ISSN 1440-771X
ISBN 0 7326 1084 2

On the Nature and Role of Hypothesis Tests

Alan McLean

Working Paper 4/2001

June 2001

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

MONASH UNIVERSITY
AUSTRALIA

On the nature and role of hypothesis tests

Alan McLean

Monash University, Australia

Abstract

Hypothesis testing is widely regarded as an essential part of statistics, but its use in research has led to considerable controversy in a number of disciplines, especially psychology, with a number of commentators suggesting it should not be used at all. A root cause of this controversy was the overenthusiastic adoption of hypothesis testing, based on a greatly exaggerated view of its role in research. A second cause was confusion between the two forms of hypothesis testing developed by Fisher on the one hand and Neyman and Pearson on the other. This paper discusses these two causes, and also proposes that there is a more general misunderstanding of the role of hypothesis testing. This misunderstanding is reflected in vocabulary such as 'the true value of the parameter'.

Keywords: hypothesis test, significance, p value, probability model, statistical model, prediction

1. Introduction

Hypothesis testing has long been a part of statistics; to many it is, as one of the faces of inference, at the core of statistics, and is taught as such. Its use dates back three centuries, but modern hypothesis testing is usually dated from the work of Fisher on the one hand, and Neyman and Pearson on the other. Their approaches had a degree of commonality (Lehmann, 1996), but the underlying views of what they were doing were very different (Gigerenzer, 1993). What is commonly taught today in fields including psychology, education, medicine and, more recently, ecology, is an amalgam of the two approaches, what Gigerenzer calls "the hybrid logic of scientific inference, the offspring of the shotgun marriage between Fisher and Neyman and Pearson".

For much of the last four decades or so there has been considerable controversy in the fields mentioned above about the use of hypothesis testing in research. Since Morrison & Henkel (1970) presented an extensive review of criticisms of hypothesis testing, the arguments for and against it have ebbed and flowed; for example, Thompson (2001), McLean and Ernest (1998). Attitudes vary from unqualified support, through recognition of the importance of its role, provided that it is supplemented by other forms of analysis, to calls for it to be banned. Neither extreme is to be countenanced, but the latter smacks of throwing the baby out with the bath water. Sterne and Davey Smith (2001) present a typical discussion, in this case in the field of medicine, with an interesting range of responses on the web site.

Much of the controversy appears to have stemmed from the simple but extensive misuse of hypothesis testing, both by researchers and by editors of journals in which the researchers wished to publish their results. Papers were published, or not, on the basis of whether their results were significant at the 1% level or not (Melton 1962) and of course researchers responded to this in the natural way. They also had the tests take the major role in determining, and presenting, their results. Not surprisingly, moves are occurring to restore hypothesis testing to something more like its proper place.

For example, the American Psychological Association (1994) encourages researchers to provide effect-size information. More recently, the APA Task Force on Statistical Inference, convened after the publication of an article by Cohen (1994), plans to recommend quite sweeping reforms, most more than acceptable but some perhaps putting that metaphorical baby at some risk (Wilkinson 1999). Similar changes are occurring within the American Education Research Association (Thompson 1996). These debates are discussed in Harlow, Mulaik & Steiger (1997).

It should be noted that the controversy appears to focus primarily on the use of univariate and bivariate tests, particularly tests on the difference between two means. This is not, I believe, coincidence.

Most of the argument is based on a common perception of what hypothesis testing is, with some differences such as those between the frequentist and Bayesian schools. Much of the disagreement can be located in varying ideas as to the **role** of hypothesis testing in statistics.

This general argument in some parts of the research community has an echo in the teaching community. Students notoriously find hypothesis testing a difficult subject to learn properly. Although it is not difficult for students to carry out a class room (or examination) hypothesis test exercise, it is difficult to convince a teacher that the students really understand what they have learnt. There is considerable research into what students understand by, for example, ‘level of significance’ which demonstrates considerable confusion; for example, Shaughnessy (1983), Falk (1986), Pollard and Richardson (1987). Given the confusion in the research community, this is not surprising.

The aim in this paper is to present a view of hypothesis testing which removes, or at least reduces, this confusion – in a sense, to step around the controversy. The discussion is couched in terms of elementary statistical applications, partly because these seem to be where most of the controversy lies, partly because the proposals seem to be implicitly followed in more advanced applications. The major contribution, apart from making explicit what seems to be the implicit approach, is to indicate that this approach can be extended to elementary statistics. It has, in particular, direct implications for statistics education.

The basic suggestions are:

1. Our knowledge of the world is in terms of models, although we observe data – so these models are based on data – plus previously held models. In particular, all statistical analyses are concerned with probability models.
2. Hypothesis testing is a form of **model selection** under particular conditions.
3. Under the predictive view of statistics, hypothesis testing is a choice between a model with parameters specified by a null hypothesis, and one with parameters specified by sample data.

The details of this approach, and some consequences, particularly with respect to this controversy, are discussed in the remainder of the paper.

2. Models and proof

2.1 Models

I pointed out recently (McLean 2001) that humans operate on the basis of models in the form of stereotypes, preconceptions, biases, myths, religious beliefs and scientific theories. These models are based on data, in the form of personal observations, information passed on to us by parents, school, mentors, advertising and the rest. Some are, to a greater or less degree, common to large numbers of people, some are very individual. Both individually and commonly held models change as new data is accepted.

Many of the individual models we operate by are probabilistic in nature, although of course very imprecise. People understand the idea of uncertainty very well indeed: since we live in an uncertain world, there has been considerable evolutionary pressure to understand it, so people have a general concept of probability in the sense of what is likely or unlikely to occur. On the other hand, people seem to have difficulty comprehending numeric probabilities, particularly very small probabilities, so it is not surprising that students find mathematical, academic probability difficult.

Scientific research consists essentially of formulating a model of some aspect of the 'real world', making predictions, on the basis of the model, which can be tested, collecting data and testing the model, then accepting, adjusting or rejecting the model according to the test. In practice, of course, the sequence is nothing like as straightforward as described here. One may think that the model is suggested by data (inductive reasoning) or that it is deduced from other models. Note that the word 'models' is used here very generally, to include the fuzzy natural models we have of the real world as well as formal scientific theories, mathematical models and the like. In this sense Fisher seems to be correct – models (hypotheses, theories) do not spring from the data alone.

Models can be classified in a number of ways. All models are developed for the purpose, at least potentially, of **prediction**. Some will be in some sense **causal**, in that they embody some mechanism. Note that the mechanism may not be 'correct'. Invoking the gods to bring rain involves a causal model of predicting rain: 'If we carry out the ceremony this will make the

gods happy and inclined to do what we want and there will be rain' is just as much a causal model as is: 'If I apply a force to this object its acceleration will be given by the force divided by the mass of the object.' Most people would consider the former to be an inferior model, in that predictions based on it are neither reliable nor accurate.

Essentially the same classification is made by Lehmann (1990). He quotes Neyman, who identified 'interpolatory formulae' and 'explanatory models', and Box, who uses the terms 'empirical' model, and 'theoretical' or 'mechanistic' model. These authors were talking of statistical models, but the same identification can be made quite generally.

Implied in the above is the idea that the quality of a model is measured by the degree to which it provides accurate **predictions**. A causal model is not better than a purely predictive model because it provides understanding. It is better because it gives better predictions. This will only happen if the hypothesised mechanism is indeed appropriate; and we only know that it is appropriate because it gives good predictions!

The **scientific method** can be viewed as an approach whereby such models are tested by comparing predictions based on them with observed data. If the data disagree with the predictions, the model is modified or discarded. As such, it is a formalisation of a mode of thought which is probably not as common as we would like to think, but is nevertheless used in daily life. People do test their models – their preconceptions – by comparing their predictions with what they observe, and modify these models. The process is called 'learning'[1].

The discussion above assumes that one can distinguish clearly between the 'theory' and the 'data' – the observational or experimental evidence. In reality the distinction is not necessarily clear. One can argue that all observational evidence is itself but a model of the real world, since it is filtered through our senses and interpreted by our brains, but to be practical, we have to operate as if 'data' really is observed fact[2]. If you like, we each operate under a super model of the world.

2.2 Probability models

A large number of the natural everyday models we use are probabilistic. As such, they provide predictions in the sense of statements of what is likely to occur. For example: 'If I walk down that street I am very likely to be mugged!' These models are of course fuzzy and ill defined, although they probably are more precise in the mind than in the language used to express them. As in the general case, these models are likely to be modified with experience. If I walk down that street a few times without being mugged I am likely to change my model.

[1] The extent to which by developments such as the Greek belief that abstract reasoning is the ideal way to develop understanding hindered the development of the scientific approach from what seems a natural origin is an interesting question.

[2] There appears to be evidence that when we 'see' a scene, most of it is a model in the sense that it is constructed from data already stored in the brain; only relevant detail is fresh.

It is well recognised that theoretical statistics is concerned with probabilistic models. Lehmann (1990) quotes Fisher as defining the principal task of statistics to be 'the reduction of data', and that 'this object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a sample. The law of distribution of this hypothetical population is specified by relatively few parameters…' (Fisher 1922). Again, the population from which a sample came is a 'product of the statistician's imagination' (Fisher 1955, p.71). On the other hand, the conclusion to be reached from elementary textbooks, particularly those directed at nontechnical students, can only be that the idea of statistics being concerned with probability models is simply not considered.

The concept of probability itself is a model. The probability of an event refers to the likelihood of that event occurring on a particular occasion. It can, in theory, be measured as a long run proportion, as a proportion of favourable cases, or as a strength of belief by identifying a breakeven bet. None of these **is** a probability. Identifying a probability as the proportion of times an event is expected to occur if the occasion is repeated (the frequentist imagery) is at best a device to help a student understand, useful in some circumstances and totally inappropriate in others.

A Bayesian may talk in terms of the probability of the true value of a parameter; for example, 'The true mean lies between 10 and 20 with probability 0.9.' First, note that this expression is a form of **shorthand**. On the face of it, the mean either is within the interval or it is not, so the probability is either 1 or 0. But probability is not about what is or is not, but is a measure of **uncertainty**, which is a property of **knowledge** about what is. We think of probability as referring to the future: we are uncertain about the result of tossing a coin because we have not yet tossed it. But probability refers to lack of knowledge about what has happened, is happening, or is yet to happen. The phrase above is a statement about the uncertainty of knowledge of the value of the parameter.

It is also a statement about a **model**. More accurately, it is 'On the basis of available knowledge, the level of certainty (probability) that the mean of the model used lies between 10 and 20 is 0.9.'

2.3 The predictive view

The **predictive view of statistics** (McLean, 2000b) suggests that the purpose in any statistical analysis is to formulate a probability model which can be used to predict what is likely to happen in the context under consideration. This is true even when data are simply recorded with no apparent use in mind, or when the ostensible use is simply to 'understand the situation'. This reflects a view that statistics are useful!

The predictive view leads to some understanding of why we do statistics the way we do – for example, why the mode and mean are so important. It also suggests changes in what we teach:

for example, prediction intervals should form part of any elementary course and, despite the importance of working with data, trying to teach statistics without probability is by its nature flawed. Naturally, the predictive approach leads to an emphasis on the use of models.

2.4 Proof

It is not true that 'You can prove anything with statistics!' The reverse is true. The role of statistics is to provide evidence based on observed data, hopefully fairly objective, and to some extent to provide an assessment of the strength of that evidence, for or against some model. This can take a form whereby a given model is tested, or where data are used to suggest a particular model, and to estimate appropriate parameters.

The word 'proof' has a variety of connotations. In mathematics its meaning is that a conclusion is 'proved' if it shown that it must be the case, given the assumptions made, that it follows logically from these premises. In everyday language it seems to be used more along the lines of: I consider the conclusion is proved if the evidence is strong enough to convince *me*. So the interpretation to be placed on the word depends on the required level of evidence implied by the user. Unfortunately, it is rarely made clear what that level of evidence is, so there is an element of ambiguity in its use.

We can define 'truth' and 'proof' pragmatically – a hypothesis or theory is 'true' if it is based on sufficiently strong observational ('factual') evidence that 'most reasonable people' accept it as a workable model. Its 'truth' has been 'proved' if this is the case. These pragmatic definitions are mental shorthand – they are not in any sense absolute statements. People have to be continually reminded of this.

It can be argued that when a researcher talks about the 'truth of the null hypothesis', for example, that he is using this mental shorthand – he really does understand that he is talking about a model which is not really 'true', but finds it easier to talk as if this is not so. From my observations, this is not generally the case.

This pragmatic view of truth describes science. The theory of evolution is accepted as 'true' because the evidence is sufficiently strong, and comes from such a wide variety of sources, that it is generally accepted as a workable model. (For most people this acceptance is probably implicit.) It is also the case in the criminal courts: to say that a person has been 'proved guilty' of a crime is merely to say that the factual evidence is sufficiently strong that 'most people' (represented by the jury) accept his or her guilt as a workable model. (People may find this an uncomfortable view, but the number of cases where later evidence has shown that the model 'X is guilty of this crime' is not sustainable is itself uncomfortably large.)

3. Hypothesis testing

3.1 What is hypothesis testing?

Hypothesis testing is a very pure example of the scientific method, in the sense that one is explicitly testing a hypothesis by experiment – that is, by looking at the data.

Hypothesis testing is a rather peculiar form of **model selection**. The particular features are:

(1) it is used to select between two complementary models, one of which, the **null** model is usually 'embedded' in the other[3], the **alternative** model;

(2) the null model is privileged in that, for external, generally nonstatistical reasons, it will be selected unless the alternative model appears to be significantly better (here the word 'significantly' is used in its ordinary meaning);

(3) the selection criterion is the *p* value for the chosen statistic – the probability according to the null model of obtaining the sample results 'or worse'; if this is sufficiently small, the alternative model will be selected.

Hypothesis testing does not say anything about the 'true value' of the parameter(s) being tested. As a form of model selection, it **does** provide a way of assessing which of the two models is likely to work better in prediction. In the usual terminology, if we **accept** the null, we conclude that the null model is likely to enable better predictions; if we reject it – that is, if we accept the alternative – we conclude that the alternative model is likely to enable better predictions.

This approach is independent of whether a traditional or Bayesian viewpoint is adopted, though it is perhaps more in sympathy with the latter.

3.2 The role of hypothesis testing

Hypothesis testing is carried out in two major contexts: on the one hand in research, on the other in what might be called industrial statistics, including particularly such methods as acceptance sampling. Commercial research, such as in marketing, might be considered as falling somewhere between. In the development of modern hypothesis testing, Fisher developed his ideas in the context of research, while Neyman and Pearson developed ideas appropriate to the industrial situation. One of the root causes of much of the confusion over the use of hypothesis testing was the mixing of the two approaches (Gingerezer, 1993), helped along by the acrimonious arguments between the principals.

In acceptance sampling we have a straightforward problem which is repeated frequently, and a process which can reasonably be automated. In this context, the hypothesis test is the whole of the process: a sample is taken, the relevant statistic computed, the null model is accepted or rejected and the corresponding action taken. Judgement is needed in setting up the process initially, but is not a factor in the operation of the process.

[3] Some authors do discuss non-embedded tests, but these seem to be distinguished by their theoretical technique rather than by their applicability.

A piece of research is an **argument** in favour of one model (or theory) over one or more others. The argument is based on evidence, generally of various types, and aims to convince the reader of the validity of the chosen model. Whether or not the reader accepts the validity is a matter of judgement. Within that argument one or many hypothesis tests may be used. Each hypothesis test is one part, perhaps a very important part, in the whole argument.

The research involves a large amount of design, perhaps a number of sources of information. It may involve collection of quantitative data, sets of values of variables measured on a sample of individuals, or it may involve qualitative research which looks deeply at a few individuals. It may involve considerable issues of definition, of variables, of populations of interest, of measurement issues, and the like. Along with all these issues, it will involve definition of the models under consideration.

By the time the researcher comes to analyse the data statistically, a large amount of model construction has been done, and a large amount of judgement exercised. To carry out the analysis, further modelling will generally be done; for example, assuming normality, or that the variables are interval rather than ordinal. Only then does hypothesis testing come in to the picture.

It is essential to remember that the models are only models. Interpretation of the result of a test involves assessing the validity of the model. There may be relevant tests to help (tests of normality, equality of variance, and the like) but there is always some element of judgement, concerning for example the validity of the measuring tool, the randomness of the sample, the nature of the population.

In an area like acceptance sampling, it is appropriate to set up a hypothesis test as an automated decision process, but in research this is not so. Although in accepting or rejecting a null model one is making a decision, it is, first, always tentative, and second, always a matter of judgement.

3.3 The null model

One misunderstanding that is fairly commonly held is that the null hypothesis is so called because it always specifies that the parameter value in the null model is zero. While this is so in many applications, it is certainly not always the case. Fisher considered the null hypothesis as that which is expected to be **nullified**. In quality control applications, it is usually identified as that which leads to **no action**.

In general, the null model is privileged for some non statistical reason (McLean 1999). In many applications, the reason amounts to Occam's razor: only choose a more complex model if it is worth the greater complexity. In statistics, from the predictive view this amounts to choosing it only if it is expected to lead to improved predictions. In acceptance sampling, the alternative requires some action, which is to be done only if necessary.

In scientific research, the null is taken to be the currently accepted wisdom – a new theory has to 'prove' itself. This is also the case in daily life, at least to those of a sceptical turn of mind. In a court case, it is a notion of fairness: under the Westminster system the null is that the charged person is innocent – guilt is to be 'proved'.

One can argue that the null should be conservative, but this is not necessarily what is done. In comparing variances, for example, the null model assumes them to be equal. This model is the simpler but more restrictive choice, not the more conservative, leading to a more powerful test.

3.4 The alternative model

Fisher considered only the null hypothesis, the test either 'nullifies' or 'disproves' it (Fisher, 1935), or 'confirms or strengthens' it (Fisher 1955, p. 73). Since the null hypothesis may be true or not, there is however an **implied** alternative hypothesis.

Neyman and Pearson introduced the concept of competing hypotheses, one of which must be true. Since they worked in terms of 'truth' and the frequentist view of probability, they could introduce the concepts of Type I and Type II errors, and the consequent logical structure.

The null model specifies a value of the model parameter of interest, and thus an expected value of the test statistic; under the alternative hypothesis the value of the parameter is usually considered to be left unspecified, but this is simply not so. If the alternative model is selected, the parameter will be estimated on the basis of the sample, so it is more honest to recognise that the test selects between two models – one suggested by the null model, the other with the parameter taking the value implied by the sample.

In short, hypothesis testing selects between the model suggested by theory and the model suggested by the sample data.

3.5 The nature of acceptance

Despite the terminology of Fisher, Pearson and Neyman and countless statisticians since, we do not 'prove' or 'disprove' either the null or the alternative hypothesis, except in the very pragmatic sense suggested in §2.4. We do **accept** one or other of the models. To talk of 'failing to reject the null' simply means that the sample evidence was not as strong as we would have liked, that our evidence was not strong enough to show what we hoped to show. Similarly, we do not only reject the null – we accept the alternative, albeit tentatively.

The nature of the acceptance of the null varies with the circumstances. In acceptance sampling, when the test is used to decide an immediate action, to accept the null means to carry out action A (often 'do nothing') while to accept the alternative means to carry out action B. If the null model is, in effect, that everything is working (adequately) well, and the alternative that it is not, we do not want to carry out action B unless we have to, so it is

natural to privilege the null. In this case, although we recognise that we are dealing with probabilities and models, we have to act on the decision.

In acceptance sampling and similar tasks we have a decision process, and it is useful to **automate** the decision. In this context it is reasonable to set up a significance level – strictly, a **threshold** significance level – and reject the null automatically if the p value is less than that.

On the other hand, in research, acceptance (of either model) is always tentative. Although the decision will affect future actions, acceptance or rejection of a null model is rarely total. Just how tentative the acceptance, and how definitive the actions, depends on the circumstances.

Fisher approached hypothesis testing from the point of view of research, while Neyman and Pearson introduced the concept of decision making – that one decides between the null and the alternative models (although all failed to recognise that they were dealing with models, and talked of the 'truth of the null hypothesis'). Fisher observed (1955) that the latter approach is applicable to situations such as acceptance sampling, where a decision has to be made. He considered that in research no decision is made, so the decision theoretical view is not applicable. This is clearly not the case – although choices are tentative, they do affect the researcher's future actions, particularly where budgets are limited and a career depends on research 'success'. Nevertheless, the distinction between what might be called 'industrial statistics' and 'research statistics' is valid, since in the former a choice must be made and acted on.

Where it is relevant confidence intervals (and prediction intervals) should be computed automatically. Using the former as an indicator of significance is simply doing a test in a roundabout way.

To conclude, when we accept the null, that 'there is no evidence for the alternative' as is sometimes done, is simply wrong. There is no evidence only if the sample data agrees completely with the null, in which case no test is carried out. A test is carried out only if there is some evidence in favour of the alternative. The purpose in carrying out the test is to assess the strength of this evidence in favour of the alternative. If the evidence is strong enough, the null model is rejected.

3.6 The p value and 'significance'

The *p* value is the criterion used to select between the null model and the alternative model. It is the probability of getting the sample result, or more extreme, if the null model is used.

The use of hypothesis testing does not in fact require the p value to be interpreted as a probability. In effect, it is a device for rescaling the difference between the null value of the parameter and the sample value of the statistic. Bearing in mind §3.4 above, this is the difference between the null and alternative values of the parameter. Its most important feature

is that its scale is universally applicable. A *p* value of 0.02, say, can validly be interpreted as indicating a significant difference between the null and alternative values regardless of the distribution used. There is some evidence that experience shows the standard levels of 0.01 and 0.05 to work reasonably well as threshold measures of 'significance'.

A *p* value is a measure of strength of the evidence against the null model and in favour of the alternative. In this sense how it is calculated and exactly what is calculated does not matter. A low *p* value means that the evidence is strong; a high value means it is weak. It is legitimate to compare the results of two tests using different statistics, and say one reflects stronger evidence than the other. (But the results should be interpreted with reservations – because the models are only models.)

A result is then 'significant' or 'highly significant' or 'not significant' according to whether we **judge** the evidence to be strong, very strong or weak, respectively. The measure, *p*, **helps** us to decide, but does not make the decision! It is **customary** to judge a result as 'significant' if the *p* value is less than 5%. It is clearly unwise to interpret this figure as a fixed measure.

To calculate the *p* value requires the null model to specify a value for the parameter. For a one sided test, for example on a mean, where the null takes the form $\mu \leq \mu_0$ or $\mu \geq \mu_0$, while the alternative hypothesis specifies an open interval, this requires the further assumption that the parameter takes the boundary value. In teaching, the ability to do this is often presented as one criterion in identifying the null hypothesis.

The reason for calculating the *p* value based on the null model is because it is privileged, as indicated in §3.3. In teaching, the subsidiary reason given is that it provides a fully specified model. Accepting the point made in §3.4, both hypotheses enable sufficiently specified models. Arguing that the null model should not be privileged, an 'alternative *p* value' which gives the probability, on the alternative model, of getting the null value. With this approach, the two *p* values will be identical if the distribution used in the test is symmetric, otherwise not.

4. Examples

Much of the controversy over the use of hypothesis testing seems to be at the elementary level. I believe that this is largely because users do not recognise that they are working with models – that they expect to find out something 'true' about the world. When more complex statistical methods are involved, people are more likely to recognise that they are working with models: more precisely, teachers and text books are more likely to talk in terms of models. To illustrate, multiple regression is more likely to be taught in terms of selection of the 'best' model, while introductory simple linear regression is likely to be couched in terms of the regression line as an estimate of 'the true regression line'. However, my impression is

that even introductory regression is more likely to be expressed in terms of models than is basic univariate statistics.

4.1 Elementary linear regression

The model used (implicitly) in much of basic statistics is

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

It is convenient in this context to write this as:

$$Y_i = \beta_0 + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma_Y^2), iid \quad \text{.....(1a)}$$

The simple linear regression model is:

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

The variable mean model (2) reduces to the constant mean model (1a) if the coefficient β_1 is zero. The standard test on the coefficient β_1 does not test whether or not there is a 'true' linear relationship, or whether *Y* (really) **is** related to *X*. All it does is to suggest which of these models is likely (on the basis of the sample data) to do a better job of forecasting values of *Y*. On grounds of simplicity the constant mean model will be selected unless the regression model performs convincingly better on the sample data: that is, unless the value of $\hat{\beta}_1$ is significantly different from zero.

Accepting the conclusion from the predictive approach that the actual alternative model has the parameter estimated from the sample, this test selects between (1a) and

$$Y_i = \beta_0 + \hat{\beta}_1 x_i + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2), iid \quad \text{.....(2a)}$$

Neither model specifies the constant term (mean or intercept) nor the variance. The test selects between the **derived** models:

$$\frac{\hat{\mathrm{B}}_1 - 0}{s_{Y|x} / \sqrt{SSX}} \sim t_{(n-2)} \qquad \text{and} \qquad \frac{\hat{\mathrm{B}}_1 - \hat{\beta}_1}{s_{Y|x} / \sqrt{SSX}} \sim t_{(n-2)}$$

where $\hat{\mathrm{B}}_1$ is the random variable of which $\hat{\beta}_1$ is the observed value, and $s_{Y|x}$ is the residual standard deviation (the misnamed 'standard error of the estimate'.)

One can extend the argument of §3.4 to observe that accepting the null model entails assuming the best estimate of it, based on the sample mean and variance:

$$Y_i = \bar{y} + \varepsilon_i; \quad \varepsilon_i \sim N(0, s_Y^2), iid \quad \text{.....(1c)}$$

where s_Y is the sample standard deviation. Accepting the alternative model again entails assuming the best estimate:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i; \quad \varepsilon_i \sim N(0, s_{Y|x}^2), iid \quad \text{.....(2c)}$$

so that the practical choice is between models (1c) and (2c).

Similarly, the standard *F* test for the multiple linear regression model selects between the constant mean model (1a) and the model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i; \quad \varepsilon_i \sim N(0,\sigma^2), iid \quad .....(3)$$

Again, model (3) is only selected if it performs significantly better than (1a) in prediction within the data set.

4.2 A *t* test on a mean

In the example above it is much clearer that hypothesis testing is selecting between models than it is in the first test that most students encounter, a (two sided) test on a mean. Here the choice is between a null model

$$Y_i = \mu_0 + \varepsilon_i; \quad \varepsilon_i \sim N(0,\sigma^2), iid \quad .....(4)$$

and an alternative model

$$Y_i = \mu + \varepsilon_i; \quad \varepsilon_i \sim N(0,\sigma^2), iid; \quad \mu \neq \mu_0 \quad .....(5)$$

This is just as much a matter of model selection as is the regression example, but it is very rarely presented, or viewed, as such.

Accepting that the alternative model actually used has the parameter estimated from the sample, this test on the mean selects between (4) and:

$$Y_i = \bar{y} + \varepsilon_i; \quad \varepsilon_i \sim N(0,\sigma^2), iid \quad .....(5a)$$

The null and alternative **derived** models are, respectively:

$$\frac{\bar{Y} - \mu_0}{s_y/\sqrt{n}} \sim t_{(n-1)} \qquad \text{and} \qquad \frac{\bar{Y} - \bar{y}}{s_y/\sqrt{n}} \sim t_{(n-1)}$$

Based on the null model, the usual *p* value is:

$$p = P\left(|t| > \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \right)$$

and based on the alternative model, the alternative *p* value is:

$$p' = P\left(|t| > \left| \frac{\mu_0 - \bar{y}}{s/\sqrt{n}} \right| \right)$$

Note that this reflects what is usually considered a very common error among students. The fact that it is so common suggests that this is a natural way to view the process.

4.3 An *F* test on a ratio of variances

In the previous example the two *p* values are equal, due to the symmetry of the *t* distribution. For some tests this is not so. To test if two variances are equal, the choice is between two composite models. For a basic example, the null is:

$$\begin{aligned} X_{1i} &= \mu_1 + \varepsilon_{1i}; \quad \varepsilon_{1i} \sim N(0,\sigma^2), iid \\ X_{2i} &= \mu_2 + \varepsilon_{2i}; \quad \varepsilon_{2i} \sim N(0,\sigma^2), iid \end{aligned} \quad \text{.....(6)}$$

and the alternative:

$$\begin{aligned} X_{1i} &= \mu_1 + \varepsilon_{1i}; \quad \varepsilon_{1i} \sim N(0,\sigma_1^2), iid \\ X_{2i} &= \mu_2 + \varepsilon_{2i}; \quad \varepsilon_{2i} \sim N(0,\sigma_2^2), iid; \quad \sigma_1^2 \neq \sigma_2^2 \end{aligned} \quad \text{.....(7)}$$

If the sample ratio is $s_1^2 / s_2^2 = \theta$, the derived null and alternative models are, respectively:

$$S_1^2 / S_2^2 \sim F_{(n_1-1,n_2-1)} \qquad \text{and} \qquad S_1^2 / S_2^2 \sim \theta F_{(n_1-1,n_2-1)}$$

so (assuming $\theta > 1$) the two *p* values are:

$$p = P\left(S_1^2 / S_2^2 > \theta \right) = P(F > \theta) \text{ .....(8a)}$$

$$p' = P\left(S_1^2 / S_2^2 < 1 \right) = P(F < \frac{1}{\theta}) \text{.....(8b)}$$

These probabilities are not in general equal.

5. Some extensions

5.1 Critical values

In the absence of computers the *p* value of a test was at best difficult to obtain, so the use of tables of critical values developed as an effective substitute. The concepts of hypothesis testing can be developed in terms of critical values rather than *p* values, and in introductory texts this is commonly still done. This approach has some pedagogical advantages, but does make for greater complexity.

A minor disadvantage of its use is terminological. The *p* value of a test is a measure of the significance of the sample data. It is, if you like, a measure of one's level of doubt about using the alternative model in preference to the null model. This does not necessarily mean that it is being interpreted as a probability: simply that a large *p* value says that we should have considerable doubt about using it, while a small *p* value says that we should have little doubt. (As always, this is subject to provisos about the appropriateness of the model in the first place.) It is natural then to refer to the *p* value as the significance level of the test.

A critical value is the value of a test statistic corresponding to a **threshold *p* value;** that is, a threshold significance level. Unfortunately of course, the phrase 'significance level' has already been commandeered for the threshold *p* value.

On the other hand, a major disadvantage is the following. Because the tables were, naturally, created for a few standard threshold p values, these have been granted much greater importance than they warrant. At the extreme, results were considered to have been established, or not, according to whether they were significant at the 5% level, or at the 1% level. This misuse of hypothesis testing has led to results with $p = 0.049$ accepted, and those with $p = 0.051$ rejected, and is one the major reasons for the controversy of the last half century[4].

The use of a fixed threshold p value and a critical value is appropriate for an application which can reasonably be automated, such as industrial sampling, but not for research.

5.2 Errors

Neyman and Pearson approached hypothesis testing as an exercise in deciding whether the null hypothesis is true or not (as did Fisher), apparently in some absolute sense. In these terms, the concept of decision errors makes sense, and so the whole baggage of type I and II errors, significance level as the conditional probability of making a type I error, and power of a test follow.

To calculate the 'probability of making a type I error' requires a previously specified threshold level of significance, usually called simply the 'level of significance'. It is then easy to set up the decision as 'reject the null model if the null p value is less than the specified (threshold) value'. The probability of rejecting the null is trivially the specified level.

The probability of a type I error makes sense in the case of acceptance sampling which is automated with a fixed threshold p value, interpreted as a measure of expected risk. It is acceptable to deduce, regardless of one's interpretation of probability, that if $P(A) = 0.2$, say, then the expected number of errors in N trials is $0.2 \times N$, so the expected costs of this repeated process can be validly estimated.

In terms of the view promulgated here, one can at most define a type I 'error' as 'choosing the alternative model when the null model would probably work better'. Note that this does **not** say that the null model is true, or correct. Neither word applies to a model, particularly a probability model! Nor does it say that the null model is the **best** model.

How do we know that 'the null model would probably work better'? In practice, of course, we generally cannot. It may eventuate from subsequent research that a different choice may have led more easily to some understanding, or that the researcher has proceeded down a wrong path. It may turn out when the production line machinery was stopped that it wasn't really necessary to stop it at that time. It may turn out, from new evidence, that the person found

[4] The use of fixed rules such as this (and the $n > 30$ rule) bedevils the teaching of statistics. Statistics is about judgement, and iron clad rules such as these discount the role of judgement.

guilty of murder was (probably) innocent. But in general the decider will not know, at any useful time, whether he or she made the 'correct' decision.

In this light, the usefulness of the concept of 'probability of making a type I error' is far from clear.

5.3 One sided tests

Based on the Neyman-Pearson concept of two alternative hypotheses, one of which must be true, the alternative hypothesis is always composite, while the null should prescribe a specific value of the parameter whose value is being tested. For a two sided test, this occurs naturally, but for a one sided test it does not. If the alternative is $\mu > \mu_0$, the null must be $\mu \leq \mu_0$. Many textbooks concentrate on the need for a specified value at the expense of the underlying idea that one of the hypotheses must be true, and simply specify the null as $\mu = \mu_0$. More careful authors will give the correct null, and observe that not only is it privileged by being assumed true unless the sample evidence is strong enough to reject it, but that rejection will be made as difficult as possible by making the strongest possible assumption about its value : that the parameter takes its boundary value, μ_0.

In terms of the modelling view presented here, regardless of the sidedness of the test, the choice is between the null model with $\mu = \mu_0$, and the alternative model with $\mu = \overline{x}$

6. The hypothesis testing controversy

The purpose in this section is to discuss some of the statements made in the course of the controversy, in the light of the approached presented in the early part of this paper. I have selected a few of the most common and /or salient points made.

6.1 "The null is always false"

This only makes **any** sense if the null is a simple statement $\mu = \mu_0$; it makes no sense for a composite null, such as $\mu \leq \mu_0$. It also makes sense only in terms of the 'true value of the parameter'. But Statistics only deals with true values in very special and generally artificial cases. To 'accept the null' means that we conclude that we should base future actions (including 'hold opinion') on the null model – but we know perfectly well that this is a model, which is at best only approximately 'true'.

6.2 The replication fallacy

What practical meaning does *p* have as a probability? One interpretation is embodied in the **replication fallacy** (Gigerenzer, 1993, p.330). This says that if a test is repeated in replications of the investigation, a significant result would be obtained 1-*p* of the time; or in nonfrequentist terms, the investigator can have a confidence level of 1-*p* that the result would be significant if the investigation were repeated.

The fallacy here is because the probability is interpreted as unconditional. The *p* value says nothing about replication other than its definition. **If** the null model provides a good description of the real world, a significant *p* value is the probability of getting a significant *p* value (and rejecting the null model in error!)

If the result is declared not significant, on the basis of the sample evidence, the null model is at least as good as the alternative model, so is chosen. The probability of obtaining an equally significant (or insignificant) result on a replication is *p*.

If the result is declared significant, the alternative model is chosen. Using this model, the probability of getting an equally significant value on the repeated test can be estimated.

To illustrate, suppose we carry out a one tailed test on a mean, assuming normality and σ known to be 5, with the null hypothesis $\mu \leq 20$. A sample of 20 gives mean 23. The effective choice is between the null model with $\mu = 20$ and the alternative with $\mu = 23$. Then $z = 2.683$, with $p = 0.0073$. This would generally be considered significant, so we would accept the alternative model $\mu = 23$.

Now suppose we replicate the test, this time with a sample of 10^5.

If the original test was carried out with a specified threshold *p* value of 5%, say, 'significant' is defined as having a *p* value less than or equal to 5%. The critical value is 1.645. For the replication to be significant in testing the same null, this critical value corresponds to a sample mean of $20 + 1.645 \times 1.5811 = 22.601$.

If the null model is valid, the probability of a significant result on a replication is 5%. However on the basis of the previous test we believe the alternative model should be used, and on the basis of this, the probability of a significant result on the replication is

$$P(\bar{X} > 22.601) = P\left(Z > \frac{22.601 - 23}{1.5811}\right) = 0.5997$$

If the threshold *p* value is identified with the observed *p* value, 'significant' is defined as having a p value less than (or equal to) 0.0073. The critical sample mean is then 23, and the probability of a significant result under the alternative model is of course 0.5.

This of course is related to the Bayesian computation of inverse probability. It is worth noting that there is no observable relationship between the probability of a significant result under the null model, *p*, and that under the alternative model.

6.2 p value as 'worth of research'

Reputedly the *p* value has sometimes been taken as a measure of the value of a piece of research. A highly significant result is better than a less significant result, is better than a nonsignificant result.

[5] This example is similar to that used in Tversky and Kahneman (1993)

The *p* value is **not** a measure of the worth of the research. A nonsignificant result is not inferior to a significant one. First, even if the test is global for the whole investigation, it is simply choosing between two models. It is true that the alternative hypothesis is often that which is expected by the researcher, but this does not make the negative result uninformative.

A significant result in research is more highly prized in that it is seen as leading to developing theory, while an insignificant result may be identifying a dead end. In this sense a *p* value may be considered a measure of satisfaction, but not of worth or importance.

Finally, in most cases each test used in an investigation is very limited in its role. Much more important to determine the worth of the research are the design, the sample selection and the design of the measuring process.

6.3 Significance, power and effect size

There is an argument (Cohen 1994) that if the sample is large enough – that is, the test has sufficient power - any effect will be significant, so that rejecting a null means only that you have a large enough sample to do so: so why carry out a test? This may be based on the notion that 'the null hypothesis is always false', but has some validity in its own right.

Although valid, this observation should be regarded as a warning against foolish practices rather than a rejection of hypothesis testing. First, experimentation costs money, and one should only take a large sample when necessary. As in any field, power costs, so use only what you need. Hence, it is wise to use hypothesis testing in its converse form – to calculate the sample size which will give the power expected to identify a significant effect of specified size.

Second, unless the sample is very large indeed, it is highly likely that it will not be known that the effect is significant, so there is an argument for 'checking it out'. Third, if an effect is sufficiently small that only a very powerful test will identify it as significant, it is hardly large enough to be useful!

A significance test helps the researcher to decide if it is reasonable to accept the alternative model on statistical grounds. But an effect has to have both statistical significance and practical significance in order to be adopted. The latter means simply that the effect size is large enough to be worth using. For example, if research shows that the difference in recovery rates between a drug and a placebo is only 1%, this will be statistically significant if the sample is large enough, but is not enough of a difference to warrant spending money on the drug. Conversely, if the difference in recovery rates is 20%, it would certainly seem reasonable to put money into the drug – but very unwise if the result is not statistically significant.

In short, it is essential to avoid confusing the statistical and nonstatistical arguments.

6.4 The modus tollens argument

Cohen (1994) presents an argument that modifying the modus tollens argument:

> If the null hypothesis is correct, the result D cannot occur.
> Result D has occurred.
> Therefore the null hypothesis is false.

to be probabilistic:

> If the null hypothesis is correct, the result D is very unlikely.
> Result D has occurred.
> Therefore the null hypothesis is very unlikely.

renders it invalid.

Cohen observes that the argument above amounts to concluding that if $p = P(data|H_0)$ is small, then the inverse probability $P(H_0|data)$ is small; further, people are likely to think that if p is less than the threshold significance level (0.01, say) then so is the inverse probability. He then uses an example, essentially a simple exercise in conditional probability, to calculate the latter as a posterior probability, and demonstrates that the inverse probability need not be small. Note that this calculation only makes sense if 'the probability that the hypothesis is true' makes sense

In terms of the predictive/modelling view of hypothesis testing, this argument becomes:

> If the null model is assumed, the result D is very unlikely.
> If the alternative model is used, the result D is very likely.
> Result D has occurred.
> Therefore the alternative model is likely to give better results than the null.

Cohen's example involved a proposed screening test for schizophrenia, for which the incidence in adults is about 2%. The test is such that if a randomly selected adult is tested, the probability of a (correct) positive result (indicating the person has schizophrenia) when he actually has it is greater than 0.95, while the probability of a correct negative result when he actually does not is about 0.97. Treating a single test as a hypothesis test, with the null model being that the person is normal and the sample result positive, we have $P(data|H_0) < 0.05$, leading one to conclude at the 5% level that the person has schizophrenia. He then computes the probability $P(H_0|data) = 0.60$ approximately. That is, with a positive result, the patient probably does not have schizophrenia.

In terms of the predictive view, the purpose of identifying a relationship such as that between the test result and the existence of schizophrenia is to improve predictability, in this case predictability of the existence of schizophrenia. For a nominal variable, the measure of predictive quality is usually taken to be the probability of error, so the best prediction is the mode (McLean 2000). With this in mind, the relevant probabilities are P(has schizophrenia) =

0.02 and P(has schizophrenia | positive result) = 0.40. So the test does improve the predictability, so is useful.

The fact that the patient is in fact more likely not to have schizophrenia simply says that the test is not as good as we would like it to be.

Cohen concluded that this example 'demonstrates how wrong one can be by considering the *p* value from a typical significance test as bearing on the truth of the null hypothesis for a set of data.' This is true, but the flaw is in the misconception as to what a hypothesis test really does.

7. Conclusions

In this paper, a view of hypothesis testing has been presented which is based on the idea that all statistical analysis is concerned with modelling, and with choosing the best model in any situation. The 'best' model is that which, on the sample evidence, is expected to give the best forecasts. Hypothesis testing is a way of identifying the better model of two alternatives, in which the model suggested by the sample data will only be selected if the sample evidence is sufficiently strong. The strength of this evidence is measured by the *p* value. This is calculated as a conditional probability, although interpreting this in applications as a probability is frequently not necessary and can lead to considerable misunderstanding.

In applications such as acceptance sampling, where the selection of the model entails an action to be carried out, it is appropriate to automate hypothesis testing by specifying a threshold *p* value and adopting the decision based approach of Neyman and Pearson. In research, this is not appropriate because, while a choice is made, it is tentative. In research, any hypothesis test is only one tool used as part of an argument; each hypothesis test mimics the total piece of research in assessing the evidence in favour of a model.

Taking the predictive view of statistics, a hypothesis test compares two specific models. This is a natural approach, consistent with the observation that informal use of probability and hypothesis testing is a natural mode of thinking, and with the reactions of students.

The confusion over the role of hypothesis testing in research appears to have sprung partly from its overenthusiastic adoption, an unwise application of the Neyman-Pearson approach to research – and to a lack of appreciation that research, and statistics, deals with models, rather than with the 'real' world'.

Acknowledgements

This project was supported by a grant from the Faculty of Business and Economics, Monash University.

References

American Psychological Association (1994). *Publication Manual of the American Psychological Association* /4th ed.). Washington, D.C.: A.P.A.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997-1003.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83-96.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A 222*, 309-368.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society* (B), *17*, 69-77.

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 behavioral 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, N.J.: Erlbaum.

Lehmann, E.L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science*, 5 (2), 160-168.

Lehmann, E.L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, *88* (424), 1242-1249.

McLean, A. L. (1998), 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. (1999). 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.

McLean, A. L. (2000). The Predictive Approach to Teaching Statistics. *Journal of Statistics Education*, 8(3).

McLean, A. L. (2001). Statistics on the catwalk – the importance of models in training researchers in statistics. In C. Batanero (Ed.), *Training Researchers in the Use of Statistics.* Granada, Spain. International Association for Statistical Education.

McLean, J.E. & Ernest, J.M. (1998). The role of statistical significance testing in educational research. *Research in the schools (Special Issue: Statistical Significance Testing) 5*(2), 15-23.

Melton, A.W., (1962) Editorial, *Journal of Experimental Psychology*, *64*, 553-557

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

Pollard, P. & Richardson, J.T.E. (1987). On the probability of making Type I errors. *Psychological bulletin*, *10*, 159-163.

Shaughnessy, J.M (1983). The psychology of inference and the teaching of probability and statistics: two sides of the same coin? in R.W. Scholz (Ed.) *Decision making under certainty: cognitive decision research, social interaction, development and epistemology.* Amsterdam: North-Holland.

Sterne, J.A.C. & Davey Smith, G. (2001), Sifting the evidence – what's wrong with significance tests?, *British Medical Journal 322*(322) 226-231.
Web: http://bmj.com/cgi/content/full/322/7280/226

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher***,** 25(2), 26–30.

Thompson, W.L. (2001). *Bill Thompson's reference list "against" and "for" the use of null hypothesis testing.* Web page: http://www.cnr.colostate.edu/~anderson/null.html.

Tversky, A. & Kahneman, D. (1993). Belief in the law of small numbers. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.