By the Same Author

Foundations of Mathematical Genetics, 1977

Pascal's Arithmetical Triangle, 1987

LIKELIHOOD

Expanded Edition

A. W. F. EDWARDS, Sc.D.

Fellow of Gonville and Caius College
Reader in Biometry, University of Cambridge

THE JOHNS HOPKINS
UNIVERSITY PRESS

Baltimore and London

CONTENTS

than N. Should this not be taken into account? The argument for not doing so is that in any simple model for the distribution of N in families, p will not appear, so that N by itself is wholly uninformative about p. Thus we may, without prejudice, condition the support on the actual value observed.

Again, confronted with an infinite sequence of binomial trials, we might decide to evaluate p by counting trials until R successes have been observed. Let n be the number counted; then the support for p is still

$$S(p) = R \ln p + (n - R) \ln (1 - p),$$

since n now has a negative binomial distribution.[21] This time R by itself is uninformative about p, and the support may therefore be conditioned on its value.

The general principle we should follow is to condition as much as possible without destroying any information about the parameter of interest. Support functions are independent of the rule for stopping the count provided they are not conditioned on any statistic which is itself informative, and this is true even if the rule is of the sequential type in which the count is stopped only when a chosen support difference between two competing hypotheses has been reached. Thus in a binomial sequential scheme the stopping rule can be represented as a boundary on the lattice diagram of the possible sample points; when a particular point is reached it would be misleading to condition the support on the achieved sample number, because this number would, by itself, generally be informative. One must simply write down the unconditional probability of reaching the sample point, and use this to derive the support. In the binomial case a sample point with r successes and $N - r$ failures will have a probability $Cp^r(1 - p)^{N-r}$, where C is a coefficient given by the number of paths to the point.[22] It will only be a binomial coefficient if the boundary does not interfere with any possible path, but the support for p induced by the sample point is in any case independent of C.

We are not here concerned with the benefits of sequential procedures, which are essentially decision procedures whose justification is to be sought in their repeated-sampling properties, but with what we ought to believe about a parameter from the knowledge that a particular sample point has been reached.[23]

The support supplies the necessary information. We are not even concerned with biasses that might arise on repeatedly taking decisions based on 'open-ended' schemes where the chosen boundary may never be reached,[24] for even if the stopping rule is 'stop when tired', the support carries the required information.

We shall meet the concept of conditional support again, when dealing with the elimination of nuisance parameters (section 6.3) and 2×2 tables (section 9.4). Example 6.3.1 illustrates the use of conditional support in the partitioning of the total information provided by a sample into its constituent parts.

3.7. EXAMPLE

A particularly simple demonstration of the Method of Support is afforded by applying it to the classic calculation which led Bernstein,[25] in 1924, to conclude that the ABO blood-groups in man were determined by three alleles (A, B, O) at a single locus rather than two alleles at each of two loci (A, a; B, b) as formerly thought.

Four blood-groups are distinguished: 'A', 'B', 'AB' and 'O', corresponding to the presence or absence of the antigens 'A' and 'B', 'AB' referring to the presence, and 'O' to the absence, of both. The two-locus hypothesis (H_2) supposed that a locus A, a controlled the antigen 'A' as follows: genotypes AA and Aa – 'A' present; genotype aa – 'A' absent. Similarly, an independent locus B, b controlled the antigen 'B' (table 1, column 3). The single-locus hypothesis (H_1) supposes that there are three alleles A, B and O, A and B conferring the corresponding antigens, and O conferring nothing (table 1, column 5). Note that an 'O' child cannot have an 'AB' parent on this hypothesis (H_1), though he can on H_2: $Aa\,Bb \times Aa\,Bb \rightarrow aa\,bb$, for example. But, remarkably enough, it was not the failure of H_2 to explain the segregation in families which led Bernstein to postulate H_1, but its failure to account for the frequencies of the four blood-groups in the population at large, as the following example shows.

EXAMPLE 3.7.1

Consider the data[26] used by Bernstein (table 1, column 2). The sample size is $n = 502$.

Bernstein observes that on H_2 we expect

$$(x_1 + x_3)(x_2 + x_3) = x_3 \qquad (3.7.1)$$

for a large sample from a population in Hardy–Weinberg equilibrium

TABLE 1. Bernstein's data on the ABO blood-groups, with genotypes and expectations on the two-locus (H_2) and single-locus (H_1) hypotheses.

(1) Group	(2) Observed proportions	(3) Genotypes on H_2	(4) Expected proportions on H_2	(5) Genotypes on H_1	(6) Expected proportions on H_1
'A'	$x_1 = 0.422$	*AAbb, Aabb*	0.358	*AA, AO*	$p(p + 2r) = 0.4112$
'B'	$x_2 = 0.206$	*aaBB, aaBb*	0.142	*BB, BO*	$q(q + 2r) = 0.1943$
'AB'	$x_3 = 0.078$	*AABB, AaBB, AABb, AaBb*	0.142	*AB*	$2pq = 0.0911$
'O'	$x_4 = 0.294$	*aabb*	0.358	*OO*	$r^2 = 0.3034$

(see below), whereas on these data the LHS is 0.142, nearly twice x_3. On H_1 he expects the relation

$$\{1 - \sqrt{(x_2 + x_4)}\} + \{1 - \sqrt{(x_1 + x_4)}\} + \sqrt{x_4} = 1, \quad (3.7.2)$$

which he finds adequately satisfied. With the Method of Support we may by-pass the consideration of these somewhat arbitrary indices, and calculate the relative support for H_1 versus H_2.

We do not need to evaluate the gene frequencies explicitly on H_2, because the class expectations may be found directly from a fourfold table, A and B being independent:

		A present	A absent	
B	present	*AB:* 0.142	*B:* 0.142	$AB + B$: 0.284
	absent	*A:* 0.358	*O:* 0.358	$A + O$: 0.716
		$AB + A$: 0.500	$B + O$: 0.500	all 1.000

The marginal totals give the proportions actually observed; the class expectations are then calculated assuming independence. Thus in a large sample we expect $x_1x_2 = x_3x_4$, which is equivalent to (3.7.1). In calculating the likelihoods for the two hypotheses, the multinomial coefficients (equation (2.4.1)) will be the same, and may therefore be discarded. For H_2 we need only calculate $S_2 = \Sigma a_i \ln p_i$ (equation (3.4.4)) over the four classes, where the p_i refer to the expectations (table 1, column 4), and $a_i = nx_i$. We find

$$S_2 = -647.50.$$

Note that we have implicitly evaluated the gene frequencies at both loci, so that H_2 involves two parameters.

On the single-locus theory the expectations in the four classes are given in table 1, column 6, where p, q and r are the frequencies of the A, B and O genes ($p + q + r = 1$). (We adhere to tne usual notation in spite of the incompatibility with the p_i that appear above.) Replacing the x_i in (3.7.2) by these expectations demonstrates Bernstein's relation.

The numerical expected proportions in column 6 are obtained by inserting the best-supported values for the gene frequencies p, q and r in the algebraic expectations. These values cannot be found explicitly, and must be obtained by iteration, as is done on the present data in example 6.8.2 where we find

$$\hat{p} = 0.2945,$$
$$\hat{q} = 0.1547,$$
and $$\hat{r} = 0.5508.$$

The support for H_1 is then

$$S_1 = -627.52,$$

TABLE 1. Bernstein's data on the ABO blood-groups, with genotypes and expectations on the two-locus (H_2) and single-locus (H_1) hypotheses.

(1) Group	(2) Observed proportions	(3) Genotypes on H_2	(4) Expected proportions on H_2	(5) Genotypes on H_1	(6) Expected proportions on H_1
'A'	$x_1 = 0.422$	*AAbb, Aabb*	0.358	*AA, AO*	$p(p + 2r) = 0.4112$
'B'	$x_2 = 0.206$	*aaBB, aaBb*	0.142	*BB, BO*	$q(q + 2r) = 0.1943$
'AB'	$x_3 = 0.078$	*AABB, AaBB, AABb, AaBb*	0.142	*AB*	$2pq = 0.0911$
'O'	$x_4 = 0.294$	*aabb*	0.358	*OO*	$r^2 = 0.3034$

and thus exceeds S_2 by 19.98 units. As with H_2, two independent parameters have been evaluated in H_1, so that the hypotheses are strictly comparable in 'simplicity'. A support difference of nearly 20 units is very substantial, corresponding to a ratio of likelihoods of over 4×10^8; Bernstein's judgement is indeed well supported.

3.8. SUMMARY

The adoption of the Likelihood Axiom leads to a method of forming opinions about statistical hypotheses which is intuitively satisfying, and which has all the properties that may reasonably be expected of a measure of support. The ensuing Method of Support is developed without reference to Bayes' Theorem, and provides a convenient way of recording, combining and assessing statistical information. Prior support is introduced as a means of quantifying prior opinions which do not justify probability statements, and the question of support functions based on conditional probabilities is discussed, with special reference to sequential procedures. Support functions are independent of the rule for 'stopping the count'. An example is given of the comparison of two hypotheses by the Method of Support.

CHAPTER 4

BAYES' THEOREM AND INVERSE PROBABILITY

4.1. INTRODUCTION

In section 3.5 there was an example of the use of support as a means of incorporating prior information into a final assessment. We had to argue solely by *induction*, that is, from the occurrence of an effect – the measured height of a column of mercury – to the presumed cause – an unknown atmospheric pressure. We were unable to argue by *analogy*, that is, by reference to a relevant series from which the case under discussion might be regarded as having been drawn at random, because we had no series of earlier measurements of atmospheric pressure.

Nowadays we could make a statement about the atmospheric pressure at noon at a particular place on a particular day without making any measurement at all, but purely by analogy. If the place is London, and the day April 1st, and we possess many years' records of the pressure in London at noon, we may adopt a probability model according to which this year's April 1st and all previous April 1sts in the series are regarded as being random examples of a population of April 1sts, with an associated population of atmospheric pressures. If the series is long enough we will be able to define the parameters of this population with some accuracy: say it is Normal with mean 760 and standard deviation 10 millimetres. Then the statement 'the pressure at noon today, April 1st, in London, exceeds 760 millimetres with probability $\frac{1}{2}$' is a valid statement of probability by analogy. We will consider at a later stage how good the analogy is.

Such a statement does not, of course, preclude us from making a measurement of the pressure today. Suppose that we do so, finding a value of 770 millimetres of mercury, with a Normal distribution of error having, as before, a standard deviation of one millimetre. The support for μ is thus $(770 - \mu)^2/2$. The question now arises as to how this information, obtained in the form of a support function by induction, is to be combined with the prior information, obtained in the form of a probability distribution by analogy.

In order to maximize this we introduce the Lagrangian multiplier λ and consider the new function

$$S'(p, \lambda) = \Sigma a_i \ln p_i + \lambda (\Sigma p_i - 1).$$

The stationary point of this new function is at

$$\frac{\partial S'}{\partial p_i} = \frac{a_i}{p_i} + \lambda = 0, \text{ (all } i\text{)}, \qquad \text{and} \qquad \frac{\partial S'}{\partial \lambda} = \Sigma p_i - 1 = 0,$$

whose solution is clearly

$$\hat{p}_i = a_i/n, \text{ (all } i\text{)}, \qquad \text{and} \qquad \lambda = -n.$$

In order to find the extended information matrix we note that

$$\frac{\partial^2 S}{\partial p_i^2} = -\frac{a_i}{p_i^2} \qquad \text{and} \qquad \frac{\partial^2 S}{\partial p_i \partial p_j} = 0, (i \neq j),$$

and hence that the matrix is

$$-\begin{pmatrix} -a_1/p_1^2 & 0 & \dots & 0 & 1 \\ 0 & -a_2/p_2^2 & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -a_k/p_k^2 & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix}.$$

The inverse of this is readily shown to be

$$-\frac{1}{\Sigma} \times \begin{pmatrix} -p_1^2\Sigma/a_1 + p_1^4/a_1^2 & \dots & p_1^2 p_k^2/a_1 a_k & p_1^2/a_1 \\ \dots & \dots & \dots & \dots \\ p_1^2 p_k^2/a_1 a_k & \dots & -p_k^2\Sigma/a_k + p_k^4/a_k^2 & p_k^2/a_k \\ p_1^2/a_1 & \dots & p_k^2/a_k & 1 \end{pmatrix}$$

where Σ stands for $(\Sigma p_i^2/a_i)$. On substituting for the observations in terms of the evaluates, $a_i = n\hat{p}_i$, and omitting the last row and column, we obtain the observed formation matrix

$$\begin{pmatrix} \frac{\hat{p}_1(1 - \hat{p}_1)}{n} & \dots & \frac{-\hat{p}_1\hat{p}_k}{n} \\ \dots & \dots & \dots \\ \frac{-\hat{p}_1\hat{p}_k}{n} & \dots & \frac{\hat{p}_k(1 - \hat{p}_k)}{n} \end{pmatrix}$$

which is the generalization of the formation of a binomial parameter (example 5.2.1).

The analogue of Newton–Raphson iteration when there are constraints is given by Aitchison and Silvey. For the above case of a single linear constraint $\Sigma\theta_i = 1$ it will be found that the effect of the constraint is to

modify the standard procedure by adding λ to each element of the unconstrained scores vector T at each iteration, and that

$$\lambda = -\frac{\sum_{i,j} F_{ij} T_i}{\sum_{i,j} F_{ij}},$$

where F is the inverse of the unconstrained information matrix.

We close this section with an example of the counting method (see example 5.7.2).

EXAMPLE 6.8.2

The genetics of the *ABO* blood-group system in man is described in section 3.7 (single-locus hypothesis, H_1) and the algebraic expectations in the four classes are given in table 1, column 6. Preserving the notation of section 3.7, the support function is

$$S(p, q, r) = n\{x_1 \ln p(p + 2r) + x_2 \ln q(q + 2r) + x_3 \ln 2pq + x_4 \ln r^2\},$$

which, omitting the constant term $nx_3 \ln 2$, reduces to

$$S(p, q, r) = n\{(x_1 + x_3) \ln p + (x_2 + x_3) \ln q + 2x_4 \ln r + x_1 \ln (p + 2r) + x_2 \ln (q + 2r)\}. \qquad (6.8.9)$$

In view of the restriction $p + q + r = 1$ the support surface may be plotted on a triangular Streng diagram. Figure 20 shows the diagram for the data given in table 1, column 2, the sample size n being 502.

In order to find the evaluates of p, q and r, iteration is necessary, and may be undertaken after the substitution of any one parameter in terms of the other two, or by using Fisher's method, or by using the Lagrangian multiplier method, or by using the counting method. Rather than following each cycle of the counting method numerically, as in example 5.7.2, we follow through a cycle algebraically, thus establishing the equivalent recurrence relations.

Let the actual proportions in the six genotype classes *AA*, *AO*, *BB*, *BO*, *AB* and *OO* be $c_1 \dots c_6$ respectively. We observe c_5 $(= x_3)$ and c_6 $(= x_4)$ directly, but the remainder are only observed in combination: $c_1 + c_2 = x_1$, $c_3 + c_4 = x_2$. If we could observe all the cs directly we would have the gene frequencies

$$\left.\begin{aligned} 2p' &= 2c_1 + c_2 + c_5 \\ 2q' &= 2c_3 + c_4 + c_5 \\ 2r' &= c_2 + c_4 + 2c_6. \end{aligned}\right\} \qquad (6.8.10)$$

Since we cannot observe $c_1 \ldots c_4$ directly, we get them from x_1 and x_2 by splitting the latter according to expectation, as the counting method prescribes:

$$\left.\begin{aligned} c_1 &= \frac{p}{p+2r}x_1 \\ c_2 &= \frac{2r}{p+2r}x_1 \\ c_3 &= \frac{q}{q+2r}x_2 \\ c_4 &= \frac{2r}{q+2r}x_2. \end{aligned}\right\} \qquad (6.8.11)$$

And also:

$$c_5 = x_3, \quad c_6 = x_4.$$

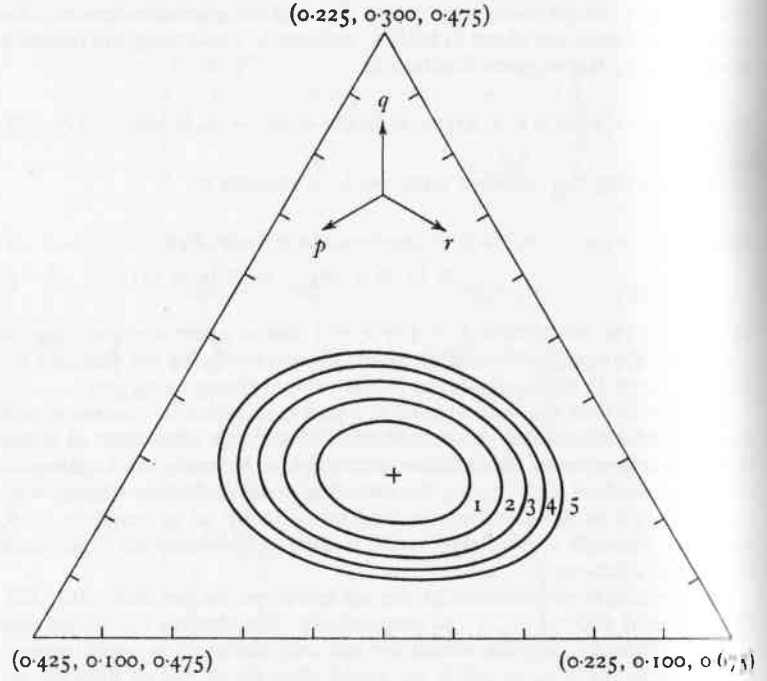

Figure 20. The support surface for the frequencies p, q and r of the A, B and O blood-group genes (example 6.8.2). Only a part of the Streng diagram is shown; the co-ordinates of the vertices of the triangle are indicated, and the scale is one division to 0.020 units of gene frequency. I am indebted to Mr C. G. Hopewell for computing the figure.

In (6.8.10) we have used primes to denote the gene frequencies because these equations are to be used to obtain new values after $c_1 \ldots c_6$ have been calculated from (6.8.11) using trial values of p, q and r.

Continuing to work algebraically, we eliminate $c_1 \ldots c_6$ between (6.8.10) and (6.8.11), obtaining

$$\left.\begin{aligned} p' &= \frac{p+r}{p+2r}x_1 + \tfrac{1}{2}x_3 \\ q' &= \frac{q+r}{q+2r}x_2 + \tfrac{1}{2}x_3 \end{aligned}\right\} \qquad (6.8.12)$$

and an equation for r'. But it is only necessary to work with p and q, and substituting $1 - p - q$ for r in (6.8.12) gives the recurrence relations

$$\left.\begin{aligned} p' &= \frac{1-q}{2-p-2q}x_1 + \tfrac{1}{2}x_3 \\ q' &= \frac{1-p}{2-2p-q}x_2 + \tfrac{1}{2}x_3. \end{aligned}\right\} \qquad (6.8.13)$$

Succeeding iterates, starting with trial values $p = q = 0.3$, for the data given in table 1, column 1, are given in table 5. The resulting evaluates are those that have been used in example 3.7.1.

TABLE 5. Successive iterates for the gene frequencies p and q obtained by using the counting method (example 6.8.2).

Iteration	p	q
—	0.3000	0.3000
1	0.3075	0.1701
2	0.2980	0.1564
3	0.2953	0.1549
4	0.2947	0.1547
5	0.2945	0.1547
6	0.2945	0.1547

$\hat{p} = 0.2945$, $\hat{q} = 0.1547$, $\hat{r} = 0.5508$

We may note that the evaluates must be the solutions to (6.8.13) when p' and q' are set equal to p and q, leading to simultaneous quadratic equations. Further, we can see – without giving a formal proof – why the counting method leads to the evaluates, for (6.8.10) gives the actual

values of the gene frequencies conditional on $c_1 \ldots c_6$, since only counting is involved, whilst (6.8.11) gives the most probable values of $c_1 \ldots$ given the gene frequencies. Thus the two sets of equations together give the most likely values for the frequencies.

The observed formation matrix is best calculated separately under the counting method, as it does not appear naturally during the course of iteration.[16] The most straightforward procedure is that discussed above in connection with the use of a Lagrangian multiplier, according to which the formation matrix is

$$\begin{pmatrix} B & -1 \\ -1 & 0 \end{pmatrix}^{-1} \tag{6.8.14}$$

with the final row and column removed, where B is the unconstrained information matrix. From (6.8.9) we readily find B to be n times

$$\begin{pmatrix} \frac{x_1+x_3}{p^2}+\frac{x_1}{(p+2r)^2} & 0 & \frac{2x_1}{(p+2r)^2} \\ 0 & \frac{x_2+x_3}{q^2}+\frac{x_2}{(q+2r)^2} & \frac{2x_2}{(q+2r)^2} \\ \frac{2x_1}{(p+2r)^2} & \frac{2x_2}{(q+2r)^2} & \frac{2x_4}{r^2}+\frac{4x_1}{(p+2r)^2}+\frac{4x_2}{(q+2r)^2} \end{pmatrix}$$

From this point on it is best to proceed numerically, inserting the evaluates $\hat{p}$, $\hat{q}$ and $\hat{r}$ in B, and then calculating (6.8.14). In the present case we find the formation matrix to be

$$\begin{pmatrix} 2.4984 & -0.4442 & -2.0542 \\ -0.4442 & 1.4246 & -0.9804 \\ -2.0542 & -0.9804 & 3.0345 \end{pmatrix} \times 10^{-4}.$$

The spans of $\hat{p}$, $\hat{q}$ and $\hat{r}$ (table 5) are thus 0.0158, 0.0119 and 0.0174 respectively.

6.9. SUMMARY

The interpretation of evaluates in multiparameter situations is considered, and a detailed method given for the case of approximately quadratic support surfaces, enabling, in particular, a formula to be given for the support for the sum of two or more parameters. The elimination of nuisance parameters is treated in depth, and shown to be possible if the likelihood factors, or if the model can be suitably restructured. The important concept of a neutral prior distribution is introduced for the probability distribution of a nuisance parameter which, if it were known, would make

no difference to the inference about the parameters of interest, and is shown not to imply the validity of the Bayesian approach. An example based on the Normal distribution is given.

General forms for the multiparameter score vector and information matrix are given, and formulae provided for the approximate combination of vectors of evaluates. After a description of Newton–Raphson iteration for many parameters, various modifications are considered, including the case in which a partial analytic solution to the support equations is available, and the case of a restriction amongst the parameters. The chapter concludes with an example of the counting method.