International Statistical Review, (1986), **54**, 1, pp. 1–13. Printed in Great Britain

Conditioned Data-Based Simulations: Some Examples from Geometrical Statistics

Alastair Young

Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge, CB2 1SB, UK

Summary

We review principles behind the use of 'data-based' or 'bootstrap' methods of testing statistical hypotheses. Distinction is drawn between the motivation for such procedures, by which representatives of the hypothesis under test are simulated by random corruption of the sample data, and the estimation bootstrap of Efron (1979). A series of geometrical examples are considered and attention is drawn to the importance in data-based inference of preserving ancillary features of the data under test.

Key words: Ancillary; Bootstrap; Conditionality; Data-based simulation; Geometrical hypothesis; Random data corruption.

1 Introduction

The aim of this paper is to review widely dispersed ideas on the use of 'data-based simulation' or 'random data corruption' methods of testing statistical hypotheses. Examples of such techniques have been considered previously by, for example, Kendall (1977), Besag & Diggle (1977) and Lotwick & Silverman (1982). We present our discussion by considering three geometrical data-analytic problems, two of them entirely new to the statistical literature. The solutions we obtain to these problems underline the importance of appropriate conditioning in this type of data analysis. In particular, we stress the concept of an 'ancillary data feature', as first discussed by Kendall & Kendall (1980). It will be argued that general theoretical development of a conditionality principle in the data-analytic context may not be possible, because of the difficulty of defining, in the absence of a fully parameterized mathematical model, what is meant by an ancillary statistic or feature.

2 Data-based simulations and the bootstrap

The bootstrap is probably most familiar as a nonparametric technique, due to Efron (1979, 1982), which is used primarily to estimate the sampling distribution of some statistic or for construction of confidence intervals for some parameter of interest. 'Bootstrap methods', however, predate Efron's work and the term is generally taken to refer to any inference which derives the reference comparison sets from the original sample data. Indeed the bootstrap approach to data analysis is often taken to encompass any statistical technique where the sample data plays an interventionalist role in determining the form of the analysis to be applied to that same data. Most statisticians, for instance, would consider cross-validation studies (Stone, 1974) and shrinkage estimation (Copas, 1983) as bootstrap methods.

In the light of such terminology, and in view of the simulations and random resampling which form their basis, the techniques discussed in this paper are certainly bootstrap

techniques. However we prefer the language of 'data-based simulations' and 'random corruption' to emphasize that ours are allied techniques, rather than methods derived directly from the Efron bootstrap.

Efron (1979) essentially defines a process by which a density estimate is constructed from a set of sample data and reference data sets are then simulated from this estimate. The bootstrap is strictly limited as a technique with which to test statistical hypotheses, because of the unconditional nature of most bootstrap simulations. In the literature, the only testing method which derives directly from the Efron bootstrap is the smoothed bootstrap procedure used by Silverman (1981) to investigate the number of modes in a population density.

The aim of Efron's methodology is the definition of *automatic* procedures which will have improved estimation characteristics over those based on standard asymptotic normality assumptions, etc. Our motivation is somewhat different. Our methodology is tailored to *specific* data sets, and our discussion will continually focus on the particular characteristics of these data sets. In the use of the Efron bootstrap it is to be observed that not every sample will yield a dependable population estimate, precisely because the density estimate constructed from the specific sample data, and used to generate reference sets, may be a poor estimate of the underlying population distribution. Within the context of testing statistical hypotheses this latter point may, in the sense of some extended conditionality principle, be put to positive use. It is argued that there is a virtue, in terms of relevance of the inference to the actual data under test, in simulating reference data sets either from a density estimate 'close' to the sample data or directly from the data, rather than from some null mathematical model.

3 Scope of data-based inference

The data-based simulation approach is an attractive one. The direct nature of the techniques, ease of interpretation of results, etc., should not, however, be considered sufficient justification for their use. The approach will not extend to all types of statistical problem. Instead, the type of procedure which we envisage may serve as something of a bridge between the exploratory and confirmatory aspects of data analysis, allowing as they do *largely* model-free inference to be made. Data-based simulation is perhaps most strongly justified in circumstances, such as the asymmetry problem discussed below, where any mathematical model of a data-generating mechanism which may underlie our analysis has no physical basis in itself, but refers instead to some projective geometrical phenomenon. It is crucial to note, however, that the test statistic which we use will often be based on a parametric model, though perhaps one suggested by the sample data itself. Then it is only implementation of the test procedure which is direct, this being based on a custom-built random representation of the null hypothesis, which takes account of uncertainties in the mathematical model.

Indeed, within the general class of problems which lend themselves to the data-based approach, we can identify three types of situation.

(i) In some statistical investigations, prior knowledge of the substantive field will yield information about the form of the phenomenon to be investigated, but will enable us to say nothing about the data-generating mechanism itself. The available sample data alone must then serve at any modelling phase. Simulation is used in the calculation of an inferential probability, because of doubts over the validity of parametric models.

(ii) In other circumstances, knowledge of the scientific background to the problem,

and the phenomenon being studied, will yield a sensible, though perhaps heuristic, form of test statistic without formal modelling. The simulation approach is used precisely because there is no parametric framework.

(iii) A third class of problem, rather different to the first two, arises where the scientific background provides all the modelling input to the analysis, but where, for some reason, it is necessary to obtain the inferential probability by simulation means.

In § 5 we consider examples of each of the above type of problem. Though structurally different, the same spirit of approach underlies all three examples and this is discussed in § 4.

4 Inferential procedure

Each of the problems considered in § 5 may be described in the following terms. We are given a set of data and we wish to test the null hypothesis that this data contains no evidence for predisposition towards some geometrical phenomenon P, this to be tested against the alternative hypothesis that the data *does* predict a frequent occurrence of P.

Statistically, we suppose our observations $Y_1, \ldots, Y_n$ to be independent, identically distributed from some unknown density $g(y)$. The scientific investigation translates into a statistical problem of testing consistency of the data with $H_0: g \in \mathcal{F}_0$, where $\mathcal{F}_0$ is a class of probability distributions over some appropriate space.

Different problems will have different detailed specifications. The data-based approach, however, may be described simply as one which bases the calculation of an inferential probability upon the generation of data sets representing $\hat{F}_0$, where $\hat{F}_0$ is that distribution consistent with H_0 which is best supported by the data, and where the sample data plays a central role in the specification of the simulation mechanism.

The class $\mathcal{F}_0$ may be highly nonspecific. For example, in the second of our problems below $\mathcal{F}_0$ would define the class of diffuse bivariate probability densities. On the other hand, H_0 may be simple, as would be the case with, say, a test of uniformity.

Often initial examination of the data will suggest a parametric family of distributions $\mathcal{F}_0 = \{f(y; \theta);\ \theta \in \Theta\}$, which may be helpful in defining a test procedure. In such circumstances it would be usual to estimate the null distribution $\hat{F}_0$ closest to the data, by, say, maximum likelihood, and then define a test statistic T by

$$T = d(\hat{F}, \hat{F}_0),$$

where $\hat{F}$ denotes the empirical distribution of the sample data and d is some specified measure of distance. The function d might, for example, define a goodness-of-fit statistic or, if the modelling phase of the analysis defines a general class $\mathcal{F}_1$ to which the unknown density g belongs, d might define the likelihood ratio or score statistic.

Depending upon the precise form of such a parametric framework (existence of nuisance parameters, etc.) inferential principles will indicate a particular type of test procedure. It is important to appreciate, however, the philosophy with regard to such parametric models. The parametric family $\{f(y; \theta); \theta \in \Theta\}$ is not regarded as a true model for the data, but instead as a mathematical convenience. Often, as in our first example below, we will prefer a more ad hoc analysis than that indicated by standard parametric theory and the estimated $\hat{F}_0$.

Suppose that we have defined some suitable test statistic T, with large values of T being evidence against H_0. Let $\mathbf{x} = (Y_1, \ldots, Y_n)$ be the observed sample data and let $T(\mathbf{x}) = t_{\text{obs}}$. Then our simulation procedure is aimed at estimation of the significance

probability

$$P_{\text{obs}} = P(T(\mathbf{X}) \geqslant t_{\text{obs}} \mid \hat{F}_0),$$

where $\mid \hat{F}_0$ indicates a conditioning on $\hat{F}_0$.

For reasons outlined above, $\hat{F}_0$ will rarely have a formal specification. The aim, therefore, is to implement a significance test by construction of numerous conditionally independent replicates $\mathbf{X}_1, \ldots, \mathbf{X}_N$ of the data under test, the conditioning being with respect to relevant ancillary features of the data. These replicates are simulated from, or by 'random corruption' of, the observed data $\mathbf{x}$. The simple idea behind such a procedure is to generate from the sample data a simulation base of test statistic values which are taken to represent $\hat{F}_0$, that is the T-distribution that would have been obtained, in similar circumstances, if in fact the null hypothesis is true. The significance probability P_{obs} is estimated by

$$\hat{P}_{\text{obs}} = \#\{i : T(\mathbf{X}_i) \geqslant t_{\text{obs}}\}/N.$$

The primary requirement of the simulation scheme must be to produce data sets $\mathbf{X}_i$ free of any intended effect P of the type being tested against. Simplistically, if we do not corrupt the sample data enough our simulation mechanism will serve only to yield data sets quantitatively rather too similar in nature to the data under test. The aim therefore is to define, perhaps in the form of some conceptualized metric, a criterion based on the phenomenon P being studied, which allows us to specify when a simulated data set is sufficiently different from the observed data configuration to be considered free of any planned P. In effect we 'stratify' the space of all possible simulation samples. Only those simulation samples which are free of any intended effect *and* which are consistent with the data are to be used to represent H_0.

Data-based inferential procedures of the above type obtain their operational substance through frequency in a series of simulations. It is often argued that frequentist inferences are inadequate because of the need to choose, in order to obtain an inferential probability, a reference set for the sample prior to the data analysis itself. Hinkley (1983) argues that the general concept of ancillarity is integral to statistical inference and that proper observance of the conditionality principle can lead to sensible frequentist procedures. Any conditional inference will, of course, allow the data to choose a reference set for the sample. The random data corruption simulation method, however, goes somewhat further. If properly modulated, the simulation mechanism, which may, if we are working in a parametric framework, have been defined in such a way as to provide exact conditioning on an ancillary statistic, will allow at least approximate conditioning on features of the data set which may, on scientific grounds, be thought of as important in defining a relevant reference set.

5 Examples

Example 1. Our first example is considered at greater length by Kendall & Young (1984) and arose from a paper by Birch (1982). We are concerned in this problem with a set of astronomical objects known as classical-double radio-sources. With each such source is associated an angular measurement or 'indirection' Δ. By the term indirection we mean the ambiguous directional information contained in an undirected line or axis. Birch's contention was that of a topographic variation in the distribution of this measurement. We can think of Δ as a point in one-dimensional real projective space P^1, though it is more convenient to work with 2Δ as a point on the unit circle S^1. Our set of observations is then $((\mathbf{p}_i, 2\Delta_i) : i = 1, \ldots, n)$, where $n = 134$ and $\mathbf{p}_i = (p_i^{(1)}, p_i^{(2)}, p_i^{(3)})$ is

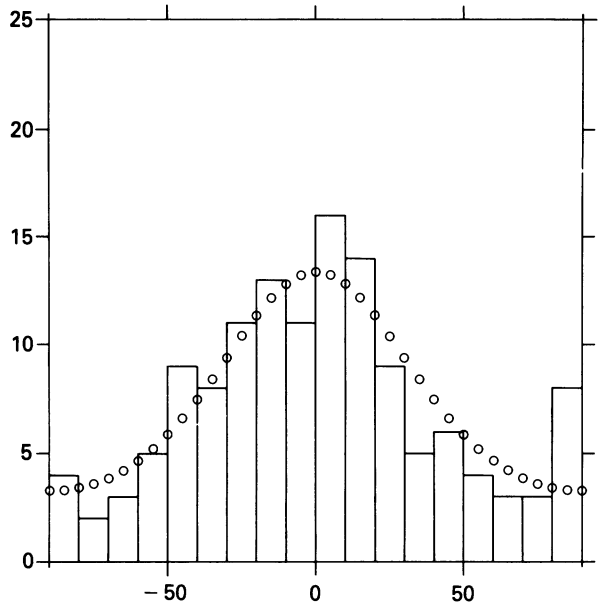

Figure 1. *Histogram of Δ-values, with fitted indirectional version of von Mises distribution shown by circles; estimated* $\alpha = 0{\cdot}703$, $n = 134$.

the unit position vector of the ith source on the celestial sphere S^2. In this problem our analysis must proceed only from the postulation of two probability measures for $(\mathbf{p}, 2\Delta)$, one corresponding to the absence of a topographic regression of 2Δ on $\mathbf{p}$ and the other to a positive Birch effect.

An adaptation of the familiar von Mises density,

$$\frac{\exp(\alpha \cos 2\Delta)}{\pi I_0(\alpha)} d\Delta \quad (-\tfrac{1}{2}\pi \leqslant \Delta \leqslant \tfrac{1}{2}\pi), \tag{1}$$

describes a simple nonuniform indirectional distribution symmetrical about $\Delta = 0$. Figure 1 shows a histogram of the overall ensemble of 134 Δ-values. Superimposed on this histogram is a fit of the indirectional version (1) of the von Mises distribution. The figure makes clear that the marginal Δ-distribution is roughly of this type and such a distribution is therefore taken as our null, position independent, model. Kendall & Young (1984) argued that, if the asymmetry is to be controlled by a topographic regression, the distribution of Δ conditional on the position of the radio-source will have density approximately of the form

$$\frac{\exp(\alpha \cos 2\Delta + (\boldsymbol{\lambda} \cdot \mathbf{p}) \sin 2\Delta)}{\pi I_0\{\sqrt{[\alpha^2 + (\boldsymbol{\lambda} \cdot \mathbf{p})^2]}\}} d\Delta. \tag{2}$$

Here, as in (1), I_0 is the zero-order modified Bessel function of the first kind, while $\mathbf{p}$ is the position of the source and $\boldsymbol{\lambda}$ a directional vector parameter whose modulus $\beta = |\boldsymbol{\lambda}|$ measures the strength of the Birch effect. In order to complete the mathematical formulation of the problem, we must incorporate into both our null and alternative models a factor $f(\mathbf{p})\, d\mathbf{p}$ to describe the distribution of source positions $\mathbf{p}$ on S^2. It is our contention, however, that it is wholly inappropriate to specify any simple form for f and that our statistical analysis should condition upon the observed assemblage of positions $\mathbf{p}_i$, for these are more descriptive of the positioning over the Earth of radio-observatories than of any astronomical phenomenon.

To test the null model (1) against the alternative model (2) we can construct the usual likelihood ratio statistic by writing down the log likelihoods L_0 and L_1 for the data with respect to the two models and maximizing each with respect to their relevant parameters. Our test statistic is then $T = \max L_1 - \max L_0$. Maximization of L_0 is trivial, while maximization on the alternative hypothesis is straightforward, though time-consuming, provided the sources do not all lie on a single great circle. If the sources do all lie on a single great circle the log likelihood L_1 may not have a unique maximum.

Wilks' Theorem tells us that the statistic T will have an asymptotic $\frac{1}{2}\chi^2_3$ distribution on the null hypothesis but it is not clear how close to the asymptote we will be in the present application, where we have 134 observations. It is convenient to make a first-order approximation to the modified Bessel function I_0: with this approximation the problem of global maximization linearizes, yielding an approximation T_1 to the test statistic T which is more readily calculated in a series of simulations. If A is the sum-of-squares and products matrix of the source positions $\mathbf{p}_i$,

$$a_{jk} = \sum p_i^{(j)} p_i^{(k)} \quad (j, k = 1, 2, 3)$$

and $B = A^{-1}$, so that B exists provided the sources do not all lie on a great circle, then the first-order statistic is given by

$$T_1 = \tfrac{1}{4}\mathbf{v}^{\mathrm{T}} B \mathbf{v}.$$

Here $\mathbf{v}$ is the vector with components $v_j = 2 \sum p_i^{(j)} \sin 2\Delta_i$ $(j = 1, 2, 3)$. For the sample data we find $T_1 = 7{\cdot}32$.

To represent the null hypothesis of no intended topographic variation we sample from the permutation distribution of the data by uniformly randomly scrambling the Δ_i amongst the source positions $\mathbf{p}_i$. In this way data sets are produced which preserve the source positions and the overall assemblage of observed Δ_i; see Fig. 1.

Figure 2 is a histogram of the T_1-values for a series of 10,000 such simulations. It is left

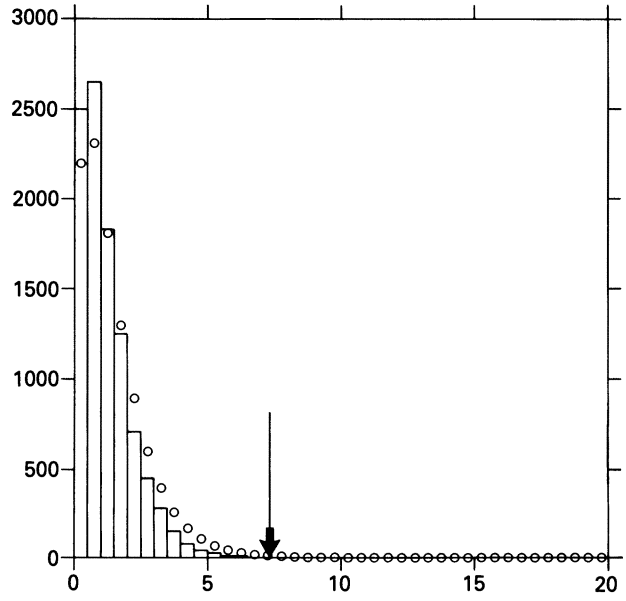

Figure 2. *Histogram of T_1-values for* 10,000 *permutations of observed data configuration. The T_1-value for the observed data is shown by an arrow and equals* 7·320. *Asymptotic distribution shown by circles. First-order solution, $n = 134$.*

to the reader to decide whether this follows the asymptotic $\frac{1}{2}\chi_3^2$ curve; our simulation test yields a $\hat{P}_{\text{obs}}$-value of the order 0·0005.

Notice that our interpretation of the test statistic value is by a standard permutation test. This procedure, and the use of a statistic T_1 which is only a very crude approximation to the log likelihood statistic, is entirely deliberate and is made in view of the unverifiable nature of our statistical models.

Of course, if we really believed the parametric model (2) a refined test procedure would be appropriate. The obvious approximation to the log likelihood statistic to use in a series of simulations, since it only requires maximization on H_0, is the score statistic (Cox & Hinkley, 1974, pp. 323–324). In the current problem this is defined by

$$S = \hat{\alpha}_0 \frac{I_0(\hat{\alpha}_0)}{I_1(\hat{\alpha}_0)} T_1,$$

where $\hat{\alpha}_0$ is the maximum likelihood estimator under $H_0: \beta = 0$ and I_1 is the first-order modified Bessel function. The data value of this score statistic is 15·53. Now, of course, S, rather than $2S$, is to be compared with χ_3^2. Notice also that, for the randomization procedure we have used in our simulations, use of T_1 is equivalent to the use of S. In addition, a simulation analysis which makes full use of the parametric model would use direct simulation from the null model, rather than data permutation. It is not clear, however, if it is possible to condition simulations from the von Mises density (1) on the sufficient statistic, $\sum \cos 2\Delta_i$, for the nuisance parameter α. A fully parametric analysis would, in any case, calculate the Bartlett correction factor to the log likelihood statistic to take account of the finite sample size, rather than use simulations.

Example 2. Our second example is already very familiar, at least in the archaeological context (Kendall & Kendall, 1980), though it certainly remains a very illuminating and instructive problem. Astronomical interest in the problem was kindled by Arp & Hazard (1980), who drew attention to a curious geometrical configuration of quasars, or QSOs, involving near collinearity of triplets of such objects.

In effect we are presented with a set of points $Y_1, Y_2, \ldots, Y_n$ in the plane. We say that three such points Y_i, Y_j and Y_k are 'ε-collinear' if they form a triangle with largest angle $> \pi - \varepsilon$. Of the ${}^nC_3$ triangles formed from the n points, some number, N_{COLL} say, will be ε-collinear. Intuitively, N_{COLL} will serve as a test statistic for a test of the null hypothesis that the observed collinearities are due to chance against the alternative hypothesis that there are too many ε-collinearities for these all to be accounted for by chance, with large values of N_{COLL} rejecting H_0.

Notice that we deliberately skirt here the problem of the value of the tolerance angle ε, by assuming it to be known. This may indeed be the case. Otherwise, if ε is not given, a more sophisticated analysis is necessary. For details see Kendall & Kendall (1980). At any rate a point to be stressed is that, in contrast to our first example, the nature of the phenomenon under study is such that it gives rise to a natural test statistic without formal modelling. The difficult part of the analysis lies in the specification of the mechanism by which we obtain our data corruptions or 'random lateral perturbations'. It is possible to define a measure of similarity between the observed data configuration $\mathbf{x}$ and a perturbed data set $\mathbf{X}$ in terms of the number of ε-collinearities the two configurations have *in common*. Random lateral perturbations ought to destroy all, or nearly all, of the existing near collinearities. There is clearly no unique way in which this can be achieved, though some corruption schemes may seem more appealing than others, in the way in which they preserve the coarse data-structure.

The most obvious corruption scheme is one in which each data point Y_i is perturbed

independently by a random quantity (ξ, η) drawn from a bivariate Gaussian distribution with density of the form

$$\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2\sigma_1^2}\xi^2-\frac{1}{2\sigma_2^2}\eta^2\right)d\xi\,d\eta. \tag{3}$$

For many practical purposes it may be suitable to take $\sigma_1 = \sigma_2$, the common value being set in order to destroy the existing collinearities.

Figure 3(a) shows a configuration of 74 points over the unit square. This data is the

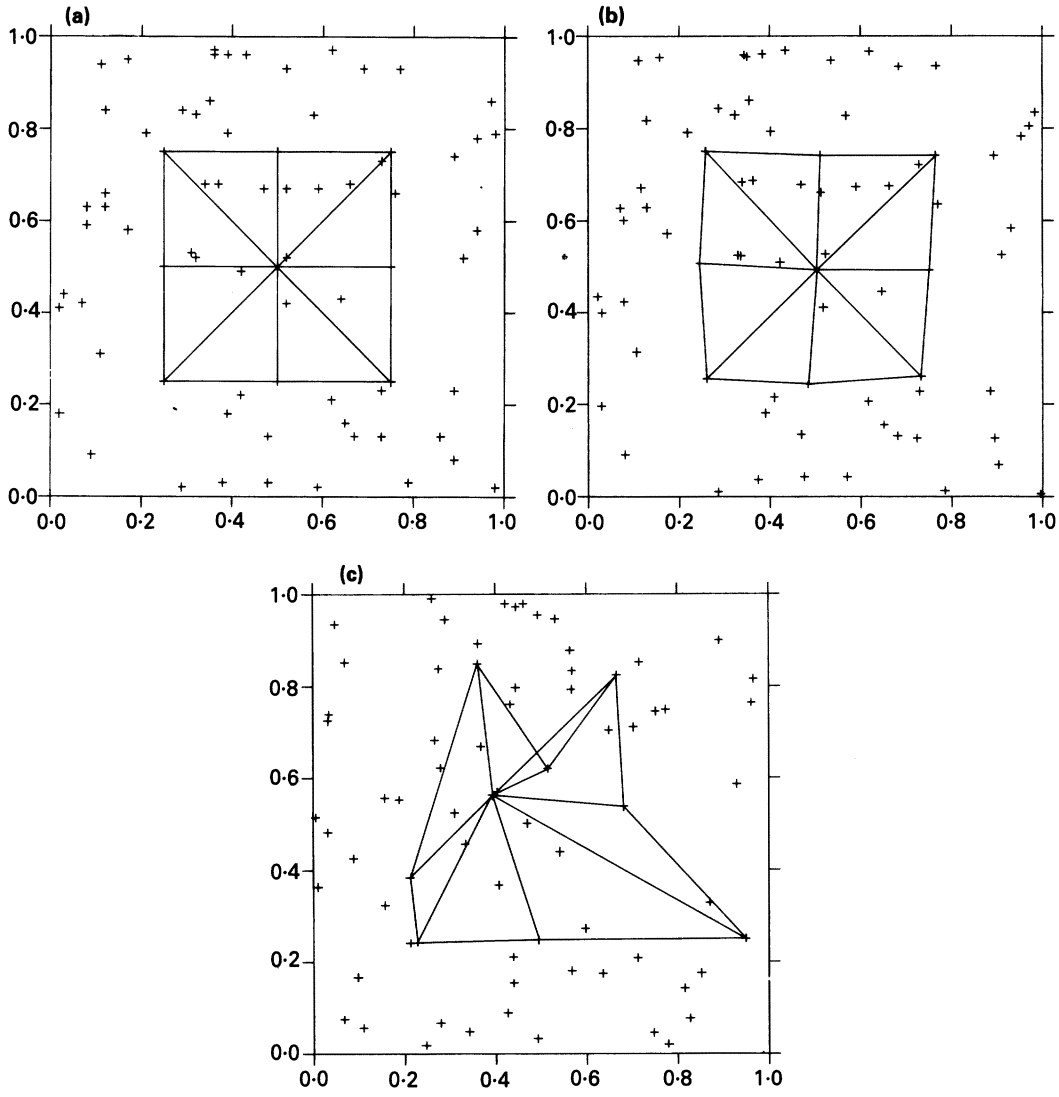

Figure 3(*a*). *Locations, on unit square, of* 65 *Japanese pine saplings* (*Diggle,* 1983), *with* 9 *artificial points on regular lattice.*

Figure 3(*b*). *Configuration obtained from Fig.* 3(a) *by independent Gaussian perturbation of each point, with* $\sigma_1 = \sigma_2 = 0{\cdot}01$. *Unit square wrapped onto torus.*

Figure 3(*c*). *As Fig.* 3(b), *with* $\sigma_1 = \sigma_2 = 0{\cdot}1$.

Japanese pine sapling data listed by Diggle (1983, p. 128), into which have been placed 9 points on a regular lattice. The artificial grid provides 8 exact three-point collinearities, as shown. In total there are some 389 1^0-collinearities in the overall configuration of 74 points. Figure 3(b) shows the same configuration after each data point has been perturbed according to the density (3) with $\sigma_1 = \sigma_2 = 0{\cdot}01$. Of the 8 planned alignments only 1 remains a 1^0-collinearity. In addition, the overall pattern remains, in its coarse structure, qualitatively the same as the original data. Figure 3(b), therefore, is a suitable data set by which to represent the null hypothesis of no planned alignments. By contrast, Fig. 3(c) illustrates what happens with $\sigma_1 = \sigma_2 = 0{\cdot}1$. The perturbations have now produced a data set which is not only geometrically, but also topologically, different from the original data: the clusters discernible in Fig. 3(a) have been destroyed. The data has been over-perturbed.

It is of interest to note that Figs. 3(b) and (c) contain 361 and 387 1^0-collinearities respectively. Comparison of these figures, and values obtained for other perturbed data sets, with the original number 389 provides a warning about this type of data-specific procedure. Had we not known about the presence of planned alignments, it is possible that our method would not have been able to detect them. The signal provided by the artificial construct may be swamped by noise from the rest of the data. In a sense, therefore, the test scheme we have described has little power against alternatives of the type provided by the 8 'faked' alignments.

In the above illustrative example the data points fall in a well-defined region. This was taken account of in the simulations by treating the unit square as a torus. In other examples, the points will fall in some undefined area. Then the region covered by the data points in a corrupted data set ought to display a similar degree of elongation to that for the actual data. This can be effected by using a Gaussian scheme of the type (3), but with component standard deviations proportional to those for the observed data set, as was done by Kendall & Kendall (1980). In this way corrupted data sets will approximately preserve the ratio of principal standard deviations for the actual data. If it is desired, after corruption the whole configuration can then be rescaled in order to preserve the value of this ancillary statistic exactly.

Example 3. Our third example arises out of a study into the orientation of spiral galaxies within a cluster of galaxies, which includes our own, known as the 'Local Supercluster', and relates to the question of galaxy cluster formation.

For our purposes spiral galaxies may be viewed as flat, circular discs. A quantity known as the 'angular momentum' or 'spin' vector of the galaxy points in a direction normal to the plane of the disc, the direction of rotation defining the sense of this normal by, say, a right-hand rule. The direction of the spin vector describes a random quantity, $\boldsymbol{\sigma}$ say, with distribution on S^2.

Under the gravitational instability model for galaxy cluster formation orientations of galaxies are completely random, so that $\boldsymbol{\sigma}$ is uniformly distributed on S^2. The cosmological turbulence theory suggests that galaxies become preferentially aligned with their discs parallel to the plane of the cluster itself, which we can identify with the equator, $Z = 0$ say, of the celestial sphere S^2. In these circumstances $\boldsymbol{\sigma}$ will have a bipolar distribution over S^2, with modes at $Z = +1$ and $Z = -1$. The theory of adiabatic fluctuations suggests that galaxies become aligned perpendicular to the major plane of the cluster, in which case $\boldsymbol{\sigma}$ has an equatorial girdle distribution over S^2.

Our problem is set apart from conventional problems in directional statistics by the projective nature of the observations. Given only the projected image of a galaxy, it is nevertheless possible to attach to the galaxy two undirected normals and to say that the

true spin vector points in some direction along one of these two indirections. Since spiral galaxies are known to rotate with their arms trailing, observation of the spiral winding direction of a galaxy (whether the galaxy appears as an S-spiral or a Z-spiral) enables us to reduce this four-fold ambiguity in spin direction to a two-fold ambiguity. We are presented, therefore, with a set of 'data' $\{(\mathbf{p}_i, \boldsymbol{\sigma}_i^1, \boldsymbol{\sigma}_i^2),\ i = 1, \ldots, n\}$ on n galaxies. Here $\mathbf{p}_i \in S^2$ describes the known position of the ith galaxy on the celestial sphere, while the true spin vector is either $\boldsymbol{\sigma}_i^1$ *or* $\boldsymbol{\sigma}_i^2$.

A possible approach to analysis in this problem is outlined below. This is not, of course, the only possible analysis, but is presented to make a number of points about data-based inference. In the analysis data-based simulations *must* be used as a means of recapturing the ambiguous nature of the sample data.

It is readily seen that the second possible spin direction $\boldsymbol{\sigma}_i^2$ is the 'reflection' of the first direction $\boldsymbol{\sigma}_i^1$ in the indirection defined by the line-of-sight $\mathbf{p}_i$ to the galaxy, so that

$$\boldsymbol{\sigma}_i^2 = 2(\mathbf{p}_i \cdot \boldsymbol{\sigma}_i^1)\mathbf{p}_i - \boldsymbol{\sigma}_i^1. \tag{4}$$

It is clearly crucial that our statistical analysis should condition upon this relationship.

The simplest probability distribution describing girdle and bipolar alternatives to uniformity on the sphere is the Dimroth–Watson distribution, which has density of the form

$$\frac{b(\kappa)}{2\pi} \exp\left(-\kappa(\mathbf{x} \cdot \boldsymbol{\mu})^2\right) \tag{5}$$

with respect to the uniform measure on S^2. For $\kappa > 0$, expression (5) represents a distribution rotationally symmetric about the unit direction $\boldsymbol{\mu}$ and concentrated around the great circle in the plane orthogonal to $\boldsymbol{\mu}$, while for $\kappa < 0$ it represents a bipolar distribution rotationally symmetric about $\boldsymbol{\mu}$. By treating the ambiguity in spin direction as an unknown parameter, to be maximized over on the null and alternative hypotheses, we may use (5) to construct a test statistic for, say, a test of uniformity of galactic orientations against the equatorial girdle alternative. In doing so we make an assumption of independence of ambiguous pairs of spin vectors. While such an assumption is questionable, it does not affect the data-based approach to significance testing.

Let α_i ($= 1$ *or* 2) be a label indicating the true (but unknown) spin vector for the ith galaxy, and write $\boldsymbol{\sigma}_i^{\alpha} = (x_i^{\alpha}, y_i^{\alpha}, z_i^{\alpha})$, $\alpha = 1, 2$. Then to test against an equatorial girdle effect a suitable test statistic is obtained by maximizing

$$T(\kappa) = -n \log J(\kappa) - \kappa S$$

with respect to κ, where

$$S = \min_{\alpha_i} \sum (z_i^{\alpha_i})^2$$

and $J(\kappa)$ is the integral $\int e^{-\kappa t^2}\, dt$ over the range $(0, 1)$. That a unique such maximizing κ exists follows from a simple application of the Schwarz inequality. If we denote the maximum likelihood estimator of κ by $\hat{\kappa}$, our test statistic is then $T = -n \log J(\hat{\kappa}) - \hat{\kappa} S$.

Existing asymptotic theory is not applicable in this example and the null distribution of T is unknown. Unless we are to develop some new theory, it seems necessary, therefore to assess the significance of the observed value of this test statistic by data-based simulation means. We believe the appropriate simulation base of T-values representing the null hypothesis to be constructed by the following scheme.

On each simulation we keep the galaxy positions $\mathbf{p}_i$ fixed, and simulate, for each galaxy, a first possible spin vector $\boldsymbol{\sigma}_i^1$ by drawing from the null (in this case uniform) distribution

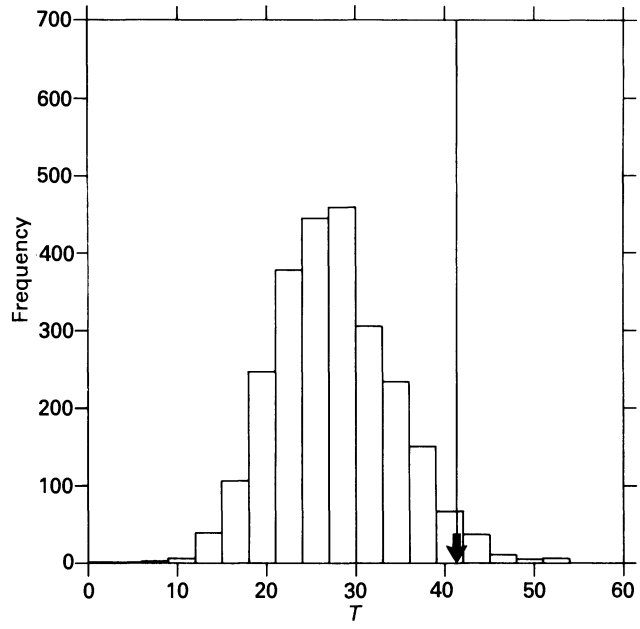

Figure 4. *Null sampling distribution of test statistic for galactic orientation problem. The value for the observed data is shown by an arrow.*

on S^2. Given this first spin vector, and the source position $\mathbf{p}_i$, a second possible spin vector is got from the 'ambiguity specification' (4). This procedure gives us a complete set of ambiguous data for which T can be calculated as before, the whole process being repeated some appropriately large number of times.

We have applied the above procedure to data on $n = 434$ spiral galaxies. This sample delimits the Local Supercluster on the basis of a magnitude criterion and gives a test statistic value of $T = 41{\cdot}3$. This value is found to be significant at about the 2·5% level when assessed against 2,500 simulations. Figure 4 gives an indication of the null sampling distribution of our test statistic in this example.

It has been pointed out that ambiguous observations $(\boldsymbol{\sigma}_i^1, \boldsymbol{\sigma}_i^2)$ corresponding to galaxies seen nearly face on in the line-of-sight are more likely to be subject to measurement error than are those for spiral galaxies seen nearly edge on. An alternative simulation scheme, which uniformly scrambles the projected galaxy images amongst the source positions $\mathbf{p}_i$, provides a means of conditioning on the number of galaxies seen face on, nearly face on, etc. However, in the current example it is precisely the number of nearly face-on galaxies which, in view of the geometrical disposition of the $\mathbf{p}_i$ over S^2, contains most information on the effect being tested against. The source positions $\mathbf{p}_i$ are themselves clustered around the equator of S^2. It should be clear that these positions, in conjunction with a real equatorial girdle effect, would be expected to produce a large number of nearly face-on galaxies in the sample. It would therefore be improper to condition on this number.

This latter point serves as a reminder that there will be circumstances where an otherwise appealing simulation or perturbation mechanism in practice only produces data sets which, in terms of evidence for the phenomenon of interest, are too similar to the actual sample data. The corruption scheme which we utilize must corrupt the specific sample data and not merely some hypothetical data set which might have been observed.

6 Discussion

Classical statistical inference is based on an approach in which evidence in a unique set of data is assessed via long-run frequency in a series of hypothetical repetitions. Within this sampling theory approach to statistics strong arguments exist which suggest that parametric inference ought to be made conditional on the observed value of some ancillary statistic. Much discussion of ancillarity, however, is clouded by problems of mathematical definition. We wish to stress here, within the data-analytic context, the role of ancillary statistics in defining a 'relevant subset' of the sample space for the problem, to which the inference should be restricted. Appropriate conditioning on ancillary statistics is to be seen as important as a means of making the inferential process as relevant as possible to the data under test.

In our bootstrap or data-based approach to testing statistical hypotheses the observed value of the test statistic for the unique sample data is now to be assessed by performing a series of actual repetitions. The conditioning arguments are still valid. Our preceding discussion and examples should make it clear that the term 'relevant' is now to have its interpretation in terms of a conditioning upon those ancillary *features* of the data under test which control the propensity for fictitious occurrences of the effect being tested against. Such features may or may not be quantifiable as ancillary *statistics*. Much of the appeal of the direct simulation methods discussed in this paper derives from their enabling such features to be preserved in a natural way. The idea is to condition, at least approximately, on those features of the data which reflect the tendency for extreme values of the test statistic to occur for artificial or frivolous reasons. This idea of approximate conditioning has also been put forward by Cox (1984), though in a different context.

In the asymmetry example the source positions clearly say nothing about the existence of a topographic regression, and it is for this reason that we condition upon them. It is less certain that we should condition upon the overall assemblage of Δ_i, rather than merely simulate data sets from the null von Mises distribution. Obvious arguments make clear the need to condition on the galaxy positions in the galaxy orientation example, whilst we have already considered reasons for conditioning in the collinearity example.

In a particular application there will be features of the data under test (specific sorts of clumpiness, clustering, etc.) which are not readily quantifiable as ancillary statistics, but which it is nevertheless felt should be conditioned on. The ambiguous nature of the data in the galactic orientation problem, and its specification by (4), are further examples of such data features. Such features may be expressible as *functionals* of the data-generating distribution (number of modes, etc.). It is worth noting that Silverman (1981) makes implicit use of the idea of conditioning on such functionals when he rescales the density estimate constructed from the data to have variance equal to the sample variance. Intuition and scientific collaboration seem the only way of deciding features to be conditioned on. This requirement of extra input to the analysis, beyond that necessary for modelling or determination of a suitable test statistic, probably means that complete formalization of this 'conditionality principle' will not, in general, be possible. Objective conditioning within certain parametric classes of problem ought to be possible, but to restrict attention purely to such problems is to ignore the flexibility of the bootstrap approach to a wide variety of otherwise intractable problems.

Acknowledgements

I should like to thank David Kendall and Bernard Silverman for many rewarding discussions on the ideas put forward in this paper and the referees for their constructive comments. This work was carried out whilst in receipt of a Research Studentship from St John's College, Cambridge.

References

Arp, H. & Hazard, C. (1980). Peculiar configurations of quasars in two adjacent areas of the sky. *Astrophys. J.* **240,** 726–736.
Besag, J. & Diggle, P.J. (1977). Simple Monte Carlo tests of spatial pattern. *Appl. Statist.* **26,** 327–333.
Birch, P. (1982). Is the universe rotating? *Nature* **298,** 451–454.
Copas, J.B. (1983). Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B **45,** 311–354.
Cox, D.R. (1984). Discussion of paper by F. Yates. *J. R. Statist. Soc.* A **147,** 451.
Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.
Efron, B. (1979). Bootstrap methods—another look at the jackknife. *Ann. Statist.* **7,** 1–26.
Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans,* monograph 38. Philadelphia: S.I.A.M.
Hinkley, D.V. (1983). Can frequentist inferences be very wrong? A conditional "yes". In *Scientific Inference, Data Analysis and Robustness,* Ed. G.E.P. Box, T. Leonard and C.F. Wu, pp. 85–103. New York: Academic Press.
Kendall, D.G. (1977). Hunting quanta. In *Proc. Symp. to honour J. Neyman,* Ed R. Bartoczynski et al., pp. 111–159. Warsaw: Polish Scientific Publishers.
Kendall, D.G. & Kendall, W.S. (1980). Alignments in two-dimensional random sets of points. *Adv. Appl. Prob.* **12,** 380–424.
Kendall, D.G. & Young, G.A. (1984). Indirectional statistics and the significance of an asymmetry discovered by Birch. *Mon. Not. R. Astr. Soc.* **207,** 637–647.
Lotwick, H.W. & Silverman, B.W. (1982). Methods for analysing spatial processes of several types of points. *J. R. Statist. Soc.* B **44,** 406–413.
Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* B **43,** 97–99.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B **36,** 111–147.

Résumé

Nous passons en revue les principes sur lesquels reposent les méthodes 'à base d'échantillons' ou 'bootstrap' de vérification d'hypothèse en statistique. Nous faisons une différence entre la motivation pour de telles procédures, où des représentants de l'hypothèse soumise à vérification sont simulés par une falsification des données d'échantillon prélevés au hasard, et les méthodes de calcul bootstrap d'Efron (1979). Des exemples géométriques sont pris en considération et nous attirons l'attention sur l'importance de garder les traits auxiliaires des échantillons soumis à verification dans l'inférence à base d'échantillons.

[*Received August* 1984, *revised August* 1985]