THE ECONOMIC RESEARCH INSTITUTE

Memorandum No 18

Comparison of the Average Critical Point and the Power Function Methods of Adjudging Relative Efficiency of Statistical Tests.

By R.C. Geary

The situation is that in which we have a random sample of n measured entities in regard to which a decision has to be made between two specific hypotheses, H_o and H_1, with regard to the populations from which the sample has been drawn, e.g. is population mean more likely to be μ_o or μ_1, these parameters having specific values? We also consider two test expressions, specific functions of the sample values, for the purpose of making decisions and we wish to ascertain which of the two tests is the more efficient. Decision in every case will be determined by the value of the test function. We prefer the test function which in the long run, i.e. after an indefinitely large number of experiments, will lead to the greater number of correct decisions.

The Power Function (PF) is on the following lines. We first consider the situation in which the hypothesis H_o is correct and we assume that we have a probability table for each of the test functions. We then fix a certain high probability level (say .95, .99, etc) and determine the corresponding "probability points", i.e. the range of values of the test functions which are such that the probability of finding values of the functions inside the range have the stated probability. Outside the range is the "rejection zone" when the hypothesis H_o obtains, clearly, <u>with both tests</u>, decision is correct in 95%, 99% etc of cases. The tests are equally efficient for this purpose.

But if H_1 is correct the efficiency of the two tests, as the following illustration will show, may be quite different. The right decision will be made with each test function in a theoretically ascertainable proportion of cases. This proportion is the PF of the test. It equals the probability, when H_1 is true, of finding the value of the test function in the H_o- hypothesis rejection zone. Of the two tests one prefers, in any particular problem, the test with the higher PF. As between a test A and a test B, if the PF of test A is greater than that of test B, then proportionately the same number of correct decisions will be made if H_o is correct but a larger number of correct decisions using A if hypothesis H_1 is correct. Hence with both hypotheses under consideration proportionately more successes will be attained using test A. This is a brief synopsis of the Neyman-Pearson theory [3].

There can be no question about the theoretical superiority of the PF approach. It unfortunately happens, however, that it is extremely difficult to apply this theory in practical applications. Very few probability distributions of test functions are known exactly and the situation often arises in which, even if one knows the distribution when H_o is true, it is not possible, given any sample size n, to derive the distribution when H_1 obtains. It is for this reason that, many years ago, the author [2] used as a compromise, what he now terms the Average Critical Point (ACP) method which is strictly applicable only in the asymptotic case, i.e. when sample size n is indefinitely large. In his work then he found, in particular applications, that conclusions as to

relative efficiency of tests which obtained in the asymptotic case did not necessarily apply when sample size was finite: the ACP method is subject to this reservation. Certain only is it that, in practice, the ACP method can be applied far more widely than PF.

The ACP Method

Suppose we have a test function $t(x_1, x_2, \ldots, x_n)$ where $x_1, x_2, \ldots, x_n$ are the measures of n random drawings from a population $f(x, \mu)$ (continuous in μ for all values of x) of defined form but with a parameter μ, the value of which is at present undetermined. We wish to decide from the sample, using the test t, whether in the population the value of μ could plausibly be taken as zero or whether μ has probably some value greater than zero. Clearly if μ is very small (without defining "smallness") no test will be sensitive enough to yield an answer (when n is finite). We propose to reject the hypothesis that in the population $\mu = 0$ when the value found for t in the particular sample is greater than, say, the .95 probability point, say τ, on the $\mu = 0$ population hypothesis, the critical value. If $\mu > 0$ we are naturally interested in the values of t which are "near" τ, some greater and some less. We therefore set

$$E\, t(\mu) = \tau, \tag{1}$$

where E is the "expected" value, or the mean of an indefinitely large number of sample values. Assuming that $E\, t(\mu)$ varies monotonically with μ, (1) is solved for μ by an identifiably unique value $\mu = M$ so that

$$M = \varphi(\tau). \tag{2}$$

Now if there are two tests t_1 and t_2 with .95 probability points τ_1 and τ_2 yielding critical values of μ, namely M_1 and M_2, t_1 is the better, or more sensitive, test if $M_1 < M_2$. It is only when sample size n is indefinitely large that this ACP method has absolute validity in the stochastic sense, for in that case all the values of t equal τ when variance t is $O(n^{-a})$, $a > 0$, as will ordinarily be the case. Note that the method involves only the calculation of E $t(\mu)$, the first moment, as distinct from deriving the frequency distribution of $t(\mu)$ (an incomparably more difficult algebraic problem), required for the PF method.

Study of a Simple Problem

A sample of n is drawn at random from a normal population with variance unity and mean μ. We wish to test in probability whether μ is zero or some positive quantity. The two tests proposed are the Gosset-Fisher t and the count of signs (+ or -) using the binomial. For the latter we require the probability $P(\mu)$ of a + sign on a single drawing; this is

$$P(\mu) = \frac{1}{\sqrt{2\pi}} \int_0^\infty dx\, e^{-(x-\mu)^2/2} = \frac{1}{\sqrt{2\pi}} \int_{-\mu}^\infty dx\, e^{-x^2/2} \tag{3}$$

Given μ, the values of $P(\mu)$ can be obtained from the normal probability table [1]. In a sample of n the probability of obtaining m + signs is

$$\binom{n}{m} [P(\mu)]^m \; [Q(\mu)]^{n-m}$$

where $Q(\mu) = 1 - P(\mu)$.

Let n = 20. On the H_o hypothesis $\mu = 0$ and $P(0)=\frac{1}{2}$. We find from the binomial table for n = 20 that while almost equal numbers of + and - signs are to be expected that the probability of a tail of 5 or fewer minus signs is .0207, an arbitrary but conveniently low value. We accordingly adopt the rule that we shall reject the nul-hypothesis (correctly if, in fact, $\mu > 0$) when we find 5 or fewer - signs.

Let μ have a few specific values. When $\mu = 1$ $P(1) = .8413$ so that $Q(1) = .1587$. The probability of getting 5 or fewer - signs with this distribution is

$$(.8413)^{20}+\binom{20}{1}(.8413)^{19}(.1587)+\ldots+\binom{20}{5}(.8413)^{15}(.1587)^{5}$$

$$= .9158$$

which is the power of the sign count test. If μ is in fact 1 decision will be correct in about 92% of cases.

The rival test should be the Gosset-Fisher t. However, since n is as large as 20 it will be assumed that the population variance is known (in fact = 1) and the test function will simply be the mean $\bar{x}$*. The known variance of $\bar{x}$ is $1/\sqrt{20} = 0.2236$. The nul-hypothesis probability point corresponding to the above probability of .0207 is 2.04: we adopt the rule that the nul-hypothesis will be rejected if $\bar{x}$ is greater than $2.04 \times 0.2236 = 0.4561$.

---

* Strictly, the power function should be based on the probability of $t = \sqrt{n}/s$ from a population $N(\mu; 1)$. This is a very complicated function and the value of the power could not differ much from the value .9925 found above.

If, however, $\mu = 1$, according to this rule, using $\bar{x}$, the probability will be that of obtaining a value greater than $-(1-0.4561)/0.2236 = -2.43$. This probability from the normal table is .9925, the power of the test $\bar{x}$. Accordingly, by reference to the two power values .9158 and .9925 the $\bar{x}$ test is far more efficient than the sign count test for discriminating between populations with means $\mu = 0$ and 1. The powers corresponding to hypotheses $\mu = 0$ versus (in succession) certain values of μ are as follows:-

	$\mu = 0$ versus $\mu =$					
Test	(1) 1.0	(2) 0.9	(3) 0.8	(4) 0.7	(5) 0.6	(6) 0.5
1 Sign	.9158	.8531	.7637	.6495	.5183	.3843
2 $\bar{x}$	.9925	.9767	.9382	.8621	.7389	.5793

(See appended diagram)

There can be no question of the superiority of $\bar{x}$ as a test compared with the sign count, for all the far greater convenience, and indeed the greater scope, of the sign count: for example, using the sign count to establish the relationship over a series of years between the unemployment and marriage rates all one has to do is to count inverse concordances between consecutive years and use the point binomial

$$(\tfrac{1}{2} + \tfrac{1}{2})^{n-1}$$

where n is number of years. Furthermore, the method can be used when only qualitative judgments are available even if actual measurements (necessary for the use of $\bar{x}$) cannot be made. Finally the writer has encountered particular cases in which the sign count was more sensitive than $\bar{x}$. When significance may be adjudged on the sign count, this test can confidently be used.

Of course there is nothing new in what has gone before. The writer merely deems it desirable to set out certain, now almost classical, considerations as a preamble to the main theme of the memorandum which is to compare this power function approach to the ACP method of making comparisons of sensitivity of tests, by reference to this simple application, one of the few in which it is practicable to derive the power functions.

The ACP treatment in this application is very simple. On the sign count test the rule, as before, will be to reject the nul-hypothesis when the count of +'s is 15 or more, corresponding to a nul-hypothesis probability of .0207. The 15 will be the mean found from an indefinitely large number of experiments of drawing samples of 20 when the population probability is 15/20 = 0.75. The corresponding value of μ, namely M_2 is found from (2) with $P(\mu) = .75$. We find $M_2 = 0.675$. With the $\bar{x}$ test the value of μ, namely M_1, is found as $M_1 = 2.04/\sqrt{20} = 0.456$, the 2.04, as before, being the nul-hypothesis probability point corresponding to probability .0207. Since $M_1 < M_2$, in the long run a positive aberration μ from the nul-hypothesis mean zero will be identified at a lower value of μ than using the count of signs method. We may regard the $\bar{x}$ test as the more "sensitive".

As implied earlier, the ACP approach, by reference to the present application, seems to exaggerate somewhat the relative superiority of $\bar{x}$ over the sign count, i.e. given any level of the power function the discrepancy between the critical levels will be somewhat less than between the $M_1 = 0.456$ and $M_2 = 0.675$ shown above for ACP. Qualitatively,

however, the findings are essentially the same. A diagram is appended illustrating the powers of the two tests for different values of the parameter μ. The power of $\bar{x}$ for $\mu = M_1 = 0.456$ is, of course, exactly $\frac{1}{2}$. The "horizontal" lines show the corresponding values of μ, given certain power levels. At power 0.5 the μ value for the sign count test is about 0.587 compared to the $M_2 = 0.675$.

References

[1] Biometrika Tables for Statisticians, edited by E. S. Pearson and H. O. Hartley, Second Edition (Cambridge University Press, 1954).

[2] Geary, R. C. (1947): Biometrika, Vol. XXXIV, Parts III and IV, 210.

[3] Neyman, J. and Pearson, E. S. (1933): Philos. Trans. A, 231, 289

August 1964

DIAGRAM

Power of Two Tests for Different Values of a Parameter μ