AN ALTERNATIVE TO SIGNIFICANCE TESTS

by

Victor Solo
Purdue University and Harvard University

Technical Report #84-14

Department of Statistics
Purdue University

1984

An alternative to significance tests

by Victor Solo

ABSTRACT

There are a number of problems with pure significance tests even within a classical view. Firstly they give no measure of magnitude of difference. Secondly and related is that for large enough sample size any null hypothesis will be rejected (since rounding errors in stating the null hypothesis will be detected): this yields the confusion of practical significance versus statistical significance. On the other hand the great advantage of the significance test is that it summarizes the data in one number. It is here suggested that this last property can also be supplied by a confidence interval on a noncentrality parameter. The advantage is that the above two problems are resolved: furthermore the user is forced to think hard at an early stage about what is a practical difference. It becomes necessary however to find a scale on which the size of the noncentrality parameter can be interpreted. This is not so simple. Various examples are discussed.

Key words: Significance, hypothesis test, noncentrality parameter, statistical inference

1. INTRODUCTION

Though all of the triad of popular classical methods of drawing conclusions from data - the pure significance test (PST), the point estimate, the confidence interval (CI) - have come under criticism from many angles (see e.g. the issue of Synthese in honour of Birnbaum (1977)) the PST has problems even within the classical setting.

Perhaps the major reason for the popularity of the PST is its ability to summarize the data the evidence for an hypothesis in a single number, a P-value. (The evidential procedure PST is to be distinguished from its behavioral or decision theory counterpart the hypothesis test see e.g. Kempthorne (1976) who draws a careful distinction and Cox and Hinkley (1974) and Birnbaum (1977)).

The classical criticism of the PST concerns two points. Firstly, if the sample size (n) is large enough any hypothesis will be found wanting. This is simply because rounding errors will be detected or tiny modelling errors will be found. On the other hand for small n acceptance may allow large practical differences. This is the practical significance versus statistical significance confusion. Secondly the PST gives no measure of the magnitude of the discrepancy from the null hypothesis. This contributes to the above confusion. Actually this point is not quite true as will be seen later. Still, as presently used the PST gives only a qualitative measure of discrepancy.

In this article an alternative to the PST is suggested. That is a CI on a relevant concentrality parameter (NCP). This idea retains the attractiveness of data summary in a single number (well an interval) while resolving the weaknesses above. The NCP does measure discrepancy and provided a scale to

refer to can be found, the user of a CI on a NCP is forced to specify practical differences before drawing a conclusion. (if not before collecting the data!) Before continuing, some words of caution and clarification are in order.

In what follows, various uses of PST's (and so CI's on NCP's) are discussed: viz preliminary inference; diagnostic tests. By no means is it suggested that CI's on NCP's replace more detailed inferential activity. Merely it is suggested that whenever, at some intermediate stage of inference, a PST would be useful then the CI on a NCP will be more so.

Next, the ensuing discussion is developed in a classical setting. However the point being made here applies equally well to a Bayesian approach. That is, calculation of the posterior distribution of a NCP (or a Bayesian credible region) is being advocated. A detailed development will need however the provision of appropriate prior distributions for NCP's. The author is not aware of much work here (see Gelfand, 1983) though Arnold Zellner has indicated that some of the calculations should be straightforward enough.

Finally by using a CI we run directly into the contentious problem of how to choose the size α. For the present discussion, traditional values of .01, .05 will be used. Another interesting choice is the 100% CI's of Robbins (1971). These entail e.g. in the Gaussian case replacing say $1.96/\sqrt{n}$ by $(2 \ln \ln n)^{1/2} (1 + \epsilon_n)/\sqrt{n}$ where ϵ_n decreases with n, starting at say .5 to agree with 1.96 at low values of n.

In the next Section uses of PST's and CI's on NCP's are reviewed. In Section 3 some discussion is given of simple t and F examples. The χ^2 and related issues is discussed in Section 4. Section 5 is concerned with C_p. In Section 6 Score tests or Lagrange multiplier tests are investigated. Section 7 looks at a problem of testing for canonical correlations. In Section 8 some connexions with exact slope are discussed. Conclusions are offered in Section 9.

II. Uses of PST's and NCP's

Cox(1977) discusses a number of uses of PST's. He observes that there are various types of null hypothesis. The 'plausible' hypothesis is one that will help interpretation a lot e.g. no interaction in a 2-way table or parallel regression lines in an analysis of covariance.

Another type of hypothesis is the 'dividing hypothesis' which serves to mark reference points e.g. all means are equal in an analysis of variance. Finally there is the scientific hypothesis, or the basic question:e.g. which drug is best?

The first two types of hypotheses particularly can be usefully investigated with PST's. Another major area of use for PST's is in model criticism or diagnostic checking e.g. normal probability plots or goodness of fit tests.

Examples of all these types will be discussed and it will be shown how a CI on a NCP may be used in each case.

We will meet two types of situations. In the first case the NCP will have a direct physical interpretation e.g. as a root mean square treatment effect measured in standard deviation units. Then there is little trouble in using the NCP. The other case is harder. Here there will be no immediate interpretation and some hard work will be needed to find one. This will usually be true of diagnostic testing. Here it will be seen there are many unanswered questions.

III. The Noncentral t and F

(a) The t

Suppose we have a random sample $X_1,\ldots X_n \sim N(\mu,\sigma^2)$ and wish to test $H_0 : \mu = \mu_0$. The standard test statistic is $t = (\bar{X} - \mu_0)\sqrt{n}/s \sim t_{n-1}$ where as usual s is the sample standard deviation.

Now in the case of a scalar parameter we can easily replace the scalar PST by a scalar CI on $\Delta = \mu - \mu_0$ namely $(\tilde{\Delta} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n})$ and so avoid the two weaknesses of the PST. Still it seems useful to start off slowly by showing how the NCP can be used in this case.

When H_0 is untrue t has a noncentral distribution (NCD) $t \sim t_{n-1}(\phi_n)$ where ϕ_n is the NCP and $\phi_n = \sqrt{n}\,\delta$ where $\delta = \Delta/\sigma = (\mu - \mu_0)/\sigma$ will be called the standardized NCP and is really the type of NCP that has been referred to in the earlier discussion.

Here we see that δ has a direct interpretation: it measures the discrepancy between the hypothesized mean μ_0 and the actual value μ on the σ-scale. The user of the CI finds it necessary to think about what values of δ are of practical consequence. A (biased) estimate of δ is clearly $\hat{\delta} = (\bar{X} - \mu_0)/s = t/\sqrt{n}$

Now since constructing a CI on a NCP is perhaps unusual, let us simply recall that to construct a CI for a parameter θ, all we need in principle is a statistic T(X) whose distribution depends only on θ. We can then, in the manner illustrated so nicely by Kendall and Stuart (1979, Sections 20.1, 20.9) construct horizontally in the $(\theta,T(X))$ plane a confidence region. We use the region vertically to get a CI for θ. (It is a great pity that the eloquent picture (Fig. 20.2) of the binomial CI given by Kendall and Stuart is not to be found in any other textbook known to the author.)

A simple example will make things entirely clear. Suppose $n = 17$, $\bar{X} = 17$, $s = 3$, $\mu_0 = 15 \Rightarrow \hat{\delta} = (\bar{X} - \mu_0)/s = .66 \Rightarrow t = \delta/\sqrt{n} = 2.749$. The Biometrika tables (Table 24) give CI's for ϕ_n; we find a 95% CI as $(\phi_-, \phi_+) = (.539, 4.891) \Rightarrow (\delta_-, \delta_+) = (.131, 1.186)$. Here $\phi_{+,-}$ have the form $\phi_{+,-} = t \pm \varepsilon_{+,-}$. It is not too surprising then that the "pseudo-interval" $(\hat{\Delta} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n})/s = (.45, 3.55)/3 = (.15, 1.18)$ gives a similar result. As a footnote we might prefer, with Mosteller and Tukey (1977 p) to use $(\hat{\Delta} \pm Z_{1-\alpha/2}\, s/\sqrt{n-2})/s \Rightarrow (\underset{\sim}{\delta}, \hat{\delta}_+) = (.16, 1.14)$

In any case we have in the 95% CI (.131, 1.186) a concrete statement of magnitude that forces the user to consider its practical meaning.

(b) The F

The generic scheme here is, of course, the linear model

$$\underline{Y} = \underline{\mu} + \underline{\varepsilon} \quad .$$

where all vectors are n-vectors, $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{I})$. Consider

$$H_0 = \underline{\mu} \in W_0 = \text{span}\,(\tilde{\underline{\xi}}_1 \ldots \tilde{\underline{\xi}}_k)$$

while in actuality $\underline{\mu} \in W = \text{span}\,(\underline{\xi}_1 \ldots \underline{\xi}_n)$ where the $\underline{\xi}$'s and $\tilde{\xi}$'s are n-vectors.

The F-statistic for testing H_0 is

$$F = \frac{n-m}{n-k} \frac{||\hat{\underline{\mu}} - \hat{\underline{\mu}}_0||^2}{||\underline{Y} - \hat{\underline{\mu}}||^2} = \frac{||\hat{\underline{\mu}} - \hat{\underline{\mu}}_0||^2}{\hat{\sigma}_m^2} / (m-k)$$

where $\hat{\sigma}_m^2 = ||Y - \hat{\underline{\mu}}||^2/(n-k)$ and $\hat{\underline{\mu}}$ = projection of $\underline{Y}$ onto W_0 etc.

Now we do have a genuine case of a vector parameter while our interest is in a preliminary scalar summary.

When H_0 is untrue F has a noncentral distribution

$$F \sim F_{n-k,n-m}(\psi)$$

$$\psi = ||\underline{\mu} - \underline{\mu}_0||^2/\sigma^2$$

we introduce now a standardized NCP

$$\delta = \sqrt{\psi/(n-k)}$$

clearly δ is a root mean square deviation between $\underline{\mu}$ and $\underline{\mu}_0$ on the σ scale, but this interpretation is not clear enough to help too much in general.

In the analysis of variance (AOV) setting we do have a clear meaning though: δ is a root mean square treatment effect (per treatment) measured on the σ-scale. This is something the user can interpret. It is important to understand here that the particular alternative $\underline{\mu}$ is not required to be specified rather δ is asked for directly. Of course δ does depend on the type of alternative but it is suggested that a direct "feeling" for δ can be aimed at. If necessary this can be done by contemplating different types of alternative and seeing what δ values result. Some further comments are offered later. An example is given below.

In the Regression case we can say δ^2 is an average signal to noise ratio per parameter. Here the meaning is less clear and some experience is needed to develop a "feel" for "how large is large" - this is the scale problem referred to earlier. Some rough (order of magnitude) arguments are now offered.

Consider the scalar regression

$$Y_k = \beta X_k + \epsilon_k \qquad k = 1,2,\ldots n$$

and suppose $\Sigma_1^n X_k^2/n \longrightarrow \sigma_X^2$ while ϵ_k are i.i.d. zero mean variance σ^2. Then the SNR is $\delta^2 = \beta^2 \Sigma_1^n X_k^2/\sigma^2$. So the average SNR per observation is

$$\lambda = \frac{\delta^2}{n} = \beta^2 \Sigma_1^n X_k^2/n\sigma^2 \longrightarrow \beta^2\sigma_X^2/\sigma^2$$

On the other hand, consider the R^2 which is

$$R^2 = \hat{\beta}^2 \Sigma_1^n X_k^2/\Sigma_1^n Y^2 \longrightarrow \rho^2 = \lambda/(1+\lambda) \qquad \text{a.s.}$$

by the strong law of large numbers. It is sometimes suggested (particularly in

Social Science application) that the smallest R^2 of interest is $\sim$ 1/3. This corresponds to $\lambda = \frac{1}{2}$. We have other pairs (ρ^2, λ) of

$$(.5, 1), (.75, 3), (.95, 19)$$

Since the NCP is a monotonic function of ρ^2 the reader may prefer to use ρ^2. Clearly there is some latitude here. In a multiple regression we use partial R^2 and corresponding partial λ. This is further discussed in Section 5.

We now offer a simple AOV example. The natural estimate of ϕ is $\hat{\phi} = \sqrt{F}$ however it is biased since

$$E(F(\psi)) = \frac{n-m}{n-m-2} \left(\frac{\psi}{m-k} + 1\right)$$

Thus $\sqrt{F-1}$ would seem to be better. Of course at this stage we could look for Stein-type estimates (see e.g. Gelfand (1983) and references) but this is avoided here to keep things simple.

Unfortunately the author is unaware of tables of the percentage points of the noncentral-F that could be used to give CI's. Instead then, the approximation technique of Patnaik (1949) will be used. We consider simple one-way AOV given in John (1971, p. 47).

The data shown in Table 1 are coded values of octane rating of 5 petrols with 4 cars for each petrol.

TABLE 1 CODED OCTANE NUMBERS John (1971), p. 47)

PETROL	OCTANE NUMBER				median	mean
A	1.7	1.2	.9	.6	1.05	1.1
B	1.7	1.9	.9	.9	1.3	1.35
C	2.4	1.2	1.6	1.0	1.4	1.55
D	1.8	2.2	2.0	1.4	1.9	1.85
E	3.1	2.9	2.4	2.4	2.65	2.70

Here $n = n_0 t$; $n_0 = 4$, $t = 5$; $\delta = \sqrt{\Sigma_1^t \Delta_i^2/(t-1)\sigma^2}$; $\Delta_i = \mu_i - \bar{\mu}$ and H_0: $\mu_i = \bar{\mu}$ $1 \leq i \leq t$; $\bar{\mu} = \Sigma_1^t \Delta_i/t$. The column of means shows a trend (hidden variable?) suggesting $\mu_i = a + ib \Rightarrow \Delta_i = (i - \frac{t+1}{2})\, b \Rightarrow \delta = \frac{t(t+1)}{12} b = 1.55b$.

Thus in this case δ can be interpreted as a slope. This little calculation shows the type of varied interpretation δ will admit.

In Table 2 is given a modified AOV table. The author prefers to use percent sum of squares rather than listing meaningless large sums of squares. Also shown are a standard deviation (SD) for treatment differences and a simple point estimate of the NCP.

TABLE 2 Modified AOV Table

SOURCE	% SS	df	MS	SD	NCP = $\sqrt{F}$
Treatment	64	4	1.53		2.6
Error	36	15	.225	.15	
Total	100	19			

The standard preliminary approach is to reason $F = 6.78 \Rightarrow p < .005$, so evidence is against H_0. The present approach is to quote instead the 95% CI on δ namely (.94, 3.6) (see Appendix A for the calculation)

Once again the user is forced to consider what magnitude of δ is of consequence here and even whether there is enough data to pin δ down sufficiently.

To be clear, this CI is not supposed to give a final inferential analysis of these data, but merely replace the first step PST with something more concrete. Finally the above idea can be extended of course to more complicated AOV designs.

IV. Non-central χ^2

Suppose $N_1 \ldots N_r \sim$ multinomial $(N,\underline{p})$: $N = \Sigma_1^r N_i$ and consider testing $H_0 = p_i = p_{i0}$ $i = 1 \ldots r$. The standard χ^2 statistic is

$$\chi^2 = N\Sigma_1^r(\hat{p}_i - p_{i0})^2/p_{i0} : \hat{p}_i = N_i/N$$

When H_0 is untrue there are two types of noncentral distribution available, both asymptotic

(i) Local Alternative. $p_i - p_{i0} = O(N^{-1/2}) = c_i/N^{1/2}$

$$\chi^2 \Rightarrow \chi^2_{r-1}(\phi) \quad \text{(to order } N^{-1})$$

$$\phi = N\Sigma_1^r(p_i - p_{i0})^2/p_{i0} = \Sigma_1^r c_i^2/p_{i0}$$

where $\chi_r^2(\phi)$ is a noncentral χ^2 with NCP ϕ. This result is due to Cochran (1952) and Patnaik (1949).

(ii) General Alternative (Broffit and Randles (1977))

$$\sqrt{N}(\hat{\delta}^2 - \delta^2) \Rightarrow N(0, t^2) \quad \text{(to order } N^{-1/2})$$

$$t^2 = 4\Sigma_1^r p_i\left(\frac{p_i}{p_{i0}} - c\right)^2 \quad : \quad c = \Sigma_1^r p_i \frac{p_i}{p_{i0}}$$

and $$\delta = \sqrt{\Sigma_1^r\left(\frac{p_i - p_{i0}}{p_{i0}}\right)^2 p_{i0}}$$

Notes. 1. δ is a root mean square weighted percent deviation of actual $\underline{p}$ from hypothesized $\underline{p}_0$. Also $\phi = N\delta^2$ and $\hat{\delta}^2 = \chi^2/N$ with $E(\hat{\delta}^2) \longrightarrow 1 + \delta^2$ so $\hat{\delta}$ is biased.

2. The order of accuracy of the limit results is not very good, so higher order terms would be useful if not mandatory for constructing CI's.

We now consider the use of these NCD's in contingency tables and goodness of fit tests.

(a) Contingency Tables

It is immediately clear that our percent interpretation of δ though concrete will not work with general contingency tables. This is because the equivalent meaning of δ as a measure of association is flawed (see Fleiss, 1981).

The author has not yet found a way around this, but two ideas suggest themselves. The first would be to define a generalized odds ratio. Bearing in mind that CI's and NCP's are never a replacement for full statistical inference the alternative is of course fitting a fully developed loglinear model. The other idea is to use one of a number of distance measures such as

$$\text{metric } \chi = \delta_M = \left[\Sigma_1^r \frac{(p_i - p_{i0})^2}{p_i^*} \right]^{1/2} : p_i^* = (p_i + p_{i0})/2$$

This is a metric on the space of probability vectors $\underline{p}$ and so does measure a discrepancy. The problem is of course the development of an intuitive feeling for the size of δ_n (for local alternatives $\delta_M \simeq \delta$) as well as relating it to the odds ratio. For these reasons we seem presently restricted to one-way tables.

This last idea is worth emphasizing. In this article NCP's have been used as measures of discrepancy between null and alternative. This is because they arise naturally in test construction and so are equated naturally with specific PST's. In general however there is nothing special about using the likelihood ratio procedure to generate such discrepancy measures. Many other measures, suited perhaps to interpretation (as above) or the type of alternative envisioned may be contemplated. This is an area where much work could be done.

A simple example is now given.

The data are taken from Plackett (1974), Table 2.1). Table 3 shows the data, proportions and percent deviations from $p_{i0} = \frac{1}{12}$

TABLE 3 Monthly cases of lymphatic leukemia, England 1946-1960

J	F	M	A	M	J	J	A	S	O	N	D	
40	34	30	44	39	58	51	55	36	48	33	38	N_i
.079	.067	.059	.087	.077	.114	.101	.109	.071	.095	.065	.075	$\hat{p}_i$
-5	-19.6	-29.2	4.4	-7.6	36.8	21.2	30.8	-14.8	14	-22	-10	$100(\hat{p}_i - p_{i0})/p_{i0}$

The null hypothesis is no monthly variation $H_0 : p_i = \frac{1}{12}$.

With $r = 12 \Rightarrow df = r-1 = 11$; $N = 506$; $X^2 = 21.3 \Rightarrow \hat{\delta} = \sqrt{X^2/N} = .205$; $\hat{\tau} = .748$.

Now to find a CI for δ we must use the percentage points of the noncentral distribution (NCD) - which one? We use both. Recall that the lower value of the CI is set by the upper percentage point of the NCD and Proffit and Randles observe that for high power the ND is best. We use the ND for the lower CI value and the NC-χ^2 for the upper CI value.

(a) The lower value is $\delta - 1.96\,\hat{\tau}/\sqrt{N} = .14$

(b) The upper value comes from Biometrika table 24 which gives percentage points of $NC\chi^2$ in terms of ϕ we find $(\chi, df) = (4.62, 11) \Rightarrow \sqrt{\phi_+} = 5.62$ $\Rightarrow \delta_+ = .250$.

Thus the 95% CI is $(\delta_-, \delta_+) = (.14, .25)$ This shows a concrete percent deviation of about 20%.

(b) Goodness of Fit Test

Suppose we have $X_1 \ldots X_n \sim g(X)$ and desire to test H_0: $X \sim f(X|\theta)$.
If we group the X values we can do this with χ^2 so that we have H_0: $p_i = p_i(\theta)$ $i = 1 \ldots r$ where

$p_i(\theta) = \int_{c_i}^{c_{i+1}} f(x|\theta)\,dx$ = cell probability and (c_i, c_{i+1}) are group boundaries.

Purely for convenience (see later) we choose here to use Y^2 to measure goodness of fit.

$Y^2 = 2N\Sigma_1^r \hat{p}_i \log(\hat{p}_i/p_i(\hat{\theta}))$; $\hat{p}_i = N_i/N$ and N_i = number of X-values in group i. While $\hat{\theta}$ is the solution to the minimum Y^2 problem i.e.

$$\Sigma_1^r \hat{p}_i \, d \log p_i/d\underline{\hat{\theta}} = 0$$

Again we find two NCD's.

(a) Local Alternative. (this is a conjecture) $p_i - p_i(\theta_0) = c_i/\sqrt{N}$

$$Y^2 \Rightarrow \chi^2_{r-m-1}(\phi) \text{ a } NC\chi^2$$

where m = dim($\underline{\theta}$) and

$$\phi = 2N\Sigma_1^r \pi_i \log(\pi_i/p_i(\theta_0))$$

and π_i are the true cell probabilities

(b) General Alternative (Moore, 1983)

$$\sqrt{N}\,(\hat{\delta}_y^2 - \delta_y^2) \Rightarrow N(0,\tau^2)$$

$$\delta_y^2 = \Sigma_1^r \pi_i \log(\pi_i/p_i(\theta_0))$$

$$\hat{\delta}_y^2 = Y^2/N \text{ and}$$

$$t^2 = 16\{\Sigma_1^r \pi_i (\log(\pi_i/p_i(\underline{\theta}_0)))^2 - (\Sigma_1^r \pi_i \log(\pi_i/p_i(\underline{\theta}_0)))^2\}$$

$$\underline{\theta}_0 \ni \Sigma_1^r (\pi_i/p_i(\underline{\theta}_0)) \, dp_i/d\underline{\theta}_0 = 0$$

Now for local alternatives $Y^2 \underline{\sim} \chi^2$ so we can still think of δ_y as a percent. But now we have a question of scaling. What is a large value of δ_y, or a small value? This is rather more subtle than the earlier examples of direct interpretation and needs some further consideration.

If the grouping is very fine we see that $\delta_y^2 \underline{\sim} 2\int g(x) \ln(g(x)/f(x|\theta_0|)) \, dx$ The idea is to evaluate this expression for various types of g and f.

As an example consider (for fixed λ for the moment) a goodness of fit for a Gamma transformed to approximate Gaussianity i.e. we suppose $Z \sim \Gamma(p,\lambda)$ and consider a fit of $X = Z_\lambda = (Z^\lambda - 1)/\lambda$ to $N(\mu,\sigma^2)$

First we calculate $\underline{\theta}_0 = (\mu_0,\sigma_0^2) \ni$

$$\int g(x) \frac{d \log f_\theta(X)}{d\theta_0} \, dx = 0$$

Hernandez and Johnson (1980) show that μ_0, σ_0^2 =true(μ,σ^2). They then calculate that

$$\frac{1}{2}\delta_y^2(\lambda) = \frac{1}{2}\log(2\pi+1) - 2\log\Gamma(p) + p(\psi(p)-1) - \lambda\psi(p)$$
$$+ \frac{1}{2}\log\{[\Gamma(p)\Gamma(2\lambda+p) - \Gamma^2(\lambda+p)]/\lambda^2\}$$

and note that $\delta_y(\lambda)$ becomes flatter near its minimum as p increases.

It is now suggested that an appropriate set of values of δ_y to scale to are those such that $\delta_y^2(\lambda) = \min$. These have been calculated by Hernandez and Johnson (1980) and are reproduced below ($\delta_{y\,\min}$ has been expressed as a percent for present use). The results do not depend on β.

TABLE 4 SCALING FOR δ

p	λ_{opt}	δ_{ymin}
3	.312	1.98%
2	.301	3.11%
1	.265	7.49%
.5	.208	15.56%

Now in the Gamma, p controls the peakiness of the distribution (while β is a scale factor). So clearly fit problems emerge for p too small. From the graphs in Hernandez and Johnson (1980) we only see major discrepancies in the transformed and fitted distributions for $p < 1$ i.e. $\delta_{ymin} > 7.5\%$. This conclusion is about correct also for other 'peaky" distributions looked at by

these authors.

A scale is thus established. If δ_y is beyond about $7\frac{1}{2}\%$ then we do not have a good fit with respect to peakiness.

Clearly for other types of deviation we may calculate other ranges of δ values. There is of course much work to be done here. Also it seems reasonable to choose the goodness of fit statistic to match the deviation of interest. It may be remarked that the NCD of the Kolmogorov Smirnoff test is not very tractable (Raghavachari, 1973) which is one reason the preceding argument was developed through χ^2.

The point being made here is that while NCP's are extremely useful, the scaling problem means that some very detailed analyses are needed showing how particular alternatives give rise to particular NCP values. In this way we can develop a feeling for order of magnitude of NCP's.

V - Model order estimation

We consider briefly the problem of variable selection in regression. A number of criteria have become popular in the last decade. For simplicity only Mallows' C_p is considered, (Mallows, 1973). It is suggested here that a CI can be constructed for MSE_p to aid the consideration of models nearby the minimum C_p model. The CI is constructed for that $\hat{p}$ which minimises C_p.

Consider then a linear regression $\underline{Y}=\underline{\mu} + \underline{\varepsilon}$, $H_o := \underline{\mu} \in$ span $(\xi_1 \ldots \xi_k) = W_k$ When H_o is false we have $RSS_k / \sigma^2 \sim \chi_{n-k} (\phi_k)$ where $RSS_k = ||Y - \hat{\mu}_k||^2$ and $\phi_k = ||\underline{\mu}_k - \mu||^2/\sigma^2 = B_k/\sigma^2$ while $\underline{\mu}_k = \underline{P}_k \underline{\mu}$ and $\underline{P}_k$ projects onto W_k. The NCP can be interpreted through $M_k = MSE_p^{(k)} / \sigma^2 = k+B_k /\sigma^2 = k + \phi_k$ where $MSE_p^{(k)}$ = mean squared error of prediction.

Thus a CI on ϕ_k from the NCX^2 yields a CI on M_k. An unbiassed estimator of M_k is $C_k = RSS_k/ \sigma^2 + 2k - n$. Mallows idea is to plot C_k versus k to suggest candidate models. These would include those with C_k near the minimum value. Also points on the C_k versus k plot not near the 45° line suggest evidence of bias. The advantage of a CI on the NCP or M_k here is that the magnitude of biases is accounted for. It seems reasonable then to calculate a CI based on the C_k value which is minimum.

First note that usually an estimate of σ^2 is provided by an overfit on say m regressors. Then C_k is biassed, however an unbiassed estimate $\tilde{C}_k$ can be based on the F distribution

$$\frac{(RSS_k - RSS_m)/(m-k)}{RSS_m / (n-m)} \sim F_{n-k,n-m} (\phi_k)$$

$$\Rightarrow \quad E(F) = \frac{n-m}{n-m-2} \left(1 + \frac{\phi_k}{m-k}\right)$$

$$\Rightarrow \quad \tilde{C}_k = \left(\frac{n-m-2}{n-m}\right) F (m-k) + 2k-m$$

We can use Patnaik's (1949) approximation to give a CI for ϕ_k. We then investigate those other models whose C_k values fall inside the CI. The behaviour of such a technique is clearly complicated and deserves a full investigation. An example is briefly sketched.

We take the cement data of Hald (1960) analysed by Draper and Smith (1981). The y variable is heat liberated in a cement setting experiment; the x's are concentrations of various chemicals. Also n=13 and an intercept (variable ϕ) and 4 other variables are used. A C_p plot is given in Figure 6.3 of Draper and Smith. The minimum C_k is at (0,1.2) with $C_3 = 2.7 \Rightarrow RSS_3/\hat{\sigma}_8^2 = C_3+13-2x3 = 9.7$. RSS_3 has 10 df and $\chi = \sqrt{9.7} = 3.2$. Biometrika Table 24 $\Rightarrow$ a 95% CI $(MSE_P^{(-)}, MSE_P^{(+)}) = (0,17)$. Thus the models plotted in the Figure may all be entertained. The use of Patnaik's approximation gives $(MSE_P^{(-)}, MSE_P^{(+)}) = (0,11)$. The conclusion is the same.

From Draper and Smith (1981).

VI. DIAGNOSTIC TESTS

Recently it has become clear that Score tests (Rao, 1973) or Lagrange multiplier tests (LM) tests (Silvey, 1959) provide a general tool for deriving diagnostic or model criticism tests. This has been made explicit in the econometric literature (Breusch and Pagan, 1980) and a growing realization has come in the statistics literature (see e.g. Pregibon (1981), Atkinson (1982)).

The basic idea is as follows. Set down an hypothesis for a particular model departure and derive the LM test (its great advantage is that it only requires model fitting under the null). Often the test statistic (see below) is a partial R^2 in a regression of residuals from the model(under the null)on a "constructed variable". Thus with the test there goes a natural plot of these residuals against the constructed variable.

It further seems that most (if not all!) popular ad-hoc diagnostic tests are in fact in LM tests. For examples see the above references. An interesting additional example is Tukey's 1 degree of freedom test for nonadditivity in a two way table and its generalisations (see Milliken and Graybill (1970) and Scheffe (1959, p. 144)). A brief review of the LM methodology is now given.

Suppose $X_1,,,X_n \sim$ iid $f(X|\underline{\theta})$ and we wish to test H_0: $\underline{h}(\underline{\theta}) = \underline{0}$; $\underline{\theta}$ is an r-vector, $\underline{h}$ is a p-vector. Denote by $\underline{\theta}$ the unrestricted maximum likelihood estimator (mle) and $\tilde{\underline{\theta}}$ the restricted one. The score test or LM test is based on Taylor series approximations to the likelihood ratio test. If

$$L(\underline{\theta}) = \Sigma_1^n \log f(X_i|\underline{\theta}) \qquad \text{then}$$

$$LR = L(\tilde{\underline{\theta}}) - L(\hat{\underline{\theta}}) \underset{\sim}{} (\partial L/\partial\tilde{\underline{\theta}})^T \tilde{J}^{-1}(\partial L/\partial\tilde{\underline{\theta}}) = LM$$

where $\tilde{\underline{J}} = E(\partial^2 L/\partial\underline{\theta}\partial\underline{\theta}^T)\,|\tilde{\underline{\theta}}$.

Under the null hypothesis usually LM $\Rightarrow \chi^2_{r-p}$ (see Aitcheson and Silvey (1958)).

Often $L(\underline{\theta})$ has a special form that helps to simplify the LM. Thus if

$$L(\underline{\theta}) \propto \frac{1}{2}\Sigma_1^n e_k^2(\underline{\theta})/\sigma^2$$

$$\Rightarrow \quad \partial L/\partial\tilde{\underline{\theta}} = \tilde{\sigma}^{-2}\Sigma_1^n \tilde{e}_k d\tilde{e}_k/d\tilde{\underline{\theta}} = \tilde{\underline{z}}'\tilde{\underline{e}}/\tilde{\sigma}^2$$

$$\tilde{J} \simeq \Sigma_1^n de_k/d\tilde{\underline{\theta}} de_k/d\tilde{\underline{\theta}}'/\tilde{\sigma}^2 = \tilde{\underline{z}}'\tilde{\underline{z}}/\sigma^2$$

$$\Rightarrow \quad LM = \tilde{\underline{e}}'\tilde{\underline{z}}(\tilde{\underline{z}}'\tilde{\underline{z}})^{-1}\tilde{\underline{z}}'\tilde{\underline{e}}/\tilde{\sigma}^2 = (n-r)\tilde{R}^2 \qquad (1)$$

where $\tilde{R}^2$ is the coefficient of determination for the regression of $\tilde{\underline{e}}_k$ on $\tilde{\underline{z}}_k = de_k/d\tilde{\underline{\theta}}$ and $\tilde{\underline{z}}' = (\tilde{\underline{z}}_1\ \tilde{\underline{z}}_2 \ldots \tilde{\underline{z}}_n)$

Furthermore, often $\underline{h}(\underline{\theta})$ takes a simple form $\underline{h}(\underline{\theta}) = [I\ 0][\underline{\theta}_1'\underline{\theta}_2']' - \underline{\theta}_{10}$ so that

$$\partial L/\partial\tilde{\underline{\theta}} = \begin{pmatrix} \partial L/\partial\tilde{\underline{\theta}}_1 \\ \partial L/\partial\tilde{\underline{\theta}}_2 \end{pmatrix} = \begin{pmatrix} \partial L/\partial\tilde{\underline{\theta}}_1 \\ \underline{0} \end{pmatrix} = \begin{pmatrix} \tilde{\underline{z}}_1'\tilde{\underline{e}} \\ \underline{z}_2'\tilde{\underline{e}} \end{pmatrix}$$

$$\Rightarrow \quad LM = \tilde{\underline{e}}'\tilde{\underline{z}}_1\ (\tilde{\underline{z}}_1'\tilde{\underline{z}}_1 - \tilde{\underline{z}}_1'\tilde{\underline{z}}_2(\tilde{\underline{z}}_2\tilde{\underline{z}}_2)^{-1}\tilde{\underline{z}}_2'\tilde{\underline{z}}_1)^{-1}\tilde{\underline{z}}_1'\tilde{\underline{e}}/\tilde{\sigma}^2$$

$$= \tilde{\underline{e}}'\tilde{\underline{z}}_{1.2}(\tilde{\underline{z}}_{1.2}'\tilde{\underline{z}}_{1.2})^{-1}\tilde{\underline{z}}_{1.2}\tilde{\underline{e}}/\sigma^2 = (n-r)\tilde{R}^2_{1.2} \qquad (2)$$

where $\tilde{\underline{z}}_{1.2}$ is the residual from the regression of $\tilde{\underline{z}}_1$ on $\tilde{\underline{z}}_2$: while $R^2_{1.2}$ is the partial R^2 for regressing $\tilde{\underline{e}}$ (which is orthogonal to $\tilde{\underline{z}}_2$) on $\tilde{\underline{z}}_{1.2}$.

Again asymptotically $(n-r)R^2_{1.2} \sim \chi^2_{r-p}$; although since now we find ourselves in a regression setting we might prefer to use (under H_0)

$$\frac{n-r}{r-p} \frac{\tilde{R}^2_{1.2}}{1-\tilde{R}^2_{1.2}} \sim F_{r-p,n-r}$$

In fact in some cases this is accurate (see Milliken and Graybill, 1970).

Incidentally the two schemes (1), (2) for generating LM may have advantages depending on the situation. In AOV, regression on $\underline{z}_2$ often involves just taking means; then (2) is useful. In ordinary regression (1) is appropriate. Note that (2) is very close to two-stage regression. It is as if we had regressed a y on $\underline{z}_2$ to give $\underline{e}$ and now we regress $\underline{e}$ on $\underline{z}_{1.2}$.

Silvey (1959, p. 399) has indicated that for local alternatives the asymptotic noncentral distribution of LM is noncentral χ^2. His argument is based on the corresponding result for the likelihood ratio test. A direct argument is sketched in Appendix B. The result is

$$LM \Rightarrow \chi^2_{r-p}(\phi)$$

$$\phi = (n-r)\delta^2 = (n-r)R^2_{1.2} = (n-r) \lim \tilde{R}^2_{1.2}$$

It is suggested that for diagnostic tests, local alternatives may be reasonable.

We now turn to interpretation for δ^2 or $R^2_{1.2}$. The standard meaning for $\tilde{R}^2_{12}$ is that it is the percentage change in RSS (=residual sum of squares) due to the addition of the constructed variable $z_1 = de/d\theta_1$.

$$\tilde{R}^2_{1.2} = \frac{RSS_2 - RSS_{12}}{RSS_2} = \frac{1-R^2_2-(1-R^2_{12})}{1-R^2_2} = \frac{R^2_{12}-R^2_2}{1-R^2_2}$$

where R^2_2 = percent of SS explained by $\tilde{z}_2$; R^2_{12} = percent explained by $\tilde{z}_1$, $\tilde{z}_2$. Another way to put this is that the percent loss in not adding z_1 is $\tilde{R}^2_{1.2}/(1-\tilde{R}^2_{1.2})$.

We can re express this in terms of MSE_p. Relabel $r = p_{12}$; $p = p_2$. The percent loss in MSE in not using $\tilde{Z}_1$ is

$$\varepsilon = (MSE_2 - MSE_{12})/MSE_{12}$$

The natural estimate of this is

$$\hat{\varepsilon} = (C_{p_2} - C_{p_{12}})/C_{p_{12}}$$

$$= [(RSS_2 - RSS_{12})/\sigma^2 + 2(p_1 - p_{12})]/(RSS_{12}/\sigma^2 + 2p_{12} - n)$$

$$= \left[\frac{\tilde{R}^2_{1.2}}{1-\tilde{R}^2_{1-2}} + \frac{2(p_1-p_{12})}{(n-p_{12})} \frac{\sigma^2}{\sigma^2_{12}}\right] / \left(\hat{\sigma}^2_{12}/\sigma^2 + \frac{2p_{12}-n}{n-p_{12}}\right)$$

where $\hat{\sigma}^2_{12} = RSS_{12}/(n-p_{12})$. For small $R^2_{1.2}$ and large n this is approximately

$$\hat{\varepsilon} = [(n-p_{12})\tilde{R}^2_{1.2} + 2(p_1-p_{12})]/p_{12}$$

We see the percent change has two parts. An increase in prediction variance of (on average) $(p_1-p_{12})/p_{12}$ offset by a reduction in bias ^{2} of (on average) $[(n-p_{12})R^2_{1.2} + (p_1-p_{12})]$. Only if this offsets the loss sufficiently do we modify the model.

As a simple example consider the use of the Box-Tidwell procedure by Cook and Weisberg (1982, p. 66, p. 83) to analyse some data on volume (V), diameter (D), height (H) of 31 trees. They consider transforming both height and

diameter (one expects of course $V \alpha HD^2$). We consider height here. They tive (p. 83) $t = .405/1.762 \Rightarrow F_{1.2} = t^2 = .053 \simeq LM$. So

$$\hat{\varepsilon} = (.053 + 2(-1))/5 = -39\%.$$

(regressors are intercept, D, H, D log D, H log H)

From this point of view clearly H should remain untransformed. However the $NC\chi^2_1$ distribution gives a very wide CI for $\hat{\varepsilon}$. From the Biometrika tables we find (starting with $\sqrt{.053} = .23$) $(\phi_-, \phi_+) = (0,4) \Rightarrow (\varepsilon_-, \varepsilon_+) = (-40\%, +40\%)$. These numbers look large, allowing a +40% loss yet it seems pointless to transform. The logical conclusion to draw here (to deal with the large numbers) is that since the bias depends on n we should look at percent loss per free observation. Then $\varepsilon/(n-p_{12}) \simeq [\tilde{R}^2_{1-2} + 2(p_1' - p_{12})/n]/p_{12}$. Thus CI on percent loss per observation is $(-40, +40)/(31 - 5) = (-1.54, + 1.54)\%$. This presents the conclusion not to transform more clearly (c.f. earlier comments about partial signal to noise ratio for regression in Section 3)

VII. CANONICAL CORRELATIONS

Suppose $(\underset{\sim}{Y}_{pxl,i}, \underset{\sim}{X}_{qxl,i})$ $i = 1,\ldots,n$ are independent and jointly Gaussian and consider

$$H_0: \text{ Rank } (\underset{\sim}{\Sigma}_{YX}) = k: \quad \underset{\sim}{\Sigma}_{YX} = \text{cov}(\underset{\sim}{Y},\underset{\sim}{X})$$

we can restate this hypothesis as

$$H_0': \quad \rho_{k+1} = 0 = \ldots = \rho_p$$

where ρ_i are singular values of $\underset{\sim}{\Sigma}_y^{-\frac{1}{2}}\underset{\sim}{\Sigma}_{YX}\underset{\sim}{\Sigma}_X^{-\frac{1}{2}}$ and $\underset{\sim}{\Sigma}_Y^{\frac{1}{2}}$ is a positive definite symmetric square root of $\underset{\sim}{\Sigma}_y = \text{var}(\underset{\sim}{Y})$ etc.

There are 3 standard test statistics

Wilks lambda	$W = -\Sigma_{k+1}^{p} \log(1-r_i^2)$
Hotellings T_0^2	$T_0^2 = \Sigma_{k+1}^{p} r_i^2/(1-r_i^2)$
Pillai's V	$V = \Sigma_{k+1}^{p} r_i^2$

where r_i are singular values of $\underset{\sim}{S}_y^{-\frac{1}{2}}\underset{\sim}{S}_{yx}\underset{\sim}{S}_x^{-\frac{1}{2}}$ where $\underset{\sim}{S}_y$ is the sample estimate of $\underset{\sim}{\Sigma}_y$ etc.

There are, as before, two types of NCD. Many workers have contributed to the development of these results but a compact collection and latest development has been given by Fujikoshi. To keep things simple only results for T_0^2 are quoted

(a) Local Alternative: Fujikoshi (1980)

$$\rho_i^2 = c_i^2/n, \quad i = k+1\ldots p$$

$$(m-\nu)\, T_0^2 \Rightarrow \chi_f^2(\phi) \qquad \text{(to order } n^{-1})$$

$$\phi = \frac{1}{2}\, n\, \Sigma_{k+1}^{p}\rho_i^2 = \frac{1}{2}\, \Sigma_{k+1}^{p}\, c_i^2$$

$$m = n-q-p-1 + \Sigma_{k+1}^{p}\rho_i^{-2}$$

ν is given by Fujikoshi; it is always $< < n$.

It should be noted that many accounts of multivariate tests omit the correction term $\Sigma_{k+1}^{p}\rho_i^{-2}$: it can make an enormous difference.

(b) General Alternative: Fujikoshi (1977)

$$\sqrt{n}\,(T_0^2 - \Sigma_{k+1}^{p}\rho_i^2/(1-\rho_i^2))/\tau \Rightarrow N(0,1) \quad \text{(to order } n^{-\frac{1}{2}}\text{)}$$

$$\tau^2 = 4\Sigma_{k+1}^{p}\,\rho_i^2/(1-\rho_i^2)^2.$$

Remark 1. Fujikoshi provides higher order terms for both (a), (b) which are actually very easy to use if tedious to state. Their use seems mandatory since the order of approximation can be improved to n^{-2} and $n^{-3/2}$ respectively.

Remark 2. It has been emphasized by Muirhead (1982) that these results are sensitive to the Gaussianity assumption. A simple correction is to model the distribution as elliptically contoured (ie a mixture of multivariate Gaussians mixed on a scalar scale parameter σ). Then if $3K$ is the kurtosis of any marginal distribution the above results hold if we replace τ^2 by $(1+K)\tau^2$ in (b) and ϕ by $\phi(1+K)^{-1}$, W by $W(1+K)^{-1}$ in (a) etc.

Much as before an interpretation of the NCP's comes from prediction. Consider the problem.

Find $\underset{\sim}{G}_{pxq}$ of Rank $k \ni E||Y-GX||^2_{\Sigma_{y.x}^{-1}}$ = min. This has

min = $p + \Sigma_{k+1}^{p}\,\rho_i^2/(1-\rho_i^2) = p+t_0^2$ (cf Brillinger, 1975).

So that t_0^2/p measures a percent loss in prediction variance on using only k linearly independent rows of X to predict Y. With this in mind and guided by the development of C_p, Fujikoshi and Veitch (1979) have provided a C_k function for choice of k namely

$$\tilde{C}_k = m_k T_{0k}^2 - 2(p-k)(q-k).$$

In view of the above NCD's it seems advisable to modify the formula. In any case the same suggestion is offered here as in the regression case. Plot $\tilde{C}_k$ for a minimum and then for that value construct a CI using both NCD's, then investigate those models whose $\tilde{C}_k$ values fall inside the CI.

VIII. CONNEXION TO EXACT SLOPE

It was observed early in the article that the P-value, $P_n(\underline{X})$ is a qualitative measure of departure between H_0 and actuality H. There is in fact a direct connexion between the P-value and the NCP which is implicit in the work of Bahadur (1971).

Let $F_n(t)$ be the distribution of a test statistic $T_n(\underline{X})$. Under H_0, $F_n(t)$ is a central distribution; under H it is the NCD. Bahadur shows that if (under H)

(a) $n^{-1/2}T_n(\underline{X}) \longrightarrow b(\underline{\theta}) = \delta$ say

i.e. there is a well defined standardised NCP

(b) There is a large deviation limit for the NCD

i.e. under H $\quad n^{-1/2}\log(1 - F_n(n^{1/2}t)) \longrightarrow - f(t)$ say

Then

$$n^{-1}\log P_n(\underline{X}) \longrightarrow f(b(\underline{\theta})) = C(\underline{\theta})$$

where $C(\underline{\theta})$ is called the exact slope.

For our purposes it is more useful to write

$$n^{-1}\log P_n(X) \longrightarrow f(\delta) \qquad (A)$$

Thus we see there is a monotonic function f (·) linking the P-value to the NCP δ.

We could use this connexion to find a CI on δ by means of the asymptotic results conveniently provided by Lambert and Hall (1982) namely (in the present setting)

$$n^{1/2}(n^{-1}\log P_n(\underline{X}) + f(\delta))/\tau \Rightarrow N(0,1) \quad \text{if} \quad \sqrt{n}(T_n - \delta)/\sigma \Rightarrow N(0,1)$$

where $\tau = \sigma f'(\delta)$; $\sigma = \sigma(\underline{\theta})$ (Since from (A) $n^{-1}\log P_n(\underline{X}) \sim f(T_n)$ this is intuitively reasonable from the delta method). This is clearly no advantage here in the width of CI over the earlier CI's presented.

IX. CONCLUSION AND SUMMARY

In this article it has been suggested that a CI on a scalar noncentrality parameter (NCP) has the same advantage as the PST of providing data summary in a single number yet avoids two weaknesses of PST's. One, of failing to make clear the difference between practical and statistical significance and second (and related) the fact that a PST is only a qualitative measure of discrepancy between null and actuality.

The use of the proposed technique has been illustrated in a number of examples. In some cases the NCP has a direct physical interpretation so its use is more or less easy. In other cases the interpretation is much more subtle and some detailed calculations are needed to establish a meaning for the scalar NCP: an example of this was given for goodness of fit tests. In this respect there is a lot of work to be done. This type of situation typically occurs in diagnostic uses of PST's.

The technique was also applied to model selection with C_p. It was suggested that candidate models be those whose C_p values fall inside, the CI on the minimum C_p value. An analysis of this procedure needs to be done.

Finally there are the noncentral distributions (NCD) themselves. It seems that mostly asymptotic results must be used; however if higher order terms are available these can be quite accurate. Still, some further results are needed, although limit NCD's for multivariate tests seems complete.

Acknowledgement. To George Tiao and Arnold Zellner for discussion.

Appendix A. Use of noncentral F to find a CI

According to the approximation given by Patnaik (see Kendall and Stuart 24.33) the percentage points of the noncentral F are obtained from

$$(\phi) \qquad F^{(+)}_{\nu_1,\nu_2}(\phi_-) = \frac{\nu_1+\phi_-}{1} F^{(+)}_{\nu_1,\nu_2}$$

$$(\phi) \qquad F^{(-)}_{\nu_1,\nu_2}(\phi_+) = \frac{\nu_1+\phi_+}{\nu_1} F^{(-)}_{\nu,\nu_2} = \frac{(\nu_1+\phi_+)}{\nu_1 F^{(+)}_{\nu_2,\nu}}$$

where + denotes an upper percentage point; $\nu = \frac{(\nu_1+\phi)^2}{\nu_1+2\phi}$

To find a CI on ϕ we solve these equations separately by trial and error e.g. below we find the (ϕ_+) point for a 95% CI.

(1) Hald data e.g. $F = (2.7-6+5)/2 = .85$; $\nu_1 = 2$, $\nu_2 = 8$

Trial (1) $\phi = 8 \Rightarrow \nu = (2+8)^2/(2+16) = 5\frac{1}{2} \sim 6 \Rightarrow F^{(+)}_{8,6} = 5.6 \Rightarrow \phi = \nu_1(F^{(+)}F-1)$

$= 2(5.6 \times .85-1) = 7.5$ which is close enough $\Rightarrow MSE^{(+)} = k+\phi_k = 3+8=11.$

(2) Octane example

(ϕ_+) $F = 6.78$; $\nu_1 = 4$, $\nu_2 = 15$

Trial (1)

$\phi_+ = 12 \Rightarrow \nu = (4+12)^2/(4+24) = 9.14 \Rightarrow 9$

$\Rightarrow F^{(+)}_{15,9} = 3.77 \Rightarrow \phi = \nu_1(F^+F-1) = 98.2$

Trial (2): $\phi_+ = 70 \Rightarrow \nu = (4+70)^2/(4+140) = 38$

$\Rightarrow F^+_{15,18} = 2.18 \Rightarrow \phi = \nu_1(F^+F-1) = 55$

Trial (3): $\phi_+ = 60 \Rightarrow \nu = (4+60)^2/(4+120) = 33$

$\Rightarrow F^+_{15,33} = 2.31 \Rightarrow \phi = 58.6 \quad \Rightarrow \phi \sim 59$ which is close enough

$\Rightarrow \delta = \sqrt{\phi/4} = 3.84$

APPENDIX B

Heuristics for the asymptotic distribution of LM under a local alternative.

We have $L(\underline{\theta}) = \Sigma_1^n \log f(X_i|\underline{\theta})$

so that if $\underline{\theta}_0$ is the true value, then under some regularity conditions the standard result

$$J = -E(\partial^2 L/\partial\underline{\theta}_0\partial\underline{\theta}_0') = E(\partial L/\partial\underline{\theta}_0 \partial L/\partial\underline{\theta}_0')$$

should follow. So that, if $\underline{H}$ is any fixed matrix

$$\text{var } (H'J^{-1}H)^{-1/2}H'J^{-1}\partial L/\partial\underline{\theta}_0 = \underline{I}$$

Also we similarly expect

$$n^{-1}J \Rightarrow \underline{\mathcal{J}}$$

$$(H'J^{-1}H)^{-1/2}\, H'J^{-1}\partial L/\partial\underline{\theta} \Rightarrow \underline{Z} = N(\underline{0},\underline{I})$$

The problem

$$\max_{\underline{\theta}} L(\underline{\theta}) \quad \ni \underline{h}(\underline{\theta}) = \underline{0}$$

has solution

$$\partial L/\partial\tilde{\underline{\theta}} + \tilde{H}\tilde{\lambda} = \underline{0}$$

where $\tilde{H} = \partial h'/\partial\tilde{\underline{\theta}}$ and $\tilde{\theta}$ is the maximising value.

Consider the joint Taylor series about $\underline{\theta}_0$

$$\underline{0} = \partial L/\partial\underline{\theta}_0 + \underline{J}(\tilde{\underline{\theta}} - \underline{\theta}_0) + \underline{\tilde{H}\tilde{\lambda}} + O_p(n^{-1/2})$$

$$\underline{h}_0 = \underline{h}(\underline{\theta}_0) = \underline{H}'(\tilde{\theta} - \underline{\theta}_0) + O_p(n^{-1/2})$$

Again under suitable regularity conditions these should be readily established.

Also we can expect $\tilde{\underline{H}} \Rightarrow \underline{H} = \partial\underline{h}'/\partial\underline{\theta}_0$.

From this we deduce (neglecting terms $O_p(n^{-1/2})$)

$$-\underline{J}^{-1}\partial L/\partial\underline{\theta}_0 = \tilde{\underline{\theta}} - \underline{\theta}_0 + \underline{J}^{-1}\underline{H}\,\tilde{\underline{\lambda}}$$

$$\Rightarrow -\tilde{\underline{H}}'\underline{J}^{-1}\partial L/\partial\underline{\theta}_0 = \tilde{\underline{H}}'\underline{J}^{-1}\tilde{\underline{H}}\tilde{\underline{\lambda}} - H'\underline{h}_0$$

$$\Rightarrow LM = \tilde{\lambda}'\tilde{\underline{H}}'\underline{J}^{-1}\tilde{\underline{H}}\,\underline{\lambda} = (\hat{\underline{z}}+\underline{\delta})'(\hat{\underline{z}}+\underline{\delta})$$

where

$$\hat{\underline{z}} = (\tilde{H}'J^{-1}\tilde{H})^{-1/2}\,\tilde{H}'J^{-1}\partial L/\partial\underline{\theta}_0 \Rightarrow \underline{Z}.$$

$$-\underline{\delta} = (\tilde{H}'J^{-1}\tilde{H})^{-1/2}\tilde{\underline{H}}'\underline{h}_0 \simeq n^{1/2}\,(\tilde{H}'\mathcal{I}^{-1}\tilde{H})^{-1/2}\,\underline{H}'\underline{h}_0$$

so if we suppose a local alternative constraint

$$\underline{h}_0 = \tilde{\underline{h}}_0\, n^{-1/2}$$

Then we find

$$LM \Rightarrow x^2_{r-p}(\phi)$$

$$\phi = \tilde{\underline{h}}_0'\underline{H}(\underline{H}'\mathcal{I}^{-1}\underline{H})^{-1}\underline{H}'\tilde{\underline{h}}_0$$

as claimed.

It seems clear that the type of regularity conditions used by Aitcheson and Silvey (1958) and Silvey (1959) will justify the present claim. Incidentally the line of argument just given is rather shorter than the corresponding null result derived in the above two articles.

REFERENCES

J. Aitcheson, D. D. Silvey (1958), Maximum likelihood estimation of parameters subject to restraints. Ann. Math. Stat., 29, p. 813-828.

A.C. ATKINSON (1982), Regression disgnostics, transformations and constructed variables (with disc) J. Roy Stat. Soc., (B), 44, 61-36.

R.R. BAHADUR (1971) Some Limit Theorems in Statistics, SIAM, Philadelphia.

A. BIRNBAUM (1977), 'The Neyman-Pearson Theory as decision theory, and as inference theory, with a criticism of the Lindley-Savage argument for Bayesian theory, Synthese.

G.E.P. BOX and P.W. TIDWELL (1962), Transformations of the independent variables, Technometrics, 4, p. 591-550.

T.S. BREUSCH and A.R. PAGAN (1980), The Lagrange Multiplier Test and its application to model specification in Econometrics, Rev. Econ. Stud., 47, p. 239.

D.R. BRILLINGER (1975), Time Series, Data Analysis and Theory, Holt, Reinhart, Winston Inc., Sydney

J.D. BROFFIT and R.H. RANDLES (1977), A power approximation for the chi-square Goodness of Fit Test: simple hypothesis case, Jl. Am. Stat. Assoc., 72, p. 604.

COCHRAN

R.D. COOK and S. WEISBERG (1982), Residuals and Influence in Regression, Chapman and Hall.

D.R. COX (1977), The role of significance tests, scand. J. Stat., 4, p. 49-70.

N.A. DRAPER and H. SMITH (1981), Applied Regression Analysis (2nd ed.), J. Wiley, New York.

J.L. FLEISS (1981), Statistical Methods for Rates and Proportions, J. Wiley, New York.

Y. FUJIKOSHI (1977), Asymptotic Expansions for the distributions of some multivariate tests, in Multivariate Analysis - IV, P.R. Krishnaiah (ed), North-Holland.

Y. FUJIKOSHI (1980), Asymptotic Expansions for the distributions of some multivariate tests under local alternatives, Tech. Report #27, Stat. Research Group, Hiroshima University, Hiroshima, Japan.

Y. FUJIKOSHI and L.G. VEITCH (1979), Estimation of dimensionality in canonical correlation analysis, Biometrika, 66, p. 345.

A.E. Gelfand (1983), Estimation in noncentral distributions, Comm. Stat. Theor. Meth. p. 463-475.

A. HALD (1960), Statistical Theory with Engineering Applications, J. Wiley, New York.

F. HERNANDEZ and R.A. JOHNSON (1980), The large-sample behavior of transformations to normality, Jl. Am. Stat. Assoc., 75, p. 855-861.

P.W.M. JOHN (1971), Statistical Design and Analysis of Experiments, MacMillan, New York.

O. KEMPTHORNE (1976), 'Of what use are tests of significance and tests of hypotheses,' Commun. Statist. - Theor. Math, A5(8), p. 763-777.

M. KENDALL and A. STUART (1979), The Advanced Theory of Statistics, Vol. II, MacMillan, New York.

D. LAMBERT and W.J. HALL (1982) Asymptotic abnormality of P-values (Ann. Stat., 10, 44-64).

C. MALLOWS (1973), Some Comments on C_p, Technometrics, 15, p. 661.

G.A. MILLIKEN and F.A. GRAYBILL (1970), Extensions of the general Linear Hypothesis model, Jl. Am. Stat. Assoc., 65, p. 797-807.

D.S. MOORE (1983), Measures of Lack of Fit from tests of chi-squared type, Tech Report #83-12, Dept. Statistics Purdue University.

F. MOSTELLER and J. TUKEY (1977), Data Analysis and Regression.

R. MUIRHEAD (1982), Aspects of Multivariate Statistical Theory, J. Wiley, Brisbane.

P.B. PATNAIK (1949), The non-central χ^2 and F-distributions and their applications, Biometrika, 36, p. 202.

R.L. PLACKETT (1981), The analysis of categorical data, MacMillan, New York, (2nd. ed.)

D. PREGIBON (1981), Logistic Regression diagnostics, Ann. Stat a p. 705.

M. RAGHAVACHARI (1973), Limiting distributions of Kolmogorov - Smirnov type statistics under the alternative, Ann. Stat., 1, p. 67-73.

C.R. RAO (1973), Linear Statistical Inference and its Application (2nd ed.) J. Wiley, New York.

H. ROBBINS (1970), Statistical Methods Related to the Law of the iterated logarithm, Ann. Math - Stat., 41, p. 1397-1409.

H. SCHEFFE (1959), The Analysis of variance, J. Wiley.

G.A.F. SEBER (1977), Linear Regression Analysis, J. Wiley, Brisbane.

S.D. SILVEY (1959), The Lagrangian Multiplier, Ann. Math., Stat., 30, p. 389-407.

M. STONE (1977), An asymptotic equivalence of choice of model by Cross-validation and Akazke's criterion, J.L. Roy, Stat. Soc. (B), 36, p. 111-147.