Batanero, C. (Ed.), Training Researchers in the Use of Statistics, 87-101.

Printed in Granada, Spain

ALAN MCLEAN

STATISTICS ON THE CATWALK: THE IMPORTANCE OF MODELS IN TRAINING RESEARCHERS IN STATISTICS

This paper emphasises the pervasive role of probability models in statistics, and the importance of the role of prediction in statistics. I argue that all thinking, including everyday decision making, is based on the use of models: theories, stereotypes, metaphors, stories, myths, equations, diagrams, blueprints. Scientific thinking, including statistical thinking, is not different from 'everyday' thinking, but is a formalisation of it. Just as we 'learn from experience', a scientific theory is tested against observed data.

Statistics is a body of techniques for developing and assessing models, particularly those involving uncertainty. This modelling process takes place at all levels of a statistical analysis, not only in 'model selection'. Examples are given to illustrate these processes. Particular attention is given to the role of hypothesis testing, showing how it is a form of model selection between two models, one of which is 'privileged'. Researchers, if they are to understand the role of statistics in scientific research, must understand the role of models in science generally and in statistics in particular.

1. INTRODUCTION

Practising statisticians are very familiar with the concept and use of statistical models, but models appear very little, if at all, in elementary courses or in the texts on which those are based. This can reasonably be considered a serious deficiency, because all statistical work involves the use of models, just as all scientific theorising does.

While one can argue for a very elementary introductory course which is data driven, which barely mentions probability, and is oriented toward everyday life applications, researchers have a strong need for a more thorough and sophisticated understanding of the role of statistical methods in their disciplines, and the role of models in statistics. The statistical training of researchers in many fields, however, is frequently little more than a traditional trek through elementary techniques, so they are likely to be only tentatively aware of the pervasiveness of models.

This paper represents an extension of my thinking. The core of the predictive or forecasting approach to statistics (McLean, 1998) is that of the pervasive role of probability models in statistics, and the importance of the role of prediction in statistics. In the paper presented at the 52nd ISI Session (McLean, 1999b) I considered some aspects of hypothesis testing. In this paper I extend this thinking in two ways: first, a more complete consideration of the nature of hypothesis testing, and second, consideration of the relationship between everyday thought processes and those involved in statistics, and science in general.

My reading connected with the well-known hypothesis testing controversy (including Morrison & Henkel, 1970; Henkel, 1976; Gingerenzer, 1993; Harlow, Mulaik, & Steiger, 1997; Batanero, 1999; and Ito, 1999) suggests that there is

considerable confusion among researchers as to the nature of testing. Discussions in email mailing lists reinforce this conclusion. The recognition of the role of probabilistic models in statistics enables a simpler, more consistent view of the nature of statistics, and of the nature and role of hypothesis testing. In this paper I present some thoughts in that direction.

Statistical and scientific thinking do not differ from everyday thinking to the extent that may commonly be thought. In this paper I discuss the ways in which all these modes of thinking involve the use of models, and consider in a very general way some of the characteristics of models. The major part of the paper deals with the way models apply in statistics at every level, from basic descriptive statistics on. As mentioned above, a substantial part of this section deals with hypothesis testing.

2. MODELS

Models constitute our knowledge of the universe

There are a great many philosophical arguments about the nature of 'reality' and our perceptions of it, and the nature of the 'scientific method', but it appears to me that the most useful simple approach is as follows.

Science, as a body of knowledge, comprises two types of information structure: observed *data,* and *theories, which* relate these data to each other. Measuring the data may involve passive observation – for example, recording the numbers of birds nesting in a sanctuary each season – or may involve experimentation – for example, recording the time taken for a sphere to travel down an inclined plane. With some philosophical qualifications, the data represent facts, the observable face of the real universe.

It is tempting to think of a theory such as the theory of relativity, of evolution, of the electron, or of covalent bonding as describing how the universe really works – the hidden mechanism which produces the results - but it is more accurate to think of each theory as a *model* of 'reality', not the real thing. For example, Newton's laws constitute one model of motion; special relativity a more general model. Both work very well. Further, special relativity can be thought of as a refinement of Newtonian mechanics.

Evolution, in the sense that living things change over generations, is an observable fact, as any animal or plant breeder can attest. So is the existence of different but clearly related species. In the terms I used above, these are data. Darwin's Theory of Evolution provides an explanation for the development of species through the mechanism of natural selection. This theory has itself developed over the decades since Darwin.

Models are not restricted to science. A religious belief or a philosophy is a model of the universe. Freud and Jung developed models of how the human mind works. According to the context, a model may variously be described as a metaphor, a story, a mathematical structure, a stereotype or a myth. In all of these, some aspect of 'reality' is described, well or badly.

Indeed, we use models unconsciously in all decision making, from deciding when to change lane when driving to selecting one's partner: *'Is that car likely to accelerate now? "Is that girl likely to say yes if I ask her out for a drink?*

We think of other people in terms of stereotypes. A stereotype is a model - the real person is replaced by a generic dummy, based on age, sex, race, etc, the details of the model depending on one's personal experience, upbringing and education: *'He is dark and different and looks dangerous, so I'd better be careful. "She is friendly and knows my parents, so what she is selling must be a good buy.'*

By their nature, stereotypes are simplistic, often wrong, and therefore prone to

offend some people when expressed. But we all use them, whether or not we 'should'. What is your first image of someone who is 'female, blonde and good looking'? 'male, dark and different'? 'middle aged, male, well-dressed in a dark suit'? 'young, bearded, hairy, in jeans and thongs'? My first picture of a person – the model on which I initially operate – is based on first impressions. As I get to know him or her better, my picture improves – I work with a better model.

Properties of models

The purpose in creating any model is to achieve some understanding of the working of the universe, and hence to gain control of some sort. This control is expressed primarily through the ability to predict what will happen under given conditions. In short, the prime characteristic of a model is that it be *predictive.*

A model may be *deterministic*: it says what will happen under specified circumstances. Newton's laws and relativity are deterministic. A model may be *probabilistic*: it predicts in the sense that it specifies what can happen, and assigns a probability to each possible outcome. A *causal model* provides predictive ability through a theoretical framework relating the variables involved such that if one or more variables are changed, the results can be predicted. The results may be deterministic or probabilistic. A scientific theory should be a causal model. Such a model is superior to a *purely predictive model*, in that it gives a description of the way the universe works, and provides a measure of control over the results of actions.

Note that the personal models described above (for example, stereotypes) are generally predictive probabilistic models. Experience (or brain washing!) perhaps has led me to believe that 'short, fat, bald men driving Mercedes tend to be aggressive, careless drivers'. There is no theory that connects 'short/fat/bald/Mercedes' with 'aggressive/careless'. Apart from an expectation that men driving Mercedes are probably wealthy, and to get that way they have probably been aggressive, the prediction is based only on an association that I have formed between the two sets of characteristics.

Models and truth

A model is only 'true' internally. *Within* a model certain things are taken to be true, in the same sense as in mathematics; assumptions are made, and 'truths' deduced from these. Externally, a model is a *good* model if it can be used to predict with adequate confidence what will happen in certain circumstances.

3. THE SCIENTIFIC APPROACH

There has been much discussion over the years about what the Scientific Method is, and whether it exists. My impression is that this argument springs primarily from an overly prescriptive definition, in an attempt to identify it as something quite different from other modes of acquiring understanding of the universe. The scientific approach grows naturally out of everyday ways of acquiring knowledge. I observe something interesting, conjecture something about it, and test my conjecture to see if it is correct. For example, I wake early, see that it is light, conjecture that it is day, and look at the clock to check this. Or I arrive at a class, find low attendance, think of possible reasons – and ask those students who are there if there is an assignment due in another subject soon. Or I check my bank balance, find it lower than I expect, think of possible reasons again – and look for a statement of transactions from the bank. Or I observe politicians

telling lies, theorise that politicians usually lie, and check this by observing politicians speaking. These examples are trivial, but they illustrate that in everyday life we develop (usually ill-defined) theories and test them using information.

In scientific research this process is formalised. Conjectures are structured as well defined models and, ideally, tested by *experiment*. The word 'experiment' has a range of meanings, but usually implies some measure of control over one or more factors. In a statistical experiment this control may be obtained by using randomisation, either through stratified sampling or random allocation of treatments. In many cases, however, this control is very limited; at worst, perhaps, the experiment consists of using whatever relevant data is available.

Desirably, the experiment should be as objective as possible and the results reproducible. Again, in many discipline areas, these properties may be limited. For example, in research in education and psychology there are likely to be severe problems of definition, Some examples are given below:

1. A model of planetary rotation may be tested by observing sunrise and sunset times.
2. A model of electromagnetic radiation may be tested by setting up a laboratory, experimenting to generate radiation under specific conditions, and measuring magnetic fields.
3. A model of bird behaviour may be tested by identifying birds living under particular conditions (or by generating those conditions) and observing their behaviour.
4. A model of bird behaviour may be tested (as with Skinner's pigeons) by endeavouring to train them and recording the effects of this training.
5. A model of variation in communication skills with gender may be tested by taking random samples of males and females and testing the people chosen.
6. A model of the effectiveness of a new drug may be tested by taking a random sample of people, randomly selecting some to be treated with the drug, the remainder to receive a placebo, and the effects measured.

Of these, the first two models are deterministic (with some measurement error to be dealt with). The third and fourth may be probabilistic. The last two are certainly probabilistic models, and the experiments described are standard statistical experiments. In 5, the model will say something like: 'A randomly chosen male is likely to have better communication skills than a randomly chosen female.' The precise interpretation of 'communication skills' will depend on the test carried out. In 5, the model will say something like: 'If a patient is treated with this drug, he or she will probably be cured.' Again, the precise meanings of 'treated' and 'cured' depend on the details of the experiment.

In research then, a model is developed and experimentally tested against observed data. If the data do not support the model, the model is rejected, or modified and retested. The researcher is not identifying the 'truth' about the universe, but simply establishing an explanation of how observable data are generated. This explanation is tentative, in that in the future further data may cause it to be rejected or modified. Models develop over time, becoming established as they survive the impact of further tests against data. A new model is more readily accepted if it is consistent with established models. This amounts to saying that the new theory is tested against the same data as was used for the earlier theories.

The scientific method is apparently very distinctive in that it concentrates on testable models and requires them to be tested. For example, religious models, and models which may be considered similar, such as 'New Age' ideas, are generally not tested, and

in many ways not testable. However, although some of these models appear to continue to be accepted despite the observable data, some do become unacceptable. For example, few people these days accept a model that says that the earth (globally, rather than locally) is flat or that it rests on a tortoise. A similar comment applies for the models that everyone uses to make decisions in daily life. People do learn by experience - although most people are in some ways 'slow learners'!

4. MODELS IN STATISTICS

Statistics is a body of techniques for developing and assessing models, particularly those involving uncertainty or 'noise'. This modelling process takes place at all levels of a statistical analysis. Note that in statistics we can talk of generic models, such as 'the linear regression model', the 'general linear model', etc. In this paper when I refer to a model, I mean a particular model; for example, a model describing the relationship between communication skills and gender. This model may be a fixed mean model, or a linear model, or a non-linear model of some sort.

One can identify three general ways in which statistics works with models. First, some form of averaging can be used to smooth out measurement error. This can be considered as part of the measurement process, so is hardly 'statistical analysis'. The measurement noise can also be considered part of a probabilistic model, and its smoothing as part of the estimation of that model.

Second, the use of causal models with a probabilistic component is one of the primary applications of statistics in many disciplines. Statistical methods are used in the development phase of such a model, in estimating its parameters on the basis of data. They are also used to verify the model, again on the basis of data; that is to 'test' the model. Some statistical methods are of course developed to deal with specific types of model.

Statistical methods are particularly applicable in environments where there is a large amount of uncertainty, such as economics, biology, psychology and sociology. In such cases, testing a theoretical structure involves techniques of identifying possible contributions to that uncertainty, and hence ways of controlling and minimising it.

Third, statistics frequently involves development of *purely predictive* models, that is, those which are not causal. These are opportunistic, with little or no underlying theory to account for the relationship between the variables, which is no more than an *association* between them. In practice, it is hoped that a predictive model is in some sense causal. Often there may be some understanding of why the variables are related, but the details are not understood.

According to Occam's Razor the best model to use is the simplest which satisfies the requirements of the situation. If the purpose of the analysis is purely predictive, a purely predictive model is adequate. If the purpose is to identify a control mechanism, a causal model is required. If there is a choice between a purely predictive model and a causal model of equivalent complexity, the latter would normally be preferred. The unconscious models that individuals use in decision making are frequently purely predictive, and based on experience – 'sample data'. As long as they enable good decisions to be made, this is adequate.

5. HOW MODELS ARE USED IN STATISTICS

The aim here is to show that *throughout* statistics we are working with models. This includes the rationale for what we do when we do descriptive statistics.

5.1. AN EXAMPLE OF MODELLING SAMPLE DATA

I have a set of incomes measured on a sample (Table 1). I could simply store the data, but to use them I must process them - I form a relative frequency distribution (Table 2), probably grouping the data, graph the distribution, calculate summary statistics. I do this because I am not really interested in the sample at all, except in that it enables me to describe the population from which the sample came. I have only a sample, so I cannot obtain a precise description. Instead, I aim for a model of the way incomes are distributed in the population.

Consequently, I smooth the data by grouping them. This filters out some of the statistical noise, to show any underlying pattern. With luck I can then expect the population to show this same pattern, so it provides the basis for the model of the population.

Table 1. Income Data ($000)

22	23	25	25
17	21	18	19
21	31	27	27
19	25	17	36
26	18	29	23

Table 2. Frequency of Incomes

Income	Frequency
10-14	0
15-19	6
20-24	8
25-29	4
30-34	1
35-39	1
Mean	23.45
SD	5.0312

Calculating summary statistics does the same thing. The distribution is commonly modelled by calculating measures of location and spread, and skewness and kurtosis if these are appropriate. An alternative is to obtain the five-figure summary (median, quartiles, maximum and minimum), often as a box plot. So a model, with the intention of applying that model to the population, approximates the sample data.

5.2. AN EXAMPLE OF MODELLING EMPIRICAL POPULATION DATA

I am interested perhaps in establishing a new business in a particular locality, and I need to know how large an area I should need to consider as my catchment area for customers, so I am interested in the income distribution for Melbourne, and how it varies across localities.

It may be possible for me to obtain this information from the most recent Census, but this information is by no means up to date; in fact it would be out of date on the day after the Census was taken, let alone by the time it is released. Further, even if I view the aggregated data that would be made available to me as 'raw data', I would almost certainly want to smooth it and approximate it further, simply for reasons of efficiency. I do not need the detail available. In fact for this example I would probably use a simpler model: one which gave only the number of households in each locality and their mean income, or one which gave the number of households with income greater than $\$x$ in each locality.

Further, it is very commonly the case, and arguably always so, that the real population of interest is not clearly defined, so that it has to be modelled using the population data that are available.

5.3. ESTIMATING MODEL PARAMETERS: A MEAN

I take a random sample in order to estimate the mean income of households in a locality in Melbourne, and will obtain a confidence interval in the standard way. It is common for people to speak of this as estimating the 'true population mean' but it is more accurate to say that what is being estimated is the mean of the model used. This model is:

$$Y_i = \mu + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma_Y^2); \quad iid \tag{1}$$

where Y_i is the income of the i-th household. This is, if you like, an incompletely specified model, but it is certainly a model. Note that it says that the only parameters of interest are the mean and variance, and the variance does not change as successive observations are taken.

Most important, the requirement for 'independent identically distributed' observations is in many practical applications a very strong assumption! When the population is not clearly defined, mistakes may be made in the selection, and some respondents do not answer or drop out, it is a very strong assumption indeed.

It is of course true that an estimate of the model mean is in turn an estimate of the population mean - but this is likely to change with time, and if the population is ill defined it is not at all clear what the 'true mean' means. So it is easy enough to estimate the model mean, with a confidence interval, but if the model is not a good reflection of the population, the confidence interval is meaningless.

5.4. ESTIMATING MODEL PARAMETERS: REGRESSION

As a real estate agent I want to identify the factors that affect the price of a house for sale. I have taken a sample of recent houses sold by my firm and recorded a number of variables, which are presented in Table 3. (This example excludes the three P's - Position, Position, Position, which refers to an old Real Estate adage in Australia - that the characteristic of a property which overwhelmingly determines its price is its location and assume that the houses are comparable in terms of this variable.) This is a standard example to demonstrate multiple regression.

As soon as we move into this topic it is clear to all that we are talking in terms of models. However, the fixed mean model is rarely mentioned and people still talk in terms of 'true' - 'the 'true' value of the slope of the regression line'. This must be

confusing to students - we talk of 'assuming a linear relationship', then talk of the true values of the parameters of this assumed relationship.

The situation is properly as follows. The basis from which we start is the fixed mean model (1) where Y_i is now the price of the i-th house in the sample. This is effectively the model for house prices in which none of the variability in prices is explained by other factors. If a model involving other factors is to be used instead of this, it must perform better - otherwise we would be using a more complex model for no benefit.

Table 3. Example

House	Price ($'000)	House size (squares)	Age (years)	Block size (000 sqft)	Heating
1	89.5	20.0	5	4.1	1
2	79.9	14.8	10	6.8	1
3	83.1	20.5	8	6.3	2
4	56.9	12.5	7	5.1	2
5	66.6	18.0	8	4.2	2
6	82.5	14.3	12	8.6	3
7	126.3	27.5	1	4.9	1
8	79.3	16.5	10	6.2	1
9	119.9	24.3	2	7.5	3
10	87.6	20.2	8	5.1	2
11	112.6	22.0	7	6.3	3
12	120.8	19.0	11	12.9	3
13	78.5	12.3	16	9.6	3
14	74.3	14.0	12	5.7	1
15	74.8	16.7	13	4.8	2
Mean	88.84	18.17	8.67	6.54	

For the house price data, we may consider a model in which the price of a house is related linearly to its size:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2_{Y|x}); \quad iid \tag{2}$$

where x_i is the size of the i-th house. The question is whether this model is likely to perform better than the fixed mean model. This is clearly a model selection process, but it is rarely presented in this light. Teachers refer to the 'true line', which is clearly meaningless, hence confusing to students. The decision is usually carried out using hypothesis testing.

Similarly, with a multiple regression model the standard F test compares the model with at least one nonzero coefficient with the fixed mean model. Again, this is a model selection process, but carried out using hypothesis testing. Again, as pointed out in 5.2, the population from which the sample was taken is likely to be ill defined. We would certainly like to apply the results to a wider population; that is, use the estimated model (carefully!) more widely.

5.5. A CROSS TABULATION EXAMPLE

I am interested in the relationship between a person's attitude to drugs and his or her age. I develop a set of statements and look for level of agreement on a standard five-point scale with each. With age, I am essentially concerned with broad age bands: 'young', 'middle' and 'old', suitably defined. I take a stratified sample and administer my

questionnaire.

The sample members are then cross-tabulated for each statement. For each of the age groups the sample frequency distribution provides an estimate of an empirical model for the pattern of agreement. There is no relationship between the variables if the three models are identical; there is a relationship if they are not identical. The choice then is between two 'super models' - one in which the three strata models are identical, with this common model estimated by pooling the data, and one in which at least two of the three differ. Again this is an exercise in model selection.

In this example there are three real subpopulations which are being modelled. In many applications this is not the case. For example, in researching the side effects of a drug, one might wish to compare the effect of the drug in two or three combinations plus a placebo. In this case random selection is used to create the effect of random selection from several *notional* subpopulations. Again, these notional subpopulations are modelled by the data, and the question is whether the models are better taken to be identical, or different.

5.6. AN ANALYSIS OF VARIANCE EXAMPLE

In one way analysis of variance the modelling process is the same as in the cross tabulation examples briefly described above. The nominal explanatory variable is used to divide the population into strata, real or notional. The question is again whether the models for the strata are identical or not. Because the dependent variable is numeric, the strata models can be expressed algebraically, and different levels of difference considered. The simplest set of models is one that differs only in mean. The choice then can be expressed in familiar terms as between the (overall) fixed mean model (1) and:

$$Y_i = \mu + \alpha_j + \varepsilon_i; \quad \varepsilon_i \sim N(0,\sigma_Y^2); \quad iid \tag{3}$$

where α_i is the stratum effect.

6. HYPOTHESIS TESTING

We construct stories of how the universe operates - we call these stories 'theories' or 'models'. Significance testing is one way in which we choose between stories as to which is (probably) more useful in a specified context.

First, hypothesis testing is not restricted to statistics or academic research. If you are told some piece of news or gossip, you automatically check it for plausibility against your knowledge and experience. If you are at a seminar, you listen to the presenter in the same way. If what you hear is consistent with your knowledge and experience you accept that it is probably true. If it is very consistent, you may accept that it IS true. If it is not consistent, you will question it, perhaps conclude that it is probably not true.

IF the news is something that requires some action on your part, you will act according to your assessment of the information. If the news is important to you, and you cannot decide what to do on the basis of prior knowledge, you will presumably go and get corroborative information, hopefully in some sense objective information.

This describes hypothesis testing almost exactly; the difference is a matter of formalism. As indicated in the examples above, a statistical hypothesis test compares two probability models of 'reality', so it is a technique for model selection. It does

however have two special characteristics.

First, one of these models is *embedded* within the other: that is, one model is a particular case of the other. Neither of these models is 'true' - but either or both may be good descriptions of the two populations, in the sense that if you do start to randomly select individuals, the results agree acceptably well with what the model predicts. The role of hypothesis testing is to help you decide which of these is (probably) the better model - or if neither is.

Second, one of these models is 'privileged' in that it is assumed 'true' - that is, if neither model is better, then you will use the privileged model. In most cases, this means the simpler model.

More accurately, if you decide that the models are equally good (or bad) you are saying that you cannot distinguish between them on the basis of the information and the statistical technique used! To decide between them you will need either to use a different technique, or, more realistically, some non statistical criterion. For example, in a court case, if you cannot decide between the models 'Guilty' and 'Innocent', you may always choose 'Innocent'. In more typical statistical applications, the choice is usually (following Occam) the simpler model, commonly the embedded model.

There is no necessary statistical reason why one model is thus privileged. In my earlier paper (McLean 99a) I stressed my belief that this approach reflects our (and Fisher's) cultural heritage rather than any inherent need for it to be that way.

Given that the null model is privileged, a test is only carried out if the sample data suggest that it should be rejected; that is, the alternative model appears to be better. The test provides a measure, the *p*-value of the test statistic, of how much better. If this measure is sufficiently 'significant' we decide that the alternative model should be used.

A commonly expressed view is that for a continuous variable 'the point null must be false'. This objection misses the point completely. It springs from the idea that a test identifies something true (or false) about the universe, and the probability that a parameter equals a particular value is effectively zero. But testing is not about 'truth' but about 'usefulness'; the null model is only a model. It is certainly usable and may be better.

6.1. TESTING IS ABOUT DECISIONS

Fisher's approach to significance testing is often described as not involving decision making: That a test is used to assess the evidence in favour of the alternative, enabling a statement about the *significance* of this result in the continuing development of scientific knowledge. In fact a significance test does entail a contingent decision, in the sense that the result tentatively establishes current knowledge. Further, this result will be used to determine the future direction of the research.

Neyman and Pearson introduced the concept of explicitly deciding which of two hypotheses is true, and consequently the concepts of type I and II errors.

Recognition that a test is a choice between two models helps us to see that the two approaches differ more in terms of their areas of application than in substantive terms. Fisher's approach emphasises the idea of *conditional rejection of the null,* so is appropriate in scientific research. The Neyman-Pearson approach applies in areas where a decision clearly must be made, such as in quality control. For example, with a bottle filling machine, which is periodically tested as to the mean contents, the null is that the machine is filling the bottles correctly. Rejecting the null entails stopping the machine; accepting it means the machine will not be stopped.

The role of decision making in testing is not confined to the test itself. In a piece of

research it is often easy enough to identify a characteristic of interest – the problem is how to measure it. If I am interested in the relationship between *ability in statistics* and *ethnic background,* for example, I measure the statistics ability using an examination of some sort; I measure ethnic background by defining a set of ethnicities. There are literally an infinite number of combinations that I can use – infinitely many different exams, all purporting to measure 'statistics ability' (even if I change only one word in an exam, I cannot be absolutely certain of its effect, so it is a different exam!) and a very large number of definitions of 'ethnicity'.

I now apply the test to a group of people of varying ethnicity, score them on the exam and analyse the results, including a hypothesis test, to decide if statistics ability is related to ethnicity. This test might be a simple ANOVA, a Kruskal-Wallis or a chi square test, depending on how I score the exam.

The point here is that the definition of the models being compared in the test includes the definition of the variables used. If I reject the null model I am NOT saying that there is a relationship between statistics ability and ethnicity, only that there is a relationship between the two variables I used.

Please note that the test is not saying this – I am. The test merely gives me a measure of the strength of the evidence provided by the data ('significant at 1%' or '*p*-value of .0135'). This measure is only relevant if the models I have used are appropriate. I can use other evidence to decide if this is so. So in a research project there are three levels at which judgement is used to make decisions:

1. Deciding what variables are to be used to measure the characteristics of interest, and how any relationship between them relates to the characteristics;
2. Deciding on the model to be used, and how to test it;
3. Deciding the conclusion for the model.

In each of these there is evidence we use to help us make the decision. Accepted theory and practice, and general experience, help with the first and the second. With the second there are also related hypothesis tests such as tests for normality. For the third, the hypothesis test provides the evidence.

It is essential to recognise that in research a theory is presented, together with observational evidence to support it. The evidence does not 'prove' the theory to be 'true'. If the evidence is accepted as sufficiently strong, the theory will (tentatively) be accepted as the current model. Further observational evidence may continue to support the theory, or may weaken its support. In any case, the judgement of the scientific community as to what is 'sufficiently strong' evidence is simply that – judgement. It is as objective as possible, but no more.

Statistical hypothesis testing is a tool, which helps with the assessment of the evidence within certain constrained and generally well-defined bounds. Within those bounds it can be extremely useful. Unfortunately those bounds are too often not understood or are ignored. The way hypothesis testing is commonly presented in the textbooks and the classrooms does not help.

6.2. A MULTIPLE REGRESSION EXAMPLE

Selection of the best model in a multiple regression application is a good example of the role of hypothesis testing for students. For the house price data (Section 5.4) the Excel output for a regression against all the variables is shown in Table 4.

The variables *oil* and *electric* are dummy variables introduced to account for type of heating.

Table 4. Excel Output for a Regression Analysis

Regression statistics	
Multiple R	0.983021
R Square	0.966331
Adj. R Square	0.957149
Standard Error	4.366877
Observations	15

ANOVA	df	SS	MS	F	Significance F
Regression	3	6020.470208	2006.823403	105.236689	0.000000
Residual	11	209.765792	19.069617		
Total	14	6230.236000			

	Coefficient	Std. Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-5.328977	7.451269	-0.715177	0.489401	-21.729118	11.071164
House size	4.086875	0.276215	14.796016	0.000000	3.478930	4.694819
oil	-11.12633	2.731337	-4.073584	0.001840	-17.137968	-5.114696
Block size	3.609421	0.572917	6.300074	0.000058	2.348438	4.870404

The *F* test compares the linear model with no restrictions on coefficients:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i ; \varepsilon_i \sim N(0, \sigma^2) \tag{4}$$

with the constant mean model (1). Each *t* test compares the above linear model with no restrictions (4) with the same model with one restriction; for the first variable:

$$Y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i ; \varepsilon_i \sim N(0, \sigma^2) \tag{5}$$

The point I wish to make here is that each of these tests provides evidence for a choice. The overall decision as to the best model – within the limits of the data and the type of model considered – is based on a number of tests. And again, it is a matter of judgement, external to the statistical analysis, although using statistical experience.

6.3. A NULL FREE APPROACH

The standard approach to hypothesis testing, recast in terms of comparing models, is to use as the measure of strength of evidence the *p*-value, that is the probability of the sample result (or 'worse') occurring *if the null model is used.* This can equivalently be expressed in terms of critical values. I suggested above that the null need not be so privileged, that the reasons why it is are cultural rather than statistical. It is true that if we formulate the models in the usual way, the null model is fully specified, so a *p*-value can be computed, while with the alternative this is not so. For example, to test

$$\begin{aligned} H_1 &: \mu = \mu_0 \\ H_2 &: \mu \neq \mu_0 \end{aligned} \tag{6a}$$

one is forced to start by using H_1.

If the test is carried out in the usual way, and the null rejected, the question must then be: 'If $\mu \neq \mu_0$, what is the best estimate of it?' Under normal circumstances the answer is 'The observed value of the sample mean.' In a very real sense, then, the true hypotheses of interest are:

$$\begin{aligned} H_1 &: \mu = \mu_0 \\ H_2 &: \mu = \mu_1 \end{aligned} \tag{6b}$$

where μ_1 is the observed value of the sample mean. That is, the choice is between two models that are identical except for the value of the mean. The first model is the best available under one theory; the second is the best available on the basis of the sample evidence.

From this point of view, neither model is privileged. A *p*-value can be computed based on either. In this example, it is reasonable to compute one sided *p*-values, and because the *t* distribution is symmetric, these will be identical:

$$P\left(\bar{x} < \mu_0 \middle| \mu = \mu_1\right) = P\left(\bar{x} > \mu_1 \middle| \mu = \mu_0\right)$$

supposing that $\mu_1 > \mu_0$. This approach appears to be more natural for students.

6.4. BAYESIAN METHODS

The Bayesian approach has not been mentioned, but it certainly forms part of the background. The predictive approach (McLean, 1998) is not unrelated to the Bayesian. For the present topic, my aim has been to clarify the nature and role of classical hypothesis testing, so these matters have been left in the background. The same comment applies to decision theoretic methods. To the best of my knowledge, nothing that I have said is in conflict with these approaches.

7. MODELS AND RESEARCHERS

Researchers using statistical techniques must understand clearly that they are working with models of the real world, not with the real world itself. They are not discovering truths, but creating a better description of the world. This description is primarily predictive. Statistical analysis is concerned with assessing the predictability of results provided by a model; whether the model is also a causal description is outside its scope.

Second, a choice between statistical models based on hypothesis testing is made within the context of a general model. The choice is between different versions of that general model. Consequently, the test is only valid if the assumptions of that general model are valid.

Third, they must understand that a statistical analysis does not, even within the context of the general model, *prove* one version is superior to the other; it simply indicates that one is likely to perform better than the other. All conclusions are tentative, although actions may be based on them.

Lastly, any statistical analysis involves considerable judgement. At the simplest level, what is an appropriate level of significance in a hypothesis test is a matter of judgement. Questions of definition of variables, wording of questions in a questionnaire,

the effect of non-response, are all involved in the definition of the models concerned, and all involve some personal judgement on the part of the researcher.

Whatever methods are adopted in teaching, these levels of understanding must be conveyed in teaching statistics to researchers – as indeed to anyone using statistics.

8. FINAL NOTES

A participant in the discussion suggested that Galileo is an example of how hard it is to defend scientific theories. I totally agree that it is important to make students realise that statistics can help them to decide between competing models.

Another question raised was what reality does a model reflect . The question of the nature of reality, or whether it even exists, is an interesting philosophical question. One can argue that not even observed data are real since they are perceived through the mind. My opinion is that we should in practice act as if there is some underlying 'reality', our observed data, if obtained carefully enough, are more or less 'real'. The important thing is that our theories and beliefs that account for the data, are models.

It was also suggested that reality is subject to change with time. In order to avoid misunderstanding the reality, we do have to look at it either as a dynamical system, or, by fixing time, to model a section of the reality. When we 'look at it as a dynamical system' we are using a model, usually more complex, since the evolution of the system has to be modelled.

Models are important, in the sense that a model is a useful fiction that reflects some aspects of the real world, that a model is our invention, which makes the problems easy to handle. However, we should keep in mind that we are sometimes liable to persist with the models themselves, disregarding the real world. Models make problems *possible* to handle. And we certainly tend to persist with some models when the 'real world 'evidence is that they should be rejected. This happens in both daily life and scientific research.

Another point debated was if I distinguish between subjective and objective probability. I do not believe 'objective probability' exists; if it does, it is inaccessible to us. In general terms, the causes of a particular result (for example, getting heads when a coin is tossed, or selecting an outcome through 'random' selection) are so complex that they have to be replaced by the notion of probability. So probability always occurs as a model. If by 'objective' probability is meant the frequentist approach: this is a widely used method of estimating the parameters of a probability model. Note that it is still subjective, in the sense that judgement is used in this process. The frequentist interpretation of a probability as literally being a (long run) proportion is simply incorrect. The probability of getting a head on the next toss is in no way the same thing as the expected proportion of heads in the next 10,000 tosses.

Models might also be seen as machines that serve to produce knowledge, rather than as images from reality, although I certainly do not see it as a static image. However, the model does not produce knowledge – it *is* the knowledge.

Another important point is students' models and how they sometime make incorrect generalisations. Learning is a process of developing in the mind models of the world, and learning to manipulate those models. Students often establish incorrect models, and must correct those through further learning. This same comment, of course, applies for all people in daily life. We all have mental models that could be improved. The process of improving models, as students or in daily life, can be painful.

REFERENCES

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Journal of Mathematics Thinking and Learning,* 2(1-2), 75-98.

Cook, T. D., & Campbell, D. T.(1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

DeGroot, M. H. (1986). A conversation with Persi Diaconis. *Statistical Science*, 1(3), 319-334.

Fisher, R. A. (1948). *Statistical methods for research workers*, 10th edition. Edimburgh: Oliver and Boyd.

Gingerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Henkel, R. E. (1976). *Tests of significance*. Beverly Hills, CA: Sage.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*, La Sale, Ill: Open Court.

Ito, P. K. (1999). Reaction to the invited papers at the meeting on statistical education and the significance tests controversy. *Proceedings of The 52nd Session of the International Statistical Institute*. *Bulletin of the International Statistical Institute* (Tome LVIII, Book 3, pp. 167-168). Helsinki: International Statistical Institute

McLean, A. L. (1998a). The forecasting voice: A unified approach to teaching statistics, In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics* (pp.1193-1199). Singapore: International Association for Statistical Education and International Statistical Institute.

McLean, A. L. (1999b). The predictive approach to teaching statistics. *Department of Econometrics and Business Statistics Working Paper* 4/99. Melbourne: Monash University.

McLean, A. L. (1999c). Hypothesis testing and the Westminster system. *Proceedings of The 52nd Session of the International Statistical Institute*, *Contributed Papers* (Book 3, pp. 287-288). Helsinki, Finland: International Statistical Institute.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance tests controversy. A reader* Chicago: Aldine.

O'Conner, M. P., & Spotilla, J. R. (1992). Consider a spherical lizard: Animals, models and approximations. *American Zoologist*, 32, 179-193.

Popper, K. R. (1977), *The logic of scientific discovery*, 14th ed. New York: Routledge.

Alan McLean
Monash University, Department of Econometrics and Business Statistics,
900 Dandenong Road, Caulfield East,
Victoria 3150, Australia.
E-mail: alan.mclean@buseco.monash.edu.au