Statistics and Reality

David and Sarah Kerridge

Introduction

W Edwards Deming distinguished between two types of statistical study, which he called "Enumerative" and "Analytic". This sounds theoretical, and it is: but it is also very practical. It affects everything we do, if we base action on figures. Nothing could be more important: it is about the way that statistics relates to reality.

Before explaining any of the details, here is a summary of the main points:

1. Sir Ronald Fisher laid the foundations of mathematical statistics, and Walter Shewhart reexamined the foundations from the point of view of physics. Both these great scientists independently saw that there was a profound difficulty with statistical theory. Unfortunately, neither of them stated it in a way that is easy to understand, so most people have no idea the problem exists. As a result, statistical theory has developed in a way that neither of these founders agreed with.

2. W Edwards Deming, who was famous as a statistician long before he was known as a management expert, knew both Fisher and Shewhart well. He restated the problem clearly, and proposed that we divide statistical problems into two kinds: those which can be solved by mathematical reasoning alone, and (by far the greater number) those in which other knowledge is required. This also tells us ***what*** other knowledge we need.

3. All experienced statisticians know that there are some circumstances in which statistical calculations can be trusted, and others in which they can only give answers of varying degrees of indefiniteness. The theory stated in textbooks does not make it clear why this is. Deming's distinction between "enumerative studies", and "analytic studies" lays the foundation

for a theory of ***using*** statistics, which makes it plain what other knowledge, beyond probability theory, is needed to form a basis for action in the real world.

4. Whether or not a statistician expresses the problem as Deming did, there is a great divide between the hack, who uses methods blindly, and the master statistician who knows why problems arise, and can find ways to avoid them in practice. The experienced statistician knows, as George Box put it, that "All models are wrong: some are useful."

Many people play safe by distrusting ***all*** statistics: this is better than being a hack. But that is losing far too much. A good way to start improving your skill in using statistics is to study Deming's solution to the problem which Fisher and Shewhart recognised.

The sections which follow will try to explain the problem, and Deming's solution, without using any mathematical theory. Those who want it can easily supply it for themselves.

A Simple Example

In the introduction we saw that leading thinkers, including Deming, recognised a problem with statistical theory.

Think of the crime rates for New York. Authorities claim that, as a result of a new policing system, murders have been reduced by 27%. Obviously, if that kind of reduction could be produced in other cities, or other countries, by using the same methods, it would be a great advance. But there are a great many questions to ask before we can act, even if the action is to design an experiment to find out more.

So long as you agree on an operational definition of murder, and stick to it, counting the number of murders in different years is an "enumerative" problem. Interpreting the change is an "analytic" problem.

The easy questions have a nice, tidy, mathematical solution. That 27% must in part depend on chance. If we imagine a set of constant conditions,

which would lead on average to 100 murders, we can, ***on the simplest mathematical model,*** expect the number we actually see to be anything between 70 and 130.

If there were 130 murders one year, and 70 the next, many people would think that there had been a great improvement: but this could be just chance. So the first question we could ask is, "Could the 27% reduction be due to chance?"

That is the least of our problems. The murders may be related, as in a war between drug barons. If so, the model is wrong, since it assumes that murders are independent. Or the methods of counting might have changed from one year to the next (are you counting all suspicious deaths, or only cases solved?). Without knowing about such things we cannot predict from these figures to what will happen next year. So if we want to draw the conclusion that the 27% reduction is a "real" one, that is, one which will continue in the future, we must use knowledge about the problem that is not given by those figures alone.

Even less can we predict accurately what would happen in a different city, or a different country. The causes of crime, or the effect of a change in policing methods, may be completely different.

So this is the context of the distinction between enumerative and analytic uses of statistics. Some things can be determined by calculation alone, others require the use of judgment or knowledge of the subject, others are almost unknowable. To the public at large, it is all just "figures". But if you understand the distinction that Deming made, you can distinguish what is precise, and what is not. Then your actions on the basis of the information available become much more effective. Even more, your actions to get more information improve, because when you understand the sources of uncertainty, you understand how to reduce it.

Sources of Uncertainty

Uncertainty has a way of "creeping in" when we least expect it. We are all familiar with the Red Bead experiment. The "willing workers" take samples of beads from a box in which there are 4000 beads, of which exactly 800 are red, and the rest white. The samples are scooped out with a "paddle", which has 50 holes in it. If the paddle is used carefully, each hole selects one bead.

Nothing could be simpler. The beads are well stirred each time, so any statistics student would immediately calculate the probabilities of any number of red beads in the sample, using standard mathematical theory, based on a random sample from the "population" of beads. Most would use the binomial distribution. We could use the slightly more "accurate" hypergeometric distribution, but the difference is trivial. The average number of red beads comes to 10, and the standard deviation the square root of 8 = 2.83.

Dr. Deming was fond of teasing the audience with this. He would ask them how many red beads there would be on average, in the sample of 50. When they said 10, would trumpet "Wrong!"

The theory is wrong, and noticeably so. In "The New Economics" (page 164) he records the results of the red bead experiment over the years. The average number of red beads changes with the paddle he used, and ranges from 9.2 to 9.6 with one set of beads, and 11.3 with a different set of beads and paddle. This difference is far too big to be due to chance. As WED says, differences like that would cost you a lot of money if you relied on the theoretical calculation. In the years before he became famous as a management theorist, he specialised in accurate sampling, and would have soon lost his reputation if he made mistakes like that.

The reason for the failure of theory here is that no mechanical process (except perhaps on the atomic scale, as Shewhart points out) is really random, or even completely stable. In this case there must be slight differences between the red an white beads. This might be a difference in size, or smoothness, or even of the tendency to pick up static electricity.

The only process that behaves ***exactly*** according to the mathematical theory is one governed by tables of random numbers, as published in statistical books. The difference from "randomness" in other cases may be small, and have no practical importance, or it may be large. Either you must test it, or estimate the extra uncertainty from subject knowledge or from experience of that kind of process.

If we number the beads from 1 to 4000, and take samples of 50 using random number tables, we can estimate the number of red beads to any desired degree of accuracy. Thus the typical problems in which the mathematical theory is

perfect are “enumerative”, that is, they correspond to counting. But if, through ignorance or for convenience, we scoop up a sample with a paddle, the problem is no longer enumerative, but “analytic” in Deming’s terminology.

The distinction does not depend on the type of data we use, or the action we will take, but on the exactness of the reasoning by which we make our estimate.

In the case of the read bead experiment, the calculation is not very far out. For some purposes the mathematical approximation may be good enough. In other cases the differences from the mathematical theory may be very large. This can easily lead to wrong conclusions, if we are unaware of the problem. Many experiments on Extra-sensory Perception, for example, use mechanical sampling, and “detect” small effects - of much the same size as the “red bead” effect.

Approximation to Reality

The example of the read beads helps us see what the deep problems of science are. The simple mathematical model, in which we treat the beads mechanically scooped up as if they were a random sample, is near enough for some purposes. But it neither gives the correct average, as Dr. Deming’s records show, nor (though it would need very large sample to prove this) the right variation.

Just as the numbers of beads in the box do not tell us what we will get in samples taken with the paddle, so the samples cannot tell us exactly what is in the box. If we really need to know ***exactly*** what the proportion of red beads is, we must not rely on the paddle: we must take all the beads out of the box. And then either count them, or use very strict sampling using random number tables.

In science, or medicine, we often want to know things that sound like the proportion of red beads in the box. We may want to know what proportion of people will be cured by a new drug, for example. ***These things really matter.*** But all we can do is to count or measure samples. Only in rare cases can we “take all the beads out of the box”: usually there is not even a definite box.

The philosopher Plato pictured the problem this way. We are like people inside a cave, watching shadows on the cave wall, and trying to work out what is “really” there, outside the cave. In our case, the samples that we take are the

shadows. There must be “something” outside, that we call reality.

It is true that if different people measure the same things, they do not get the same answer. But they get ***approximately*** the same answer: and better operational definitions, or better measurement techniques, make the agreement ever closer. So we are not as helpless as the cave-dwellers in Plato’s fable: we can improve and focus the shadows to make the picture clearer.

Sometimes a dramatic improvement in the way we observe is almost the same as taking the beads out of the box, or escaping from the cave. For example, at one time all that was known about genetics came from studying samples of mice, or peas, or fruit-flies, and the results of different matings. From these samples biologists tried to work out how genes are arranged on chromosomes. Then first the electron microscope was invented, and now DNA sequencing. The arrangements are now visible, instead of guessed at: and the results inferred by statistical methods are seen to be correct.

The results were reliable largely because R A Fisher, who created the statistical tools for this, and many other fields, knew exactly what he was doing. But although he saw that mathematics alone was not enough, he did not explain this in a way that others could follow.

Contributions of Fisher and Shewhart

In the previous posts, we have seen that there are problems that mathematics alone cannot solve. The first to say this clearly was R A Fisher.

He drew a distinction between ***scientific*** problems, and ***decision making*** problems. For example, suppose that we want to know which of two antibiotics is better in treating typhoid. We cannot take a random sample of all the people who will be treated in the future: there is no "bead-box" of people waiting to be sampled, because we do not know who will get typhoid in the future.

R A Fisher still used the mathematics of random sampling, but he made it plain that it was a different kind of problem. He described it as sampling from an ***imaginary*** population. His actual words were "a hypothetical infinite population" but that is just another way of saying imaginary. The practical difference as he saw it, is that we must not rely on what happens in any one experiment: we must repeat the experiment under as many different circumstances as we can. If the results under different circumstances are consistent, believe them. If they disagree, think again.

This is in very strong contrast with what is normally taught in most statistics textbooks, where it describes the problem as one of "accepting" or "rejecting" hypotheses. No scientist ever does this: science works by accumulating evidence of different kinds, not decision making, as Fisher emphasised.

The next to throw light on the problem was Walter Shewhart. He stated the difference by means of an example (The Economic page 55):

"You go to your tailor for a suit of clothes and the first thing that he does is to make some measurements; you go to your physician because you are ill and the first thing he does is to make some measurements. The objects of making measurements in these two cases are different. They typify the two general objects of making measurements. They are:

(a) To obtain quantitative information

(b) To obtain a causal explanation of observed phenomena"

This is very like Fisher's distinction between scientific and decision making problems. The difference is that Shewhart relates it to the ***action to be taken***, rather than the kind of problem.

But the most important contribution of Walter Shewhart was to point out how completely the analogy with random sampling breaks down. Most physical processes are not under statistical control. So no theory based on sampling can give the correct answer. It may be a guide, but no more.

These comments on the contributions of R A Fisher and Walter Shewhart are drawn from a paper by Beth Blankenship and Pete Petersen, to appear in the Journal of Management History.

Prediction

As we have seen, the distinction between enumerative and analytic studies emerged very slowly. R A Fisher recognised that, in scientific problems, we must look for repeatability over many different populations. Walter Shewhart added the new concept of statistical control, which defines repeatability over time. His thinking relates to sampling from a process, rather than a population.

Most mathematical statisticians state statistical problems in terms of repeated sampling from the ***same*** population. This leads to a very simple mathematical theory, but does not relate to the real needs of the statistical user. You ***cannot*** take repeated samples from the exactly the same population, except in the rare cases, like the red bead experiment. There the total set of beads remains exactly the same, because you put the beads you have sampled back into the box.

This sounds like a rather rarified and theoretical argument, but it has very important practical consequences. Suppose that we compare two antibiotics in the treatment of some infection. We conclude that one did better in our tests. How does that help us?

Suppose that all our testing was done in one hospital in New York in 1993. But we may want to use the antibiotic in Africa in 1997. It is quite possible that the best antibiotic in New York is not the same as the best in a refugee camp in Zaire. In New York the strains of bacteria may be different: and the problems of transport and storage really are different. If the antibiotic is freshly made and stored in efficient refrigerators, it may be excellent. It may not work at all if transported to a camp with poor storage facilities.

And even if the same antibiotic works in both places, how long will it go on working? This will depend on how carefully it is used, and how quickly resist strains of bacteria build up.

This may seem an extreme case, and it is. But in every application of statistics we have to decide how far we can trust results obtained at one time, and under one set of circumstances, as a guide to what will happen at some other time, and under new circumstances. Statistical theory, as it is stated in most textbooks, simply analyses what would happen if we took repeated, strictly random samples, from the same population, under circumstances in which nothing changes with time.

This does tell us something. It tells us what would happen under the most favourable imaginable circumstances. If a method would be useless under these circumstances, it would be hopeless in all real problems. But the estimate of uncertainty we get is a gross and misleading underestimate of the real uncertainty.

In almost all applications we do not want a description of the past, but a prediction of the future. And for this we must rely on theoretical knowledge of the subject, at least as much as on the theory of probability.

Implications for Design of Studies

An enumerative study always focuses on the actual state of something at one point in the past. An analytic study usually focuses on predicting the results of action in the future, in circumstances we cannot fully know. This predictive way of thinking is fundamental to Dr. Deming's theory of statistics. A key paper is "On probability as a basis for action" Journal of the American Statistical Association volume 29, 1975: pages 146-152.

Most of the mathematical theory of statistics concentrates on "finding the best estimate" of the "true value of a parameter". The methods are optimised to precisely this end. But when we want to predict the future, we should ***design*** our studies differently: to take account of real uncertainties that at the outset we can only guess at. What is best for one purpose will seldom be best for another.

We use subject knowledge and previous experience to see whether we need to allocate more effort to seeing if the process is stable over time, or to seeing whether it changes noticeably with other factors. Then we design a study to give the most useful and reliable results.

The effectiveness of a drug may depend on the age of the patient, or previous treatment, or the stage of the disease. Ideally we want one treatment that works well in all foreseeable circumstances, but we may not be able to get it. Once we recognise that the aim of the study is to ***predict,*** we can see what range of possibilities are most important. We not only design studies to cover a wide range of circumstances, but to make the "inference gap" as small as possible.

By the inference gap we mean the gap between the circumstances under which the observations were collected, and the circumstances in which the treatment will be used. This gap has to be bridged by assumptions, based on theoretical medical knowledge, about the importance of the differences. So if we know that a treatment will be used mainly in refugee camps, we should try to get evidence from a similar circumstances as we can find. Well planned and designed experiments are rarely possible in emergencies, so the gap may be quite large. Most testing is in fact done in hospitals.

The examples so far have come mainly from medicine. This is because the need for reliable ***prediction*** is so obvious. And differences between circumstances where samples are collected, and results used, can be so great. So studies designed by the best medical statisticians already take account of these problems. That it should be so is all the more remarkable when the theory, as stated, still uses the old concepts, though the influence of R A Fisher is very strong in medical statistics.

In management the need for prediction is at least as great as in medicine. "All management is prediction." And the "inference gap" may be huge. So there are many opportunities for those who understand the difference between enumerative and analytic studies to make a great contribution. There are equally great

opportunities for hacks to sell slick, mathematically impressive, but misleading techniques. These will lead to wrong actions.

Unknown and Unknowable

There remain a number of practical problems. We have already discussed the way that an analytic study should be designed: but there are also questions about the way we analyse the results, and present them.

There are no difficulties with an enumerative study. If we use methods designed to estimate the "true value of the mean", we must remember to replace "true value" of the mean by the value operationally defined to be the "true" value. The rest of the methodology is satisfactory, if we use pure random sampling.

We often use random sampling in analytic studies, but it is not the same as that in an enumerative study. For example, we may take a group of patients who attend a particular clinic, and suffer from arthritis. We then choose at random, or in some complicated way involving random numbers, who is to get which treatment. But the resulting sample is ***not*** a random sample of the patients who will be treated in the future at that same clinic. Still less are they a random sample of the patients who will be treated in any other clinic.

In fact the patients who will be treated in the future will depend on choices that we and others have not yet made. And those choices will depend on the results of the study we are doing, and on studies by other people that may be carried out in the future.

So with an analytic study, there are two distinct sources of uncertainty:

1. Uncertainty due to sampling, just as in an enumerative study. This can be expressed numerically, by standard statistical theory.

2. Uncertainty due to the fact that we are predicting what will happen at some time in the future, and to some group that is different from our original sample. This uncertainty is ***"unknown and unknowable"***. We rarely know how the results we produce will be used, and so all we can do is to warn the potential user of the range of uncertainties which will affect different actions.

This is rarely done. Even when statisticians are fully aware of this extra uncertainty, they do not know how to express it. But the uncertainties of this kind will in most circumstances be an order of magnitude greater than the uncertainty due to sampling.

There may be examples in some subjects, such as the physical sciences, in which the uncertainties of this second kind may be smaller than the sampling uncertainty. There may be others, perhaps including engineering, in which the circumstances in which the product is likely to be used are well understood. This may justify a single numerical estimate of the added uncertainty. Every case has to be considered on its merits. In management, as in medicine, the second type of uncertainty is usually large and difficult to quantify, especially since things change rapidly. But we must certainly know that it is there.

Many people feel uncomfortable with uncertainty that is, to a great extent, "unknown and unknowable". Of course, we would rather be certain if we can. But it is very dangerous to pretend to be more certain than we are. Suppose that we have done a study on the likely reaction of consumers to a new product. We always have choices: get more information, prepare for the worst, or just wait and see. Or we may decide that the risks too great, and abandon the project.

False certainty leads to wrong choices. If we present results as more certain than they are, events may prove us wrong. Then, like the boy who kept shouting "Wolf!" we will not be believed when we ***can*** calculate chances precisely.

Aim and Method

Does all this matter? Clearly Dr Deming thought it did. He held special seminars on the topic every year until 1992. And this is what he says on page 100 of "The New Economics" (page 103 in the first edition):

"Use of data requires knowledge about the different sources of uncertainty. Measurement is a process. Is the system of measurement stable or unstable? Use of data requires also understanding of the distinction between enumerative studies and analytic problems."

And a little further on:

"The interpretation of results of a test or experiment is something else. It is prediction that a specific change in a process or procedure will be a wise choice, or that no change would be better. Either way the choice is prediction. This is known as an analytic problem, or a problem of inference, prediction."

Another type of analytic problem is that of exploration, or ***discovering*** a problem. This is noted on page 182 of "Out of the Crisis". Confusion between different types of problem leads to waste and inefficiency. "What is your aim?" is always the first question. Here are some different aims, calling for different methods.

Aim 1: Describe accurately the state of things at one point in time and place.

Method: Define precisely the population to be studied, and use very exact random sampling.

Aim 2: Discover problems and possibilities, to form a new theory.

Method: Look for interesting groups, where new ideas will be obvious. These may be focus groups, rather than random samples. The accuracy and rigour required in the first case is wasted. But this assumes that the possibilities discovered will be tested by other means, before making any prediction.

Aim 3: Predict the future, to test a general theory.

Method: Study extreme and atypical samples, with great rigour and accuracy.

Aim 4: Predict the future, to help management.

Method: Get samples as close as possible to the foreseeable range of circumstances in which the prediction will be used in practice.

Aim 5: Change the future, to make it more predictable.

Method: Use SPC to remove special causes, and experiment using the PDSA cycle to reduce common cause variation.

The first case is enumerative, all the rest are analytic. Do you know of any statistics textbook that makes these obviously necessary distinctions?

I have seen dozens of statistical studies that were virtually useless, because they did not make these distinctions, and were not well fitted to any one aim.

What Now?

Why have these obvious ideas been overlooked for so long? As we have seen, eminent statisticians from Fisher to Deming have done their best. But there has been an interesting historical pattern.

R A Fisher developed most of the useful statistical techniques. But then the mathematical statisticians Jerzy Neyman and E S Pearson wrote what they believed was a clarification of Fisher's ideas. In so doing they eliminated every uncertainty that could not be expressed mathematically. The result was much easier to teach, but it ignored the real problems that Fisher, however imperfectly, was trying to solve. What Neyman and Pearson did was helpful in enumerative problems, but not analytic.

Walter Shewhart introduced the revolutionary idea of statistical control. This did not assume the existence of mathematical distributions, but showed how to change processes, bringing them under control. This makes it possible (among other things) to use the mathematical theory as an ***approximation*** to reality. Just as in the case of Fisher's theory, the mathematical statisticians who came later thought they could improve on Shewhart's methods by eliminating the non-mathematical part of the theory, and reformulating it in terms of "exact" probability distributions.

In both cases a very similar misunderstanding arose. An unrealistic theory, which offers a cut-and dried solution to every problem is easier to teach than realistic theory, which must deal with things that cannot be measured. So the bad theory was, and is, widely taught. There has been a similar problem in Dr. Deming's management theory: there is great resistance to the "unknown and unknowable". People ***want*** tidy solutions, even when they do not exist.

What of the future? One way might be to develop a general theory which

includes uncertainty that cannot be expressed in numbers. There is a theory of geometry (topology) that is still valid when nothing can be measured. In the same way we could have a theory of uncertainty that is more general than probability theory, because it includes all types of uncertainty, whether measurable or not.

We have one such theory, the Bayesian theory of statistics, which expresses all kinds of uncertainty in numbers, regardless of whether the uncertainty can be measured exactly, or just guessed at. The mathematical theory called "Chaos Theory" also deals with non-random variation, but it seems too limited to provide a theory of prediction and action.

Most statisticians used a Bayesian approach until R A Fisher and others rejected it. They rejected it because they believed that we must distinguish between what can be measured exactly, and what depends on opinion. But the result, in most statistics courses, has been a theory in which the unmeasured uncertainty has just been ignored.

What matters is that the different kinds of uncertainty will not go away. We must deal with them in practice, however we express them. Fisher and Shewhart laid the foundations for a new, all-embracing theory. Deming made the essential step of bringing the apparently conflicting insights together. But there may still be more to be done to clarify these ideas before they will be widely used.

Based on a series of postings to the Deming Electronic Network. The sections correspond to the ten original messages.