic.

To verify that these are indeed the appropriate estimators of the generalized cumulants, we observe first that the generalized k-statistics are symmetric functions of $Y_1, \ldots, Y_n$. Also, on taking expectations, we have that

$$E\{k^{r,st}\} = n^{-1}\phi^{ij}\{\kappa^{r,st}\delta_{ij} + \kappa^r \kappa^{st}\delta_i \delta_j\} = \kappa^{r,st}.$$

Similarly for $k^{rs,tu}$ and $k^{r,s,tu}$. More generally, if we define $Y_i^{rs} = Y_i^r Y_i^s$, it is immediately evident that (4.13) is an unbiased estimate of $\text{cov}(Y_i^{rs}, Y_i^{tu})$. Evidently, the sample moments are special cases of generalized k-statistics, for we may write

$$k^{rst} = n^{-1} \sum_i \phi^i Y_i^r Y_i^s Y_i^t,$$

which is the same as the expression given in Section 4.2.2. In fact, unbiased estimates exist for all generalized cumulants of order (α, β), provided only that $\alpha \le n$. In particular, unbiased estimates exist for all moments of all orders but only for ordinary cumulants whose order does not exceed n.

The importance of generalized k-statistics stems from the following properties:

(i) The generalized k-statistics are linearly independent (but not functionally independent).

(ii) Every polynomial symmetric function can be expressed uniquely as a *linear* combination of generalized k-statistics.

(iii) Any polynomial symmetric function whose expectation is independent of n can be expressed as a *linear* combination of generalized k-statistics with coefficients independent of n.

These properties are not sufficient to identify the generalized k-statistics uniquely. In fact, as will be shown in the sections that follow, there are alternative systems of symmetric functions that are in some ways more convenient than generalized k-statistics. All such systems are invertible linear functions of generalized k-statistics with coefficients independent of the sample size, and the three properties listed above are preserved under such transformations.

In view of the results derived earlier, particularly in Section 3.8, the proofs of the above assertions are fairly elementary. Only broad outline proofs are provided here. To establish (i), we suppose that there exists a linear combination of generalized k-statistics that is identically zero for all Y and show that this assumption leads to a contradiction. The expectation of such a linear combination is the same linear combination of generalized cumulants, which, by assumption, must be zero for all distributions. But this is known to be impossible because the generalized cumulants are linearly independent. Hence (i) follows.

To prove (ii), we note first that if a linear combination exists, it must be unique because the generalized k-statistics are linearly independent. For the remainder of (ii), we need to show that there are enough generalized k-statistics to span the space of polynomial symmetric functions. Without loss of generality, we may restrict attention to homogeneous symmetric functions of degree one in each of the variables $Y^1, \ldots, Y^p$. Any such polynomial may be written in the form

$$a^{i_1 i_2 \cdots i_p} Y_{i_1}^1 \cdots Y_{i_p}^p.$$

This is one of the few examples in this book of an array of coefficients that is not symmetric under index permutation. The following discussion is given for $p = 4$ but generalizes in an obvious way to arbitrary p. For $p = 4$, symmetry implies that the array a^{ijkl} can have at most $B_4 = 15$ distinct values, namely a^{1111}, $a^{1112}[4]$, $a^{1122}[3]$, $a^{1123}[6]$ and a^{1234}. To see that this is so, we note the quartic $\sum a^{ijkl}Y_i^rY_j^sY_k^tY_l^u$ must be invariant under permutation of $Y_1, \ldots, Y_n$. Hence, for any $n \times n$ permutation matrix π_r^i, we must have

$$a^{ijkl} = \pi_r^i\pi_s^j\pi_t^k\pi_u^l a^{rstu} = a^{\pi_i\pi_j\pi_k\pi_l},$$

where $\pi_1, \ldots, \pi_n$ is a permutation of the first n integers. Hence, if i,j,k,l are distinct integers, then

$$a^{iiij} = a^{1112} \neq a^{2111}$$
$$a^{iijk} = a^{1123} \neq a^{1233}$$

and so on. It follows that there are exactly 15 linearly independent symmetric functions of degree one in each of four distinct variables. Any convenient basis of 15 linearly independent symmetric functions, each of degree one in the four variables, is adequate to span the required space. For example, one possibility would be to set each of the distinct as to unity in turn, the remainder being kept at zero. However, the generalized cumulants provide an alternative and more convenient basis. Of course, if the four variables were not distinct, as, for example, in univariate problems, the number of linearly independent symmetric functions of total degree four and of specified degree in each of the component variables would be reduced. In the univariate case, there are five linearly independent symmetric functions of degree four, one for each of the partitions of the *number* four.

The argument just given holds for homogeneous symmetric functions of degree one in an arbitrary number of *distinct* random variables. Even when the number of distinct variables is less than the degree of the polynomial, it is often convenient in the algebra to sustain the fiction that there are as many distinct variables as the total degree of the polynomial. Effectively, we replicate an existing variable and the algebra treats the replicate as a distinct variable. This device of algebraic pseudo-replication often simplifies the algebra and avoids the need to consider numerous special cases.

The final assertion (iii) follows from completeness of the set of generalized k-statistics together with the fact that generalized k-statistics are unbiased estimates of the corresponding cumulant.

4.3.2 *Symmetric means and poly*

The symmetric means have previously been introduced in Section 4.2.2 as the unique polynomial symmetric functions that are unbiased estimates of products of moments. Symmetric means, also called power products in the combinatorial literature dealing with the univariate case (Dressel, 1940), are most conveniently indexed in the multivariate case by a partition of a set of indices. Of course, this must be done in such a way that the notation does not give rise to confusion with generalized k-statistics, which are indexed in a similar manner. Our proposal here is to write $k^{(rs)(tuv)}$ for the estimate of the moment product $\kappa^{rs}\kappa^{tuv}$. The bracketing of superscripts is intended to suggest that some kind of product is involved. In a sense, the letter k is redundant or may be inferred from the context, and we might well write $(rs)(tuv)$ or $\langle(rs)(tuv)\rangle$ corresponding more closely to the conventions used in the univariate case where $k^{(11)(111)}$ would typically be written as 23 or $\langle 2\,3\rangle$ (MacMahon, 1915; Dressel, 1940; Tukey, 1950, 1956a). In this chapter, we use the notations $k^{(rs)(tuv)}$ and $(rs)(tuv)$ interchangeably: the latter notation has the advantage of greater legibility.

The expressions for the symmetric means are rather simple. For instance, it is easily verified that

$$(rs)(tuv) \equiv k^{(rs)(tuv)} = \sum_{i\neq j} Y_i^rY_i^sY_j^tY_j^uY_j^v/n^{(2)}$$

and

$$(rs)(tu)(v) \equiv k^{(rs)(tu)(v)} = \sum_{i\neq j\neq k} Y_i^rY_i^sY_j^tY_j^uY_k^v/n^{(3)},$$

where the sum extends over distinct subscripts and the divisor is just the number of terms in the sum. The extension to arbitrary partitions is immediate and need not be stated explicitly. Additional symmetric means are listed in Section 4.2.2.

The polykays are the unique polynomial symmetric functions that are unbiased estimates of cumulant products. It is natural therefore to write $k^{(r,s)(t,u,v)}$ or $(r,s)(t,u,v)$ to denote that unique symmetric function whose expectation is the cumulant product $\kappa^{r,s}\kappa^{t,u,v}$. Again, the bracketing suggests multiplication, and the commas indicate that cumulant products rather than moment products are involved. We first give a few examples of polykays and then show how the three systems of symmetric functions, generalized k-statistics, symmetric means and polykays are related.

It was shown in Section 4.2.2 that

$$k^{(r)(s)} = \sum \phi^{i|j} Y_i^r Y_j^s / n^{(2)}$$

is an unbiased estimate of the product $\kappa^r\kappa^s$, where $\phi^{i|j} = 1$ if $i \neq j$ and zero otherwise. By extension,

$$k^{(r)(s,t)} = \sum \phi^{i|jk} Y_i^r Y_j^s Y_k^t / n^{(2)},$$

with suitably chosen coefficients $\phi^{i|jk}$, is an unbiased estimate of $\kappa^r\kappa^{s,t}$. The required coefficients are

$$\phi^{i|jk} = \begin{cases} 0 & \text{if } i = j \text{ or } i = k \\ 1 & \text{if } j = k \neq i \\ -1/(n-2) & \text{otherwise,} \end{cases}$$

as can be seen by writing $k^{(r)(s,t)} = k^{(r)(st)} - k^{(r)(s)(t)}$ in the form

$$\sum\nolimits^{\#} Y_i^r Y_j^s Y_j^t / n^{(2)} - \sum\nolimits^{\#} Y_i^r Y_j^s Y_k^t / n^{(3)}.$$

In addition, we may write

$$k^{(r,s)(t,u)} = \sum \phi^{ij|kl} Y_i^r Y_j^s Y_k^t Y_l^u / n^{(2)}$$

for the unbiased estimate of the product $\kappa^{r,s}\kappa^{t,u}$, where

$$\phi^{ij|kl} = \begin{cases} 0 & i \text{ or } j = k \text{ or } l \\ 1 & i = j \text{ and } k = l \neq i \\ -1/(n-2) & i = j,\ k \neq l \neq i \text{ or reverse} \\ 1/\{(n-2)(n-3)\} & \text{all distinct.} \end{cases}$$

Also,

$$k^{(r)(s)(t,u)} = \sum \phi^{i|j|kl} Y_i^r Y_j^s Y_k^t Y_l^u / n^{(3)}$$

is an unbiased estimate of the product $\kappa^r\kappa^s\kappa^{t,u}$, where the coefficients are given by

$$\phi^{i|j|kl} = \begin{cases} 0 & \text{same value occurs in different blocks} \\ 1 & k = l,\ i \neq j \neq k \\ -1/(n-3) & \text{all indices distinct.} \end{cases}$$

Evidently, the coefficients for the polykays are more complicated than those for ordinary k-statistics, symmetric means or generalized k-statistics. These complications affect the algebra but do not necessarily have much bearing on computational difficulty. Neither the formulae given above nor the corresponding ones for generalized k-statistics are suitable as a basis for computation. Computational questions are discussed in Section 4.5.

Since the generalized k-statistics, symmetric means and polykays are three classes of symmetric functions indexed in the same manner, it is hardly surprising to learn that any one of the three classes can be expressed as a linear function of any other. In fact, all of the necessary formulae have been given in Section 3.6.2: we need only make the obvious associations of symmetric means with moment products, polykays with cumulant products and generalized k-statistics with generalized cumulants. The following are a few simple examples of symmetric means expressed in terms of polykays.

$$\begin{aligned}(rs)(t) &= (r,s)(t) + (r)(s)(t) = \{(r,s) + (r)(s)\}(t)\\ (rs)(tu) &= (r,s)(t,u) + (r,s)(t)(u) + (r)(s)(t,u) + (r)(s)(t)(u)\\ &= \{(r,s) + (r)(s)\}\{(t,u) + (t)(u)\}\\ (rs)(tuv) &= \{(r,s) + (r)(s)\}\{(t,u,v) + (t)(u,v)[3] + (t)(u)(v)\}.\end{aligned}$$

Of course, $(rs) = (r,s) + (r)(s)$ and similarly

$$(tuv) = (t,u,v) + (t)(u,v)[3] + (t)(u)(v)$$

but this does not imply that $k^{(rs)(tuv)}$ is the same as $k^{(rs)}k^{(tuv)}$. The above multiplication formulae for the indices are purely symbolic and the multiplication must be performed first before the interpretation is made in terms of polykays. Thus $(rs)(tuv)$ is expressible as the sum of the 10 polykays whose indices are sub-partitions of $rs|tuv$.

The corresponding expressions for polykays in terms of symmetric means may be written symbolically as

$$\begin{aligned}(r,s)(t) &= \{(rs) - (r)(s)\}(t)\\ (r,s)(t,u) &= \{(rs) - (r)(s)\}\{(tu) - (t)(u)\}\\ (r,s)(t,u,v) &= \{(rs) - (r)(s)\}\{(tuv) - (t)(uv)[3] + 2(t)(u)(v)\}.\end{aligned} \qquad (4.15)$$

Again, it is intended that the indices should be multiplied algebraically before the interpretation is made in terms of symmetric means. Thus, $(r,s)(t,u,v)$ is a linear combination of the 10 symmetric means whose indices are sub-partitions of $rs|tuv$. The coefficients in this linear combination are values of the Möbius function for the partition lattice.

From the identity connecting generalized cumulants with products of ordinary cumulants, it follows that we may express generalized k-statistics in terms of polykays using (3.3) and conversely for polykays in terms of generalized cumulants using (3.18). By way of illustration, we find using (3.3) that

$$k^{rs,tu} = k^{(r,s,t,u)} + k^{(r)(s,t,u)}[4] + k^{(r,t)(s,u)}[2] + k^{(r)(t)(s,u)}[4],$$

where $k^{(r,s,t,u)} \equiv k^{r,s,t,u}$ and the sum extends over all partitions complementary to $rs|tu$. The inverse expression giving polykays in terms of generalized cumulants is a little more complicated but fortunately it is seldom needed. Application of (3.18) gives, after some arithmetic,

$$6k^{(r,s)(t,u)} = k^{r,stu}[4] - k^{rs,tu}[3] - 2k^{r,s,tu}[2] + k^{r,t,su}[4] - k^{r,s,t,u}.$$

This identity can be read off the sixth row of the 15 × 15 matrix given at the end of Section 3.6.2. The remaining expressions involving four indices are

$$\begin{aligned}6k^{(r)(s,t,u)} &= -k^{r,stu}[4] + k^{rs,tu}[3] + 2k^{rs,t,u}[3] - k^{r,s,tu}[3] - 2k^{r,s,t,u}\\ 6k^{(r)(s)(t,u)} &= -k^{r,stu}[2] + 2k^{rst,u}[2] - 2k^{rs,tu} + k^{rt,su}[2]\\ &\qquad + 2k^{r,s,tu} - k^{r,t,su}[4] - k^{rs,t,u} + k^{r,s,t,u}\\ 6k^{(r)(s)(t)(u)} &= 6k^{rstu} - 2k^{r,stu}[4] - k^{rs,tu}[3] + k^{r,s,tu}[6] - k^{r,s,t,u}.\end{aligned}$$

These examples emphasize that the relationship between the polykays and the generalized k-statistics is linear, invertible and that the coefficients are independent of the sample size. Because of linearity, unbiasedness of one set automatically implies unbiasedness of the other.

4.4 Derived scalars

To each of the derived scalars discussed in Section 2.8 there corresponds a sample scalar in which the κs are replaced by ks. Denote by $k_{i,j}$ the matrix inverse of $k^{i,j}$ and write

$$\begin{aligned} p\bar{r}_{13}^2 &= k^{r,s,t}k^{u,v,w}k_{r,s}k_{t,u}k_{v,w} \\ p\bar{r}_{23}^2 &= k^{r,s,t}k^{u,v,w}k_{r,u}k_{s,v}k_{t,w} \\ p\bar{r}_4 &= k^{r,s,t,u}k_{r,s}k_{t,u} \end{aligned}$$

for the sample versions of $p\bar{\rho}_{13}^2$, $p\bar{\rho}_{23}^2$ and $p\bar{\rho}_4$ defined by (2.14)–(2.16). Although these three statistics are in a sense, the obvious estimators of the invariant parameters, they are not unbiased because, for example, $k_{r,s}$ is not unbiased for $\kappa_{r,s}$.

By their construction, the three statistics listed above are invariant under affine transformation of the components of X. Hence their expectations and joint cumulants must also be expressible in terms of invariants. To obtain such expressions, it is convenient to expand the matrix inverse $k_{r,s}$ in an asymptotic expansion about $\kappa_{r,s}$. If we write

$$k^{r,s} = \kappa^{r,s} + \epsilon^{r,s}$$

it follows that $\epsilon^{r,s} = O_p(n^{-1/2})$ in the sense that $n^{1/2}\epsilon^{r,s}$ has a non-degenerate limiting distribution for large n. The matrix inverse may therefore be expanded as

$$k_{r,s} = \kappa_{r,s} - \epsilon_{r,s} + \epsilon_{r,i}\epsilon_{s,j}\kappa^{i,j} - \epsilon_{r,i}\epsilon_{j,k}\epsilon_{l,s}\kappa^{i,j}\kappa^{k,l} + \cdots$$

where $\epsilon_{r,s} = \kappa_{r,i}\kappa_{s,j}\epsilon^{i,j}$ is not the matrix inverse of $\epsilon^{r,s}$. The advantage of working with this expansion is that it involves only $\epsilon^{r,s}$ whose cumulants, apart from the first, are the same as those of $k^{r,s}$. In the case of the scalar $p\bar{r}_4$, we may write

$$\begin{aligned} p\bar{r}_4 = k^{r,s,t,u}&(\kappa_{r,s} - \epsilon_{r,s} + \epsilon_{r,i}\epsilon_{s,j}\kappa^{i,j} - \cdots) \\ &\times(\kappa_{t,u} - \epsilon_{t,u} + \epsilon_{t,i}\epsilon_{u,j}\kappa^{i,j} - \cdots). \end{aligned}$$

On taking expectation and including terms up to order $O(n^{-1})$ only, we find using the identity (3.3) that

$$\begin{aligned} E(p\bar{r}_4) = p\bar{\rho}_4 &- 2\kappa_k(r,s,t,t|r,s) + \kappa_k(r,s,t,u)\kappa_k(r,s|t,u) \\ &+ 2\kappa_k(r,s,t,t)\kappa_k(r,u|s,u), \end{aligned} \tag{4.16}$$

where, for example, $\kappa_k(r,s,t,t|r,s)$ is a convenient shorthand notation for the scalar

$$\kappa_{r,v}\kappa_{s,w}\kappa_{t,u}\,\mathrm{cov}(k^{r,s,t,u}, k^{v,w})$$

and $\kappa_k(r,s,t,u) = \kappa^{r,s,t,u}$ by construction. Simplification of this particular scalar gives

$$\begin{aligned} \kappa_{r,v}\kappa_{s,w}\kappa_{t,u}\{\kappa^{r,s,t,u,v,w}/n &+ (\kappa^{r,s,t,v}\kappa^{u,w}[6] + \kappa^{r,t,u,v}\kappa^{s,w}[2])/(n-1) \\ &+ (\kappa^{r,s,v}\kappa^{t,u,w}[4] + \kappa^{r,t,w}\kappa^{s,u,v}[2])/(n-1)\} \\ = p\bar{\rho}_6/n&+(6p\bar{\rho}_4 + 2p^2\bar{\rho}_4 + 4p\bar{\rho}_{13}^2 + 2p\bar{\rho}_{23}^2)/(n-1). \end{aligned}$$

After simplification of the remaining terms in (4.16), we are left with

$$\begin{aligned} E(\bar{r}_4) = \bar{\rho}_4(1 - 8/n - 2p/n) &- 2\bar{\rho}_6/n - 8\bar{\rho}_{13}^2/n - 4\bar{\rho}_{23}^2/n \\ &+ 2\bar{\rho}_{14}^2/n + \bar{\rho}_{24}^2/n + O(n^{-2}), \end{aligned}$$

where $p\bar\rho^2_{14}$ and $p\bar\rho^2_{24}$ are defined in Section 2.8.

Similar expressions may be found for higher-order cumulants, though such expressions tend to be rather lengthy especially when carried out to second order. To first order we have, for example, that

$$\begin{aligned}\mathrm{var}(p\bar r_4) = \kappa_k(r,r,s,s|t,t,u,u) - 4\kappa_k(r,r,s,t)\kappa_k(u,u,v,v|s,t)&\\ + O(n^{-2}),&\end{aligned}$$

both terms being of order $O(n^{-1})$. This variance can be expressed directly in terms of invariants but the expression is rather lengthy and complicated and involves invariants of a higher degree than those so far discussed.

In the case of jointly normal random variables, the second term above vanishes and the first reduces to

$$\mathrm{var}(p\bar r_4) = (8p^2 + 16p)/n + O(n^{-2}). \tag{4.17}$$

In fact, it can be shown (Exercise 4.10), that the limiting distribution of $n^{1/2}p\bar r_4$ is normal with zero mean and variance $8p^2 + 16p$. The normal limit is hardly surprising because all the k-statistics are asymptotically normal: the joint cumulants behave in the same way as the cumulants of a straightforward average of independent random variables.

In the case of the quadratic scalar, $\bar r^2_{13}$, we may make the following expansion

$$\begin{aligned}p\bar r^2_{13} = k^{r,s,t}k^{u,v,w}&(k_{r,s} - \epsilon_{r,s} + \epsilon_{r,i}\epsilon_{s,j}\kappa^{i,j} - \cdots)\\ &\times(k_{t,u} - \epsilon_{t,u} + \epsilon_{t,i}\epsilon_{u,j}\kappa^{i,j} - \cdots)\\ &\times(k_{v,w} - \epsilon_{v,w} + \epsilon_{v,i}\epsilon_{w,j}\kappa^{i,j} - \cdots)\end{aligned}$$

On taking expectation, and including terms up to order $O(n^{-1})$, we find

$$\begin{aligned}E(p\bar r^2_{13}) = p\bar\rho^2_{13} &+ \kappa_k(r,r,s|s,t,t) - 2\kappa_k(r,s,t)\kappa_k(t,u,u|r,s)\\ &- 2\kappa_k(u,v,v)\kappa_k(r,s,u|r,s) - 2\kappa_k(r,r,s)\kappa_k(u,t,t|s,u)\\ &+ 2\kappa_k(r,r,t)\kappa_k(t,v,w)\kappa_k(v,u|u,w)\\ &+ \kappa_k(r,r,t)\kappa_k(u,u,w)\kappa_k(t,s|s,w) + O(n^{-2}).\end{aligned}$$

Again, the above formula may be expressed directly in terms of invariants. For example, the second term may be written as

$$p(\bar\rho_6 + p\bar\rho_4 + 8\bar\rho_4 + 5\bar\rho^2_{13} + 4\bar\rho^2_{23} + 2p + 4)/n + O(n^{-2}).$$

In the case of normal random variables, only the second term contributes, giving

$$E(p\bar r^2_{13}) = 2p(p+2)/n + O(n^{-2}).$$

In fact, it may be shown (Exercise 4.11), that for large n,

$$\frac{(n-1)(n-2)p\bar r^2_{13}}{2n(p+2)} \sim \chi^2_p$$

under the assumption of normality.

Similar calculations for $\bar r^2_{23}$ give

$$\frac{(n-1)(n-2)p\bar r^2_{23}}{6n} \sim \chi^2_{p(p+1)(p+2)/6}$$

for large n under the assumption of normality (Exercise 4.11).

Finally, it is worth pointing out that, although $\bar{r}_{13}^2$, $\bar{r}_{23}^2$ and $\bar{r}_4$ are the most commonly used scalars for detecting multivariate non-normality, they are not the only candidates for this purpose. Other scalars that have an equal claim to be called the sample versions of $p\bar{\rho}_{13}^2$ and $p\bar{\rho}_{23}^2$ include

$$k^{(r,s,t)(u,v,w)}k_{r,s}k_{t,u}k_{v,w} \quad \text{and} \quad k^{(r,s,t)(u,v,w)}k_{r,u}k_{s,v}k_{t,w}.$$

Another class of invariant scalars that has considerable geometrical appeal despite its incompleteness, may be defined by examining the directional standardized skewness and kurtosis and choosing the directions corresponding to maxima and minima. In two dimensions, if $\bar{\rho}_{13}^2 = 0$, there are three directions of equal maximum skewness separated by $2\pi/3$, and zero skewness in the orthognal direction. On the other hand, if $4\bar{\rho}_{23}^2 = 3\bar{\rho}_{13}^2$ there is one direction of maximum skewness and zero skewness in the orthogonal direction. More generally, a complete picture of the directional skewness involves a combination of the above: see Exercises 2.36 and 2.37. In the case of the directional kurtosis, it is necessary to distinguish between maxima and minima. Machado (1976, 1983) gives approximate percentage points of the distribution under normality of the sample versions of the maximized directional skewness and kurtosis and the minimized directional kurtosis.

4.5 Computation

4.5.1 *Practical issues*

Practical considerations suggest strongly that we are seldom likely to require k-statistics or polykays of degree more than about four. Without further major assumptions, this is enough to enable the statistician to make approximate interval estimates for the first two cumulants only and to make approximate point estimates for cumulants of orders three and four. As a general rule, the higher-order k-statistics tend to have large sampling variability and consequently, large quantities of data are required to obtain estimates that are sufficiently precise to be useful. By way of example, under the optimistically favourable assumption that the data are approximately normally distributed with unit variance, approximately 40 000 observations are required to estimate the fourth cumulant accurately to one decimal place. In the case of observations distributed approximately as Poisson with unit mean, the corresponding sample size is just over half a million. More realistically, to estimate κ_4 accurately to the nearest whole number, we require 400 observations if circumstances are favourable, and more than 5000 if they are only a little less favourable. These rough calculations give some idea of the kind of precision achievable in practice.

In addition to the statistical considerations just mentioned, it should be pointed out that the number of k-statistics and, more emphatically, the number of polykays, grows rapidly with the number of variables and with the order of k-statistic or polykay considered. For example, the number of distinct k-statistics of total order k in up to p variables is $\binom{k+p-1}{k}$, while, if we include those of order less than k, the number becomes $\binom{k+p}{k}$. The latter formula includes the sample size itself as the k-statistic of order zero. On the other hand, the number of polykays of total order exactly k in up to q variables is $p_q(k)$, where, for example, $p_1(k)$ is the number of partitions of the number k and

$$p_2(k) = \sum_{\substack{a+b=k \\ a,b\geq 0}} p(a,b),$$

where $p(a,b)$ is the number of distinct partitions of a set containing a objects of type A and b objects of type B. For the purposes of this discussion, objects of the same type are assumed to be indistinguishable. In an obvious notation,

$$p_3(k) = \sum_{\substack{a+b+c=k \\ a,b,c\geq 0}} p(a,b,c)$$

Table 4.2 *Numbers of k-statistics and polykays of various orders*

Number of distinct k-statistics

	Number of variables			
Order	1	2	3	4
1	1	2	3	4
2	1	3	6	10
3	1	4	10	20
4	1	5	15	35
5	1	6	21	56
k	1	$k+1$	$\binom{k+2}{2}$	$\binom{k+3}{3}$

Number of distinct polykays

1	1	2	3	4
2	2	6	12	20
3	3	14	38	80
4	5	33	117	305
5	7	70	336	1072
k	$p_1(k)$	$p_2(k)$	$p_3(k)$	$p_4(k)$

involves a sum over various tri-partite partition numbers of three integers. Some values for these totals are given in Table 4.2.

Partly for the reasons just given, but mainly to preserve the sanity of author and reader alike, we discuss computation only for the case $k \le 4$.

4.5.2 *From power sums to symmetric means*

As a first step in the calculation, we compute the following $\binom{p+4}{4}$ 'power sums' and 'power products': $k^0 = 1$,

$$k^r = n^{-1}\sum_i Y_i^r \qquad k^{rs} = n^{-1}\sum_i Y_i^r Y_i^s$$
$$k^{rst} = n^{-1}\sum_i Y_i^r Y_i^s Y_i^t \qquad k^{rstu} = n^{-1}\sum_i Y_i^r Y_i^s Y_i^t Y_i^u .$$

One simple way of organizing these calculations for a computer is to augment the data by adding the dummy variable $Y_i^0 = 1$ and by computing k^{rstu} for $0 \le r \le \cdots \le u$. This can be accomplished in a single pass through the data. The three-index, two-index and one-index quantities can be extracted as k^{0rst}, k^{00rs} and k^{000r} as required.

The above quantities are special cases of symmetric means. All subsequent $k$statistics and polykays whose total degree does not exceed 4 may be derived from these symmetric functions. No further passes through the data matrix are required. The remaining symmetric means are $k^{(r)} = k^r$,

$$\begin{aligned}
k^{(r)(s)} &= \{nk^r k^s - k^{rs}\}/(n-1) \\
k^{(r)(s)(t)} &= \{n^2 k^r k^s k^t - nk^r k^{st}[3] + 2k^{rst}\}/(n-1)^{(2)} \\
k^{(r)(s)(t)(u)} &= \{n^3 k^r k^s k^t k^u - n^2 k^r k^s k^{tu}[6] + 2nk^r k^{stu}[4] \\
&\qquad + nk^{rs} k^{tu}[3] - 6k^{rstu}\}/(n-1)^{(3)}. \qquad (4.18)
\end{aligned}$$

Included, essentially as special cases of the above, are the following symmetric means:

$$\begin{aligned}
k^{(r)(st)} &= \left\{nk^r k^{st} - k^{rst}\right\}/(n-1)\\
k^{(rs)(tu)} &= \left\{nk^{rs} k^{tu} - k^{rstu}\right\}/(n-1)\\
k^{(r)(s)(tu)} &= \left\{n^2 k^r k^s k^{tu} - nk^r k^{stu} - nk^s k^{rtu}\right.\\
&\qquad \left. - nk^{rs} k^{tu} + 2k^{rstu}\right\}/(n-1)^{(2)}.
\end{aligned} \tag{4.19}$$

Of course, if the computations are organized as suggested in the previous paragraph, then the above formulae may all be regarded as special cases of expression (4.18) for $k^{(r)(s)(t)(u)}$. By way of example, direct substitution gives

$$\begin{aligned}
k^{(0)(r)(s)(t)} &= \left(n^3 k^r k^s k^t - 3n^2 k^r k^s k^t - n^2 k^r k^{st}[3]\right.\\
&\quad \left. + 2nk^{rst} + 2nk^r k^{st}[3] + nk^r k^{st}[3] - 6k^{rst}\right)/(n-1)^{(3)}\\
&= k^{(r)(s)(t)}.
\end{aligned}$$

For $p = 2$, there are 55 such terms, all linearly independent but functionally dependent on the 15 basic power sums.

4.5.3 *From symmetric means to poly*

In going from symmetric means to polykays, the expressions are all linear and the coefficients are integers independent of the sample size. These properties lead to simple recognizable formulae. Some examples are as follows:

$$\begin{aligned}
k^{(r,s)} &= k^{(rs)} - k^{(r)(s)}\\
k^{(r)(s,t)} &= k^{(r)(st)} - k^{(r)(s)(t)}\\
k^{(r,s,t)} &= k^{(rst)} - k^{(r)(st)}[3] + 2k^{(r)(s)(t)}\\
k^{(r,s)(t,u)} &= k^{(rs)(tu)} - k^{(r)(s)(tu)} - k^{(rs)(t)(u)} + k^{(r)(s)(t)(u)}.
\end{aligned}$$

More generally, the Möbius coefficients that occur in the above formulae may be obtained using the symbolic index multiplication formulae (4.15). For example, in the final expression above, the indices and the coefficients are given by the expression

$$\{(rs) - (r)(s)\}\{(tu) - (t)(u)\}.$$

Similarly, in the expression for $k^{(r)(s,t,u)}$, the indices and the coefficients are given by

$$(r)\{(stu) - (s)(tu)[3] + 2(s)(t)(u)\}.$$

The foregoing two-stage operation, from power sums to symmetric means to polykays, produces the symmetric means as an undesired by-product of the computation. For most statistical purposes, it is the polykays that are most useful and, conceivably, there is some advantage to be gained in going from the power sums to the polykays directly. The machinery required to do this will now be described. First, we require the 15 × 15 upper triangular matrix $\mathbf{M} = m(\Upsilon_i, \Upsilon_j)$ whose elements are the values of the Möbius function for the lattice of partitions of four items. In addition, we require the vector $\mathbf{K}$ whose 15 elements are products of power sums.

In practice, since $\mathbf{M}$ is rather sparse, it should be possible to avoid constructing the matrix explicitly. To keep the exposition as simple as possible, however, we suppose here that $\mathbf{M}$ and $\mathbf{K}$ are constructed explicitly as shown in Table 4.3.

Table 4.3 *The Möbius matrix and the vector of power sums used in* (4.20) *to compute the polykays*

							$\mathbf{M}$								$\mathbf{K}$
1	−1	−1	−1	−1	−1	−1	−1	2	2	2	2	2	2	−6	k^{rstu}
	1							−1	−1		−1			2	$k^{rst}k^{u}$
		1						−1		−1		−1		2	$k^{rsu}k^{t}$
			1						−1	−1			−1	2	$k^{rtu}k^{s}$
				1							−1	−1	−1	2	$k^{stu}k^{r}$
					1			−1					−1	1	$k^{rs}k^{tu}$
						1			−1			−1		1	$k^{rt}k^{su}$
							1			−1	−1			1	$k^{ru}k^{st}$
								1						−1	$k^{rs}k^{t}k^{u}$
									1					−1	$k^{rt}k^{s}k^{u}$
										1				−1	$k^{ru}k^{s}k^{t}$
											1			−1	$k^{st}k^{r}k^{u}$
												1		−1	$k^{su}k^{r}k^{t}$
													1	−1	$k^{tu}k^{r}k^{s}$
														1	$k^{r}k^{s}k^{t}k^{u}$

Evidently, (4.18) and (4.19) amount to a statement that the vector of symmetric means is given by the matrix product

$$\mathbf{D}_1\mathbf{M}^T\mathbf{D}_2\mathbf{K}$$

where

$$\mathbf{D}_1 = \mathrm{diag}\{1/n^{(\nu_j)}\}, \quad \mathbf{D}_2 = \mathrm{diag}\{n^{\nu_j}\}$$

and ν_j is the number of blocks in the jth partition. The polykays are obtained by Möbius inversion of the symmetric means, giving the vector

$$\mathbf{P} = \mathbf{M}\mathbf{D}_1\mathbf{M}^T\mathbf{D}_2\mathbf{K}. \tag{4.20}$$

By this device, the computation of polykays is reduced to a simple linear operation on matrices and vectors. Despite this algebraic simplicity, it is necessary to take care in the calculation to avoid rounding error, particularly where n is large or where some components have a mean value that is large.

The inverse relationship, giving $\mathbf{K}$ in terms of $\mathbf{P}$, though not of much interest for computation, involves a matrix having a remarkably elegant form. The (i, j) element of the inverse matrix is n raised to the power $|\Upsilon_i \vee \Upsilon_j| - |\Upsilon_i|$, where $|\Upsilon|$ is the number of blocks in the partition. See Exercise 4.18. It follows that the inverse matrix has a unit entry where $\mathbf{M}$ is non-zero and negative powers of n elsewhere.

4.6 Application to sampling

4.6.1 *Simple random sampling*

In a population containing N units or individuals, there are $\binom{N}{n}$ distinct subsets of size n. The subsets are distinct with regard to their labels though not necessarily in their values. A sample of size n chosen in such a way that each of the distinct subsets occurs with equal probability, namely $\binom{N}{n}^{-1}$, is said to be a simple random sample of size n taken without replacement from the population. In particular, each unit in the population occurs in such a sample with probability

n/N; each distinct pair occurs in the sample with probability $n(n-1)/\{N(N-1)\}$, and similarly for triplets and so on.

Suppose that on the ith unit in the sample, a p-dimensional variable $Y_i = Y_i^1, \ldots, Y_i^p$, $(i = 1, \ldots, n)$ is measured. Most commonly, $p = 1$, but it is more convenient here to keep the notation as general as possible. Let k^r be the rth component of the sample average and let K^r be the corresponding population average. Each unit in the population occurs in exactly $\binom{N-1}{n-1}$ of the $\binom{N}{n}$ distinct samples. Thus, if we denote by ave(k^r), the average value of the sample mean, averaged over all possible samples, we have that

$$\begin{aligned}\text{ave}(k^r) &= \binom{N}{n}^{-1} n^{-1} \sum_{\text{all subsets}} Y_1^r + \cdots + Y_n^r \\ &= \binom{N}{n}^{-1} n^{-1} \binom{N-1}{n-1} \{Y_1^r + \cdots + Y_N^r\} \\ &= N^{-1}\{Y_1^r + \cdots + Y_N^r\} = K^r.\end{aligned}$$

In other words, ave(k^r), averaged over all possible samples, is just the same function computed for the whole population of N units, here denoted by K^r.

The same argument immediately gives ave$(k^{rs}) = K^{rs}$, where

$$K^{rs} = N^{-1}\{Y_1^r Y_1^s + \cdots + Y_N^r Y_N^s\}$$

is a population average of products. It follows by direct analogy that ave$(k^{rst}) = K^{rst}$ and so on.

In order to show that symmetric means have the same property, namely

$$\begin{aligned}\text{ave}\{k^{(r)(s)}\} &= \sum\nolimits^{\#} Y_i^r Y_j^s / \{N(N-1)\} = K^{(r)(s)} \\ \text{ave}\{k^{(r)(st)}\} &= \sum\nolimits^{\#} Y_i^r Y_j^s Y_j^t / \{N(N-1)\} = K^{(r)(st)},\end{aligned}$$

where summation runs from 1 to N over unequal indices, we need only replace each occurrence of 'unit' in the previous argument with 'pair of distinct units' and make appropriate cosmetic changes in the formulae. For example, each pair of distinct units in the population occurs in $\binom{N-2}{n-2}$ of the $\binom{N}{n}$ distinct samples. It follows that the average value of the symmetric mean, $k^{(r)(s)}$ is

$$\begin{aligned}\text{ave}\{k^{(r)(s)}\} &= \binom{N}{n}^{-1} \frac{1}{n(n-1)} \sum_{\text{all subsets}} Y_1^r Y_2^s + \cdots + Y_{n-1}^r Y_n^s \\ &= \binom{N}{n}^{-1} \frac{1}{n(n-1)} \binom{N-2}{n-2} \sum_{\substack{1 \le i,j \le N \\ i \ne j}} Y_i^r Y_j^s \\ &= K^{(r)(s)}.\end{aligned}$$

This argument easily extends to any symmetric mean and hence to any polykay. The conclusion can therefore be summarized as follows.

> *Under simple random sampling, the average value of any polykay or symmetric mean, averaged over all $\binom{N}{n}$ possible samples, is just the same function computed in the population of N values.*

Tukey (1950), carefully avoiding the more conventional terminology of unbiasedness, refers to this property as 'inheritance on the average'. To see that this property does not hold for arbitrary symmetric functions, see Exercise 4.13.

The main advantage in the present context of working with polykays, rather than with the less numerous power sums and products, is that variances and higher-order cumulants of computed statistics are more aesthetically appealing when expressed linearly in terms of population polykays.

4.6.2 *Joint cumulants of k-statistics*

In this section, no attempt will be made at achieving total generality. Instead, we concentrate mainly on the joint cumulants likely to be of most use in applications, usually where the degree does not exceed four. First, we give a few simple derivations of the more useful formulae. Since $\operatorname{ave}(k^r) = K^r$, it follows that the covariance of k^r and k^s is

$$\operatorname{ave}\{k^r k^s - K^r K^s\},$$

averaged over all possible samples. From (4.18) it follows that

$$\begin{aligned} k^r k^s &= k^{(r)(s)} + k^{(r,s)}/n \\ K^r K^s &= K^{(r)(s)} + K^{(r,s)}/N. \end{aligned}$$

Hence, since $\operatorname{ave}(k^{(r)(s)}) = K^{(r)(s)}$, we have

$$\operatorname{cov}(k^r, k^s) = K^{r,s}\left(\frac{1}{n} - \frac{1}{N}\right),$$

a well known result easily derived in other ways.

Similarly, the covariance of k^r and $k^{s,t}$ may be written as

$$\operatorname{ave}\{k^r k^{s,t} - K^r K^{s,t}\}.$$

From the multiplication formula

$$\begin{aligned} k^r k^{s,t} &= k^{(r)(s,t)} + k^{r,s,t}/n \\ K^r K^{s,t} &= K^{(r)(s,t)} + K^{r,s,t}/N, \end{aligned}$$

it follows that

$$\operatorname{cov}(k^r, k^{s,t}) = K^{r,s,t}\left(\frac{1}{n} - \frac{1}{N}\right).$$

In the case of the covariance of two sample variances or covariances, we use the multiplication formula

$$k^{r,s}k^{t,u} = k^{(r,s)(t,u)} + k^{r,s,t,u}/n + k^{(r,t)(s,u)}[2]/(n-1),$$

together with an identical expression for a product of Ks. This gives

$$\begin{aligned} \operatorname{cov}(k^{r,s}, k^{t,u}) = K^{r,s,t,u}&\left(\frac{1}{n} - \frac{1}{N}\right) \\ &+ K^{(r,t)(s,u)}[2]\left(\frac{1}{n-1} - \frac{1}{N-1}\right), \end{aligned} \tag{4.21}$$

which should be compared with the corresponding infinite population expression

$$\kappa_k(r,s|t,u) = \kappa^{r,s,t,u}/n + \kappa^{r,t}\kappa^{s,u}[2]/(n-1).$$

Note that if (4.21) were expressed in terms of products of population k-statistics such as $K^{r,t}K^{s,u}[2]$, it would be necessary to introduce the additional product $K^{r,s}K^{t,u}$ thereby involving a partition not satisfying the connectivity condition in (3.3).

The key to the derivation of formulae such as those given above, is evidently to express multiple products of k-statistics and polykays as a *linear* combination of polykays. Formulae for multiple products of ordinary k-statistics are easy to write down, particularly with the help of the expressions in Exercises 4.5 and 4.7 for the joint cumulants of k-statistics. The following example helps to illustrate the method.

Consider, in an infinite population, the mean value of the product of three covariances. We find

$$\begin{aligned}
E(k^{r,s}k^{t,u}k^{v,w}) = & \\
& \kappa^{r,s}\kappa^{t,u}\kappa^{v,w} + \kappa^{r,s}\kappa_k(t,u|v,w)[3] + \kappa_k(r,s|t,u|v,w) \\
=& \kappa^{r,s}\kappa^{t,u}\kappa^{v,w} + \kappa^{r,s}\{\kappa^{t,u,v,w}/n + \kappa^{t,v}\kappa^{u,w}[2]/(n-1)\}[3] \\
& \quad + \kappa^{r,s,t,u,v}/n^2 + \kappa^{r,t}\kappa^{s,u,v,w}[12]/\{n(n-1)\} \\
& \quad + \kappa^{r,t,v}\kappa^{s,u,w}[4](n-2)/\{n(n-1)^2\} \\
& \quad + \kappa^{r,t}\kappa^{s,v}\kappa^{u,w}[8]/(n-1)^2.
\end{aligned}$$

Evidently, the combination

$$\begin{aligned}
& k^{(r,s)(t,u)(v,w)} + k^{(r,s)(t,u,v,w)}[3]/n + k^{(r,s)(t,v)(u,w)}[6]/(n-1) \\
& + k^{r,s,t,u,v,w}/n^2 + k^{(r,t)(s,u,v,w)}[12]/\{n(n-1)\} \\
& + k^{(r,t,v)(s,u,w)}[4](n-2)/\{n(n-1)^2\} + k^{(r,t)(s,v)(u,w)}[8]/(n-1)^2
\end{aligned} \tag{4.22}$$

is an unbiased estimate of $E(k^{r,s}k^{t,u}k^{v,w})$. It follows immediately by linear independence that (4.22) is identical to the product $k^{r,s}k^{t,u}k^{v,w}$, giving the required multiplication formula. Although it is certainly possible to write down general formulae for the product of two arbitrary polykays, such expressions tend to be rather complicated. For most purposes, a multiplication table is more useful: see Table 4.4, which gives complete products up to fourth degree and selected products up to sixth degree. In the univariate case, more extensive tables and general formulae are given by Wishart (1952), Dwyer & Tracy (1964) and Tracy (1968).

By definition, the third-order joint cumulant of $k^{r,s}$, $k^{t,u}$ and $k^{v,w}$ is equal to

$$\text{ave}(k^{r,s}k^{t,u}k^{v,w}) - K^{r,s}\,\text{ave}(k^{t,u}k^{v,w})[3] + 2K^{r,s}K^{t,u}K^{v,w}.$$

On substituting

$$\text{ave}(k^{t,u}k^{v,w}) = K^{(t,u)(v,w)} + K^{t,u,v,w}/n + K^{(t,v)(u,w)}[2]/(n-1)$$

and simplifying using (4.22), we find that the third cumulant may be written in the form

$$\begin{aligned}
& \alpha_1\left(K^{(r,s)(t,u,v,w)} - K^{(r,s)}K^{(t,u,v,w)}\right)[3] \\
& + \beta_1\left(K^{(r,s)(t,v)(u,w)} - K^{(r,s)}K^{(t,v)(u,w)}\right)[6] \\
& \quad + (n^{-2} - N^{-2})K^{r,s,t,u,v,w} + \gamma_1 K^{(r,t)(s,u,v,w)}[12] \\
& \quad + \left(\frac{n-2}{n(n-1)^2} - \frac{N-2}{N(N-1)^2}\right)K^{(r,t,v)(s,u,w)}[4] \\
& \quad + \left((n-1)^{-2} - (N-1)^{-2}\right)K^{(r,t)(s,v)(u,w)}[8].
\end{aligned} \tag{4.23}$$

The coefficients in this formula are given by

$$\begin{aligned}
\alpha_1 &= n^{-1} - N^{-1}, & \alpha_2 &= \alpha_1 - N^{-1}, \\
\beta_1 &= (n-1)^{-1} - (N-1)^{-1}, & \beta_2 &= \beta_1 - (N-1)^{-1}, \\
\gamma_1 &= \{n(n-1)\}^{-1} - \{N(N-1)\}^{-1}, & \gamma_2 &= \gamma_1 - \{N(N-1)\}^{-1}.
\end{aligned}$$

Table 4.4 *Some multiplication formulae for polykays*

Product	*Linear expression in polykays*
$k^r k^s$	$k^{(r)(s)} + k^{(r,s)}/n$
$k^r k^{s,t}$	$k^{(r)(s,t)} + k^{r,s,t}/n$
$k^r k^{(s)(t)}$	$k^{(r)(s)(t)} + \{k^{(s)(r,t)} + k^{(t)(r,s)}\}/n$
$k^r k^s k^t$	$k^{(r)(s)(t)} + k^{(r)(s,t)}[3]/n + k^{r,s,t}/n^2$
$k^r k^s k^t k^u$	$k^{(r)(s)(t)(u)} + k^{(r)(s)(t,u)}[6]/n + k^{(r,s)(t,u)}[3]/n^2$
	$+ k^{(r)(s,t,u)}[4]/n^2 + k^{r,s,t,u}/n^3$
$k^r k^s k^{t,u}$	$k^{(r)(s)(t,u)} + k^{(r,s)(t,u)}/n + k^{(r)(s,t,u)}[2]/n$
	$+ k^{r,s,t,u}/n^2$
$k^r k^{s,t,u}$	$k^{(r)(s,t,u)} + k^{r,s,t,u}/n$
$k^r k^s k^{(t)(u)}$	$k^{(r)(s)(t)(u)} + k^{(r)(t)(s,u)}[4]/n + k^{(t)(u)(r,s)}/n$
	$+ k^{(r,t)(s,u)}[2]/n^2 + k^{(t)(r,s,u)}[2]/n^2$
$k^r k^{(s)(t)(u)}$	$k^{(r)(s)(t)(u)} + k^{(r,s)(t)(u)}[3]/n$
$k^r k^{(s)(t,u)}$	$k^{(r)(s)(t,u)} + k^{(r,s)(t,u)}/n + k^{(s)(r,t,u)}/n$
$k^{r,s} k^{(t)(u)}$	$k^{(r,s)(t)(u)} + k^{(t)(r,s,u)}[2]/n - k^{(r,t)(s,u)}[2]/\{n(n-1)\}$
$k^{(r)(s)} k^{(t)(u)}$	$k^{(r)(s)(t)(u)} + k^{(r,t)(s)(u)}[4]/n$
	$+ k^{(r,t)(s,u)}[2]/\{n(n-1)\}$
$k^{r,s} k^{t,u}$	$k^{(r,s)(t,u)} + k^{r,s,t,u}/n$
	$+ \{k^{(r,t)(s,u)} + k^{(r,u)(s,t)}\}/(n-1)$
$k^{r,s} k^{t,u} k^{v,w}$	$k^{(r,s)(t,u)(v,w)} + k^{(r,s)(t,u,v,w)}[3]/n$
	$+ k^{(r,s)(t,v)(u,w)}[6]/(n-1) + k^{r,s,t,u,v,w}/n^2$
	$+ k^{(r,t)(s,u,v,w)}[12]/\{n(n-1)\} + k^{(r,t)(s,v)(u,w)}[8]/(n-1)^2$
	$+ k^{(r,t,v)(s,u,w)}[4](n-2)/\{n(n-1)^2\}$
$k^{r,s} k^{t,u,v,w}$	$k^{(r,s)(t,u,v,w)} + k^{r,s,t,u,v,w}/n$
	$+ k^{(r,t)(s,u,v,w)}[8]/(n-1) + k^{(r,t,u)(s,v,w)}[6]/(n-1)$
$k^{r,s} k^{(t,u)(v,w)}$	$k^{(r,s)(t,u)(v,w)} + k^{(r,t)(s,u)(v,w)}[4]/(n-1)$
	$+ k^{(t,u)(r,s,v,w)}[2]/n - k^{(r,t,u)(s,v,w)}[2]/\{n(n-1)\}$

On replacing all products by population polykays, we find the following gratifyingly simple expression involving only connecting partitions, namely

$$\begin{aligned}\operatorname{cum}(k^{r,s}, k^{t,u}, k^{v,w}) = {} & K^{r,s,t,u,v,w}\alpha_1\alpha_2 + K^{(r,t)(s,u,v,w)}[12]\alpha_1\beta_2 \\ & + K^{(r,t,v)(s,u,w)}[4]\{\alpha_1\beta_2 - \beta_1\gamma_2\} \\ & - K^{(r,s,t)(u,v,w)}[6]\alpha_1/(N-1) \\ & + K^{(r,t)(s,v)(u,w)}[8]\beta_1\beta_2. \end{aligned} \tag{4.24}$$

The simplicity of this formula, together with the corresponding one (4.21) for the covariance of $k^{r,s}$ and $k^{t,u}$, should be sufficient to justify the emphasis on polykays.

Wishart (1952), dealing with the univariate case, does not make this final step from (4.23) to

(4.24). As a result, his formulae do not bring out fully the simplicity afforded by expressing the results in terms of polykays alone.

In this final step, it was necessary to find a linear expression in polykays for the product $K^{(r,s)}K^{(t,v)(u,w)}$, which appears in (4.23). The required multiplication formula is given in Table 4.4. Notice that as $N \to \infty$, the fourth term on the right vanishes and the remaining terms converge to the corresponding expression in the infinite population cumulant $\kappa_k(r, s|t, u|v, w)$.

Table 4.5 *Joint cumulants of k-statistics for finite populations*

k-statistics	*Joint cumulant*
k^r, k^s	$\alpha_1 K^{r,s}$
$k^r, k^{s,t}$	$\alpha_1 K^{r,s,t}$
k^r, k^s, k^t	$\alpha_1\alpha_2 K^{r,s,t}$
k^r, k^s, k^t, k^u	$\alpha_1\big(n^{-2} - 6\alpha_1/N\big)K^{r,s,t,u} - 2\alpha_1^2 K^{(r,s)(t,u)}[3]/(N-1)$
$k^r, k^{s,t,u}$	$\alpha_1 K^{r,s,t,u}$
$k^{r,s}, k^{t,u}$	$\alpha_1 K^{r,s,t,u} + \beta_1 K^{(r,t)(s,u)}[2]$
$k^r, k^s, k^{t,u}$	$\alpha_1\alpha_2 K^{r,s,t,u} - \alpha_1 K^{(r,t)(s,u)}[2]/(N-1)$
$k^{r,s}, k^{t,u,v}$	$\alpha_1 K^{r,s,t,u,v} + \beta_1 K^{(r,t,u)(s,v)}[6]$
$k^r, k^{s,t}, k^{u,v,w}$	$\alpha_1\alpha_2 K^{r,s,t,u,v,w} + \alpha_1\beta_2 K^{(r,s,u)(t,v,w)}[6]$
	$+ \alpha_1\beta_2 K^{(r,s,u,v)(t,w)}[6] - \alpha_1 K^{(r,s)(t,u,v,w)}[2]/(N-1)$
	$- \alpha_1\{K^{(r,u)(s,t,v,w)}[3] + K^{(r,u,v)(s,t,w)}[3]\}/(N-1)$
$k^{r,s}, k^{t,u}, k^{v,w}$	Equation (4.24)
$k^{r,s,t}, k^{u,v,w}$	$\alpha_1 K^{r,s,t,u,v,w} + \beta_1 K^{(r,u)(s,t,v,w)}[9]$
	$+ \beta_1 K^{(r,s,u)(t,v,w)}[9]$
	$+ \{n/(n-1)^{(2)} - N/(N-1)^{(2)}\}K^{(r,u)(s,v)(t,w)}[6]$

Table 4.5 gives a selection of joint cumulants of k-statistics for finite populations, including all combinations up to fourth order and selected combinations up to sixth order.

There is, naturally, a fundamental contradiction involved in using formulae such as those in Table 4.5. The purpose of sampling is presumably to learn something about the population of values, perhaps the average value or the variance or range of values, and the first two k-statistics are useful as point estimators. However, in order to set confidence limits, it becomes necessary to know the population k-statistics – something we had hoped simple random sampling would help us avoid computing. The usual procedure is to substitute the estimated k-statistic for the population parameter and to hope that any errors so induced are negligible. Such procedures, while not entirely satisfactory, are perfectly sensible and are easily justified in large samples. However, it would be useful to have a rule-of-thumb to know roughly what extra allowance for sampling variability might be necessary in small to medium samples.

Note that if the objective is to test some fully specified hypothesis, no such contradiction arises. In this case, all population k-statistics are specified by the hypothesis and the problem is to test whether the sample k-statistics are in reasonable agreement with the known theoretical values.

4.7 k-statistics based on least squares residuals

4.7.1 *Multivariate linear regression*

In the previous sections where we dealt with independent and identically distributed observations, it was appropriate to require summary statistics to be invariant under the permutation group. The rationale for this requirement is that the joint distribution is unaffected by permuting the observations and therefore any derived statistics should be similarly unaffected. This argument leads directly to consideration of the symmetric functions. Of course, if departures from the model are suspected, it is essential to look beyond the symmetric functions in order to derive a suitable test statistic. For example, suspected serial correlation might lead the statistician to examine sample autocorrelations: suspected dependence on an auxiliary variable might lead to an examination of specific linear combinations of the response variable. Neither of these statistics is symmetric in the sense understood in Section 4.2. In this section, we consider the case where the mean value of the response is known to depend linearly on given explanatory variables. We must then abandon the notion of symmetric functions as understood in Section 4.2 and look for statistics that are invariant in a different, but more appropriate, sense.

In previous sections we wrote Y_i^r for the rth component of the ith independent observation. Subscripts were used to identify the observations, more out of convenience and aesthetic considerations, than out of principle or necessity. Only permutation transformations were considered and the k-statistics are invariant under this group. Thus, there is no compelling argument for using superscripts rather than subscripts. In this section, however, we consider arbitrary linear transformations and it is essential to recognize that Y and its cumulants are contravariant both with respect to transformation of the individual observations and with respect to the components. In other words, both indices must appear as superscripts.

Suppose then, that the rth component of the ith response variable has expectation

$$E(Y^{r;i}) = \omega^{\alpha;i}\kappa_\alpha^r,$$

where $\omega^{\alpha;i}$ is an $n \times q$ array of known constants and κ_α^r is a $q \times p$ array of unknown parameters to be estimated. When we revert to matrix notation we write

$$E(\mathbf{Y}) = \mathbf{X}\beta,$$

this being the more common notation in the literature on linear models. Here, however, we choose to write κ_α^r for the array of regression coefficients in order to emphasize the connection with cumulants and k-statistics. In addition to the above assumption regarding the mean value, we assume that there are available known arrays $\omega^{i,j}$, $\omega^{i,j,k}$ and so on, such that

$$\begin{aligned}\operatorname{cov}(Y^{r;i}, Y^{s;j}) &= \kappa^{r,s}\omega^{i,j},\\ \operatorname{cum}(Y^{r;i}, Y^{s;j}, Y^{t;k}) &= \kappa^{r,s,t}\omega^{i,j,k}\end{aligned}$$

and so on for the higher-order cumulants. The simplest non-trivial example having this structure arises when the ith observation is the sum of m_i independent and identically distributed, but unrecorded, random variables. In this case, the ω-arrays of order two and higher take the value m_i on the main diagonal and zero elsewhere.

The notation carries with it the implication that any nonsingular linear transformation

$$\bar{Y}^{r;i} = a_j^i Y^{r;j} \tag{4.25}$$

applied to Y has a corresponding effect on the arrays $\omega^{\alpha;i}$, $\omega^{i,j}$, $\omega^{i,j,k}$,..., namely

$$\bar{\omega}^{\alpha;i} = a_j^i\omega^{\alpha;j}, \quad \bar{\omega}^{i,j} = a_k^i a_l^j\omega^{k,l}, \quad \bar{\omega}^{i,j,k} = a_l^i a_m^j a_n^k\omega^{l,m,n}$$

and so on. In other words, the ω-arrays are assumed to behave as contravariant tensors under the action of the general linear group (4.25).

In what follows, it is assumed that the estimation problem is entirely invariant under the group of transformations (4.25). In some ways, this may seem a very natural assumption because, for example,

$$E(\bar{Y}^{r;i}) = \bar{\omega}^{\alpha;i}\kappa^r_\alpha, \qquad \mathrm{cov}(\bar{Y}^{r;i}, \bar{Y}^{s;j}) = \bar{\omega}^{i,j}\kappa^{r,s}$$

and so on, so that the definition of the cumulants is certainly invariant under the group. However, it is important to emphasize the consequences of the assumption that *all* linear transformations of the observations are to be treated on an equal footing and that no particular scale has special status. We are forced by this assumption to abandon the concepts of interaction and replication, familiar in the analysis of variance, because they are not invariant under the group (4.25). See Exercise 4.22. This is not to be construed as criticism of the notions of interaction and replication, nor is it a criticism of our formulation in terms of the general linear group rather than some smaller group. Instead, it points to the limitations of our specification and emphasizes that, in any given application, the appropriateness of the criteria used here must be considered carefully. In simple cases, it may be possible to find a better formulation in terms of a group more carefully tailored to the problem.

Often it is appropriate to take $\omega^{i,j} = \delta^{ij}$, $\omega^{i,j,k} = \delta^{ijk}$ and so on, but, since no particular significance attaches to the diagonal arrays, we shall suppose instead that $\omega^{i,j}$, $\omega^{i,j,k}, \ldots$ are arbitrary arrays and that $\omega^{i,j}$ has full rank.

Our aim, then, is to construct unbiased estimates of κ^r_α, $\kappa^{r,s}$, $\kappa^{r,s,t}, \ldots$ that are invariant under transformations (4.25). Note that it was necessary to introduce the arrays $\omega^{\alpha;i}$, $\omega^{i,j}, \ldots$ as tensors in order that the definition of the cumulants be unaffected by linear transformation. Otherwise, if the cumulants were not invariant, it would not be sensible to require invariant estimates.

Let $\omega_{i,j}$ be the matrix inverse of $\omega^{i,j}$. In matrix notation, $\omega_{i,j}$ is written as $\mathbf{W}$. By assumption, $\omega^{i,j}$ has full rank, but even in the rank deficient case, the choice of inverse does not matter (see Exercise 2.24). Now define

$$\omega^\alpha_i = \omega_{i,j}\omega^{\alpha;j}, \qquad \omega_{i,j,k} = \omega_{i,l}\omega_{j,m}\omega_{k,n}\omega^{l,m,n}$$

and so on, lowering the indices in the usual way. In the calculations that follow, it is convenient to have a concise notation for the $q \times q$ array

$$\lambda^{\alpha,\beta} = \omega^{\alpha;i}\omega^{\beta;j}\omega_{i,j} = \omega^\alpha_i\omega^\beta_j\omega^{i,j},$$

and the $q \times q \times q$ array

$$\lambda^{\alpha,\beta,\gamma} = \omega^\alpha_i\omega^\beta_j\omega^\gamma_k\omega^{i,j,k}.$$

In matrix notation, $\lambda^{\alpha,\beta}$ is usually written as $\mathbf{X}^T\mathbf{W}\mathbf{X}$, and $\lambda_{\alpha,\beta}$ as $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$, but there is no convenient matrix notation for $\lambda^{\alpha,\beta,\gamma}$.

A straightforward calculation shows that

$$E(\omega^\alpha_i Y^{r;i}) = \lambda^{\alpha,\beta}\kappa^r_\beta$$

and hence, by matrix inversion, that

$$k^r_\alpha = \lambda_{\alpha,\beta}\omega^\beta_i Y^{r;i} \tag{4.26}$$

is an unbiased estimate of κ^r_α, invariant under linear transformations (4.25). In matrix notation, (4.26) becomes

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

which is instantly recognized as the weighted least squares estimate of the regression coefficients. Note that the equally weighted estimate is not invariant under linear transformation.

The notation used in (4.26) is chosen to emphasize that the weighted least squares estimate is in fact a k-statistic of order one. Although it is in a sense the obvious invariant estimate of κ^r_α, the k-statistic, k^r_α is not unique except in rather special cases. See Exercise 4.19.

The second- and third-order cumulants of k^r_α are

$$\begin{aligned}\operatorname{cov}(k^r_\alpha, k^s_\beta) &= \lambda_{\alpha,\beta}\kappa^{r,s}\\ \operatorname{cum}(k^r_\alpha, k^s_\beta, k^t_\gamma) &= \lambda_{\alpha,\beta,\gamma}\kappa^{r,s,t}\end{aligned}$$

where the indices of $\lambda^{\alpha,\beta,\gamma}$ have been lowered in the usual way by multiplying by $\lambda_{\alpha,\beta}$ three times. Similar expressions hold for the higher-order cumulants.

In the case of quadratic expressions, we write

$$\begin{aligned}E(\omega_{i,j}Y^{r;i}Y^{s;j}) &= n\kappa^{r,s} + \lambda^{\alpha,\beta}\kappa^r_\alpha\kappa^s_\beta\\ E(\omega^\alpha_i\omega^\beta_j Y^{r;i}Y^{s;j}) &= \lambda^{\alpha,\beta}\kappa^{r,s} + \lambda^{\alpha,\gamma}\lambda^{\beta,\delta}\kappa^r_\gamma\kappa^s_\delta.\end{aligned}$$

These equations determine the k-statistic $k^{r,s}$ and the polykays $k^{(r)(s)}_{(\alpha)(\beta)}$, which are unbiased estimates of the products $\kappa^r_\alpha\kappa^s_\beta$. There are as many equations as there are k-statistics and polykays. We find that

$$(n-q)k^{r,s} = \{\omega_{i,j} - \lambda_{\alpha,\beta}\omega^\alpha_i\omega^\beta_j\}Y^{r;i}Y^{s;j},$$

which is the usual weighted residual sum of squares and products matrix, and

$$k^{(r)(s)}_{(\alpha)(\beta)} = k^r_\alpha k^s_\beta - \lambda_{\alpha,\beta}k^{r,s}.$$

These k-statistics and polykays are not unique except in very special cases: see Exercise 4.19. Among the estimates that are invariant under the general linear group, it appears that the ks do not have minimum variance unless the observations are normally distributed and, even then, additional conditions are required for exact optimality. See Exercise 4.23. Under a smaller group such as the permutation group, which admits a greater number of invariant estimates, they are neither unique nor do they have minimum variance among invariant estimates. Conditions under which $k^{r,s}$ has minimum variance in the univariate case are given by Plackett (1960, p. 40) and Atiqullah (1962). See also Exercises 4.22–4.26 for a more general discussion relating to higher-order ks.

When dealing with higher-order k-statistics, it is more convenient to work with the unstandardized residuals

$$\begin{aligned}R^{r;i} = Y^{r;i} - \omega^{\alpha;i}k^r_\alpha &= \{\delta^i_j - \omega^{\alpha;i}\lambda_{\alpha,\beta}\omega^\beta_j\}Y^{r;j}\\ &= \rho^i_j Y^{r;j}.\end{aligned}$$

In matrix notation, this becomes

$$\mathbf{R} = (\mathbf{I}-\mathbf{H})\mathbf{Y}, \quad \text{where} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W},$$

the projection matrix producing fitted values, is not symmetrical unless $\mathbf{W}$ is a multiple of $\mathbf{I}$. Working with residuals is entirely analogous to working with central moments, rather than moments about the origin, in the single sample case. Note, however, that polykays having a unit part are not estimable from the residuals because the distribution of $R^{r;i}$ does not depend on κ^r_α.

The least squares residuals have cumulants

$$\begin{aligned}E(R^{r;i}) &= 0,\\ \operatorname{cov}(R^{r;i}, R^{s;j}) &= \rho^i_k\rho^j_l\omega^{k,l}\kappa^{r,s} = \rho^{i,j}\kappa^{r,s}\\ \operatorname{cum}(R^{r;i}, R^{s;j}, R^{t;k}) &= \rho^i_l\rho^j_m\rho^k_n\omega^{l,m,n}\kappa^{r,s,t} = \rho^{i,j,k}\kappa^{r,s,t}\end{aligned}$$

and so on. Effectively, the arrays $\omega^{i,j}$, $\omega^{i,j,k}, \ldots$ are transformed to $\rho^{i,j}$, $\rho^{i,j,k}, \ldots$ by projection on to the residual space and its associated direct products. Note that $\omega_{i,j}$ is a generalized inverse of $\rho^{i,j}$, and hence we may write

$$\rho_{i,j} = \omega_{i,k}\omega_{j,l}\rho^{k,l}, \qquad \rho_{i,j,k} = \omega_{i,l}\omega_{j,m}\omega_{k,n}\rho^{l,m,n}$$

and so on, in agreement with the definitions given previously. Evidently, $\rho_{i,j}$ is the Moore-Penrose inverse of $\rho^{i,j}$. In the discussion that follows, we may use $\rho_{i,j}$ and $\omega_{i,j}$ interchangeably to lower indices, both being generalized inverses of $\rho^{i,j}$. Thus, the invariant unbiased estimates of the second- and third-order cumulants are

$$\begin{aligned} n_2 k^{r,s} &= \rho_{i,j} R^{r;i} R^{s;j} = \omega_{i,j} R^{r;i} R^{s;j} \\ n_3 k^{r,s,t} &= \rho_{i,j,k} R^{r;i} R^{s;j} R^{t;k} = \omega_{i,j,k} R^{r;i} R^{s;j} R^{t;k} \end{aligned} \tag{4.27}$$

where $n_2 = n - q$, $n_3 = \omega_{i,j,k}\omega^{l,m,n}\rho^i_l \rho^j_m \rho^k_n$, which reduces to $\sum_{ij}(\rho^{i,j})^3$ when $\omega^{i,j} = \delta^{ij}$ and $\omega^{i,j,k} = \delta^{i,j,k}$. In matrix notation, $n_2 k^{r,s}$ is just $\mathbf{R}^T\mathbf{W}\mathbf{R}$ or equivalently, $\mathbf{R}^T\mathbf{W}\mathbf{Y}$, but there is no simple matrix notation for $n_3 k^{r,s,t}$, the residual sum of cubes and products array.

In the case of cumulants of degree four, we write

$$\begin{aligned} E(\omega_{i,j,k,l} R^{r;i} R^{s;j} R^{t;k} R^{u;l}) &= n_4 \kappa^{r,s,t,u} + n_{22}\kappa^{r,s}\kappa^{t,u}[3] \\ E(\omega_{i,j}\omega_{k,l} R^{r;i} R^{s;j} R^{t;k} R^{u;l}) &= n_{22}\kappa^{r,s,t,u} + n_2^2 \kappa^{r,s}\kappa^{t,u} \\ &\qquad + n_2 \kappa^{r,t}\kappa^{s,u}[2], \end{aligned}$$

where the coefficients are defined by

$$\begin{aligned} n_4 &= \rho^{i,j,k,l}\rho_{i,j,k,l} = \omega_{i,j,k,l}\rho^{i,j,k,l} \\ n_{22} &= \rho^{i,j,k,l}\rho_{i,j}\rho_{k,l} = \omega_{i,j,k,l}\rho^{i,j}\rho^{k,l}. \end{aligned}$$

In the particular case where $\omega^{i,j} = \delta^{ij}$, the coefficients reduce to $n_4 = \sum_{ij}(\rho^{i,j})^4$ and $n_{22} = \sum_i (\rho^{i,i})^2$.

Matrix inversion gives the following invariant unbiased estimates of $\kappa^{r,s,t,u}$ and $\kappa^{r,s}\kappa^{t,u}$.

$$\begin{aligned} \Delta_1 k^{r,s,t,u} &= \{n_2(n_2+2)\omega_{i,j,k,l} - n_{22}\omega_{i,j}\omega_{k,l}[3]\} R^{r;i} R^{s;j} R^{t;k} R^{u;l} \\ \Delta_2 k^{(r,s)(t,u)} &= \{-n_2(n_2-1)n_{22}\omega_{i,j,k,l} + (n_2(n_2+1)n_4 - 2n_{22}^2)\omega_{i,j}\omega_{k,l} \\ &\qquad + (n_{22}^2 - n_2 n_4)\omega_{i,k}\omega_{j,l}[2]\} R^{r;i} R^{s;j} R^{t;k} R^{u;l} \end{aligned}$$

where

$$\Delta_1 = n_2 n_4 (n_2 - 1) + 3(n_2 n_4 - n_{22}^2) \quad \text{and} \quad \Delta_2 = n_2(n_2-1)\Delta_1.$$

The final expression above simplifies to some extent when $r = s = t = u$, as is always the case in univariate regression problems. In particular, the final two terms can be combined and $n_2(n_2 - 1)$ may be extracted as a factor.

Note that, in the case where the observations are independent with identical cumulants of order two and higher, these k-statistics and polykays are symmetric functions of the least squares residuals, even though the residuals are neither independent nor identically distributed. It is this feature that is the source of non-uniqueness of k-statistics and polykays in regression problems.

4.7.2 *Univariate linear regression*

Some condensation of notation is inevitable in the univariate case, but here we make only the minimum of adjustment from the notation of the previous section. However, we do assume that the observations are independent with cumulants

$$\begin{aligned}\mathrm{cov}(Y^i, Y^j) &= \kappa_2\delta^{ij}\\ \mathrm{cum}(Y^i, Y^j, Y^k) &= \kappa_3\delta^{ijk}\end{aligned}$$

and so on. It follows that the projection matrix producing fitted values, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, unusually for a projection matrix, is symmetrical and $\kappa_2\mathbf{H}$ is also the covariance matrix of the fitted values. The residual projection matrix is also symmetrical with elements $\rho_{i,j} = \delta_{ij} - h_{ij}$ and it is immaterial whether we use subscripts or superscripts.

If we write

$$S_r = \sum (R^i)^r$$

for the rth power sum of the residuals, it follows that

$$\begin{aligned}k_2 &= S_2/n_2\\ k_3 &= S_3/n_3\\ k_4 &= \{n_2(n_2+2)S_4 - 3n_{22}S_2^2\}/\Delta_1\\ k_{22} &= \{n_4S_2^2 - n_{22}S_4\}/\Delta_1\end{aligned}$$

where

$$n_2 = n - q, \quad n_r = \sum_{ij}(\rho_{i,j})^r \quad \text{and} \quad n_{22} = \sum_i (\rho_{i,i})^2,$$

and Δ_1 is given at the end of the previous section. These are the unbiased estimates of residual cumulants and products of residual cumulants up to fourth order. In other words, $E(k_4) = \kappa_4$ and $E(k_{22}) = \kappa_2^2$.

Little simplification of the above formulae seems to be possible in general. However, in the case of designs that are quadratically balanced, the residuals have identical marginal variances, though they are not usually exchangeable, and $h_{ii} = q/n$, $\rho_{i,i} = 1 - q/n$. It follows then that $n_{22} = n(1-q/n)^2$. In addition, if n is large, we may wish to avoid the extensive calculations involved in computing n_3 and n_4. The following approximations are often satisfactory if n is large and if the design is not grossly unbalanced:

$$n_3 \simeq n(1-q/n)^3, \quad n_4 \simeq n(1-q/n)^4, \tag{4.28}$$

the errors of approximation being typically of order $O(n^{-1})$ and $O(n^{-2})$ respectively. Evidently, the above approximation is a lower bound for n_4, but not for n_3 The upper bounds, $n_3, n_4 \le n - q$, are attained if $n-q$ of the observations are uninformative for the regression parameters. In practice, n_3 and n_4 are typically closer to the lower limits than to the upper limits. It follows then that

$$n_2n_4 - n_{22}^2 \simeq -qn_4$$

and hence that

$$\Delta_1 \simeq n_4(n_2(n_2-1) - 3q).$$

The above approximations are often adequate even if the design is not quadratically balanced and even if n is not particularly large. For example, in simple linear regression, with 20 observations and with a single quantitative covariate whose values are equally spaced, the error incurred in using the approximations is less than 2%. In less favourable circumstances, other approximations might be preferred. For example, for $r \geq 3$,

$$n_r \simeq \sum (\rho_{i,i})^r, \tag{4.29}$$

which reduces to the earlier approximation if the design is quadratically balanced. On the basis of limited numerical comparisons, it appears that, on balance, (4.29) is usually more accurate than (4.28), at least when q is small relative to n.

For further details, including more efficient estimates of κ_3 and κ_4, see McCullagh & Pregibon (1987) and also, Exercise 4.22.

4.8 Bibliographic notes

The estimation of means and variances has been a concern of statisticians from the earliest days, and consequently the literature is very extensive. No attempt will be made here to survey the early work concerned with sample moments and the joint moments of sample moments.

The idea of simplifying calculations by estimating cumulants rather than moments is usually attributed to Fisher (1929), although similar ideas were put forward earlier by Thiele (1897). Thiele's book was published in English in 1903 and the work was cited by Karl Pearson. It is curious that, in 1929, Fisher was apparently unaware of Thiele's earlier contribution.

Following the publication of Fisher's landmark paper, subsequent work by Wishart (Fisher & Wishart, 1931) and Kendall (1940a,b,c), seems directed at demonstrating the correctness of Fisher's combinatorial method for evaluating pattern functions. Fisher's algorithm must have appeared mystifying at the time: it has weathered the passage of half a century extraordinarily well.

Cook (1951) gives extensive tables of the joint cumulants of multivariate k-statistics. Kaplan (1952), using a form of tensor notation, gives the same formulae, but more succinctly. For further discussion, including explicit formulae in the univariate and bivariate case, see Kendall & Stuart (1977, Chapters 12,13).

The connection with finite population sampling was made by Tukey (1950, 1956a), who also bears responsibility for the term *polykay*: see Kendall & Stuart (1977, Section 12.22), where the less informative term l-statistic is used. Calculations similar to Tukey's were previously given by Dressel (1940), though without any suggestion that the results might be useful for sampling theory calculations. See also Irwin & Kendall (1943–45) For later developments, see Dwyer & Tracy (1964) and Tracy (1968).

The reader who finds the notation used in this chapter cumbersome or impenetrable should be warned that notational problems get much worse for symmetric functions derived from two-way or higher-order arrays. In a one-way layout, we talk of the *between groups variance* and the *within groups variance.* A similar decomposition exists for the higher-order statistics, so that we may talk of the *between groups skewness*, the *within groups skewness* and also the *cross skewness.* These are the three k-statistics of the third degree. In the case of a two-way design with one observation per cell, there are three variance terms and five third-order k-statistics, one of which is Tukey's (1949) non-additivity statistic. Such statistics are called *bi-kays.* There are of course, the corresponding *bi-polykays*, also derived from a two-way array, not to mention *poly-bi-kays* and *poly-poly-kays*, which are much worse. Any reader who is interested in such matters should read Hooke (1956a,b) or Tukey (1956b). For those intrepid souls whose appetites are whetted by this fare, we heartily recommend Speed (1986a,b).

Anscombe (1961) discusses the computation of and the applications of k-statistics in detecting departures from the usual linear model assumptions. There has been much subsequent work in this vein. See, for example, Bickel (1978), Hinkley (1985) or McCullagh & Pregibon (1987) for some recent developments.

4.9 Further results and exercises 4

4.1 Show that

$$k^{r,s} = n^{-1}\sum\nolimits_{ij}\phi^{ij}Y_i^rY_j^s$$

is the usual sample covariance matrix based on n independent and identically distributed observations. Show that $k^{r,s}$ is unbiased for $\kappa^{r,s}$.

4.2 By applying the criteria of unbiasedness and symmetry to $k^{r,s,t}$ and $k^{r,s,t,u}$, derive the expression (4.7) for the coefficients ϕ^{ijk} and ϕ^{ijkl}.

4.3 Find the mean value of the following expressions:

$$\sum_i Y_i^rY_i^sY_i^t, \qquad \sum_{i\neq j} Y_i^rY_j^sY_j^t, \quad \text{and} \quad \sum_{i\neq j\neq k} Y_i^rY_j^sY_k^t.$$

Hence find the symmetric functions that are unbiased estimates of κ^{rst}, $\kappa^r\kappa^{st}$ and $\kappa^r\kappa^s\kappa^t$. By combining these estimates in the appropriate way, find the symmetric functions that are unbiased estimates of $\kappa^{r,s,t}$, $\kappa^{r,st}$ and $\kappa^r\kappa^{s,t}$.

4.4 Show that the expressions for ordinary k-statistics in terms of power sums are

$$\begin{aligned}
k^{r,s} &= \{k^{rs} - k^rk^s\}n/(n-1)\\
k^{r,s,t} &= \big\{k^{rst} - k^rk^{st}[3] + 2k^rk^sk^t\big\}n^2/(n-1)^{(2)}\\
k^{r,s,t,u} &= \big\{(n+1)k^{rstu} - (n+1)k^rk^{stu}[4] - (n-1)k^{rs}k^{tu}[3]\\
&\quad + 2nk^rk^sk^{tu}[6] - 6nk^rk^sk^tk^u\big\}n^2/(n-1)^{(3)},
\end{aligned}$$

Kaplan (1952).

4.5 Show that

$$\sum\nolimits_{ij}(\phi^{ij})^2 = n^2/(n-1).$$

Hence derive the following joint cumulants of k-statistics:

$$\begin{aligned}
\kappa_k(r|s) &= \kappa^{r,s}/n, \qquad\qquad \kappa_k(r|s,t) = \kappa^{r,s,t}/n,\\
\kappa_k(r|s|t) &= \kappa^{r,s,t}/n^2, \qquad\qquad \kappa_k(r|s,t,u) = \kappa^{r,s,t,u}/n,\\
\kappa_k(r,s|t,u) &= \kappa^{r,s,t,u}/n + \kappa^{r,t}\kappa^{s,u}[2]/(n-1),\\
\kappa_k(r|s|t,u) &= \kappa^{r,s,t,u}/n^2, \qquad \kappa_k(r|s,t,u,v) = \kappa^{r,s,t,u,v}/n,\\
\kappa_k(r,s|t,u,v) &= \kappa^{r,s,t,u,v}/n + \kappa^{r,t,u}\kappa^{s,v}[6]/(n-1),\\
\kappa_k(r|s|t,u,v) &= \kappa^{r,s,t,u,v}/n^2,\\
\kappa_k(r|s,t|u,v) &= \kappa^{r,s,t,u,v}/n^2 + \kappa^{r,s,u}\kappa^{t,v}[4]/\{n(n-1)\}\\
\kappa_k(r|s|t|u,v) &= \kappa^{r,s,t,u,v}/n^3
\end{aligned}$$

Show explicitly that the term $\kappa^{r,t,u}\kappa^{s,v}$ does not appear in the third cumulant, $\kappa_k(r|s|t,u,v)$.

4.6 By considering separately the five distinct index patterns, show that

$$\begin{aligned}
\sum_{ijk}\phi^{ij}\phi^{ik}\phi^{jk} &= n + \frac{3n(n-1)}{(n-1)^2} - \frac{n(n-1)(n-2)}{(n-1)^3}\\
&= n^3/(n-1)^2.
\end{aligned}$$

4.7 Using the pattern functions given in Table 4.1, derive the following joint cumulants:

$$\begin{aligned}\kappa_k(r,s|t,u,v,w) &= \kappa^{r,s,t,u,v,w}/n + \kappa^{r,t}\kappa^{s,u,v,w}[8]/(n-1)\\ &\quad + \kappa^{r,t,u}\kappa^{s,v,w}[6]/(n-1),\\ \kappa_k(r,s|t,u|v,w) &= \kappa^{r,s,t,u,v,w}/n^2 + \kappa^{r,t}\kappa^{s,u,v,w}[12]/\{n(n-1)\}\\ &\quad + \kappa^{r,t,v}\kappa^{s,u,w}[4](n-2)/\{n(n-1)^2\}\\ &\quad + \kappa^{r,t}\kappa^{s,v}\kappa^{u,w}[8]/(n-1)^2,\\ \kappa_k(r,s,t|u,v,w) &= \kappa^{r,s,t,u,v,w}/n + \kappa^{r,u}\kappa^{s,t,v,w}[9]/(n-1)\\ &\quad + \kappa^{r,s,u}\kappa^{t,v,w}[9]/(n-1)\\ &\quad + \kappa^{r,u}\kappa^{s,v}\kappa^{t,w}[6]n/(n-1)^{(2)}\end{aligned}$$

Kaplan (1952). Compare these formulae with those given in Table 4.5 for finite populations.

4.8 Show that joint cumulants of ordinary k-statistics in which the partition contains a unit block, e.g. $\kappa_k(r|s,t|u,v,w)$, can be found from the expression in which the unit block is deleted, by dividing by n and adding the extra index at all possible positions. Hence, using the expression given above for $\kappa_k(s,t|u,v,w)$, show that

$$\begin{aligned}\kappa_k(r|s,t|u,v,w) = \kappa^{r,s,t,u,v,w}/n^2 &+ \kappa^{r,s,u,v}\kappa^{t,w}[6]/\{n(n-1)\}\\ &+ \kappa^{r,s,u}\kappa^{t,v,w}[6]/\{n(n-1)\}.\end{aligned}$$

4.9 Show that as $n \to \infty$, every array of k-statistics of fixed order has a limiting joint normal distribution when suitably standardized. Hence show in particular, that, for large n,

$$\begin{aligned}n^{1/2}(k^r - \kappa^r) &\sim N_p(0, \kappa^{r,s})\\ n^{1/2}(k^{r,s} - \kappa^{r,s}) &\sim N_{p^2}(0, \tau^{rs,tu}),\end{aligned}$$

where $\tau^{rs,tu} = \kappa^{r,s,t,u} + \kappa^{r,t}\kappa^{s,u}[2]$ has rank $p(p+1)/2$.

4.10 Under the assumption that the data have a joint normal distribution, show that $n^{1/2}k^{r,s,t,u}$ has a limiting covariance matrix given by

$$n\,\mathrm{cov}(k^{i,j,k,l}, k^{r,s,t,u}) \;\to\; \kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa^{l,u}[4!].$$

Hence show that $n^{1/2}p\bar{r}_4$ has a limiting normal distribution with mean zero and variance $8p^2 + 16p$.

4.11 Under the assumptions of the previous exercise, show that

$$n\ \mathrm{cov}(k^{i,j,k}, k^{r,s,t}) \;\to\; \kappa^{i,r}\kappa^{j,s}\kappa^{k,t}[3!]$$

as $n \to \infty$. Hence show that, for large n,

$$\begin{aligned}\frac{np\bar{r}_{13}^2}{2(p+2)} &\sim \chi^2_p\\ \frac{np\bar{r}_{23}^2}{6} &\sim \chi^2_{p(p+1)(p+2)/6}.\end{aligned}$$

4.12 By considering the symmetric group, i.e. the group comprising all $n \times n$ permutation matrices, acting on $Y_1, \ldots, Y_n$, show that every invariant polynomial function of degree k is expressible as a *linear* combination of polykays of degree k.

4.13 Show that

$$\sum_{i,j=1}^{n} Y_i^r Y_j^s = nk^{rs} + n(n-1)k^{(r)(s)}.$$

Hence deduce that, under simple random sampling, the average value over all samples of $n^{-2}\sum_{i,j=1}^{n} Y_i^r Y_j^s$ is

$$n^{-1}K^{rs} + (1-n^{-1})K^{(r)(s)} = n^{-1}K^{r,s} + K^{(r)(s)}$$

while the same function calculated in the population is

$$N^{-2}\sum_{i,j=1}^{N} Y_i^r Y_j^s = N^{-1}K^{r,s} + K^{(r)(s)}.$$

4.14 Derive the following multiplication formulae for k-statistics and polykays

$$\begin{aligned}
k^r k^s &= k^{(r)(s)} + k^{(r,s)}/n \\
k^r k^{s,t} &= k^{(r)(s,t)} + k^{r,s,t}/n \\
k^r k^{(s)(t)} &= k^{(r)(s)(t)} + \{k^{(s)(r,t)} + k^{(t)(r,s)}\}/n \\
k^r k^s k^t &= k^{(r)(s)(t)} + k^{(r)(s,t)}[3]/n + k^{r,s,t}/n^2
\end{aligned}$$

4.15 Derive the following multiplication formulae for k-statistics and polykays

$$\begin{aligned}
k^{r,s}k^{t,u} &= k^{(r,s)(t,u)} + k^{r,s,t,u}/n \\
&\qquad + \{k^{(r,t)(s,u)} + k^{(r,u)(s,t)}\}/(n-1) \\
k^{r,s}k^{(t,u)(v,w)} &= k^{(r,s)(t,u)(v,w)} + k^{(r,t)(s,u)(v,w)}[4]/(n-1) \\
&\qquad + k^{(t,u)(r,s,v,w)}[2]/n - k^{(r,t,u)(s,v,w)}[2]/\{n(n-1)\}
\end{aligned}$$

4.16 Using the multiplication formulae given in the previous two exercises, derive the finite population joint cumulants listed in Table 4.3.

4.17 By considering the sample covariance matrix of the observed variables and their pairwise products, show that

$$k^{rs,tu} - k^{rs,i}k^{tu,j}k_{i,j},$$

regarded as a $p^2 \times p^2$ symmetric matrix, is non-negative definite. Hence deduce that

$$k^{r,s,t,u} - k^{r,s,i}k^{t,u,j}k_{i,j} + \{k^{(r,t)(s,u)} + k^{(r,u)(s,t)}\}n/(n-2)$$

is also non-negative definite when regarded as a matrix whose rows are indexed by (r,s) and columns by (t,u). Hence derive lower bounds for $\bar{r}_4 - \bar{r}_{13}^2$ and $\bar{r}_4 - \bar{r}_{23}^2$.

4.18 By expanding the product of several power sums along the lines suggested in Section 4.6.2 for products of k-statistics, show that

$$\mathbf{K} = \mathbf{HP}$$

in the notation of (4.20), where $\mathbf{P}$ is a vector of polykays, $\mathbf{K}$ is a vector of power sum products and $\mathbf{H}$ has components

$$h_{ij} = n^l \qquad (l = |\Upsilon_i \vee \Upsilon_j| - |\Upsilon_i|),$$

where $|\Upsilon|$ denotes the number of blocks of the partition. Show also that $l \leq 0$ and $l = 0$ if and only if Υ_j is a sub-partition of Υ_i. Hence prove that as $n \to \infty$, the components of $\mathbf{P}$ are the Möbius transform of the components of $\mathbf{K}$.

4.19 Show, in the notation of Section 4.7.1, that if the scalar $n_{22} = \rho_{i,j,k,l}\rho^{i,j}\rho^{k,l}$ is non-zero, then

$$n_{22}^{-1}\rho_{i,j,k,l}\rho^{k,l}R^{r;i}R^{s;j}$$

is an invariant unbiased estimate of $\kappa^{r,s}$. Under what circumstances is the above estimate the same as $k^{r,s}$?

4.20 Show that the covariance of $k^{r,s}$ and $k^{t,u}$, as defined in (4.27), is

$$\mathrm{cov}(k^{r,s}, k^{t,u}) = n_{22}\kappa^{r,s,t,u}/n_2^2 + \kappa^{r,t}\kappa^{s,u}[2]/n_2,$$

which reduces to

$$\kappa^{r,s,t,u}/n + \kappa^{r,t}\kappa^{s,u}[2]/(n-q)$$

for quadratically balanced designs. Hence show that, in the univariate notation of Section 4.7.2,

$$n_{22}k_4/n_2^2 + 2k_{22}/n_2$$

is an unbiased estimate of $\mathrm{var}(k_2)$.

4.21 Show that, for the ordinary linear regression problem with an intercept and one dependent variable, x, that

$$n_{22} = (n-1)(n-3)/n + \sum(x_i - \bar{x})^4 / \Big(\sum(x_i - \bar{x})^2\Big)^2$$

and that, for equally spaced x-values, this reduces to

$$n_{22} = (n-2)^2/n + 4/(5n) + O(n^{-2})$$

in reasonable agreement with the approximations of Section 4.7.2. Show more generally, that the discrepancy between n_{22} and the approximation $(n-2)^2/n$, is a function of the standardized fourth cuumulant of the x-values. Find this function explicitly.

4.22 In the notation of Section 4.7.2, show that, when third- and higher-order cumulants are neglected, the cubes of the least squares residuals have covariance matrix $\mathrm{cov}(R^iR^jR^k,\ R^lR^mR^n)$ given by

$$\kappa_2^3\{\rho^{i,j}\rho^{k,l}\rho^{m,n}[9] + \rho^{i,l}\rho^{j,m}\rho^{k,n}[6]\},$$

here taken to be of order $n^3 \times n^3$. Show that, if $\nu = n - p$ is the rank of $\rho^{i,j}$, then

$$w_{ijk,lmn} = \rho_{i,l}\rho_{j,m}\rho_{k,n}[6]/36 - \rho_{i,j}\rho_{k,l}\rho_{m,n}[9]/\{18(\nu+4)\}$$

is the Moore-Penrose inverse matrix. Hence prove that

$$l_3 = \frac{k_3 - 3k_2\bar{R}\,n(n-p)/\{n_3(n-p+4)\}}{1 - 3\sum_{ij}\rho_{i,j}\rho_{i,i}\rho_{j,j}/\{n_3(n-p+4)\}}$$

where $n\bar{R} = \sum\rho_{i,i}R^i$, is unbiased for κ_3 and, under normality, has minimum variance among homogeneous cubic forms in the residuals, (McCullagh & Pregibon, 1987).

4.23 Deduce from the previous exercise that if the vector having components $\rho_{i,i}$ lies in the column space of the model matrix $\mathbf{X}$, then $l_3 \equiv k_3$. More generally, prove that if the constant vector lies in the column space of $\mathbf{X}$, then

$$n^{1/2}(l_3 - k_3) = O_p(n^{-1})$$

for large n under suitably mild limiting conditions on $\mathbf{X}$. Hence, deduce that k_3 is nearly optimal under normality.

4.24 Suppose $n = 4$, $p = 1$, $y = (1.2, 0.5, 1.3, 2.7)$, $x = (0, 1, 2, 3)$. Using the results in the previous two exercises, show that $k_2 = 0.5700$, $k_3 = 0.8389$, with variance $6.59\sigma^6$ under normality, and $l_3 = 0.8390$ with variance $4.55\sigma^6$ under normality. Compare with Anscombe (1961, p.15) and Pukelsheim (1980, p.110). Show also that $k_{22} = 0.6916$ and $k_4 = 0.7547$.

4.25 Repeat the calculations of Exercise 4.22, but now for l_4 and l_{22}, the optimal unbiased estimates of the fourth cumulant and the square of the second cumulant. Show that l_4 and l_{22} are linear combinations of

$$S_2^2 \quad \text{and} \quad S_4 - 6S_2 \sum \rho_i^i R_i^2/(n_2 + 6).$$

Deduce that, in the quadratically balanced case, k_4 and k_{22} are the optimal unbiased estimates. [Hint: It might be helpful to consult McCullagh & Pregibon (1987) to find the inverse of the $n^4 \times n^4$ covariance matrix of the quartic functions of the residuals.] Note that, in the case of the fourth cumulant, this calculation is different from Pukelsheim (1980, Section 4), who initially assumes κ_2 to be known and finally replaces κ_2 by k_2.

4.26 Let $\mathbf{Y}$ have components Y^i satisfying

$$E(Y^i) = \mu^i = \omega^{\alpha;i}\beta_\alpha$$

or, in matrix notation, $E(\mathbf{Y}) = \mathbf{X}\beta$, where $\mathbf{X}$ is $n \times q$ of rank q. Let $\omega^{i,j}, \omega^{i,j,k}, \ldots$ be given tensors such that $Y^i - \mu^i$ has cumulants $\kappa_2\omega^{i,j}$, $\kappa_3\omega^{i,j,k}$ and so on. The first-order interaction matrix, $\mathbf{X}^*$, is obtained by appending to $\mathbf{X}$, $q(q+1)/2$ additional columns having elements in the ith row given by $\omega^{i;r}\omega^{i;s}$ for $1 \le r \le s \le q$. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$ and let $\mathbf{H}^*$, defined analogously, have rank $q^* < n$. Show that

$$k_2 = \mathbf{Y}^T\mathbf{W}(\mathbf{I} - \mathbf{H})\mathbf{Y}/(n - q)$$
$$k_2^* = \mathbf{Y}^T\mathbf{W}(\mathbf{I} - \mathbf{H}^*)\mathbf{Y}/(n - q^*)$$

are both

(i) unbiased for κ_2, (ii) invariant under the symmetric group (applied simultaneously to the rows of $\mathbf{Y}$ and $\mathbf{X}$),

(iii) invariant under the general linear group applied to the columns of $\mathbf{X}$ (i.e. such that the column space of $\mathbf{X}$ is preserved).

Show also that k_2 is invariant under the general linear group (4.25) applied to the rows of $\mathbf{X}$ and $\mathbf{Y}$, but that k_2^* is not so invariant.

4.27 Justify the claims made in Section 4.7.1 that interaction and replication are not invariant under the general linear group (4.25). Show that these concepts are preserved under the permutation group.

4.28 In the notation of Section 4.7.2, let

$$F_i = h_{ij}Y^j, \quad R_i = \rho_{ij}Y^j = (\delta_{ij} - h_{ij})Y^j$$

be the fitted value and the residual respectively. Define the derived statistics

$$T_1 = \sum R_j F_j^2 \quad \text{and} \quad T_2 = \sum R_j^2 F_j.$$

Show that

$$E(T_1) = \kappa_3(n_2 - 2n_{22} + n_3)$$
$$E(T_2) = \kappa_3(n_2 - n_3).$$

Show also, under the usual normal theory assumptions, that conditionally on the fitted values,

$$\operatorname{var}(T_1) = \kappa_2 \sum_{ij} \rho_{i,j} F_i^2 F_j^2$$

$$\operatorname{var}(T_2) = 2\kappa_2^2 \sum \rho_{i,j}^2 F_i F_j.$$

4.29 A population of size $N = N_0 + N_1$ comprises N_1 unit values and N_0 zero values. Show that the population Ks are

$$\begin{aligned}
K_1 &= N_1/N, \\
K_2 &= N_0N_1/N^{(2)}, \\
K_3 &= N_0N_1(N_0 - N_1)/N^{(3)}, \\
K_4 &= N_0N_1(N(N+1) - 6N_0N_1)/N^{(4)}
\end{aligned}$$

Hence derive the first four cumulants of the central hypergeometric distribution (Barton, David & Fix, 1960).

4.30 In a population consisting of the first N natural numbers, show that the population k-statistics are

$$\begin{aligned}
K_1 &= (N+1)/2, \\
K_2 &= N(N+1)/12, \\
K_3 &= 0, \\
K_4 &= -N^2(N+1)^2/120.
\end{aligned}$$

(Barton, David & Fix, 1960).

CHAPTER 5

Edgeworth series

5.1 Introduction

While the lower-order cumulants of X are useful for describing, in both qualitative and quantitative terms, the shape of the joint distribution, it often happens that we require either approximations for the density function itself or, more commonly, approximations for tail probabilities or conditional tail probabilities. For example, if X^1 is a goodness-of-fit statistic and $X^2, \ldots, X^p$ are estimates of unknown parameters, it would often be appropriate to assess the quality of the fit by computing the tail probability

$$\mathrm{pr}(X^1 \geq x^1 | X^2 = x^2, \ldots, X^p = x^p),$$

small values being taken as evidence of a poor fit. The cumulants themselves do not provide such estimates directly and it is necessary to proceed via an intermediate step where the density, or equivalently, the cumulative distribution function, is approximated by a series expansion. Series expansions of the Edgeworth type all involve an initial first-order approximation multiplied by a sum of correction terms whose coefficients are simple combinations of the cumulants of X. In the case of the Edgeworth expansion itself, the first-order approximation is based on the normal density having the same mean vector and covariance matrix as X. The correction terms then involve cumulants and cumulant products as coefficients of the Hermite polynomials. More generally, however, in order to minimize the number of correction terms, it may be advantageous to use a first-order approximation other than the normal. The correction terms then involve derivatives of the approximating density and are not necessarily polynomials.

In Section 5.2 we discuss the nature of such approximations for arbitrary initial approximating densities. These series take on different forms depending on whether we work with the density function, the log density function, the cumulative distribution function or some transformation of the cumulative distribution function. For theoretical calculations the log density is often the most convenient: for the computation of significance levels, the distribution function may be preferred because it is more directly useful.

The remainder of the chapter is devoted to the Edgeworth expansion itself and to derived expansions for conditional distributions.

5.2 A formal series expansion

5.2.1 *Approximation for the density*

Let $f_X(x;\kappa)$ be the joint density function of the random variables $X^1, \ldots, X^p$. The notation is chosen to emphasize the dependence of the density on the cumulants of X. Suppose that the initial approximating density, $f_0(x) \equiv f_0(x;\lambda)$ has cumulants λ^i, $\lambda^{i,j}$, $\lambda^{i,j,k}, \ldots$, which differ from the cumulants of X by

$$\eta^i = \kappa^i - \lambda^i, \quad \eta^{i,j} = \kappa^{i,j} - \lambda^{i,j}, \quad \eta^{i,j,k} = \kappa^{i,j,k} - \lambda^{i,j,k}$$

and so on. Often, it is good policy to choose $f_0(x;\lambda)$ so that $\eta^i = 0$ and $\eta^{i,j} = 0$, but it would be a tactical error to make this assumption at an early stage in the algebra, in the hope that substantial simplification would follow. While it is true that this choice makes many terms vanish, it does so at the cost of ruining the symmetry and structure of the algebraic formulae.

The cumulant generating functions for $f_X(x;\kappa)$ and $f_0(x;\lambda)$ are assumed to have their usual expansions

$$K_X(\xi) = \xi_i\kappa^i + \xi_i\xi_j\kappa^{i,j}/2! + \xi_i\xi_j\xi_k\kappa^{i,j,k}/3! + \cdots$$

and

$$K_0(\xi) = \xi_i\lambda^i + \xi_i\xi_j\lambda^{i,j}/2! + \xi_i\xi_j\xi_k\lambda^{i,j,k}/3! + \cdots.$$

Subtraction gives

$$K_X(\xi) = K_0(\xi) + \xi_i\eta^i + \xi_i\xi_j\eta^{i,j}/2! + \xi_i\xi_j\xi_k\eta^{i,j,k}/3! + \cdots$$

and exponentiation gives

$$M_X(\xi) = M_0(\xi)\{1 + \xi_i\eta^i + \xi_i\xi_j\eta^{ij}/2! + \xi_i\xi_j\xi_k\eta^{ijk}/3! + \cdots\} \tag{5.1}$$

where

$$\begin{aligned} \eta^{ij} &= \eta^{i,j} + \eta^i\eta^j, \\ \eta^{ijk} &= \eta^{i,j,k} + \eta^i\eta^{j,k}[3] + \eta^i\eta^j\eta^k \\ \eta^{ijkl} &= \eta^{i,j,k,l} + \eta^i\eta^{j,k,l}[4] + \eta^{i,j}\eta^{k,l}[3] + \eta^i\eta^j\eta^{k,l}[6] + \eta^i\eta^j\eta^k\eta^l \end{aligned}$$

and so on, are the formal 'moments' obtained by treating η^i, $\eta^{i,j}$, $\eta^{i,j,k}$, ... as formal 'cumulants'. It is important here to emphasize that $\eta^{ij} \neq \kappa^{ij} - \lambda^{ij}$, the difference between second moments of X and those of $f_0(x)$. Furthermore, the ηs are not, in general, the cumulants of any real random variable. For instance, $\eta^{i,j}$ need not be positive definite.

To obtain a series approximation for $f_X(x;\kappa)$, we invert the approximate integral transform, (5.1), term by term. By construction, the leading term, $M_0(\xi)$, transforms to $f_0(x)$. The second term, $\xi_i M_0(\xi)$, transforms to $f_i(x) = -\partial f_0(x)/\partial x^i$ as can be seen from the following argument. Integration by parts with respect to x^r gives

$$\begin{aligned} \int_{-\infty}^{\infty} \exp(\xi_i x^i)\frac{\partial f_0(x)}{\partial x^r}\,dx^r &= \exp(\xi_i x^i) f_0(x) \Bigg|_{x^r=-\infty}^{x^r=\infty} \\ &\quad - \xi_r \int_{-\infty}^{\infty} \exp(\xi_i x^i) f_0(x)\,dx^r. \end{aligned}$$

For imaginary ξ, $\exp(\xi_i x^i)$ is bounded, implying that the first term on the right is zero. Further integration with respect to the remaining $p-1$ variables gives

$$\int_{R^p} \exp(\xi_i x^i) f_r(x) dx = \xi_r M_0(\xi).$$

The critical assumption here is that $f_0(x)$ should have continuous partial derivatives everywhere in R^p. For example, if $f_0(x)$ had discontinuities, these would give rise to additional terms in the above integration.

By the same argument, if $f_0(x)$ has continuous second-order partial derivatives, it may be shown that $\xi_r\xi_s M_X(\xi)$ is the integral transform of $f_{rs}(x) = (-1)^2\partial^2 f_0(x)/\partial x^r \partial x^s$ and so on. In other words, (5.1) is the integral transform of

$$\begin{aligned} f_X(x) = f_0(x) &+ \eta^i f_i(x) + \eta^{ij} f_{ij}(x)/2! \\ &+ \eta^{ijk} f_{ijk}(x)/3! + \cdots \end{aligned} \tag{5.2}$$

where the alternating signs on the derivatives have been incorporated into the notation.

Approximation (5.2) looks initially very much like a Taylor expansion in the sense that it involves derivatives of a function divided by the appropriate factorial. In fact, this close resemblance can be exploited, at least formally, by writing

$$X = Y + Z$$

where Y has density $f_0(y)$ and Z is a pseudo-variable independent of Y with 'cumulants' η^i, $\eta^{i,j}$, $\eta^{i,j,k}, \ldots$ and 'moments' η^i, η^{ij}, $\eta^{ijk}, \ldots$. Conditionally on $Z = z$, X has density $f_0(x - z)$ and hence, the marginal density of X is, formally at least,

$$f_X(x) = E_Z\{f_0(x - Z)\}.$$

Taylor expansion about $Z = 0$ and averaging immediately yields (5.2). This formal construction is due to Davis (1976).

To express (5.2) as a multiplicative correction to $f_0(x)$, we write

$$\begin{aligned} f_X(x;\kappa) = f_0(x)\{1 + \eta^i h_i(x) + \eta^{ij} h_{ij}(x)/2! \\ + \eta^{ijk} h_{ijk}(x)/3! + \cdots\} \end{aligned} \tag{5.3}$$

where

$$h_i(x) = h_i(x;\lambda) = f_i(x)/f_0(x), \quad h_{ij}(x) = h_{ij}(x;\lambda) = f_{ij}(x)/f_0(x)$$

and so on. In many instances, $h_i(x)$, $h_{ij}(x), \ldots$ are simple functions with pleasant mathematical properties. For example, if $f_0(x) = \phi(x)$, the standard normal density, then $h_i(x)$, $h_{ij}(x), \ldots$ are the standard Hermite tensors, or polynomials in the univariate case. See Sections 5.3 and 5.4.

5.2.2 *Approximation for the log density*

For a number of reasons, both theoretical and applied, it is often better to consider series approximations for the log density $\log f_X(x;\kappa)$ rather than for the density itself. One obvious advantage, particularly where polynomial approximation is involved, is that $f_X(x;\kappa) \geq 0$, whereas any polynomial approximation is liable to become negative for certain values of x. In addition, even if the infinite series (5.3) could be guaranteed positive, there is no similar guarantee for the truncated series that would actually be used in practice. For these reasons, better approximations may be obtained by approximating $\log f_X(x;\kappa)$ and exponentiating.

Expansion of the logarithm of (5.3) gives

$$\begin{aligned} \log f_X(x;\kappa) \simeq \log f_0(x) &+ \eta^i h_i(x) \\ &+ \{\eta^{i,j} h_{ij}(x) + \eta^i \eta^j h_{i,j}(x)\}/2! \\ &+ \{\eta^{i,j,k} h_{ijk}(x) + \eta^i \eta^{j,k} h_{i,jk}(x)[3] + \eta^i \eta^j \eta^k h_{i,j,k}(x)\}/3! \\ &+ \{\eta^{i,j,k,l} h_{ijkl} + \eta^i \eta^{j,k,l} h_{i,jkl}[4] + \eta^{i,j} \eta^{k,l} h_{ij,kl}[3] \\ &\quad + \eta^i \eta^j \eta^{k,l} h_{i,j,kl}[6] + \eta^i \eta^j \eta^k \eta^l h_{i,j,k,l}\}/4! \end{aligned} \tag{5.4}$$

where

$$\begin{aligned} h_{i,j} &= h_{ij} - h_i h_j \\ h_{i,j,k} &= h_{ijk} - h_i h_{jk}[3] + 2h_i h_j h_k \\ h_{i,jk} &= h_{ijk} - h_i h_{jk} \\ h_{ij,kl} &= h_{ijkl} - h_{ij} h_{kl} \end{aligned}$$

and so on. Compare equations (2.7) and (3.2). Note that, apart from sign, the hs with fully partitioned indices are derivatives of the log density

$$\begin{aligned} h_i(x) &= -\partial \log f_0(x)/\partial x^i \\ h_{i,j}(x) &= \partial^2 \log f_0(x)/\partial x^i \partial x^j \\ h_{i,j,k}(x) &= -\partial^3 \log f_0(x)/\partial x^i \partial x^j \partial x^k \end{aligned}$$

and so on.

The h-functions with indices partially partitioned are related to each other in exactly the same way as generalized cumulants. Thus, for example, from (3.2) we find

$$h_{i,jk} = h_{i,j,k} + h_j h_{i,k} + h_k h_{i,j}$$

and

$$h_{i,jkl} = h_{i,j,k,l} + h_j h_{i,k,l}[3] + h_{i,j} h_{k,l}[3] + h_{i,j} h_k h_l[3],$$

using the conventions of Chapter 3.

Often it is possible to choose $f_0(x)$ so that its first-order and second-order moments agree with those of X. In this special case, (5.4) becomes

$$\begin{aligned} \log f_X(x) - \log f_0(x) &= \eta^{i,j,k} h_{ijk}(x)/3! + \eta^{i,j,k,l} h_{ijkl}(x)/4! \\ &+ \eta^{i,j,k,l,m} h_{ijklm}(x)/5! + \eta^{i,j,k,l,m,n} h_{ijklmn}(x)/6! \\ &+ \eta^{i,j,k} \eta^{l,m,n} h_{ijk,lmn}(x)[10]/6! + \cdots . \end{aligned} \tag{5.5}$$

This simplification greatly reduces the number of terms required but at the same time it manages to conceal the essential simplicity of the terms in (5.4). The main statistical reason for keeping to the more general expansion (5.4) is that, while it is generally possible to chose $f_0(x)$ to match the first two moments of X under some null hypothesis H_0, it is generally inconvenient to do so under general alternatives to H_0, as would be required in calculations related to the power of a test.

5.2.3 *Approximation for the cumulative distribution function*

One of the attractive features of expansions of the Edgeworth type is that the cumulative distribution function

$$F_X(x;\kappa) = \mathrm{pr}(X^1 \le x^1, \ldots, X^p \le x^p; \kappa)$$

can easily be approximated by integrating (5.2) term by term. If $F_0(x)$ is the cumulative distribution function corresponding to $f_0(x)$, this gives

$$\begin{aligned} F_X(x;\kappa) \simeq F_0(x) + \eta^i F_i(x) + \eta^{ij} F_{ij}(x)/2! \\ + \eta^{ijk} F_{ijk}(x)/3! + \cdots \end{aligned} \tag{5.6}$$

where

$$F_i(x) = (-1)\partial F_0(x)/\partial x^i; \quad F_{ij}(x) = (-1)^2 \partial^2 F_0(x)/\partial x^i \partial x^j$$

and so on with signs alternating.

Of course, we might prefer to work with the multivariate survival probability

$$S_X(x) = \mathrm{pr}(X^1 \ge x^1, X^2 \ge x^2, \ldots, X^p \ge x^p; \kappa)$$

and the corresponding probability, $S_0(x)$, derived from $f_0(x)$. Integration of (5.2) term by term gives

$$S_X(x;\kappa) \simeq S_0(x) + \eta^i S_i(x) + \eta^{ij} S_{ij}(x)/2! + \eta^{ijk} S_{ijk}(x)/3! + \cdots$$

where

$$S_i(x) = -\partial S_0(x)/\partial x^i; \qquad S_{ij}(x) = (-1)^2\partial^2 S_0(x)/\partial x^i \partial x^j$$

and so on with signs alternating as before.

In the univariate case, $p = 1$, but not otherwise, $F_X(x;\kappa) + S_X(x;\kappa) = 1$ if the distribution is continuous, and the correction terms in the two expansions sum to zero giving $F_0(x)+ S_0(x) = 1$.

Other series expansions can be found for related probabilities. We consider here only one simple example, namely

$$\bar{S}_X(x;\kappa) = \mathrm{pr}(X^1 \le x^1 \text{ or } X^2 \le x^2 \text{ or } \cdots \text{ or } X^p \le x^p;\kappa)$$

as opposed to $F_X(x;\kappa)$, whose definition involves replacing 'or' with 'and' in the expression above. By definition, in the continuous case, $\bar{S}_X(x;\kappa) = 1 - S_X(x;\kappa)$. It follows that

$$\bar{S}_X(x;\kappa) \simeq \bar{S}_0(x) + \eta^i \bar{S}_i(x) + \eta^{ij}\bar{S}_{ij}(x)/2! + \eta^{ijk}\bar{S}_{ijk}(x)/3! + \cdots$$

where

$$\bar{S}_i(x) = -\partial \bar{S}_0(x)/\partial x^i; \quad \bar{S}_{ij}(x) = (-1)^2\partial^2 \bar{S}_0(x)/\partial x^i \partial x^j$$

and so on, again with alternating signs on the derivatives.

5.3 Expansions based on the normal density

5.3.1 *Multivariate case*

Suppose now, as a special case of (5.2) or (5.3) that the initial approximating density is chosen to be $f_0(x) = \phi(x;\lambda)$, the normal density with mean vector λ^i and covariance matrix $\lambda^{i,j}$. The density may be written

$$\phi(x;\lambda) = (2\pi)^{-p/2}|\lambda^{i,j}|^{-1/2}\exp\{-\tfrac{1}{2}(x^i-\lambda^i)(x^j-\lambda^j)\lambda_{i,j}\}$$

where $\lambda_{i,j}$ is the matrix inverse of $\lambda^{i,j}$ and $|\lambda^{i,j}|$ is the determinant of the covariance matrix. The functions $h_i(x) = h_i(x;\lambda)$, $h_{ij}(x) = h_{ij}(x;\lambda), \ldots$, obtained by differentiating $\phi(x;\lambda)$, are known as the Hermite tensors. The first six are given below

$$\begin{aligned}
h_i &= \lambda_{i,j}(x^j - \lambda^j)\\
h_{ij} &= h_i h_j - \lambda_{i,j}\\
h_{ijk} &= h_i h_j h_k - h_i \lambda_{j,k}[3]\\
h_{ijkl} &= h_i h_j h_k h_l - h_i h_j \lambda_{k,l}[6] + \lambda_{i,j}\lambda_{k,l}[3]\\
h_{ijklm} &= h_i h_j h_k h_l h_m - h_i h_j h_k \lambda_{l,m}[10] + h_i \lambda_{j,k}\lambda_{l,m}[15]\\
h_{ijklmn} &= h_i \cdots h_n - h_i h_j h_k h_l \lambda_{m,n}[15] + h_i h_j \lambda_{k,l}\lambda_{m,n}[45]\\
&\quad - \lambda_{i,j}\lambda_{k,l}\lambda_{m,n}[15]
\end{aligned} \tag{5.7}$$

The general pattern is not difficult to describe: it involves summation over all partitions of the indices, unit blocks being associated with h_i, double blocks with $-\lambda_{i,j}$ and partitions having blocks of three or more elements being ignored.

In the univariate case, these reduce to the Hermite polynomials, which form an orthogonal basis with respect to $\phi(x;\lambda)$ as weight function, for the space of functions continuous over $(-\infty,\infty)$. The Hermite tensors are the polynomials that form a similar orthogonal basis for functions continuous over R^p. Their properties are discussed in detail in Section 5.4.

The common practice is to chose $\lambda^i = \kappa^i$, $\lambda^{i,j} = \kappa^{i,j}$, so that

$$\begin{gathered}\eta^i = 0, \quad \eta^{i,j} = 0, \quad \eta^{i,j,k} = \kappa^{i,j,k}, \quad \eta^{i,j,k,l} = \kappa^{i,j,k,l} \\ \eta^{ijk} = \kappa^{i,j,k}, \quad \eta^{ijkl} = \kappa^{i,j,k,l}, \quad \eta^{ijklm} = \kappa^{i,j,k,l,m}, \\ \eta^{ijklmn} = \kappa^{i,j,k,l,m,n} + \kappa^{i,j,k}\kappa^{l,m,n}[10]\end{gathered}$$

and so on, summing over all partitions ignoring those that have blocks of size 1 or 2. Thus (5.3) gives

$$\begin{aligned}f_X(x;\kappa) \simeq \phi(x;\kappa)\{&1 + \kappa^{i,j,k}h_{ijk}(x)/3! + \kappa^{i,j,k,l}h_{ijkl}(x)/4! \\ &+ \kappa^{i,j,k,l,m}h_{ijklm}(x)/5! \\ &+ (\kappa^{i,j,k,l,m,n} + \kappa^{i,j,k}\kappa^{l,m,n}[10])h_{ijklmn}(x)/6! + \cdots\}.\end{aligned} \tag{5.8}$$

This form of the approximation is sometimes known as the Gram-Charlier series. To put it in the more familiar and useful Edgeworth form, we note that if X is a standardized sum of n independent random variables, then $\kappa^{i,j,k} = O(n^{-1/2})$, $\kappa^{i,j,k,l} = O(n^{-1})$ and so on, decreasing in power of $n^{-1/2}$. Thus, the successive correction terms in the Gram-Charlier series are of orders $O(n^{-1/2})$, $O(n^{-1})$, $O(n^{-3/2})$ and $O(n^{-1})$ and these are not monotonely decreasing in n. The re-grouped series, formed by collecting together terms that are of equal order in n,

$$\begin{aligned}\phi(x;\kappa)\Big[&1 + \kappa^{i,j,k}h_{ijk}(x)/3! \\ &+ \{\kappa^{i,j,k,l}h_{ijkl}(x)/4! + \kappa^{i,j,k}\kappa^{l,m,n}h_{ijklmn}(x)[10]/6!\} \\ &+ \{\kappa^{i,j,k,l,m}h_{ijklm}(x)/5! + \kappa^{i,j,k}\kappa^{l,m,n,r}h_{i\cdots r}(x)[35]/7! \\ &\quad + \kappa^{i,j,k}\kappa^{l,m,n}\kappa^{r,s,t}h_{i\cdots t}(x)[280]/9!\} + \cdots\Big]\end{aligned} \tag{5.9}$$

is called the Edgeworth series and is often preferred for statistical calculations. The infinite versions of the two series are formally identical and the main difference is that truncation of (5.8) after a fixed number of terms gives a different answer than truncation of (5.9) after a similar number of terms.

At the origin of x, the approximating density takes the value

$$\phi(0;\kappa)\;[1 + 3\rho_4/4! - \{9\rho_{13}^2 + 6\rho_{23}^2\}/72 + O(n^{-2})] \tag{5.10}$$

where ρ_4, ρ_{13}^2 and ρ_{23}^2 are the invariant standardized cumulants of X. Successive terms in this series decrease in whole powers of n as opposed to the half-powers in (5.9).

5.3.2 *Univariate case*

The univariate case is particularly important because the notion of a tail area, as used in significance testing, depends on the test statistic being one-dimensional. In the univariate case it is convenient to resort to the conventional power notation by writing

$$\begin{aligned}h_1(x;\kappa) &= \kappa_2^{-1}(x-\kappa_1) \\ h_2(x;\kappa) &= h_1^2 - \kappa_2^{-1} \\ h_3(x;\kappa) &= h_1^3 - 3\kappa_2^{-1}h_1\end{aligned}$$

and so on for the Hermite polynomials. These are derived in a straightforward way from the Hermite tensors (5.7). The standard Hermite polynomials, obtained by putting $\kappa_1 = 0$, $\kappa_2 = 1$ are

$$\begin{aligned}h_1(z) &= z \\ h_2(z) &= z^2 - 1 \\ h_3(z) &= z^3 - 3z \\ h_4(z) &= z^4 - 6z^2 + 3 \\ h_5(z) &= z^5 - 10z^3 + 15z \\ h_6(z) &= z^6 - 15z^4 + 45z^2 - 15.\end{aligned} \tag{5.11}$$

The first correction term in (5.9) is

$$\begin{aligned}\kappa_3 h_3(x;\kappa)/6 &= \kappa_3\{\kappa_2^{-3}(x-\kappa_1)^3 - 3\kappa_2^{-2}(x-\kappa_1)\}/6 \\ &= \rho_3 h_3(z)/6 = \rho_3(z^3-3z)/6\end{aligned}$$

where $z = \kappa_2^{-1/2}(x-\kappa_1)$ and $Z = \kappa_2^{-1/2}(X-\kappa_1)$ is the standardized version of X. Similarly, the $O(n^{-1})$ correction term becomes

$$\rho_4 h_4(z)/4! + 10\rho_3^2 h_6(z)/6!$$

Even in theoretical calculations it is rarely necessary to use more than two correction terms from the Edgeworth series. For that reason, we content ourselves with corrections up to order $O(n^{-1})$, leaving an error that is of order $O(n^{-3/2})$.

For significance testing, the one-sided tail probability may be approximated by

$$\begin{aligned}\mathrm{pr}(Z \geq z) \simeq{}& 1 - \Phi(z) \\ &+ \phi(z)\{\rho_3 h_2(z)/3! + \rho_4 h_3(z)/4! + \rho_3^2 h_5(z)/72\},\end{aligned} \tag{5.12}$$

which involves one correction term of order $O(n^{-1/2})$ and two of order $O(n^{-1})$. The two-sided tail probability is, for $z > 0$,

$$\begin{aligned}\mathrm{pr}(|Z| \geq z) ={}& 2\{1-\Phi(z)\} \\ &+ 2\phi(z)\{\rho_4 h_3(z)/4! + \rho_3^2 h_5(z)/72\}\end{aligned} \tag{5.13}$$

and this involves only correction terms of order $O(n^{-1})$. In essence, what has happened here is that the $O(n^{-1/2})$ corrections are equal in magnitude but of different sign in the two tails and they cancel when the two tails are combined.

5.3.3 *Regularity conditions*

So far, we have treated the series expansions (5.8) to (5.13) in a purely formal way. No attempt has been made to state precisely the way in which these series expansions are supposed to approximate either the density or the cumulative distribution function. In this section, we illustrate in an informal way, some of the difficulties encountered in making this notion precise.

First, it is not difficult to see that the infinite series expansion (5.3) or (5.8) for the density is, in general, divergent. For example, if the cumulant differences, η^i, $\eta^{i,j}$, $\eta^{i,j,k}$,... are all equal to 1, then $\eta^{ij} = 2$, $\eta^{ijk} = 5$, $\eta^{ijkl} = 15, \ldots$, otherwise known as the Bell numbers. These increase exponentially fast so that (5.3) fails to converge. For this reason we need a justification of an entirely different kind for using the truncated expansion, truncated after some fixed number of terms.

To this end, we introduce an auxiliary quantity, n, assumed known, and we consider the formal mathematical limit $n \to \infty$. The idea now is to approximate $K_X(\xi)$ by a truncated series such that after $r+2$ terms the error is $O(n^{-r/2})$, say. Formal inversion gives a truncated series expansion for the density or distribution function which, under suitable smoothness conditions, has an error also of order $O(n^{-r/2})$. In all the applications we have in mind, X is a standardized sum and $f_0(x)$ is the normal density having the same mean and variance as X. Then $\eta^i = 0$, $\eta^{i,j} = 0$, $\eta^{i,j,k} = O(n^{-1/2})$, $\eta^{i,j,k,l} = O(n^{-1})$ and so on. Thus the truncated expansion for $K_X(\xi)$, including terms up to degree 4 in ξ, has error $O(n^{-3/2})$. To obtain similar accuracy in the approximations for $M_X(\xi)$ or $f_X(x)$ it is necessary to go as far as terms of degree 6.

This rather informal description ignores one fairly obvious point. The truncated series approximation (5.5) for $F_X(x;\kappa)$ is continuous. If X has a discrete distribution, $F_X(x;\kappa)$ will have discontinuities with associated probabilities typically of order $O(n^{-1/2})$. Any continuous approximation for $F_X(x)$ will therefore have an error of order $O(n^{-1/2})$, however many correction terms are employed.

In the univariate case, if X is the standardized sum of n independent and identically distributed random variables, Feller (1971, Chapter XVI) gives conditions ensuring the validity of expansions (5.9) for the density and (5.12) for the cumulative distribution function. The main condition, that

$$\limsup_{\zeta\to\infty} |M_X(i\zeta)| < 1,$$

excludes lattice distributions but ensures that the error in (5.12) is $o(n^{-1})$ uniformly in z. With weaker conditions, but again excluding lattice distributions, the error is known to be $o(n^{-1/2})$ uniformly in z (Esseen, 1945; Bahadur & Ranga-Rao, 1960). In the lattice case, (5.12) has error $o(n^{-1/2})$ provided that it is used with the usual correction for continuity. An asymptotic expansion given by Esseen (1945) extends the Edgeworth expansion to the lattice case: in particular, it produces correction terms for (5.12) so that the renaining error is $O(n^{-3/2})$ uniformly in z.

Similar conditions dealing with the multivariate case are given by Barndorff-Nielsen & Cox (1979).

The case where X is a standardized sum of non-independent or non-identically distributed random variables is considerably more complicated. It is necessary, for example, to find a suitably strengthened version of the Lindeberg condition involving, perhaps, the higher-order cumulants. In addition, if X is a sum of discrete and continuous random variables, it would be necessary to know whether the discrete or the continuous components dominated in the overall sum. No attempt will be made to state such conditions here.

For details of regularity conditions, see Bhattacharya & Rao (1976) or Skovgaard, (1981a,b, 1986b).

5.4 Some properties of Hermite tensors

5.4.1 *Tensor properties*

We now investigate the extent to which the arrays of polynomials $h_i(x;\lambda)$, $h_{ij}(x;\lambda)$, $h_{ijk}(x;\lambda), \ldots$, defined apart from choice of sign, as the coefficient of $\phi(x;\lambda)$ in the partial derivatives of $\phi(x;\lambda)$, deserve to be called tensors. To do so, we must examine the effect on the polynomials of transforming from x to new variables $\bar{x}$. In doing so, we must also take into account the induced transformation on the λs, which are the formal cumulants of x.

Consider the class of non-singular affine transformations on x

$$\bar{x}^i = a^i + a^i_r x^r \tag{5.14}$$

together with the induced transformation on the λs

$$\bar{\lambda}^i = a^i + a^i_r \lambda^r \quad \text{and} \quad \bar{\lambda}^{i,j} = a^i_r a^j_s \lambda^{r,s}.$$

It is easily checked that

$$(x^i - \lambda^i)(x^i - \lambda^j)\lambda_{i,j} = (\bar{x}^i - \bar{\lambda}^i)(\bar{x}^j - \bar{\lambda}^j)\bar{\lambda}_{i,j} \tag{5.15}$$

so that this quadratic form is invariant under nonsingular affine transformation. If a^i_r does not have full rank, this argument breaks down. Apart from the determinantal factor $|\lambda^{i,j}|^{-1/2}$, which has no effect on the definition of the Hermite tensors, $\phi(x;\lambda)$ is invariant under the transformation

(5.14) because it is a function of the invariant quadratic form (5.15). It follows therefore that the derivatives with respect to $\bar{x}$ are

$$\frac{\partial\phi}{\partial x^r}\frac{\partial x^r}{\partial\bar{x}^i}, \qquad \frac{\partial^2\phi}{\partial x^r\partial x^s}\frac{\partial x^r}{\partial\bar{x}^i}\frac{\partial x^s}{\partial\bar{x}^j}+\frac{\partial\phi}{\partial x^r}\frac{\partial^2 x^r}{\partial\bar{x}^i\partial\bar{x}^j}$$

and so on. Since $\partial^2 x^r/\partial\bar{x}^i\partial\bar{x}^j\equiv 0$, these derivatives may be written in the form

$$b_i^r\phi_r, \quad b_i^r b_j^s\phi_{rs}, \quad b_i^r b_j^s b_k^t\phi_{rst}$$

and so on where b_i^r is the matrix inverse of a_r^i. Thus the derivatives of ϕ are Cartesian tensors and they transform in the covariant manner. It follows immediately that $h_i(x;\lambda)$, $h_{ij}(x;\lambda)$ and so on are also Cartesian tensors, justifying the terminology of Section 5.3.

The above result applies equally to the more general case discussed in Section 5.2 where the form of the initial approximating density $f_0(x;\lambda)$ is left unspecified. See Exercise 5.5.

5.4.2 *Orthogonality*

The principal advantage of index notation is its total transparency: the principal disadvantage, nowhere more evident than in Chapter 1 where tensors were defined, is that general expressions for tensors of arbitrary order are difficult to write down without introducing unsightly subscripted indices. For this reason, we have avoided writing down general expressions for Hermite tensors, even though the general pattern is clear and is easily described in a few words (Section 5.3.1). To prove orthogonality it is necessary either to devise a suitable general notation or to use a less direct method of proof.

Our method of proof uses the generating function

$$\begin{aligned}\exp\{\xi_i(x^i-\lambda^i)-\tfrac{1}{2}\xi_i\xi_j\lambda^{i,j}\} = 1+\xi_i h^i+\xi_i\xi_j h^{ij}/2!\\ +\xi_i\xi_j\xi_k h^{ijk}/3!+\cdots\end{aligned} \tag{5.16}$$

where

$$h^i=\lambda^{i,r}h_r(x;\lambda), \qquad h^{ij}=\lambda^{i,r}\lambda^{j,s}h_{rs}(x;\lambda),$$
$$h^{ijk}=\lambda^{i,r}\lambda^{j,s}\lambda^{k,t}h_{rst}(x;\lambda)$$

and so on, are the contravariant expressions of the Hermite tensors. To show that (5.16) is indeed a generating function for the hs, we observe that

$$\phi(x^i-\lambda^{i,j}\xi_j;\lambda)=\phi(x;\lambda)\exp\{\xi_i(x^i-\lambda^i)-\xi_i\xi_j\lambda^{i,j}/2\}.$$

Taylor expansion about $\xi=0$ of $\phi(x^i-\lambda^{i,j}\xi_j;\lambda)$ gives

$$\phi(x;\lambda)\{1+\xi_i h^i+\xi_i\xi_j h^{ij}/2!+\xi_i\xi_j\xi_k h^{ijk}/3!+\cdots\},$$

which we have simplified using the definition of Hermite tensors. This completes the proof that $\exp\{\xi_i(x^i-\lambda^i)-\xi_i\xi_j\lambda^{i,j}/2\}$ is the generating function for the Hermite tensors.

To prove orthogonality, consider the product

$$\begin{aligned}&\phi(x^i-\lambda^{i,j}\zeta_j;\lambda)\exp\{\xi_i(x^i-\lambda^i)-\xi_i\xi_j\lambda^{i,j}/2\}\\ &=\phi(x;\lambda)\{1+\zeta_i h^i+\zeta_i\zeta_j h^{ij}/2!+\cdots\}\{1+\xi_i h^i+\xi_i\xi_j h^{ij}/2!+\cdots\}.\end{aligned}$$

Simplification of the exponent gives

$$\phi(x^i-\lambda^{i,j}\xi_j-\lambda^{i,j}\zeta_j;\lambda)\exp(\xi_i\zeta_j\lambda^{i,j})$$

and integration with respect to x over R^p gives $\exp(\xi_i\zeta_j\lambda^{i,j})$. Orthogonality follows because $\exp(\xi_i\zeta_j\lambda^{i,j})$ involves only terms of equal degree in ξ and ζ. Moreover, from the expansion

$$\exp(\xi_i\zeta_j\lambda^{i,j}) = 1 + \xi_i\zeta_j\lambda^{i,j} + \xi_i\xi_j\zeta_k\zeta_l\lambda^{i,k}\lambda^{j,l}/2! + \cdots,$$

it follows that the inner products over R^p are

$$\begin{aligned} \int h^i h^j \phi\, dx &= \lambda^{i,j} \\ \int h^{ij} h^{kl} \phi\, dx &= \lambda^{i,k}\lambda^{j,l} + \lambda^{i,l}\lambda^{j,k} \\ \int h^{ijk} h^{lmn} \phi\, dx &= \lambda^{i,l}\lambda^{j,m}\lambda^{k,m}[3!] \end{aligned} \tag{5.17}$$

and so on for tensors of higher order.

In the univariate case where the hs are standard Hermite polynomials, the orthogonality relations are usually written in the form $\int h_r(x)h_s(x)\phi(x)\,dx = r!\delta_{rs}$.

The extension of the above to scalar products of three or more Hermite polynomials or tensors is given in Exercises 5.9–5.14.

5.4.3 *Generalized Hermite tensors*

From any sequence of arrays h^i, h^{ij}, $h^{ijk}, \ldots$, each indexed by an unordered set of indices, there may be derived a new sequence h^i, $h^{i,j}$, $h^{i,j,k}, \ldots$, each indexed by a fully partitioned set of indices. If $M_h(\xi)$ is the generating function for the first set, then $K_h(\xi) = \log M_h(\xi)$ is the generating function for the new set and the relationship between the two sequences is identical to the relationship between moments and cumulants. In the case of Hermite tensors, we find on taking logs in (5.16) that

$$K_h(\xi) = \xi_i(x^i - \lambda^i) - \xi_i\xi_j\lambda^{i,j}/2 \tag{5.18}$$

so that $h^i = x^i - \lambda^i$, $h^{i,j} = -\lambda^{i,j}$ and all other arrays in the new sequence are identically zero. In fact, apart from choice of sign, the arrays in this new sequence are the partial derivatives of $\log\phi(x;\lambda)$ with respect to the components of x.

Just as ordinary moments and ordinary cumulants are conveniently considered as special cases of generalized cumulants, so too the sequences h^i, h^{ij}, h^{ijk}, $\cdots$ and h^i, $h^{i,j}$, $h^{i,j,k}$ are special cases of generalized Hermite tensors. Some examples with indices fully partitioned are given below.

$$\begin{aligned} h^{i,jk} &= h^{ijk} - h^i h^{jk} \\ &= h^{i,j,k} + h^j h^{i,k} + h^k h^{i,j} \\ &= -(x^j - \lambda^j)\lambda^{i,k} - (x^k - \lambda^k)\lambda^{i,j} \\ h^{i,jkl} &= h^{ijkl} - h^i h^{jkl} \\ &= h^{i,j,k,l} + h^j h^{i,k,l}[3] + h^{i,j} h^{k,l}[3] + h^{i,j} h^k h^l[3] \\ &= -h^k h^l \lambda^{i,j}[3] + \lambda^{i,j}\lambda^{k,l}[3] \\ h^{ij,kl} &= h^{ijkl} - h^{ij} h^{kl} \\ &= h^{i,j,k,l} + h^i h^{j,k,l}[4] + h^{i,k} h^{j,l}[2] + h^{i,k} h^j h^l[4] \\ &= -h^j h^l \lambda^{i,k}[4] + \lambda^{i,k}\lambda^{j,l}[2] \\ h^{i,j,kl} &= h^{ijkl} - h^i h^{jkl}[2] - h^{ij} h^{kl} + 2h^i h^j h^{kl} \\ &= h^{i,j,k,l} + h^k h^{i,j,l}[2] + h^{i,k} h^{j,l}[2] \\ &= \lambda^{i,k}\lambda^{j,l}[2] \end{aligned}$$

In these examples, the first line corresponds to the expressions (3.2) for generalized cumulants in terms of moments. The second line corresponds to the fundamental identity (3.3) for generalized cumulants in terms of ordinary cumulants and the final line is obtained on substituting the coefficients in (5.18).

A generalized Hermite tensor involving β indices partitioned into α blocks is of degree $\beta-2\alpha+2$ in x provided that $\beta-2\alpha+2\geq 0$. Otherwise the array is identically zero.

The importance of these generalized Hermite tensors lies in the formal series expansion (5.4) for the log density, which is used in computing approximate conditional cumulants.

5.4.4 *Factorization of Hermite tensors*

Suppose that X is partitioned into two vector components $X^{(1)}$ and $X^{(2)}$ of dimensions q and $p-q$ respectively. We suppose also that the normal density $\phi(x;\lambda)$, from which the Hermite tensors are derived, is expressible as the product of two normal densities, one for $X^{(1)}$ and one for $X^{(2)}$. Using the indices $i,j,k,\ldots$ to refer to components of $X^{(1)}$ and $r,s,t,\ldots$ to refer to components of $X^{(2)}$, we may write $\lambda^{i,r}=0$ and $\lambda_{i,r}=0$. Since the two components in the approximating density are uncorrelated, there is no ambiguity in writing $\lambda_{i,j}$ and $\lambda_{r,s}$ for the matrix inverses of $\lambda^{i,j}$ and $\lambda^{r,s}$ respectively. More generally, if the two components were correlated, this notation would be ambiguous because the leading $q\times q$ sub-matrix of

$$\begin{bmatrix}\lambda^{i,j} & \lambda^{i,s}\\ \lambda^{r,j} & \lambda^{r,s}\end{bmatrix}^{-1},$$

namely $\{\lambda^{i,j}-\lambda^{i,r}\{\lambda^{r,s}\}^{-1}\lambda^{s,j}\}^{-1}$, is not the same as the matrix inverse of $\lambda^{i,j}$.

Since $h^{i,r}=-\lambda^{i,r}=0$ and all higher-order generalized Hermite tensors, with indices fully partitioned, are zero, it follows immediately from (3.3) that

$$\begin{aligned}h^{i,rs}=0;\quad h^{ij,r}=0;\quad h^{i,rst}=0;\quad h^{ij,rs}=0\\ h^{ijk,r}=0;\quad h^{i,j,rs}=0\end{aligned}$$

and so on for any sub-partition of $\{(i,j,k,\ldots),\ (r,s,t,\ldots)\}$. This is entirely analogous to the statement that mixed cumulants involving two independent random variables and no others are zero. The corresponding statement for moments involves multiplication or factorization so that

$$\begin{aligned}&h^{irs}=h^ih^{rs};\qquad h^{ijr}=h^{ij}h^r;\qquad h^{ijrs}=h^{ij}h^{rs};\\ &h^{i,jrs}=h^{i,j}h^{rs};\qquad h^{i,r,js}=h^{i,j}h^{r,s};\\ &h^{ir,js}=h^{i,j}h^{r,s}+h^ih^jh^{r,s}+h^{i,j}h^rh^s=h^{ij}h^{r,s}+h^{i,j}h^rh^s\end{aligned}$$

and so on.

5.5 Linear regression and conditional cumulants

5.5.1 *Covariant representation of cumulants*

For calculations involving conditional distributions or conditional cumulants, it is often more convenient to work not with the cumulants of X^i directly but rather with the cumulants of $X_i=\kappa_{i,j}X^j$, which we denote by κ_i, $\kappa_{i,j}$, $\kappa_{i,j,k},\ldots$. We refer to these as the covariant representation of X and the covariant representation of its cumulants. This transformation may, at first sight, seem inconsequential but it should be pointed out that, while the notation X^i is unambiguous, the same cannot be said of X_i because the value of X_i depends on the entire set of variables $X^1,\ldots,X^p$. Deletion of X^p, say, leaves the remaining components $X^1,\ldots,X^{p-1}$ and their joint cumulants unaffected but the same is not true of $X_1,\ldots,X_{p-1}$.

5.5.2 *Orthogonalized variables*

Suppose now that X is partitioned into two vector components $X^{(1)}$ and $X^{(2)}$ of dimensions p and $q-p$ respectively and that we require the conditional cumulants of $X^{(1)}$ after linear regression on $X^{(2)}$. This is a simpler task than finding the conditional cumulants of $X^{(1)}$ given $X^{(2)} = x^{(2)}$, because it is assumed implicitly that only the conditional mean of $X^{(1)}$ and none of the higher-order conditional cumulants depends on the value of $x^{(2)}$. Further, each component of the conditional mean is assumed to depend linearly on $x^{(2)}$.

We first make a non-singular linear transformation from the original $(X^{(1)}, X^{(2)})$ to new variables $Y = (Y^{(1)}, Y^{(2)})$ in such a way that $Y^{(2)} = X^{(2)}$ and $Y^{(1)}$ is uncorrelated with $Y^{(2)}$. Extending the convention established in Section 5.4.4, we let the indices $i, j, k, \ldots$ refer to components of $X^{(1)}$ or $Y^{(1)}$, indices $r, s, t, \ldots$ refer to components of $X^{(2)}$ or $Y^{(2)}$ and indices $\alpha, \beta, \gamma, \ldots$ refer to components of the joint variable X or Y. The cumulants of X are κ^α partitioned into κ^i and κ^r, $\kappa^{\alpha,\beta}$ partitioned into $\kappa^{i,j}$ $\kappa^{i,r}$ and $\kappa^{r,s}$ and so on. The same convention applies to the covariant representation, κ_i, κ_r, $\kappa_{i,j}$, $\kappa_{i,r}$ and $\kappa_{r,s}$.

At this stage, it is necessary to distinguish between $\kappa_{r,s}$, the $(p-q) \times (p-q)$ sub-matrix of $[\kappa^{\alpha,\beta}]^{-1}$ and the $(p-q) \times (p-q)$ matrix inverse of $\kappa^{r,s}$, the covariance matrix of $Y^{(2)}$. The usual way of doing this is to distinguish the cumulants of Y from those of X by means of an overbar. Thus $\bar\kappa^{r,s} = \kappa^{r,s}$ is the covariance matrix of $Y^{(2)}$. In addition, $\bar\kappa^{i,j}$ is the covariance matrix of $Y^{(1)}$ and $\bar\kappa^{i,r} = 0$ by construction. The inverse matrix has elements $\bar\kappa_{r,s}$, $\bar\kappa_{i,j}$, $\bar\kappa_{i,r} = 0$, where $\bar\kappa^{r,s}$ $\bar\kappa_{s,t} = \delta^r_t$.

The linear transformation to orthogonalized variables $Y^{(1)}$ and $Y^{(2)}$ may be written

$$Y^i = X^i - \beta^i_r X^r; \qquad Y^r = X^r \tag{5.19}$$

where $\beta^i_r = \kappa^{i,s}\bar\kappa_{r,s}$ is the regression coefficient of X^i on X^r. This is the contravariant representation of the linear transformation but, for our present purposes, it will be shown that the covariant representation is the more convenient to work with. The covariance matrix of $Y^{(1)}$ is

$$\bar\kappa^{i,j} = \kappa^{i,j} - \kappa^{i,r}\kappa^{j,s}\bar\kappa_{r,s} = \{\kappa_{i,j}\}^{-1}. \tag{5.20}$$

Further, using formulae for the inverse of a partitioned matrix, we find

$$\begin{aligned} \kappa_{i,j} &= \bar\kappa_{i,j} \\ \kappa_{i,r} &= -\kappa_{i,j}\beta^j_r \\ \kappa_{r,s} &= \bar\kappa_{r,s} + \beta^i_r\beta^j_s\kappa_{i,j}. \end{aligned}$$

Hence the expressions for Y_i and Y_r are

$$\begin{aligned} Y_i &= \bar\kappa_{i,j}Y^j = \kappa_{i,j}Y^j = \kappa_{i,j}(X^j - \beta^j_r X^r) \\ &= \kappa_{i\alpha}X^\alpha = X_i \end{aligned}$$

and

$$\begin{aligned} Y_r &= \bar\kappa_{r,s}Y^s = \bar\kappa_{r,s}X^s = \bar\kappa_{r,s}\kappa^{s,\alpha}X_\alpha \\ &= X_r + \beta^i_r X_i. \end{aligned}$$

Thus, the covariant representation of the linear transformation (5.19) is

$$Y_i = X_i; \quad Y_r = X_r + \beta^i_r X_i. \tag{5.21}$$

It follows that the joint cumulants of $(Y^{(1)}, Y^{(2)})$ expressed in covariant form in terms of the cumulants of $(X^{(1)}, X^{(2)})$, are

$$\begin{array}{lll} \bar\kappa_i = \kappa_i, & \bar\kappa_{i,j} = \kappa_{i,j}, & \bar\kappa_{i,j,k} = \kappa_{i,j,k}, \ldots \\ \bar\kappa_{i,r} = 0, & \bar\kappa_{r,s} = \kappa_{r,s} + \beta^i_r\kappa_{i,s}[2] + \beta^i_r\beta^j_s & \\ \kappa_{i,j} = \kappa_{r,s} - \beta^i_r\beta^j_s\kappa_{i,j} & & \end{array}$$

$$
\begin{aligned}
\bar\kappa_{i,j,r} &= \kappa_{i,j,r} + \beta_r^k \kappa_{i,j,k} \\
\bar\kappa_{i,r,s} &= \kappa_{i,r,s} + \beta_r^j \kappa_{i,j,s}[2] + \beta_r^j \beta_s^k \kappa_{i,j,k} \\
\bar\kappa_{r,s,t} &= \kappa_{r,s,t} + \beta_r^i \kappa_{i,s,t}[3] + \beta_r^i \beta_s^j \kappa_{i,j,t}[3] + \beta_r^i \beta_s^j \beta_r^k \kappa_{i,j,k}.
\end{aligned}
\tag{5.22}
$$

The main point that requires emphasis here is that, in covariant form, the cumulants of $Y^{(1)}$ are the same as those of $X^{(1)}$ and they are unaffected by the Gram-Schmidt orthogonalization (5.19).

The contravariant expressions for the cumulants of $(Y^{(1)}, Y^{(2)})$ are

$$
\begin{aligned}
\bar\kappa^{i} &= \kappa^{i} - \beta_r^i \kappa^{r} \\
\bar\kappa^{i,j} &= \kappa^{i,j} - \beta_r^i \kappa^{r,j}[2] + \beta_r^i \beta_s^j \kappa^{r,s} = \kappa^{i,j} - \beta_r^i \beta_s^j \kappa^{r,s} \\
\bar\kappa^{i,j,k} &= \kappa^{i,j,k} - \beta_r^i \kappa^{r,j,h}[3] + \beta_r^i \beta_s^j \kappa^{r,s,k}[3] - \beta_r^i \beta_s^j \beta_r^k \kappa^{r,s,t} \\
\bar\kappa^{i,r} &= 0, \qquad\qquad \bar\kappa^{r,s} = \kappa^{r,s} \\
\bar\kappa^{i,j,r} &= \kappa^{i,j,r} - \beta_s^j \kappa^{i,r,s}[2] + \beta_s^i \beta_r^j \kappa^{r,s,t} \\
\bar\kappa^{i,r,s} &= \kappa^{i,r,s} - \beta_r^i \kappa^{r,s,t} \\
\bar\kappa^{r,s,t} &= \kappa^{r,s,t}
\end{aligned}
\tag{5.23}
$$

It is important to emphasize at this stage that $Y^{(1)}$ and $Y^{(2)}$ are not independent unless the strong assumptions stated in the first paragraph of this section apply. Among the cumulants given above, exactly two of the mixed third-order expressions are non-zero, namely $\kappa^{i,r,s}$ and $\kappa^{i,j,r}$. A non-zero value of $\kappa^{i,r,s}$ means that, although $E(Y^i|X^{(2)})$ has no linear dependence on $X^{(2)}$, there is some quadratic dependence on products $X^r X^s$. There is a close connection here with Tukey's (1949) test for non-additivity. Similarly, a non-zero value of $\kappa^{i,j,r}$ implies heterogeneity of covariance, namely that $\mathrm{cov}(Y^i, Y^j|X^{(2)})$ depends linearly or approximately linearly on $X^{(2)}$. See (5.28).

5.5.3 *Linear regression*

Suppose now, in the notation of the previous section, that the orthogonalized variables $Y^{(1)}$ and $Y^{(2)}$ are independent. It follows that the second- and higher-order conditional cumulants of $X^{(1)}$ given $X^{(2)}$ are the same as the unconditional cumulants of $Y^{(1)}$. These are given in covariant form in the first set of equations (5.22) and in contravariant form in (5.23). To complete the picture, we require only the conditional mean of $X^{(1)}$. We find from (5.19) that

$$E\{X^i|X^{(2)}\} = \kappa^i + \beta_r^i(x^r - \kappa^r) = \kappa^i + \kappa^{i,r} h_r(x^{(2)}; \bar\kappa)$$

where $h_r(x^{(2)}; \bar\kappa) = \bar\kappa_{r,s}(x^s - \kappa^s)$ is the first Hermite polynomial in the components of $X^{(2)}$.

The above expression may be written in covariant form as

$$E(X^i|X^{(2)}) = \bar\kappa^{i,j}\{\kappa_j - \kappa_{j,r}x^r\}$$

so that $\kappa_j - \kappa_{j,r}x^r$ is the covariant representation of the conditional mean of $X^{(1)}$.

These results may be summarized by a simple recipe for computing the conditional cumulants of $X^{(1)}$ after linear regression on $X^{(1)}$

(i) First compute the cumulants of $\kappa_{\alpha,\beta}X^\beta$, giving κ_i κ_r; $\kappa_{i,j}$, $\kappa_{i,r}$, $\kappa_{r,s}$; $\kappa_{i,j,k}$, ...
(ii) Compute $\bar\kappa^{i,j}$ the $p \times p$ matrix inverse of $\kappa_{i,j}$.
(iii) Replace κ_i by $\kappa_i - \kappa_{i,r}x^r$.
(iv) Raise all indices by multiplying by $\bar\kappa^{i,j}$ as often as necessary, giving

$$
\begin{aligned}
&\bar\kappa^{i,j}\{\kappa_j - \kappa_{j,r}x^r\} \\
&\bar\kappa^{i,k}\bar\kappa^{j,l}\kappa_{k,l} = \bar\kappa^{i,j} \\
&\bar\kappa^{i,j,k} = \bar\kappa^{i,l}\bar\kappa^{j,m}\bar\kappa^{k,n}\kappa_{l,m,n}
\end{aligned}
$$

and so on for the transformed cumulants.

5.6 Conditional cumulants

5.6.1 *Formal expansions*

In the previous section we considered the simple case where only the conditional mean of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ and none of the higher-order cumulants depends on the value of $x^{(2)}$. In this special case it was possible to compute exact expressions for the conditional cumulants. More generally, all of the conditional cumulants may depend, to some extent, on the value of $x^{(2)}$. Our aim in this section is to use the series expansion (5.4), together with the results of the previous three sections, to find formal series expansions for at least the first four conditional cumulants and, in principle at least, for all of the conditional cumulants.

To simplify matters, we assume that the initial approximating density for $(X^{(1)}, X^{(2)})$ is the product of two normal densities, $\phi_1(x^{(1)}; \lambda^{(1)})\phi_2(x^{(2)}; \lambda^{(2)})$. The advantage derived from this choice is that the Hermite tensors factor into products as described in Section 5.4.4. In practice, it is often sensible to choose the argument $\lambda^{(1)}$ of ϕ_1 to be the mean vector and covariance matrix of $X^{(1)}$ but, for reasons given in Section 5.2.1, we shall not do so for the moment.

Expansion (5.4) for the logarithm of the joint density of $X^{(1)}$ and $X^{(2)}$ may be written

$$
\begin{aligned}
&\log\phi_1(x^{(1)}; \lambda^{(1)}) + \log\phi_2(x^{(2)}; \lambda^{(2)}) + \eta^i h_i + \eta^r h_r \\
&\quad + \bigl\{\eta^{i,j} h_{ij} + \eta^{i,r} h_i h_r[2] + \eta^{r,s} h_{rs} + \eta^i\eta^j h_{i,j} + \eta^r\eta^s h_{r,s}\bigr\}/2! \\
&\quad + \bigl\{\bigl(\eta^{i,j,k} h_{ijk} + \eta^{i,j,r} h_{ij} h_r[3] + \eta^{i,r,s} h_i h_{rs}[3] + \eta^{r,s,t} h_{rst}\bigr) \\
&\qquad + \bigl(\eta^i\eta^{j,k} h_{i,jk} + \eta^i\eta^{j,r} h_{i,j} h_r[2] + \eta^r\eta^{i,s} h_i h_{r,s}[2] + \eta^r\eta^{s,t} h_{r,st}\bigr)[3]\bigr\}/3! \\
&+ \bigl\{\bigl(\eta^{i,j,k,l} h_{ijkl} + \eta^{i,j,k,r} h_{ijk} h_r[4] + \eta^{i,j,r,s} h_{ij} h_{rs}[6] + \eta^{i,r,s,t} h_i h_{rst}[4] \\
&\qquad\qquad\qquad + \eta^{r,s,t,u} h_{rstu}\bigr) \\
&\quad + \bigl(\eta^i\eta^{j,k,l} h_{i,jkl} + \eta^i\eta^{j,k,r} h_{i,jk} h_r[3] + \eta^i\eta^{j,r,s} h_{i,j} h_{rs}[3] \\
&\qquad + \eta^r\eta^{i,j,s} h_{ij} h_{r,s}[3] + \eta^r\eta^{i,s,t} h_i h_{r,st}[3] + \eta^r\eta^{s,t,u} h_{r,stu}\bigr)[4] \\
&\quad + \bigl(\eta^{i,j}\eta^{k,l} h_{ij,kl} + \eta^{i,j}\eta^{k,r} h_{ij,k} h_r[4] + \cdots + \eta^{r,s}\eta^{t,u} h_{rs,tu}\bigr)[3] \\
&\quad + \bigl(\eta^i\eta^j\eta^{k,l} h_{i,j,kl} + \eta^i\eta^r\eta^{j,s} h_{i,j} h_{r,s}[4] + \eta^r\eta^s\eta^{t,u} h_{r,s,tu}\bigr)[6]\bigr\}/4! \\
&+ \cdots . \qquad (5.24)
\end{aligned}
$$

This series may look a little complicated but in fact it has a very simple structure and is easy to extrapolate to higher-order terms. Such terms will be required later. Essentially every possible combination of terms appears, except that we have made use of the identities $h_{i,j,k} = 0$, $h_{i,j,k,l} = 0$ and so on.

On subtracting the logarithm of the marginal distribution of $X^{(2)}$, namely

$$
\log\phi_2(x^{(2)}; \lambda^{(2)}) + \eta^r h_r + \{\eta^{r,s} h_{rs} + \eta^r\eta^s h_{r,s}\}/2! + \cdots
$$

we find after collecting terms that the conditional log density of $X^{(1)}$ given $X^{(2)}$ has a formal Edgeworth expansion in which the first four cumulants are

$$
\begin{aligned}
&E(X^i | X^{(2)}) = \kappa^i \\
&\quad + \eta^{i,r} h_r + \eta^{i,r,s} h_{rs}/2! + \eta^{i,r,s,t} h_{rst}/3! + \eta^{i,r,s,t,u} h_{rstu}/4! + \cdots \\
&\quad + \eta^r\eta^{i,s} h_{r,s} + \eta^r\eta^{i,s,t} h_{r,st}/2! + \eta^r\eta^{i,s,t,u} h_{r,stu}/3! + \cdots \\
&\quad + \eta^{i,r}\eta^{s,t} h_{r,st}/2! + \eta^{i,r}\eta^{s,t,u} h_{r,stu}/3! + \eta^{r,s}\eta^{i,t,u} h_{rs,tu}/(2!\,2!) + \cdots \\
&\quad + \eta^r\eta^{s,t}\eta^{i,u} h_{r,st,u}/2! + \cdots
\end{aligned}
$$

$$\begin{aligned}\operatorname{cov}(X^i, X^j|X^{(2)}) &= \kappa^{i,j} + \eta^{i,j,r}h_r + \eta^{i,j,r,s}h_{rs}/2! + \eta^{i,j,r,s,t}h_{rst}/3! + \cdots \\ &\quad + \eta^r\eta^{i,j,s}h_{r,s} + \eta^r\eta^{i,j,s,t}h_{r,st}/2! + \cdots \\ &\quad + \eta^{i,r}\eta^{j,s}h_{r,s} + \eta^{i,r}\eta^{j,st}[2]h_{r,st}/2! + \eta^{r,s}\eta^{i,j,t}h_{rs,t}/2! + \cdots\end{aligned}$$

$$\begin{aligned}\operatorname{cum}(X^i, X^j, X^k|X^{(2)}) &= \kappa^{i,j,k} + \eta^{i,j,k,r}h_r + \eta^{i,j,k,r,s}h_{rs}/2! + \cdots \\ &\quad + \eta^r\eta^{i,j,k,s}h_{r,s} + \eta^r\eta^{i,j,k,s,t}h_{r,st}/2! + \cdots \\ &\quad + \eta^{i,r}\eta^{j,k,s}[3]h_{r,s} + \eta^{i,r}\eta^{j,k,s,t}[3]h_{r,st}/2! + \eta^{r,s}\eta^{i,j,k,t}h_{rs,t}/2! + \cdots \\ &\quad + \eta^{r,s,i}\eta^{j,k,t}[3]h_{rs,t}/2! + \cdots\end{aligned}$$

$$\begin{aligned}\operatorname{cum}(X^i, X^j, X^k, X^l|X^{(2)}) &= \kappa^{i,j,k,l} + \eta^{i,j,k,l,r}h_r + \eta^{i,j,k,l,r,s}h_{rs}/2! + \cdots \\ &\quad + \eta^r\eta^{i,j,k,l,s}h_{r,s} + \eta^{i,r}\eta^{j,k,l,s}[4]h_{r,s} + \eta^{i,j,r}\eta^{k,l,s}[3]h_{r,s} + \cdots \\ &\quad + \cdots\,. \end{aligned} \tag{5.25}$$

Of course, $\eta^{r,s,t} = \kappa^{r,s,t}$, $\eta^{r,s,i} = \kappa^{r,s,i}$, $\eta^{r,s} = \kappa^{r,s} - \lambda^{r,s}$ and so on, but the ηs have been retained in the above expansions in order to emphasize the essential simplicity of the formal series. We could, in fact, choose $\lambda^r = \kappa^r$, $\lambda^{r,s} = \kappa^{r,s}$ giving $\eta^r = 0$, $\eta^{r,s} = 0$. This choice eliminates many terms and greatly simplifies computations but it has the effect of destroying the essential simplicity of the pattern of terms in (5.25). Details are given in the following section.

5.6.2 *Asymptotic expansions*

If X is a standardized sum of n independent random variables, we may write the cumulants of X as 0, $\kappa^{\alpha,\beta}$, $n^{-1/2}\kappa^{\alpha,\beta,\gamma}$, $n^{-1}\kappa^{\alpha,\beta,\gamma,\delta}$ and so on. Suppose for simplicity that the components $X^{(1)}$ and $X^{(2)}$ are uncorrelated so that $\kappa^{i,r} = 0$. Then, from (5.25), the conditional cumulants of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ have the following expansions up to terms of order $O(n^{-1})$.

$$\begin{aligned}E(X^i|X^{(2)}) &= \kappa^i + n^{-1/2}\kappa^{i,r,s}h_{rs}/2! \\ &\qquad + n^{-1}\{\kappa^{i,r,s}\kappa^{t,u,v}h_{rs,tuv}/(3!\,2!) + \kappa^{i,r,s,t}h_{rst}/3!\}\end{aligned}$$

$$\begin{aligned}\operatorname{cov}(X^i, X^j|X^{(2)}) &= \kappa^{i,j} + n^{-1/2}\kappa^{i,j,r}h_r + n^{-1}\{\kappa^{i,j,r,s}h_{rs}/2! \\ &\qquad + \kappa^{i,j,r}\kappa^{s,t,u}h_{r,stu}/3! + \kappa^{i,r,s}\kappa^{j,t,u}h_{rs,tu}/(2!\,2!)\}\end{aligned}$$

$$\begin{aligned}\operatorname{cum}(X^i, X^j, X^k|X^{(2)}) &= n^{-1/2}\kappa^{i,j,k} \\ &\qquad + n^{-1}\{\kappa^{i,j,k,r}h_r + \kappa^{i,j,r}\kappa^{k,s,t}[3]h_{r,st}/2!\}\end{aligned}$$

$$\operatorname{cum}(X^i, X^j, X^k, X^l|X^{(2)}) = n^{-1}\{\kappa^{i,j,k,l} + \kappa^{i,j,r}\kappa^{k,l,s}[3]h_{r,s}\}.$$

In the above expansions, the Hermite tensors are to be calculated using the exact mean vector κ^r and covariance matrix $\kappa^{r,s}$ of $X^{(2)}$. Note that, to the order given, the conditional mean is a cubic function of $X^{(2)}$, the conditional covariances are quadratic, the conditional skewnesses are linear and the conditional kurtosis is constant, though not the same as the unconditional kurtosis. All higher-order cumulants are $O(n^{-3/2})$ or smaller.

5.7 Normalizing transformation

In the multivariate case, there is an infinite number of smooth transformations, $g(.)$, that make the distribution of $Y = g(X)$ normal to a high order of approximation. Here, in order to ensure a unique solution, at least up to choice of signs for the components, we ask that the transformation be *triangular*. In other words, Y^1 is required to be a function of X^1 alone, Y^2 is required to be a function of the pair X^1, X^2, and so on. Algebraically, this condition may be expressed by writing

$$\begin{aligned} Y^1 &= g_1(X^1) \\ Y^2 &= g_2(X^1, X^2) \\ Y^r &= g_r(X^1, \ldots, X^r) \qquad (r = 1, \ldots, p). \end{aligned}$$

To keep the algebra as simple as possible without making the construction trivial, it is assumed that X is a standardized random variable with cumulants

$$0, \qquad \delta^{ij}, \qquad n^{-1/2}\kappa^{i,j,k}, \qquad n^{-1}\kappa^{i,j,k,l}$$

and so on, decreasing in powers of $n^{1/2}$. Thus, X is standard normal to first order: the transformed variable, Y is required to be standard normal with error $O(n^{-3/2})$.

The derivation of the normalizing transformation is unusually tedious and rather unenlightening. For that reason, we content ourselves with a statement of the result, which looks as follows.

$$\begin{aligned} Y^i = X^i &- n^{-1/2}\left\{3\kappa^{i,r,s}h_{rs} + 3\kappa^{i,i,r}h_i h_r + \kappa^{i,i,i}h_{ii}\right\}/3! \\ &- n^{-1}\left\{4\kappa^{i,r,s,t}h_{rst} + 6\kappa^{i,i,r,s}h_{rs}h_i + 4\kappa^{i,i,i,r}h_{ii}h_r + \kappa^{i,i,i,i}h_{iii}\right\}/4! \\ &+ n^{-1}\left\{(36\kappa^{\alpha,i,r}\kappa^{\alpha,s,t} + 18\kappa^{i,i,r}\kappa^{i,s,t})h_{rst}\right. \\ &\quad + (18\kappa^{\alpha,i,i}\kappa^{\alpha,r,s} + 12\kappa^{i,i,i}\kappa^{i,r,s} + 36\kappa^{\alpha,i,r}\kappa^{\alpha,i,s} + 27\kappa^{i,i,r}\kappa^{i,i,s})h_{rs}h_i \\ &\quad + (36\kappa^{\alpha,i,i}\kappa^{\alpha,i,r} + 30\kappa^{i,i,i}\kappa^{i,i,r})h_{ii}h_r \\ &\quad \left. + (9\kappa^{\alpha,i,i}\kappa^{\alpha,i,i} + 8\kappa^{i,i,i}\kappa^{i,i,i})h_{iii}\right\}/72 \\ &+ n^{-1}\left\{(36\kappa^{\alpha,\beta,i}\kappa^{\alpha,\beta,r} + 36\kappa^{\alpha,i,i}\kappa^{\alpha,i,r} + 12\kappa^{i,i,i}\kappa^{i,i,r})h_r\right. \\ &\quad \left. + (18\kappa^{\alpha,\beta,i}\kappa^{\alpha,\beta,i} + 27\kappa^{\alpha,i,i}\kappa^{\alpha,i,i} + 10\kappa^{i,i,i}\kappa^{i,i,i})h_i\right\}/72 \end{aligned} \tag{5.26}$$

In the above expression, all sums run from 1 to $i - 1$. Greek letters repeated as superscripts are summed but Roman letters are not. In other words, $\kappa^{\alpha,\beta,i}\kappa^{\alpha,\beta,i}$ is a shorthand notation for

$$\sum_{\alpha=1}^{i-1}\sum_{\beta=1}^{i-1}\kappa^{\alpha,\beta,i}\kappa^{\alpha,\beta,i}.$$

In addition, the index i is regarded as a fixed number so that

$$\kappa^{i,i,i,r}h_{ii}h_r = \sum_{r=1}^{i-1}\kappa^{i,i,i,r}h_{ii}h_r.$$

Fortunately, the above polynomial transformation simplifies considerably in the univariate case because most of the terms are null. Reverting now to power notation, we find

$$Y = X - \rho_3(X^2 - 1)/6 - \rho_4(X^3 - 3X)/24 + \rho_3^2(4X^3 - 7X)/36$$

to be the polynomial transformation to normality. In this expression, we have inserted explicit expressions for the Hermite polynomials. In particular, $4X^3 - 7X$ occurs as the combination $4h_3 + 5h_1$.

5.8 Bibliographic notes

The terms ‘Edgeworth series’ and ‘Edgeworth expansion’ stem from the paper by Edgeworth (1905). Similar series had previously been investigated by Chebyshev, Charlier, and Thiele (1897): Edgeworth’s innovation was to group the series inversely by powers of the sample size rather than by the degree of the Hermite polynomial.

For a historical perspective on Edgeworth’s contribution to Statistics, see the discussion paper by Stigler (1978).

Jeffreys (1966, Section 2.68) derives the univariate Edgeworth expansion using techniques similar to those used here.

Wallace (1958) gives a useful discussion in the univariate case, of Edgeworth series for the density and Cornish-Fisher series for the percentage points. See also Cornish & Fisher (1937).

Proofs of the validity of Edgeworth series can be found in the books by Cramér (1937) and Feller (1971). Esseen (1945) and Bhattacharya & Ranga-Rao (1976) give extensions to the lattice case. See also Chambers (1967) or Bhattacharya & Ghosh (1978).

Skovgaard (1981a) discusses the conditions under which a transformed random variable has a density that can be approximated by an Edgeworth series.

Michel (1979) discusses regularity conditions required for the validity of Edgeworth expansions to conditional distributions.

The notation used here is essentially the same as that used by Amari & Kumon (1983): see also Amari (1985). Skovgaard (1986) prefers to use coordinate-concealing notation for conceptual reasons and to deal with the case where the eigenvalues of the covariance matrix may not tend to infinity at equal rates.

5.9 Further results and exercises 5

5.1 Show, under conditions to be stated, that if

$$f_X(x;\kappa) = f_0(x) + \eta^i f_i(x) + \eta^{ij} f_{ij}(x)/2! + \eta^{ijk} f_{ijk}(x)/3! + \cdots$$

then the moment generating function of $f_X(x;\kappa)$ is

$$M_0(\xi)\{1 + \xi_i\eta^i + \xi_i\xi_j\eta^{ij}/2! + \xi_i\xi_j\xi_k\eta^{ijk}/3! + \cdots\}$$

where $M_0(\xi)$ is the moment generating function of $f_0(x)$.

5.2 Using expansion (5.2) for the density, derive expansion (5.4) for the log density.

5.3 Give a heuristic explanation for the formal similarity of expansions (5.2) and those in Section 5.2.3.

5.4 Show that any generalized Hermite tensor involving β indices partitioned into α blocks, is of degree $\beta - 2\alpha - 2$ in x or is identically zero if $\beta - 2\alpha - 2 < 0$.

5.5 If $f_X(x)$ is the density function of $X^1, \ldots, X^p$, show that the density of

$$Y^r = a^r + a^r_i X^i$$

is

$$f_Y(y) = J f_X\{b^i_r(y^r - a^r)\},$$

where $b^i_r a^r_j = \delta^i_j$ and J is the determinant of b^i_r. Hence show that the partial derivatives of $\log f_X(x)$ are Cartesian tensors.

5.6 Show that the mode of a density that can be approximated by an Edgeworth series occurs at

$$\hat{x}^i = -\kappa^{i,j,k}\kappa_{j,k}/2 + O(n^{-3/2}).$$

5.7 Show that the median of a univariate density that can be approximated by an Edgeworth series occurs approximately at the point

$$\hat{x} = \frac{-\kappa_3}{6\kappa_2}.$$

Hence show that, to the same order of approximation, in the univariate case,

$$\frac{(\text{mean} - \text{median})}{(\text{mean} - \text{mode})} = \frac{1}{3}$$

(Haldane, 1942). See also Haldane (1948) for a discussion of medians of multivariate distributions.

5.8 Let X be a normal random variable with mean vector λ^r and covariance matrix $\lambda^{r,s}$. Define

$$h^r = h^r(x;\lambda), \quad h^{rs}(x;\lambda), \ldots$$

to be the Hermite tensors based on the same normal distribution, i.e.,

$$\begin{aligned} h^r &= x^r - \lambda^r \\ h^{rs} &= h^r h^s - \lambda^{r,s} \end{aligned}$$

and so on as in (5.7). Show that the random variables

$$h^r(X), \quad h^{rs}(X), \quad h^{rst}(X), \ldots$$

have zero mean and are uncorrelated.

5.9 Using the notation established in the previous exercise, show that

$$\begin{aligned} \operatorname{cum}\left(h^{rs}(X), h^{tu}(X), h^{vw}(X)\right) &= \lambda^{r,r}\lambda^{s,v}\lambda^{t,w}[8] \\ \operatorname{cum}\left(h^{r}(X), h^{s}(X), h^{tu}(X)\right) &= \lambda^{r,t}\lambda^{s,u}[2] \\ \operatorname{cum}\left(h^{r}(X), h^{st}(X), h^{uvw}(X)\right) &= \lambda^{r,u}\lambda^{s,v}\lambda^{t,w}[6]. \end{aligned}$$

Give an expression for the cumulant corresponding to an arbitrary partition of the indices.

5.10 Suppose now that X has cumulants κ^r, $\kappa^{r,s}$, $\kappa^{r,s,t}, \ldots$, and that the Hermite tensors are based on the normal density with mean λ^r and covariance matrix $\lambda^{r,s}$. Show that

$$\begin{aligned} E\{h^r(X)\} &= \eta^r \\ E\{h^{rs}(X)\} &= \eta^{rs} \\ E\{h^{rst}(X)\} &= \eta^{rst} \end{aligned}$$

and so on, where the ηs are defined in Section 5.2.1.

5.11 Using the notation established in the previous exercise, show that

$$\begin{aligned} \operatorname{cov}\left(h^r(X), h^s(X)\right) &= \kappa^{r,s} \\ \operatorname{cov}\left(h^r(X), h^{st}(X)\right) &= \kappa^{r,s,t} + \eta^s\kappa^{r,t}[2] \\ \operatorname{cov}\left(h^{rs}(X), h^{tu}(X)\right) &= \kappa^{r,s,t,u} + \eta^r\kappa^{s,t,u}[4] + \kappa^{r,t}\kappa^{s,u}[2] \\ &\quad + \eta^r\eta^t\kappa^{s,u}[4] \\ \operatorname{cov}\left(h^r(X), h^{stu}(X)\right) &= \kappa^{r,s,t,u} + \eta^s\kappa^{r,t,u}[3] + \kappa^{r,s}\eta^{t,u}[3] \\ &\quad + \eta^s\eta^t\kappa^{r,u}[3]. \end{aligned}$$

5.12 Generalize the result of the previous exercise by showing that the joint cumulant corresponding to an arbitrary set of Hermite tensors involves a sum over connecting partitions. Describe the rule that determines the contribution of each connecting partition.

5.13 Show that

$$\int h_1(x)h_2(x)h_3(x)\phi(x)\,dx = 6$$

$$\int h_1(x)h_2(x)h_3(x)h_4(x)\phi(x)\,dx = 264$$

where $h_r(x)$ is the standard univariate Hermite polynomial of degree r and $\phi(x)$ is the standard normal density. [Hint: use the tables of connecting partitions.]

5.14 More generally, using the notation of the previous exercise, show that, for $i > j > k$,

$$\int h_i(x)h_j(x)h_k(x)\phi(x)\,dx = \frac{i!\,j!\,k!}{\{\frac{1}{2}(j+k-i)\}!\,\{\frac{1}{2}(i+k-j)\}!\,\{\frac{1}{2}(i+j-k)\}!}$$

when $j + k - i$ is even and non-negative, and zero otherwise, (Jarrett, 1973, p. 26).

5.15 Using (5.26) or otherwise, show that in the univariate case, where X is a standardized sum with mean zero, unit variance and so on, then

$$Y^* = X - \rho_3 X^2/6 - \rho_4 X^3/24 + \rho_3^2 X^3/9$$

has mean $-\rho_3/6$ and standard deviation

$$1 - \rho_4/8 + 7\rho_3^2/36$$

when terms of order $O(n^{-3/2})$ are ignored. Show also that

$$\frac{Y^* + \rho_3/6}{1 - \rho_4/8 + 7\rho_3^2/36} \sim N(0,1) + O(n^{-3/2})$$

5.16 Taking the definition of Y^* as given in the previous exercise, show that

$$W/2 = (Y^*)^2 = X^2/2 - \rho_3 X^3/3! - \{\rho_4 - 3\rho_3^2\}X^4/4!$$

has mean given by

$$E(W) = 1 + (5\rho_3^2 - 3\rho_4)/12 = 1 + b/n.$$

Deduce that

$$\frac{W}{1 + b/n} \sim \chi_1^2 + O(n^{-3/2}).$$

5.17 Using the equation following (5.26), show by reversal of series, that X may be expressed as the following polynomial in the normally distributed random variable Y

$$X = Y + \rho_3(Y^2 - 1)/6 + \rho_4(Y^3 - 3Y)/24 - \rho_3^2(2Y^3 - 5Y)/36.$$

Hence, express the approximate percentage points of X in terms of standard normal percentage points (Cornish & Fisher, 1937).

5.18 Let $X = Y + Z$ where Y has density function $f_0(y)$ and Z is independent of Y with moments η^i, η^{ij}, η^{ijk}, ... Show formally, that the density of X is given by

$$f_X(x) = E_Z\{f_0(x - Z)\}.$$

Hence derive the series (5.2) by Taylor expansion of $f_0(x)$. By taking $\eta^i = 0$, $\eta^{i,j} = 0$, and $f_0(x) = \phi(x;\kappa)$, derive the usual Edgeworth expansion for the density of X, taking care to group terms in the appropriate manner. (Davis, 1976).

CHAPTER 6

Saddlepoint approximation

6.1 Introduction

One difficulty that arises as a result of approximating the density function or log density function is that the density is not invariant under affine transformation of X. Any approximation, therefore, ought to have similar non-invariant properties and this requirement raises difficulties when we work exclusively with tensors. For example, if $Y^r = a^r + a^r_i X^i$ is an affine transformation of X, then the density function of Y at y is

$$f_Y(y) = |A|^{-1} f_X(x)$$

where $x^i = b^i_r(y^r - a^r)$, $b^i_r a^r_j = \delta^i_j$ and $|A|$ is the determinant of a^r_i, assumed to be non-zero. In the terminology of Thomas (1965), the density is said to be an invariant of *weight* 1: ordinary invariants have weight zero.

One way to exploit the advantages of working with invariants, at least under affine transformation, is to work with probabilities of sets rather than with probability densities. The difficulty then is to specify the sets in an invariant manner, for example as functions of the invariant polynomials

$$(x^i - \kappa^i)(x^j - \kappa^j)\kappa_{i,j}, \quad (x^i - \kappa^i)(x^j - \kappa^j)(x^k - \kappa^k)\kappa_{i,j,k},$$
$$(x^i - \kappa^i)\kappa^{j,k}\kappa_{i,j,k}, \quad (x^i - \kappa^i)(x^j - \kappa^j)\kappa^{k,l}\kappa_{i,j,k,l}$$

and so on. Another way, more convenient in the present circumstances, is to specify the probability density with respect to a so-called carrier measure on the sample space. This can always be done in such a way that the approximating density is invariant and the carrier measure transforms in such a way as to absorb the Jacobian of the transformation.

To be more explicit, suppose that the density function of X is written as the product

$$f_X(x) = |\kappa^{i,j}|^{-1/2} g(x).$$

Now let Y be an affine transformation of X as before. The covariance matrix, being a contravariant tensor, transforms to

$$\bar\kappa^{r,s} = a^r_i a^s_j \kappa^{i,j}.$$

Then the density of Y at y is simply

$$f_Y(y) = |\bar\kappa^{r,s}|^{-1/2} g(x).$$

Thus, with the inverse square root of the determinant of the covariance matrix playing the role of carrier measure, the density $g(x)$ is invariant under affine transformation of coordinates.

From this viewpoint, the usual normal-theory approximation uses the constant carrier measure $(2\pi)^{-p/2}|\kappa^{i,j}|^{-1/2}$ together with the invariant quadratic approximation

$$(x^i - \kappa^i)(x^j - \kappa^j)\kappa_{i,j}/2$$

for the negative log density. The Edgeworth approximation retains the carrier measure but augments the approximation for the negative log density by the addition of further invariant polynomial terms, namely

$$-\kappa^{i,j,k}h_{ijk}/3! - \kappa^{i,j,k,l}h_{ijkl}/4! - \kappa^{i,j,k}\kappa^{l,m,n}h_{ijk,lmn}[10]/6! - \cdots$$

In the Edgeworth system of approximation, the carrier measure is taken as constant throughout the sample space. Thus the whole burden of approximation lies on the invariant series approximation. Furthermore, this property of constancy of the carrier measure is preserved only under transformations for which the Jacobian is constant, i.e. under affine transformation alone. It is therefore appropriate to investigate the possibility of using alternative systems of approximation using non-polynomial invariants together with carrier measures that are not constant throughout the sample space. In all cases considered, the carrier measure is the square root of the determinant of a covariant tensor or, equivalently so far as transformation properties are concerned, the inverse square root of the determinant of a contravariant tensor.

6.2 Legendre transformation of $K(\xi)$

6.2.1 *Definition*

In what follows, it is convenient to consider the set of ξ-values, denoted by Ξ, for which $K(\xi) < \infty$ as being, in a sense, complementary or dual to the sample space of possible averages of identically distributed Xs. Thus, $\mathcal{X}$ is the interior of the convex hull of the sample space appropriate to a single X. In many cases, the two sample spaces are identical, but, particularly in the case of discrete random variables there is an important technical distinction. Both $\mathcal{X}$ and Ξ are subsets of p-dimensional space. It is essential, however, to think of the spaces as distinct and qualitatively different: if we are contemplating the effect of linear transformation on X, then vectors in $\mathcal{X}$ are contravariant whereas vectors in Ξ are covariant and inner products are invariant. To keep the distinction clear, it is sometimes helpful to think of Ξ as a parameter space even though we have not yet introduced any parametric models in this context.

Corresponding to the cumulant generating function $K(\xi)$ defined on Ξ, there is a dual function $K^*(x)$ defined on $\mathcal{X}$ such that the derivatives $K^r(\xi)$ and $K^*_r(x)$ are functional inverses. In other words, the solution in ξ to the p equations

$$K^r(\xi) = x^r \tag{6.1}$$

is

$$\xi_i = K^*_i(x), \tag{6.2}$$

where $K^*_i(x)$ is the derivative of $K^*(x)$. The existence of a solution to (6.1) has been demonstrated by Daniels (1954); see also Barndorff-Nielsen (1978, Chapter 5), where $K^*(x)$ is called the *conjugate function* of $K(\xi)$. Uniqueness follows from the observation that $K(\xi)$ is a strictly convex function (Exercise 6.2).

The function $K^*(x)$ is known as the Legendre or Legendre-Fenchel transformation of $K(\xi)$: it occurs in the theory of large deviations (Ellis, 1985, p.220), where $-K^*(x)$ is also called the *entropy* or *point entropy* of the distribution $f_X(x)$.

In the literature on convex analysis, the term convex conjugate is also used (Fenchel, 1949; Rockafellar, 1970, Sections 12, 26).

An alternative, and in some ways preferable definition of $K^*(x)$ is

$$K^*(x) = \sup_\xi \{\xi_i x^i - K(\xi)\}. \tag{6.3}$$

To see that the two definitions are mutually consistent, we note that (6.1) or (6.2) determines the stationary points of the function in (6.3). Now write $h(x)$ for the maximum value, namely

$$h(x) = x^i K^*_i(x) - K(K^*_i(x))$$

Differentiation with respect to x^i gives $h_i(x) = K^*_i(x)$, showing that $K^*(x) = h(x)$+const. The constant is identified by (6.3) but not by the previous definition. The turning point gives a

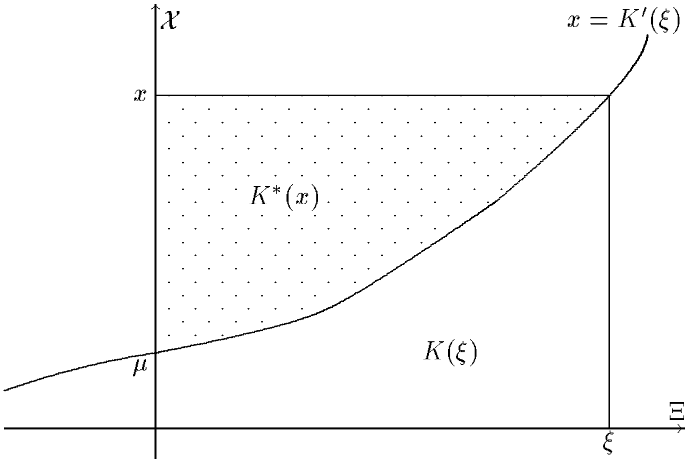

Figure 6.1: *Graphical construction of Legendre transformation in the univariate case.*

maximum because $K(\xi)$ is convex on Ξ. Note also that $K^*(x)$ is convex on $\mathcal{X}$ and achieves its minimum value of zero at $x^i = \kappa^i$.

Figure 6.1 illustrates graphically the construction of the Legendre transformation in the univariate case. The solid line gives the graph of $K'(\xi)$ against ξ and has positive gradient since $K''(\xi) > 0$. In the illustration, the intercept, which is equal to $E(X)$ is taken to be positive. The area under the curve from the origin to ξ is equal to the cumulant generating function $K(\xi) = \int_0^\xi K'(t)dt$. For any given value of x, ξx is the area of the rectangle whose opposite corners are at the origin and (ξ, x). Evidently, from the geometry of the diagram, $\xi x - K(\xi)$ is maximized at the value ξ satisfying $K'(\xi) = x$. The shaded area above the curve is equal to $\xi K'(\xi) - K(\xi)$. Regarded as a function of x, this is the Legendre transformation of $K(\xi)$. Equivalently, the graph reflected about the 45° line gives the inverse function of $K'(\xi)$. Integration from μ to x gives the area shaded above the line, showing that the two definitions given earlier are equivalent.

A similar geometrical description applies in the p-dimensional case with the two axes in Figure 6.1 replaced by the appropriate p-dimensional dual spaces. The graph is replaced by a mapping from Ξ to $\mathcal{X}$, but unfortunately, this is awkward to visualize even for $p = 2$ because four dimensions are involved.

Equations (6.1) or (6.2) identify corresponding points in Ξ and $\mathcal{X}$. In fact, every point in Ξ has a unique image in $\mathcal{X}$ given by (6.1) and conversely, every point in $\mathcal{X}$ has a unique image in Ξ given by (6.2). For example, the point $x^i = \kappa^i$ in $\mathcal{X}$ is identified with the point $\xi = 0$ in Ξ. The values of the two functions at these points are $K(0) = 0$ and $K^*(\kappa^i) = 0$ respectively.

Schematically, the relation between the first few derivatives of $K(\xi)$ and the dual function, $K^*(x)$ is as follows.

$$\begin{tikzcd}
K(\xi) \arrow[d, "\partial/\partial\xi_i"] & & K^*(x) \arrow[d, "\partial/\partial x^i"] \\
K^i(\xi) \arrow[rr, leftrightarrow, "\text{inverse function}"] \arrow[d, "\partial/\partial\xi_j"] & & K^*_i(x) \arrow[d, "\partial/\partial x^j"] \\
K^{ij}(\xi) \arrow[rr, leftrightarrow, "\text{matrix inverse}"] & & K^*_{ij}(x)
\end{tikzcd}$$

Apart, therefore, from the choice of constant at the integration step, it is evident that the procedure can be reversed and that the Legendre transformation of $K^*(x)$ is just $K(\xi)$, i.e. $K^{**} = K$.

Equivalently, we may show algebraically that

$$K(\xi) = \sup_x \{x^i \xi_i - K^*(x)\},$$

which is obvious from the diagram in Figure 6.1.

6.2.2 *Applications*

The following is a brief informal description mostly without proofs of the role of the Legendre transformation in approximating distributions. For a more formal and rigorous account of the theory in the context of large deviations, see Ellis (1985). For the connection with exponential family models, see Barndorff-Nielsen (1978).

Large deviations: The following inequality, which is central to much of the theory of large deviations in the univariate case, helps to explain the role played by the Legendre transformation. Let X be a real-valued random variable with mean μ and let $x > \mu$ be a given number. We show that

$$\mathrm{pr}(X \geq x) \leq \exp\{-K^*(x)\}. \tag{6.4}$$

To derive this inequality, we first write the required probability in the form

$$\mathrm{pr}(X - x \geq 0) = \mathrm{pr}\{\exp(\xi(X - x)) \geq 1\},$$

which is valid for all $\xi > 0$. Thus, since $\exp(\xi x) \geq H(x)$, where $H(.)$ is the Heaviside function, it follows that

$$\begin{aligned} \mathrm{pr}(X \geq x) &\leq \inf_{\xi > 0} \exp\{-\xi x + K(\xi)\} \\ &= \exp\{-K^*(x)\} \end{aligned}$$

and the inequality is proved.

More generally, for vector-valued X, it may be shown that, for any set A in $\mathcal{X}$,

$$\mathrm{pr}(X \in A) \leq \exp\{I(A)\}$$

where

$$I(A) = \sup_{x \in A} \{-K^*(x)\}$$

is called the entropy of the set A. Evidently, this is a generalization of the univariate inequality. The multivariate inequality follows from the univariate inequality together with the observation that $A \subset B$ implies $I(A) \leq I(B)$.

To see the relevance of the above inequalities to the theory of large deviations, let $X_1, \ldots, X_n$ be independent and identically distributed with cumulant generating function $K(\xi)$ and let $\bar{X}_n$ be the sample average. Large deviation theory is concerned mainly with approximations for the probability of the event $\bar{X}_n \geq x$ where $x > \mu$ is a fixed value independent of n. It is not difficult to see from the law of large numbers that the event in question has negligible probability for large n: in fact, for fixed x, the probability decreases exponentially fast as $n \to \infty$. The central result due to Cramér (1938) and Chernoff (1952), which determines the exponential rate of decrease, is that

$$n^{-1} \log \mathrm{pr}(\bar{X}_n \geq x) \to -K^*(x) \tag{6.5}$$

as $n \to \infty$. Note that $nK^*(x)$ is the Legendre transformation of $\bar{X}_n$, implying, in effect, that the inequality (6.4) becomes increasingly sharp as $n \to \infty$. As a numerical device for approximating tail probabilities however, the above limit is rarely of sufficient accuracy for statistical purposes. Better approximations are described in Section 6.2.6.

The Cramér-Chernoff key limit theorem may be extended to vector-valued random variables in the following way. Let $\bar{X}_n \in R^p$ and $A \subset R^p$ be open and convex. The condition that A be open can often be dropped. Then the required limit may be written

$$n^{-1} \log \mathrm{pr}(\bar{X}_n \in A) \to I(A) \tag{6.6}$$

as $n \to \infty$. Note that if $E(X)$ lies in the interior of A then $I(A)$ is zero, which is consistent with the law of large numbers, $\mathrm{pr}(\bar{X}_n \in A) \to 1$. Similarly, if $E(X)$ lies on the boundary of A, $I(A)$ is zero and $\mathrm{pr}(\bar{X}_n \in A) \to$const$\leq \frac{1}{2}$. In other words, if $\mathrm{pr}(\bar{X}_n \in A)$ tends to a finite limit, the limiting value cannot be deduced from the entropy limit (6.6).

Proofs of the limits (6.5) and (6.6) may be found in the books by Bahadur (1971) and Ellis (1985). Bahadur & Zabell (1979) give a number of generalizations.

Likelihood ratio statistics: Consider now the exponential family of distributions parameterized by θ in the form

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_X(x). \tag{6.7}$$

First note that since $K(.)$ is the cumulant generating function of the distribution $f_X(x)$, it follows that $f_X(x;\theta)$ is a probability distribution on $\mathcal{X}$ for all θ in Ξ. The cumulant generating function of $f_X(x;\theta)$ is $K(\xi+\theta) - K(\theta)$.

Suppose we wish to test the hypothesis that $\theta = 0$ based on an observed value x on X. The log likelihood ratio is

$$\log f_X(x;\theta) - \log f_X(x) = \theta_i x^i - K(\theta).$$

The maximized log likelihood ratio statistic, maximized over θ, gives

$$\sup_\theta \{\theta_i x^i - K(\theta)\} = K^*(x)$$

where the maximum occurs at $\hat{\theta}_i = K_i^*(x)$. In this setting, the Legendre transformation is none other than the maximized log likelihood ratio statistic. It follows then, by the usual asymptotic property of likelihood ratio statistics, that

$$2nK^*(\bar{X}_n) \sim \chi_p^2 + o(1)$$

and typically the error of approximation is $O(n^{-1})$. For a derivation of this result including an explicit expression for the $O(n^{-1})$ term, see Section 6.2.4.

Saddlepoint approximation: One reason for considering the Legendre transformation of $K(\xi)$ is that it is a function on $\mathcal{X}$, invariant under affine transformation, that is useful for approximating the density of X, and hence, the density of any one-to-one function of X. In fact, the saddlepoint approximation, derived in Section 6.3, for the density of X may be written in the form

$$(2\pi)^{-p/2} |K_{rs}^*(x)|^{1/2} \exp\{-K^*(x)\}. \tag{6.8}$$

This approximation arises from applying the inversion formula to $K(\xi)$, namely

$$f_X(x) = (2\pi i)^{-p} \int_{c-i\infty}^{c+i\infty} \exp\{K(\xi) - \xi_i x^i\}\, d\xi.$$

The saddlepoint of the integrand is given by (6.1) and occurs at a point for which ξ_i is real. Approximation of the exponent in the neighbourhood of the saddlepoint by a quadratic function and integrating gives (6.8). The approximation can be justified as an asymptotic approximation if X is an average or standardized sum of n independent random variables where n is assumed to be large.

From the point of view discussed in Section 6.1, approximation (6.8) uses the carrier measure $(2\pi)^{-p/2}|K^*_{rs}(x)|^{1/2}$ on $\mathcal{X}$, together with the invariant approximation $K^*(x)$ for the negative log density. Note that the carrier measure in this case is not constant on $\mathcal{X}$. This property permits us to consider non-linear transformations of X incorporating the non-constant Jacobian into the determinant and leaving the exponent in (6.8) unaffected. Of particular importance is the distribution of the random variable $Y_i = K^*_i(X)$. Direct substitution using (6.8) gives as the required approximation

$$(2\pi)^{-p/2}|K^{rs}(y)|^{1/2}\exp\{-K^*(x)\}.$$

In this case, $Y_i = \hat{\theta}_i$, the maximum likelihood estimate of the parameter in the exponential family derived from $f_X(x)$.

Greater accuracy in the asymptotic sense can be achieved by retaining the carrier measure and replacing the exponent in (6.8) by

$$-K^*(x) - (3\rho^*_4(x) - 4\rho^{*2}_{23}(x))/4! \tag{6.9}$$

where

$$\rho^*_4 = K^*_{ijkl}K^{*ij}K^{*kl}$$
$$\rho^{*2}_{23} = K^*_{ijk}K^*_{rst}K^{*ir}K^{*js}K^{*kt}$$

and K^{*ij}, the matrix inverse of K^*_{ij}, is identical to $K^{ij}(\xi)$ with ξ satisfying (6.2). The additional correction terms in (6.9) arise from expanding the exponent, $K(\xi) - \xi_i x^i$, about the saddlepoint as far as terms of degree four in ξ. These correction terms may alternatively be expressed in terms of the derivatives of $K(\xi)$ evaluated at the saddlepoint. Thus (6.9) becomes

$$-K^*(x) + (3\rho_4(\xi) - 3\rho^2_{13}(\xi) - 2\rho^2_{23}(\xi))/4!$$

where, for example,

$$\rho^2_{13}(\xi) = K_{ijk}K_{rst}K^{ij}K^{kr}K^{st}$$

and the functions are evaluated at the saddlepoint. Evidently, $\rho^2_{13}(\xi)$ evaluated at $\xi = 0$ is identical to ρ^2_{13}, and similarly for the remaining invariants.

6.2.3 *Some examples*

In general, it is not easy to compute the Legendre transformation of an arbitrary cumulant generating function in terms of known functions. The following examples show that the calculation is feasible in a number of cases.

Multivariate normal density: The cumulant generating function for the normal distribution is

$$K(\xi) = \kappa^i\xi_i + \kappa^{i,j}\xi_i\xi_j/2!.$$

The derivative, which is linear in ξ, may be inverted giving the Legendre transformation, which is

$$K^*(x) = (x^i - \kappa^i)(x^j - \kappa^j)\kappa_{i,j}/2!.$$

Evidently, since $K^*(x)$ is quadratic in x, K^*_{rs} is a constant equal to $\kappa_{r,s}$ and the saddlepoint approximation (6.8) is exact.

Gamma distribution: Suppose that X has the univariate gamma distribution with mean μ and variance μ^2/ν, where ν is the index or precision parameter. The cumulant generating function is

$$K(\xi) = -\nu\log(1 - \mu\xi/\nu).$$

Taylor expansion about $\xi = 0$ yields the higher-order cumulants $\kappa_r = (r-1)!\,\mu^r/\nu^{r-1}$. On solving the equation $K'(\xi) = x$, we find

$$\xi(x) = \frac{\nu}{\mu} - \frac{\nu}{x}.$$

Integration gives

$$K^*(x) = \nu\left\{\frac{x-\mu}{\mu} - \log\left(\frac{x}{\mu}\right)\right\},$$

which is exactly one half of the deviance contribution for gamma models (McCullagh & Nelder, 1983, p.153).

Since the second derivative of $K^*(x)$ is ν/x^2, it follows that the saddlepoint approximation for the density is

$$\frac{\left(\frac{\nu x}{\mu}\right)^{\nu} \exp\left(\frac{-\nu x}{\mu}\right)\frac{1}{x}\,dx}{(2\pi)^{1/2}\nu^{\nu-1/2}\exp(-\nu).}$$

This approximation differs from the exact density only to the extent that Stirling's approximation is used in the denominator in place of $\Gamma(\nu)$. In this example, this kind of defect can be corrected by choosing a multiplicative constant in such a way that the total integral is exactly one. This re-normalized saddlepoint approximation is exact for all gamma distributions.

Poisson distribution: The cumulant generating function is $\mu(\exp(\xi) - 1)$ showing that all cumulants are equal to μ. Following the procedure described above, we find

$$K^*(x) = x\log(x/\mu) - (x-\mu).$$

In the literature on log linear models, $2K^*(x)$ is more familiar as the deviance contribution or contribution to the likelihood ratio test statistic. The second derivative of K^* is just x^{-1}. Thus, the saddlepoint approximation gives

$$\frac{\exp(-\mu)\mu^x}{(2\pi)^{1/2}x^{x+1/2}\exp(-x).}$$

Again, this differs from the exact distribution in that Stirling's approximation has been used in place of $x!$. Unlike the previous example, however, re-normalization cannot be used to correct for this defect because x is not a constant and the *relative* weights given to the various values of x by the approximation are not exact.

6.2.4 *Transformation properties*

The Legendre transformation possesses a number of invariance properties that help to explain its role in approximating densities. First, let $Y^r = a^r + a^r_i X^i$ be an affine transformation of X. Any point in $\mathcal{X}$ can be identified either by its x-coordinates or by its y-coordinates, and the two are, in a sense, equivalent. The inverse transformation may be written $X^i = b^i_r(Y^r - a^r)$. It is easy to show that the Legendre transformation of $K_Y(\xi)$ is identical to the Legendre transformation of $K_X(\xi)$. To demonstrate this fact, we note first that

$$K_Y(\xi) = K_X(\xi_i a^i_j) + \xi_i a^i$$

Differentiation with respect to ξ_i and equating the derivative to y^i gives

$$K^i_X(\xi_i a^i_j) = b^i_j(y^j - a^j) = x^i.$$

Hence

$$\xi_i = b^j_i K^*_j(x),$$

which is the derivative with respect to y^i of $K^*(x)$.

For an alternative proof using definition (6.3) directly, see Exercise 6.9.

More generally, if a_i^r does not have full rank, it is easily shown that

$$K_X^*(x) \geq K_Y^*(y)$$

for all $x \in \mathcal{X}$ and y in the image set. To derive this inequality, we work directly from definition (6.3), giving

$$\begin{aligned} K_Y(y) &= \sup_\zeta \{\zeta_r y^r - K_Y(\zeta)\} \\ &= \sup_\zeta \{(\zeta_r a_i^r) x^i - K_X(\zeta_r a_i^r)\} \\ &\leq \sup_\xi \{\xi_i x^i - K_X(\xi)\} = K^*(x). \end{aligned}$$

This inequality is intuitively obvious from the interpretation of $K^*(x)$ as a log likelihood ratio statistic or *total deviance* in the sense of McCullagh & Nelder (1983).

A second important property of the Legendre transformation concerns its behaviour under exponential tilting of the density $f_X(x)$. The exponentially tilted density is

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_X(x)$$

where θ is the tilt parameter, otherwise known as the canonical parameter of the exponential family. The effect of exponential tilting on the negative log density is to transform from $-\log f_X(x)$ to

$$-\log f_X(x) - \theta_i x^i + K(\theta).$$

To see the effect on the Legendre transform, we write

$$K^*(x;\theta) = \sup_\xi \{\xi_i x^i - K(\xi+\theta) + K(\theta)\} \tag{6.10}$$

where $K(\xi+\theta) - K(\theta)$ is the cumulant generating function of the tilted density (6.7). An elementary calculation gives

$$K^*(x;\theta) = K^*(x) - \theta_i x^i + K(\theta) \tag{6.11}$$

so that, under this operation, the Legendre transformation behaves exactly like the negative log density.

Table 6.1 *Transformation properties of the Legendre transform*

Transformation	*Cumulant generating function*	*Legendre transform*
Identity	$K(\xi)$	$K^*(y)$
Convolution (sum)	$nK(\xi)$	$nK^*(y/n)$
Average	$nK(\xi/n)$	$nK^*(y)$
Location shift	$K(\xi) + \xi_i a^i$	$K^*(y-a)$
Exponential tilt	$K(\xi+\theta) - K(\theta)$	$K^*(y) - \theta_i y^i + K(\theta)$
Affine	$K(a_j^i \xi_i) + \xi_i a^i$	$K^*(b_j^i(y^j - a^j))$

The above transformation properties have important consequences when the Legendre transform is used to approximate the negative log density function. In particular, whatever the error incurred in using the saddlepoint approximation to $f_X(x)$, the same error occurs uniformly for all θ in the saddlepoint approximation to $f_X(x;\theta)$, the exponentially tilted density.

Table 6.1 provides a list of some of the more important transformation properties of the Legendre transformation.

6.2.5 *Expansion of the Legendre transformation*

In this section, we derive the Taylor expansion of $K^*(x)$ about $x^i = \kappa^i$. The expansion is useful for finding the approximate distribution of the random variables $K^*(X)$ and $K_i^*(X)$. As shown in Section 6.2.2, the first of these is a likelihood ratio statistic: the second is the maximum likelihood estimate of the canonical parameter in the exponential family generated by $f_X(x)$.

The first step is to expand the derivative of the cumulant generating function in a Taylor series about the origin giving

$$K^r(\xi) = \kappa^r + \kappa^{r,i}\xi_i + \kappa^{r,i,j}\xi_i\xi_j/2! + \kappa^{r,i,j,k}\xi_i\xi_j\xi_k/3! + \cdots .$$

On solving the equation $K^r(\xi) = x^r$ by reversal of series and substituting $z^i = x^i - \kappa^i$, we find

$$\begin{aligned}\xi_i = {}& \kappa_{i,r}z^r - \kappa_{i,r,s}z^r z^s/2! \\ & - \{\kappa_{i,r,s,t} - \kappa_{i,r,u}\kappa_{s,t,v}\kappa^{u,v}[3]\}z^r z^s z^t/3! + \cdots .\end{aligned}$$

After integrating term by term and after applying the boundary condition $K^*(\kappa^r) = 0$, we find

$$\begin{aligned}K^*(x) = {}& \kappa_{i,j}z^i z^j/2! - \kappa_{i,j,k}z^i z^j z^k/3! \\ & - \{\kappa_{i,j,k,l} - \kappa_{i,j,r}\kappa_{k,l,s}\kappa^{r,s}[3]\}z^i z^j z^k z^l/4! \\ & - \{\kappa_{i,j,k,l,m} - \kappa_{i,j,k,r}\kappa_{l,m,s}\kappa^{r,s}[10] \\ & \qquad + \kappa_{i,j,r}\kappa_{k,l,s}\kappa_{m,t,u}\kappa^{r,t}\kappa^{s,u}[15]\}z^i \cdots z^m/5! \\ & - \cdots .\end{aligned} \tag{6.12}$$

Note that the invariance of $K^*(x)$ follows immediately from the above expansion.

The close formal similarity between (6.12) and expansion (5.5) for the negative log density is remarkable: in fact, the two expansions are identical except that certain polynomial terms in (6.12) are replaced by generalized Hermite tensors in (5.5). For example, in (5.5) the coefficient of $\kappa_{i,j,k}$ is $-h^{ijk}/3!$ while the coefficient of $\kappa_{i,j,r}\kappa_{k,l,s}$ is $-h^{ijr,kls}[10]/6!$, which is quartic in x.

If $\bar{X}_n$ is the mean of n identically distributed Xs, then $Z^r = \bar{X}_n^r - \kappa^r$ is $O_p(n^{-1/2})$. A straightforward calculation using (6.12) reveals that the likelihood ratio statistic has expectation

$$\begin{aligned}E(2nK^*(\bar{X})) &= p\{1 + (3\bar{\rho}_{13}^2 + 2\bar{\rho}_{23}^2 - 3\bar{\rho}_4)/(12n)\} + O(n^{-2}) \\ &= p\{1 + b/n\} + O(n^{-2}),\end{aligned} \tag{6.13}$$

where the ρs are the invariant cumulants of X_1. It is evident from expansion (6.12) that $2nK^*(\bar{X}_n)$ has a limiting χ_p^2 distribution because the terms beyond the first are negligible. With a little extra effort, it can be shown that all cumulants of $\{1 + b/n\}^{-1}2nK^*(\bar{X}_n)$ are the same as those of χ_p^2 when terms of order $O(n^{-2})$ are neglected. This adjustment to the likelihood ratio statistic is known as the Bartlett factor.

The key idea in the proof is to write the likelihood ratio statistic as a quadratic form in derived variables Y. Thus

$$2nK^*(\bar{X}_n) = Y^r Y^s \kappa_{r,s}$$

where Y is a polynomial in Z. Rather intricate, but straightforward calculations then show that Y has third and fourth cumulants of orders $O(n^{-3/2})$ and $O(n^{-2})$ instead of the usual $O(n^{-1/2})$ and $O(n^{-1})$. Higher-order cumulants are $O(n^{-3/2})$ or smaller. To this order of approximation, therefore, Y has a normal distribution and $2nK^*(\bar{X}_n)$ has a non-central χ_p^2 distribution for which the rth cumulant is

$$\kappa_r = \{1 + b/n\}^r 2^{r-1}p(r-1)! + O(n^{-2}).$$

Thus the correction factor $1 + b/n$, derived as a correction for the mean, corrects all the cumulants simultaneously to the same order of approximation. See also Exercises 6.16 and 6.17.

Details of the proof are not very interesting but can be found in McCullagh (1984b, Section 7.1).

6.2.6 *Tail probabilities in the univariate case*

All of the calculations of the previous section apply equally to the univariate case, particularly (6.12) and (6.13). In order to find a suitably accurate approximation to the distribution function $\mathrm{pr}(X \le x)$, or to the tail probability $\mathrm{pr}(X \ge x)$, there are several possible lines of attack. The first and most obvious is to attempt to integrate the saddlepoint approximation directly. This method has the advantage of preserving the excellent properties of the saddlepoint approximation but it is cumbersome because the integration must be carried out numerically. An alternative method that is less cumbersome but retains high accuracy is to transform from X to a new scale defined by

$$T(x) = \pm\{2K^*(x)\}^{1/2}$$

where the sign of $T(x)$ is the same as that of $x - \mu$. The random variable $T(X)$ is sometimes called the *signed likelihood ratio statistic* although the term is appropriate only in the context of the exponential family of densities (6.7).

From the discussion in the previous section and from Exercises 6.16, 6.17, it may be seen that $T(X)$ is nearly normally distributed with mean and variance

$$\begin{aligned} E(T(X)) &\simeq -\rho_3/6 \\ \mathrm{var}(T(X)) &\simeq 1 + (14\rho_3^2 - 9\rho_4)/36, \end{aligned}$$

where $\rho_3 = \kappa_3/\kappa_2^{3/2}$ is the usual univariate standardized measure of skewness of X. In fact, the cumulants of $T(X)$ differ from those of the normal distribution having the above mean and variance, by $O(n^{-3/2})$ when X is a mean or total of n independent observations. Since $T(x)$ is increasing in x, it follows that

$$\begin{aligned} \mathrm{pr}(X \ge x) &= \mathrm{pr}\{T(X) \ge T(x)\} \\ &\simeq 1 - \Phi\left(\frac{T(x) + \rho_3/6}{1 + (14\rho_3^2 - 9\rho_4)/72}\right). \end{aligned} \tag{6.14}$$

A similar, but not identical, formula was previously given by Lugannani & Rice (1980). In fact, Daniels's (1987) version of the Lugannani-Rice formula may be written

$$\mathrm{pr}(X \ge x) \simeq 1 - \Phi(T(x)) + \phi(T(x))\left(\frac{1}{S(x)} - \frac{1}{T(x)}\right), \tag{6.15}$$

where $K'(\hat\xi) = x$ defines the saddlepoint and $S(x) = \hat\xi\{K''(\hat\xi)\}^{1/2}$ is a kind of Wald statistic. Note that the standard error is calculated under the supposition that the mean of X is at x rather than at μ.

Admittedly, approximation (6.14) has been derived in a rather dubious fashion. Neither the nature of the approximation nor the magnitude of the error have been indicated. In fact, the approximation may be justified as an asymptotic expansion if X is a sum or average of n independent random variables, in which case ρ_3 is $O(n^{-1/2})$ and ρ_4 is $O(n^{-1})$. The error incurred in using (6.14) for normal deviations is typically $O(n^{-3/2})$. By normal deviations, we mean values of x for which $K^*(x)$ is $O(1)$, or equivalently, values of x that deviate from $E(X)$ by a bounded multiple of the standard deviation. This range includes the bulk of the probability. Further adjustments to (6.14) are required in the case of discrete random variables: it often helps, for example, to make a correction for continuity.

For large deviations, the approximation of Lugannani & Rice has relative error of order $O(n^{-1})$. On the other hand, the relative error of (6.14) is $O(1)$, but despite this substantial asymptotic inferiority, (6.14) is surprisingly accurate even for moderately extreme tail probability calculations of the kind that occur in significance testing.

For further discussion of the above and related approximations, the reader is referred to Daniels (1987).

6.3 Derivation of the saddlepoint approximation

The most direct derivation of the saddlepoint approximation, by inversion of the cumulant generating function, was described in Section 6.2.2, if only briefly. A simpler method of derivation is to apply the Edgeworth approximation not to the density $f_X(x)$ directly, but to an appropriately chosen member of the conjugate family or exponential family

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_X(x). \tag{6.16}$$

We aim then, for each value of x, to chose the most advantageous value of θ in order to make the Edgeworth approximation to $f_X(x;\theta)$ as accurate as possible. This is achieved by choosing $\theta = \hat{\theta}(x)$ in such a way that x is at the mean of the conjugate density under $\hat{\theta}$. In other words, we chose $\hat{\theta}$ such that

$$K^r(\hat{\theta}) = x^r \quad \text{or} \quad \hat{\theta}_r = K^*_r(x).$$

As the notation suggests, $\hat{\theta}$ is the maximum likelihood estimate of θ based on x in the family (6.16).

From (5.5) or (5.10), the Edgeworth approximation for the log density at the mean may be written

$$-\tfrac{1}{2}p\log(2\pi) - \tfrac{1}{2}\log|\kappa^{r,s}| + (3\rho_4 - 3\rho_{13}^2 - 2\rho_{23}^2)/4! + \cdots .$$

Taking logs in (6.16) gives

$$\log f_X(x) = -K^*(x) + \log f_X(x;\hat{\theta}).$$

Applying the Edgeworth expansion to the second term on the right gives

$$\begin{aligned}
\log f_X(x) &= -\tfrac{1}{2}p\log(2\pi) - \tfrac{1}{2}\log|K^{rs}(\hat{\theta})| \\
&\qquad - K^*(x) + (3\rho_4(\hat{\theta}) - 3\rho_{13}^2(\hat{\theta}) - 2\rho_{23}^2(\hat{\theta}))/4! + \cdots \\
&= -\tfrac{1}{2}p\log(2\pi) + \tfrac{1}{2}\log|K^*_{rs}(x)| \\
&\qquad - K^*(x) - (3\rho^*_4(x) - 4\rho^{*2}_{23}(x))/4! + \cdots .
\end{aligned}$$

The first three terms above constitute the saddlepoint approximation for the log density: the fourth term is a correction term that is $O(n^{-1})$ in large samples. Terms that are ignored are $O(n^{-2})$.

Alternatively, and sometimes preferably, we may write the approximate density function as

$$(2\pi c)^{-p/2}|K^*_{rs}|\exp\{-K^*(x)\} \tag{6.17}$$

where c is a constant chosen to make the integral equal to one. To a first order of approximation, we may write

$$\log(c) = (3\bar{\rho}_{13}^2 + 2\bar{\rho}_{23}^2 - 3\bar{\rho}_4)/12.$$

Thus, $\log(c)$ is just the Bartlett adjustment term defined in (6.13). The error of approximation is $O(n^{-3/2})$ when the above approximation is used for the constant of integration.

The advantages of the saddlepoint approximation over the Edgeworth series are mainly connected with accuracy. Although both approximations are asymptotic, the saddlepoint approximation is often sufficiently accurate for statistical purposes even when n is small, less than 10, say. In addition, the saddlepoint approximation retains high relative accuracy over the whole range of possible values of x. The Edgeworth approximation, on the other hand, is valid only for values of $\bar{X}$ that deviate from $E(\bar{X})$ by $O(n^{-1/2})$. The implication of this restriction is that the Edgeworth series may not be of adequate accuracy to judge the probability of unusual events.

On the negative side, the saddlepoint approximation applies to the density, and if tail probability calculations are required, integration is necessary. Unfortunately, the saddlepoint approximation, unlike the Edgeworth series, cannot usually be integrated analytically. Numerical integration is one answer, but this is often cumbersome. An alternative and more convenient solution is described in Section 6.2.6. A second argument against the saddlepoint approximation in favour of the Edgeworth series is that in order to compute the saddlepoint, it is necessary to have an explicit formula for the cumulant generating function. To use the Edgeworth series, on the other hand, it is necessary only to know the first few cumulants, and these can often be computed without knowing the generating function. There may, in fact, be no closed form expression for the generating function: see, for example, Exercise 2.30. In short, the Edgeworth series is often easier to use in practice but is usually inferior in terms of accuracy, particularly in the far tails of the distribution.

6.4 Approximation to conditional distributions

6.4.1 *Conditional density*

Suppose that we require an approximation for the conditional distribution of a statistic X_2 given that $X_1 = x_1$. Both components may be vector valued. Calculations of this kind arise in a number of important areas of application. The following are a few examples.

(i) Elimination of nuisance parameters by conditioning, particularly where matched retrospective designs are used to study factors that influence the incidence of rare diseases (Breslow & Day, 1980, Chapter 7).

(ii) Conditioning to take account of the observed value of an ancillary statistic (Cox, 1958).

(iii) Testing for goodness of fit when the model to be tested contains unknown parameters (McCullagh, 1985).

The simplest and most natural way to proceed from the cumulant generating functions $K_{X_1X_2}(.)$ and $K_{X_1}(.)$ is to compute the corresponding Legendre transformations, $K^*_{X_1X_2}(x_1, x_2)$ and $K^*_{X_1}(x_1)$. The saddlepoint approximation is then used twice, once for the joint density and once for the marginal density of X_1. Thus,

$$\begin{aligned} f_{X_1X_2}(x_1, x_2) &\simeq c_{12}|K^*_{X_1X_2;rs}|^{1/2}\exp\{-K^*_{X_1X_2}(x_1, x_2)\} \\ f_{X_1}(x_1) &\simeq c_1|K^*_{X_1;rs}|^{1/2}\exp\{-K^*_{X_1}(x_1)\}, \end{aligned}$$

where c_{12} and c_1 are normalizing constants. On subtracting the approximate log densities, we find

$$\begin{aligned} \log f_{X_2|X_1}(x_2|x_1) \simeq{} & \log c_{12} - \log c_1 \\ & + \tfrac{1}{2}\log|K^*_{X_1X_2;rs}| - \tfrac{1}{2}\log|K^*_{X_1;rs}| \\ & - K^*_{X_1X_2}(x_1, x_2) + K^*_{X_1}(x_1). \end{aligned} \tag{6.18}$$

In large samples, the error of approximation is $O(n^{-3/2})$ provided that the constants of integration c_1 and c_{12} are appropriately chosen.

Approximation (6.18) is sometimes called the *double saddlepoint approximation*. It is not the same as applying the saddlepoint approximation directly to the conditional cumulant generating function of X_2 given $X_1 = x_1$. For an example illustrating the differences, see Exercises 6.18–6.20.

6.4.2 *Conditional tail probability*

Suppose now that X_2 is a scalar and that we require an approximation to the conditional tail probability

$$\text{pr}(X_2 \geq x_2|X_1 = x_1).$$

Expression (6.15) gives the required unconditional tail probability: the surprising fact is that the same expression, suitably re-interpreted, applies equally to conditional tail probabilities.

In the double saddlepoint approximation, there are, of course, two saddlepoints, one for the joint distribution of (X_1, X_2) and one for the marginal distribution of X_1. These are defined by

$$K^r(\hat{\xi}_1, \hat{\xi}_2) = x_1^r, \quad r = 1, \ldots, p-1; \quad K^p(\hat{\xi}_1, \hat{\xi}_2) = x_2$$

for the joint distribution, and

$$K^r(\tilde{\xi}_1, 0) = x_1^r, \quad r = 1, \ldots, p-1$$

for the marginal distribution of the $p-1$ components of X_1. In the above expressions, ξ_1 has $p-1$ components and ξ_2 is a scalar, corresponding to the partition of X.

The signed likelihood ratio statistic $T = T(x_2|x_1)$ is most conveniently expressed in terms of the two Legendre transformations, giving

$$T = \operatorname{sign}(\hat{\xi}_2)\{K^*_{X_1X_2}(x_1, x_2) - K^*_{X_1}(x_1)\}.$$

Further, define the generalized conditional variance, $V = V(x_2|x_1)$, by the determinant ratio

$$V = \frac{|K^{rs}(\hat{\xi}_1, \hat{\xi}_2)|}{|K^{rs}(\tilde{\xi}_1, 0)|} = \frac{|K^*_{X_1;rs}|}{|K^*_{X_1X_2;rs}|}.$$

Using these expressions, the double saddlepoint approximation becomes

$$cV^{-1/2}\exp(-T^2/2).$$

In the conditional sample space, $V^{-1/2}$ plays the role of carrier measure and the exponent is invariant. The conditional tail probability is given by (6.15) using the Wald statistic

$$S(x) = \hat{\xi}_2 V^{1/2}.$$

This important result is due to I. Skovgaard (1986, personal communication) and is given here without proof.

6.5 Bibliographic notes

The following is a very brief description of a few key references. Further references can be found cited in these papers.

The Cramér-Chernoff large deviation result is discussed in greater detail by Bahadur (1971) and by Ellis (1985).

The derivation of the saddlepoint approximation by using an Edgeworth expansion for the exponentially tilted density goes back to the work of Esscher (1932), Cramér (1938), Chernoff (1952) and Bahadur & Ranga-Rao (1960). In a series of important papers, Daniels (1954, 1980, 1983) develops the saddlepoint method, derives the conditions under which the re-normalized approximation is exact and finds approximations for the density of a ratio and the solution of an estimating equation. Barndorff-Nielsen & Cox (1979) discuss double saddlepoint approximation as a device for approximating to conditional likelihoods.

The Legendre transformation plays an important role in the literature on large deviations, and is emphasized by Ellis (1985), Bahadur & Zabell (1979) and also to an extent, Daniels (1960).

The relationship between the Bartlett adjustment factor and the normalization factor in the saddlepoint formula is discussed by Barndorff-Nielsen & Cox (1984).

Tail probability calculations are discussed by Lugannani & Rice (1980), Robinson (1982) and by Daniels (1987).

6.6 Further results and exercises 6

6.1 Show that the array

$$M^{ij}(\xi) = E\{X^i X^j \exp(\xi_r x^r)\}$$

is positive definite for each ξ. Hence deduce that the function $M(\xi)$ is convex. Under what conditions is the inequality strict?

6.2 By using Hölder's inequality, show for any $0 \leq \lambda \leq 1$, that

$$K(\lambda\xi_1 + (1-\lambda)\xi_2) \leq \lambda K(\xi_1) + (1-\lambda)K(\xi_2)$$

proving that $K(\xi)$ is a convex function.

6.3 Prove directly that $K^{rs}(\xi)$ is positive definite for each ξ in Ξ. Hence deduce that $K^*(x)$ is a convex function on $\mathcal{X}$.

6.4 Prove that $K^*(x) \geq 0$, with equality only if $x^i = \kappa^i$.

6.5 Prove the following extension of inequality (6.4) for vector-valued X

$$\mathrm{pr}(X \in A) \leq \exp\{I(A)\}$$

where $I(A) = \sup_{x \in A}\{-K^*(x)\}$.

6.6 From the entropy limit (6.6), deduce the law of large numbers.

6.7 Show that the Legendre transformation of $Y = X_1 + \cdots + X_n$ is $nK^*(y/n)$, where the Xs are *i.i.d.* with Legendre transformation $K^*(x)$.

6.8 Show that, for each θ in Ξ,

$$\exp(\theta_i x^i - K(\theta))f_X(x)$$

is a distribution on $\mathcal{X}$. Find its cumulant generating function and the Legendre transformation.

6.9 From the definition

$$K^*_Y(y) = \sup_{\xi}\{\xi_i y^i - K_Y(\xi)\}$$

show that the Legendre transformation is invariant under affine transformation of coordinates on $\mathcal{X}$.

6.10 By writing ξ_i as a polynomial in z

$$\xi_i = a_{ir}z^r + a_{irs}z^r z^s/2! + a_{irst}z^r z^s z^t/3! + \cdots,$$

solve the equation

$$\kappa^{r,i}\xi_i + \kappa^{r,i,j}\xi_i\xi_j/2! + \kappa^{r,i,j,k}\xi_i\xi_j\xi_k/3! + \cdots \quad = z^r$$

by series reversal. Hence derive expansion (6.12) for $K^*(x)$.

6.11 Using expansion (6.12), find the mean of $2nK^*(\bar{X}_n)$ up to and including terms that are of order $O(n^{-1})$.

6.12 Show that the matrix inverse of $K^{rs}(\xi)$ is $K^*_{rs}(x)$, where $x^r = K^r(\xi)$ corresponds to the saddlepoint.

6.13 Using (6.12) or otherwise, show that, for each x in $\mathcal{X}$,

$$\begin{aligned} K^*_{ijk}(x) &= -K^{rst}K_{ri}K_{sj}K_{tk} \\ K^*_{ijkl}(x) &= -\{K^{rstu} - K^{rsv}K^{tuw}K_{vw}[3]\}K_{ri}K_{sj}K_{tk}K_{ul} \end{aligned}$$

where all functions on the right are evaluated at $\xi_r = K^*_r(x)$, the saddlepoint image of x.

6.14 Using the results given in the previous exercise, show, using the notation of Section 6.3, that

$$\rho_{13}^2(\hat\theta) = \rho_{13}^{*2}(x) \qquad \rho_{23}^2(\hat\theta) = \rho_{23}^{*2}(x)$$
$$\rho_4(\hat\theta) = -\rho_4^*(x) + \rho_{13}^{*2}(x) + 2\rho_{23}^{*2}(x).$$

6.15 By using the expansion for ξ_i given in Section 6.2.4, show that the maximum likelihood estimate of θ based on $\bar{X}_n$ in the exponential family (6.14) has bias

$$E(n^{1/2}\hat\theta_r) = -\tfrac{1}{2}n^{-1/2}\kappa^{i,j,k}\kappa_{i,r}\kappa_{j,k} + O(n^{-3/2}).$$

6.16 Show that if $Z^r = \bar{X}_n^r - \kappa^r$ and

$$\begin{aligned} n^{-1/2}Y^r &= Z^r - \kappa^{r,s,t}\kappa_{s,i}\kappa_{t,j}Z^iZ^j/6 \\ &+ \{8\kappa^{r,s,t}\kappa^{u,v,w}\kappa_{s,i}\kappa_{t,u}\kappa_{v,j}\kappa_{w,k} - 3\kappa^{r,s,t,u}\kappa_{s,i}\kappa_{t,j}\kappa_{u,k}\}Z^iZ^jZ^k/72 \end{aligned}$$

then $Y = O_p(1)$ and

$$2nK^*(\bar{X}_n) = Y^rY^s\kappa_{r,s} + O(n^{-2}).$$

6.17 Show that Y^r defined in the previous exercise has third cumulant of order $O(n^{-3/2})$ and fourth cumulant of order $O(n^{-1})$. Hence show that $2nK^*(\bar{X}_n)$ has a non-central χ_p^2 distribution for which the rth cumulant is

$$\{1 + b/n\}^r 2^{r-1}(r-1)!p + O(n^{-2}).$$

Find an expression for b in terms of the invariant cumulants of X.

6.18 Show that, in the case of the binomial distribution with index m and parameter π, the Legendre transformation is

$$y\log\left(\frac{y}{\mu}\right) + (m-y)\log\left(\frac{m-y}{m-\mu}\right)$$

where $\mu = m\pi$. Hence show that the saddlepoint approximation is

$$\frac{\pi^y(1-\pi)^{m-y}m^{m+1/2}}{(2\pi)^{1/2}y^{y+1/2}(m-y)^{m-y+1/2}}.$$

In what circumstances is the saddlepoint approximation accurate? Derive the above as a double saddlepoint approximation to the conditional distribution of Y_1 given $Y_1 + Y_2 = m$, where the Ys are independent Poisson random variables.

6.19 Let X_1, X_2 be independent exponential random variables with common mean μ. Show that the Legendre transformation of the joint cumulant generating function is

$$K^*(x_1, x_2; \mu) = \frac{x_1 + x_2 - 2\mu}{\mu} - \log\left(\frac{x_1}{\mu}\right) - \log\left(\frac{x_2}{\mu}\right).$$

Show also that the Legendre transformation of the cumulant generating transformation of $\bar{X}$ is

$$K^*(\bar{x}; \mu) = 2\left(\frac{\bar{x} - \mu}{\mu}\right) - 2\log\left(\frac{\bar{x}}{\mu}\right).$$

Hence derive the double saddlepoint approximation for the conditional distribution of X_1 given that $X_1 + X_2 = 1$. Show that the re-normalized double saddlepoint approximation is exact.

6.20 Extend the results described in the previous exercise to gamma random variables having mean μ and indices ν_1, ν_2. Replace $\bar{X}$ by an appropriately weighted mean.

6.21 In the notation of Exercise 6.18, show that the second derivative of $K^*(x, 1-x; 1/2)$ at $x = 1/2$ is 8, whereas the conditional variance of X_1 given that $X_1 + X_2 = 1$ is 1/12. Hence deduce that the double saddlepoint approximation to the conditional density of X_1 is not the same as applying the ordinary saddlepoint approximation directly to the conditional cumulant generating function of X_1.

6.22 Using the asymptotic expansion for the normal tail probability

$$1 - \Phi(x) \simeq \frac{\phi(x)}{x} \qquad\qquad x \to \infty$$

and taking $x > E(X)$, show, using (6.14), that

$$n^{-1} \log \mathrm{pr}\{\bar{X}_n > x\} \to -K^*(x)$$

as $n \to \infty$, where $\bar{X}_n$ is the average of n independent and identically distributed random variables. By retaining further terms in the expansion, find the rate of convergence to the entropy limit (6.5).

6.23 Using (6.11), show that the Legendre transform $K^*(x;\theta)$ of the exponentially tilted density satisfies the partial differential equations

$$\frac{\partial K^*(x;\theta)}{\partial x^r} = \hat{\theta}_r(x) - \theta_r$$

$$-\frac{\partial K^*(x;\theta)}{\partial \theta_i} = x^i - K^i(\theta).$$

Hence show that in the univariate case,

$$K^*(x;\theta) = \int_\mu^x \frac{x-t}{v(t)}\,dt,$$

where $\mu = K'(\theta)$ and $v(\mu) = K''(\theta)$, (Wedderburn, 1974; Nelder & Pregibon, 1986).

6.24 By using Taylor expansions for $S(x)$ and $T(x)$ in (6.15), show that, for normal deviations, the tail probability (6.15) reduces to

$$1 - \Phi(T) + \phi(T)\left(-\frac{\rho_3}{6} + \frac{5\rho_3^2 - 3\rho_4}{24}T\right) + O(n^{-3/2}).$$

Hence deduce (6.14).

6.25 Let X be a random variable with density function

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_0(x)$$

depending on the unknown parameter θ. Let θ have Jeffreys's prior density

$$\pi(\theta) = |K^{rs}(\theta)|^{1/2}.$$

Using Bayes's theorem, show that the posterior density for θ given x is approximately

$$\pi(\theta|x) \simeq c \exp\{\theta_i x^i - K(\theta) - K^*(x)\}|K^{rs}(\theta)|^{1/2},$$

whereas the density of the random variable $\hat{\theta}$ is approximately

$$p(\hat{\theta}|\theta) \simeq c \exp\{\theta_i x^i - K(\theta) - K^*(x)\}|K^{rs}(\hat{\theta})|^{1/2}.$$

In the latter expression x is considered to be a function of $\hat{\theta}$.

Find expressions for the constants in both cases.

6.26 Consider the conjugate density $f_X(x;\theta)$ as given in the previous exercise, where $K(\theta)$ is the cumulant generating function for $f_0(x)$ and $K^*(x)$ is its Legendre transform. Show that

$$E_\theta\left\{\log\left(\frac{f_X(X;\theta)}{f_0(X)}\right)\right\} = K^*(E_\theta(X))$$

where $E_\theta(.)$ denotes expectation under the conjugate density. [In this context, $K^*(E_\theta(X))$ is sometimes called the *Kullback-Leibler distance* between the conjugate density and the original density.]

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

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

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

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

we find the following expansions.

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

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,

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

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

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

so that the log likelihood derivatives transform to $\bar U_r$, $\bar U_{rs}$, $\bar U_{rst}$ satisfying

$$\begin{aligned} U_r &= a_r^i\bar U_i \\ U_{rs} &= a_r^i a_s^j \bar U_{ij} + a_{rs}^i \bar U_i \\ U_{rst} &= a_r^i a_s^j a_t^k \bar U_{ijk} + a_r^i a_{st}^j \bar U_{ij}[3] + a_{rst}^i \bar U_i \end{aligned} \tag{7.7}$$

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

$$\mathbf{U} = \mathbf{B}\mathbf{V}. \tag{7.8}$$

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 $a_i^r = \delta_i^r$, 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 $a_r^i \neq \delta_r^i$, we may write $\mathbf{B} = \mathbf{A}\bar{\mathbf{B}}\mathbf{A}^{*-1}$, where $\mathbf{A}^*$ is a direct product matrix involving a_r^i alone (Exercise 7.8). Premultiplication of (7.8) by $\mathbf{A}^{-1}$ gives

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

Since, by definition, the transformed Vs satisfy $\bar{\mathbf{U}} = \bar{\mathbf{B}}\bar{\mathbf{V}}$, it follows that if $a_i^r = \delta_i^r$, then $\bar{\mathbf{V}} = \mathbf{V}$. Thus, the Vs are unaffected by non-linear transformation in which the leading coefficient is δ_i^r. 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_{rs}^i, $\tau_{rst}^i, \ldots$ that transforms like the βs will generate a sequence of symmetric tensors when inserted into (7.6). One possibility is to define

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

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_{st}^r, $\beta_{stu}^r, \ldots$ or the alternative sequence τ_{st}^r, $\tau_{stu}^r, \ldots$ 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

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

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

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

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

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

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

In the univariate case we may write

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

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

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

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

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

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

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

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, $U_rU_s\kappa^{r,s}$, 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.

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

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 $\mathbf{W} = \operatorname{diag}(\mu_i)$. 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

$$n\kappa_{r,s} = \sum_i x_r^i x_s^i \mu_i = \mathbf{X}^T\mathbf{W}\mathbf{X}, \qquad \kappa^{r,s} = n(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$
$$n\kappa_{r,s,t} = \sum_i x_r^i x_s^i x_t^i \mu_i, \qquad n\kappa_{r,s,t,u} = \sum_i x_r^i x_s^i x_t^i x_u^i \mu_i.$$

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

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

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

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

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

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

where $\mathbf{Y} = \mathrm{diag}\{y_1, \ldots, y_n\}$ 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

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

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,

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

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

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

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.

CHAPTER 8

Ancillary statistics

8.1 Introduction

In problems of parametric inference, an ancillary statistic is one whose distribution is the same for all parameter values. This is the simplest definition and the one that will be used here. Certain other authors, e.g. Cox & Hinkley (1974), require that the statistic in question be a component of the minimal sufficient statistic, and this is often a reasonable requirement if only to ensure uniqueness in certain special cases. More complicated definitions are required to deal satisfactorily with the case where nuisance parameters are present. For simplicity, nuisance parameters are not considered here.

Ancillary statistics usually arise in one of two ways. One simple example of each will suffice to show that no frequency theory of inference can be considered entirely satisfactory or complete if ancillary statistics are ignored. Even an approximate theory must attempt to come to grips with conditional inference if approximations beyond the first order are contemplated.

The first and most common way in which an ancillary statistic occurs is as a random sample size or other random index identifying the experiment actually performed. Suppose that N is a random variable whose probability distribution is

$$\mathrm{pr}(N = 1) = 0.5, \quad \mathrm{pr}(N = 100) = 0.5.$$

Conditionally on $N = n$, the random variables $Y_1, \ldots, Y_n$ are normally distributed with mean θ and unit variance. On the basis of the observed values n and $y_1, \ldots, y_n$, probabilistic statements concerning the value of θ are required.

The same problem can be posed in a number of ways. For instance, either of two machines can be used to measure a certain quantity θ. Both machines are unbiased, the first with variance 1.0, the second with variance 1/100. A machine is selected at random; a measurement is taken and the machine used is recorded.

In the first case N is ancillary: in the second case the indicator variable for the machine used is an ancillary statistic. In either case, the ancillary statistic serves as an index or indicator for the experiment actually performed.

Whatever the distribution of N, few statisticians would quarrel with the statement that $\hat{\theta} = \bar{Y}_N$ is normally distributed with mean θ and variance $1/N$, even though N is a random variable. Implicitly, we invoke the conditionality principle, which states that values of the ancillary other than that observed are irrelevant for the purposes of inference. The unconditional variance of $\hat{\theta}$, namely $E(1/N) = 1/2 + 1/200 = 0.505$, is relevant only if N is not observed and if its distribution as stated above is correct.

In fact, ancillary statistics of this kind arise frequently in scientific work where observations are often censored or unavailable for reasons not connected with the phenomenon under investigation. Accidents will happen, so the saying goes.

The second type of ancillary arises in the context of a translation model or, more generally, a group transformation model. To take a very simple example, suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed uniformly on the interval $(\theta, \theta + 1)$. This is an instance of a non-regular problem in which the sample space is parameter dependent. The sufficient statistic is the pair of extreme values, $(Y_{(1)}, Y_{(n)})$, or equivalently, the minimum $Y_{(1)}$ and the range $R = Y_{(n)} - Y_{(1)}$. Clearly, by invariance, the distribution of R is unaffected by the value of θ and hence R is ancillary. Given that $R = r$, the conditional distribution of $Y_{(1)}$ is uniform on the interval $(\theta,\ \theta{+}1{-}r)$. Equivalently, the conditional distribution of $Y_{(n)} = Y_{(1)} + r$ is uniform over the interval$(\theta{+}r,\ \theta{+}1)$. Either conditional distribution leads to conditional equi-tailed confidence intervals

$$\{y_{(1)} - (1-r)(1-\alpha/2), \quad y_{(1)} - (1-r)\alpha/2\} \tag{8.1}$$

at level $1-\alpha$. Although these intervals also have the required coverage probability unconditionally as well as conditionally, their unconditional properties seem irrelevant if the aim is to summarise the information in the data concerning the value of θ.

On the other hand, if the ancillary is ignored, we may base our inferences on the pivot $Y_{(1)} - \theta$ whose density is $n(1-y)^{n-1}$ over the interval $(0,1)$. If $n = 2$,

$$(Y_{(1)} - 0.7764, \quad Y_{(1)} - 0.0253) \tag{8.2}$$

is an exact equi-tailed 90% confidence interval for θ. Even though it is technically correct and exact, the problem with this statement is that it is inadequate as a summary of the data actually observed. In fact, if $R = 0.77$, not a particularly extreme value, the conditional coverage of (8.2) is approximately 96% rather than the 90% quoted. At the other extreme, if $R \geq 0.975$, the conditional coverage of (8.2) is exactly zero so that the unconditional 90% coverage is quite misleading.

These examples demonstrate the poverty of any frequency theory of inference that ignores ancillary statistics. Having made that point, it must be stated that the difficulties involved in establishing a coherent frequency theory of conditional inference are formidable. Current procedures, even for an approximate conditional theory, are incomplete although much progress has been made in the past decade. Two difficulties are connected with the existence and uniqueness of ancillary statistics. In general, exactly ancillary statistics need not exist. Even where they do exist they need not be unique. If two ancillary statistics exist, they are not, in general, jointly ancillary and an additional criterion may be required to choose between them (Cox, 1971). One device that is quite sensible is to require ancillaries to be functions of the minimal sufficient statistic. Even this does not guarantee uniqueness. Exercises 8.2–8.4 discuss one example where the maximal ancillary is not unique but where the conclusions given such an ancillary are unaffected by the choice of ancillary.

Given the overwhelming magnitude of these difficulties, it appears impossible to devise an exact frequency theory of conditional inference to cover the most general cases. For these reasons, we concentrate here on approximate theories based on statistics that are ancillary in a suitably defined approximate sense. It is tempting here to make a virtue out of necessity, but the approximate theory does have the advantage of greater 'continuity' between those models for which exact analysis is possible and the majority of models for which no exact theory exists.

8.2 Joint cumulants

8.2.1 *Joint cumulants of A and U*

Suppose that A has a distribution not depending on the parameter θ. The moments and cumulants do not depend on θ and hence their derivatives with respect to θ must vanish. Thus, differentiation of

$$E(A(Y)) = \int A(y) f_Y(y;\theta)dy = \text{const}$$

with respect to θ gives

$$\int A(y) u_r(\theta; y) f_Y(y;\theta)dy = \text{cov}(A, U_r) = 0.$$

This final step assumes the usual regularity condition that it is legitimate to interchange the order of differentiation and integration. In particular, the sample space must be the same for all parameter values.

Further differentiation gives

$$\operatorname{cov}(A, U_{[rs]}) = 0; \quad \operatorname{cov}(A, U_{[rst]}) = 0, \ldots$$

and so on.

If A is ancillary, all functions of A are also ancillary. Hence it follows that the joint cumulants

$$\begin{aligned} \kappa_3(A, A, U_r) &= 0, \qquad \kappa_4(A, A, A, U_r) = 0, \ldots \\ \kappa_3(A, A, U_{[rs]}) &= 0, \quad \kappa_4(A, A, A, U_{[rs]}) = 0, \ldots \end{aligned}$$

also vanish. These results extend readily to higher-order joint cumulants provided that these cumulants involve one of U_r, $U_{[rs]}$, $U_{[rst]}, \ldots$ exactly once.

In general, although U_r is uncorrelated with A and with all functions of A, the conditional variance of U_r given A may be heavily dependent on the value of A. For instance, if A is a random sample size, the conditional covariance matrix of U_r given $A = a$ is directly proportional to a.

These results apply to *any* exact ancillary whatsoever: they also apply approximately to any approximate ancillary suitably defined.

8.2.2 *Conditional cumulants given A*

The conditional distribution of the data given $A = a$ is

$$f_{Y|A}(y|a;\theta) = f_Y(y;\theta)/f_A(a).$$

Hence, the conditional log likelihood differs from the unconditional log likelihood by a function not involving θ. Derivatives with respect to θ of the conditional log likelihood are therefore identical to derivatives of the unconditional log likelihood. With some exceptions as noted below, the conditional cumulants are different from the unconditional cumulants. In particular, it follows from differentiating the identity

$$\int f_{Y|A}(y|a;\theta) = 1$$

that the following identities hold

$$\begin{aligned} E(U_r|A) &= 0 \\ E(U_{[rs]}|A) &= 0 \\ E(U_{[rst]}|A) &= 0. \end{aligned}$$

Thus, the Bartlett identities (7.2) for the null cumulants of log likelihood derivatives are true conditionally on the value any ancillary statistic as well as unconditionally. Other cumulants or combinations of cumulants are affected by conditioning in a way that depends on the particular ancillary selected.

8.3 Local ancillarity

Exactly ancillary statistics arise usually in one of two ways, either as a random sample size or other index of the experiment actually performed, or as as a configuration statistic in location problems. These instances of exactly ancillary statistics are really rather special and in the majority of problems, no exactly ancillary statistic exists. In some cases it is known that no exact ancillary exists: in other cases no exact ancillary is known and it is suspected that none exists. For that reason, we concentrate our attention here on statistics that are approximately ancillary in some suitable sense.

Let θ_0 be an arbitrary but specified parameter value and let $A = A(Y;\theta_0)$ be a candidate ancillary statistic. If the distribution of A is approximately the same for all parameter values in some suitably chosen neighbourhood of θ_0, we say that A is approximately ancillary or locally ancillary in the vicinity of θ_0. More precisely, if the distribution of A under the density evaluated at $\theta_0 + n^{-1/2}\delta$ satisfies

$$f_A(a;\theta_0 + n^{-1/2}\delta) = f_A(a;\theta_0)\{1 + O(n^{-q/2})\}, \tag{8.3}$$

we say that A is qth-order locally ancillary in the vicinity of θ_0 (Cox, 1980). Two approximations are involved in this definition of approximate ancillarity. First, there is the $O(n^{-q/2})$ tolerance in (8.3), where n is the sample size or other measure of the extent of the observational record. Second, the definition is local, applying only to parameter values in an $O(n^{-1/2})$ neighbourhood of the target value, θ_0. The latter restriction is reasonable, at least in large samples, because if θ_0 is the true parameter point, the likelihood function eventually becomes negligible outside this $O(n^{-1/2})$ neighbourhood.

From (8.3) it can be seen that the log likelihood based on A satisfies

$$l_A(\theta_0 + n^{-1/2}\delta) = l_A(\theta_0) + O(n^{-q/2}). \tag{8.4}$$

Since $\partial l/\partial\theta = n^{1/2}\partial l/\partial\delta$, it follows that the Fisher information based on A for θ at θ_0 must be $O(n^{1-q/2})$ at most. This criterion is also used as a definition of approximate ancillarity.

If the distribution of A can be approximated by an Edgeworth series in which the cumulants are κ_1, κ_2, $\kappa_3, \ldots$, condition (8.3) is equivalent to

$$\nabla\kappa_r \equiv \kappa_r(\theta_0 + n^{-1/2}\delta) - \kappa_r(\theta_0) = O(n^{-q/2}), \quad r = 1, 2, \ldots \tag{8.5}$$

provided that A has been standardized so that $A - \kappa_1(\theta_0)$ is $O_p(1)$. Since κ_r is $O(n^{1-r/2})$, it is necessary only to check the first q cumulants of A in order to verify the property of local ancillarity. Usually (8.5) is much easier to verify than (8.3). Note, however, that A is usually a vector or array of random variables so that the rth cumulant array has a large number of components, each of which must satisfy (8.5).

We now proceed to show that approximately ancillary statistics exist and that they may be constructed from log likelihood derivatives. Since there is little merit in considering derivatives that are not tensors, we work with the tensorial derivatives V_r, $V_{rs}, \ldots$ at θ_0, as defined in (7.6). These derivatives are assumed to be based on n independent observations, to be asymptotically normal and to have all cumulants of order $O(n)$.

Under θ_0, the expectation of V_{rs} is $n\nu_{rs}$. By (7.4), the expectation of V_{rs} under the density at $\theta_0 + n^{-1/2}\delta$ is

$$n\{\nu_{rs} + \nu_{rs,i}\delta^i/n^{1/2} + \nu_{rs,[ij]}\delta^i\delta^j/(2n) + \cdots\}.$$

In terms of the standardized random variables $Z_{rs} = n^{-1/2}(V_{rs} - n\nu_{rs})$, which are $O_p(1)$ both under θ_0 and under the density at $\theta_0 + \delta/n^{1/2}$, we have

$$E(Z_{rs};\theta_0 + \delta/n^{1/2}) = \nu_{rs,i}\delta^i + n^{-1/2}\nu_{rs,[ij]}\delta^i\delta^j/2 + O(n^{-1}).$$

Thus, since $\nu_{rs,i} = 0$, it follows from (8.5) that Z_{rs} is first-order locally ancillary in the vicinity of θ_0. Hence V_{rs}, the tensorial array of second derivatives, is first-order locally ancillary.

A similar argument shows that V_{rst} is also first-order locally ancillary but not ancillary to second order.

To construct a statistic that is second-order locally ancillary, we begin with a first-order ancillary statistic, Z_{rs}, that is $O_p(1)$ and aim to make a suitable adjustment of order $O_p(n^{-1/2})$. Thus, we seek coefficients $\beta_{rs}^{i,j}$, $\beta_{rs}^{i,jk}$ such that the cumulants of

$$A_{rs} = Z_{rs} - n^{-1/2}\beta_{rs}^{i,j} Z_i Z_j - n^{-1/2}\beta_{rs}^{i,jk} Z_i Z_{jk} \tag{8.6}$$

satisfy the ancillarity conditions up to second order. Since ancillarity is preserved under transformation, it is unnecessary in (8.6) to include quadratic terms in Z_{rs}. The differences between the cumulants of Z_r, Z_{rs} under θ_0 and under $\theta_0 + n^{-1/2}\delta$ are given by

$$\begin{aligned}
\nabla E(Z_r) &= \nu_{r,i}\delta^i + n^{-1/2}\nu_{r,[ij]}\delta^i\delta^j/2 + O(n^{-1}) \\
\nabla E(Z_{rs}) &= n^{-1/2}\nu_{rs,[ij]}\delta^i\delta^j/2 + O(n^{-1}) \\
\nabla \operatorname{cov}(Z_r, Z_s) &= n^{-1/2}\nu_{r,s,i}\delta^i + O(n^{-1}) \\
\nabla \operatorname{cov}(Z_r, Z_{st}) &= n^{-1/2}\nu_{r,st,i}\delta^i + O(n^{-1}).
\end{aligned}$$

It follows that the mean of A_{rs} changes by an amount

$$\nabla E(A_{rs}) = n^{-1/2}\{\nu_{rs,[ij]} - 2\beta_{rs}^{k,l}\nu_{i,k}\nu_{j,l}\}\delta^i\delta^j + O(n^{-1}).$$

Thus, if we choose the coefficients

$$\beta_{rs}^{i,j} = \nu_{rs,[kl]}\nu^{i,k}\nu^{j,l}/2, \tag{8.7}$$

it follows that $\nabla E(A_{rs}) = O(n^{-1})$ as required. The coefficients $\beta_{rs}^{i,j}$ are uniquely determined by second-order ancillarity.

To find the remaining coefficients, it is necessary to compute $\nabla \operatorname{cov}(A_{rs}, A_{tu})$ and to ensure that this difference is $O(n^{-1})$. Calculations similar to those given above show that

$$\nabla \operatorname{cov}(A_{rs}, A_{tu}) = n^{-1/2}\{\nu_{rs,tu,i}\delta^i - \beta_{rs}^{i,jk}\nu_{jk,tu}\nu_{i,l}\delta^l[2]\} + O(n^{-1}).$$

The coefficients $\beta_{rs}^{i,jk}$ are not uniquely determined by the requirement of second-order ancillarity unless the initial first-order statistic is a scalar. Any set of coefficients that satisfies

$$\{\beta_{rs}^{i,jk}\nu_{jk,tu} + \beta_{tu}^{i,jk}\nu_{jk,rs}\}\nu_{i,v} = \nu_{rs,tu,v} \tag{8.8}$$

gives rise to a statistic that is locally ancillary to second order. If, to the order considered, all such ancillaries gave rise to the same sample space conditionally, non-uniqueness would not be a problem. In fact, however, two second-order ancillaries constructed in the above manner need not be jointly ancillary to the same order.

So far, we have not imposed the obvious requirement that the coefficients $\beta_{rs}^{i,jk}$ should satisfy the transformation laws of a tensor. Certainly, (8.8) permits non-tensorial solutions. With the restriction to tensorial solutions, it might appear that the only solution to (8.8) is

$$\beta_{rs}^{i,jk} = \nu_{rs,tu,v}\nu^{tu,jk}\nu^{i,v}/2,$$

where $\nu^{rs,tu}$ is a generalized inverse of $\nu_{rs,tu}$. Without doubt, this is the most obvious and most 'reasonable' solution, but it is readily demonstrated that it is not unique. For instance, if we define the tensor $\epsilon_{rs}^{i,jk}$ by

$$\epsilon_{rs}^{i,jk}\nu_{jk,tu} = \{\nu_{rs,t,u,v} - \nu_{r,s,tu,v}\}\nu^{i,v},$$

it is easily seen that $\beta_{rs}^{i,jk} + \epsilon_{rs}^{i,jk}$ is also a solution to (8.8). In fact, any scalar multiple of ϵ can be used here.

It is certainly possible to devise further conditions that would guarantee uniqueness or, alternatively, to devise criteria in order to select the most 'relevant' of the possible approximate ancillaries. In the remainder of this section, however, our choice is to tolerate the non-uniqueness and to explore its consequences for conditional inference.

The construction used here for improving the order of ancillarity is taken from Cox (1980), who considered the case of one-dimensional ancillary statistics, and from McCullagh (1984a), who dealt with the more general case. Skovgaard (1986c) shows that, under suitable regularity conditions, the order of ancillarity may be improved indefinitely by successive adjustments of decreasing orders. Whether it is desirable in practice to go much beyond second or third order is quite another matter.

8.4 Stable combinations of cumulants

In Section 8.2.2, it was shown that, conditionally on any ancillary however selected, certain combinations of cumulants of log likelihood derivatives are identically zero. Such combinations whose value is unaffected by conditioning may be said to be *stable*. Thus, for instance, ν_i and $\nu_{ij} + \nu_{i,j}$ are identically zero whether we condition on A or not. It is important to recognize stable combinations in order to determine the effect, if any, of the choice of ancillary on the conclusions reached.

In this section, we demonstrate that, for the type of ancillary considered in the previous section, certain cumulants and cumulant combinations, while not exactly stable, are at least stable to first order in n. By way of example, it will be shown that $\nu_{i,j,k,l}$ and $\nu_{ij,kl}$ are both unstable but that the combination

$$\nu_{i,j,k,l} - \nu_{ij,kl}[3] \tag{8.9}$$

is stable to first order.

Suppose then that A, with components A_r, is ancillary, either exactly, or approximately to some suitable order. It is assumed that the joint distribution of (A, V_i, V_{ij}) may be approximated by an Edgeworth series and that all joint cumulants are $O(n)$. The dimension of A need not equal p but must be fixed as $n \to \infty$. The mean and variance of A are taken to be 0 and $n\lambda_{r,s}$ respectively. Thus, $A_r = O_p(n^{1/2})$ and $A^r = n^{-1}\lambda^{r,s}A_s$ is $O_p(n^{-1/2})$.

From the results given in Section 5.6, the conditional covariance of V_i and V_j is

$$\begin{aligned}\kappa_2(V_i, V_j|A) &= n\{\nu_{i,j} + \nu_{r;i,j}A^r + O(n^{-1})\}\\ &= n\nu_{i,j} + O(n^{1/2}),\end{aligned}$$

where $\nu_{r;i,j}$ is the third-order joint cumulant of A_r, V_i, V_j. Note that the ancillarity property

$$\mathrm{cov}(A_r, V_i) = \nu_{r;i} = 0$$

greatly simplifies these calculations. Similarly, in the case of the third cumulant, we have

$$\begin{aligned}\kappa_3(V_i, V_j, V_k|A) &= n\{\nu_{i,j,k} + \nu_{r;i,j,k}A^r + O(n^{-1})\}\\ &= n\nu_{i,j,k} + O(n^{1/2}),\end{aligned}$$

Thus, to first order at least, $\nu_{i,j}$ and $\nu_{i,j,k}$ are unaffected by conditioning. We say that these cumulants are *stable to first order*.

On the other hand, from the final equation in Section 5.6.2, we find that the conditional fourth cumulant of V_i, V_j, V_k, V_l given A is

$$\kappa_4(V_i, V_j, V_k, V_l|A) = n\{\nu_{i,j,k,l} - \nu_{r;i,j}\nu_{s;k,l}\lambda^{r,s}[3] + O(n^{-1/2})\}.$$

In this case, unlike the previous two calculations, conditioning has a substantial effect on the leading term. Thus, $\nu_{i,j,k,l}$ is unstable to first order: its interpretation is heavily dependent on the conditioning event.

Continuing in this way, it may be seen that the conditional covariance matrix of V_{ij} and V_{kl} is

$$\kappa_2(V_{ij}, V_{kl}|A) = n\{\nu_{ij,kl} - \nu_{r;ij}\nu_{s;kl}\lambda^{r,s}[3] + O(n^{-1/2})\}.$$

Again, this is an unstable cumulant. However, from the identity $\nu_{r;ij} = -\nu_{r;i,j}$ it follows that the combination

$$\kappa_4(V_i, V_j, V_k, V_l|A) - \kappa_2(V_{ij}, V_{kl}|A)[3] = n\{\nu_{i,j,k,l} - \nu_{ij,kl}[3] + O(n^{-1/2})\}$$

is stable to first order. Similarly, the conditional third cumulant

$$\kappa_3(V_i, V_j, V_{kl}|A) = n\{\nu_{i,j,kl} - \nu_{r;i,j}\nu_{s;kl}\lambda^{r,s} + O(n^{-1/2})\},$$

is unstable, whereas the combination

$$\nu_{i,j,kl} + \nu_{ij,kl} \tag{8.10}$$

is stable to first order.

These calculations are entirely independent of the choice of ancillary. They do not apply to the random sample size example discussed in Section 8.1 unless $\mu = E(N) \to \infty$, $\text{var}(N) = O(\mu)$ and certain other conditions are satisfied. However, the calculations do apply to the approximate ancillary constructed in the previous section. Note that, conditionally on the ancillary (8.6), the conditional covariance of V_{ij} and V_{kl} is reduced from $O(n)$ to $O(1)$, whereas the third-order joint cumulant of V_i, V_j and V_{kl} remains $O(n)$. This is a consequence of the stability of the combination (8.10).

In this context, it is interesting to note that the Bartlett adjustment factor (7.13) is a combination of the four stable combinations derived here. It follows that, up to and including terms of order $O(n^{-1})$, the likelihood ratio statistic is independent of all ancillary and approximately ancillary statistics.

8.5 Orthogonal statistic

Numerical computation of conditional distributions and conditional tail areas is often a complicated unappealing task. In many simple problems, particularly those involving the normal-theory linear model, the problem can be simplified to a great extent by 'regressing out' the conditioning statistic and forming a 'pivot' that is independent of the conditioning statistic. In normal-theory and other regression problems, the reasons for conditioning are usually connected with the elimination of nuisance parameters, but the same device of constructing a pivotal statistic can also be used to help cope with ancillary statistics. The idea is to start with an arbitrary statistic, V_i say, and by making a suitable minor adjustment, ensure that the adjusted statistic, S_i, is independent of the ancillary to the order required. Inference can then be based on the marginal distribution of S_i: this procedure is sometimes labelled 'conditional inference without tears' or 'conditional inference without conditioning'.

In those special cases where there is a complete sufficient statistic, S, for the parameters, Basu's theorem (Basu, 1955, 1958) tells us that all ancillaries are independent of S. This happy state of affairs means that inferences based on the marginal distribution of S are automatically conditional and are safe from conditionality criticisms of the type levelled in Section 8.1. By the same token, the result given at the end of the previous section shows that the maximized likelihood ratio statistic is similarly independent of all ancillaries and of approximate ancillaries to third order in n. Inferences based on the marginal distribution of the likelihood ratio statistic are, in large measure, protected against criticism on grounds of conditionality.

The score statistic, V_i, is asymptotically independent of all ancillaries, but only to first order in n. In other words, $\mathrm{cov}(V_i, A) = 0$ for all ancillaries. First-order independence is a very weak requirement and is occasionally unsatisfactory if n is not very large or if the ancillary takes on an unusual or extreme value. For this reason we seek an adjustment to V_i to make the adjusted statistic independent of all ancillaries to a higher order in n. For the resulting statistic to be useful, it is helpful to insist that it have a simple null distribution, usually normal, again to the same high order of approximation.

Thus, we seek coefficients $\gamma_r^{i,j}$, $\gamma_r^{i,jk}$ such that the adjusted statistic

$$S_r = Z_r + n^{-1/2}\{\gamma_r^{i,j}(Z_i Z_j - \nu_{i,j}) + \gamma_r^{i,jk} Z_i Z_{jk}\}$$

is independent of A to second order and also normally distributed to second order. Both of these calculations are made under the null density at θ_0.

For the class of ancillaries (8.6) based on the second derivatives of the log likelihood, we find

$$\begin{aligned} \mathrm{cov}(S_r, A_{st}) &= O(n^{-1}) \\ \kappa_3(S_r, S_s, A_{tu}) &= n^{-1/2}\{\nu_{r,s,tu} - 2\beta_{tu}^{i,j}\nu_{i,r}\nu_{j,s} \\ &\qquad + \gamma_r^{i,jk}\nu_{jk,tu}\nu_{i,s}[2]\} + O(n^{-1}) \end{aligned}$$

On using (8.7), we find

$$(\gamma_r^{i,jk}\nu_{i,s} + \gamma_s^{i,jk}\nu_{i,r})\nu_{jk,tu} = \nu_{rs,tu} \tag{8.11}$$

as a condition for orthogonality to second order. One solution, but by no means the only one, unless $p = 1$, is given by

$$\gamma_r^{i,jk}\nu_{i,s} = \delta_{rs}^{jk}/2. \tag{8.12}$$

The remaining third-order joint cumulant

$$\kappa_3(S_r, A_{st}, A_{uv}) = n^{-1/2}\{\nu_{r,st,uv} - \beta_{st}^{i,jk}\nu_{i,r}\nu_{jk,uv}[2]\} + O(n^{-1})$$

is guaranteed to be $O(n^{-1})$ on account of (8.8). The choice of ancillary among the coefficients satisfying (8.8) is immaterial.

Finally, in order to achieve approximate normality to second order, we require that the third-order cumulant of S_r, S_s and S_t be $O(n^{-1})$. This condition gives

$$\nu_{r,s,t} + \gamma_r^{i,j}\nu_{i,s}\nu_{j,t}[6] = 0.$$

Again, the solution is not unique unless $p = 1$, but it is natural to consider the 'symmetric' solution

$$\gamma_r^{i,j} = -\nu_{r,s,t}\nu^{i,s}\nu^{j,t}/6. \tag{8.13}$$

For the particular choice of coefficients (8.12) and (8.13), comparison with (7.15) shows that

$$S_r = W_r + \nu_{r,s,t}\nu^{s,t}/(6n^{1/2}) + O_p(n^{-1}),$$

where W_r are the components in the tensor decomposition of the likelihood ratio statistic given in Section 7.4.5.

In the case of scalar parameters, and using the coefficients (8.12) and (8.13), $W_r/i_{20}^{1/2}$ is equal to the signed square root of the likelihood ratio statistic. If we denote by ρ_3 the third standardized cumulant of $\partial l/\partial\theta$, then the orthogonal statistic may be written in the form

$$S = \pm\{2l(\hat\theta) - 2l(\theta_0)\}^{1/2} + \rho_3/6. \tag{8.14}$$

This statistic is distributed as $N(0,1) + O(n^{-1})$ independently of all ancillaries. The sign is chosen according to the sign of $\partial l/\partial\theta$.

8.6 Conditional distribution given A

In the previous sections it was shown that ancillary statistics, whether approximate or exact, are, in general, not unique, but yet certain useful formulae can be derived that are valid conditionally on any ancillary, however chosen. Because of the non-uniqueness of ancillaries, the most useful results must apply to as wide a class of *relevant* ancillaries as possible. This section is devoted to finding convenient expressions for the distributions of certain statistics such as the score statistic U_r or the maximum likelihood estimate $\hat{\theta}^r$, given the value of A. At no stage in the development is the ancillary specified. The only condition required is one of relevance, namely that the statistic of interest together with A should be sufficient to high enough order. Minimal sufficiency is not required.

It turns out that there is a particularly simple expression for the conditional distribution of $\hat{\theta}$ given A and that this expression is either exact or, if not exact, accurate to a very high order of asymptotic approximation. This conditional distribution, which may be written in the form

$$p(\hat{\theta}; \theta | A) = (2\pi c)^{-p/2} |\hat{\jmath}_{rs}|^{1/2} \exp\{l(\theta) - l(\hat{\theta})\}\{1 + O(n^{-3/2})\}$$

is known as Barndorff-Nielsen's formula (Barndorff-Nielsen, 1980, 1983). One peculiar aspect of the formula is that the ancillary is not specified and the formula appears to be correct for a wide range of ancillary statistics. For this reason the description 'magical mystery formula' is sometimes used. In this expression $\log c = b$, the Bartlett adjustment factor, $\hat{\jmath}$ is the *observed* information matrix regarded as a function of $\hat{\theta}$ and A, and the formula is correct for any relevant ancillary.

It is more convenient at the outset to work with the score statistic with components $U_r = \partial l / \partial \theta$ evaluated at $\theta = 0$, a value chosen here for later convenience of notation. All conditional cumulants of U are assumed to be $O(n)$ as usual. In what follows, it will be assumed that A is locally ancillary to third order and that the pair (U, A) is jointly sufficient to the same order. In other words, for all θ in some neighbourhood of the origin, the conditional log likelihood based on U satisfies

$$l_{U|A}(\theta) - l_{U|A}(0) = l_Y(\theta) - l_Y(0) + O(n^{-3/2}). \tag{8.15}$$

Ancillary statistics satisfying these conditions, at least for θ in an $O(n^{-1/2})$ neighbourhood of the origin, can be constructed along the lines described in Section 8.3, but starting with the second and third derivatives jointly. In the case of location models, or more generally for group transformation models, exactly ancillary statistics exist that satisfy the above property for all θ. Such ancillaries typically have dimension of order $O(n)$. Since no approximation will be used here for the marginal distribution of A, it is not necessary to impose restrictions on its dimension. Such restrictions would be necessary if Edgeworth or saddlepoint approximations were used for the distribution of A.

The first step in the derivation is to find an approximation to the log likelihood function in terms of the conditional cumulants of U given $A = a$. Accordingly, let $K(\xi)$ be the conditional cumulant generating function of U given A at $\theta = 0$. Thus, the conditional cumulants $\kappa_{r,s}$, $\kappa_{r,s,t}, \ldots$ are functions of a. By the saddlepoint approximation, the conditional log density of U given A at $\theta = 0$ is

$$\begin{aligned} -K^*(u) + \log|K^{*rs}(u)|/2 - p\log(2\pi)/2 - (3\rho_4^* - 4\rho_{23}^{*2})/4! \\ + O(n^{-3/2}), \end{aligned} \tag{8.16}$$

where $K^*(u)$ is the Legendre transformation of $K(\xi)$. This is the approximate log likelihood at $\theta = 0$. To obtain the value of the log likelihood at an arbitrary point, θ, we require the conditional cumulant generating function $K(\xi; \theta)$ of U given A at θ.

For small values of θ, the conditional cumulants of U have their usual expansions about $\theta = 0$ as follows.

$$\begin{aligned} E(U_r; \theta) &= \kappa_{r,s}\theta^s + \kappa_{r,[st]}\theta^s\theta^t/2! + \kappa_{r,[stu]}\theta^s\theta^t\theta^u/3! + \cdots \\ \operatorname{cov}(U_r, U_s; \theta) &= \kappa_{r,s} + \kappa_{r,s,t}\theta^t + \kappa_{r,s,[tu]}\theta^t\theta^u/2! + \cdots \\ \kappa_3(U_r, U_s, U_t; \theta) &= \kappa_{r,s,t} + \kappa_{r,s,t,u}\theta^u + \cdots . \end{aligned}$$

These expansions can be simplified by suitable choice of coordinate system. For any given value of A, we may choose a coordinate system in the neighbourhood of the origin satisfying $\kappa_{r,st} = 0$, $\kappa_{r,stu} = 0$, so that all higher-order derivatives are conditionally uncorrelated with U_r. This property is achieved using the transformation (7.5). Denoting the cumulants in this coordinate system by ν with appropriate indices, we find after collecting certain terms that

$$\begin{aligned} E(U_r;\theta) &= K_r(\theta) + \nu_{r,s,tu}\theta^s\theta^t\theta^u[3]/3! + \cdots \\ \operatorname{cov}(U_r, U_s;\theta) &= K_{rs}(\theta) + \nu_{r,s,tu}\theta^t\theta^u/2! + \cdots \\ \kappa_3(U_r, U_s, U_t;\theta) &= K_{rst}(\theta) + \cdots \, . \end{aligned} \tag{8.17}$$

Thus, the conditional cumulant generating function of U under θ is

$$\begin{aligned} K(\xi;\theta) = K(\theta+\xi) - K(\theta) &+ \nu_{r,s,tu}\xi^r\theta^s\theta^t\theta^u[3]/3! \\ &+ \nu_{r,s,tu}\xi^r\xi^s\theta^t\theta^u/4 + \cdots \, . \end{aligned}$$

The final two terms above measure a kind of departure from simple exponential family form even after conditioning. To this order of approximation, an arbitrary model cannot be reduced to a full exponential family model by conditioning.

To find the Legendre transformation of $K(\xi;\theta)$, we note first that

$$K^*(u) - \theta^r u_r + K(\theta)$$

is the Legendre transformation of $K(\theta+\xi) - K(\theta)$. To this must be added a correction term of order $O(n^{-1})$ involving $\nu_{r,s,tu}$. Note that $\nu_{rs,tu}$, the so-called 'curvature' tensor or covariance matrix of the residual second derivatives, does not enter into these calculations. A straightforward calculation using Taylor expansion shows that the Legendre transformation of $K(\xi;\theta)$ is

$$\begin{aligned} K^*(u;\theta) = K^*(u) - \theta^r u_r + K(\theta) &- \nu_{r,s,tu}u^r u^s\theta^t\theta^u/4 \\ &+ \nu_{r,s,tu}\theta^r\theta^s\theta^t\theta^u/4 + O(n^{-3/2}). \end{aligned}$$

It may be checked that $K^*(u;\theta)$, evaluated at the mean of U under θ given by (8.17), is zero as it ought to be.

We are now in a position to use the saddlepoint approximation for a second time, but first we require the log determinant of the array of second derivatives. In subsequent calculations, terms that have an effect of order $O(n^{-3/2})$ on probability calculations are ignored without comment. Hence, the required log determinant is

$$\log\det K^{*rs}(u;\theta) = \log\det K^{*rs}(u;0) - \nu_{r,s,tu}\nu^{r,s}\theta^t\theta^u/2.$$

Exercise 1.16 is useful for calculating log determinants.

The log likelihood function or the log density function may be written in terms of the Legendre transformation as

$$\begin{aligned} l(\theta) - l(0) &= -K^*(u;\theta) + K^*(u) \\ &\quad + \tfrac{1}{2}\log\det K^{*rs}(u;\theta) - \tfrac{1}{2}\log\det K^{*rs}(u) \\ &= \theta^r u_r - K(\theta) + \nu_{r,s,tu}u^r u^s\theta^t\theta^u/4 \\ &\quad - \nu_{r,s,tu}\theta^r\theta^s\theta^t\theta^u/4 - \nu_{r,s,tu}\nu^{r,s}\theta^t\theta^u/4 \end{aligned} \tag{8.18}$$

from which it can be seen that the conditional third cumulant $\nu_{r,s,tu}$ governs the departure from simple exponential form. This completes the first step in our derivation.

It is of interest here to note that the change in the derivative of $l(\theta)$, at least in its stochastic aspects, is governed primarily by the third term on the right of (8.18). It may be possible to interpret $\nu_{r,s,tu}$ as a curvature or torsion tensor, though its effect is qualitatively quite different from Efron's (1975) curvature, which is concerned mainly with the variance of the residual second derivatives. The latter notion of curvature is sensitive to conditioning.

In many instances, the likelihood function and the maximized likelihood ratio statistic can readily be computed either analytically or numerically whereas the conditional cumulant generating function, $K(\xi)$ and the conditional Legendre transformation, $K^*(u)$ cannot, for the simple reason that A is not specified explicitly. Thus, the second step in our derivation is to express $K^*(u)$ and related quantities directly in terms of the likelihood function and its derivatives.

The likelihood function given above has its maximum at the point

$$\hat{\theta}^r = K^{*r}(u) + \epsilon^r(u) + O(n^{-2}) \tag{8.19}$$

where ϵ^r is $O(n^{-3/2})$ given by

$$\epsilon^r = -\nu_{st,u,v}\nu^{r,s}\nu^{u,v}u^t/2 - \nu_{s,t,uv}\nu^{r,s}u^t u^u u^v/2.$$

On substituting $\hat{\theta}$ into (8.18), further simplification shows that the log likelihood ratio statistic is

$$l(\hat{\theta}) - l(0) = K^*(u) - \nu_{r,s,tu}\nu^{r,s}u^t u^u/4 + O(n^{-3/2}),$$

which differs from the conditional Legendre transformation of $K(\xi)$ by a term of order $O(n^{-1})$. Each of these calculations involves a small amount of elementary but tedious algebra that is hardly worth reproducing.

We now observe that

$$-\{l(\hat{\theta}) - l(0)\} + \tfrac{1}{2}\log\det K^{*rs}(u;\hat{\theta}) = -K^*(u) + \tfrac{1}{2}\log\det K^{*rs}(u), \tag{8.20}$$

which is the dominant term in the saddlepoint approximation for the conditional log density of U given A. It now remains to express the left member of the above equation solely in terms of the log likelihood function and its derivatives.

On differentiating (8.18), we find that

$$\begin{aligned} u_{rs}(\theta) = -K_{rs}(\theta) - \nu_{rs,t,u}\nu^{t,u}/2 - \nu_{r,s,tu}\theta^t\theta^u/2 - \nu_{r,t,su}[2]\theta^t\theta^u\,. \\ + \nu_{rs,t,u}u^t u^u/2 - \nu_{rs,t,u}\theta^t\theta^u/2. \end{aligned}$$

Hence, the observed Fisher information at $\hat{\theta}$ is

$$\hat{j}_{rs} = -u_{rs}(\hat{\theta}) = K_{rs}(\hat{\theta}) + \nu_{r,s,tu}\hat{\theta}^t\hat{\theta}^u/2 + \nu_{rs,t,u}\nu^{t,u}/2 + \nu_{r,t,su}[2]\hat{\theta}^t\hat{\theta}^u$$

and the observed information determinant is given by

$$\log\det\hat{j}_{rs} = -\log\det K^{*rs}(u;\hat{\theta}) + 2\nu_{r,t,su}\nu^{r,s}\hat{\theta}^t\hat{\theta}^u + \nu_{r,s,tu}\nu^{r,s}\nu^{t,u}/2.$$

On substituting into (8.20), it is seen that the saddlepoint approximation with one correction term for the conditional log density of U given A is

$$-\{l(\hat{\theta}) - l(\theta)\} - \tfrac{1}{2}\log\det\hat{j}_{rs} - \tfrac{1}{2}p\log(2\pi) + \nu_{r,t,su}\nu^{r,s}\hat{\theta}^t\hat{\theta}^u - \tfrac{1}{2}pb(\theta), \tag{8.21}$$

where $b(\theta)$ is the Bartlett adjustment, given in this instance by

$$pb(\theta) = (3\rho_{13}^2 + 2\rho_{23}^2 - 3\rho_4)/12 - \nu_{r,s,tu}\nu^{r,s}\nu^{t,u}/2.$$

As pointed out in Section 8.4, $b(\theta)$ can be computed from the unconditional cumulants using the formula (7.13). Similarly, $\nu_{r,t,su}$ in (8.21) can be computed from the unconditional cumulants using (8.10).

Expression (8.21) can be simplified even further. The derivative of the transformation (8.19) from u to $\hat\theta$ is

$$\begin{aligned}\hat\theta^{rs} &= K^{*rs}(u) - \nu^{r,t}\nu^{s,u}\nu^{v,w}\nu_{tu,v,w}/2 - \nu^{r,t}\nu^{s,u}u^v u^w \nu_{t,u,vw}/2 \\ &\qquad - \nu^{r,t}\nu^{v,s}u^u u^w \nu_{tu,v,w}[2]/2 \\ &= K^{*rs}(u;\hat\theta) - \nu^{r,t}\nu^{s,u}\nu^{v,w}\nu_{tu,v,w}/2 - \nu^{r,t}\nu^{v,s}u^u u^w \nu_{tu,v,w}.\end{aligned}$$

Hence, the log determinant of the transformation is

$$\begin{aligned}\log\det\hat\theta^{rs} &= \log\det K^{*rs}(u;\hat\theta) - \nu^{r,s}\nu^{t,u}\nu_{rs,t,u}/2 - \nu^{r,s}u^t u^u \nu_{rt,s,u} \\ &= -\log\det\hat{\jmath}_{rs} + \nu_{r,t,su}\nu^{r,s}u^t u^u.\end{aligned} \tag{8.22}$$

Hence, under the assumption that $\theta = 0$ is the true parameter point, the conditional log density of $\hat\theta$ given A is

$$-\{l(\hat\theta) - l(\theta)\} + \tfrac{1}{2}\log\det\hat{\jmath}_{rs} - p\log(2\pi) - pb(\hat\theta)/2.$$

More generally, if the true parameter point is θ, the conditional density of $\hat\theta$ as a function of θ and the conditioning variable may be written

$$p(\hat\theta;\theta|A) = (2\pi\hat c)^{-p/2}\exp\{l(\theta) - l(\hat\theta)\}|\hat{\jmath}|^{1/2} \tag{8.23}$$

where $\hat c = \log b(\hat\theta)$. The log likelihood function $l(\theta)$, in its dependence on the data, is to be regarded as a function of $(\hat\theta, a)$. Similarly for the observed information determinant. Thus, for example, $l(\theta) = l(\theta;\hat\theta,a)$ is the log likelihood function, $u_{rs}(\theta) = u_{rs}(\theta;\hat\theta,a)$ and $\hat{\jmath}_{rs} = u_{rs}(\hat\theta;\hat\theta,a)$ is the observed information matrix.

Approximation (8.23) gives the conditional density of the maximum likelihood estimate of the canonical parameter corresponding to the transformation (7.5). On transforming to any other parameter, the form of the approximation remains the same. In fact, the approximation is an invariant of *weight* 1 under re-parameterization in the sense of Section 6.1. Thus, (8.23) is equally accurate or inaccurate in all parameterizations and the functional form of the approximation is the same whatever parameterization is chosen.

8.7 Bibliographic notes

It would be impossible in the limited space available here to discuss in detail the various articles that have been written on the subject of ancillary statistics and conditional inference. What follows is a minimal set of standard references.

Fisher (1925, 1934) seems to have been first to recognize the need for conditioning to ensure that probability calculations are relevant to the data observed. His criticism of Welch's test (Fisher, 1956) was based on its unsatisfactory conditional properties. Other important papers that discuss the need for conditioning and the difficulties that ensue are Cox (1958, 1971), Basu (1964), Pierce (1973), Robinson (1975, 1978), Kiefer (1977), Lehmann (1981) and Buehler (1982). The book by Fraser (1968) is exceptional for the emphasis placed on group structure as an integral part of the model specification.

The question of the existence or otherwise of ancillary statistics (or similar regions) was first posed by R.A. Fisher as the 'problem of the Nile' in his 1936 Harvard Tercentenary lecture. It is now known that, in the continuous case, exactly ancillary statistics always exist if the parameter space is finite or if attention is restricted to any finite set of parameter values, however numerous.

This conclusion follows from Liapounoff's theorem (Halmos, 1948), which states that the range of a vector measure is closed and, in the non-atomic case, convex. On the other hand, it is also known that no such regions, satisfying reasonable continuity conditions, exist for the Behrens-Fisher problem (Linnik, 1968). It seems, then, that as the number of parameter points under consideration increases, regions whose probability content is exactly the same for all parameter values are liable to become increasingly 'irregular' in some sense. However, acceptable regions whose probability content is approximately the same for all parameter values do appear to exist in most instances.

Kalbfleisch (1975) makes a distinction, similar to that made in Section 8.1, between experimental ancillaries and mathematical ancillaries. Lloyd (1985ab), on the other hand, distinguishes between internal and external ancillaries. The former are functions of the minimal sufficient statistic.

Much of the recent work has concentrated on the notion of approximate ancillarity or asymptotic ancillarity. See, for example, Efron & Hinkley (1978), Cox (1980), Hinkley (1980), Barndorff-Nielsen (1980) for further details. The results given in Section 8.5 are taken from McCullagh (1984a).

Formula (8.23) was first given by Barndorff-Nielsen (1980) synthesizing known exact conditional results for translation models due to Fisher (1934), and approximate results for full exponential family models based on the saddlepoint approximation. In subsequent papers, (Barndorff-Nielsen, 1983, 1984, 1985), the formula has been developed and used to obtain conditional confidence intervals for one-dimensional parameters. More recent applications of the formula have been to problems involving nuisance parameters. The formulae in Section 6.4 are special cases of what is called the 'modified profile likelihood'.

The derivation of Barndorff-Nielsen's formula in Section 8.6 appears to be new.

8.8 Further results and exercises 8

8.1 Show that if (X_1, X_2) has the bivariate normal distribution with zero mean, variances σ_1^2, σ_2^2 and covariance $\rho\sigma_1\sigma_2$, then the ratio $U = X_1/X_2$ has the Cauchy distribution with median $\theta = \rho\sigma_1/\sigma_2$ and dispersion parameter $\tau^2 = \sigma_1^2(1-\rho^2)/\sigma_2^2$. Explicitly,

$$f_U(u;\theta,\tau) = \tau^{-1}\pi^{-1}\{1 + (u-\theta)^2/\tau^2\}^{-1},$$

where $-\infty < \theta < \infty$ and $\tau > 0$. Deduce that $1/U$ also has the Cauchy distribution with median $\theta/(\tau^2+\theta^2)$ and dispersion parameter $\tau^2/(\tau^2+\theta^2)^2$. Interpret the conclusion that θ/τ is invariant.

8.2 Let $X_1, \ldots, X_n$ be independent and identically distributed Cauchy random variables with unknown parameters (θ, τ). Let $\bar{X}$ and s_X^2 be the sample mean and sample variance respectively. By writing $X_i = \theta + \tau\epsilon_i$, show that the joint distribution of the *configuration statistic* A with components $A_i = (X_i - \bar{X})/s_X$ is independent of the parameters and hence that A is ancillary. [This result applies equally to any location-scale family where the ϵ_i are *i.i.d.* with known distribution.]

8.3 Using the notation of the previous exercise, show that for any constants a, b, c, d satisfying $ad - bc \neq 0$,

$$Y_i = (a + bX_i)/(c + dX_i) \qquad i = 1, \ldots, n$$

are independent and identically distributed Cauchy random variables. Deduce that the derived statistic A^* with components $A_i^* = (Y_i - \bar{Y})/s_Y$ has a distribution not depending on (θ, τ). Hence conclude that the maximal ancillary for the problem described in Exercise 8.2 is not unique. Demonstrate explicitly that two such ancillaries are not jointly ancillary. [This construction is specific to the two-parameter Cauchy problem.]

8.4 Suppose, in the notation previously established, that $n = 3$. Write the ancillary in the form $\{\text{sign}(X_3 - X_2),\ \text{sign}(X_2 - X_1)\}$, together with an additional component

$$A_X = (X_{(3)} - X_{(2)})/(X_{(2)} - X_{(1)}),$$

where $X_{(j)}$ are the ordered values of X. Show that A_X is a function of the sufficient statistic, whereas the first two components are not. Let $Y_i(t) = 1/(X_i - t)$ and denote by $A(t)$ the corresponding ancillary computed as a function of the transformed values. Show that the function $A(t)$ is continuous except at the three points $t = X_i$ and hence deduce that the data values may be recovered from the set of ancillaries $\{A(t), -\infty < t < \infty\}$. [In fact, it is enough to know the values of $A(t)$ at three distinct points interlacing the observed values. However, these cannot be specified in advance.]

8.5 Suppose that (X_1, X_2) are bivariate normal variables with zero mean, unit variance and unknown correlation ρ. Show that $A_1 = X_1$ and $A_2 = X_2$ are each ancillary, though not jointly ancillary, and that neither is a component of the sufficient statistic. Let $T = X_1 X_2$. Show that

$$\begin{aligned} T|A_1 = a_1 &\sim N\{\rho a_1^2, (1-\rho^2)a_1^2\} \\ T|A_2 = a_2 &\sim N\{\rho a_2^2, (1-\rho^2)a_2^2\}. \end{aligned}$$

8.6 In the notation of the previous exercise, suppose that it is required to test the hypothesis $H_0 : \rho = 0$, and that the observed values are $x_1 = 2$, $x_2 = 1$. Compute the conditional tail areas $\text{pr}(T \geq t | A_1 = a_1)$ and $\text{pr}(T \geq t | A_2 = a_2)$. Comment on the appropriateness of these tail areas as measures of evidence against H_0.

8.7 Suppose that (X_1, X_2, X_3) have the trivariate normal distribution with zero mean and intraclass covariance matrix with variances σ^2 and correlations ρ. Show that $-\frac{1}{2} \leq \rho \leq 1$. Prove that the moments of $X_1 + \omega X_2 + \omega^2 X_3$ and $X_1 + \omega X_3 + \omega^2 X_2$ are independent of both parameters, but that neither statistic is ancillary $[\omega = \exp(2\pi i/3)]$.

8.8 Suppose in the previous exercise that $\sigma^2 = 1$. Show that this information has no effect on the sufficient statistic but gives rise to ancillaries, namely X_1, X_2, X_3, no two of which are jointly ancillary.

8.9 Show that the tensorial decomposition of the likelihood ratio statistic in (7.15) is not unique but that all such decompositions are orthogonal statistics in the sense used in Section 8.5 above.

8.10 Show that the Legendre transformation of

$$\begin{aligned} K(\xi; \theta) = K(\theta + \xi) - K(\theta) &+ \nu_{r,s,tu}\xi^r\theta^s\theta^t\theta^u[3]/3! \\ &+ \nu_{r,s,tu}\xi^r\xi^s\theta^t\theta^u/4 + \cdots \end{aligned}$$

with respect to the first argument is approximately

$$\begin{aligned} K^*(u; \theta) = K^*(u) - \theta^r u_r + K(\theta) &- \nu_{r,s,tu}u^r u^s\theta^t\theta^u/4 \\ &+ \nu_{r,s,tu}\theta^r\theta^s\theta^t\theta^u/4. \end{aligned}$$

8.11 Show that the Legendre transformation, $K^*(u; \theta)$, evaluated at

$$u_r = K_r(\theta) + \nu_{r,s,tu}\theta^s\theta^t\theta^u[3]/3!,$$

is zero to the same order of approximation.

8.12 Show that the maximum of the log likelihood function (8.18) is given by (8.19).

8.13 Beginning with the canonical coordinate system introduced at (8.17), transform from θ to ϕ with components

$$\phi_r = K_r(\theta) + \nu_{r,s,tu}\theta^s\theta^t\theta^u/2 + \nu_{r,s,tu}\theta^s\nu^{t,u}/2.$$

Show that, although $E(U_r;\theta) \neq \phi_r$, nevertheless $\hat{\phi}_r = U_r$. Show also that the observed information determinant with respect to the components of ϕ satisfies

$$\tfrac{1}{2}\log\det K^{*rs}(u;\hat{\theta}) = \tfrac{1}{2}\log\det \hat{j}_{\phi}^{rs} + \nu_{r,s,tu}\nu^{r,s}\nu^{t,u}$$

at the maximum likelihood estimate. Hence deduce (8.23) directly from (8.20).

8.14 Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed on the interval $\theta,\ \theta{+}1$. Show that the likelihood function is constant in the interval $(y_{(n)} - 1,\ y_{(1)})$ and is zero otherwise. Hence, interpret $r = y_{(n)} - y_{(1)}$ as an indicator of the shape of the likelihood function.