Appl. Statist. (1996)
45, *No.* 4, *pp.* 401–410

An Applied Statistician's Creed

By MARKS R. NESTER†

Queensland Forestry Research Institute, Gympie, Australia

[Received March 1994. Final revision June 1996]

SUMMARY

Hypothesis testing, as performed in the applied sciences, is criticized. Then assumptions that the author believes should be axiomatic in all statistical analyses are listed. These assumptions render many hypothesis tests superfluous. The author argues that the image of statisticians will not improve until the nexus between hypothesis testing and statistics is broken.

Keywords: Analysis of variance; Confidence limits; Hypothesis testing; Statistical assumptions

1. Introduction

According to Hacking (1965), Arbuthnott (1710) was the first to publish a test of a statistical hypothesis. Hogben (1957), chapter 14, p. 324, attributed to Gavarret (1840) the earliest use of the probable error as a form of significance test in the biological arena. Hogben (1957), chapter 14, p. 325, also stated that Venn (1888) was one of the earliest users of the terms 'test' and 'significant'. The form of the χ^2-distribution in the context of goodness-of-fit tests was published by Pearson (1900). W. S. Gosset, using the pseudonym 'Student' (1908), developed the t-distribution. The foundations of modern hypothesis testing were laid by Fisher (1925), although the modifications propounded by Neyman and Pearson (1933) are the generally accepted norm.

Today, the ritual testing of hypotheses is performed in many applied sciences and their respected journals. Not surprisingly, many statistics journals are replete with applications of null hypothesis tests or with theory that permits new applications of hypothesis tests. Tests of hypotheses are seemingly performed because

(a) they appear to be objective and exact,
(b) they are readily available and easily invoked in many commercial statistics packages,
(c) everyone else seems to use them,
(d) students, statisticians and scientists are taught to use them and
(e) some journal editors and thesis supervisors demand them.

Although many practising statisticians eventually come to the realization that these tests are highly overrated, there are still many scientists and statisticians

†*Address for correspondence*: Queensland Forestry Research Institute, MS 483, Fraser Road, Gympie 4570, Australia.
E-mail: marks@qfri.se2.dpi.qld.gov.au

0035–9254/96/45401

who promote the statistical testing of hypotheses as being integral to the scientific method and scientific argument. This paper is yet another attack on the ubiquitous hypothesis test, and several issues are discussed:

(a) the attitude of many scientists towards hypothesis tests;
(b) the fact that many people equate statistical methods with hypothesis testing;
(c) a 'creed' for the practising statistician.

2. Null Hypothesis Testing

Consider the simple case of comparing two treatment means. A researcher will often have some idea of which size difference between the treatment means has practical importance, but invariably the analysis will be a test of H_0: $\mu_2 = \mu_1$ *versus* H_1: $\mu_2 \neq \mu_1$. If the researcher has no preferred outcomes for the experiment then Table 1 depicts the researcher's possible states of mind. However, if the researcher is actually testing some pet theory then an additional layer of complexity should be superimposed on Table 1.

If the researcher is allowed some latitude then the discussion of the results will be along one of the following themes.

(a) *If a significant result is obtained*
  (i) This is good. It fits in with the researcher's hopes and/or expectations. The true difference between the treatment means is equated to the observed difference.
  (ii) This is bad. The researcher did not want this, but it is satisfactory because the difference between the observed means is of no practical importance anyway.
  (iii) This is bad. The researcher did not want this, and does not know how to explain it. Maybe it is just one of those type I errors.

(b) *If a non-significant result is obtained*
  (i) This is good. It fits in with the researcher's hopes and/or expectations. The true difference between the treatment means is now equated to 0. The researcher's great pleasure that a type I error did not occur is not revealed, nor the great pleasure that a larger sample could not be afforded.
  (ii) This is bad. The researcher did not want this, and does not know how to explain it. The difference between the observed means is of considerable practical importance. Maybe it is one of those type II situations, and the

TABLE 1
States of mind of a null hypothesis tester

Practical importance of observed difference	*Statistical significance of difference*	
	Not significant	*Significant*
Not important	Happy	Annoyed
Important	Very sad	Elated

researcher would have obtained a significant result had more funds been available to increase the sample size.

Observe that many researchers equate the true difference to the observed difference, or to 0, depending on the outcome of the significance test and the prior expectations. However, if a statistically significant result is achieved then there are some researchers who will incorporate confidence limits or standard errors in their discussions.

A researcher who is given latitude will generally defend a scientific hypothesis by appealing to scientific principles and rational argument. An accompanying statistical analysis either provides support for the researcher's position, which is naturally regarded as good, or does not provide support, which causes the analysis to be regarded as a nuisance and to be effectively dismissed and ignored.

In contrast, a researcher who has been constrained to adopt a formal hypothesis testing approach is simply not permitted to assert that a treatment difference exists unless it is accompanied by a statistically significant result. This has led to the absurd situation in which a journal editor has confided that an author's thesis is undoubtedly true, but the editor must reject the paper because the author's ideas are not supported by statistically significant results. Even ignoring the implications for publication bias, is this good science, or is it statistics gone mad? Conjecture and speculation, the life-blood of science, may be carefully expunged from a scientific paper, and yet a purely mathematical paper in the field of number theory may be riddled with conjectures.

Sadly, the discussion in many scientific papers revolves around the statistical analysis instead of the science. The following problems exist.

(a) Hypothesis tests accompany most statistical analyses and often constitute the sole analysis.
(b) Poor understanding of hypothesis tests often leads to their misinterpretation. In particular, the roles of sampling variation and sample size are often obscured.
(c) The hypothesis tests usually deal with silly null hypotheses which assert no differences between treatments.

With regard to this last point, many scientists would like to see the roles of the null hypothesis and alternative hypothesis reversed, i.e. they desire tests in which the null hypothesis is H_0: $\mu_2 \neq \mu_1$ and the alternative hypothesis is H_1: $\mu_2 = \mu_1$.

3. The Creed

Many analyses of variance indicate no significant differences between treatments, yet the known biology, physics or chemistry of the situation suggests that the treatments cannot have identical effects. Whether or not the reasons for a lack of statistical significance are attributed to insufficient replication, should the researcher adopt a formal statistical approach and not reject the null hypothesis? Or can the researcher bypass the test of significance and report confidence limits, say, for the treatment differences? Of course the confidence limits will include 0, but the upper confidence limit may be the key number since it is an indication of the maximum plausible difference between the treatment means. Despite the apparent reasonableness of the latter approach, there are some who either do not understand statistics or do not understand science, or do not understand either, who rigidly adopt the formal approach. It is frustrating to inform a scientific client that statistically the client cannot assert that there are differences between treatments when any reasonable person knows that differences must exist.

A natural question is whether different treatments can ever have identical effects. My answer is presented in the following creed which also incorporates some well-known statistical wisdom.

Each statement of the nonalogue has an associated keyword:

(a) TREATMENTS—all treatments differ;
(b) FACTORS—all factors interact;
(c) CORRELATIONS—all variables are correlated;
(d) POPULATIONS—no two populations are identical in any respect;
(e) NORMALITY—no data are normally distributed;
(f) VARIANCES—variances are never equal;
(g) MODELS—all models are wrong;
(h) EQUALITY—no two numbers are the same;
(i) SIZE—many numbers are very small.

4. Explanation of the Creed

The TREATMENTS, FACTORS, CORRELATIONS and POPULATIONS beliefs apply to hypothesis testing of the ‘zero difference’ or ‘nil existence’ variety. Throughout this paper such tests are referred to as zero hypothesis tests, being a subset of more general null hypothesis tests which hypothesize a non-zero difference.

Many tests of hypotheses assume normality, linearity or constant variance. NORMALITY, MODELS and VARIANCES simply assert that such assumptions are always false. The NORMALITY and VARIANCES statements are special cases of the MODELS belief.

EQUALITY and SIZE are really statements of mathematical fact rather than subjective beliefs. In many respects EQUALITY is a synopsis of TREATMENTS, FACTORS, CORRELATIONS and POPULATIONS. For those who consider the creed to be too extreme, a measure theoretic interpretation can be given to EQUALITY. From this viewpoint, if f and g are real numbers then it is impossible for f to equal g, in the sense that $\Pr(f = g) = 0$. An explanation of how the numbers f and g are obtained is avoided. On this interpretation, two treatments can indeed have the same effect, in the sense that such treatment pairs exist, but that the probability of encountering such a pair of treatments is 0!

Continuity and non-finiteness are essential ingredients of the creed. Thus the creed applies to situations where it is assumed that all observations derive from continuous scales of measurement, or that observations derive from hypothetically infinite populations. For example, the POPULATIONS belief is not applicable if population D is defined to be the group of people who work in a particular building, population E is defined to be the group of people who work in a different building and it is the percentages of blind people which are being compared. In contrast, POPULATIONS would be operative if the same populations were being compared for mean height, assuming that height is a continuous variate. Note that continuity and hypothetically infinite population sizes are assumptions which are made in many statistical estimation, testing and modelling situations.

5. In Defence of the Creed

Surely there is no one among us who believes that a sample of data from a normal distribution has ever existed. One hopes that the MODELS belief is also universally held to be true. Any analyst who has fitted a straight line through some data has either done so knowing that it was only a reasonable approximation to the true relationship or has remembered the dire text-book warnings of extrapolation beyond the range of the data.

VARIANCES refers to population variances about regression surfaces, or to variances of error random variables in the analysis of variance, and so on. Presumably a large proportion of statisticians already adhere to this belief as a result of their training and/or experience. The remainder may be swayed by considering

(a) what might happen to their regression variances if the data ranges of the independent variables are expanded,
(b) whether or not treatments really have no effect on the variances of observations,
(c) whether or not machine components wear through time and consequently cause variances to change through time,
(d) whether or not soil properties vary through space and consequently affect variances of the observations, and so on.

In any hypothesis testing situation, a reason can be found why a small difference might exist or a small effect may have taken place. The creed merely asserts that the *a priori* existence of these effects should be the foundation of any statistical analysis.

With regard to randomization in his tea tasting experiment, Fisher (1935) wrote

> 'It is no sufficient remedy to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation.'

Fisher is stating that no two experimental units can be exactly alike. Presumably no reasonable statistician or scientist has ever disputed this, yet it seems somewhat irrational to accept Fisher's claim and yet to assume that different treatments might have identical effects.

6. Applying the Creed

Acceptance of the creed forces a data analyst to focus on the important issues, and it reminds the analyst that there are many assumptions which must be examined to ensure that the analysis is sound and appropriate.

Here are some examples of the creed in action.

(a) How many replications should be used in an experiment to test whether two treatments have different effects?

Response. Obviously one need not use any replications, nor even do the experiment, since the treatments must have different effects.

(b) A client wants to establish a relationship for predicting one variable from another.

Response. CORRELATIONS, NORMALITY, VARIANCES and MODELS all impact on this situation and many books on regression analysis deal with the issues very well. Of course a major question is why does anyone ever bother to test whether or not the slope, assuming linearity, is 'significantly different' from 0? Does it matter if the client wants to know whether or not the slope is different from 1, say? No, for we can invoke EQUALITY and immediately assert that the slope is different from 1. Often the client will simply want to make predictions and our main task should be to inform the client about the sizes of the likely prediction errors for some range of values of the predictor.

(c) A client wants to know whether two regression equations are significantly different.

Response. Clearly the true (unknown) regression equations cannot possibly have the same slope (POPULATIONS or EQUALITY), or the same intercept, so who cares whether or not the equations are significantly different when a statistical test is applied? However, one should persuade the client to examine the consequences of using a combined equation instead of two separate equations. Perhaps the simplest way of doing this is to choose an appropriate range of predictor values and to compute the maximum absolute difference between predictions obtained from the single combined equation and from one or other of the two separate equations.

(d) You have just completed an analysis of variance. Because of TREATMENTS and FACTORS, you know that the exercise was futile, but it gave you the base information to initiate a multiple-comparison procedure. How do you continue?

Solution. If all that you want to do is to ascertain which treatments differ from certain other treatments, then obviously you do not do anything because all treatments differ from each other. Unless you are investigating matters such as sizes of differences then you are wasting your time. One thinks back to the poor old least significant difference (LSD) test, much maligned because it predicts too many significant differences when the true treatment means are all the same. This theory of too many significant differences is sound, but the problem is that, in practice, no two treatment means are the same, if one accepts the creed. If the LSD procedure generates a greater number of significant differences than does any other procedure, then it must arguably be the best procedure!

(e) The POPULATIONS belief renders many population tests and comparisons obsolete. Do men and women have the same average intelligence? The answer is no. Do smokers and non-smokers have identical lung cancer rates? The answer is no. Tests that address these and many similar questions are pointless.

Clearly, point hypothesis testing has no place in statistical practice if we adopt the creed. This means that most paired and unpaired t-tests, analyses of variance (except for estimating variance components), linear contrasts and multiple comparisons, and tests of significance for correlation and regression coefficients should be avoided by statisticians and discarded from the scientific literature. Standard deviations and confidence limits are some of the concepts that should be promoted and that the scientific community should be publishing. There is absolutely nothing wrong with mathematical statisticians studying and writing about null hypothesis testing; it is just that statisticians and others should realize that many tests have either limited or no worthwhile application in practice.

7. Some Arguments against Hypothesis Testing

As the examples in the previous section show, the creed bypasses kinds of questions such as 'Are the treatments different?' and proceeds immediately to questions such as 'How different?'. Since this is merely normal statistical practice but with zero hypothesis testing excluded, I shall raise some issues that I hope will validate this approach.

(a) To proceed one way on the basis of a statistically significant result, and a different way on the basis of a non-significant result, is to react to a result which depends not only on the size of a true difference, say, but also on inherent variation and sample size. The probability associated with a significance test is nothing more than a reflection of the power of the test. Surely all of us have encountered many situations where we believe that we can turn a non-significant result into a significant result merely by sufficiently increasing the sample size. Rejectors of the creed will entertain some situations where they think that this will not happen—so be it. However, rejectors should realize that, if a non-zero difference or effect exists, then when they say

> 'I reject the null hypothesis because my probability is small'

they imply

> 'I reject the null hypothesis because my test is powerful'.

Similarly,

> 'I do not reject the null hypothesis because my probability is large'

implies

> 'I do not reject the null hypothesis because my test is not powerful'.

The tragedy can be heightened with a little rewording.

> 'I reject the null hypothesis because I collected enough samples.'
>
> 'I accept the null hypothesis because I did not collect enough samples.'

(b) To reverse many scientific conclusions reached during the past 60 years of hypothesis testing is trivially easy. Take any small number—10^{-10} will suffice—and substitute it for 0 in all zero hypothesis tests. Now, in almost every instance where a zero difference or effect was concluded, we shall have substantiated the presence of a non-zero effect. It may be claimed that a negligible effect has merely been substituted for a zero effect—that is true, but it works both ways—why cling to a zero effect when a negligible effect may be possible and plausible?

(c) In a situation where a zero hypothesis test is to be performed on conceptually continuous data, consider the maximum likelihood estimate, *e* say, of the effect under investigation. The estimate *e* can reasonably be assumed to be different from 0. If the likelihood function is symmetrical about its maximum, then there is a greater 'likelihood' that the true effect lies in the inverval $e \pm e/2$ than in the interval $0 \pm e/2$. Why, then, should anyone performing a zero hypothesis test conclude that an effect is 0 when it is more 'likely' that the true effect lies in some interval which excludes 0? The case for non-symmetrical density functions entails different intervals, but the same principle applies.

(d) Every null hypothesis test represents a loss of information, or at least is a waste of information. Incorporating probability levels in the significance statements does not dramatically enhance the information content relative to that contained in confidence limits or standard errors. In fact the latter quantities enable the researcher to ascertain directly the sensitivity of the data, i.e. the size of error relative to the size of estimate.

8. Statisticians, Statistics and Hypothesis Testing

Statistics as an art or science or mere collection of figures will have a generally bad image for as long as statistics is placed in a category worse than 'damned lies'. The valid cry that 'you can't prove anything with statistics' certainly does not enhance the image of statistics or statisticians. The popular cry that 'you can prove anything with statistics' also does not help us! In much scientific work, the formal tests of zero hypotheses are the only statistical analyses that researchers are expected to perform. Many researchers regard such tests as time wasting irrelevant burdens that they must endure to ensure publication. Since statisticians apparently provide the tests, and since the tests are regarded as aspects of statistics, it is little wonder that the image of statistics and statisticians continues to suffer.

Although 'significant' has a special technical meaning in the context of hypothesis testing, it is most unfortunate that, because of a hypothesis test, a treatment difference of no practical import whatever is described as being significant. Who promotes this unfortunate choice of words?: statisticians!

Scientists were the first to initiate hypothesis testing, and I think that it would be wise to lay all credit and blame for hypothesis testing squarely at their feet. Fisher and others who laid the mathematical foundations of hypothesis testing in response to requests from scientists deserve the highest praise for their mathematical prowess, but the continued promotion of such tests by generations of statisticians should be deplored. With the awakening of many scientists and statisticians to the problems that are inherent in hypothesis testing, those wearing hypothesis tester hats may eventually regret their stances. We statisticians should involve ourselves with the design of experiments, parameter and error estimation, and model building. Continued association with hypothesis testing is not in our own best interest. I believe that statisticians would be unwise to seek the limelight in any forthcoming 75th anniversary, centennial or tricentennial celebrations of hypothesis testing.

9. Quotations

Hypothesis testing is so entrenched in many applied sciences, statistical text-books and statistics courses that yet another attack on hypothesis testing can only do good. The following quotations in chronological order are offered as support for the views expressed in this paper. A more extensive compilation of quotations is available from the author.

Neyman and Pearson (1933): 'if x is a continuous variable . . . then any value of x is a singularity of relative probability equal to zero. We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.'

Geary (1947): 'Normality is a myth; there never was, and never will be, a normal distribution'.

Yates (1951): 'the emphasis given to formal tests of significance . . . has resulted in . . . an undue concentration of effort by mathematical statisticians on investigations of tests of significance applicable to problems which are of little or no practical importance . . . and . . . it has caused scientific research workers to pay undue attention to the results of the tests of significance . . . and too little to the estimates of the magnitude of the effects they are investigating'.

Savage (1957): 'Null hypotheses of no difference are usually known to be false before the data are collected . . . when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science'.

Kish (1959): 'Significance should stand for meaning and refer to substantive matter. . . . I would recommend that statisticians discard the phrase "test of significance".'

Rozeboom (1960): 'the stranglehold that conventional null hypothesis significance testing has clamped on publication standards must be broken'.

Bakan (1967), chapter 1, p. 7: 'there is really no good reason to expect the null hypothesis to be true in any population . . . . Why should any correlation coefficient be exactly .00 in the population? . . . why should different drugs have exactly the same effect on any population parameter . . . ?'

Nelder (1971): 'multiple comparison methods have no place at all in the interpretation of data'.

Box (1976): 'all models are wrong'.

Chew (1980): 'I have tried to steer them [agricultural researchers] away from testing H_0. I maintain that on *a priori* physical, chemical and biological grounds, H_0 is always false in all realistic experiments, and H_0 will always be rejected given enough replication.'

Nelder (1985): 'the grotesque emphasis on significance tests in statistics courses of all kinds . . . is taught to people, who if they come away with no other notion, will remember that statistics is about tests for significant differences. . . . The apparatus on which their statistics course has been constructed is often worse than irrelevant, it is misleading about what is important in examining data and making inferences.'

Pearce (1992): 'In a biological context interactions are common, so it is better to play safe and regard any appreciable interaction as real whether it is significant or not'.

Wang (1993): 'the tyranny of the N–P [Neyman–Pearson] theory in many branches of empirical science is detrimental, not advantageous, to the course of science'.

Acknowledgements

I thank Tony Costantini, Patricia Nester, Syd Perry, Paul Ryan, Jane Siebuhr, Debbie Street, Ross Wylie, two referees and two editors of this journal for their comments on various drafts of this paper. Neither these reviewers nor the Queensland Forestry Research Institute necessarily endorse any of the views expressed herein.

References

Arbuthnott, J. (1710) An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Phil. Trans. R. Soc.*, **23**, 186–190.

Bakan, D. (1967) *On Method.* San Francisco: Jossey-Bass.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.

Chew, V. (1980) Testing differences among means: correct interpretation and some alternatives. *HortScience*, **15**, 467–470.

Fisher, R. A. (1925) *Statistical Methods for Research Workers.* London: Oliver and Boyd.

——— (1935) *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Gavarret, J. (1840) *Principes Généraux de Statistique Médicale.* Paris.

Geary, R. C. (1947) Testing for normality. *Biometrika*, **34**, 209–242.

Hacking, I. (1965) *Logic of Statistical Inference.* Cambridge: Cambridge University Press.

Hogben, L. (1957) *Statistical Theory.* London: Allen and Unwin.

Kish, L. (1959) Some statistical problems in research design. *Am. Sociol. Rev.*, **24**, 328–338.

Nelder, J. A. (1971) Discussion on the papers by Wynn and Bloomfield, and O'Neill and Wetherill. *J. R. Statist. Soc.* B, **33**, 244–246.

——— (1985) Discussion on The initial examination of data (by C. Chatfield). *J. R. Statist. Soc.* A, **148**, 238.

Neyman, J. and Pearson, E. S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc.* A, **231**, 289–337.

Pearce, S. C. (1992) Data analysis in agricultural experimentation: II, Some standard contrasts. *Exptl Agric.*, **28**, 375–383.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated systems of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag. Ser. V*, **1**, 157–175.

Rozeboom, W. W. (1960) The fallacy of the null hypothesis significance test. *Psychol. Bull.*, **57**, 416–428.

Savage, I. R. (1957) Nonparametric statistics. *J. Am. Statist. Ass.*, **52**, 331–344.

Student (1908) The probable error of a mean. *Biometrika*, **6**, 1–25.

Venn, J. (1888) Cambridge anthropometry. *J. Anth. Inst.*, **18**, 140–154.

Wang, C. (1993) *Sense and Nonsense of Statistical Inference.* New York: Dekker.

Yates, F. (1951) The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *J. Am. Statist. Ass.*, **46**, 19–34.