# THE FORECASTING VOICE: A UNIFIED APPROACH TO TEACHING STATISTICS

Alan McLean, Monash University, Australia

*The typical text in business statistics claims, in one way or another, that the use of statistics is to aid in decision making. In practice, very few texts do much to illustrate this, concentrating on a conventional development of techniques. Putting this reservation aside: how does statistical analysis 'aid in decision making'?*

*Statistical analysis provides forecasts of what is likely to happen, based on probability models, with estimates of how good the forecasts are likely to be. Decisions can then be made on the basis of the forecasts. A statistical analysis provides the basis for choosing an appropriate model. Statistics is not about finding 'truth' but about finding useful ways of describing 'reality'.*

*It follows that statistics teaching at any level should be intimately concerned with models and forecasting.*

## INTRODUCTION

A typical introductory statistics text is made up of five parts. A coverage of descriptive statistics techniques is followed by probability theory, with the emphasis on the standard distributions (particularly, of course, the normal). Third is a development of the basic concepts of inference, with some work done on the assumptions underlying the development of sampling distributions. Fourth is a selection of work on comparison of means, regression, ANOVA and cross tabulation. Finally, there may be some chapters on other topics.

The treatment is directed toward the student understanding how to carry out the techniques, even when the development is quite nonmathematical. Authors use a variety of examples to illustrate, usually more or less piecemeal, how techniques are applied. Rarely does an author present a unified approach to the application of statistical methods. It is toward the development of such a unified approach that this paper is directed.

## THE THEME

It is my contention that the ultimate aim of any statistical analysis is to forecast, and that this determines which techniques apply in particular circumstances. The word 'forecast' is used here not in the restricted sense of extrapolating a time series, but in the more general everyday sense of predicting the future in a situation of uncertainty. All forecasts are based on the use of probabilistic models. The type of forecast is determined

by the choice of model, and the quality of a forecast is determined by the validity or otherwise of the model.

The idea that statistics is all about making forecasts based on probabilistic models of 'reality' provides a unified approach to the subject. In the literary sense, it provides a consistent authorial 'voice'.

## PROBABILITY

The core idea of probability is that of the *probability distribution of a random variable*. The two ideas expressed in this phrase are inextricably entangled. If one identifies a set of possible outcomes for an 'experiment', a listing of these outcomes, with their associated probabilities, forms a probability distribution. The outcomes may be identified by name - in which case we would speak of a nominal random variable - or numerically - in which case we would speak of a numeric random variable.

## PROBABILITY MODELS

In specifying a probability distribution, we specify a *model for the future*: these are the outcomes which we accept as possible, and these probabilities measure, according to the model, how likely each outcome is to occur. When we say that the probability of a coin coming up heads is 0.5, we are not expressing some absolute truth. We are describing a model which experience has shown to be reasonably valid.

The concept of a probability model becomes more explicit when we introduce the standard distributions. Each of these is a model which can be applied to any particular situation with greater or less validity. Our students learn the conditions, for example, under which a binomial model is applicable.

For the normal model we are likely to rely on 'experience shows that this type of variable has a normal distribution…..'. In the real world however there is no such thing as a normal distribution, nor any other of the standard distributions. Each is a mathematical fiction. We can say that a variable is *approximately* normal, by which we would mean that past experience shows that the normal model is a useful, if partial, description of reality. Indeed, with this example, our students are likely to force us into talking of models, when they ask about behaviour at infinity, and we answer that normality 'is a good model near the centre of the distribution'.

## SOURCES OF MODELS

The typical textbook introduction to probability identifies a spectrum of ways in which probabilities are arrived at, with ‘subjective probability’ at one end, and ‘objective probability’ based on long run proportions and exemplified in games of chance at the other. The middle part of the spectrum is occupied by ‘frequentist probability’, based on short run proportions. Rarely is it pointed out, first, that there is always some element of subjectivity – probability models are always subjective – and second, that in using short run proportions we are inferring on the basis of sample data.

With subjective probabilities it is clear that a model is being formulated, and whether or not it is valid can be discussed. With a gambling game, it has to be emphasised that the game is *fair*; that is, the die is balanced, the cards are well shuffled. Then it can reasonably be assumed that all outcomes are equally likely – and hey presto! we have our model.

Despite the rise of the casino culture, the importance of this equal probability model is that it underlies random sampling. If we have a variable measured on a population, the ‘probability distribution of the variable’ refers to the probability of each value when a member of the population is selected randomly.

If we know the probability distribution, having measured the variable for all members of the population, we can use this, perhaps with some simplification through grouping of data, as an ‘empirical model’. Alternatively, we can approximate the empirical model by some standard model. Conceptually the empirical model is simpler, but is likely to be computationally more intense. Further, the probability distribution is likely to change marginally in the future, when the errors in using the standard model may be no more than the errors due to change.

Whatever the source of the model, there is always a subjective element in its choice. And whether or not the choice was good is eventually determined by whether or not it works!

## PROBABILITY AND FORECASTING

A probability distribution is used to predict what will happen when the experiment is carried out. The prediction may simply take the form of ‘It will come down heads with probability 0.5’, but in practical applications we are likely to have to put our money where out mouth is, so to speak. Neglecting the obvious example of gambling games we for example forecast demand, on the basis of which we will determine our inventory; we

forecast how many people will accept an invitation, which will affect our catering arrangements; we forecast rainfall, which will affect our crop planting or our holiday plans.

## THE 'BEST' FORECAST

What is the 'best' forecast depends on the type of variable. For a nominal variable, it is reasonable to define the 'best' forecast as the outcome which is most likely to eventuate; that is, the *mode*. The criterion of 'best' is to minimise the probability of error. This has nothing to do with .centrality. - the concept is meaningless with a nominal variable. Note that there may not be a single 'best forecast'. For a nominal variable, it is necessary therefore to know the probability of each outcome, in order to determine the mode. For a variable defined on a population, under random sampling, we therefore need to know the proportion of the population for each value of the variable.

With a numeric variable, if the number of different outcomes is small, we can also use the mode. However, the numeric scale gives the option of using the concept of *error*, and if there are many outcomes, we have to take this option. The best forecast can then be defined as the one which in some way minimises the likely error.

The almost invariable choice is to minimise both the absolute expected error and the expected squared error; this is achieved by using the *mean* as the forecast. If the mean is used as the forecast, the absolute expected error indeed is zero, ensuring that the forecast has no bias built in, and the expected squared error is just the variance. In comparing forecasts across variables, models or populations, if in each case the mean is used, the forecasts will have zero expected error. The best forecast will then be the one with the smallest variance.

The reason for the choice of squared errors is usually explained in terms of 'getting rid of the negative signs'; that this is better than using absolute errors because it is 'mathematically more amenable'. These reasons are plausible, but the real reason is that our mathematics is based on models in which variables are assumed to be mutually orthogonal. There is no reason, in principle, why some other criterion, or *loss function*, is not used; for example, the expected absolute error.

## QUALITY OF FORECASTS

With a nominal variable, the quality of the forecast is automatically specified by giving the probability of it being correct. This is for most people a meaningful way of

expressing the result. ‘I predict that it will rain tomorrow. The probability of my being correct is 0.8.’

Note that this is not the same as: ‘The probability of it raining tomorrow is 0.8.’ In this case, I am forecasting the probability of rain.

For a numeric variable, the quality is specified by giving the expected squared error. This is generally not meaningful in practical terms. However, we can do something which corresponds to the nominal case. We can specify a *prediction interval* – an interval in which the result will lie with specified probability.

Prediction intervals can be calculated for any distribution, but this is rarely done in the textbooks; although this type of calculation is typically done as an example for the normal distribution, the concept of a prediction interval is not developed.

## DESCRIPTIVE STATISTICS AND PROBABILITY

In descriptive statistics for snapshot data we distil information from a set of data from a population or sample by obtaining frequency distributions, usually with grouping of data to remove some of the noise, and by calculating summary statistics. This enables us to ‘describe the sample or population’. The results ‘describe the population’ in two senses. First, in a static sense: ‘20 percent of the population use Noxia soap’. The second sense is dynamic: ‘the probability that a randomly chosen person from the population uses Noxia is 0.20’, or: ‘the expected proportion of a random sample of people from the population who use Noxia is 20%’. The first usage always implies the second. That is, when we describe a population, we describe the probability distribution.

If the data are collected on the population we directly observe the probability distribution – the empirical model - for the variable being measured. This is then used to make probability statements, as discussed previously.

If the data are collected on a sample, we infer the probability distribution for the population from the sample results. At the simplest level the assumption is made that the sample results describe the population very closely, so forecasts are based simply on the sample results: ‘20% of the sample use Noxia, so the probability that a person uses Noxia is 20%.’ This is not restricted to newspaper reports of surveys: in introductory probability using the ‘frequentist’ approach we do precisely this.

More rigorously, we use the sample statistics as estimates for the population, but we also estimate how well the sample describes the population. To do this we obtain

*confidence intervals* for the population parameters. For a nominal variable we must estimate the probability of each outcome, since the best forecast is the mode, so for a snapshot of a population under random selection, this probability is the population proportion. In order to forecast a numeric value we must estimate the population mean, since this is the best forecast. We also have to estimate the variability, as measured by the standard deviation, since this is used in computing the prediction interval.

The sample mean is the best estimate for the population mean on the criteria of unbiasedness and minimum variance. These criteria correspond to those for 'best forecast'.

## DESCRIPTIVE STATISTICS AND FORECASTING

To forecast an individual observation – to provide a prediction interval for an individual value - based on sample data for a numeric variable, use the sample mean to estimate the population mean, then use it to provide the forecast.

The best forecast for a numeric variable, under the criterion of zero absolute expected error and minimum expected squared error, is $\mu$. If $\mu$ is 'estimated' by some arbitrarily chosen number we can test it by calculating the absolute mean error and mean squared error for the sample values. The sample mean is then the best forecast under the same criteria, tested on the sample, as the population mean, tested on the population. To obtain a prediction interval for this forecast we combine the uncertainty in the estimate of the mean with the variability assumed in the model for $X$.

Prediction intervals are frequently introduced in the 'typical introductory text' under simple linear regression, when it is required to predict the value of $Y$ for a given value of $x$: it is common to find both a confidence interval on the mean and a prediction interval on the individual value. It is understandable that prediction should be raised in regression – the reason for a regression analysis is to predict the dependent variable. On the other hand, analysis of variance and cross tabulation are seen as extensions of basic statistics, in which subpopulations are compared in terms of a numeric or nominal variable respectively, and the predictive usage ignored.

## CONCLUDING REMARKS

I have argued in this paper that the underlying purpose, often implicit rather than explicit, of every statistical analysis is to forecast future values of a variable. These

forecasts are based on probability models for the variables, in turn based on sample data. Using natural criteria, the ‘best forecasts’ for nominal and numeric variables are respectively the mode and mean. For a numeric variable, the quality of a forecast is specified using a prediction interval.

If it is accepted that this view of the underlying thrust of statistics is correct, then it is reasonable that texts should reflect this view. The predictive use of probability models, and the use of prediction intervals, should be emphasised.

In a future paper I plan to discuss further the use of prediction concepts in multiple variable models, particularly with nominal variables. For example, is a relationship which ‘exists’ in the sense of being significant, but which has very weak predictive power, at all useful?