Testing the Approximate Validity of Statistical Hypotheses

By J. L. Hodges, Jr. and E. L. Lehmann*

University of California, Berkeley

[Received March, 1954]

SUMMARY

The distinction between statistical significance and material significance in hypotheses testing is discussed. Modifications of the customary tests, in order to test for the absence of material significance, are derived for several parametric problems, for the chi-square test of goodness of fit, and for Student's hypothesis. The latter permits one to test the hypothesis that the means of two normal populations of equal variance, do not differ by more than a stated amount.

1. *Introduction*

When testing statistical hypotheses, we usually do not wish to take the action of rejection unless the hypothesis being tested is false to an extent sufficient to matter. For example, we may formulate the hypothesis that a population is normally distributed, but we realize that no natural population is ever exactly normal. We would want to reject normality only if the departure of the actual distribution from the normal form were great enough to be material for our investigation. Again, when we formulate the hypothesis that the sex ratio is the same in two populations, we do not really believe that it could be exactly the same, and would only wish to reject equality if they are sufficiently different. Further examples of the phenomenon will occur to the reader.

In practice, this imprecision in the formulation does not usually cause much trouble, since the tests employed are sufficiently lacking in power that we do not run serious risk of rejecting the hypothesis unless it is false to a considerable extent. But whenever the available data are extensive, the tests may become embarrassingly powerful, leading to a paradox enunciated by Berkson (1938):

"I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small. Having observed this, and on reflection, I make the following dogmatic statement, referring for illustration to the normal curve: 'If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large—for instance, on the order of 200,000—the chi-square P will be small beyond any usual limit of significance.'

"This dogmatic statement is made on the basis of an extrapolation of the observation referred to and can also be defended as a prediction from *a priori* considerations. For we may assume that it is practically certain that any series of real observations does not actually follow a normal curve *with absolute exactitude* in all respects, and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square P will be small if the sample has a sufficiently large number of observations in it.

"If this be so, then we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all!"

It seems to us that this difficulty can be avoided by making a clear distinction, in the formulation of the problem, between "statistical significance" and what might be called "material significance".

* This paper was prepared with the partial support of the Office of Naval Research, U.S.A.

In the space of the parameters we may distinguish a set H_0 of values, representing the idealized hypothesis as it is customarily formulated. About the set H_0 we may then distinguish a larger set H_1 of values, representing situations close enough to H_0 that the difference is not materially significant in the problem at hand. If we knew that some distribution θ in H_1 were the true one, we should still wish to accept the hypothesis idealized by H_0. The size of the test should be the maximum value of the power function $\beta(\theta)$ in H_1, not H_0. Outside of H_1, however, we want to reject, so that consistency outside of H_1 is no longer undesirable. We reject as soon as there is statistically significant evidence that the departure from H_0 is materially significant.

It might be objected that there is nothing novel in the point of view just presented. We may just forget about H_0 and let H_1 play the role of the hypothesis in the classical formulation. There is however often a practical advantage in keeping H_0, as it is mathematically simple and corresponds to the idea underlying the situation to be tested. The boundaries of H_1 will be less precise in the experimenter's mind. It will usually be best to introduce into the space of parameters a measure, say $\Delta(\theta)$ of the "distance" of θ from H_0 on a scale reflecting at least roughly the materiality of departures from H_0, and then define H_1 as the set of those θ for which $\Delta(\theta)$ does not exceed a specified value Δ_0. The choice of Δ_0 will present problems similar to those encountered in choosing the alternative at which specified power is to be obtained.

If such a formulation is adopted, it is necessary to modify the customary tests, so as to adjust them to the new situation. This will be done briefly in section 2, for a few simple parametric situations which require only trivial modifications of the standard tests. In section 3 a modification of Student's problem is treated, and in section 4 a discussion from the present point of view is given of the χ^2-test for goodness of fit.

2. *Some Parametric Problems*

The simplest situation of the kind described in the introduction arises when a single parameter is involved, as for example in the case of a sample from a binomial or Poisson variable. We then have a (vector-valued) random variable X, with generalized probability density p_θ. For testing $H : \theta = \theta_0$ we have available a test statistic T such that

$$P_\theta(a < T < b) = G_{a,b}(\theta) \qquad . \quad . \quad . \quad . \quad . \quad (2.1)$$

is a unimodal function of θ. For testing $H : \theta = \theta_0$ we would adjust a and b such that $G_{a,b}(\theta)$ takes on its maximum at $\theta = \theta_0$ and that the value of this maximum is $1 - \alpha$. If now instead we test $H^* : \theta_1 \leqq \theta \leqq \theta_2$ we adjust a, b so that $G_{a,b}(\theta_1) = G_{a,b}(\theta_2) = 1 - \alpha$. It then follows that the power function $\beta(\theta) = 1 - G_{a,b}(\theta)$ satisfies $\beta(\theta) \leqq \alpha$ for $\theta_1 \leqq \theta \leqq \theta_2$ and $\beta(\theta) > \alpha$ for $\theta < \theta_1$ and $\theta_2 < \theta$.

The following are some examples of this approach :

(i) X has a binomial distribution with probability θ of success. We can then take $T = X$.

(ii) $X_1, \ldots, X_n$ is a sample from a Poisson distribution with $E(X_i) = \theta$. We take $T = \Sigma X_i$.

(iii) $X_1, \ldots, X_n$ is a sample from a normal distribution, mean ξ, variance σ^2. We wish to test $H^* : \sigma_1 \leqq \sigma \leqq \sigma_2$, and take $T = \Sigma(X_i - \bar{X})^2$.

(iv) $X_1, \ldots, X_n$ is a sample from a normal distribution, mean ξ, variance σ^2, and we wish to test $H^* : \delta_0 \leqq \xi/\sigma \leqq \delta_1$. We may then take

$$T = \frac{\bar{X}}{\sqrt{\{\Sigma(X_i - \bar{X})^2\}}}.$$

This is a special case of the general linear hypothesis which we consider later in this section.

Certain two-sample problems immediately reduce to situations of the kind just considered. In particular, let $X_1, \ldots, X_m$; $Y_1, \ldots, Y_n$ be samples from two normal populations with means ξ, η and variances σ^2 and τ^2 respectively. Then we can test

(v) the hypothesis $H^* : a \leqq \sigma^2/\tau^2 \leqq b$ by putting $T = \Sigma(X_i - \bar{X})^2/\Sigma(Y_j - \bar{Y})^2$.

If we assume in addition that $\sigma = \tau$ we can base a test of the hypothesis

(vi) $H^* : \delta_0 \leqq (\eta - \xi)/\sigma \leqq \delta_1$ on the statistic

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{\{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_j - \bar{Y})^2\}}}.$$

The case of two Poisson or binomial populations also can be reduced to a one-parameter problem of the kind already considered.

(vii) Suppose that X, Y are two independent Poisson variables with means λ and μ respectively, and that we wish to test H^*: $a \leqq \lambda/\mu \leqq b$. As is well known and easily checked, the conditional distribution of Y given $X + Y = m$ is the binomial distribution of m trials and success probability $p = \mu/(\lambda + \mu) = 1/(1 + \lambda/\mu)$. Thus we obtain a test of H^* by making on each line $X + Y = m$ a conditional binomial test of the hypothesis $1/(1 + b) \leqq p \leqq 1/(1 + a)$.

In a completely analogous manner we can test $H^* : a \leqq (p_1/q_1)/(p_2/q_2) \leqq b$ where X, Y are independent binomial variables with probabilities p_1, p_2 of success respectively, and where $q_i = 1 - p_i$, since again the conditional distribution of Y given $X + Y = m$ depends on p_1, p_2 only through $\theta = (p_1/q_1)/(p_2/q_2)$.

As a last example consider the general univariate linear hypothesis, which we shall take in the canonical form, according to which $Y_1, \ldots, Y_r$; $Y_{r+1}, \ldots, Y_s$; $Y_{s+1}, \ldots, Y_n$ are independently distributed with common (unknown) variance σ^2 and means

$$E(Y_i) = \eta_i \quad i = 1, \ldots, s; \quad E(Y_i) = 0 \quad i = s + 1, \ldots, n.$$

The usual hypothesis $H : \eta_1 = \ldots = \eta_r = 0$ is tested by rejecting when

$$W = \frac{\sum_{i=1}^{r} Y^2_i/r}{\sum_{i=s+1}^{n} Y^2_i/(n - s)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (2.2)$$

is too large. Here W has a noncentral F-distribution with noncentrality parameter $\lambda = \sum_{i=1}^{r} \eta^2_i/\sigma^2$ and in fact the probability $P(W \geqq C)$ is a strictly increasing function of λ. From this it is obvious how to test $H^* : \Sigma\eta^2_i/\sigma^2 \leqq \lambda_0$. We again reject when W exceeds a constant C, but instead of determining C from the central F-distribution, we now determine it from the equation

$$P(W \geqq C \mid \lambda = \lambda_0) = \alpha.$$

It is interesting to note that this is the uniformly most powerful test for testing H^* among all those based on W (or equivalently that it is the most powerful invariant test for testing H^*; see, for example, Lehmann (1950)). To obtain the test having this property we must consider the probability ratio test based on W for testing $\Sigma\eta^2_i/\sigma^2 = \lambda_0$ against the alternative $\Sigma\eta^2_i/\sigma^2 = \lambda_1$, which rejects when

$$\frac{p_{\lambda_1}(w)}{p_{\lambda_0}(w)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2.3)$$

is too large. (Here p_λ denotes the density of W for parameter value λ). As we shall show below, (2.3) is an increasing function of w and this test is therefore equivalent to rejecting when $w \geqq C$. From this it is then seen that the probability $P\{p_{\lambda_1}(W)/p_{\lambda_0}(W) \geq k\}$ is an increasing function of the true parameter value λ, and the result follows.

It only remains to show that the ratio (2.3) is an increasing function of w. (This fact has been known for some time. The proof which follows is essentially the same as that given by Paul L. Meyer in *An Application of the Invariance Principle to the Student Hypothesis*, Technical Report No. 24, Department of Statistics, Stanford University, California.) We have

$$p_\lambda(w) = e^{-\lambda^2/2} \sum_{k=0}^{\infty} c_k \frac{(\lambda^2/2)^k}{k!} \frac{w^{r/2-1+k}}{(1 + w)^{(r+n-s)/2+k}}$$

where $0 < w$ and the c_k are positive constants. Hence, if we put $z = w/2(1 + w)$, $a_k = c_k \lambda_0^{2k}/k!$ and $b_k = c_k \lambda_1^{2k}/k!$, we have

$$\frac{p_{\lambda_1}(w)}{p_{\lambda_0}(w)} = \frac{\Sigma b_k z^k}{\Sigma a_k z^k} = f(z), \text{ say.}$$

It is easy to show that the derivative of $f(z)$ is given by

$$(\Sigma a_k z^k)^2 f'(z) = \sum_{k<n} (n-k)(a_k b_n - a_n b_k) z^{k+n-1},$$

which is positive since $b_k/a_k < b_n/a_n$ for $k < n$. It follows that $p_{\lambda_1}(w)/p_{\lambda_0}(w)$ is an increasing function of z and hence of w.

3. *Student's Problem*

We next consider the modification of the classical Student testing problem in which absolute units are used, rather than σ-units as was done in example (iv) above. Let X have a normal distribution with $E(X) = \xi$, $V(X) = \sigma^2$, and let S, independent of X, be such that S/σ has the chi distribution with m degrees of freedom. The problem of testing the hypothesis that ξ equals a specified value ξ_0, with σ playing the role of a nuisance parameter, is often formulated. For example, it arises when testing whether two normal populations, of the same but unknown variance, have the same mean. In line with the arguments advanced above, we feel that in many applications it would be more realistic to formulate as our hypothesis the assertion that ξ does not differ from ξ_0 by more than a specified quantity chosen in the light of the material problem. For instance, instead of testing that two normal populations have the same mean, we would test that their means do not differ by more than an amount specified to represent the smallest difference of practical interest.

One possible test of the modified hypothesis that $\xi_1 \leqq \xi \leqq \xi_2$, which uses only existing tables, consists in performing separate one-tailed t-tests of the two one-sided hypotheses:

$$H : \xi \leqq \xi_2 \text{ against alternative } \xi > \xi_2$$

and

$$H : \xi \geqq \xi_1 \text{ against alternative } \xi < \xi_1.$$

We then reject if either of these separate tests rejects. The size of this composite t-test is the sum of their separate sizes, as is seen by letting σ tend to ∞. (By the size of a test we mean as usual the upper bound of its probability of first kind error.) Thus, we would obtain a test of size α if each of the one-sided t-tests was carried out at level $\alpha/2$. The main objection to the composite t-test solution of our problem lies in the fact that, when σ is small, the test has the power only of the t-test with level $\alpha/2$, although we are paying for a test of size α. The obvious remedy is to seek an unbiased test. Such a test will of necessity have power identically α when $\xi = \xi_1$ or $\xi = \xi_2$; that is, it will be "similar on the boundary". For simplicity we shall assume throughout that $\alpha \leqq \frac{1}{2}$, and make a scale change to insure $\xi_1 = -1$, $\xi_2 = 1$.

In the present paragraph we consider the right boundary, that is, we assume $\xi = 1$. We may specify the location of the sample point (x, s) by the polar coordinates (r, θ) defined by $r^2 = (x-1)^2 + s^2$, $\sin\theta = s/r$. It is well known that the corresponding random variables R, Θ are independent and that $P[\Theta \leqq \theta] = \frac{1}{2} I_{\sin^2\theta}(\frac{1}{2}m, \frac{1}{2})$ for $0 \leqq \theta \leqq \pi/2$, where I is the Incomplete Beta function. The distribution of Θ is further symmetric about $\pi/2$. It is known (Lehmann and Scheffé, 1950) that $(X-1)^2 + S^2 = R^2$ is a complete sufficient statistic for σ, so that the only tests of size α which are similar for $\xi = 1$ are those whose critical regions intersect each semicircle $R = r$ in a set of points whose conditional probability is α.

As the analogous argument goes through for $\xi = -1$, we have a double condition on the test region. We also impose the intuitively reasonable requirements that the test region be symmetric about the s-axis, and that for given values of R we want to reject for extreme values of θ. The problem now formulated is to construct a test region satisfying these conditions. We shall carry through this construction in detail in the simple case $m = 1$, where an explicit solution can be given.

When $m = 1$, Θ has the uniform distribution over $(0, \pi)$, so that we must take from each semicircle an arc of size $\pi\alpha$. The construction is made inductively on increasing values of r. From semicircles of radius $r \leqq 2 = r_1$, we take arcs of size $\pi\alpha$ adjacent to the positive x-axis (see Fig. 1). This gives the segment P_0P_1 as part of the boundary of the rejection region. By the symmetry requirement, we must also reject at points below the segment Q_0Q_1, obtained by reflecting P_0P_1

about the s-axis. Let r_2 denote the distance from P_0 to Q_1, and next consider the semicircles with radii $r_1 \leqq r \leqq r_2$. From these we have already taken as rejection points the arcs below Q_0Q_1, which are the same size as the arcs below the segment R_0R_1 obtained by translating P_0P_1 two units to the right. To bring the rejection arcs up to size $\pi\alpha$, we cut off the additional rejection arcs below P_1P_2. The notable fact, obvious from the symmetry of the picture, is that the boundary portion P_1P_2 is a horizontal straight line segment. The construction proceeds inductively, producing a boundary of the rejection region consisting of straight line segments of length 2, alternately horizontal and of slope $\pi\alpha$, together of course with its reflection about the s-axis.

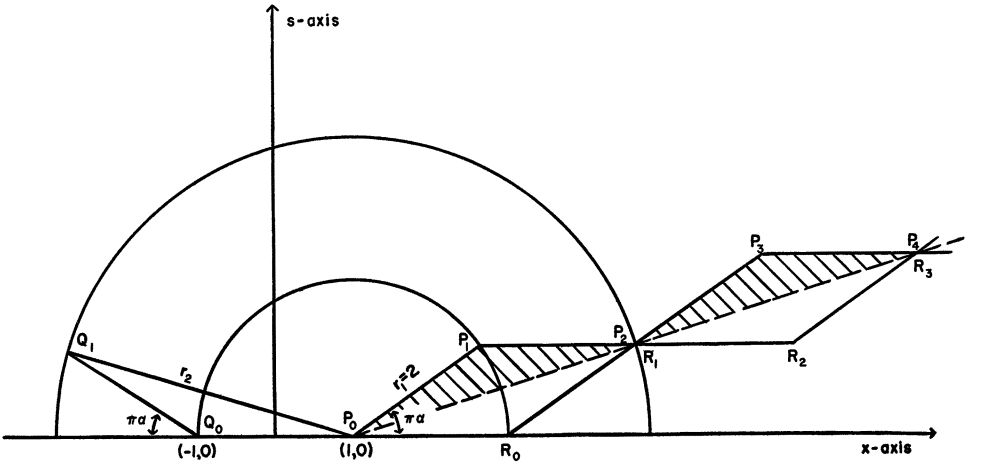

FIG. 1.—Construction of the modified Student test when $m = 1$.

The test we have obtained is by its construction similar on the boundary $\xi = \pm 1$; to verify that it is in addition unbiased, we need only observe that for each given value of s, we accept for x in an interval centered at 0. Therefore, the conditional probability of rejection, given s, is a monotonely increasing function of $|\xi|$, whatever s may be. The same must then hold for the power function as a function of $|\xi|$ for any given value of σ.

For comparison, we show as the dashed line of the Figure the boundary of the composite t-test of size α. The shaded region represents the additional rejection points which the new test affords without increase in size. It is these points which provide our test with its greater power for finite σ.

So far we have been considering only the case $m = 1$, where the problem is simple and we can give an explicit solution. When we turn to the general case things are not so easy. The inductive construction may still be employed, alternately translating boundary segments to the right by two units and then generating new segments by the requirement that a conditional probability α be taken on each arc. This leads to the relation

$$\left.\begin{aligned} (r')^2 &= 4r\cos\theta + r^2 + 4 \\ I(\sin^2\theta') &= 2\alpha - I\left[\frac{r^2\sin^2\theta}{r'^2}\right] \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (3.1)$$

where (r, θ) and (r', θ') are two boundary points, and we write $I(x)$ for $I_x(\frac{1}{2}m, \frac{1}{2})$. These equations, together with the initial values $\theta = \arcsin\sqrt{\{I^{-1}(2\alpha)\}}$ for $0 \leqq r \leqq 2$, permit the effective computation of a boundary curve C. It appears from the numerical computations underlying the chart that C has positive slope everywhere, but we have not found a proof of this. From the assumption that C has positive slope it would follow (as for $m = 1$) that the test region with boundary C is unbiased, and at least approximate unbiasedness is in any case guaranteed by the numerical results.

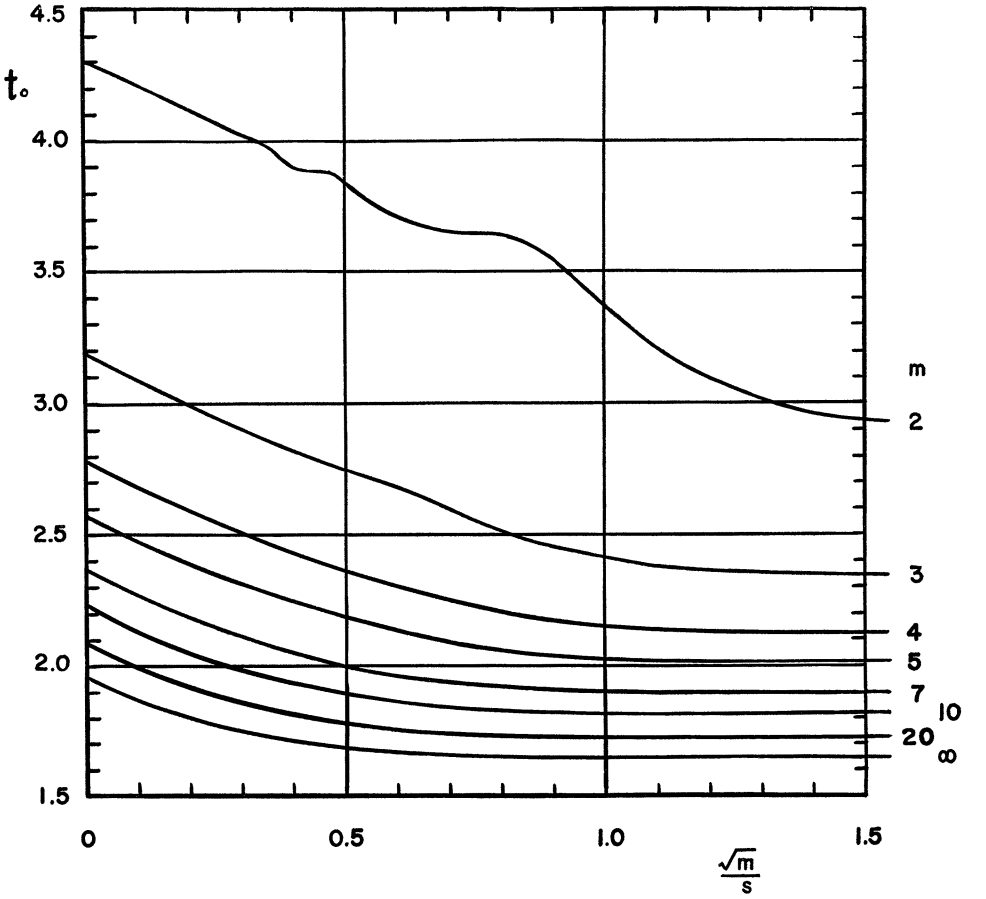

FIG. 2.—5 per cent. critical values for the modified Student test.

Fig. 2 shows critical values of the 5 per cent. test for various degrees of freedom m. Suppose X is normal, $E(X) = \xi$, $V(X) = \sigma^2$, S^2/σ^2 has the chi-square distribution of m degrees of freedom, X and S are independent, and we wish to test $|\xi| \leqq 1$. If $|X| \leqq 1$, accept. Otherwise, let t be $(X - 1)\sqrt{m}/S$ if $X > 1$, or $(-1 - X)\sqrt{m}/S$ if $X < -1$. Read the critical value t_0 from Fig. 2 corresponding to the value of m and of $\sqrt{m}/S$. Reject if $t > t_0$.

To illustrate the test, suppose $Y_1, \ldots, Y_k$ is a sample from the normal population of expectation η and variance τ^2, while $Z_1, \ldots, Z_n$ is a sample from the normal population of expectation ζ and variance τ^2. We wish to test the hypothesis that the absolute difference $|\eta - \zeta|$ of the population means does not exceed Δ. If we let $k\bar{Y} = \Sigma Y_i$, $n\bar{Z} = \Sigma Z_i$, $X = (\bar{Y} - \bar{Z})/\Delta$, $\xi = (\eta - \zeta)/\Delta$, $\sigma^2 = (k + n)\tau^2/kn\Delta^2$, $S^2 = (k + n)\{\Sigma(Y_i - \bar{Y})^2 + \Sigma(Z_i - \bar{Z})^2\}/kn\,\Delta^2$, and $m = k + n - 2$, then the assumptions of the test are satisfied.

In the computations for the chart good use was made of the approximate value $t_{\alpha/2} - \sqrt{m}/S$ for $t_0(\sqrt{m}/S)$. This approximation rests on the fact that as $r \to \infty$, the distance of the point (r, θ) in the first quadrant from the line $y = x\sqrt{m}/t_{\alpha/2}$ tends to 0. The recursion formulae (3.1) were used until the error of this approximation was less than 0.1, and thereafter the error was interpolated quadratically.

Inspection of Fig. 2 shows that the test has an intuitively reasonable form. When $s/\sqrt{m}$ is very small, it is plausible that σ is small, and the two sides do not interfere with each other. We may then safely use separate one-tailed t-tests each of level α; this corresponds to the flat

portions of the curves. When $s/\sqrt{m}$ is large, our hypothesis is not essentially different from the classical Student hypothesis, and the two-tailed t-test of size α is appropriate; this is given in the left margin of the chart. When m is very large, we may safely take the consistent estimate $s/\sqrt{m}$ as if it were σ, and use the obvious normal test available when σ is known; this is the lowest curve of the chart. For any given finite value of m, the critical value of t changes in a more or less regular way, as $s/\sqrt{m}$ is increased, from one extreme to the other. The waves, visible for $m = 2$ and to a lesser extent for $m = 3$, are to be expected in the light of the situation for $m = 1$.

We conclude by remarking that our test also provides confidence intervals in the usual way. For instance, it will give an upper confidence limit for the absolute amount by which the mean of a normal population differs from a given quantity, or for the absolute difference of the means of two normal populations of the same unknown variance.

4. *Testing for Goodness of Fit*

It is in connection with testing for goodness of fit that the existence of the problem under consideration has been pointed out by Berkson and others. We are concerned with a multinomial distribution with, say, r classes and with the hypothesis that the point $p = (p_1, \ldots, p_r)$, where p_i is the probability of the i^{th} class, lies on a specified surface $\mathscr{S}$. The simplest case is that in which the hypothesis specifies the p's completely so that $\mathscr{S}$ consists of a single point. Let $\Delta(p, p')$ be a distance function in the sense that $\Delta(p, p') \geqq 0$ for all p, p' and $\Delta(p, p') = 0$ if and only if $p = p'$. A typical example is ordinary Euclidean distance or more generally $\Delta(p, p') = \Sigma\, w_i(p'_i - p_i)^2$ where the weights w_i may be functions of p, p'. This includes in particular the usual χ^2-measure of distance. Let $d(p)$ denote the smallest distance of a point p from $\mathscr{S}$ so that

$$d(p) = \inf_{p' \varepsilon S} \Delta(p, p'). \qquad (4.1)$$

We shall then test instead of H the hypothesis H^* that the true point p lies within a given distance of $\mathscr{S}$, that is, the hypothesis

$$H^* : d(p) \leqq c. \qquad (4.2)$$

In order to obtain a test of H^* we shall assume that $\mathscr{S}$ and Δ are such that the function $d(p)$ possesses continuous first and second partial derivatives. We shall not discuss conditions on $\mathscr{S}$ and Δ which would insure this since such conditions do not appear to be particularly simple and since in applications it will always be necessary to obtain the function $d(p)$, and the regularity assumption is then easily checked directly. Of course in the particular case that $\mathscr{S}$ consists of a single point, say p^0, we simply have $d(p) = \Delta(p, p^0)$ and the condition on d immediately reduces to one on Δ.

Under these assumptions we propose the following procedure for testing H^*. Let $q_i\,(i = 1, \ldots, r)$ denote the relative frequency in the i^{th} class and let $q = (q_1, \ldots, q_r)$. Then if $d(q) \leqq c$ accept H^*. If $d(q) > c$ test the hypothesis

$$H' : d(p) = c \qquad (4.3)$$

by means of minimum χ^2 (or some asymptotically equivalent test) in the usual manner. For example, we may reject when the minimum (modified) χ^2

$$n \sum_{i=1}^{r} \frac{(q_i - \hat{p}_i)^2}{q_i} > K \qquad (4.4)$$

where n is the sample size, where the $\hat{p}_i$ are BAN estimates (Neyman, 1949) of the p_i subject to H', and where K is determined from the distribution of χ_1^2, that is, χ^2 with 1 degree of freedom (See Neyman (1949), p. 267. The one degree of freedom arises because the only restriction on $\hat{p}_i$ is $d(\hat{p}_i) = 0$). However, since we are subjecting the rejection region to the additional restriction $d(q) > c$, the cut-off point K is determined so that

$$P(\chi_1^2 > K) = 2\alpha \qquad (4.5)$$

where α is the desired level of significance. This differs from the usual procedure in which the right-hand side of (4.5) is taken to be α.

We shall now prove that the power function $\beta_n(p)$ of this test has the following property:

$$\lim_{n\to\infty} \beta_n(p) = \begin{matrix}0\\ \alpha\\ 1\end{matrix} \text{ as } d(p) \begin{matrix}<\\ =\\ >\end{matrix} c. \qquad (4.6)$$

This result follows easily from the theory of minimum-χ^2 tests. (For details see Cramér (1946) and Neyman (1949).) In particular, if $d(p) < c$, it follows from the continuity of d and the fact that $q \to p$ in probability that $d(q) < c$ with probability tending to one, so that $\beta_n(p) \to 0$. Suppose next that $d(p) > c$. Then it follows similarly that $d(q) > c$ with probability tending to one. Also it follows from the consistency of the χ^2 test that the probability of (4.4) tends to one. Hence $\beta_n(p)$ is the probability of the simultaneous occurrence of two events, say A_n and B_n, for which $\lim P(A_n) = \lim P(B_n) = 1$. But this implies that $\lim \beta_n(p) = \lim P(A_n \cap B_n) = 1$.

Let us finally consider the case in which the true probability point, say p^0, lies on the surface $\mathscr{S}'$: $d(p) = c$. Then it is known that $n\Sigma\{(q_i - \hat{p}_i)^2/q_i\}$ has the same limiting distribution as $[\Sigma\gamma_i\sqrt{n}\,(q_i - p_i^0)]^2$ where the γ's are a set of coefficients such that

(i) The equation of the tangent plane of the surface $\mathscr{S}'$ at p^0 is $\Sigma\gamma_i\,(p_i - p_i^0) = 0$, and

(ii) $\Sigma p_i^0\,\gamma_i = 0$, $\Sigma p_i^0\,\gamma_i^2 = 1$.

Therefore, $d(q) = d(p^0) + \Sigma\,\gamma_i(q_i - p_i^0) + 0_p\,(1/n)$ and hence, since $d(p^0) = c$,

$$d(q) - c \gtreqless 0 \text{ implies } \sqrt{n}\,\Sigma\gamma_i\,(q_i - p_i^0) + 0_p\,(1/\sqrt{n}) \gtrless 0.$$

Therefore

$$\beta_n(p^0) = P\{n\Sigma\{(q_i - \hat{p}_i)^2/q_i\} > K,\ d(q) > c\}$$

becomes in the limit equal to

$$P\{[\Sigma\,\gamma_i\sqrt{n}(q_i - p_i^0)]^2 > K,\ \Sigma\gamma_i\sqrt{n}(q_i - p_i^0) > 0\}.$$

But if we let $Y = \Sigma\gamma_i\sqrt{n}(q_i - p_i^0)$, then Y has a limiting normal distribution with zero mean and unit variance, and the result follows.

The test is of course very simple to carry out since in the minimization of χ^2 we have only the single condition $d(p) = c$. If, following Neyman, we replace this by the asymptotically equivalent condition $d(q) + \Sigma a_i(q)\,(p_i - q_i) = c$ where $a_i = a_i(q) = \dfrac{\partial}{\partial q_i}\,d(q)$, we get the solution

$$q_i - \hat{p}_i = (a_i - \bar{a})\,q_i\,[d(q) - c]/\sigma_a^2 \qquad (4.7)$$

where $\bar{a} = \Sigma a_i q_i$, $\sigma_a^2 = \Sigma q_i(a_i - \bar{a})^2$.

As an example, suppose that we wish to test that all r cell probabilities are equal, and that we set $\Delta(p, p') = \Sigma(p_i' - p_i)^2$. Then the modified hypothesis becomes H^*: $\sum_{i=1}^{r} (p_i - 1/r)^2 \leqq c$, and we must determine the $\hat{p}_i$ subject to the condition $\Sigma(p_i - 1/r)^2 = c$. We then have $a_i = 2(q_i - 1/r)$ and can compute $q_i - \hat{p}_i$ from (4.7).

References

BERKSON, J. (1938), *J. Amer. Statist. Ass.*, **33**, 526–542.
CRAMÉR, H. (1946), *Mathematical Methods of Statistics*. Princeton University Press.
LEHMANN, E. L. (1950), *Ann. Math. Statist.*, **21**, 1–26.
—— & SCHEFFÉ, H. (1950), *Sankhyā*, **10**, 305–340.
NEYMAN, J. (1949), *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley and Los Angeles: University of Calif. Press.