SAMPLING FOR POSSIBILITIES

Draft of article published in *Quality & Quantity, 33*, 185-202, 1999.

Michael Wood & Richard Christy, Portsmouth Business School,

6 October 1998

Abstract

This paper views empirical research as a search for illustrations of interesting possibilities which have occurred, and the exploration of the variety of such possibilities in a sample or universe. This leads to a definition of "illustrative inference" (in contrast to statistical inference), which, we argue, is of considerable importance in many fields of inquiry - ranging from market research and qualitative research in social science, to cosmology. Sometimes, it may be helpful to model illustrative inference quantitatively, so that the size of a sample can be linked to its power (for illustrating possibilities): we outline one model based on probability theory, and another based on a resampling technique.

SAMPLING FOR POSSIBILITIES

Michael Wood & Richard Christy

6 October 1998

Abstract

This paper views empirical research as a search for illustrations of interesting possibilities which have occurred, and the exploration of the variety of such possibilities in a sample or universe. This leads to a definition of "illustrative inference" (in contrast to statistical inference), which, we argue, is of considerable importance in many fields of inquiry - ranging from market research and qualitative research in social science, to cosmology. Sometimes, it may be helpful to model illustrative inference quantitatively, so that the size of a sample can be linked to its power (for illustrating possibilities): we outline one model based on probability theory, and another based on a resampling technique.

Introduction

This paper concerns inferences from data in empirical research in areas where there is substantial uncertainty, so that precise predictions, exact understanding and universal laws are not a realistic expectation.

The usual methods for handling this uncertainty are those of statistical inference. Typically, this involves extrapolating the value of a population parameter, such as a mean or proportion, from a sample statistic, and then using significance levels, Bayesian posterior probabilities or confidence intervals to indicate a level of "confidence" - in a sense depending on the statistical formalism adopted - in these extrapolations. Essentially the same theory can also be used to estimate, in advance, the sample size required to achieve a given level of confidence in a given type of result.

Statistical inferences are clearly not the only type of inference which can be drawn from empirical data. For example, the attempts at universal inferences in some physical sciences (eg Newton's laws of motion, which were assumed to be universally valid), and many "qualitative" inferences in the social sciences, do not fit in any obvious way into the statistical category. Despite this, there is a strong tradition in many areas of the social

sciences that statistical inference is the only legitimate form of inference from empirical data.

This paper argues that there is an important category of inferences which are not statistical, but which are of substantial importance, and are at least as rigorous as statistical inferences. These are inferences about *what* is possible, as distinct from statistical inferences about how prevalent each of the possibilities is. Exploratory research into an unfamiliar new market may seek to uncover and *illustrate* the *variety of possibilities*: ie all the different types of users, uses and contexts of use for a product or service. Research into plant life in a tropical rain-forest may seek to find illustrations of, and catalogue, as much of the diversity of plant life as possible. In each case the prevalence of each of the possibilities may be of minor concern (initially, at least); the important task is to find as many possibilities as possible so that those of particular interest can be investigated further.

We describe this as *illustrative inference*, and argue that it represents an important mode of inference. We go on to give more examples of its use - including its implicit use in much *qualitative* research in social science.

Despite this, books and articles on sampling theory, and computer programs implementing this theory (eg Konijn, 1973; Thompson, 1992; Tryfos, 1996; Maisel and Persell, 1996; Nowack, 1990; nQuery Advisor, 1995) are almost always based on the assumption that samples must be analysed statistically. There are a few alternatives mentioned: the loosely defined, purposive approach adopted by qualitative researchers - particularly with small samples (Miles and Hubermann, 1994), and occasional references to discovery sampling - mainly with reference to auditing (Smith, 1976). Discovery samples are designed to find errors (or other items) with a given probability based on given assumptions, and are, in the terms introduced in the present paper, samples designed for illustrative inferences in a fairly limited domain.

If we are prepared to make a number of assumptions, it is possible to model illustrative inference quantitatively. This can provide estimates of sample sizes necessary to achieve particular goals, or levels of confidence of a given sample having covered a given proportion of the variety in the universe. We discuss two such models: one based on probability theory and another on a resampling approach. These models are the equivalent of the statistical models for estimating "confidence" (in whatever sense) and appropriate sample sizes - but on the assumption that the goal of the research is to draw illustrative inferences. It is important, however, to stress that the concept of illustrative inference may be relevant even

when neither of these models is useful: we can still infer that something is possible from an empirical datum, even without any means of estimating such quantitative parameters as the likelihood of finding such data, or the number of similar data we have failed to find.

Illustrative inference

We will start by defining illustrative inference in general terms, and then give some examples to clarify the definition.

> Suppose that an empirical observation *O* illustrates a general possibility *P* which is relevant to a universe *U*. Then we can infer from *O* that *P* is an *empirically demonstrated possibility* relevant to the universe *U*, and describe this inference as an *illustrative inference*.

The possibility *P* may be constructed on the basis of *O* after the observation (an *inductive* illustrative inference), or it may be derived from a prior *hypothetical possibility*: a possibility on theoretical or conceptual grounds only. Once a hypothetical possibility has been observed it becomes an empirically demonstrated possibility.

Clearly, any observation *O* is likely to illustrate a multitude of different possibilities (see below for an example): those which are considered by a researcher will depend on the researcher's perspective and the motivation for the research. The possibilities which are demonstrated by illustrative inferences are *possibilities of interest to the researcher*.

The use of the term "observation" is not intended to imply an objective, realist interpretation. Observations, like the possibilities they may or may not illustrate, and the universes in which they are relevant, are at least in part constructed by observers with particular perspectives. All we are assuming here is that it is always clear whether a particular observation does illustrate a particular possibility.

The more restricted the universe, *U*, is the *stronger* the inference is in one sense, but the *weaker* it is in another sense. Knowing that there is an instance of a possibility in England is *more* informative than simply knowing it has happened somewhere in the world. On the other hand, the larger U is the *greater* the potential applicability of the possibility: the whole world, as opposed to England only. The choice of U is inevitably arbitrary: extending U too far means that *P* may be too remote a possibility to be interesting.

Illustrative inferences are inferences which can be drawn from data. Clearly the same data may, on occasions be used to draw statistical inferences; many research projects produce

both illustrative and statistical inferences.

Some examples from marketing and management

Illustrative inferences in these fields are very common: they occur whenever an example is cited to demonstrate that something can happen or to explain how it might work. The examples below seem fairly typical.

Penn and Christy (1994) elicited the comment that the Côtes de Duras wine production region in France had "the problem of establishing itself as something other than a cheap Bordeaux / Bergerac alternative" from an open ended question in a questionnaire sent to 10 major UK wine retailers. The observation, *O*, is the gathering and interpretation of this comment; the possibility, *P*, is the perceived problem to which the comment refers; and the universe, *U*, may be that of major UK wine retailers, or customers for the wine. The illustrative inference is the inference that this possibility (ie the perception of this problem) exists in the sense that it has been empirically demonstrated. The response was elicited from an open question, so it was not a possibility which the researchers had hypothesised in advance. The value of inferences such as this should be obvious: they enable people in the wine trade to appreciate (some of) the *variety* of opinions about Côtes de Duras wine. Statistical questions about the frequency of these opinions may (or may not) be of interest, but a simple list of opinions which have been expressed is of value independently of any statistical data.

The researchers' interests are also crucial to the definition of *P*: if the research had concerned handwriting, for example, the meaning of the comment may have been ignored in favour of the handwriting style and a very different set of possibilities would have been derived.

A series of case studies of the use of statistical methods for industrial quality management (Wood and Preece, 1992) illustrated the anticipated possibility of serious misinterpretations of the methods. In addition, this research provided illustrations of more specific possible modes of misinterpretation; these were not anticipated in advance but were derived inductively from the data gathered. In conclusion, Wood and Preece (1992) put forward some general recommendations for avoiding the counterproductive possibilities illustrated by the empirical evidence.

The value of illustrative inference

We can distinguish five senses in which illustrative inferences are useful.

1 A real example of a possibility has been found: this may be useful to explore the possibility in more detail (particularly if further access is possible) or to bring the idea "alive" with a real-life story. This is part of the motivation behind the case studies in Wood and Preece (1992).

2 Illustrative inferences demonstrate that the possibility illustrated is a genuine, empirically demonstrated, possibility.

3 A list of different possibilities which encompasses the variety in a sample may be extremely valuable - eg to see customers' differing attitudes to wine, or businesses' differing uses of statistical methods.

The next two uses relate to attempts to infer general laws.

4 If a possibility is illustrated which contradicts a general law, then, in principle, this general law has been shown to be false. The classic example here is that the observation of a black swan falsifies the "law" that all swans are white. (In practice this falsification process is slightly hazier than this might suggest - see Lakatos, 1981). The final use is the converse of this.

5 If, despite repeated, well-directed attempts, a hypothetical possibility has *not* been empirically illustrated, this provides evidence that it may actually be an impossibility: this is the basis of Popper's (Popper, 1980) description of science as a search for general hypotheses which can withstand serious attempts at falsification. To take a practical example, if, despite repeated attempts, Wood and Preece had failed to find an illustration of a misinterpretation of statistical methods, this would have provided strong support for the proposition that statistical methods are always used correctly.

The contrast with statistical inference

It is worth briefly contrasting illustrative inference with statistical inference. We have not managed to find in the literature a *general* definition of statistical inference, which distinguishes it from other modes of inference, and is independent of any particular approach to statistics (such as the Bayesian school). The following seems to us to summarise the essence of the concept:

> The essential feature of a statistical inference from a sample of data is that the conclusion depends on the *prevalence* or *frequency* of particular types of individual or ranges of measurements found in the sample. Furthermore, methods of statistical inference are typically applied to phenomena that are expected to occur *sometimes*,

rather than always or never.

The values of aggregating statistics such as means or correlation coefficients, or order statistics such as medians, are all dependent on the prevalence of different categories or values in the sample: if the frequencies were different the statistics calculated may change. Statistical inferences typically involve extrapolating patterns found in the data (eg "smokers are more likely to develop lung cancer") to a wider context.

Data gathering for illustrative and statistical inference

If the main purpose of the research is to derive statistical inferences, it is clearly necessary to try to ensure that the sample is as *representative* as possible. (This is one of two requirements for the application of inferential statistics listed by Shvyrkov, 1997.) Statistical results depend - by the definition above - on the frequencies of various categories of individual in the sample, so the sample will clearly be of little use if these frequencies do not correspond reasonably closely to the proportions in the underlying universe. The sample is designed to represent the universe in this sense. Random and stratified sampling are approaches which are normally designed to achieve representative samples in this sense.

If, on the other hand, the main purpose of the research is to derive illustrative inferences, then representativeness in this sense is *not* necessary. In these circumstances *non-representative*, or deliberately "biased", samples may be of more use than representative ones - if the bias is in favour of interesting possibilities - although illustrative inferences can certainly be drawn from representative samples. Qualitative researchers typically take small purposive samples which are designed to uncover illustrations of interesting and relevant possibilities.

Stratification is a potentially powerful tactic for improving both types of samples. With a suitable choice of strata stratified samples are likely to be more representative than simple random samples. Stratification may also be helpful for improving the usefulness of illustrative inferences if strata are chosen to ensure the inclusion in the sample of different categories of individual which are likely to illustrate different types of possibilities. On the other hand, a random sample may throw up new possibilities, precisely because it is not chosen on the basis of the researcher's preconceptions.

Sometimes, data is not collected from discrete units: eg observation studies based on video evidence. In these cases the issues involving representativeness, purposiveness and

stratification are identical, although the practical approaches to designing the sample will obviously differ.

The scope of illustrative inference

The examples above illustrate the way in which illustrative inferences can be useful in marketing and management. In this section we list a few more general possibilities.

Qualitative research in the social sciences (including management and education)

Qualitative research is typically based on "detailed descriptions of situations, events, people, interactions and observed behaviours; direct quotations from people about their experience, attitudes, beliefs and thoughts ...." (Sykes, 1991). Such data is of limited use for statistical inferences since each case is unique; on the other hand it is clearly the basis of illustrative inferences yielding detailed analysis of specific possibilities.

Despite this, qualitative researchers often make inferences from their samples about what "most people" do, or about phenomena which "tend to" cause other phenomena: these are, in effect, statistical inferences (Wood, 1997).

Case study research

A detailed analysis of a single case, or a few cases, is an established method in areas such as management and education (Yin, 1993). Case studies are useful because they demonstrate what is possible - perhaps so that these possibilities can be emulated or avoided elsewhere.

Risk management

Statistical analysis concentrates on the likelihood of occurrence of particular anticipated risks. The prior, more fundamental, problem is that of compiling a list of everything that could go wrong: eg it is clearly important that there should be a systematic search for the possible side-effects of new drugs.

Case-based reasoning

Case-based reasoning is an AI (artificial intelligence) technique for automating reasoning which is based on the idea of finding a similar past case as a model on which to base recommendations (Kolodner 1993). This involves searching for empirical cases which illustrate particular possibilities.

Informal arguments

Many informal arguments rely on illustrative inference. One of us was giving a paper recently about some new possibilities for statistical education. A question was raised, very

reasonably, about whether the ideas proposed had been tried out. The questioner was not asking about a statistical survey, but a simple demonstration that the new possibilities would work in practice. Another illustration of this principle is the present paper: as part of our argument we are providing illustrations of the use of illustrative inference in order to convince the reader that it is a real and useful possibility.

Other possibilities

These are just a few examples. Illustrative inference is also obviously relevant in experimental design (to see what is possible under the different treatments, as opposed to comparing average performance), cosmology (finding illustrations of the theoretical possibility of black holes and extra-terrestrial intelligence), biology (investigating the diversity in a population) and many other fields.

In many of these areas, the instinctive reaction of many people - perhaps especially academics - would be to look for statistical evidence. Clearly statistical evidence has its uses, but we also believe that there is a very important role for illustrative inferences - for the search for possibilities which have been empirically demonstrated. Particularly if we are interesting in change, in improvement, or in exploring new areas, finding new possibilities, and understanding the variety of possibilities, may be much more valuable than finding out how likely or prevalent the known possibilities are under existing conditions.

Quantitative models of illustrative inference

In some (but by no means all) contexts it may be useful to build a quantitative model of illustrative inference to link parameters such as the size of a sample and the number of possibilities it illustrates. This section outlines two such models. Both depend on a number of assumptions. The first group of assumptions (1-5 below) are necessary for both of the models; Assumptions 6-12 are necessary for one of the models only and are listed in the subsections on these models. These assumptions may be justified in two senses: firstly they may be deemed sufficiently realistic, or secondly, while not strictly realistic, they may be useful for exploring possible scenarios on a "what if?" basis.

Assumption 1. We can always decide unambiguously whether a particular observation illustrates a particular possibility.

Assumption 2. The universe is composed of discrete, individual items of a similar kind: people or organisations, for example. Each individual is then "observed" (ie observed,

or interviewed, etc) once, so we can talk loosely of a sample of individuals instead of a sample of observations.

Assumption 3. The possibilities are discrete. In practice, for example, the possibility of misunderstanding statistics may be difficult to distinguish from the possibility of carelessness interpreting graphs. The assumption here is that these are distinct possibilities - an observation may illustrate one, or the other, or neither, or both of them.

Assumption 4. One observation of a possibility provides all the information of interest about that possibility: further observations are of no interest. If, for example, we are interested in the possibility of someone taking a holiday in Greece, all the information of interest is provided by one illustration: no value is added by further illustrations. In practice, this assumption can be made reasonable by defining the possibilities in fairly restricted terms - eg not Greece in general, but perhaps particular Greek resorts.

Assumption 5. If possibilities are derived inductively from observations, this is done in a way which depends only on the observations and the researcher's perspective - which is assumed to be stable. (If, on the other hand, researchers can generate as many possibilities as they choose from a given observation, any attempt to model the number of possibilities illustrated is obviously doomed to failure.)

Illustrative inference: a probability model

How large does a sample need to be to provide an adequate picture of the variety in the universe? One response to this problem is to carry on sampling additional cases until sufficient possibilities have emerged. However, this raises the questions of how "sufficient" can be defined, and of providing an initial estimate of the likely sample size. Just as for statistical sampling, we need a way of balancing the costs of a sample against the information it is likely to provide. In this section we will set up a probability model to approach these questions. We need to make a further four assumptions in addition to the five above.

Assumption 6. The researcher is able to make a statement, before taking the sample, about how many possibilities there are in the universe: we will call this the *variety*, *v*. This may be a "what-if" assumption (suppose there are 10 interesting possibilities), or because there are a number of hypothesised categories of possibilities as formalised, for example, in a "tick the box" question on a questionnaire.

Assumption 7. Assumption 6 holds, and it is also assumed that all of these possibilities are equally prevalent in the universe, that this *prevalence* can be measured as a probability, *p*

(ie the probability of a randomly chosen individual illustrating a particular possibility), and that *p* is known in advance. This is unlikely to be realistic, but it is necessary to keep the model to a manageable level of complexity, and may be a useful basis for a "what-if" analysis (see below).

Ideally we would like to use samples which allow us 100% *confidence* (*c*) of 100% *coverage* (*g*) of all *v* possibilities. In practice this is not possible and compromises need to be made. A simple probability model can be built on the basis of two further assumptions:

Assumption 8. Making assumption 2 (discrete units), the sample is selected at random from the universe.

Assumption 9. The possibilities are spread randomly and independently throughout the universe; they are not, for example, strongly clustered.

The mathematical relationships between the six variables - universe size (*N* - which may be infinite), sample size (*n*), variety (*v*), prevalence (*p*), coverage (*g*) and confidence (*c*) - are outlined the Appendix. This appendix includes some spreadsheet expressions for calculating some of these variables from the others. (Discovery sampling (Smith, 1976) is, in effect, a particular case of this model with $v = 1$ and $g = 100\%$.)

Numerical tables can deal comfortably with three variables (one tabulated, one across the top of the table, and one down the side). Six, however, is not feasible, so a general set of tables summarising the model is not viable. Furthermore, as the reader may care to verify, some of the calculations necessitated by the model are not trivial. This means that the only fully satisfactory implementation of the model is via a computer program. However, Table 1 can be used to answer some specific questions.

TABLE 1 HERE

As an example consider the case of a very simple questionnaire to ask respondents to list the features they would particularly like to see on a battery operated electric car. The initial coding scheme had twelve categories of response, three of which were:

speed, features such as electric windows, "other".

The first of these covers the possibility of a respondent being concerned that the car will go fast enough. Different respondents may have different speeds in mind, but it seemed reasonable (to us) to view this concern over speed as a single possibility. This was not the case for the other two categories above both of which may encompass several, very different, requirements. Accordingly we made a rough estimate of the variety of possibilities in the

universe: between 20 and 40. The next step was to decide on a "cut-off" prevalence level. We made a decision that we were prepared to ignore possibilities which occurred to fewer than 10% of people in the universe. Table 1 can now be used: it shows that if we want to be 95% confident of collecting data on all possibilities with a prevalence above this 10% cut-off level we needed a sample of 57-64 people (depending on the exact value used for the variety). This suggests that our objective would be achieved by a sample of 64 people.

The assumptions on which this conclusion is based are the 9 assumptions above and the values of v and p used. Assumption 6, that the variety is known in advance, is, in practice, not so important as it may seem, because the final conclusion is relatively insensitive to the value used: $v = 20$ giving a sample size of 57, and $v = 40$ giving a sample size of 64. A slightly larger sample, 72, would be sufficient for a variety of 100 possibilities. Assumption 7, about the equal prevalences, is implausible as a description of reality, but useful if viewed as a cut-off mechanism for possibilities whose prevalence is very low. The fact that many of the hypothesised possibilities are almost certain to have prevalences greater than 10% means that actual confidence of achieving 100% coverage is likely to be *more* than 95%. Furthermore, if the estimate of variety is an upper bound on what is likely, the actual variety is likely to be lower, so this is another reason for supposing that the confidence is actually *more* than 95%. (This suggests the idea of using a lower value of c, say 80%, for Table 1; we will however stick with the conventional 95% here.)

Illustrative inference: a resampling model

The model presented in this section makes fewer assumptions about the universe: it is not necessary to make Assumptions 6-9 on which the probability model depends. This model starts from a position of ignorance about how many possibilities there are, and about their prevalence, and about how they are distributed in the universe. It is also not necessary to assume that the sample is selected at random.

Given that we are assuming no prior knowledge of the universe, there is obviously nothing that can be deduced in advance. Accordingly the model can only be applied *after* a given number of units have been sampled and analysed; it will then provide some help with analysing the performance of the sample and the likely benefits of sampling further individuals from the universe. (To avoid confusion, this section follows a different example.)

We assume that we have observed a sample of n individuals and have a list of the possibilities illustrated by each individual. Table 2 gives an example of such a set of data.

(The data refer to comments on software made by respondents to a questionnaire about a statistics course: the first two possibilities are labelled "frustrating" and "easy".)

TABLE 2 HERE

The first individual in Table 2 illustrates two possibilities (3, 8); the second illustrates none, and the third three. However, one of these three possibilities (8) has been illustrated by individual 1, so the additional value (using Assumption 4) from individual 3 is two possibilities (6, 7).

The question we wish to answer is that of deciding how many additional possibilities we are likely to illustrate by sampling further individuals. Clearly, the 13th individual is likely to yield less benefit than the 12th, and the 14th less than the 13th, because the more individuals that have been sampled already the greater the chance that any possibilities illustrated by the new individual have already been illustrated by an earlier individual. At this point it is worth formalising:

Assumption 10. The value of the information derived from a sample can be measured by the number of different possibilities illustrated: the greater this number the greater the value of the information. This is only reasonable to the extent to which all possibilities are of roughly equal value.

Predicting the value, in this sense, of extending a sample is clearly a difficult question because we are making no *a priori* assumptions about the universe. On the other hand there are situations where some extrapolations do seem plausible. Tables 3 and 4 below each show three individuals and three possibilities, but the pattern suggests that the additional value from the next individual is likely to be close to 0 in Table 3 but 1 in Table 4. Clearly this pattern may not continue. Table 4 may result from a universe with 3 possibilities each occurring in 33% of the universe - in which case there are no new possibilities to be illustrated, or it may be the result of 100 possibilities each occurring in 1% of the universe - in which case there are another 97 to be found and sampling further is likely to be very profitable. But despite this, Table 4 is more hopeful than Table 3 in terms of the likelihood of more possibilities emerging from further sampling.

TABLES 3 AND 4 HERE

We have no basis for a probability model like the one in the previous section, and yet there is a sense in which it is meaningful to extrapolate patterns. This suggests the possibility of simply writing down the additional possibilities illustrated by the first individual, the

second individual, the third individual and so on, and then simply extrapolating the sequence. This sequence is inevitably somewhat irregular (2, 0, 2, 7, 1, ... from Table 2), and clearly depends on the arbitrary order in which the individuals were selected. If the individuals were selected in another order the sequence would clearly be quite different (eg if individual 4 was first, the value from the first individual would be 8; if individual 2 were first it would be 0). As no order is any more likely than any other, the obvious thing to do is to take the mean over all possible orders of the 12 individuals. Unfortunately there are just over 479 million such orders, so a reasonable compromise is to simulate say, a few thousand, of these orders chosen at random, and then to take the mean of these. Clearly, the greater the number of orders simulated, the more reliable the answers: in what follows we have used 3000 different orders, which is reasonably reliable in the sense that two runs produce roughly similar results. This procedure relies on:

Assumption 11. Making assumption 2 (discrete units), essentially the same strategy is used to select each individual in the sample and to interpret the resulting observation: otherwise it will not make sense to reorder the sample in this way. This strategy may involve random sampling (ie Assumption 8), or it may be a deliberately biased strategy. This assumption rules out the possibility that the method by which the later individuals are chosen may depend on what is learned from the earlier individuals in the sample.

This simulation procedure is an example of the general approach of *resampling* (Simon, 1992; Noreen, 1989) - which is often useful when analytic models are unrealistic or impossible to derive. Table 5 gives the output from a simple computer program which performs this resampling procedure.

TABLE 5 HERE

In Table 5, the top left entry (2.07) indicates the mean - over 3000 randomly simulated orders of the 12 individuals in the sample - number of possibilities illustrated by the first individual in the sample. The entry below this (1.75) indicates the mean number of additional possibilities illustrated by the second individual (ie the value of the second individual using Assumption 10), and so on.

To go beyond this, to predict the value of sampling a 13th and a 14th individual, it is necessary to extend the pattern in Table 5. Clearly, the function used should be decreasing but should never be negative: there are obviously an infinite number of such functions, but there are perhaps two "obvious" ones: an exponential decay (Equation 1) and an inverse

linear relationship (Equation 2). Both of these are as "simple" as a straight line in that they involve two constants: we will make the value corresponding to individual 1, V_1, the first of these constants in each case.

$$V_n = a^{n-1}V_1 \qquad \text{(Equation 1)}$$

$$V_n = V_1/\{1+(n-1)b\} \qquad \text{(Equation 2)}$$

Equation 1 means that the values attributable to successive individuals form a geometric series: the sum to infinity of such a series is finite. This sum would correspond to the (finite) number of possibilities in the universe (ie the variety in the terminology introduced above). It is easy to prove that Equation 1 is a consequence of the assumptions made to set up the probability model above.

The sum to infinity of the series defined by Equation 2 is infinite. This would be consistent with the assumption that there is no absolute limit to the number of possibilities which can be found. This may be plausible for universes comprising an infinite number of individuals.

FIGURE 1 HERE

Figure 1 shows the results in Table 5, and the extrapolations produced by fitting Equations 1 and 2 using the least squares criterion (implemented by the Optimiser Tool on the spreadsheet Quattro Pro). In this case, Equation 2 appears to fit better; the predicted value of the 13th individual is 0.6 from Equation 1 and 0.7 from Equation 2. The total number of possibilities to which the extrapolations from Equation 1 converge is 19.8.

Alternatively, we could, of course, extrapolate the pattern of the resample results simply by drawing an intuitively derived line on Figure 1. The final, assumption, *Assumption 12*, is that the method of extrapolation used is an "appropriate" one.

There is a danger that the relatively neat pattern of Figure 1 may mislead readers into believing that the predictions made are more definite than they in fact are. We are making an empirical prediction about novel possibilities which have not yet been observed - which is obviously a task for which a high degree of accuracy should not be expected. The assumptions which it has been necessary to make to arrive at the predictions should alert the reader to the fact that they are *very* rough estimates indeed. (In principle, it would be possible to try to construct a *confidence interval* of some kind instead of a point estimate.)

Bearing in mind their likely inaccuracies, these extrapolations are obviously relevant to decisions about extending the sample further. Is the cost - in time and other resources - of

adding another individual to the sample justified by the estimated value in terms of new possibilities illustrated (ie 0.6 - 0.7 possibilities, or about 30% of the mean value of the first individual in a sample)? This is a judgement for the researcher to make, bearing in mind the costs and benefits of the survey.

Conclusions

In this paper we have defined illustrative inference as a type of inference from empirical data, and demonstrated its importance in a wide range of areas. These include qualitative research in the social and management sciences, risk management, case based reasoning and informal arguments. This is not to deny the importance of other forms of inference: eg statistical inference, and inferences about possibilities which are derived from the imagination (eg thought experiments or fiction) or from a conceptual, theoretical or mathematical analysis.

We then proposed two models for relating the size and usefulness of samples for illustrative inference. The first, probability, model entails assumptions about the number, prevalence and (uniform) distribution of possibilities in the universe. The second, resampling, model is only relevant after some data has been collected: this model has the advantage that it requires no *a priori* assumptions about the universe although it does require a number of other assumptions. Both models are useful for answering questions such as "Is it likely to be worthwhile or cost-effective to extend the sample?" and "How large a sample do we need to achieve particular goals?". The assumptions of these models are inevitably rather restrictive: the relevance of the underlying concept of illustrative inference extends beyond the scope of both of these models.

What is the practical value of this analysis? The first, essentially negative point, is that the distinction between the different kinds of inference, and the argument that representative samples are only necessary for statistical inferences, imply that the standard advice on sampling which statisticians are likely to give to qualitative researchers and others, is irrelevant if the research is aiming for illustrative, as opposed to statistical, inferences.

The second point is to reiterate the value of illustrative inferences. It is often more important to gather data on the range of possibilities, without imposing preconceptions on the data, than it is to estimate the current prevalence of these possibilities. Some of these possibilities may then be studied in depth. If the situation is changing fast, so that the past is a poor guide to the future, or if the prevalence of various possibilities in the sample is

influenced by factors which are of little long term significance (current market conditions, source of the sample, etc) a statistical analysis to the effect that 20% fall in this category and 30% in that may be of little interest. On the other hand, the knowledge that there are, say, five market segments, and that detailed illustrative examples of each are available, may be very valuable indeed.

The third point is that sample size and effectiveness issues for illustrative inferences can be analysed. This should reveal if samples are too small to provide a reasonable coverage of the variety of interesting possibilities in the universe.

Appendix: A probability model of illustrative inference

The probability of possibility P_i occurring at least once in the (random) sample, c_i, is

$$c_i = 1 - (1 - p)^n$$

if the universe is infinite or very large. (The symbols used are defined in the section on the probability model above.) The probability, c, of the sample containing at least one of all v possibilities is

$$c = c_i^v = (1 - (1-p)^n)^v$$

This equation gives the confidence that we may have that the sample will achieve 100% coverage if the universe is infinite or large. It depends on the assumptions (8 and 9) that the possibilities are independently distributed in the universe and that the sample is randomly selected. If, on the other hand, there is a tendency for the possibilities to cluster together the formulae will not be accurate. Note also that the confidence here is a probability, whereas the confidence level implicit in a confidence interval, strictly, is not a probability (Sprent, 1981, p. 92).

This equation can easily be rearranged to give an expression for n, p or v (remembering that n and v must take integer values):

$$n = roundup(log(1-c\^(1/v))/log(1-p),0)$$

$$p = 1-(1-c\^(1/v))\^(1/n)$$

$$v = int(log(c)/log(1-(1-p)\^n))$$

These expressions are in the format required by the spreadsheet Excel - except that obviously the variables should be replaced by the appropriate cell references.

If we are interested in less than 100% coverage (g), or the universe is finite (of size N), the corresponding formulae are slightly more complex. The probability (c) of the sample

illustrating at least a given proportion - the *coverage* (*g*) - of these *v* possibilities can be calculated by using the cumulative binomial distribution with *v* "trials" and a probability of success on each trial of c_i. The finite universe means that the expression for c_i becomes:

c sub i = 1 - {N(1-p)}over N {N(1-p)-1} over{N-1}...{N(1-p)-n+1}over{N-n+1}
The resulting expression for *c* in Excel format is
*BINOMDIST(v-g*v,v,FACT(N*(1-p))*FACT(N-n)/(FACT(N*(1-p)-n)*FACT(N)),TRUE)*

References

Kolodner, J. (1993). Case-based reasoning. San Mateo, California: Morgan Kaufmann.

Konijn, H. S. (1973). Statistical theory of sample survey design and analysis. Amsterdam: North-Holland.

Lakatos, I. (1981). Science and pseudo-science. in S. Brown, J. Fauvel, & R. Finnegan (eds), Conceptions of inquiry, (pp. 114- 121). London: Methuen.

Maisel, R., & Persell, C. H. (1996). How sampling works. Thousand Oaks, California: Pine Forge Press.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis (2nd edition). London: Sage.

Noreen, E. W. (1989). Computer intensive methods for testing hypotheses. Chichester: Wiley.

Nowack, K. M. (1990, April). Getting them out and getting them back. Training and Development Journal, 82-85.

nQuery Advisor (1995). Cork, Ireland: Statistical Solutions Ltd.

Penn, J., & Christy, R. (1994). Marketing by smaller wine producers and penetration of new distribution channels. International Journal of Wine marketing, 6(3/4), 20-31.

Popper, K. R. (1980). The logic of scientific discovery. London: Hutchinson.

Sprent, P. (1981). Quick statistics. Harmondsworth: Penguin.

Shvyrkov, V. V. (1997). The new statistical thinking. Quality & Quantity, 31(2), 155-171.

Simon, J. L. (1992). Resampling: the new statistics. Arlington, VA: Resampling Stats, Inc.

Smith, T. M. F. (1976). Statistical sampling for accountants. London: Accountancy Age Books.

Sykes, W. (1991). Taking stock: issues from the literature on validity and reliability in

qualitative research. Journal of the Market Research Society, 33(1), 3-12.

Thompson, S. K. (1992). Sampling. New York: Wiley.

Tryfos, P. (1996). Sampling methods for applied research. New York: Wiley.

Wood, M. (1997). The statistical paradigm in the social sciences - and its alternatives. Paper presented at the Conference: Uncertainty, knowledge and skill, Hasselt, Belgium.

Wood, M., & Preece, D. (1992). Using quality measures: practice, problems and possibilities. International Journal of Quality and Reliability Management, 9(7), 42-53.

Yin, R. K. (1993). Applications of case study research. Newbury Park, CA: Sage.

Table 1: Minimum sample sizes necessary for 95% confidence of achieving 100% coverage (infinite universe)

	Minimum prevalence (%)				
Variety	1	5	10	20	50
1	299	59	29	14	5
2	366	72	35	17	6
3	406	80	39	19	6
4	435	86	42	20	7
5	457	90	44	21	7
6	475	93	46	22	7
7	490	96	47	23	8
8	503	99	48	23	8
9	515	101	50	24	8
10	525	103	51	24	8
11	535	105	51	25	8
12	543	107	52	25	8
13	551	108	53	25	8
14	559	110	54	26	9
15	566	111	54	26	9
20	594	117	57	27	9
30	635	125	61	29	10
40	663	130	64	30	10
50	685	135	66	31	10
60	703	138	68	32	11
70	719	141	69	33	11
80	732	144	70	33	11
90	744	146	71	34	11
100	754	148	72	34	11

Table 2: Dataset 1

	Individual											
Possibility	1	2	3	4	5	6	7	8	9	10	11	12
1 Frustrating	0	0	0	1	0	0	0	0	0	0	0	0
2 Easy	0	0	0	0	0	0	0	0	1	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	1	0	0	1	0
5	0	0	0	1	0	0	0	0	0	0	1	0
6	0	0	1	1	0	0	0	0	1	0	0	0
7	0	0	1	0	0	0	1	0	0	0	0	0
8	1	0	1	0	1	0	1	1	0	0	1	0
9	0	0	0	1	0	0	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	0	0	0	0
11	0	0	0	1	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	1	0	0	0	0	0
13	0	0	0	0	1	0	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0	0	0	0	0

1 indicates that an individual illustrates a possibility, 0 that it does not.

Table 3: Dataset 2

	Individual		
Possibility	1	2	3
1	1	1	1
2	1	1	1
3	1	1	1

1 indicates that an individual illustrates a possibility, 0 that it does not.

Table 4: Dataset 3

	Individual		
Possibility	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

1 indicates that an individual illustrates a possibility, 0 that it does not.

Table 5: Output from resampling procedure based on Dataset 1 (3000 simulated sample orders)

Individual	Value	Total value	% of value from Individual 1
1	2.07	2.07	100%
2	1.75	3.83	85%
3	1.45	5.27	70%
4	1.31	6.59	63%
5	1.20	7.78	58%
6	1.08	8.86	52%
7	1.02	9.88	49%
8	0.91	10.79	44%
9	0.88	11.67	43%
10	0.80	12.47	39%
11	0.77	13.24	37%
12	0.75	13.99	36%

Figure 1: Extrapolations from Table 5