J. R. Statist. Soc. A (1993)
156, *Part* 1, *pp.* 9–37

Exchangeability and Data Analysis

By DAVID DRAPER, JAMES S. HODGES, COLIN L. MALLOWS and DARYL PREGIBON†

University of California, Los Angeles, USA — *Rand Corporation, Santa Monica, USA* — *AT&T Bell Laboratories, Murray Hill, USA*

[*Read before* The Royal Statistical Society *on Wednesday, April 15th, 1992, the President,* Professor T. M. F. Smith, *in the Chair*]

SUMMARY

The term 'exchangeability' was introduced by de Finetti in the context of personal probability to describe a particular sense in which quantities treated as random variables in a probability specification are thought to be similar. We believe, however, that judgments of similarity are more primitive than those of probability and are at the heart of all statistical activities, including those for which probability specifications are absent or contrived. In this paper, we give a definition of exchangeability in a descriptive context, which extends de Finetti's concept to a wider domain. Our objective is to analyse the logic of judgments of exchangeability (or similarity, or homogeneity), to clarify the roles of context and data analysis in these judgments. We give several examples to illustrate the nature of these judgments in description, inference and prediction. We use this discussion to clarify the extent to which judgments of similarity in inference and prediction can be based on data, and the extent to which they must rely on pure faith. Our discussion is a contribution to the emerging theory of data analysis, the as yet largely atheoretical and informal process that precedes and supports formal statistical activities.

Keywords: BAYESIAN MODELLING; BORROWING STRENGTH; EXPLORATORY DATA ANALYSIS; HOMOGENEITY; POOLING; SIMILARITY; SUBJECTIVE JUDGMENTS

1. INTRODUCTION

Statistical methods are concerned with combining information from different observational units, and with making inferences from the resulting summaries to prospective measurements on the same or other units. These operations will be useful only when the units to be combined are judged to be *similar* (comparable or homogeneous). This paper focuses on the *logic* of judgments of similarity, particularly the interacting roles of contextual information and data manipulation or display. To clarify their roles, we propose a modest formalism that identifies the elements of a judgment of similarity. Although this formalism may remind readers of significance tests, it is distinct, and we use it to analyse the logic of similarity judgments, not to construct procedures.

The term 'exchangeability' was introduced by de Finetti (1930, 1974), who defined it in the context of personalistic probability specifications. He used it to describe a sense in which random quantities in such a specification are judged to be similar. We believe, however, that judgments of similarity involve concepts more primitive than probability, and that these judgments are central to preliminary activities that all statisticians must perform, even though probability specifications are absent or contrived at such a preliminary stage. We propose to extend the use of the term

†*Address for correspondence*: AT&T Bell Laboratories, Murray Hill, NJ 07974, USA.

 0035-9238/93/156009 $2.00

'exchangeability' to cover these more primitive judgments. Although we use a term from probability theory, our subject is not part of probability theory or statistical theory or methodology, but rather an approach to a theory of data analysis.

Section 2 gives a motivating example. Section 3 gives our definition of exchangeability. Section 4 shows how the definition illuminates the case that de Finetti considered, in which at least some units have not been observed, as well as other cases. Section 5 illustrates aspects of the definition by considering how it works in three examples. Section 6 discusses the relation between this paper and previous work.

2. REPEATERS STUDY

Judgments of the type that we are exploring will typically be based on a substantial amount of data, but it is not practical for us to present a large body of data here. We begin with a real problem that is sufficiently simple to allow a fair amount of the context and some of the data to be presented in a reasonable space.

Fibre optic cables for data and voice communication contain complicated assemblies called 'repeaters' at regular intervals to intercept and regenerate digital signals. When such cables are in inaccessible locations, such as under the ocean, it is vitally important that the repeaters perform well for long periods, because replacing a defective repeater is very expensive. Repeaters contain various semiconducting devices. These devices are tested carefully, and some devices are not installed in the cable because they are judged to be of slightly lower quality. Some of these latter devices are used in on-going experiments intended to provide information about the reliability of the devices that were installed. The objective of these experiments is to estimate the useful life of the cable and to identify deficiencies so that they can be eliminated from future production.

One series of such experiments studied the frequency of certain errors that seem to be unavoidable. Every so often, an incoming '1' bit (out of the millions that arrive every second) is read as a '0' or vice versa (the phenomenon may be asymmetric). The causes of these events are not completely understood, and may include γ-rays, decay of trace amounts of radioactive elements in the devices and imperfections during manufacture. As part of this series of studies, at the specific request of one of us, detailed records were kept of the occurrence of errors in several devices maintained in an environment as close as possible to field conditions. Previously, only aggregate data (number of errors per hour) had been recorded, on many devices, for many hours, and Poisson models had been assumed to be relevant, but persistent small anomalies suggested examining whether errors tended to occur in bursts.

Fig. 1 displays the occurrence of errors for each of three devices over 10 hours. The following are among the questions of interest here (for now we use the terms 'like' and 'similar to', but in the next section we shall give a precise definition using the term 'exchangeable with').

(a) Are the records for each device free of any time trend?
(b) Are these records like what we would expect from a Poisson model?
(c) Are these three devices similar to one another?
(d) Are these laboratory devices similar to the devices in the cable?
(e) Will future experience with the laboratory devices be like what we see here?
(f) Will future experience with the devices in the cable be like what we see here?

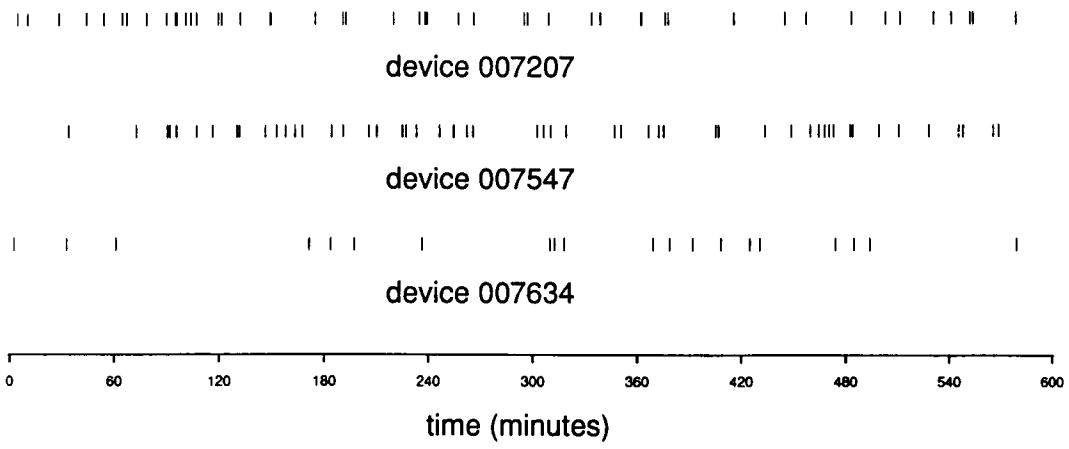

Fig. 1. Repeaters: times at which errors were observed in three repeaters over 10 hours (see Section 2)

Questions like these are typical in the application of statistical methods and are of various kinds. Answers to the first three can be based directly on the data in hand. Such judgments are basic, and we should understand how they are made before proceeding to consideration of more speculative judgments such as those required by questions (d)–(f). Judgments concerning these last three questions cannot be based solely on the records immediately at hand; various other data are available and relevant. (We present the devices' identification numbers to emphasize that such additional data could be accessed by using the identifiers.) The answer to question (d) depends on our judgment—based on the data that we have—about how these devices and those in the cable were manufactured and selected; question (e) depends on our judgment—based on the data that we have—about the mechanism generating the errors; question (f) is a combination of (d) and (e).

This example raises some general questions. How do analysts use contextual information—including the purpose of the work—and data analysis to evaluate, to modify and eventually to justify judgments of exchangeability? How are such judgments used in inference and prediction? Lindley and Novick (1981) give a superb discussion of how exchangeability judgments, once formed, are central in the proper formulation of inference and prediction problems, but they do not discuss how to use data to make those judgments. This paper is intended to complement their work by considering situations in which data are available to inform exchangeability judgments, which then allow inference or prediction from units that are seen to others that are unseen.

We use Bayesian terminology, but the questions that we address can be expressed in frequentist terms. In our example, the analyst is concerned with defining relevant populations of devices and potential measurements on them, and with whether the available data can be regarded as a realization of some random sampling mechanism applied to the relevant populations. Such judgments must precede the application of standard probability-based methodology. Few would disagree that judgments of this type should be based on data; the question is precisely how data and contextual information are used to make them. Much of classical experimental design and sampling theory is about ways to simplify such judgments, and the theory of

observational studies is about how to make them in the absence of randomization. In such cases, the role of formal statistical methodology is often minimal, with the difficulty lying in the decision about what is relevant. An example of a problem of this type is given in Section 5.3.

Implicit in the above discussion is the concept of a *unit*. We use this term in the way that de Finetti uses 'event', to denote an entity bearing a quantity or vector that has been or can be observed. Measurements on a unit are either of primary interest or concomitant—perhaps bearing contextual information—so the measurements are generally multidimensional. 'Repeated measurements' may correspond to distinct units, or may be components of a multidimensional observation on a single unit. Judgments involved in defining units and measurements rely on data and context, and logically precede judgments of exchangeability.

For questions (c)–(f) in the repeaters study, it is natural to take an individual device as a unit, but for questions (a) and (b) different units are needed, e.g. hours or possibly shorter periods within devices, or intervals between successive errors within devices. In a given analysis we may study several possibilities and the choice may depend on a preliminary analysis of the data. This choice is illustrated further in Section 5.1.

3. EXCHANGEABILITY

Questions (a)–(f) concern ways in which units can be grouped into sets. Some groupings are meaningful in context, others are not. Here, it may be meaningful to classify intererror times according to the length of the preceding interval, which could allow for a transient fatigue effect, or to differentiate devices by their date of manufacture, or by whether they are observed in the laboratory or in the cable.

The basic idea of this paper is to formalize how analysts judge whether some meaningful sets of units (not all sets!) are 'similar to an adequate approximation'—in our terminology, exchangeable. Suppose that we have a set U of units, for each of which some measurements or attributes are available. Consider a collection of ordered subsets of U, namely $S_1, S_2, \ldots$. These need not be disjoint. Define a *comparison* to be an unordered pair (S_i, S_j) of these subsets of U. Let Ω be a set of such comparisons. The notion 'to an adequate approximation' must involve three distinct things: a summary *descriptor* $d(\ )$ that is a function of the measurements on the units in a subset of U, a *norm* $|\ |$ applied to the difference between two such summaries and a *caliper* C to which such differences are compared. If a group of units is to be treated as exchangeable, no two meaningful subsets can differ much on interesting characteristics. The comparison set Ω specifies the collection of meaningful subsets. The descriptor function $d(\ )$ specifies the interesting characteristics. The norm $|\ |$ measures the difference between two subsets of units. Descriptions of meaningful subsets need not be identical, only sufficiently close for the purpose at hand. The caliper C specifies 'sufficiently close'.

3.1. *Definition*

The units in U are $(\Omega, d(\ ), |\ |, C)$ exchangeable if for all comparisons (S_1, S_2) in Ω

$$|d(S_1) - d(S_2)| \leqslant C.$$

The context must inform the choice of each of the components of this definition.

A specification of the comparison set Ω describes which aspects of the structure of the data may be considered. Choosing the descriptor d requires specifying which aspects of the units need to be compared. Specifying the norm $|\ |$ and the caliper C prescribes what differences are sufficiently big to matter. These choices may be elicited, in part, by data analysis.

Context is essential in defining permissible comparisons, and it is important that *some possible pairs of subsets be excluded*. If Ω included all pairs of subsets, S_1 and S_2 could be chosen to maximize $|d(S_1)-d(S_2)|$, and the units in U would be judged exchangeable only when all differences among them were trivial. Usually the comparison that maximizes $|d(S_1)-d(S_2)|$ will be contextually vacuous. The subsets S_i are *ordered*, so that different permutations of the units can be compared. Judging that all permutations of a set of units are exchangeable is equivalent to judging that the identifying labels on the units can be ignored for the purpose at hand, although the labels may be important for other purposes.

Context is also important in choosing the descriptor d. Often ancillary quantities are available, and the analyst must make an explicit judgment to ignore them. In the repeaters problem, ancillary quantities such as ambient temperature and humidity were recorded. No strong relationships were found between these measurements and the error series, so the tentative decision was made to ignore them. Notice that d can be multidimensional and can be a *robust* descriptor, so that two sets of units can be judged exchangeable even when either or both contains outliers.

Choosing the norm $|\ |$ is usually not critical, unless d is multidimensional, when the decision has to be made about which aspects of the difference between S_1 and S_2 are most important. (Actually it need not be a norm, merely a measure of distance.)

The caliper C specifies which differences between the subsets matter for the purpose at hand, so C must also depend strongly on the context. It may sometimes be appropriate to use significance tests to select the caliper C, even in cases without explicit stochastic content (Freedman and Lane, 1983; Finch, 1979), but C is not limited to this use. For example, Ehrenberg (1968) discusses 'law-like' relationships—relationships between two or more quantities that hold to acceptable accuracy across a wide range of circumstances. His main example concerns the relation between height and log-weight of children, which closely follows the same linear relationship for children of different ages, nationalities, social classes and genders, and in data sets collected by many experimenters. As Ehrenberg's collection of data sets has grown, some systematic deviations from the linear 'law' have been found, but they are small compared with the variation captured in the law. As Ehrenberg put it after noting that French boys tended to be somewhat heavier than other boys as they grew older (Ehrenberg (1968), pages 293–294),

> 'it is descriptively convenient to keep to a single numerical generalization that $\log \bar{w} = 0.8\bar{h} + 0.4$ holds to within 0.01, together with the verbal statement that the older French boys are about 0.01 heavier. If the deviations subsequently generalize for other data one would often formalize such sub-laws.'

Even though the French boys are known to differ, the descriptive statement '$\log \bar{w} = 0.8\bar{h} + 0.4$ holds to within 0.01' is still accurate and may be adopted in 'a deliberate act of oversimplification' (Ehrenberg (1968), p. 288). In our terms, contextual considerations lead to a C sufficiently large that French boys are exchangeable with

other children with respect to how their heights and weights are related, even though a smaller caliper would lead them to be treated as not exchangeable in this way.

3.2. *Further Comments*

Should exchangeability of the units in a data set be defined as an absolute property that is present or not irrespective of the purpose of the analysis? This seems inappropriate. A set of units may be exchangeable for some purposes but not for others, depending on what is measured (the descriptor d) and the questions of interest (the comparison set Ω).

An obvious and necessary generalization of exchangeability is *conditional exchangeability* of units given concomitant quantities, such as ambient temperature in the repeaters example, or given values of a function of the quantities of interest. (This is also known as 'partial exchangeability'; see de Finetti (1938, 1972) and Diaconis (1988).) Informally, the units are not similar (to adequate tolerance for the purpose at hand) overall but instead are only similar within groups defined by various levels of the concomitant quantities. If some of the concomitant quantities are continuous, we can model the dependence between them and the quantities of interest smoothly, for example with regression, and apply the definition to the residuals from this fit. Statistical modelling—tentatively entertaining a structure for the data, examining residuals from fitting this structure (model), modifying the current model, and so on, eventually arriving at a final model—may be viewed as the iterative process of entertaining, modifying and finally making conditional exchangeability judgments. In this way we are led to choices about how to pool or to 'borrow strength' (Tukey, 1986) from similar units. At some point in this process we must make a judgment of unconditional exchangeability.

A technical difficulty arises in applying our definition when structure (depending on concomitant quantities) has to be allowed for. It may not be appropriate simply to choose a descriptor d that is a function of empirical residuals, and to ignore that some structure has already been fitted to the data. However, in principle it is possible to define more appropriate ds or more appropriate residuals so that the definition can still be applied. A related point comes up briefly in Section 6 under the heading of diagnostic methods in regression.

Finally, consider an alternative definition: let $d(\ )$, $|\ |$ and C be as before, and let Ω^* be some collection of subsets of units.

> The units in U are $(\Omega^*, d(\ ), |\ |, C)^*$ exchangeable if and only if, for some description d_0, for all subsets S^* in Ω^*
>
> $$|d(S^*) - d_0| \leqslant C,$$

i.e. all relevant subsets of the data have the same description, to within C. But this definition is equivalent to a special case of that given earlier: with $d(\ )$ and $|\ |$ held fixed throughout, $(\Omega^*, C)^*$-exchangeability implies $(\Omega, 2C)$-exchangeability with $\Omega = (\Omega^*, \Omega^*)$, and (Ω, C)-exchangeability implies $(\Omega^*, C)^*$-exchangeability, where Ω^* contains exactly all the sets that appear as components of pairs in Ω.

4. THREE SENSES OF EXCHANGEABILITY

We have used the term 'exchangeable' in an unfamiliar sense, so to clarify we

distinguish three related concepts encompassed by the definition. Consider a set of n units, on each of which is measured a vector of concomitant quantities x_j, $j=1, 2, \ldots, n$, and on each of which a quantity of primary interest has been or could be measured. When it has been measured we use a lower case y, and when it has not been measured we use the upper case Y.

The first sense of exchangeability is *descriptive exchangeability*, which is appropriate when the y_j have been observed for $j=1, \ldots, n$. In this case, the observations (both y_j and the concomitants x_j) are used to construct the set of permissible comparisons Ω, and then this Ω is used with appropriate d, $|\ |$ and C to judge whether the units are $(\Omega, d, |\ |, C)$ exchangeable. If they are, a simple description of the set of units suffices.

The second sense of exchangeability is *exchangeability of measured and unmeasured units*, which is appropriate when y_j, $j=1, \ldots, m$, have been observed and Y_j, $j=m+1, \ldots, n$, have not. In this case, judging all the units exchangeable and specifying a probability model linking the observed x_j explicitly to the unobserved Y_j requires three distinct applications of the above definition and a leap of faith, as follows.

First, the observations (x_j, y_j) are used to define a comparison set Ω_m that is suitable for judging the descriptive exchangeability of the first m units, and the definition is applied as before to decide whether the first m units are $(\Omega_m, d, |\ |, C)$ exchangeable.

Second, if the first m units are exchangeable, and if the concomitants x_j for the last $n-m$ units are judged exchangeable with those for the first m units, then the descriptive exchangeability of the first m units may be extended by a leap of faith to a judgment of exchangeability of all n units. We note in passing (although we shall return to this in Section 5) that such leaps of faith may be more or less firmly grounded.

Third, with the seen and unseen units judged exchangeable, it may then make sense to specify an explicit probability model for the Y_j. This involves judging that $y_1, \ldots, y_m$ are 'like' a random sample from a given distribution and extending that 'likeness' to the unseen units by the previous judgment of exchangeability. We formalize the notion 'being like a sample from a given distribution' (see question (b) in Section 2) by viewing that distribution as a notional infinite data set. To apply the definition to this formalization, let U be the union of the infinite data set S_1 and the observed data set S_2, and let Ω consist of the singleton (S_1, S_2). Take the description $d(S)$ to be the empirical measure of the data set S, and let $|\ |$ be some suitable measure of distance between two such measures. If the caliper C is chosen appropriately, being like a sample from a given distribution corresponds to the judgment that the observed and notional data sets S_2 and S_1 are exchangeable according to the definition. This permits a probability specification, within which one can employ familiar inferential and predictive procedures.

The third sense of exchangeability is the sense used by de Finetti and his school; it relates to the case where all the Y_j are unseen. We could express the concept in terms of notional infinite data sets as before, but we have chosen to use more formal language to make the relation with de Finetti's concept as clear as possible. To de Finetti,

> a sequence (finite or infinite) of random variables $Y_1, Y_2, \ldots$ taking values in some space Y is *exchangeable* if, for all $k \geqslant 1$, for all sets of distinct indices

$i_1, i_2, \ldots, i_k$ and for all measurable sets A in Y^k, we have

$$P(\{Y_{i_1}, Y_{i_2}, \ldots, Y_{i_k}\} \in A) = P(\{Y_1, Y_2, \ldots, Y_k\} \in A).$$

This can be regarded as an application of our definition as follows. Take U to be the set $\{Y_1, Y_2, \ldots\}$. For each k, take the subsets $S_{k,i}$ to be a list of all k-element ordered sets of distinct indices. For each k, include in Ω all pairs $(S_{k,i}, S_{k,i'})$. Take the description of a k-dimensional quantity W to be the probability measure of that quantity. Take the norm, measuring the distance between two measures μ_1 and μ_2, as

$$\sup_A |\mu_1(A) - \mu_2(A)|,$$

and take $C = 0$.

de Finetti's writings suggest that he thought of exchangeability in at least two ways: as a property enjoyed by a probability assignment (de Finetti (1974), chapter 11) and as a judgment made during the construction of a probability assignment (e.g. his frequent coin flipping examples). Many subsequent researchers have concentrated on the first interpretation, which lends itself to the formulation of clean mathematical problems (e.g. the extension of the concept to finite sets of random variables; see Diaconis (1988)). As far as we can tell, for de Finetti the choice of a probability specification was a primitive operation that he did not analyse into component stages. He used the term 'recognition of analogy' to justify exchangeability judgments among future observations; in our language he used observed concomitants x_j to set up Ω and then asserted exchangeability among the Y_j by a leap of faith. Whenever de Finetti mentioned the use of data to inform exchangeability judgments (as for example towards the end of de Finetti (1938), quoted by Diaconis (1988)), he always appealed to Bayes's theorem, assuming that probabilities had been assigned at an earlier stage.

We view the third sense of exchangeability as describing a property of a probability specification. We think that *choosing* such a specification involves thinking about exchangeability in the other two senses.

5. EXAMPLES

The three examples that follow illustrate several points. All parts of the definition of exchangeability come into play in each example, even when they are not specified directly, and the comparison set Ω, the descriptor d and the caliper C are all to a greater or lesser extent determined by the context of each problem. All the examples illustrate the iteration between looking at data and examining the context for meaningful comparisons, ds or Cs. Finally, the examples illustrate the steps needed to connect measured units to future observations on unmeasured units and the role of leaps of faith.

5.1. *Repeaters Study*

Question (a) (is there a time trend?) of the repeaters study of Section 2 requires a judgment of exchangeability among multiple observations on the same physical device. We might start by taking an hour to be a meaningful unit. In any one hour (for one device), let k be the number of errors. Table 1 gives numerical values for k for each of the three devices in Fig. 1. We need to decide what sets of hours should be considered and how to describe the relevant aspects of any chosen set. For the first point, if we are concerned about a possible time trend, and in the absence of

TABLE 1
Summary of errors k by hour for the repeaters data

Device	*Hourly counts of errors*									
007207	5	10	5	6	4	3	5	2	4	4
007547	1	6	7	8	5	6	5	7	5	4
007634	2	1	1	3	0	3	4	3	2	1

concomitant information, it is reasonable to consider only consecutive sets of hours. If other information distinguished some hours from others, we could group the hours in other ways.

Some possibilities for descriptor functions of a subset $S=(h_1, \ldots, h_n)$ of hours include the following:

$$d_1 = \text{average}(k_j \mid j \in S);$$
$$d_2 = \text{variance}(k_j \mid j \in S);$$
$$d_3 = d_2/d_1.$$

It is easy to extend this list of descriptors; for example, we could try regression on time to look for polynomial or sinusoidal trend. Because the data record occurrences of relatively rare events, we have a strong interest in comparing the observed error counts with their predicted behaviour under a Poisson model. To address this (our question (b)) thoroughly, we shall need to invent descriptors that are sensitive to various kinds of departures. One possible descriptor is the variance-to-mean ratio d_3 on the list above (under a Poisson model $d_3 = 1$). But we might have been led to consider d_3 simply by analysis of the data: we might have observed that this quantity is approximately 1 for many subsets of hours, so that large departures from 1 for some subsets might indicate something unusual.

One way to approach question (a) is to compare early with later hours. We could take Ω to be the set of all pairs $\{(1, \ldots, h), (h+1, \ldots, 10)\}$ with $1 \leqslant h \leqslant 9$. We must also choose a norm and a caliper for each d. Because each d is real valued, there is no reason not to use absolute differences. Before choosing the calipers, it is useful to look at some numerical values. Taking $h = 5$ as an example, Table 2 compares the first five hours with the last five.

TABLE 2
Values of the descriptors d_1 (average), d_2 (variance) and d_3 (variance-to-mean ratio), for the repeaters data†

Device		d_1	d_2	d_3
007207	S_1	6.0	5.5	0.92
	S_2	3.6	1.3	0.36
007547	S_1	5.4	7.3	1.35
	S_2	5.4	1.3	0.24
007634	S_1	1.4	1.3	0.93
	S_2	2.6	1.3	0.50

†The set of comparisons Ω consists of the pairs $(S_1, S_2) \equiv$ (first 5 hours, last 5 hours).

We see that for the second device d_1 is exactly the same for the two subsets of hours, whereas for the other two devices there are differences. What size difference should we take as important? (Not 'significant' in the technical sense: we do not have a probability model yet.) If, for example, all differences between these d_1 values were judged too small to worry about, we might take $C_1 = 4$ and conclude that all the hourly observations for a given device are sufficiently similar (as far as d_1 is concerned). Or if a smaller caliper were substantively more reasonable ($C_1 = 1.5$, say), we would not regard the first and last five hours for the first device as exchangeable but would for the other two devices. Given average error rates per hour for the three devices of 4.8, 5.4 and 2.0, a useful summary based on this smaller caliper that begins to address question (c) in Section 2—are the devices similar?—might simply state that the error rates are about 5 per hour for the first two devices and about 2 per hour for the third device.

If we desired detailed modelling, given our strong *a priori* bias towards the Poisson model in this case, we might want to compute P-values under a simple Poisson model with constant rate parameter over time. With this amount of data, we shall not be able to reject such a model, but with a data set of more realistic size we would probably find small but persistent departures from the time homogeneous Poisson model. We would then have to decide whether to ignore the departures, or, in our terms, whether to relax the Poisson-based caliper to something less stringent. This decision would have to be based on both the data and the prospective uses for the Poisson model. For example, if we wanted to do hypothesis tests comparing groups of devices, the difference between the Poisson and negative binomial distributions could be important, whereas for descriptive purposes the Poisson model might be adequate.

Another way to approach question (b) is to take a gap between successive errors to be a unit and to ask whether these are exchangeable. If they are, we might consider a stationary renewal model for the errors. For any one device, we can take S_1 to be the set of gaps that are preceded by a short gap and S_2 to be those gaps preceded by a long gap. If we take short to mean less than 10 min and long to mean more than 10 min, we can compare S_1 and S_2 by means of a Q–Q-plot as in Fig. 2. This shows some departure from the 'null' configuration represented by the line $y = x$. With more data, it would be easy to judge the configuration's linearity, and any of

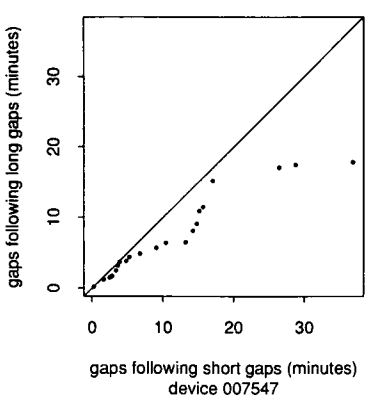

Fig. 2. Repeaters: Q–Q-plot of (intererror) gaps following long gaps *versus* gaps following short gaps, for the second repeater (see Section 5.1)

several measures of departure from linearity could be used. To put this analysis into the framework that we have set up, we would have to express each such measure in the form $|d(S_1) - d(S_2)|$. This would present only technical difficulty. Fig. 2 constitutes a description of (one aspect of) the device considered and might also be used in approaching question (c) where now the units are the devices.

Answering questions (d)–(f) about the devices involves an extrapolation from devices seen in the laboratory to those unseen, in the cable. This requires either a judgment that the devices in the two places are exchangeable or a more complicated judgment of exchangeability conditional on quantities that have been measured for all devices. If the devices in the laboratory are judged descriptively exchangeable, and our study of the available data leads us to judge that the units in the cable do not differ from those in the laboratory in any material way, the judgment of exchangeability may be extended to include the devices in the cable.

This latter leap of faith may be more or less firmly grounded. The devices installed in the cables were selected because they performed better in tests done before installation. Thus exchangeability would be hard to argue in this case, although a weaker *a fortiori* argument might be useful and defensible.

More generally, if we have abundant experience of similar units and have had no recent surprises, this leap may require less faith than if we have had little experience with the units. This case arises in careful experimentation with standard equipment. In Ehrenberg's (1968) problem regarding heights and weights of children, the fact that a single relationship has been found to hold in many different environments gives encouragement in extrapolating to a new environment. Tukey (1986), p. 277, distinguishes between 'borrowing strength', where one uses data from other situations even though one is interested only in the present case, and 'broadening the basis', where one attempts to generalize a finding by drawing on a wider variety of data. Ehrenberg's example is of the latter type; such arguments can provide grounding for a leap of faith from seen units to unseen.

This example illustrates several points. Several comparisons were made, each driven by the context, and the caliper was determined by the purpose of the analysis (i.e. hypothesis tests or description). Many descriptors were suggested by the context: no 'sufficiency' concept is available to prune them (see the discussion in Mallows and Pregibon (1987)).

5.2. *Old Faithful*

Consider the data given by Weisberg (1980, 1985) on eruptions of the Old Faithful Geyser. An exploratory analysis of these data has been published by Denby and Pregibon (1987) and will be discussed here. The data record intervals between eruptions and durations of eruptions, gathered on each of eight consecutive days between 6 a.m. and midnight. The ultimate purpose was to provide a system that the park rangers could use to predict the time of the next eruption.

In this example, the definition of units is unclear. One possibility is to say that there is only one unit, Old Faithful, on which repeated observations are made. Another is to take days as units, with interval and duration measurements regarded as multivariate observations on the units. Or individual eruptions, with associated durations and (following) intervals, could be units. This last choice is supported by the contextual judgment that the 'interval' may depend on the duration of the preceding

eruption. If the objective were to predict durations rather than intervals, the units might be chosen to correspond to durations and the *preceding* intervals.

Denby and Pregibon (1987) examined stem-and-leaf diagrams of both intervals and durations. Both exhibited strong bimodality. This is counter-intuitive in this context and suggests that some grouping of the data is present, i.e. that the eruptions as described by their preceding intervals should not all be considered exchangeable. In this case, Ω was defined after looking at the data: the presence of multiple modes violated intuition about measurements of this sort, in the absence of grouping, and the permissible comparisons included a partition made by splitting the intervals according to the mode to which they were closest.

So how can we arrive at an exchangeable set of units? Staying with interval as the only relevant measurement on an eruption, a first conjecture is that some days tend to have long intervals and other days short intervals, i.e. that the units might be exchangeable within days but not between days. Box plots of intervals, one for each day, showed this to be false. Days, although different, were not sufficiently different according to the norm and caliper applied to this box plot summary, so that the bimodality must have occurred within days, i.e. eruptions within days were not exchangeable. Plots of intervals in the time sequence in which they were collected show distinct saw-tooth patterns, strongly suggesting that (if only intervals are to be studied) the description of an eruption should include a reference to the interval preceding the previous eruption.

Turning now to the duration measurements, a scatterplot of interval against preceding duration (Fig. 3) shows a distinct clumping of points in the upper right-hand and lower left-hand quadrants of the picture, indicating that the bimodality of the marginals carries over to the joint distribution. But some points are in the upper left-hand and lower right-hand quadrants—all from day 7. When time-order plots of interval and duration are superimposed (Fig. 4), the saw-tooth patterns of the two series correspond closely except for day 7, for which the saw-tooth patterns are one time period out of phase. This apparent failure of exchangeability of days suggests a data transcription error that, when corrected, makes day 7 like the other days.

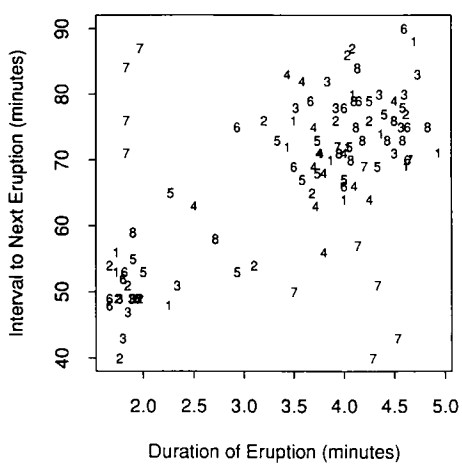

Fig. 3. Old Faithful: scatterplot of intervals between eruptions *versus* durations of previous eruptions of the Old Faithful Geyser (see Section 5.2)

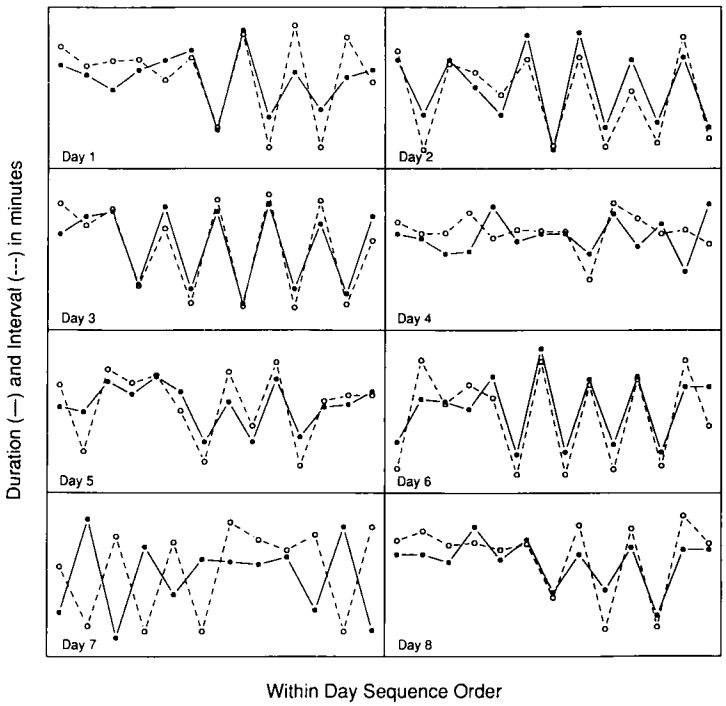

Fig. 4. Old Faithful: superimposed time sequence plots of intervals between eruptions and durations of previous eruptions of the Old Faithful Geyser over an eight-day period

The ultimate judgment was that eruptions are not exchangeable within days, because of the serial correlation, but that days are exchangeable, and residuals from a model that allows for serial correlation are also exchangeable.

Weisberg (1985) used standard regression methods (ignoring the serial structure) to obtain nominal 95% prediction intervals. He also regressed interval on previous duration and the two previous intervals. Azzalini and Bowman (1990) analysed a more extensive set of data and proposed a second-order Markov chain model. We thus have several distinct ways of obtaining predictions. In this case, all analyses, except that based on the original data with day 7 uncorrected, give predictions of about the same accuracy.

This example illustrates three points. First, much thought and data analysis may be needed to arrive at suitable definitions of the units and the four components of our definition. In particular, the descriptor function d may evolve from the analysis, as it did here, and, interestingly, as it did in the analysis of Azzalini and Bowman (1990). Second, the example provides an instance in which the analysis came to a reasonably satisfying conclusion, but this need not always happen. Suppose that the anomalous behaviour of day 7 had not had a ready explanation, and that no contextual information could be found to differentiate day 7 from the other days. Then the data analyst could not have established any defensible groups of exchangeable observations. When description of the data is the only purpose, this outcome simply means that the description cannot be as succinct as we might hope but (as we shall see in the next section), when the purpose is inference or prediction, such an outcome can create

serious difficulty. If the search for exchangeability does not terminate successfully, we may have to abandon the statistical approach altogether. Third, in this case Weisberg's analysis of the original data—which relied on an exchangeability judgment (between days) that is grossly violated in the data—yields a predictive accuracy that is 50% worse than the analysis presented here. But choosing to ignore another failure of exchangeability (within days) leads to models with predictive accuracy that is very similar to those that did account for it. Thus, although a particular failure of exchangeability may be descriptively important in a given case, it may not matter predictively (i.e. for that purpose the implied caliper C is wider).

5.3. *Quality of Health Care*

Consider using mortality rates to assess the quality of care provided by hospitals for specific diseases. For the moment, consider the narrower problem of predicting a hospital's future performance from its past performance. By phrasing the problem in this way we have implicitly defined units to be (hospital, time period) pairs. Suppose that we have ample past data on a given hospital's history of treatment of, say, acute myocardial infarction (AMI). Suppose further that we wish to assess this hospital's likely performance for AMI patients next year, to decide whether to audit its care more thoroughly or to intervene in its management. As Lindley and Novick (1981) make plain, to do this we must make some connection between the past (seen) and the future (unseen).

One of us (Draper *et al.*, 1990; Rogers *et al.*, 1990) has studied several years of patient records at a large sample of American hospitals. Each patient record includes characteristics of the patient on admission to the hospital, details of his or her care and whether he or she was alive 30 days after admission. For the present discussion, the summary measure of performance at a hospital is a mortality rate, which can be computed for all AMI patients or for any interesting subgroups. The simplest judgment would be that next year's mortality rate and those in previous years are exchangeable. If we had next year's rate we could apply our definition descriptively as in Section 3, but we do not. In the spirit of Section 4 we can ask whether we would treat as exchangeable the mortality rates that we have observed. If so, we may be willing to extend that judgment to next year, but if we do not consider the observations in hand to be descriptively exchangeable we are hard pressed to justify extending such a judgment to the future period.

Substantive considerations suggest that time trends in mortality rates may be present, arising from changes in how sick AMI patients are when they arrive at the hospital (measured by an illness severity index), from trends in the quality of the care that they receive (measured by an index evaluating the processes of care), or both. A plot of the severity index against time (Fig. 5) shows an increasing trend, but a similar plot of the process index shows that quality of care is also rising (Fig. 6). These two trends would have opposite effects on mortality: other things being equal, sicker patients mean higher mortality, but better process of care means lower mortality. Not surprisingly, time trend plots of mortality at individual hospitals show varying patterns: for some hospitals, mortality remains about the same; for others it goes down; for others yet it goes up. The judgment of unconditional exchangeability of mortality rates across years within hospitals appears untenable. Instead, years are exchangeable in mortality only after conditioning on severity and process.

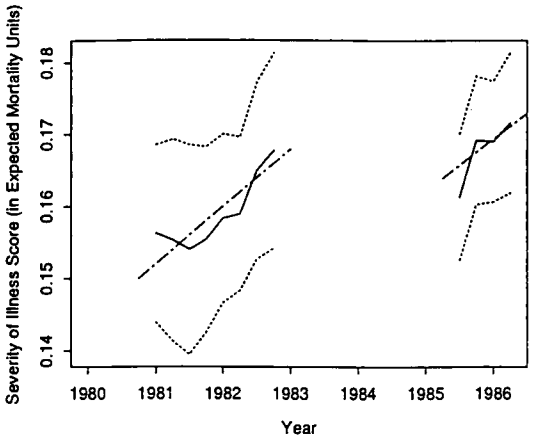

Fig. 5. Quality of health care: time series plot of index of severity of illness on admission to hospital (see Section 5.3); data representative of the US population 65 years of age and older were gathered at 297 hospitals for patients with one of five diseases, including AMI (the plot for AMI is similar but exhibits more random variation); severity of illness has been scaled in this plot to correspond to expected mortality of a patient assuming average quality of care (———, severity of illness observed quarterly; · · · · ·, two standard errors each way from the full curves; - - - -, linear trends in the observation periods 1981–82 and 1985–86)

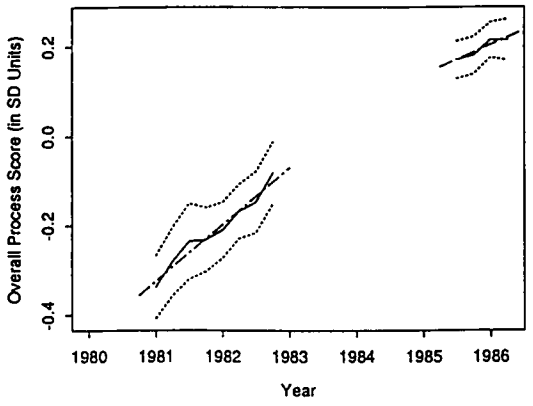

Fig. 6. Quality of health care: time series plot of index of quality of care in the hospitals study; quality of care was measured by examining the appropriateness of the processes of care given to patients; the resulting index was scaled to have mean 0 and standard deviation 1 (———, process observed quarterly; · · · · ·, two standard errors each way from the full curves to indicate sampling uncertainty; - - - -, linear fits to the observed process values in the two observation periods 1981–82 and 1985–86)

To check this by using our definition of exchangeability, Ω could be defined by grouping the years according to the values of the hospitals' severity and process indices, d the mortality rate, $|\ |$ Euclidean distance and C something small like 0.025 (for the types of patients in this study, AMI has a 30-day mortality rate of about 25%, and policy makers regard differences of the order of a tenth of this overall rate—or larger—to be substantively meaningful). Suppose that descriptive exchangeability holds for this set-up, and that further reflection yields no other observed factors on which to condition. Then to make a Lindley–Novick-style

exchangeability judgment for a prediction to the next year, the logic is as follows:

(a) past mortality rates are descriptively exchangeable, conditional on severity and process, and

(b) we can find no other contextually meaningful ways to differentiate between years or between patients within years.

This much is purely descriptive. To extend this to the following year, we must add

(c) we judge next year's mortality rate to be exchangeable with previous years', conditional on severity and process.

The descriptive part is justified by data; the extension is pure judgment, a leap of faith. In this case, evidence of the sort described here is all that can be supplied by statistics or data analysis to buttress a judgment of exchangeability.

6. DISCUSSION

Nelder (1986), p. 113, described *homogeneity* as a fundamental notion in statistics and the sciences that it serves. But we have few tools for declaring things significantly the same (Nelder, personal communication). Significance tests cannot do this. If an observed difference between two sets of units could be noise, we may not necessarily conclude that the two sets should be treated as if indistinguishable; conversely, if an observed difference is too large to be noise, it does not follow that we should differentiate the two sets. Context and purpose are indispensable in such choices, but significance tests are non-contextual. A decision that some units are homogeneous involves assessing the substantive or practical significance of an observed difference rather than its statistical significance.

Decision theory may appear suited to declaring things significantly the same. But the use of decision theory requires fully specified probabilistic machinery, and it is clear that statisticians must do much learning and deciding before any plausible probabilistic specification can be stated (see Smith (1986)). This learning and deciding has come to be called, after Tukey, data analysis. But data analysis is largely without theory (though some tentative steps have been made, e.g. Good (1983a), Mallows and Walley (1980) and Mallows and Pregibon (1987)), to the extent that some with a formal bent (e.g. Smith (1986)) admit that they do it but have no language in which to discuss it. Can a more formal language be devised for data analysis in this and the other roles that it plays?

Certainly, previous work on data analysis is not devoid of formal structure. Treatments following Tukey (1977) are guided by themes such as the idea that

$$\text{data} = \text{fit} + \text{residual}$$

or that data should be summarized to reveal structure. However, this work has focused so little on contextual matters that critics have—unfairly, we think—caricatured exploratory data analysis as concerned, for example, with searching for interesting patterns in the heights of the 20 tallest volcanoes in the world. Cox and Snell (1981) and Chatfield (1985) emphasize the role of data analysis in particular phases of a statistical analysis, shifting attention from techniques to their use in context. Both of these treatments emphasize sequences of stages of statistical analyses, and although both often note the importance of context, each finds it 'difficult . . . to convey the interplay between subject-matter considerations and statistical analysis that is essential

for fruitful work' (Cox and Snell (1981), p. 3), so this interplay is exemplified but not really analysed. Chatfield (1985), pages 218–219, notes

> 'the need [in the initial examination of data] to collaborate with appropriate experts, to incorporate as much background theory as possible, to look at the data and recognize their important features, to check that a model formulated on empirical and/or theoretical grounds is capable of reproducing the main characteristics of the data, and to look for improvements as necessary'.

The present paper is an attempt to analyse how these activities are performed, particularly in the initial examination of data and in model formulation, using as vehicles the notion of exchangeability and the formalism introduced in Section 3.

Although we believe that our formalism is novel, we could hardly claim to be the first to consider the use of data to inform exchangeability judgments. For example, the related notion of 'recognizable subpopulations' is due to Fisher (see, for example, Fisher (1956)). Among other early researchers who offer tantalizing comments on this paper's topic, Shewhart (1939) is especially noteworthy. We have discussed what we can infer about de Finetti's position in Section 4. Other researchers make allusions that suggest that they thought more about this matter than they wrote (e.g. Savage (1967) on necessarians, or Good (1983b,c) on events that have never occurred), but we have been unable to find any direct discussion of the role of data in informing exchangeability judgments in those most associated with the idea of exchangeability (de Finetti, Savage, Good or kindred researchers), although students of Savage say that he was interested in it.

It is striking that Bayesians, who originated the notion of exchangeability, have had relatively little to say about how to use data to inform exchangeability judgments. A strict reading of personalists might suggest that we are discussing forbidden ways to use data, but Savage's last papers show a high regard for 'puttering about with the data' (Savage, 1977), which we construe as learning by means other than Bayes's theorem. Subsequent Bayesian writers (e.g. Smith (1986) and Hill (1990)) acknowledge a need for informal methods preceding the use of formal Bayesian methods, but reject significance tests without proposing another way to understand the informal activity. The present paper can be viewed as an attempt to use a hitherto Bayesian idea as the basis for analysis of such informal activity.

We have emphasized the *logic* of a determination of similarity, not procedures. For this, our definition lays out the elements of a judgment of exchangeability, and the surrounding discussion indicates how each element is influenced by context and by data analysis. Thus, despite a superficial similarity to procedures like classification and regression trees (CART) (Breiman *et al.* (1984), especially p. 28) and cluster analysis (Hartigan, 1975), we have not simply renamed the parts of existing procedures. Although CART and cluster analysis may be used to search for groupings of the units that may turn out to be contextually meaningful—as part of the iteration towards a final judgment of conditional exchangeability—the actual application of these methods is typically context free, or nearly so.

In addition to its key role in description and data analysis, the notion of homogeneity is also central in many activities of inference and prediction. In these areas, the question is how we may justify the leap from the seen to the unseen. To take a highly stylized case, some (e.g. Kempthorne (1986)) justify such leaps by randomization arguments. Lindley and Novick, in contrast, differentiate (as do researchers like Kish (1965) or

Holland (1988)) between randomization arguments in large and small sample situations. In large sample situations, if

> 'it has been possible to take random samples from a population . . . then complete exchangeability is available'

(Lindley and Novick (1981), p. 56). In small samples,

> 'an allocation found by a random mechanism will always be confounded with some effect: one can do no better than what the personalistic view suggests, use an allocation which You think is unlikely to have important confounding effects'

(Lindley and Novick (1981), p. 52). As with other exchangeability judgments, Lindley and Novick are not specific about how to determine what the important confounding effects are.

This difference reflects an old element of the classical–Bayesian dispute, over whether probabilities computed before observing the data (the randomization probabilities) are relevant after the data are observed. However, the difference disappears and all accept the randomization probabilities as relevant *if* a judgment is added that the seen and unseen units are exchangeable. Conversely, we think all would agree that the randomization argument would be an unsatisfying treatment of a random sample of Americans that happened to contain 90% females. If a judgment of exchangeability is the ultimate rationale for the use of randomization probabilities, that judgment is subject to the same verification as any other such judgment. We see this in techniques like post-stratification: a random sample is taken, concomitants are measured and if the groups of interest differ importantly on some concomitant they are stratified *post hoc* and analysed as if the original sample were stratified. Our position is like Lindley and Novick's: the cogency of randomization as an argument for exchangeability grows with the number of units subjected to the random allocation. By itself, for small samples, it is cold comfort.

To generalize away from the stylized case of randomization, many (all?) leaps from the seen to the unseen involve similar judgments of exchangeability. Ehrenberg (1968) emphasizes the importance of determining the conditions under which such judgments are well founded.

There is a close relation between our formalism and the topic of diagnostic techniques, which attempt to find deviations in data from postulated or fitted models that, in turn, specify the conditions given which the future and past are exchangeable. For example, the statistic DFBETAS (Belsley *et al.*, 1980), which measures the effect on a regression coefficient of deleting a single observation, can readily be expressed in our terms. We emphasize that most writers on diagnostic methods overtly discourage the use of significance tests to assess the importance of a diagnostic indication. Reference values determined from appeal to sampling distributions 'are intended as aids to interpretation and not as foundations for accept–reject rules or P-values' (Cook and Weisberg, 1982). Thus in our terms analysts using such reference values are appealing to significance tests to set the caliper C. In practice other choices for C may be as good or better.

Our formalism will not be relevant when there is an alternative and sufficient foundation for inference, such as a known physical mechanism generating the data. However, judging whether a particular random mechanism is indeed sufficiently well understood may fall under our purview. Study of a physical system such as a roulette

wheel raises questions like those in the repeaters problem: is the system stable?; is the wheel rigged? Study of a pseudorandom number generator raises different questions; some of these questions may respond to mathematical analysis, whereas others (such as whether to use this particular pseudorandom generator on this particular simulation problem) involve extrapolation to the problem at hand from past performance on problems judged to be similar. Such judgments might not be based explicitly on statistical data, but on less quantitative concepts such as 'analogy'; they are then outside the realm of our discussion.

In this paper we have introduced a modest formalism and some intellectual scaffolding to build part of a theory of data analysis and statistical modelling. By laying out the role of data analysis in judgments of exchangeability, we can begin to incorporate into formal statistical theory insights that are currently outside it. Perhaps when we understand things better it will be easier to teach others how to practise the art of statistics.

ACKNOWLEDGEMENTS

Clearly we have been influenced by many earlier researchers. We apologize to any who feel slighted. We thank the many people who have commented on previous drafts of this paper, particularly G. A. Barnard, D. R. Cox, A. P. Dempster, P. Diaconis, A. S. C. Ehrenberg, S. P. Ellis, V. P. Godambe, W. B. Joyce, D. V. Lindley, L. E. Moses, F. Mosteller, T. C. Redman, S. M. Stigler, J. W. Tukey and three referees. We despair of making all our readers happy.

REFERENCES

Azzalini, A. and Bowman, A. W. (1990) A look at some data on the Old Faithful Geyser. *Appl. Statist.*, **39**, 357–365.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics*. New York: Wiley.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Chatfield, C. (1985) The initial examination of data (with discussion). *J. R. Statist. Soc.* A, **148**, 214–253.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*, p. 174. New York: Chapman and Hall.

Cox, D. R. and Snell, E. J. (1981) *Applied Statistics: Principles and Examples*. London: Chapman and Hall.

Denby, L. and Pregibon, D. (1987) An example of the use of graphics in regression. *Am. Statistn*, **41**, 33–38.

Diaconis, P. (1988) Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith). Oxford: Oxford University Press.

Draper, D., Kahn, K. L., Reinisch, E. J., Sherwood, M. J., Carney, M. F., Kosecoff, J., Keeler, E. B., Rogers, W. H., Savitt, H., Allen, H., Wells, K. B., Reboussin, D. and Brook, R. H. (1990) Studying the effects of the DRG-based Prospective Payment System on quality of care. *J. Am. Med. Ass.*, **264**, 1956–1961.

Ehrenberg, A. S. C. (1968) The elements of lawlike relationships. *J. R. Statist. Soc.* A, **131**, 280–302.

Finch, P. D. (1979) Description and analogy in the practice of statistics. *Biometrika*, **66**, 195–208.

de Finetti, B. (1930) Funzione caratteristica di un fenomeno aleatorio. *Mem. R. Acad. Linc.*, **4**, 86–133.

——(1938) Sur la condition d'equivalence partielle. *Act. Sci. Ind.*, **739**. (Translated in R. Jeffrey (ed.) (1980) *Studies in Inductive Logic and Probability II*. Berkeley: University of California Press.)

——(1972) *Probability, Induction and Statistics*. New York: Wiley.

——(1974) *Theory of Probability*. New York: Wiley.

Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.
Freedman, D. A. and Lane, D. (1983) Significance testing in a nonstochastic setting. In *A Festschrift for E. L. Lehmann* (eds P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr). Belmont: Wadsworth.
Good, I. J. (1983a) The philosophy of exploratory data analysis. *Phil. Sci.*, **50**, 283–294.
——(1983b) Random thoughts about randomness. In *Good Thinking*, pp. 83–94. Minneapolis: University of Minnesota Press.
——(1983c) The Bayesian influence, or how to sweep subjectivism under the carpet. In *Good Thinking*, pp. 22–55. Minneapolis: University of Minnesota Press.
Hartigan, J. (1975) *Clustering Algorithms*. New York: Wiley.
Hill, B. M. (1990) A theory of Bayesian data analysis. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (eds S. Geisser, J. S. Hodges, S. J. Press and A. Zellner), pp. 49–73. Amsterdam: North-Holland.
Holland, P. W. (1988) Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology 1988* (ed. C. C. Clogg), pp. 449–484. San Francisco: Jossey-Bass.
Kempthorne, O. (1986) Randomization II. In *Encyclopedia of Statistical Sciences* (eds S. Kotz and N. L. Johnson), vol. 7, pp. 519–524. New York: Wiley.
Kish, L. (1965) *Survey Sampling*. New York: Wiley.
Lindley, D. V. and Novick, M. R. (1981) The role of exchangeability in inference. *Ann. Statist.*, **9**, 45–58.
Mallows, C. L. and Pregibon, D. (1987) Some principles of data analysis. *46th Sess. Int. Statist. Inst.*, invited paper 26.1. The Hague: International Statistical Institute.
Mallows, C. L. and Walley, P. (1980) A theory of data analysis? *Proc. Bus. Econ. Sect. Am. Statist. Ass.*, 8–14.
Nelder, J. A. (1986) Statistics, science and technology. *J. R. Statist. Soc.* A, **149**, 109–121.
Rogers, W. H., Draper, D., Kahn, K. L., Keeler, E. B., Rubenstein, L. V., Kosecoff, J. and Brook, R. H. (1990) Quality of care before and after implementation of the DRG-based Prospective Payment System: a summary of effects. *J. Am. Med. Ass.*, **264**, 1989–1994.
Savage, L. J. (1967) Implications of personal probability for induction. *J. Phil.*, **64**, 593–607.
——(1977) The shifting foundations of statistics. *Logic, Laws, and Life* (ed. R. Colodny), pp. 3–18. Pittsburgh: University of Pittsburgh Press.
Shewhart, W. A. (1939) *Statistical Method from the Viewpoint of Quality Control*. Washington DC: Graduate School of the US Department of Agriculture.
Smith, A. F. M. (1986) Some Bayesian thoughts on modelling and model choice. *Statistician*, **35**, 97–102.
Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading: Addison-Wesley.
——(1986) Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In *The Collected Works of John W. Tukey* (ed. L. W. Jones), vol. III. Monterey: Wadsworth.
Weisberg, S. (1980) *Applied Linear Regression*. New York: Wiley.
——(1985) *Applied Linear Regression*, 2nd edn. New York: Wiley.

DISCUSSION OF THE PAPER BY DRAPER, HODGES, MALLOWS AND PREGIBON

C. Chatfield (University of Bath): I welcome this paper which considers an important, but neglected, area of statistical activity. My experience with real life statistical problems (Chatfield, 1988) suggests that issues such as 'problem formulation' and 'the initial examination of data' are often at least as consequential as statistical inference based on a probability model. This is well illustrated by the Old Faithful data where the detection of the data transcription error is an essential prerequisite to any inferential analysis. It may seem sensible to ask which is the more important, either

(a) spotting the transcription error or
(b) being able to carry out techniques like regression and Markov chain model fitting.

I would prefer to say that (a) and (b) are both important and are *complementary* to one another. Yet the statistical literature and our teaching of statistics concentrate heavily on inferential problems. This may be partly because it is easier to set up statistical inference within a formal, mathematical apparatus, which people like to think is 'objective', although in practice many subjective judgments still need to be made. The imbalance of the literature may also be because informal methods depend so much on context that it is difficult to set up general principles and procedures.