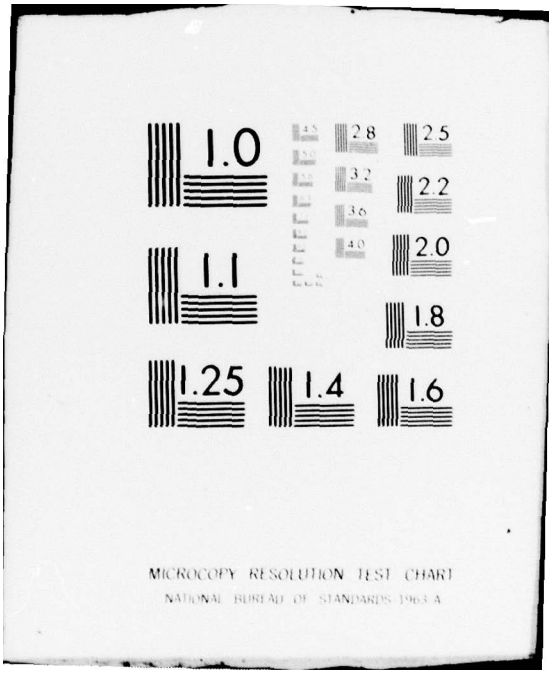
AD-A079 739 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
WHY DO WE NEED SIGNIFICANCE LEVELS.(U)
OCT 79 T LEONARD
UNCLASSIFIED MRC-TSR-2004 DAAG29-75-C-0024
NL

| OF |
ADA
079 739

END
DATE
FILMED
2-80
DDC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

MRC Technical Summary Report #2004

(3)

WHY DO WE NEED SIGNIFICANCE LEVELS?

Tom Leonard

LEVEL II

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

October 1979

(Received August 7, 1979)

Sponsored by

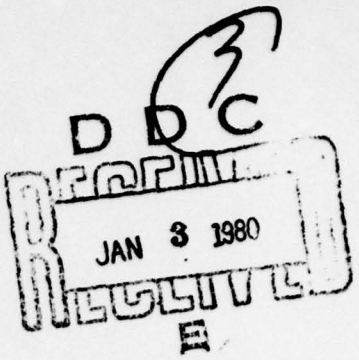
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

WHY DO WE NEED SIGNIFICANCE LEVELS?

Tom Leonard*

Technical Summary Report #2004
October 1979



ABSTRACT

Classical significance tests depend upon a choice of significance level and also seem overready to reject the null hypothesis when the sample size is large. A plausible alternative to significance testing has been suggested by Schwarz (1978) in the context of model discrimination. A simpler and more general formulation is discussed here; this leads to more precise approximations. For a single parameter and when the sample size n is large we recommend viewing the data as supporting a simple null hypothesis versus a completely composite alternative whenever the maximum likelihood estimate lies within an adjustment to $\sqrt{\log n}$ approximate standard deviations of the null hypothesis. This criterion indeed appears to provide an attractive rule of thumb for all sample sizes: It removes the need for tables of significance levels and becomes less keen to reject the null hypothesis for large sample sizes. The ideas are extended to provide alternatives to multivariate likelihood ratio tests, and to the chi-squared goodness of fit test.

AMS(MOS) Subject Classifications: Primary 62G10, Secondary 62A15

Key Words: Significance, Bayes, Asymptotic normality, Likelihood ratio, Bayes factor, Chi-squared

Work Unit Number 4 - Probability, Statistics, and Combinatorics

* Department of Statistics, University of Warwick, Coventry, CV4 7AL Warwickshire.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

SIGNIFICANCE AND EXPLANATION

Significance tests are commonly used in many application areas as attempts to formally confirm or refute specific conclusions. For example, in the social sciences (e.g. psychology, sociology, and econometrics) there is often much more emphasis on data-fitting and seeking "significant" results than on developing proper mathematical models which relate in an inductively sensible way to the real-life problem. However, significance tests do not possess too much formal justification in the literature for making specific decisions as to whether a particular hypothesis is true.

In the present paper a new formulation is used to demonstrate that significance tests tend to be much too ready to reject the null hypothesis for large sample sizes. It is recommended that the usual percentage points should be replaced by quantities depending in a particular way upon sample size, but not upon a choice of significance level. The phenomena discussed would appear to be particularly relevant to the area of scientific reporting. For example, many results in applied journals which might have been viewed as "significant," because they yield a low p-value, may in fact serve to detract from the very scientific theory which they claim to substantiate.

For large sample sizes, the techniques proposed in this paper permit a larger range of viable null hypotheses than experienced under fixed-size significance testing. It should therefore be easier to use them to find a data-credible model which is also reasonable in real-life terms.

WHY DO WE NEED SIGNIFICANCE LEVELS?

Tom Leonard*

1. Introduction

Consider firstly the situation where a single parameter θ possesses a likelihood function $\ell(\theta|\underset{\sim}{x})$ depending upon a vector $\underset{\sim}{x} = (x_1, \ldots, x_n)$ of n observations, and where θ assumes values in a continuous parameter space Θ. We assume that it is desired to test the simple null hypothesis $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \neq \theta_0$ or to find some viable alternative to this procedure.

There are two main questions which the statistician may wish to ask himself in this situation, namely

(A) Should the information in the data have a positive or negative influence upon his judgement about the truth of H_0?

(B) Is the information in the data in sufficient conflict with H_0 to suggest that he should take the step of rejecting H_0?

It appears to us that question (B) can only be adequately answered within a decision-making framework where the statistician assigns utilities to θ, under H_0 and H_1; see for example Dickey (1968, 1975) and one of the methods discussed in section 4. We would indeed go so far as to view it as unfair to expect classical significance tests to formally cope with (B). It should be remembered that significance tests were originally introduced as inductive tools for the working statistician and were viewed as alternatives to the decision-theoretic approach of Wald. For example, the sampling probabilities obtained could be interpreted in the context of the real-life problem. Significance tests possess little sensible justification in the literature for

* Department of Statistics, University of Warwick, Coventry, CV4 7AL Warwickshire.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

formally coping with decisions regarding actual acceptance or rejection, and should therefore not be expected to provide an adequate formal answer to the decision-making problem. In section 4 it will indeed be noted that they provide very different answers to those suggested by a sensibly-formulated decision-theoretic approach.

Question (A) seems to be of frequent importance to statisticians thinking inductively about their data sets; this is probably one question which many statisticians would like to answer when they employ significance tests. The main claim of this paper is however that classical significance testing should not be used to formally answer (A) or (B) unless the size of the test is permitted to depend upon the size n of the sample in a particular way. This might lead the reader to question the usefulness of significance tests. In our opinion the latter assume an over-prominent position in statistical methodology. In the long term, substantial changes to the teaching of significance tests, and to their widespread applications e.g. in the social sciences, might perhaps be beneficial.

A simple criterion is now introduced for answering (A). Suppose that the statistician possesses a prior density $\pi(\theta)$ for θ and denote the corresponding posterior density by $\pi(\theta|\underset{\sim}{x})$. Then consider the definition

Definition: The data support H_0 with respect to π if

$$\pi(\theta_0|\underset{\sim}{x}) > \pi(\theta_0) \tag{1.1}$$

This provides a natural criterion for answering (i), as long as the prior can be specified; note that if the data support H_0 then the probability of θ lying within a small neighbourhood of θ_0 will be increased by the information provided by the data.

One of the few criticisms that can be levelled at (1.1) by supporters of any philosophy of statistics is that it is dependent upon the choice of prior

π for θ. We will however demonstrate that as n increases the effect of the prior decreases, leading to a sensible approximate procedure which is completely free from the choice of prior distribution. For small sample sizes we feel that our asymptotic procedure will still provide a useful rule of thumb in situations where the prior information about θ is fuzzy and difficult to specify.

2. Asymptotic Results

As n gets large the posterior distribution of θ will be (typically) asymptotically normal with mean equal to the maximum likelihood estimate $\hat{\theta}$ of θ and variance v/n where

$$nv^{-1} = \left. - \frac{\partial^2 \log \ell(\theta|\underset{\sim}{x})}{\partial\theta^2} \right|_{\theta=\hat{\theta}} \tag{2.1}$$

Precise regularity conditions for this approximation, based upon the idea of supercontinuous likehoods, are described by De Groot (1970, p. 210). Substituting the corresponding approximation for $\pi(\theta_0|\underset{\sim}{x})$ in (1.1) tells us that as $n \to \infty$ the data support H_0 whenever

$$\frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}v^{\frac{1}{2}}} \exp\{-\frac{1}{2} nv^{-1}(\theta_0 - \hat{\theta})^2\} > \pi(\theta_0) \tag{2.2}$$

A slight rearrangement leads to

$$\frac{|\hat{\theta} - \theta_0|}{n^{-\frac{1}{2}}v^{\frac{1}{2}}} < \{\log n - \log(2\pi v) - 2\log \pi(\theta_0)\}^{\frac{1}{2}}$$

$$\sim \{\log n - \log(2\pi v)\}^{\frac{1}{2}} \tag{2.3}$$

$$(n \to \infty)$$

The data will therefore support H_0 with respect to any prior distribution π if n is large enough and θ lies within an adjustment to $\sqrt{\log n}$ approximate standard deviations $\sqrt{(v/n)}$ of the hypothesised value θ_0. Whilst the adjustment term $\log(2\pi v)$ will often tend to a constant as $n \to \infty$ it should usually be included as it may be quite sizeable.

The condition in (2.3) provides an interesting alternative to classical significance tests when the sample size is large. For example, if $n = 1000$ we have $\sqrt{\log n} = 2.63$, whilst for $n = 10{,}000$, $\sqrt{\log n}$ increases to 3.05. Therefore, for sample sizes up to 10,000, and in special cases where $\log(2\pi v)$

is negligible, this procedure effectively fixes the significance level of the standard test (based upon the normal approximation described above) to a value depending upon n but less than the 0.005% level. For sample sizes higher than 10,000 we are really saying that none of the standard significance levels are appropriate to this situation. Indeed, standard test procedures will frequently reject H_0 at any sensible significance level in situations where our procedure suggests that the data support H_0. Therefore, as well as showing that standard test procedures do not sensibly answer question (A) according to our own criterion, we have demonstrated in a simple and direct way that they should not really be expected to adequately answer (B).

When n is small, the condition in (2.3) becomes less adequate in a formal sense. We however feel that it still provides a useful "rule of thumb" for the inductive statistician in situations where his prior π is difficult to specify. It certainly seems no worse than the standard convention which requires him to investigate, for any sample size, whether $\hat{\theta}$ lies within two standard deviations of θ_0. The criterion enables him to make a judgement which, whilst not completely precise for any particular sample size, at least varies in a sensible way for different sample sizes.

The adequacy of the approximation in (2.3) is open to some speculation on the grounds that the prior density π may be heaped around $\theta = \theta_0$, in which case the term $\log \pi(\theta_0)$ cannot be neglected. However, in section 4 we will employ an analogy with Dickey's "sharp null hypothesis testing" to show that a spike at $\theta = \theta_0$ would in fact have negligible effect upon the accuracy of our approximate criterion.

3. Likelihood Ratio Tests

We next modify the results of the previous section to cover any likelihood ratio procedure for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Expanding the log-likelihood of θ in a Taylor Series about $\theta = \hat{\theta}$, truncating after the quadratic term, and taking exponentials, yields the asymptotic approximation

$$\ell(\theta|\underset{\sim}{x}) \sim \ell(\theta|\underset{\sim}{x}) \Big|_{\theta=\hat{\theta}} \exp\{-\tfrac{1}{2} n v^{-1} (\theta - \hat{\theta})^2\} \qquad (n \to \infty) \qquad (3.1)$$

where v is defined in (2.1).

The left hand side of (2.2) represents our asymptotic approximation to the posterior density $\pi(\theta|\underset{\sim}{x})$ of θ. Hence (3.1) gives, after some elementary manipulation

$$\lambda(\theta|\underset{\sim}{x}) \sim \frac{(2\pi)^{\frac{1}{2}} v^{\frac{1}{2}}}{n^{\frac{1}{2}}} \pi(\theta|\underset{\sim}{x}) \qquad (3.2)$$

where

$$\lambda(\theta|\underset{\sim}{x}) = \ell(\theta|\underset{\sim}{x})/\ell(\theta|\underset{\sim}{x}) \Big|_{\theta=\hat{\theta}} \qquad (3.3)$$

represents the likelihood ratio. Note that (3.2) provides, as a subsidiary result, a simple demonstration of the asymptotic behaviour of the likelihood ratio in terms of the posterior density. It follows immediately that the condition in (1.1) is asymptotically equivalent to

$$-2 \log \lambda(\theta_0|\underset{\sim}{x}) < \log n - \log(2\pi v) \qquad (n \to \infty) \qquad (3.4)$$

This result may be compared with the standard likelihood ratio test which, under H_0, takes $-2 \log \lambda$ to possess a distribution which is asymptotically chi-squared with a single degree of freedom. The $\log n$ contribution is similar in spirit to a result described by Schwarz (1978) in the context of estimating the dimension of a model. Our formulation is however much broader, and the proofs simpler. Schwarz's work is also related to the approach described by Lindley (1961).

In the special case where $x_1, \ldots, x_n$ constitute a random sample from a distribution with parameter θ, it is well-known that as $n \to \infty$ the likelihood ratio test will reject H_0 with sampling probability one. For any fixed large n, the criterion in (3.4) becomes less inclined to recommend against H_0 then under a fixed-sized test. However, in the extreme limit as $n \to \infty$ it will retain a property which is similar in spirit to that of the likelihood ratio test, i.e. as $n \to \infty$ a random sample will support H_0 with sampling probability zero (this follows essentially because $n^{-1} \log n$ approaches zero in the limit and v approaches a constant value in this special case). Our criterion therefore enables us to replace the classical significance level by values which depend upon sample size in a conservative enough manner to preserve a sensible property as $n \to \infty$. This fits in with our general philosophy that all models are ultimately wrong i.e. given an arbitrarily large amount of available data, in the form of a random sample, any particular model will ultimately become inadequate.

4. Sharp Null Hypothesis Testing

Suppose now that the prior density π possesses a high concentration around the null hypothesis $\theta = \theta_0$. Such densities may be approximated by supposing that the statistician possesses a positive prior probability ϕ that $H_0 : \theta = \theta_0$ is true, and, given that $\theta \neq \theta_0$, that he possesses conditional prior density $q(\theta)$ for θ. Prior distributions of this special nature, assigning a positive prior probability to a "sharp null hypothesis" have been discussed by a number of authors, notably Dickey (1968, 1975), Lindley (1957), and Schwarz (1978). The definition in (1.1) may be adjusted to this type of formulation by saying that the data support H_0 with respect to the prior if the posterior probability, that H_0 is true, is greater than the prior probability ϕ. This posterior probability is given by

$$\text{prob}\,(H_0|\underset{\sim}{x}) = \frac{\phi B}{1 + \phi B} \tag{4.1}$$

where

$$B = \ell(\theta_0|\underset{\sim}{x}) \,/\, \int_\Theta q(\theta)\ell(\theta|\underset{\sim}{x})d\theta \tag{4.2}$$

is referred to as the "Bayes factor."

Note from (4.1) and (4.2) that the data support H_0 if and only if $B > 1$, and that this condition is equivalent to

$$q(\theta_0|\underset{\sim}{x}) > q(\theta_0) \tag{4.3}$$

where

$$q(\theta_0|\underset{\sim}{x}) = q(\theta_0)\ell(\theta_0|\underset{\sim}{x}) \,/\, \int_\Theta q(\theta)\ell(\theta|\underset{\sim}{x})d\theta \tag{4.4}$$

denotes the limit as $\theta \to \theta_0$ of the conditional posterior density of θ, given that $\theta \neq \theta_0$. Owing to the similarity between (4.3) and (1.1) the results of sections 2 and 3 may be applied directly to this conditional situation. They tell us that as $n \to \infty$ the condition in (4.3) is asymptotically equivalent to

either (2.3) or (3.4). Therefore, whatever the value of ϕ, we see that as $n \to \infty$ the data will support H_0 if (2.3), or (3.4), is satisfied. The adequacy of these approximations depends upon q but not upon ϕ. In other words, if our prior density π possesses a high concentration at $\theta = \theta_0$, then this will not affect the adequacy of our asymptotic approximations involving log n. This property noticeably increases the viability of our approximations.

Note that many of the previous results in the literature of Bayes factors may be approximated by observing whether $\hat{\theta}$ lies within the adjustment in (2.3) to $\surd \log n$ approximate standard deviations of H_0. This highlights Lindley's paradox in a very general way - note that Lindley (1957) describes particular examples where overwhelming rejection of H_0 via significance testing is complemented by high Bayes factors in favour of H_0.

The theory of sharp null hypothesis testing can also be used for answering question (B) as described in section 1. Following Dickey (1968) suppose that, when H_0 is true, the statistician incurs extra loss M_0 by rejecting rather than accepting H_0, and that, when H_1 is true, he incurs a loss M_1 by accepting rather than rejecting H_0. Then the Bayes decision would tell him to accept H_0 whenever the Bayes factor B satisfies

$$B > c \tag{4.5}$$

where

$$c = (1 - \phi)M_1/\phi M_0 \tag{4.5}$$

and to reject H_0 otherwise. The condition in (4.5) is equivalent to $q(\theta_0|\underset{\sim}{x}) > cq(\theta_0)$ where $q(\theta_0|\underset{\sim}{x})$ is defined in (4.4). The asymptotic arguments of sections 2 and 3 are again appropriate since the constant c is readily absorbed as $n \to \infty$. As a refinement, we conclude that we should accept H_0 if and only if

$$\frac{|\hat{\theta} - \theta_0|}{n^{-\frac{1}{2}} v^{\frac{1}{2}}} < \{\log n - 2 \log c - \log (2\pi v)\}^{\frac{1}{2}} \qquad (n \to \infty) \qquad (4.7)$$

We therefore recommend that, when the statistician is deciding whether to take the step of rejecting H_0, he should refer to a criterion which depends upon a constant c. This constant is different in spirit from a significance level; its specification should be based upon two costs and a prior probability. More importantly, once c is specified we see that the right hand side of (4.2) depends upon n and therefore differs from standard significance testing procedures. The latter do not therefore agree in asymptotic terms with this sensibly formulated procedure for answering question (B).

5. Multvariate Procedures

Assume now that a vector $\underset{\sim}{\theta}_q = (\theta_1, \ldots, \theta_q)^T$ of q parameters possess likelihood function $\ell(\underset{\sim}{\theta}_q | \underset{\sim}{x})$, given $\underset{\sim}{x} = (x_1, \ldots, x_n)$. Suppose that for $\ell \leq q$ the classical significance tester wishes to test the null hypothesis $H_0 : \underset{\sim}{\theta}_e = \underset{\sim}{\xi}_e$ against the alternative hypothesis $H_1 : \underset{\sim}{\theta}_e \neq \underset{\sim}{\xi}_e$ where $\underset{\sim}{\theta}_e$ denotes the subvector of the first e elements of q, and none of the nuisance parameters $\theta_{e+1}, \ldots, \theta_q$ are specified or restricted under either H_0 or H_1.

We extend the definition in (1.1) by taking the data to support H_0 with respect to a prior density $\pi(\underset{\sim}{\theta}_q)$ for $\underset{\sim}{\theta}_q$ if the posterior density $\pi(\underset{\sim}{\theta}_e | \underset{\sim}{x})$ of $\underset{\sim}{\theta}_e$ is greater than the prior density $\pi(\underset{\sim}{\theta}_e)$ when evaluated at $\underset{\sim}{\theta}_e = \underset{\sim}{\xi}_e$.

As $n \to \infty$, the posterior density $\pi(\underset{\sim}{\theta}_q | \underset{\sim}{x})$ of θ_q is asymptotically approximated by a multivariate normal density with mean vector equal to the maximum likelihood vector $\hat{\underset{\sim}{\theta}}_q$ and covariance matrix equal to the likelihood dispersion matrix $n^{-1}\underset{\sim}{V}_q$ which possesses inverse

$$n\underset{\sim}{V}_q^{-1} = \left. - \frac{\partial^2 \log \ell(\underset{\sim}{\theta}_q | \underset{\sim}{x})}{\partial(\underset{\sim}{\theta}_q \underset{\sim}{\theta}_q)^T} \right|_{\underset{\sim}{\theta}_q = \hat{\underset{\sim}{\theta}}_q} \tag{5.1}$$

Integrating out the nuisance parameters $\theta_{e+1}, \ldots, \theta_q$ we find that the posterior density $\pi(\underset{\sim}{\theta}_e | \underset{\sim}{x})$ of $\underset{\sim}{\theta}_e$ is asymptotically multivariate normal with mean vector $\hat{\underset{\sim}{\theta}}_e$ and covariance matrix $n^{-1}\underset{\sim}{V}_e$ where $\hat{\underset{\sim}{\theta}}_e$ is the maximum likelihood vector of $\underset{\sim}{\theta}_e$, and $n^{-1}\underset{\sim}{V}_e$ is the first $e \times e$ submatrix on the diagonal of $n^{-1}\underset{\sim}{V}_q$. Hence, by analogy with the method of section 2, we find that as $n \to \infty$ the data will support H_0 whenever

$$n(\underset{\sim}{\xi}_e - \hat{\underset{\sim}{\theta}}_e)^T \underset{\sim}{V}_e^{-1} (\underset{\sim}{\xi}_e - \hat{\underset{\sim}{\theta}}_e) < \ell \log n - \ell \log (2\pi |\underset{\sim}{V}_e|) \tag{5.2}$$

The statistic on the left hand side of (5.2) is seldom employed by classical testers, unless $e = q$ or $\underset{\sim}{V}_q$ is diagonal, since the matrix $\underset{\sim}{V}_e^{-1}$ does not

otherwise occur under a likelihood ratio approach. The likelihood of $\underset{\sim}{\theta}_q$ may be asymptotically approximated by

$$\ell(\underset{\sim}{\theta}_q|\underset{\sim}{x}) = \ell(\underset{\sim}{\theta}_q|\underset{\sim}{x}) \Big|_{\underset{\sim}{\theta}_q=\hat{\underset{\sim}{\theta}}_q} \exp\{-\tfrac{1}{2} n(\underset{\sim}{\theta}_q - \hat{\underset{\sim}{\theta}}_q)^T \underset{\sim}{V}_q^{-1}(\underset{\sim}{\theta}_q - \hat{\underset{\sim}{\theta}}_q)\} \qquad (n \to \infty) \tag{5.3}$$

Replacing $\underset{\sim}{\theta}_e$ in (5.3) by $\underset{\sim}{\xi}_e$ and the remaining parameters by their maximum likelihood estimates, then dividing through by the first term on the right hand side, we find that the likelihood ratio for the test defined at the beginning of this section possesses the asymptotic behaviour

$$L(\underset{\sim}{\xi}_e|\underset{\sim}{x}) \sim \exp\{-\tfrac{1}{2} n(\underset{\sim}{\xi}_e - \hat{\underset{\sim}{\theta}}_e)^T \underset{\sim}{W}_e^{-1}(\underset{\sim}{\xi}_e - \hat{\underset{\sim}{\theta}}_e) \qquad (n \to \infty) \tag{5.4}$$

where $n\underset{\sim}{W}_e^{-1}$ represents the first $e \times e$ submatrix on the diagonal of the information matrix $n\underset{\sim}{V}_q^{-1}$. By similar arguments to those described in section 2 it is straightforward to relate the asymptotic behaviour of the likelihood ratio to the posterior density of $\underset{\sim}{\theta}_e$ by

$$L(\underset{\sim}{\xi}_e|\underset{\sim}{x}) \sim \frac{(2\pi)^{\frac{1}{2}e} |\underset{\sim}{V}_e|^{\frac{1}{2}}}{n^{\frac{1}{2}e}} \exp\{-\tfrac{1}{2} nA_e\} \pi(\underset{\sim}{\xi}_e|\underset{\sim}{x}) \qquad (n \to \infty) \tag{5.5}$$

where

$$A_e = (\underset{\sim}{\xi}_e - \hat{\underset{\sim}{\theta}}_e)(\underset{\sim}{W}_e^{-1} - \underset{\sim}{V}_e^{-1})(\underset{\sim}{\xi}_e - \hat{\underset{\sim}{\theta}}_e) \tag{5.6}$$

Note that the expression for A_e in (5.6) will always be non-negative since the matrix $\underset{\sim}{W}_e^{-1} - \underset{\sim}{V}_e^{-1}$ is positive semi-definite. This follows from the representation

$$\underset{\sim}{W}_e^{-1} - \underset{\sim}{V}_e^{-1} = \underset{\sim}{H}\underset{\sim}{G}^{-1}\underset{\sim}{H}^T$$

based upon the partition

$$\underset{\sim}{V}_q^{-1} = \left(\begin{array}{c|c} \underset{\sim}{W}_e^{-1} & \underset{\sim}{H} \\ \hline \underset{\sim}{H} & \underset{\sim}{G} \end{array}\right)$$

of the matrix satisfying (5.1).

It follows immediately from (5.5) that, as $n \to \infty$, the data will support H_0 whenever

$$- 2 \log L(\underset{\sim}{\xi}_e | \underset{\sim}{x}) < \ell \log n - \ell \log (2\pi |\underset{\sim}{V}_e|) + nA_e \qquad (5.7)$$

Under H_0, the quantity on the left hand side of (5.1) possesses a sampling distribution which is asymptotically chi-squared with e degrees of freedom. We however see from the right hand side of (5.7) that there is an increased problem in working with the likelihood ratio rather than the quadratic term in (5.2). This is because of the addition of the extra term nA_e whenever nuisance parameters are present. In many examples A_e will tend to a constant positive value as $n \to \infty$. Therefore for large n the contribution nA_e could dominate even the term $\ell \log n$. We conclude that fixed-size likelihood ratio tests with nuisance parameters present may be even more overready to reject the null hypothesis than the tests discussed in section 2 for single parameter situations.

6. An alternative to the chi-squared goodness of fit test

Suppose that all frequencies $x_1, \ldots, x_s$ possess a multinomial distribution with corresponding all probabilities $\theta_1, \ldots, \theta_s$, summing to unity, and sample size n, and suppose also that we wish to compare $H_0 : \theta_1 = \xi_1, \ldots, \theta_s = \xi_s$ with an unrestricted alternative hypothesis. Then a straightforward application of the result in (5.2) for the $s - 1$ distinct parameters $\theta_1, \ldots, \theta_{s-1}$ tells us that as $n \to \infty$ and the $P_j = x_j/n$ remain fixed and positive the data will support H_0 whenever

$$n \sum_{j=1}^{s} P_j^{-1}(P_j - \xi_j)^2 < (s - 1) \log n + (s - 1) \log \{2\pi \sum_{j,k:j\neq k} P_j^{-1} P_k^{-1}\} \qquad (6.1)$$

Note firstly that the standard chi-squared statistic should be replaced by its modification on the left hand side of (6.1) and that for large enough n the data will support H_0 whenever the modified statistic is less than $\log n$ times the usual degrees of freedom. It might be interesting to compare this with a limiting result by Leonard (1977) which suggests that in preliminary testing situations the critical value should be twice the degrees of freedom, though the purpose of the analysis is then very different. Other aspects of significance testing are discussed by Leonard and Ord (1976) and Leonard (1978).

General Conclusions

The phenomena discussed here would seem to be particularly relevant to the area of scientific reporting. For example, many results in applied journals which might have been viewed as "significant," because they yield a low p-value, may indeed yield evidence in support of the null hypothesis and in fact serve to detract from the very scientific theory which they claim to substantiate. Efforts should perhaps be made to reduce the role of significance tests in application areas of statistics (e.g. psychology, medicine, and sociology). It seems evident that low p-values should no longer be viewed as a prerequisite for the acceptance of a scientific theory.

All too often, significance tests are employed by applied workers either in an attempt to formally confirm results which one already thought to be true intuitively speaking, or to assist in a search for a model which provides a good fit to the data, or to help extract a few "significant" conclusions from a large and maybe noisy data set. We feel that much more emphasis should be placed in applied circles on developing a model which is meaningful in the real-life context of the practical problem at hand. The techniques involving $\sqrt{\log n}$ described in this paper may be useful in checking that the data gives credibility to the model. When n is large, they permit a larger range of viable null hypotheses than experienced under fixed size significance testing. It should therefore be easier to find a data-credible model which is also reasonable in real-life terms.

Acknowledgment

The author wishes to thank Professor J. M. Dickey for making available some of his preliminary results on the topic, and for assisting with the analysis in section 5.

REFERENCES

[1] DeGroot, M. (1970) Optimal Statistical Decisions, McGraw-Hill, New York.

[2] Dickey, J. M. (1968) A Bayesian Hypothesis Decision Procedure, Ann. Math. Stat. 19 pp. 367-369.

[3] Dickey, J. M. (1975) Bayesian alternatives to the F-test and least squares estimates in the normal linear model, in Studies in Bayesian Econometrics and Statistics (ed. by A. Zellner and S. E. Feinberg) pp. 515-554, Amsterdam, North Holland.

[4] Lindley, D. V. (1957) A Statistical Paradox, Biometrika 44 pp. 187-192.

[5] Lindley, D. V. (1961) The use of prior distributions in statistical inference and decision, Proc. Fourth Berkeley Sympos. 1 pp. 453-468.

[6] Leonard, T. and Ord, J. K. (1976) An investigation of the F-test procedure as an estimation short-cut, J. Roy. Statist. Soc. B. 38 pp. 95-98.

[7] Leonard, T. (1977) A Bayesian approach to some multinomial estimation and pre-testing problems, J. Amer. Statist. Assoc. 72 pp. 869-876.

[8] Leonard, T. (1978) Density estimation, stochastic processes, and prior intermation (with discussion), J. Ray. Statist. Soc. B. 40 pp 113-146.

[9] Schwarz, G. (1978) Estimating the dimension of a model, Annals of Statistics 6 pp. 461-464.

TL/ck

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2004	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: Technical
4. TITLE (and Subtitle): WHY DO WE NEED SIGNIFICANCE LEVELS?	5. TYPE OF REPORT & PERIOD COVERED: Summary Report, no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Tom Leonard	8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Probability, Statistics, and Combinatorics
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709	12. REPORT DATE: October 1979
	13. NUMBER OF PAGES: 16
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

MRC-TSR-2004

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Significance, Bayes, Asymptotic normality, Likelihood ratio, Bayes factor
Chi-squared

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Classical significance tests depend upon a choice of significance level and also seem overready to reject the null hypothesis when the sample size is large. A plausible alternative to significance testing has been suggested by Schwarz (1978) in the context of model discrimination. A simpler and more general formulation is discussed here; this leads to more precise approximations. For a single parameter and when the sample size n is large we recommend viewing the data as supporting a simple null hypothesis versus a completely composite alternative whenever the maximum likelihood estimate lies

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

221200

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (Continued)

within an adjustment to $\sqrt{\log n}$ approximate standard deviations of the null hypothesis. This criterion indeed appears to provide an attractive rule of thumb for all sample sizes: It removes the need for tables of significance levels and becomes less keen to reject the null hypothesis for large sample sizes. The ideas are extended to provide alternatives to multivariate likelihood ratio tests, and to the chi-squared goodness of fit test.