304

MODEL SELECTION IN MULTINOMIAL EXPERIMENTS

JAN DE LEEUW
DEPARTMENT OF DATA THEORY FSW
UNIVERSITY OF LEIDEN
MIDDELSTEGRACHT 4
2312 TW LEIDEN, THE NETHERLANDS

Multinomial experiments and models

We first define what we mean by a multinomial experiment. The first component of the definition is a set $\mathcal{P}$, the *population*. Elements of $\mathcal{P}$ are called *objects*. We shall assume in this paper that $\mathcal{P}$ is finite, and that there is a list of $\mathcal{N}$ objects defining this population. The objects are of m different types. Suppose there are $\mathfrak{n}_j$ objects of type j, and define the *theoretical proportions* $\pi_j = \mathfrak{n}_j/\mathcal{N}$. Now consider the set of all sequences of length N, with elements from $\mathcal{P}$. There are $\mathcal{N}^N$ such sequences, but if types are indistinguishable not all of them are different. In particular there are $\prod_{j=1}^{m} \mathfrak{n}_j^{n_j}$ indistinguishable sequences with *frequencies* n_j, and if the order of the objects in the sequence is irrelevant there are even N! $\prod_{j=1}^{m} \{\mathfrak{n}_j^{n_j}/n_j!\}$ such sequences. The proportion of sequences with given frequencies, where order is not taken into account, is consequently prop$(n_1,...,n_m)$ = N! $\prod_{j=1}^{m} \{\pi_j^{n_j}/n_j!\}$. With each sequence of length N we associate the *vector of proportions* p, with $p_j = n_j/N$, and the *multinomial probability* prop$(n_1,...,n_m)$. Another way of expressing this is that for each N we have defined a random vector $\underline{p}_N$, taking values in S^{m-1}, the unit simplex in $\Re^m$. We use the convention of underlining random variables in this paper. Thus $\underline{p}_N$ is a sequence of random vectors.

This describes the multinomial experiment. We have the pair $(\mathcal{N},\pi)$ and the sequence $\underline{p}_N$, and generally $(\mathcal{N},\pi)$ is unknown. The first and most basic problem we shall study is the *estimation* of π. This means, informally, that we construct a new sequence of random variables $\Phi(\underline{p}_N)$ which is (in some sense) as close as possible to the unknown π. Observe that we restrict our attention to functions of the proportions $\underline{p}_N$. Precise definitions of closeness and optimality will be given below.

In order to make our estimates accurate, we must try to take as much prior information into account as possible. Science is cumulative, and presumably we already know something about the subject area in question. Prior information takes the form of a *model* in this paper. A model Ω is a subset of S^{m-1}. If we say that model Ω is *true*, then we mean that $\pi \varepsilon \Omega$. The second problem we shall investigate in this paper, if assuming that a model is true will help us to compute more precise estimates, even in those cases in which the model actually is only approximately true. This could be formulated as deciding whether a model is *useful*. Moreover, and finally, we shall study a similar problem for the case in which we have a finite number of models $\Omega_1,...,\Omega_t$. This is one version of the problem of *model choice*, in which we want to find out which model helps us best to improve our estimates, i.e. which model is most useful.

Up to now we have defined a mathematical structure called a multinomial experiment, but we have not yet specified how this structure is connected with empirical data. Suppose we have selected N objects from the population. We can compute the m quantities n_j, which are the *observed frequencies* of objects with type j, and $p_j = n_j/N$, which are the *observed proportions*. We assume now, that p is a *realization* of the random variable $\underline{p}_N$. This means that we compare statistics computed from p with statistics computed from the other possible realizations of this random variable. This defines our *framework of replication* (De Leeuw, 1984), the set of all possible outcomes with which we compare our results. Because our framework gives all sequences the same probability this means that we act as if that our sample is a simple random sample, drawn with replacement from a finite population. It is important to realize that all statistical statements are about this framework $\underline{p}_N$, not about the data p, and also not about the true value π. It is also important to see that a framework is never true or untrue (for a particular empirical situation), but it is either relevant or irrelevant. We do not assume, in any sense of the word, that our sample is indeed a random sample. Our results have a simple combinatorial interpretation. The counting of samples, familiar from combinatorics, is done by using asymptotic approximations. Thus the 'foundational' and 'inferential' aspects of statistics, which are both controversial and problematical, are not relevant for our discussion. We go 'back to the Laplace definition' (Hemelrijk, 1968).

It is perhaps worth mentioning that our results, which are true for multinomial situations, can be generalized with little effort to product-multinomial situations in which we deal with more than one population (or with a stratified population). With a bit more effort they can be extended to more complicated sampling designs. And with a considerable amount of technical effort they can also be extended to infinite dimensional problems (functions of empirical distribution functions, regression functions, or density estimates). The basic methodological ideas and interpretations remain the same in these alternative situations.

Models are never true

There has recently been much discussion in statistics about the role of models. Compare McCullagh and Nelder (1983, section 1.1), De Leeuw (1984), or Nelder (1984) for fairly modern introductions. The general consensus seems to be that, contrary to the practice of classical statistics, we must not routinely assume that our models are true, i.e. that $\pi \varepsilon \Omega$. Models are approximations, which are summaries of the prior scientific information we have, but which can still be quite far off the mark in some cases. We need models to increase the stability of our estimators, descriptors, and predictors. Using a model which is perhaps not true, but based on only a few parameters, means that we trade statistical stability for unbiasedness. Making too few assumptions means instability, and may not be very cumulative from the scientific point of view. Too many assumptions means a great deal of stability, but possibly a very large bias. We need to find a compromise between the two, and for such a compromise models are necessary. The fact that models are approximations, and that approximation errors may be more important than sampling errors, is stressed by various schools of data analysis. We mention Tukey (1980), Guttman (1985), Benzécri et al. (1973), Gifi (1981), Box (1979), Verbeek (1984), Kalman (1983), Willems (1986).

If we compare the description of multinomial experiments above, for example, with actual data collection or experimentation, then the description involves various *idealizations*. In the first place populations are usually not given by finite lists. The current population of the Netherlands, for instance, is not exactly well defined. We must decide whether we mean all human beings currently within our borders, or all human beings with dutch citizenship, or all human beings registered by the proper authorities as living in Netherlands. It is clear that a little bit of creative thinking shows that there are various borderline cases and exceptions, which make it difficult to construct lists, even in theory. And, more seriously, these populations are changing every day because of births, deaths, naturalizations, immigrations, tourism, and so on.

The second idealization is the idea of sampling with replacement. Of course this can be carried out formally only if we actually have a list, while we have just decided that such lists do not exist. The actual use of 'with replacement' is very rare, except in artificial sampling experiments. In many cases *hypergeometric experiments*, which do not use replacement, are more realistic. Fortunately most of what we say applies directly to hypergeometric experiments as well, and if N is large the difference between the two random variables is small anyway. In actual experiments with human subjects simple comparison with all possible subsets of size N is often not the most interesting comparison, because stratification and clustering are the rule rather than the exception. This means, in other words, than simple random sampling (with or without replacement) does not provide the most relevant replication framework.

And, as we have already seen, our notion of a model itself is an idealization. Models are never true, they are at best good approximations, where 'good' means good enough for practical purposes. Classical mechanics is not true, it is an approximation which is good enough for most purposes. Relativity theory improves the approximation, but this still does not make it true. There is no reason to suppose that we shall ever find a model which is 'true', in the sense that it gives perfect predictions of all phenomena under consideration, or even predictions which can never be improved any more. This notion of truth is not really needed, if only because of the omnipresence of measurement error. That 'truth' exists is, in itself, a model, which may be useful for guiding our actions, but which is not part of science.

The fact that even the simple multinomial model involves many idealizations and approximations which are, at least in many cases, not really appropriate, need not disturb us greatly. The model is meant as an approximation, and this is not only true for the model Ω but also for the framework modelling the sampling procedure. The question is whether these idealizations make it possible to give descriptions and predictions which are useful and good enough for practical purposes, or which are as good as possible under the circumstances. There are many situations in the social sciences in which the idea of random sampling from a well-defined population is much more far fetched than in the simple demographic or survey situations we have in mind here. In order to apply basic statistical ideas much more far reaching idealizations, such as super-populations or infinite hypothetical populations, are needed. One can seriously wonder in many of these cases if the classical statistical approach is really fruitful, although not many systematic alternatives have been proposed, at least for prediction. In such cases a satisfactory description of the data is perhaps all that can be realistically expected. Techniques that try to go beyond mere description must make many assumptions that are often untested or even untestable. This entails that the conclusions in these cases are based largely on prejudice.

Example: twins

In the first part of the paper we use a simple multinomial example, taken from Andersen (1980). Suppose that we study the gender of twins. There are three outcomes: girl-girl, boy-boy, and girl-boy. We start with simple binomial models for the probability of a girl and the probability of a boy. Thus we ignore the fact that twins come in pairs, and we only count the sexes. The saturated binomial model leaves prob(♀) and prob(♂) = 1 - prob(♀) free, the restricted binomial model sets prob(♀) = prob(♂) = 1/2.

The situation becomes a bit more complicated if we study the multinomial experiment with the three possible combinations of boy and girl as outcomes. The first model one may think of, is that of independence (model A in the sequel). If the probability of a girl is ω, then

$$\text{prob}(♀♀) = \omega^2, \tag{1a}$$

$$\text{prob}(♂♂) = (1 - \omega)^2, \tag{1b}$$

$$\text{prob}(♀♂) = 2\omega(1 - \omega). \tag{1c}$$

Is this a useful model ? In order to study this question we first need data. In Figure 1 we have drawn the simplex S^2, and the one-dimensional quadratic manifold defined by model A. We have also drawn in the data points for samples from six birth centres. These data are given in Table 1, which is taken from Andersen (1980, page 93).

	2 boys	2 girls	1 boy, 1 girl
Melbourne	29	36	33
Sao Paulo	61	69	81
Santiago	88	77	76
Alexandria	116	114	161
Hong Kong	45	46	34
Zagreb	20	32	30

Table 1: twin data

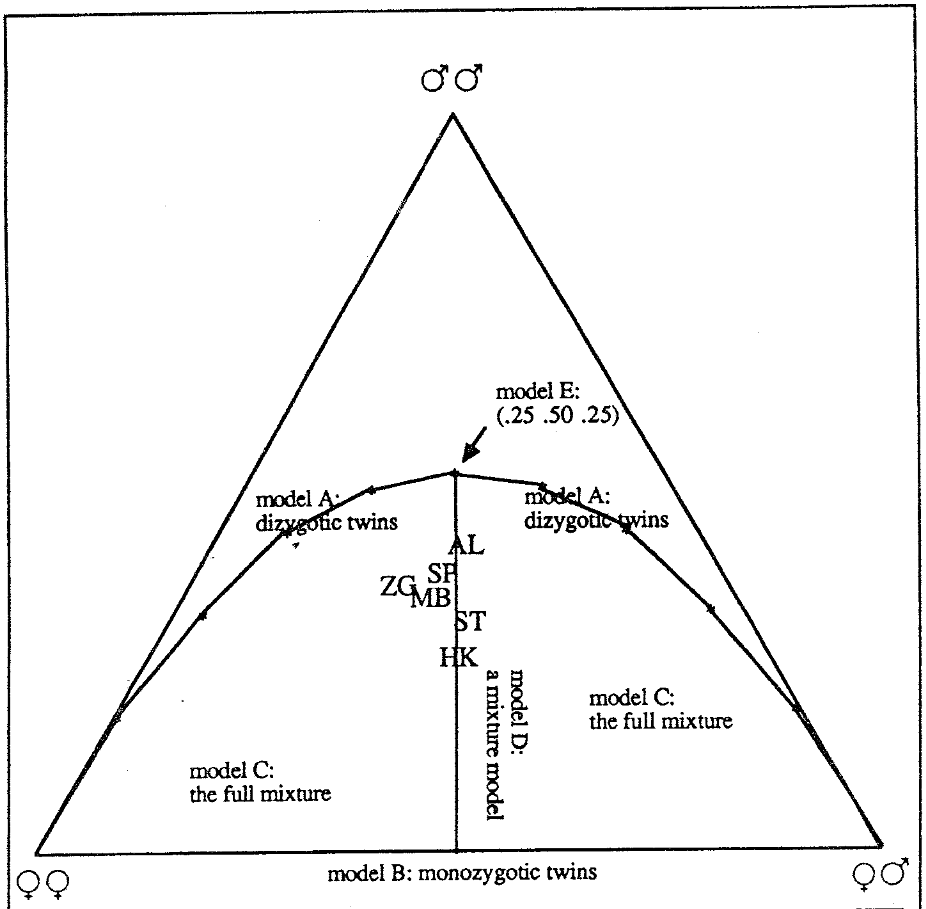

Figure 1: Models for the twin data.

It seems clear from the data that there are fewer girl-boy twins than expected on the basis of independence. The explanation is that twins are either monozygotic or dizygotic, and that monozygotic twins are always of the same sex. For monozygotic twins only we consequently hav

$$\mathrm{prob}(\text{♀♀}) = \theta, \tag{2a}$$
$$\mathrm{prob}(\text{♂♂}) = 1 - \theta, \tag{2b}$$
$$\mathrm{prob}(\text{♀♂}) = 0. \tag{2c}$$

This is, of course, the basis of the triangle in Figure 1. Model B fits badly, which is not very surprising because certainly not all twins are monozygotic.

What we really need is a mixture of models A and B, in which we suppose that the proportion of monozygotic twins is λ, an additional parameter. This mixture model C is the convex region between the 'monozygotic' basis of the triangle and the 'dizygotic' quadratic of model A, and clearly all data points are in this region. A more specific mixture model D supposes that $\omega = \theta = .5$. Then

$$\mathrm{prob}(\text{♀♀}) = (1 + \lambda)/4, \tag{3a}$$
$$\mathrm{prob}(\text{♂♂}) = (1 + \lambda)/4, \tag{3b}$$
$$\mathrm{prob}(\text{♀♂}) = (1 - \lambda)/2. \tag{3c}$$

This is the vertical line segment in Figure 1, which seems to describe the data quite well.

Models A, B, and D are all one-dimensional. Model C is two-dimensional, but still restrictive. For the sake of completeness we also define the zero-dimensional model E, which is the intersection of A and D. It consists of the single point (0.25 0.50 0.25). The question is if these models help us to find better estimates of π From the figure it would seem that model D is best, but we would like to have one or more procedures that make it possible to make a choice if the situation is less clear. For this we need some formal statistical theory, which we introduce in a somewhat unconventional way. The methods we use are inspired by the treatment of multinomial experiments in the book of Rao (1973), but also by the recent emphasis on geometrical methods in statistical estimation theory, and by the equally recent work on resampling methods.

Minimum distance methods

An *estimator* for a multinomial experiment is a continuous mapping Φ of S^{m-1} into S^{m-1}. The idea is, that we try to estimate the population value or true value π. We do this by associating an *estimate* $\Phi(p)$ with each vector of observed frequencies p. The estimator Φ can be thought of as a random variable, the estimate $\Phi(p)$ is a realization of this variable. The identity mapping is an obvious estimator, but in some circumstances it may be possible to improve on this estimator by taking prior information into account. In our case the prior information takes the form of a model Ω, in other cases, which we do not study here, it may take the form of a prior distribution. The procedures we develop have a frequency interpretation even if they use prior distributions, because our replication framework simply remains the counting of samples.

The problem we study is how to estimate π 'optimally'. For this we first have to define optimality, of course. The theory we shall explain is more limited in scope than other forms of statistical large sample theory, but sufficiently general to cover many situations of practical interest. It uses the currently popular geometrical terminology of distance, projection, and manifolds. Introduce a *multinomial separator*, which is a function Δ on $S^{m-1} \times S^{m-1}$ with the properties that $\Delta(p,q) \geq 0$ for all p,q, and $\Delta(p,q) = 0$ if and only if $p = q$. Throughout the paper we shall not be concerned with regularity conditions. We simply assume that Δ is sufficiently many times differentiable for our results to be true. Given the separator Δ, and the model Ω, we now define the estimator Φ by $\Phi(p) = \text{argmin} \{\Delta(p,q) \mid q \varepsilon \Omega\}$, where we assume that for each $p \varepsilon S^{m-1}$ the minimum exists, and is unique. Thus $\Phi(p)$ is the 'Δ-projection' of p on Ω. Compare Figure 2.

It follows directly that $\Phi(p) = p$ for all $p \varepsilon \Omega$. This condition is known as *F-consistency*, where the F stands for Fisher, who introduced the concept in the twenties. If Φ is F-consistent for a given model Ω containing π, and Φ is also differentiable, it follows that the distribution of $N^{1/2}(\Phi(\underline{p}) - \pi)$ converges to a multivariate normal distribution. We say that Φ is CAN, or *consistent asymptotically normal.*

Our comparison of models will be based on the following two statistics. Compare Figure 2. In the first place $\Delta(\pi,\Phi(\underline{p}))$, the *estimation error*, i.e. the distance from $\Phi(\underline{p})$ to the true value π. It is obvious how the estmation error must be interpreted. It tells us how far off we are if the observed value is $\underline{p}$ and the true value is π. In the second place we study the *prediction error* $\Delta(\underline{q},\Phi(\underline{p}))$, with $\underline{q}$ an independent vector of proportions from the same multinomial distribution. This distance shows how well the estimate $\Phi(\underline{p})$ predicts an independent replication. It is used to *cross validate* the estimate. Observe that both the estimation error and the prediction error are in general nondegenerate random variables, and that we consequently study their distribution.

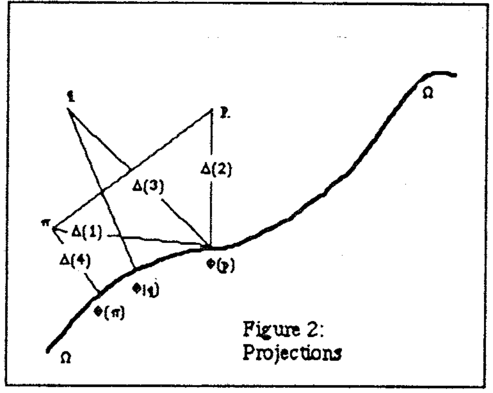

Figure 2: Projections

The third variable we use, among other things to compute our estimates, is the *projection distance* $\Delta(\underline{p},\Phi(\underline{p})) = \min \{\Delta(\underline{p},q) \mid q \varepsilon \Omega\}$. We are not really interested in the projection distance as such, but it is a very important intermediate quantity. The reasons for this are clear. In the usual cases in data analysis we do not have a realization available of $\Delta(\pi,\Phi(\underline{p}))$, because π is unknown, and of $\Delta(\underline{q},\Phi(\underline{p}))$, because we do not have a second independent sample. We only have a single realization of $\Delta(\underline{p},\Phi(\underline{p}))$.

It is instructive to find out what happens with the statistics we have introduced in the 'extreme' cases $T\Omega = S^{m-1}$ and $\Omega = \{\pi_0\}$, where π_0 is not necessarily equal to the true value π. For the zero dimensional model π_0 the estimation error is nonrandom, and equal to *approximation error* $\Delta(\pi,\Phi(\pi))$, while the prediction error and the projection distance are two independent copies $\Delta(\underline{p},\pi_0)$ and $\Delta(\underline{q},\pi_0)$ of the same random variable. For the *saturated model* S^{m-1} we have a zero projection distance, and a zero approximation error. The estimation error is $\Delta(\pi,\underline{p})$ and the prediction error is $\Delta(\underline{q},\underline{p})$.We shall now proceed to compute asymptotic distributions of these separator-statistics, emphasizing the limit of their expectations. The expected value of the estimation error is called the *bias*, the expected value of the prediction error we call the *distortion*.

The first result is very simple. Clearly the expected values of estimation error, prediction error, and projection distance are asymptotically equal to the approximation error $\Delta(\pi,\Phi(\pi))$. Thus $E(\Delta(\underline{p},\Phi(\underline{p})) = \Delta(\pi,\Phi(\pi)) + o(1)$, and we have the same expression for $E(\Delta(\pi,\Phi(\underline{p}))$ and $E(\Delta(\underline{q},\Phi(\underline{p}))$. This means that we can estimate the approximation error, the bias, and the distortion consistently by using the projection distance. This result is not very satisfactory, however, because the projection distance is, by definition, smallest for high-dimensional models, and is zero for $\Omega = S^{m-1}$. Thus 'large' models always give small estimates of bias and distortion, the larger the model the smaller the estimate. This is contrary to the well-known phenomenon that prediction error increases if our models are too large. If we want to obtain more precise comparisons, we need more precise approximations.

Let us introduce the following short-hand notation for the partial derivatives. We use $\lambda(p,q)$ for $\mathbf{D}_1\Delta(p,q)$, and $\eta(p,q)$ for $\mathbf{D}_2\Delta(p,q)$. For the second derivatives we use $A(p,q) = \mathbf{D}_{11}\Delta(p,q)$, $B(p,q) = \mathbf{D}_{22}\Delta(p,q)$, and $C(p,q) = \mathbf{D}_{12}\Delta(p,q)$. If the arguments are not indicated the derivatives are evaluated at $(\pi,\Phi(\pi))$, which is equal to (π,π) if $\pi \in \Omega$. For $\mathbf{D}\Phi(p)$ we use $G(p)$, G without argument is evaluated at $\Phi(\pi)$. The matrix $T_j(p)$ contains the second partials $\mathbf{D}^{(2)}\phi_j(p)$. With T_j we again mean $T_j(\Phi(\pi))$. Finally we define $\Gamma = \Sigma_{j=1}^{m} \eta_j T_j$. It is not difficult to see that the properties assumed for Δ imply that $\lambda(p,p) = \kappa u$ fo some κ, where u has all elements equal to $+1$, and $\eta(p,p) = -\kappa u$. Moreover $B(p,p) = A(p,p) = -C(p,p) = -C'(p,p)$ for all p in S^{m-1}. These four matrices with second order partials are all positive semidefinite.

Now let $\underline{\delta} = N^{1/2}(\underline{p} - \pi)$. Then $\underline{\delta}$ is asymptotically normal with mean zero, and variance $V = \Pi - \pi\pi'$, $\Pi = \mathrm{diag}(\pi)$. We write

$$\begin{aligned}\Delta(\pi,\Phi(\underline{p})) &= \Delta(\pi,\Phi(\pi + N^{-1/2}\underline{\delta})) = \Delta(\pi,\Phi(\pi)) + N^{-1/2}\eta' G\underline{\delta} + \\ &\quad + 1/2\, N^{-1}\underline{\delta}'\{\Gamma + G'BG\}\underline{\delta} + o_p(N^{-1}).\end{aligned} \tag{4a}$$

In the same way

$$\Delta(\underline{p},\Phi(\underline{p})) = \Delta(\pi + N^{-1/2}\underline{\delta},\Phi(\pi + N^{-1/2}\underline{\delta})) = \Delta(\pi,\Phi(\pi)) + N^{-1/2}(\lambda'\underline{\delta} + \eta'G\underline{\delta}) + \\ + 1/2\, N^{-1}\underline{\delta}'\{A + C'G + G'C + \Gamma + G'BG\}\underline{\delta} + o_p(N^{-1}), \quad (4b)$$

and

$$\Delta(\underline{q},\Phi(\underline{p})) = \Delta(\pi + N^{-1/2}\underline{\delta}_1,\Phi(\pi + N^{-1/2}\underline{\delta}_2)) = \Delta(\pi,\Phi(\pi)) + N^{-1/2}(\lambda'\underline{\delta}_1 + \eta'G\underline{\delta}_2) + \\ + 1/2\, N^{-1}\{\underline{\delta}_1'A\underline{\delta}_1 + 2\,\underline{\delta}_1'CG\underline{\delta}_2 + \underline{\delta}_2'(\Gamma + G'BG)\underline{\delta}_2\} + o_p(N^{-1}). \quad (4c)$$

It follows that

$$N\{\, E\{\, \Delta(\pi,\Phi(\underline{p})) - \Delta(\pi,\Phi(\pi))\}\} = 1/2 \operatorname{tr} \{\Gamma + G'BG\}V + o(1), \quad (5a)$$

$$N\{E\{\, \Delta(\underline{p},\Phi(\underline{p})) - \Delta(\pi,\Phi(\pi))\}\} = 1/2 \operatorname{tr} \{A + C'G + G'C + \Gamma + G'BG\}V + o(1), \quad (5b)$$

$$N\{E\{\, \Delta(\underline{q},\Phi(\underline{p})) - \Delta(\pi,\Phi(\pi))\}\} = 1/2 \operatorname{tr} \{A + \Gamma + G'BG\}V + o(1). \quad (5c)$$

In deriving these expressions we have not used the fact that Φ is the projection of p in the metric Δ on Ω. The formulae (4) and (5) are true for any function of the proportions (which has sufficiently many derivatives). If we use the fact that Φ is a projection, then we can derive a formula for the matrix of partial derivatives G, using the implicit function theorem (compare Wolfe, 1976, or Abatzoglou, 1979, for closely related results). First introduce a local coordinate system $\Psi(\theta)$, with $\theta \in \Re^r$, in the point $\Phi(\pi)$. The matrix H contains the partial derivatives of Ψ with respect to the r coordinates, i.e. the columns of H are a basis for the tangent space of Ω in $\Phi(\pi)$. Also define the matrix ∇ with elements

$$\alpha_{st} = \Sigma_{k=1}^{m}\, \eta_k\, \partial^2\psi_k/\partial\theta_s\partial\theta_t, \quad (6)$$

where the second order derivatives are evaluated at $\Phi(\pi)$. Now the partials are given by

$$G = -H\{H'BH + \nabla\}^{-1}H'C. \quad (7)$$

Formula (7) can be substituted in (4) and (5), but we do not present the resulting expressions. We do point out that in the extreme cases S^{m-1} and π_0 we have G = I or G = 0. In both cases the map Φ is linear, and $\Gamma = 0$. This obviously leads to dramatic simplifications in (4) and (5).

In the general case it is somewhat tedious to use (5), because it involves the complicated matrix Γ of second derivatives of Φ. But it follows from (5) that

$$N\{\, E\{\, \Delta(\pi,\Phi(\underline{p})) - E\{\, \Delta(\underline{p},\Phi(\underline{p}))\}\} = -1/2 \operatorname{tr} \{A + C'G + G'C\}V + o(1), \quad (8a)$$

$$N\{E\{\,\Delta(\underline{q},\Phi(\underline{p})) - E\{\,\Delta(\underline{p},\Phi(\underline{p}))\}\} = -\operatorname{tr} G'CV + o(1). \qquad (8b)$$

Thus if we define $\vartheta(p) = \{N\Delta(p,\Phi(p)) - \operatorname{tr} C'GV\}/N$, with C, G, and V evaluated at $(p,\Phi(p))$, then

$$E\{\Delta(\underline{q},\Phi(\underline{p}))\} = E\{\vartheta(\underline{p})\} + o(N^{-1}). \qquad (9a)$$

In the same way, with $\wp(p) = \{N\Delta(p,\Phi(p)) - 1/2 \operatorname{tr} AV - \operatorname{tr} C'GV\}/N$, we have

$$E\{\Delta(\pi,\Phi(\underline{p}))\} = E\{\,\wp(\underline{p})\} + o(N^{-1}). \qquad (9b)$$

This means that bias and distortion can be estimated consistently by using $\wp(\underline{p})$ and $\vartheta(\underline{p})$, for which we do not need the second order derivatives of Φ.

Best asymptotically normal theory

Now suppose the model is true. Then $\Phi(\pi) = \pi$, and expansions (4) simplify to

$$\Delta(\pi,\Phi(\underline{p})) = 1/2\, N^{-1}\underline{\delta}'G'BG\underline{\delta} + o_p(N^{-1}), \qquad (10a)$$

$$\Delta(\underline{p},\Phi(\underline{p})) = 1/2\, N^{-1}\underline{\delta}'(I - G)'B(I - G)\underline{\delta} + o_p(N^{-1}), \qquad (10b)$$

$$\Delta(\underline{q},\Phi(\underline{p})) = 1/2\, N^{-1}(\underline{\delta}_{-1} - G\underline{\delta}_2)'B(\underline{\delta}_{-1} - G\underline{\delta}_2) + o_p(N^{-1}). \qquad (10c)$$

Thus $N\Delta(\pi,\Phi(\underline{p}))$, $N\,\Delta(\underline{p},\Phi(\underline{p}))$, and $N\,\Delta(\underline{q},\Phi(\underline{p}))$ are asymptotically quadratic forms in normal variables. We can compute their moments, and in particular

$$NE\{\Delta(\pi,\Phi(\underline{p}))\} = 1/2 \operatorname{tr} G'BGV + o(1), \qquad (11a)$$

$$NE\{\,\Delta(\underline{p},\Phi(\underline{p}))\} = 1/2 \operatorname{tr} (I - G)'B(I - G)V + o(1), \qquad (11b)$$

$$NE\{\,\Delta(\underline{q},\Phi(\underline{p}))\} = 1/2 \operatorname{tr} (V + G'VG)B + o(1). \qquad (11c)$$

We now study the concept of *optimality* a bit more in detail, using bias and expected prediction error as criteria. The expected values in (11) still depend on B, which is a property of the separator Δ, and on G, which is a property of the estimator Φ. We know that $\Phi(p) = p$ for all $p \in \Omega$. Using a local coordinate system Ψ around π, it follows that $\Phi(\Psi(\theta)) = \Psi(\theta)$, with $\Psi(\theta_0) = \pi$. By differentiating this we see that $GH = H$, where $H = \mathbf{D}\Psi(\theta_0)$, i.e. H is again a basis for the tangent space of Ω at π. But $GH = H$ can also be written as $G = H(H'\Pi^{-1}H)^{-1}H'\Pi^{-1} + TH_{\perp}'$, with $H_{\perp}$ a basis for the null space of H, i.e. the normal space to Ω at π. Now $\mathrm{tr}\ BGVG' = \mathrm{tr}\ BH(H'\Pi^{-1}H)^{-1}H' + \mathrm{tr}\ BTH_{\perp}'VH_{\perp}T' \geq \mathrm{tr}\ BH(H'\Pi^{-1}H)^{-1}H'$. Here we have used the fact that B is positive definite, while $V\Pi^{-1}H = H$. Estimators Φ for which we have that $G = H(H'\Pi^{-1}H)^{-1}H'\Pi^{-1}$ are called BAN, or *best asymptotically normal*. Compare Neyman (1949), Wijsman (1959), Berkson (1980), Bemis and Bhapkar (1983), and LeCam (1986, section 11.10) for various aspects of the general theory of BAN estimation. The estimates are asymptotically normal, and they are are best in the class of F-consistent estimators, in the sense that their variance $(H'\Pi^{-1}H)^{-1}$ is minimal. If we apply this to (11) we find that

$$NE\{\Delta(\pi,\Phi(\underline{p}))\} = 1/2\ \mathrm{tr}\ (H'\Pi^{-1}H)^{-1}H'BH + o(1), \tag{12a}$$

$$NE\{\Delta(\underline{p},\Phi(\underline{p}))\} = 1/2\ \mathrm{tr}\ VB - 1/2\ \mathrm{tr}\ (H'\Pi^{-1}H)^{-1}H'BH + o(1), \tag{12b}$$

$$NE\{\Delta(\underline{q},\Phi(\underline{p}))\} = 1/2\ \mathrm{tr}\ VB + 1/2\ \mathrm{tr}\ (H'\Pi^{-1}H)^{-1}H'BH + o(1), \tag{12c}$$

if Φ is BAN, and this is optimal, no matter how we choose B (i.e. no matter what separator we use).

The interesting question remains which separators give BAN estimates when minimized (also compare Taylor, 1953). From (8) we have, if the model is true, $G = H\{H'BH\}^{-1}H'B$.We see that Φ is BAN, for any model, if $B = \kappa\Pi^{-1}$ for some $\kappa \neq 0$. In this case we say that Δ is an *optimal multinomial separator*. It is *normalized* if $\kappa = 2$. A number of these normalized optimal multinomial separators are given by Rao (1973, section 5d.2). Thus if we derive our F-consistent estimator by projection using a separator, then we must choose an optimal separator to get BAN estimates. If the model is true, then the combination of an arbitrary normalized optimal separator and an arbitrary BAN estimate gives a technique for which the asymptotic distributions of the three statistics $N\ \Delta(\pi,\Phi(\underline{p}))$, $N\ \Delta(\underline{p},\Phi(\underline{p}))$, and $N\ \Delta(\underline{q},\Phi(\underline{p}))$ are central chi squares, with the appropriate numbers of degrees of freedom. In addition

$$NE\{\Delta(\pi,\Phi(\underline{p}))\} = r + o(1), \tag{13a}$$

$$NE\{\Delta(\underline{p},\Phi(\underline{p}))\} = (m - r - 1) + o(1), \tag{13b}$$

$$NE\{\Delta(\underline{q},\Phi(\underline{p}))\} = (m + r - 1) + o(1). \tag{13c}$$

This implies that $\{N\Delta(\underline{p},\Phi(\underline{p})) + 2r\}/N$ is a consistent estimate of the distortion, which seems a very simple way to justify the AIC-criterion of Akaike (1977). Compare also Sakamoto, Ishiguro, and Kitagawa (1986). Also $\{N\Delta(\underline{p},\Phi(\underline{p})) - (m - 2r - 1)\}/N$ is a consistent estimate of the bias.

Model comparison if the models are true

Let us now briefly review some of the techniques that have been proposed in statistics to compare models. These comparisons have usually been in terms of significance tests, and they try to answer the question 'is the model Ω true ?'. We have emphasized from the start that the answer to this question is simply 'no'. One does not need statistics to answer it, the answer follows from general considerations about the role of models. Our suggestion is to try and answer the question 'is the model Ω helpful in improving our predictions and descriptions ?'. This is a different question, and it is consequently not surprising that it may have a different answer. But a review of the classical procedures from our more general point of view is still useful.

Testing if a single model Ω is satisfactory can be interpreted as a problem of model choice: it compares Ω and S^{m-1}. The classical statistical technique for comparing Ω and S^{m-1} is to use the statistic $N\Delta(\underline{p},\Phi(\underline{p}))$, with Φ a BAN-estimator, and with Δ an optimal normalized multinomial separator. We already know that $N\Delta(\underline{p},\Phi(\underline{p}))$ is asymptotically a chi square with m - r - 1 degrees of freedom. The classical statistical procedure, due for r = 0 to Pearson and for general r to Fisher, tells us to reject the model Ω if $N\Delta(\underline{p},\Phi(\underline{p}))$ is too large, because this implies that 'either the model is untrue, or a very improbable event has occurred'. There is no reason to worry if you have trouble understanding the syllogism in the previous sentence. Nobody understands it. But the procedure can be interpreted quite simply in our terms. We evaluate a model by computing the integral from $N\Delta(\underline{p},\Phi(\underline{p}))$ to $+\infty$ of the χ^2_{m-r-1} distribution. The resulting 'P-value' must be as large as possible. It is difficult for most models to compete in P-value with the saturated model, because this has a P-value of one.

We now illustrate the procedure on our twin example. As the separator we take the one associated with the method of maximum likelihood. This is

$$\Delta(p,q) = 2\,\Sigma_{j=1}^{m}\{p_j \ln p_j - p_j \ln q_j\}. \tag{14}$$

Thus $\lambda_j(p,q) = 2(1 + \ln p_j - \ln q_j)$ and $\eta_j(p,q) = -2p_j/q_j$. Moreover $A(p,q) = 2P^{-1}$, $C(p,q) = -2Q^{-1}$, and $B(p,q) = 2PQ^{-2}$. Here P = diag(p) and Q = diag(q). If p = q then $\lambda(p,p) = 2u$, $\eta(p,p) = -2u$, where u has all elements equal to +1, and $A(p,p) = B(p,p) = -C(p,p) = 2P^{-1}$, which shows that the maximum likelihood estimator is BAN.

We start with the simple binomial model, which says that the probability of a boy is equal to the probability of a girl. The projection distance is $2N\{\underline{p} \ln \underline{p} + (1 - \underline{p}) \ln (1 - \underline{p}) + \ln 2\}$, which has a chi square distribution with one degree of freedom if the model is true. Of course for the saturated binomial model, which does not restrict the probability of a boy or girl, the projection distance is zero. Table 2 lists the projection distances, together with the P-values, for the six centres. According to the usual criteria we would accept the restrictive model for all centres, except perhaps for Zagreb. If we use the distortion as a criterion (assuming that the model is true) we see from (13) that the restrictive model is better than the saturated model if the projection distance is less than two. Again this is the case for all centres, except

	chi	P-value
Melbourne	1.00	.317
Sao Paolo	0.61	.438
Santiago	1.00	.317
Alexandria	0.02	.888
Hong Kong	0.02	.903
Zagreb	3.52	.061

Table 2: binomial model tests

Zagreb. This means that in all centres we estimate the probability of a boy to be equal to .5, but in Zagreb we estimate it to be 70/134 = .5224. We see the damping effect of our use of models, which is basically the same as the shrinkage in empirical Bayes procedures, in ridge regression, or in Morris-Stein estimation.

Now we analyze the five models A-E for twin-pairs in the classical way. Consider model A, which is a one-dimensional quadratic manifold. The parametrisation in (1) is valid for all $0 \leq \omega \leq 1$. The maximum likelihood estimator of ω is, after some easy computation, given by $\underline{p}_1 + \underline{p}_3/2$. Thus $\phi_1(\underline{p}) = (\underline{p}_1 + \underline{p}_3/2)^2$, $\phi_2(\underline{p}) = (\underline{p}_2 + \underline{p}_3/2)^2$, and $\phi_3(\underline{p}) = 2(\underline{p}_1 + \underline{p}_3/2)(\underline{p}_2 + \underline{p}_3/2)$. For model D, which is a line segment, the maximum likelihood estimator of λ is $\underline{p}_1 + \underline{p}_2 - \underline{p}_3$. Thus $\phi_1(\underline{p}) = \phi_2(\underline{p}) = (1 + \underline{p}_1 + \underline{p}_2 - \underline{p}_3)/4$, and $\phi_3(\underline{p}) = (1 - \underline{p}_1 - \underline{p}_2 + \underline{p}_3)/2$. For model E the only F-consistent estimator is $\phi_1(\underline{p}) = \phi_2(\underline{p}) = .25$ and $\phi_3(\underline{p}) = .50$. Models B and C are somewhat problematical from the classical viewpoint. Model B says that there are no pairs of twins with different gender in the population. Thus they also cannot occur in any subset, and if only one occurs in the sample we reject B. No statistics is needed, only logic. Model C gives a maximum likelihood estimator equal to p if p is in the convex region defining the model, and equal to the estimate under model A if p is outside the region. For any π interior to the model the probability that $\underline{p}$ is interior tends to one, and thus the projection distance tends to zero with probability one. If π is on the model A boundary, then the projection distance has the mixture-distribution $\text{prob}(\underline{p} \notin C)\chi_1^2 + \text{prob}(\underline{p} \in C)\chi_0^2 = \text{prob}(\underline{p} \notin C)\chi_1^2$.

In Table 3a we have listed N times the projection distances for the models A, D, and E. The corresponding quantitites for model B are all 'infinite', those of model C are all zero. According to classical statistical theory the projection distances have chi square distributions with one (models A and D) or two (model E) degrees of freedom. Table 3b uses these asymptotic distributions to convert the chi square values to probabilities. This puts them on a convenient scale, and makes them comparable. The estimate of the distance between the observed and the expected value under the (true) model is thus corrected for the dimensionality of the model. Table 3 shows clearly that according to the classical analysis model D is the best one. It also shows that on the probability scale model E is better than A. One can interprete this according to the classical, although mysterious, syllogism quoted above. One can also think in terms of distance estimates between models and data, transformed to a convenient scale. But an interpretation of this scale in our replication framework is possible only if the model is true.

	model A	model D	model E
Melbourne	10.40	0.76	11.40
Sao Paulo	11.37	0.49	11.98
Santiago	33.39	0.73	34.39
Alexandria	12.24	0.02	12.26
Hong Kong	26.97	0.01	26.99
Zagreb	5.24	2.79	8.77

Table 3a: chi squares

	model A	model D	model E
Melbourne	.001	.383	.003
Sao Paulo	.001	.484	.003
Santiago	.001	.393	.001
Alexandria	:001	.888	.002
Hong Kong	.001	.920	.001
Zagreb	.022	.095	.012

Table 3b: probability transform

Another comparison is based on Akaike's AIC. As we have seen this amounts to adding two to the chi squares of models A and D, and zero to those of model E. The resulting quantity estimates the distortion, but again only if the model is true. Clearly this does not change the conclusions a great deal. Model D is still the best, by far, and on the AIC scale E is also slightly better than A, except for Zagreb.

Model evaluation using separator statistics

We now continue with our evaluation of models, but it is no longer assumed that the model, or one of the models, is true. Within classical statistics there has been at least one major development which applies in this more general situation. This the theory of *Pitman powers*. In this theory one does not assume that π ϵ Ω. The assumption is that there is a sequence Ω_N of models, giving rise to a sequence Φ_N of projections. We assume that the approximation errors $\Delta(\pi,\Phi_N(\pi))$ tend to zero in such a way that $N\Delta(\pi,\Phi_N(\pi))$ tends to a constant Δ_0. More precisely the assumptions guarantee that $N\Delta(\underline{p},\Phi_N(\underline{p}))$ converges in law to a noncentral chi square with m - r - 1 degrees of freedom, and noncentrality parameter Δ_0. Thus $NE\{\Delta(\underline{p},\Phi_N(\underline{p}))\} = (m - r - 1) + \Delta_0 + o(1)$. Verbeek (1984) has suggested to use the *noncentrality*, given by $\{N\Delta(\underline{p},\Phi_N(\underline{p})) - (m - r - 1)\}/N$, as an index of model fit. It is convenient to think of the noncentrality as an estimate of the approximation error $\Delta(\pi,\Phi_N(\pi))$ for models which are not too false. Nevertheless the use of Pitman powers in actual practical work seems a bit difficult. Who actually works with a sequence of models ?

A more direct approach is to go back to the formulas (5) to (9). In the case of the maximum likelihood separator (14) we find that $1/2\ \mathrm{tr}\ AV = (m - 1)$ and $\mathrm{tr}\ C'GV = -2\ \mathrm{tr}\ H'BH(H'BH + \nabla)^{-1}$. Thus we can estimate the bias by

$$\wp(\underline{p}) = \{N\Delta(\underline{p},\Phi(\underline{p})) + 2\ \mathrm{tr}\ (H'BH + \nabla)^{-1}H'BH\ - (m - 1)\}/N, \tag{15a}$$

and the distortion by

$$\vartheta(\underline{p}) = \{N\Delta(\underline{p},\Phi(\underline{p})) + 2\ \mathrm{tr}\ (H'BH + \nabla)^{-1}H'BH\}/N. \tag{15b}$$

In (15) $B = 2PQ^{-2}$ is evaluated in (p,q) equal to $(\underline{p},\Phi(\underline{p}))$. The basis H is evaluated in $\underline{\theta}$, where $\Psi(\underline{\theta}) = \Phi(\underline{p})$. Matrix ∇ is evaluated at the same point. If $\nabla = 0$, which happens for instance if Ω is an affine manifold, then $\mathrm{tr}\ (H'BH + \nabla)^{-1}H'BH = r$, and (15) simplifies accordingly. This gives a justification for using the AIC, even if the model is not true.

We can use formula (5) to estimate the approximation error. Using the simplifications resulting from the use of the maximum likelihood method we find

$$\begin{aligned}\aleph(\underline{p}) = \{&N\Delta(\underline{p},\Phi(\underline{p})) - (m - 1) + 2\ \mathrm{tr}\ (H'BH + \nabla)^{-1}H'BH\ - \\ &- \mathrm{tr}\ (H'BH + \nabla)^{-1}H'BH(H'BH + \nabla)^{-1}H'BH -\ 1/2\ \mathrm{tr}\ \Gamma\nabla\}/N\end{aligned} \tag{16}$$

This satisfies $E(\aleph(\underline{p})) = \Delta(\pi,\Phi(\pi)) + o(N^{-1})$. It is related, but not identical, to the noncentrality proposed by Verbeek, and it becomes identical to it in the case that ∇ and Γ are zero.

Let us now apply formulas (15) and (16) to the twin example. For the binomial model (and for saturated and zero-dimensional models in general) we find that indeed ∇ and Γ are zero. Thus $\wp(\underline{p}) = \{N\Delta(\underline{p},\Phi(\underline{p}))\ - (m - 2r - 1)\}/N$, and $\vartheta(\underline{p}) = \{N\Delta(\underline{p},\Phi(\underline{p})) + 2r\}/N$. Moreover $\aleph(\underline{p}) = \{N\Delta(\underline{p},\Phi(\underline{p})) - (m - r - 1)\}/N$. For models D and E we also have ∇ and Γ equal to zero, and the same formulas apply. For model A the situation is more complicated. Table 4a has the relevant information, but the computations needed to arrive at the numbers in the table, especially those for estimating the approximation error, are not simple. They will be even more complicated for models with many more cells than the three in our example. Table 4a shows that the bias and distortion are larger for model A than for the saturated model. In Table 4b we show the comparable statistics for (zero-dimensional) model E. Clearly E is preferable to A, even in terms of estimated approximation error (although of course the true approximation error for A is lower).

	bias	prediction	approximation
Melbourne	.113	.133	.099
Sao Paulo	.056	.066	.050
Santiago	.142	.150	.136
Alexandria	.032	.037	.029
Hong Kong	.223	.239	.211
Zagreb	.070	.094	.055

Table 4a: separator statistics model A.

	bias	prediction	approximation
Melbourne	.086	.106	.086
Sao Paulo	.044	.054	.044
Santiago	.130	.139	.130
Alexandria	.026	.031	.026
Hong Kong	.200	.216	.200
Zagreb	.040	.064	.040

Table 4b: separator statistics model E.

Use of Jackknife-type methods

The separator statistics used in the previous section are fairly difficult to compute, even if we use the simplifications that result from using the maximum likelihood method. For complicated models evaluating the second derivatives may be a painful process. A possible way out is to use resampling methods such as the Bootstrap and the Jackknife. Compare Efron (1982) for a nice review of these methods. In this paper we do not go into the philosophy of resampling, we merely use the methods as computational tools that use finite difference approximations to the derivatives, and that consequently can be used to approximate the separator statistics. We also restrict our attention to Jackknife (i.e. leave-one-out) methods. These are computationally far less demanding than the Bootstrap methods, and in the cases in which we have compared the two they give virtually identical results.

Define $q_{[j]} = p + (N-1)^{-1}(p - e_j)$, with e_j the j^{th} unit vector. Then

$$\Delta(q_{[j]},\Phi(p)) = \Delta(p,\Phi(p)) + (N-1)^{-1}\lambda'(p - e_j) + 1/2\,(N-1)^{-2}\,(p - e_j)'A(p - e_j) + o((N-1)^{-2}), \qquad (17)$$

and thus

$$\Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(p)) = \Delta(p,\Phi(p)) + 1/2\ (N-1)^{-2}\,\mathrm{tr}\ AV + o((N-1)^{-2}), \tag{18}$$

or

$$(N-1)^2\{\Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(p)) - \Delta(p,\Phi(p))\} = 1/2\ \mathrm{tr}\ AV + o(1). \tag{19a}$$

In the same way

$$(N-1)^2\{\Sigma_{j=1}^{m} p_j\Delta(p,\Phi(q_{[j]})) - \Delta(p,\Phi(p))\} = 1/2\ \mathrm{tr}\ (\Gamma + G'BG)V + o(1), \tag{19b}$$

and

$$(N-1)^2\{\Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(q_{[j]})) - \Delta(p,\Phi(p))\} =$$

$$= 1/2\ \mathrm{tr}\ (A + \Gamma + G'BG + G'C + C'G)V + o(1). \tag{19c}$$

It is clear that, under suitable regularity assumptions, we can use the quantities on the left of (19) to estimate the quantities on the right. If we combine (19) with (5), we see that estimating the approximation error, the bias and the distortion becomes quite simple. We first show this for the approximation error. Combining (5b) and (19c) shows that

$$\mathbf{E}\{N\Delta(\underline{p},\Phi(\underline{p})) - (N-1)\Sigma_{j=1}^{m} \underline{p}_j\Delta(\underline{q}_{[j]},\Phi(\underline{q}_{[j]}))\} = \Delta(\pi,\Phi(\pi)) + o((N-1)^{-1}). \tag{20}$$

This is the classical multinomial Jackknife result, but it is remarkable (and very satisfactory) that it can be generalized easily to deal with bias and distortion. Indeed, from (19b) and (19c),

$$(N-1)^2\{\Sigma_{j=1}^{m} p_j\Delta(p,\Phi(q_{[j]})) - \Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(q_{[j]}))\} = -1/2\ \mathrm{tr}\ (A + G'C + C'G)V + o(1), \tag{21}$$

and thus

$$\mathbf{E}\{\Delta(\underline{p},\Phi(\underline{p})) + (N-1)\{\Sigma_{j=1}^{m} \underline{p}_j\Delta(\underline{p},\Phi(\underline{q}_{[j]})) - \Sigma_{j=1}^{m} \underline{p}_j\Delta(\underline{q}_{[j]},\Phi(\underline{q}_{[j]}))\}\} =$$

$$= \mathbf{E}\{\Delta(\pi,\Phi(\underline{p}))\} + o((N-1)^{-1}). \tag{22}$$

In the same way, from (19a) and (21),

$$(N-1)^2\{\Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(p)) + \Sigma_{j=1}^{m} p_j\Delta(p,\Phi(q_{[j]})) - \Delta(p,\Phi(p)) - \Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(q_{[j]}))\}$$

$$= - \mathrm{tr}\, G'CV + o(1), \tag{23}$$

and thus

$$\mathbf{E}\{N\Delta(\underline{p},\Phi(\underline{p})) + (N-1)\{\Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(p)) + \Sigma_{j=1}^{m} p_j\Delta(p,\Phi(q_{[j]})) - \Sigma_{j=1}^{m} p_j\Delta(q_{[j]},\Phi(q_{[j]}))\}\} =$$

$$= \mathbf{E}\{\Delta(\underline{q},\Phi(\underline{p}))\} + o((N-1)^{-1}). \tag{24}$$

These formulae may look complicated, but they are actually quite easy to use. The computational burden is that we must evaluate our estimate Φ not once, but m + 1 times, with m the number of cells in the multinomial experiment. This could be a large amount of work, especially in very large multiway tables, but in such very large tables using asymptotic theory is of doubtful value anyway. Observe that these formulas are perfectly general, in the sense that they use no properties of the likelihood separator and of the maximum likelihood estimate. They also do not assume, of course, that the model is true.

We shall illustrate our formulae by applying (20) to our example, in particular to models A, D, and E. Table 5 lists the relevant results. Table 5a actually presents the Jackknife result derived from (20), and

	model A	model D	model E
Melbourne	.099	-.003	.096
Sao Paulo	.050	-.002	.047
Santiago	.136	-.001	.134
Alexandria	.029	-.003	.026
Hong Kong	.211	-.008	.192
Zagreb	.055	.022	.082

Table 5a: Estimated approximation errors
using the Jackknife

	model A	model D	model E
Melbourne	.099	-.003	.096
Sao Paulo	.050	-.002	.047
Santiago	.136	-.001	.134
Alexandria	.029	-.002	.026
Hong Kong	.211	-.008	.200
Zagreb	.055	.022	.082

Table 5b: Estimated approximation errors
using the Bootstrap

Table 5b gives the corresponding Bootstrap estimates (from results not presented here). It is clear that the results are very close indeed, and because the Bootstrap is computationally so much more demanding it does not seem to be a wise choice in this context. We can also compare the results for models A and E in Table 5 with the last columns of Tables 4a and 4b. The correspondence is very satisfactory.

Conclusion

It is clear from our discussion above that the expansions we have derived are valuable tools to answer the simple question whether the model is useful or not. But the meaning of 'useful' must be specified in at last two ways before we can actually answer the question. In the first place a model can be useful for bias reduction and for prediction. As we have seen above these two forms of usefulness tend to be contradictory, in the sense that 'large' models are good for bias reduction and bad for prediction, while for 'small' models it is the other way around. The solution of classical statistics is to choose from the class of true models only. This approach does not make sense to us, although it is possible to interpret the classical procedures in our somewhat wider framework. But in this wider framework the classical procedures may not give good estimates of bias and expected prediction error.

A second specification we have to choose before we can answer the question about the usefulness of a model is the choice of a separator. The one suggested by the method of maximum likelihood is convenient, the class of methods suggested by BAN-theory is also attractive, but essentially the choice is open. Small bias in terms of one separator does not necessarily imply small bias in terms of another one. It is quite possible that results can be derived, in our general framework, which show that some separators are asymptotically or uniformly preferable to others (think of second order efficiency), but we have no results in this direction.

The question if a model is useful or not can be coupled with the question of computing a good estimate of the unknown probabilities in the multinomial model. If we decide that a particular model is preferable to the saturated model, then we can replace the sample proportion by the estimates under the model restrictions. We have seen, in our simple example, that considering a number of models works in the same way as a prior distribution, and produces shrinkage estimators very much like the empirical Bayes estimators familiar from other statistical work. No explicit prior distribution is considered, but a discrete class of models must be chosen for consideration, and of course this choice is to some extent 'subjective'. It is a useful subject of study to find out if these shrunken estimators are in some sense better than the usual estimators, although the method with which they are constructed already seems to answer this question in the affirmative.

We think that the primary contribution of this paper is, or should be, that it tries to make people more careful about assuming models to be true (whatever that means), about using standard statistical reasoning (as in hypothesis testing), and about using probability models without clearly specifying the framework of replication. Whether specific models are actually false or not is not interesting. All models are false. The question is how false, and whether their being false actually implies that they are not useful. And whether the resulting statements are consequently not interesting. The independence assumption, for example, which is at the basis of most work in statistics, cannot really be falsified. As we have seen, the independence assumption merely corresponds with a particular framework of replication, for which we have to decide whether it is relevant or not. We make statements about all samples with replacement from the population of the Netherlands, for instance. If such statements are really relevant for the particular policy issue we are studying is something which certainly must be decided with a great deal of care. For the more complicated frameworks, such as those with continuous variables and non-identically distributed observations, it may be most difficult to convince a sceptical user that our statements are indeed relevant for his problem.

Another more specific contribution of the paper is that it shows that looking at the projection distance is only a first crude apprximation. The interesting statistics, bias and distortion, can easily be approximated more precisely, both by using additional terms in the expansions and by using resampling methods. In particular we have argued that the distortion, which is the expected value of the prediction error, is at least as interesting as the bias, the expected value of the estimation error. In general the two statistics lead to a different ordering of models, and consequently also to different estimates of π. We have tried to justify the AIC statistic, used to estimate distortion, and the deviance statistic, used to estimate approximation error used by Verbeek, without assuming the model to be even approximately true.

References

Abatzoglou,T, The Metric Projection on C^2 Manifolds in Banach Spaces. **Journal of Approximation Theory**, 26, 1979, 204-211.

Akaike, H., On Entropy Maximization Principle. In P.R. Krishnaiah (ed.), **Applications of Statistics**, Amsterdam, North Holland Publishing Company, 1977

Andersen. E.B., **Discrete Statistical Models with Social Science Applications.** Amsterdam, North Holland Publishing Co, 1980.

Bemis, K.G. & Bhapkar, V.P., On BAN Estimators for Chi Squared Test Criteria, **The Annals of Statistics**, 11, 1983, 183-196.

Benzécri, J.P. et al., **L'Analyses des Données.** Paris, Dunod, 1973.

Berkson, J., Minimum Chi-Square, not Maximum Likelihood (with Discussion). **The Annals of Statistics**, 8, 1980, 457-487.

Box, G.E.P., Some problems of Statistics and Everyday Life. **Journal of the American Statistical Association**, 74, 1979, 1-4.

De Leeuw, J., Models for Data. **Kwantitatieve Methoden**, 5, 1984, 17-30.

Efron, B. **The Bootstrap, the Jackknife, and other Resampling Plans.** Philadelphia, SIAM, 1982.

Gifi, A., **Nonlinear Multivariate Analysis.** Leiden, Department of Data Theory FSW/RUL, 1981

Guttman, L., The Illogic of Statistical Inference for Cumulative Science. **Applied Stochastic Models and Data Analysis**, 1, 1985, 3-10.

Hemelrijk, J., Back to the Laplace definition. **Statistica Neerlandica**, 22, 1968, 13-21.

Kalman, R.E., Identifiability and Modeling in Econometrics. In P.R. Krishnaiah (ed.), **Developemnts in Statistics**, Amsterdam, North Holland Publishing Company, 1983.

Le Cam, L., **Asymptotic Methods in Statistical Decision Theory.** Berlin, Springer, 1986.

McCullagh , P. & Nelder, J.A., **Generalized Linear Models.** London, Chapman and Hall, 1983.

Nelder, J.A., The Role of Models in Official Statistics. **Eurostat News**, 1984, special number on Recent Developments in the Analysis of Large-Scale Data Sets.

Neyman, J., Contributions to the Theory of the χ^2 Test. **Proceedings Berkeley Symposium**, 1, 1949, 239-273.

Rao, C.R., **Linear Statistical Inference and its Applications**, New York, Wiley, 1973.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G., **Akaike Information Criterion Statistics**. Dordrecht, Reidel, 1986.

Taylor, W.F., Distance Functions and Regular Best Asymptotically Normal Estimates. **Annals of Mathematical Statistics**, 24, 1953, 85-92.

Tukey, J.W., We need both Exploratory and Confirmatory. **American Statistician**, 34, 1980, 23-25.

Verbeek, A. The Geometry of Model Selection in Regression. In T.K. Dijkstra (ed.), **Misspecification Analysis**, Berlin, Springer, 1984.

Willems, J.C., **From Time Series to Linear System**, Mathematical Institute, University of Groningen, 1986

Wolfe, J.M., Differentiability of Nonlinear Best Approximation Operators in a Real Inner Product Space. **Journal of Approximation Theory**, 16, 1976, 341-346.

Wijsman, R.A., On the Theory of B.A.N. estimates. **Annals of Mathematical Statistics**, 30, 1959, 185-191. Correction idem, 1268-1270.