Sankhyā : The Indian Journal of Statistics -
1979, Volume 41, Series A, Pts. 1 and 2, pp. 1–21.

$C(\alpha)$ TESTS AND THEIR USE

By JERZY NEYMAN
University of California, Berkeley

SUMMARY. This review paper was presented after the Inauguration of the New Delhi Campus of the Indian Statistical Institute.* $C(\alpha)$ tests were developed to deal with societal problems that invariably involve non-standard conditions : nonstandard composite hypotheses, nonstandard alternatives and distributions. However, if the societal problem is really important, one can presume enough observations for the use of central limit theorem.

1. Introduction

Whoever has participated in non-trivial research in any domain of science involving statistical problems must have encountered the difficulty that none of the statistical procedures found in the books fits exactly the practical situation. In particular, this applies to the uniformly most powerful tests. Most usually, the hypotheses that one encounters are composite and refer to non-standard distributions. Next, the alternative hypotheses which it is important to guard against are also non-standard. Confronted with this situation, the statistician is likely to recur to non-parametric tests. Unfortunately, while maintaining, at least approximately, the desired level of significance, it is only rarely that these tests have proved properties of optimality with respect to interesting classes of alternatives.

Considerations of the above kind caused me to seek tests that have a compromise but a clear property of optimality and that are relatively easy to deduce. The ease in deducing these tests must be paid for and the price I paid is composed of two items. One is my definition of optimality, which is only a "local" and an "asymptotic" optimality. The other is that the tests deduced are optimal only within a certain class of tests, labeled $C(\alpha)$ tests. However, Lucien Le Cam proved that, quite frequently, this local asymptotic optimality is general.

With reference to the many questions I heard, I must explain the origin of the symbol $C(\alpha)$. Here α refers to the possibility of prescribing an arbitrarily chosen level of significance α. The letter C refers to Harald Cramér, whose work I greatly admire. Particularly I value greatly Cramér's book, the *Mathematical Methods of Statistics*, which, first published in 1946, had a very

*December, 1974.

strong and very beneficial influence on our discipline ; it made mathematical statistics considerably more mathematical than it was before. In 1959 a jubilee volume was published in honor of Cramér. The basic theory of $C(\alpha)$ tests is published in that volume. Originally, I thought of calling these tests the Cramér tests. This idea was abandoned because the use of Cramér's name could have been interpreted as loading Cramér with the responsibility for all the deficiencies of these particular tests. In the following, some of these deficiencies will be mentioned. My intention was, and is, merely to honor Cramér.

2. General idea of $C(\alpha)$ tests

Consider a sequence of independent random variables $\{X_n(\xi, \theta)\}$, possibly vectors, all following the same distribution, with its density $p(x \mid \xi, \theta)$ depending upon two parameters, a scalar ξ and possibly a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$ of some $s \geqslant 1$ components. Without loss of generality it will be assumed that ξ can have values in an interval containing zero, either as one of the boundaries or as an interior point. This parameter ξ will be the test parameter. Specifically, we shall be concerned with the hypothesis H that asserts $\xi = 0$. The alternatives may specify $\xi < 0$ or $\xi > 0$ or, more generally, $\xi \neq 0$. The parameter θ is a nuisance parameter.

The vector parameter θ will be assumed to have an unknown value within some open set Θ. The following discussion presupposes certain properties of regularity of $p(x \mid \xi, \theta)$, including the possibility of two differentiations under the sign of integral taken over all the sample space of X, say W.

In order to define a test of class $C(\alpha)$, consider an arbitrary measurable function $f(x)$, defined for all $x \in W$, and that, for $\xi = 0$, the random variable $f[X(0, \theta)]$ has a finite variance $\sigma^2(\theta)$. Let $f_1(\theta)$ be the expectation of $f[X(0, \theta)]$. Then, by the central limit theorem, the function

$$Z_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f[X_i(0, \theta)] - f_1(\theta)}{\sigma(\theta)} \qquad \ldots \quad (1)$$

will tend to be normally distributed $N(0, 1)$ as $n \to \infty$. Hence, if one deals with an important problem involving the test of the hypothesis H, the problem so important as to justify a large number n of observations, the statistic $Z_n(\theta)$ can be used to test H. For example, the test procedure may consist in the rule of rejecting H when $Z_n > \nu(\alpha)$ or when $Z_n < -\nu(\alpha)$, etc. where

$$\frac{1}{\sqrt{2\pi}} \int_{\nu(\alpha)}^{\infty} e^{-u^2/2} du = \alpha \qquad \ldots \quad (2)$$

More generally, in order to test H, we may select a set $S(\alpha)$ on the real line (for the mild restriction, see Neyman, 1959) such that the integral over $S(\alpha)$ of the normal density functions $N(0, 1)$ is equal to the chosen α, and make a rule to reject H whenever the observations yield $Z_n(\theta) \epsilon S(\alpha)$.

Obviously, with a substantial n, such procedure would guarantee that, if H is true, it will be rejected only with the relative frequency nearly equal to the chosen α. The trouble is that the criterion $Z_n(\theta)$ cannot be calculated from the observations alone, because generally it will depend on the value of the nuisance parameter θ, the value of which is unknown. Thus the question arises about the asymptotic distribution of $Z_n(\hat{\theta})$ which is the result of substituting into (1) an estimate $\hat{\theta}$ for θ. For the theory to be useful, the estimate $\hat{\theta}$ must be allowed to be "moderately" good. Any requirement of being unbiased and/or consistent "in the large," that is, for all possible values of ξ, might lead to prohibitive difficulties.

My own choice of limitation on $\hat{\theta}$ was that it be "locally root n consistent." This means that, for each possible ξ, the variable

$$|\hat{\theta}-\theta-A\xi|\sqrt{n} \qquad \dots .\ (3)$$

must be bounded in probability. Here A means a constant and $A\xi$ stands for bias in the estimate $\hat{\theta}$. If $A = 0$, then there is no bias and the estimator $\hat{\theta}$ is labelled "consistent in the large." Otherwise, it is only "locally" consistent.

For the case $\xi = 0$, one of the basic theorems in Neyman (1959) indicates that, for $Z_n(\hat{\theta})$ to have the same asymptotic distribution as $Z_n(\theta)$, that is normal $N(0, 1)$, this irrespective of the value of the nuisance parameter θ, it is necessary and sufficient that the function f be orthogonal to all the logarithmic derivatives, say

$$\phi_j(x, \theta) = \frac{\partial \log p}{\partial \theta_j}\Big|_{\xi=0}, \quad j = 1, 2, \dots, s \qquad \dots \quad (4)$$

(Further on, this result is referred to as Theorem 1.)

Starting with an arbitrary function f, it is easy to replace it by one orthogonal to the ϕ_j, namely, say

$$g(x, \theta) = f(x, \theta)-\sum_{j=1}^{s} a_j\phi_j(x, \theta) \qquad \dots \quad (5)$$

where $f(x, \theta) = f(x) - f_1(x)$ and where the coefficients a_j are solutions of a system of linear equations and represent partial regression coefficients of $f[X(0, \theta)] - f_1(\theta)$ on $\phi_1, \phi_2, \ldots, \phi_s$. All the difficulty connected with this, usually not a great difficulty, consists in using the given density $p(x \mid 0, \theta)$ to compute the variances and covariances of f and the ϕ_j. Once this is done, one computes the variance of g. Dividing (5) by the square root of this variance, we obtain the "normed" form of g, say $g^*(x, \theta)$. This is now inserted into formula (5) to obtain the criterion

$$Z_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g^*[X_i(\xi, \theta), \theta] \qquad \ldots \quad (6)$$

which is now certain to have the normal $N(0, 1)$ asymptotic distribution when $\xi = 0$.

Now we are in the position to define the class of $C(\alpha)$ tests.

Definition : The rule of rejecting the hypothesis H that $\xi = 0$ whenever the computed value $Z(\theta)$ of formula (6) *falls within the set $S(\alpha)$ is called a $C(\alpha)$ test.*

Thus, each $C(\alpha)$ test is determined by two (more or less) arbitrary choices : first we select the function f which eventually determines g^*, and second, we select the rejection set $S(\alpha)$.

The next step is to deduce a formula yielding an approximate value of the power

$$\beta(\xi) = P\{Z_n(\theta) \in S(\alpha) \mid \xi\}. \qquad \ldots \quad (7)$$

In order to write this formula we must consider the derivative

$$\phi_\xi = \frac{\partial \log p}{\partial \xi}\Big|_{\xi=0} \qquad \ldots \quad (8)$$

and also the expression, say

$$\psi(x, \theta) = \phi_\xi - \sum_{j=1}^{s} b_j \phi_j \qquad \ldots \quad (9)$$

where, as in formula (5), the coefficients b_j are partial regression coefficients of ϕ_ξ on the ϕ_j. Finally, let σ_ψ^2 stand for the variance of ψ, and ρ for the correlation coefficient of g^* and ψ. With this notation, the asymptotic formula for the power function $\beta(\xi)$ is

$$\beta(\xi) \sim \frac{1}{\sqrt{2\pi}} \int_{S(\alpha)} e^{-(u-\xi\sqrt{n}\rho\sigma_\psi)^2/2} \quad du. \qquad \dots \quad (10)$$

The particular passage to the limit which determines this formula is to let $n \to \infty$ and, simultaneously, $\xi \to 0$.

Formula (10) allows to select the optimal $C(\alpha)$ test. If ξ and ρ have the same sign then the optimal rejection region $S(\alpha)$ extends from $\nu(\alpha)$ to infinity. On the other hand, if the product $\xi\rho$ is negative, the optimal rejection region extends from $-\infty$ to $-\nu(\alpha)$. When this is noticed, the selection of the optimal f is immediate. Obviously, the asymptotic power (10) depends upon f only through the correlation coefficient ρ whose extreme values are ± 1. It follows that the greatest power for any given ξ is attained if $g^* = \psi$, which represents the solution of the problem of the optimal $C(\alpha)$ test : the optimal test criterion is, say

$$Z_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, \theta)/\sigma_\psi. \qquad \dots \quad (11)$$

Against the alternatives $\xi > 0$ the hypothesis H is rejected when (11) exceeds $\nu(\alpha)$. Against the alternatives $\xi < 0$, this hypothesis is rejected when (11) is less than $-\nu(\alpha)$. Finally, if the alternatives are $\xi \neq 0$, the rejection occurs when the absolute value of Z^* exceeds $\nu(\alpha/2)$. In each case, the asymptotic power function of the optimal $C(\alpha)$ test is obtained from (10) by putting $\rho = 1$ and by specifying the requisite $S(\alpha)$.

With reference to the above sketch the reader will notice a brief remark that the tentative test function f must satisfy certain conditions of regularity. The same applies to the probability density (or frequency function) $p(x \mid \xi, \theta)$. All these conditions would take more time than is available at the present meeting and the audience is referred to the original publication (Neyman, 1959). Briefly, the restrictions imposed on the probability density $p(x \mid \xi, \theta)$ are similar to those used by Cramér in his book already quoted, under which he proves the consistency of the maximum likelihood estimates. In "ordinary" cases encountered in applied work these conditions are satisfied.

3. AN IMPORTANT PARTICULAR CASE

Formula (11), giving the optimal $C(\alpha)$ test criterion, is basic. As mentioned before, the use of this formula involves the evaluation of the logarithmic derivatives ϕ_ξ and ϕ_j, for $j = 1, 2, \ldots, s$, the calculation of the variance-covariance matrix of these $s+1$ variables and then the finding of the coefficients b_j in formula (9) that minimize the variance of ψ. Granting the conditions of regularity, including the independence of all the parameters, the above procedure is always applicable — even though, occasionally, it is somewhat laborious. In a more recent paper in 1965, authored jointly with Elizabeth L. Scott, we isolated a particular case in which the labor in calculating ψ is greatly reduced. In addition, this particular case presents an independent interest, especially in problems of experimentation.

Treating a case somewhat less general than in Neyman and Scott, 1965, we consider a sequence of units of observation $\{U_n\}$, each characterized by some measurements which we denote by a single letter X_n referring to U_n. It is assumed that the typical X is a random variable with known density $p(x|\theta)$ where θ is a nuisance parameter, usually a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_s) \epsilon \Theta$. A randomized experiment is being performed on units U. For each of them a figurative coin is tossed with known probability π of falling heads. If the coin falls heads, then the "randomizing variable" T is assigned the value $t = 1$ and the corresponding unit U is subjected to some experimental treatment. If the coin falls tails, then T is given the value $t = 0$ and the unit U is assigned to go without treatment and to serve as a control. (Here, the letter T connotes "treatment.")

In consequence of the above, to each unit of observation there correspond two observable random variables, the randomizing variable T and the experimental variable X, possibly a vector. If $T = 0$, then the distribution of X is the given function $p(x|\theta)$. Now we must specify our assumptions regarding the distribution of X if $T = 1$, that is, if the given unit of observation is subject to the treatment studied. Not always, but frequently, the statistician concerned is prepared to admit that for treated units U, the variable X follows the same type of distribution as for the untreated units, but with an altered value of the nuisance parameter θ. To use a "bookish" example, the distribution of X for control units U may be assumed normal with an unspecified mean μ and an unspecified variance σ^2. For treated units one frequently assumes that the distribution of X continues to be normal, but possibly with a different mean, say $\mu(\xi) = \mu + \xi$ and, possibly with a modified variance, say $\sigma^2(\xi)$, where ξ stands for a conventional measure of the effect of treatment.

Returning to the general discussion, we use the letter ξ to denote a conventional measure of the effect of the treatment studied. We assume that the statistician wishes to have a test especially directed toward the discovery of the effect of treatment ξ specified in the following manner. Whatever be the nuisance parameter point $\theta \epsilon \Theta$ corresponding to control units U, the use of the treatment shifts this parameter point from the original position θ to another position $\vartheta(\xi, \theta) \epsilon \Theta$, where the function ϑ is specified by the statistician's expectations,- undoubtedly derived from the discussions with the experimenter. If the treatment studied has no effect, so that $\xi = 0$, then the function $\vartheta(0, \theta)$ assumes the value of the nuisance parameter corresponding to the untreated units.

Certain discussions we had with some applied statisticians lead me to emphasize that the use of any particular function $\vartheta(\xi, \theta)$ does not imply that the statistician and/or the experimenter *believe* that the possible effects of treatment *must* be accurately described by this function. Contrary to this, the choice of a particular ϑ corresponds to the *desire* of the statistician to guard against the effects of the treatment specified by the chosen ϑ. In many cases, the specification of an appropriate $\vartheta(\xi, \theta)$ requires some soul-searching which may appear troublesome. There is no law requiring the statistician (or the experimenter) to go through this trouble. However, if his interest in the study is compelling, then the following results might be useful.

The only specific requirements from the function $\vartheta(\xi, \theta)$ are that at $\xi = 0$ we have $\vartheta(0, \theta) = \theta$ and that, for each component θ_j of the vector θ, we have

$$\frac{\partial \vartheta_j(\vartheta, \theta)}{\partial_\xi} \Big|_{\xi=0} = \vartheta_j' \qquad \dots \quad (12)$$

a known function of θ.

With reference to the above "bookish" example with the normal distribution of X the statistician may be particularly interested in the effect of treatment which multiplies the original mean $\mu > 0$ and the original standard deviation σ by the same positive factor, so that the coefficient of variation of X is not altered. If so, he may specify

$$\mu(\xi, \mu) = \mu \exp\{\xi\}, \qquad \dots \quad (13)$$

$$\sigma(\xi, \sigma) = \sigma \exp\{\xi\}, \qquad \dots \quad (14)$$

in which case the partial derivatives at zero as in (12) are $\mu'(\mu) = \mu$ and $\sigma'(\sigma) = \sigma$. On the other hand, if the statistician's $\mu(\xi, \mu) = \mu+\xi$ and $\sigma(\xi) = \sigma$, then $\mu' = 1$ and $\sigma' = 0$, etc.

With the above notation, the frequency function of the two random variables T and X can be written as, say

$$\pi^t(1-\pi)^{1-t}p[x\,|\,\vartheta(t\xi, \theta)] \qquad \ldots \quad (15)$$

for $t = 0$ or 1. Then it was found in Neyman and Scott (1965) that no effort is needed to calculate the coefficients b in formula (9) and that formula (11) reduces to

$$Z_n^* = \frac{\sum_{i=1}^{n} (t_i-\pi) \sum_{j=1}^{s} \vartheta_j' \phi_j(x_i, \hat{\theta})}{\left\{n\pi(1-\pi) \operatorname{var}\left(\sum_{j=1}^{s} \vartheta_j'\phi_j\right)\right\}^{\frac{1}{2}}}. \qquad \ldots \quad (16)$$

Here t_i equals unity or zero depending on whether the i-th unit of observation is assigned to be treated or not. Frequently, the '"randomizing probability" π is one-half, which ensures the greatest precision of the experiment. In this case formula (16) simplifies somewhat.

This completes the sketch of the theory of $C(\alpha)$ tests possible to give in this paper. A number of generalizations have been published, frequently in joint papers, by several authors including Bartoo, Buhler, Davies and Puri (1966, 1967). Some time ago Professor Piotr Mikulski, present at this conference, investigated the conditions under which the optimal $C(\alpha)$ test coincides with the uniformly most powerful test when such exists, but I do not remember seeing his results published.

The defects of $C(\alpha)$ tests are that their "optimality" is only "local" and that it is only asymptotic. In particular, the all-important formula (10), supposed to give an approximation to the power function $\beta(\xi)$, is based on the passage to the limit as $n \to \infty$ while ξ tends to zero in such a way that the product $\xi\sqrt{n}$ tends to some fixed limit different from zero. Of course, in any actual case the test parameter ξ has some fixed value. If $\xi = 0$, then through the use of the central limit theorem and of Theorem 1 above, the theory guarantees the approximate maintenance of the chosen level of significance α. If ξ is not zero, it may be quite large or small and the only thing that is under our control is the number of observations n, which can be made large. Until further investigations are made (Buhler and Puri, 1966), the information about the performance of optimal $C(\alpha)$ tests must come from the Monte Carlo simulation experiments. In many cases studied thus far this performance appears quite good and this justifies the present paper. However, as described below, this is not a general rule. Certain categories of cases were identified in which the actual power of the optimal $C(\alpha)$ test is not comparable to predictions of formula (10). Here, then, there is a challenging new field for further theoretical studies. Some such research is already in progress.

Most of the remaining material in this paper is concerned with examples. Some of these examples illustrate the use of the two basic formulas (11) and (16), and are somewhat "bookish." The other examples illustrate some "non-bookish" problems for which the $C(\alpha)$ technique provides an easily attainable solution.

4. Two bookish examples

Example 1 : *Test of the hypothesis of independence of two variables X and Y, known to have a joint bivariate Poisson distribution.* Even though this problem originated from a study related to ecology, it is really bookish. It is discussed in Neyman (1959). The observable variables are n vectors $\{X_i, Y_i\}$. The joint probability generating function of X and Y can be written as

$$G(u, v) = \exp\{-\xi(1-uv)-\theta_1(1-u)-\theta_2(1-v)\}. \qquad \dots \quad (17)$$

If $\xi = 0$, the two variables X and Y are independent Poisson variables with expectations θ_1 and θ_2. In all cases the marginal distributions of X and Y are Poisson, with expectations $\theta_1+\xi$ and $\theta_2+\xi$, respectively. It follows that the means, say $\bar{X}_n$ and $\bar{Y}_n$, of n observations on X and Y are locally root n consistent estimates of θ_1 and θ_2, respectively : the bias in each is equal to ξ. For any non-negative integers x and y, we have

$$P\{X = x,\ Y = y \,|\, \theta_1, \theta_2, \xi\} = e^{-\xi-\theta_1-\theta_2} \sum_{k=0}^{\min(x,y)} \frac{\xi^k \theta_1^{x-k} \theta_2^{y-k}}{k!(x-k)!(y-k)!}. \qquad \dots \quad (18)$$

The first thing to consider is whether, in the present case, one can use the formula (16) which is simpler than (11). The answer is in the negative and the development of the optimal $C(\alpha)$ test requires the calculation of all three logarithmic derivatives of (18), namely

$$\phi_\xi = \frac{xy}{\theta_1\theta_2} - 1 \qquad \dots \quad (19)$$

$$\phi_1 = \frac{x}{\theta_1} - 1 \qquad \dots \quad (20)$$

$$\phi_2 = \frac{y}{\theta_2} - 1. \qquad \dots \quad (21)$$

Because of familiarity with the Poisson distribution, the calculation of the variance of the variance-covariance matrix of (19), (20) and (21) presents no difficulty (remember : these calculations have to be performed on the assumption $\xi = 0$!), and the final result is the criterion

$$Z_n^* = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{n\{\bar{X}_n \bar{Y}_n\}^{\frac{1}{2}}}. \qquad \dots \quad (22)$$

The illustrative quality of this example depends on the fact that the criterion (22) appeals to intuition and, perhaps, could have been guessed. The numerator divided by n is a moment estimate of the covariance of X and Y. If it is divided by the square root of the product of variances of the two variables, then the whole could be considered as an estimate of the correlation coefficient. Actually in the denominator there is the square root of $n\bar{X}_n\bar{Y}_n$. If X and Y are really independent Poisson variables, then the means $\bar{X}_n$ and $\bar{Y}_n$ are actually estimates of the two variances and it appears that the optimal $C(\alpha)$ test based on (22) is a test of the hypothesis that the correlation of X and Y is zero.

I should mention that, quite frequently, when the optimal $C(\alpha)$ criterion is deduced, a little thought leads one to exclaim, more or less : I should have guessed it !

Example 2 : Test of the hypothesis that the treatment does not increase the mean of a normal random variable X with an unspecified mean μ and with an unspecified variance, supposed not to be affected by the treatment. Apart from randomization, this is one of the classical bookish cases for which a uniformly most powerful test exists, namely the Student-Fisher t-test. We assume $\pi = \frac{1}{2}$.

Clearly, in this case the optimal $C(\alpha)$ test criterion is determined by formula (16). With obvious notation, easy manipulations yield

$$Z_n^* = \frac{2\sqrt{n_0 n_1}}{n} \, \frac{x_1 - \bar{x}_0}{\partial\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}. \qquad \dots \quad (23)$$

Here, as n grows, the first factor has a stochastic limit equal to unity. On the other hand, if σ is estimated by the usual mean square, the second factor is exactly the Student-Fisher t. As n is increased, t tends to be normally distributed, and so does Z_n^*.

5. Two not so bookish examples

Example 3 : *Does a treatment affect the learning ability of rats?* In a randomized training experiment with rats the observable variable X represents the time before a rat makes "a mistake." Rats are observed for a fixed limited period of time which we can take as unity, so that $0 < X \leqslant 1$. No information on the distribution of X is available, but the experimenter expects, or hopes, that the treatment given to randomly selected rats tends to increase the time to error, X. The hypothesis to test, using observations on n rats, some treated and some not, is that the treatment has no effect on the distribution of X. How should one test this hypothesis? Obviously, several non-parametric tests are available. Also, by dividing the unit interval of variation of X into several sub-intervals, one might use the χ^2 test. Each of these procedures will ensure at least an approximate maintenence of the chosen level of significance.

With reference to the chi square test, we note that, if the number n of observations is large, the number of sub-intervals can be considerable. If n is not very large, the number of sub-intervals must be limited, perhaps only to three. The following discussion of the possibility of using the $C(\alpha)$ technique will be limited to this particular case. The procedure is as follows.

We divide the unit interval into three arbitrary disjoint sub-intervals and denote by θ_i the unknown probability that the control rat X will fall into the i-th of them, $i = 1, 2, 3$. Because the three probabilities θ_i must add up to unity, the distribution of X considered involves two nuisance parameters, say θ_1 and θ_2. Now it is convenient to introduce three random variables $I_k(X)$, for $k = 1, 2, 3$, which we shall call indicators. For each k, the indicator $I_k(X)$ is defined to be equal to unity if X falls into the k-th sub-interval and to zero otherwise. Now we have to define the functions $\vartheta_j(\xi, \theta)$, for $j = 1, 2$, so as to incorporate into the definition the experimenter's vague hope that the value of X for a treated rat is, generally, somewhat larger than that for a control rat. Clearly, this can be done in a great variety of ways, each adding something to the vague idea that the treatment tends to increase X.

One possibility, discussed in Neyman (1970), is as follows. We presume that the treatment effect on the least "trainable" rats, whose X would normally fall within the first of the three sub-intervals, can increase X to fall into the second sub-interval but hardly ever so that it falls within the third. Similarly, we presume that for some of the "middle" trainable rats, with X normally falling within the middle of the three sub-intervals, the application of the treatment would make X fall into the last sub-interval. In order to approximate these expectations, we write

$$\vartheta_1(\xi, \theta) = \theta_1(1-\xi), \qquad \dots \quad (24)$$

with ξ positive and not exceeding unity, and

$$\vartheta_2(\xi, \theta) = \theta_2+\xi(\theta_1-\theta_2). \qquad \dots \quad (25)$$

Denoting by $\vartheta_3 = 1-\vartheta_1-\vartheta_2$ the frequency function of the randomizing random variable T and of the three indicators can now be written as

$$\pi^t(1-\pi)^{1-t} \prod_{k=1}^{3} [\vartheta_k(t\xi, \theta)]^{I_k(x)} \qquad \dots \quad (26)$$

and the use of formula (16) leads to the optimal $C(\alpha)$ test criterion. The formula for the asymptotic power provides indications as to the way of dividing the unit time interval into three sub-intervals so as to increase the sensitivity of the experiment. Through a set of Monte Carlo simulation experiments it was found that the power of the optimal $C(\alpha)$ test deduced as above exceeds substantially that of the corresponding chi square test.

Example 4 : ***Is it true that, as suggested by some meteorologists, the seeding of clouds during certain unidentified types of stormy days tends to increase the rainfall, while on some other days, the effect is negative, so that the average effect is zero*** ? Suggestions of the above kind seem to have been first made by E. J. Smith of CSIRO, Sydney, Australia (Smith, 1967). Later they were echoed by other authors. The kind of evidence supporting these suggestions is illustrated in Figure 1, which is based on a cloud seeding experiment performed in Quebec (Godson, Crozier and Holland, 1966). Three equal square areas 37 miles on the side were involved, labeled North, Buffer, and South. Seeding of clouds was performed on all "suitable" days either over the North or over the South area, chosen at random (this is the so-called cross-over design). It was hoped that the large Buffer area would prevent contamination. The evaluation by the three authors indicated that the average apparent effect of seeding was practically zero. If the true effect were so, then the joint distribution of X = rain in the South and Y = rain in the North on "North-seeded days" would have been the same as on "South-seeded days." A glance at the scatter diagrams in Figure 1 tends to contradict this assumption and to support the suggestion of E. J. Smith. The difficulty is that this support depends on the presumption that the presence of the Buffer area did, in fact, prevent the contamination of rain over one of the two

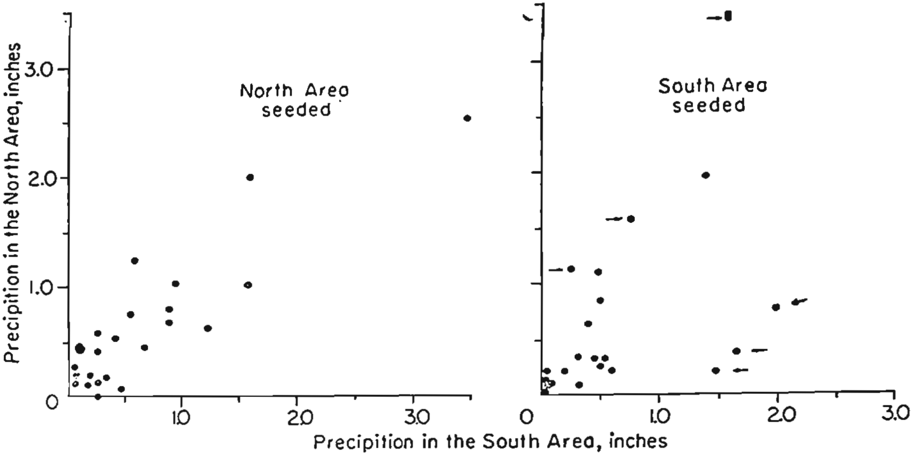

Figure 1. Mean areal rainfall, Western Quebec, Canada, 1960-63

experimental areas by seeding over the other. Since this is subject to uncertainty, the judgment about Smith's suggestion must be based on experiments with a simple randomized seed-no-seed design. Here, then, the problem arises of deducing a test of the hypothesis H_0 of no effect of seeding, promising to be particularly powerful against the somewhat vague alternatives that yes, the true effect of seeding exists but is occasionally positive and occasionally negative, averaging out to zero. (The first test of this kind, but referring to a cross-over design and assuming a joint normal distribution of X and Y, was the subject of a thesis by Kulkarni. His results were reported in Smith (1967). The application of this test to Quebec data leaves little doubt that either there was a variable effect of treatment or there was some contamination phenomenon.)

In the following a procedure is described leading to the optimal $C(\alpha)$ test of the hypothesis H_0. The basic assumptions will be (a) that the non-zero rainfall on both seeded and not-seeded experimental units (days, storms, etc.) follows a Gamma distribution, and (b) that the effect of seeding, if any, alters the scale parameter, not the shape parameter, of the Gamma density. (There is no certainty about this point, only some little evidence in its favor. However, the adequacy of this presumption is not the subject of the present discussion.)

The Gamma density can be written in the form, say

$$p(x \mid \theta_1, \theta_2) = \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} x^{\theta_1 - 1} e^{-\theta_2 x} \qquad \ldots \quad (27)$$

with two nuisance parameters, the shape parameter θ_1 and the reciprocal of the scale θ_2, both positive numbers. When trying to deduce the optimal $C(\alpha)$ test of H_0 against the alternative, asserting what may be called a fixed effect, we assume that θ_1 is not affected by the treatment, but that θ_2 may be, and introduce $\vartheta_2(\xi, \theta) = \theta_2 e^{-\xi}$. If ξ is positive, the fixed effect is also positive, that is, the treatment tends to increase the rainfall, and vice versa. However, in the present case the contemplated alternatives do not assert the presence of a fixed effect. They assert that for some unspecified experimental units the effect is positive and on some others it is negative. In fact, it is admitted, some of these effects may be quite large and some others small. Obviously, in order to incorporate such possibilities into the theory underlying the deduction of an optimal $C(\alpha)$ test, one must assume that for treated experimental units the scale parameter (and hence its reciprocal θ_2) is a random variable with some unspecified distribution. This can be done in several different ways, but the way that I found convenient is as follows.

The basic assumption is that the random variation in θ_2 due to the random effect of treatment is such that the logarithm of θ_2 has a finite variance which we denote by $\xi \geqslant 0$. Also, we assume that this distribution is regular enough to allow differentiation under the sign of certain integrals. Denoting by $\log \theta_2^0$ the mean of $\log \theta_2$ we can write

$$\frac{\log \theta_2 - \log \theta_2^0}{\sqrt{\xi}} = U. \qquad \dots \quad (28)$$

Here, then, U is a random variable, with expectation zero and unit variance. Denote by $F(u)$ the distribution of U which is unspecified but is uniquely determined by the postulated distribution of $\log \theta_2$. It follows that

$$\theta_2 = \theta_2^0 \, e^{U\sqrt{\xi}}. \qquad \dots \quad (29)$$

For any particular treated experimental unit the variable U has some fixed value. Given that value, the probability density of the rainfall X is given by (27) with θ_2 replaced by (29). For any experimental unit, whether treated or not, the joint distribution of the randomizing T and of the non-zero rainfall X is given by

$$\left[(1-\pi)\frac{x^{\theta_1-1}\theta_2^{0\theta_1}e^{-x\theta_2^0}}{\Gamma(\theta_1)}\right]^{1-t}\left[\pi\frac{x^{\theta_1-1}\theta_2^{0\theta_1}}{\Gamma(\theta_1)}\int \exp\{\theta_1 u\sqrt{\xi} - x\theta_2^0 e^{u\sqrt{\xi}}\}dF(u)\right]^{t} \qquad \dots \quad (30)$$

where the integral extends from $-\infty$ to $+\infty$.

It is seen that the joint distribution of T and X for $\xi > 0$ differs from that corresponding to $\xi = 0$ by more than just a change in the nuisance parameter θ. It follows that the deduction of the optimal $C(\alpha)$ test must be based on formula (11) rather than on the simpler formula (16).

Taking the logarithm of (30) and differentiating (the differentiation under the integral sign being assumed legitimate), we gradually find

$$\phi_1(x) = \log x - \log \theta_2^0 - \frac{\Gamma'(\theta_1)}{\Gamma(\theta_1)} \qquad \dots \quad (31)$$

$$\phi_2(x) = \frac{\theta_1 - \theta_2^0 x}{\theta_2^0} \qquad \dots \quad (32)$$

$$\phi_\xi = t[(\theta_1 - \theta_2^0 x)^2 - \theta_2^0 x]/2. \qquad \dots \quad (33)$$

Further calculations depend upon the nature of the estimators of the two nuisance parameters to be used. In order to gain an intuitive feeling of the working of the test, let us assume that the two parameters are estimated by the method of maximum likelihood. In other words the estimators θ_1 and θ_2^0 will be roots of the equations

$$\sum_{i=1}^{n} \phi_1(x_i) = \sum_{i=1}^{n} \phi_2(x_i) = 0 \qquad \ldots \quad (34)$$

where the summations extend over all the n observations, those with treatment and controls. In consequence, the numerator in formula (11) will reduce to the sum of terms depending upon ϕ_3 alone and we shall have

$$Z_n^* = \sum_{i=1}^{n} t_i[(\theta_1 - \theta_2^0 x_i)^2 - \theta_2^0 x_i]/\sigma\sqrt{n}. \qquad \ldots \quad (35)$$

Comparing (34) with (32) it is seen that

$$\theta_1 = \theta_2^0 \bar{x} \qquad \ldots \quad (36)$$

where $\bar{x}$ stands for the mean of all the n observations, under treatment and controls. Substituting (36) in (35), simplifying, and remembering that $t_i = 0$ for all the control observations, it is easy to see that the use of the criterion Z_n^* amounts to checking whether for observations under treatment the sum of squares $\Sigma(x_i - \bar{x})^2$ exceeds significantly what would be expected with no variable effect of treatment.

6. General problem of variability in response to treatments

The above discussion depends explicitly on certain preconceptions connected specifically with rain stimulation and with the idea that in a certain situation when the average effect of cloud seeding is zero, the seeding may be increasing rain in some cases and in some other cases it may be decreasing it. In other words, the solution described refers to a very particular case, involving gamma density, with only the scale parameter being subjected to possible change due to treatment, etc. It is of some interest to generalize the problem.

Consider a randomized experiment intended to discover whether the experimental units (storms, mice, etc.) exhibit a variable response to the treatment studied. For example, to quote Professor Kempthorne : is it true that penicillin is occasionally helpful and occasionally harmful ? As formerly, the randomizing variable T can have only two values : unity, with known probability π, and zero. If $T=1$, the particular experimental unit is subjected to the treatment, but not otherwise. The statistician is prepared to act on the assumption that with $T=0$, the observable variable X, possibly a vector, has a known distribution depending on some nuisance parameters. There may be several of them but for our purposes here it will be sufficient to assume that there are only two nuisance parameters θ_1 and θ_2. Symbol $p(x|\theta_1, \theta_2)$ will represent the corresponding density or frequency function.

If $T=1$, so that the particular experimental unit is subjected to the treatment, the statistician is prepared to act on the assumption that the distribution of the same X may be different. More specifically, he is prepared to assume that this distribution is characterized by a function, say $q(x|\theta_1, \theta_3, \xi, u)$ depending upon four parameters. The first of them θ_1 is the same as that in the distribution corresponding to $T=0$. The next nuisance parameter θ_3 is a new one. The third parameter $\xi \geqslant 0$ is a conventional measure of the variability of response to the treatment. If $\xi = 0$, then all the experimental units respond to the treatment similarly, in the sense that the corresponding distribution is fully characterized by $q(x|\theta_1, \theta_3, 0, 0)$. On the other hand, if $\xi > 0$ then the distribution of X *for a particular* experimental unit depends also on a number *u characterizing this particular unit.* This number u is considered as a particular value of a random variable U with an unspecified distribution $F(u)$, except that it satisfies the conditions $EU = 0$ and $EU^2 = 1$. (In addition, of course, all the distributions mentioned satisfy certain conditions of regularity, rather obvious but too long to describe in this paper). The particular values of U are not observable. In consequence, for a randomly selected experimental unit the distribution of X corresponding to $T=1$ is characterized by the marginal,

$$Q(x|\theta_1, \theta_3, \xi) = \int q(x|\theta_1, \theta_3, \xi, u)dF(u), \qquad \ldots \quad (37)$$

where the integral extends from $-\infty$ to $+\infty$.

With these assumptions, the joint distribution of the two observable random variables T and X is given by the formula

$$[(1-\pi_1)p(x|\theta_1, \theta_2)]^{1-t}\,[\pi_1 Q(x|\theta_1, \theta_3, \xi)]^t, \qquad \ldots \quad (38)$$

where t stands for the value of T either zero or unity.

The reader may enjoy the deduction of an optimal $C(\alpha)$ test for the presence of variability of response to treatment in a situation different from that in Section 5. One possibility is to continue considering the general set-up referring to cloud seeding, including Gamma density in formula (27), etc., but just drop the assumption that, apart from the possible variable effects, the treatment does not influence the average rainfall. The essential difference between the earlier approach and the one now suggested consists in admitting that the "typical" value of the reciprocal of the scale paramter for control units is an unknown positive number θ_2, while for the treated units it is possibly a different positive number θ_3 with a consequent slight change in formula (30). In the first factor θ_2^0 would be replaced by θ_2 and in the second factor by θ_3. The three logarithmic derivations (31), (32), and (33) would have to be replaced by four. Also, there would be three rather than two nuisance parameters to estimate. If the maximum likelihood method is used, then θ_1 would be estimated using all the n observations, with and without treatment, θ_2 using only the observations on controls and θ_3 only those with treatment.

Another interesting application of the theory might consist in considering experiments other than those with rain stimulation and involving distributions of the observable X other than Gamma. Also, it is likely to be important to consider a different specification of the conventional measure ξ of variability of response to treatment. Some remarks to this effect will be found in the next section.

7. Some difficulties and a challenge

The treatment of practical problems requires not only an easily computable test criterion having an intelligible property of optimality, but also a reasonable approximation to its power function. When approaching a significant study, it is important to be able to compute how many observations must be made in order to have a reasonable chance to "detect" the effect of the treatment if it exists and is of specified magnitude. Here an appropriate specialization of formula (10) comes to one's mind. But how good is the approximation provided by this formula ?

One source of uncertainty in this respect is the method by which it has been deduced. Invented in 1936 (Neyman, 1937) and since used in countless studies of asymptotic relative efficiency, this method utilizes a "double" passage to limit : The number n of observations is increased while the "error" ξ in the hypotheses tested goes to zero, so that the product $\xi\sqrt{n}$ tends to a finite limit $\tau \neq 0$. The expression on the right in formula (10) is just the limit of the power with τ replaced by $\xi\sqrt{n}$.

The adequacy of the formula was first checked in Buhler and Puri (1966) and many times later with varying results. The method of verification was empirical; through Monte Carlo simulation. In a number of situations the working of the formula proved satisfactory, but in a number of others it was hopelessly bad. One category of cases in which formula (10) proved useless was found in an unpublished work by Robert Traxler, then our Ph.D. student, currently at the University of Maryland.

Traxler considered a problem closely related to the problem of variability of response, but a simpler one. This problem was to test the hypotheses that an observable X follows a Poisson distribution with an unspecified parameter λ_0, rather than an unspecified mixture of Poisson distributions. His set-up and the definition of ξ were very similar to those discussed in Section 5. Each observation on X was supposed to be Poisson distributed with expectation

$$\lambda_0 \exp\{u\sqrt{\xi}\} \qquad \ldots \quad (39)$$

where u is a particular value of a random variable U, such that $EU = 0$ and $EU^2 = 1$. The easily deducible optimal $C(\alpha)$ criterion happens to be equivalent to that proposed long ago by R. A. Fisher, with obvious notation $S^2/\bar{X}$. The thing that interested Traxler was the power of the test corresponding to a fixed value of ξ and to several different distributions U. One possibility was that U is uniformly distributed in an interval centered at zero. The other possibilities studied were two point distributions of U, one value being negative and the other positive. All the results obtained through Monte Carlo simulations proved interesting but one of them was rather distressing. Also, it was instructive.

When Traxler assumed that U could have only two values $\underline{a} < 0$ and $\underline{b} > 0$ with $\underline{a}$ rather small and frequent, while $\underline{b}$ had to be large and rare, the power of the optimal $C(\alpha)$ test, corresponding to a substantial number of observations, was found to be rather close to α, the intended level of significance. At the same time, formula (10) predicted a reasonable probability of detecting the falsehood of the hypotheses tested !

While at first disappointing, this result proved instructive. It showed that the conventional measure ξ of the error in the hypotheses tested, defined in (39) is not appropriate. The motivation behind this formula, as well as that behind formula (29), was based on a rather universal custom of using variance as a measure of variability. However, the tests under discussion are not really dealing with variability of the scale parameter in one case and the Poisson parameter in the other. All these tests could do is to distinguish between two distributions of the observable X, one of them fixed and the other a mixture. If the mixture of distributions is very similar to the one fixed, then a power of a test cannot be large. And it happens that, with a fixed variance ξ, and Traxler's $\underline{a}$ very small, the resulting mixture of distributions is very close to a fixed Poisson distribution.

The moral of these findings is that, in order to have a reasonable hope for the conformity with the asymptotic power function (10) a new definition of what we call the conventional measure of the error in the hypothesis tested must be invented. Could one somehow develop the theory by defining

$$\xi = \sup_{x} | F_f(x\,|\,0) - F_m(x\,|\,\xi) |$$

where F_f and F_m designate the fixed and the mixture distributions of the observable X, each with some nuisance parameters ?

This is one of the problems awaiting solution. Here is another. It consists in developing reliable methods of estimating the power of the already existing optimal $C(\alpha)$ test. One remark is obvious (Buhler and Puri, 1966). This is that the actual power must depend upon the identity of the locally root n consistent estimators of the nuisance parameters. Recently, I tried my hand in this direction, but with little success. Now my hopes center on two papers (Singh and Zhurbenko, 1975; Zhurbenko, 1976) symbolizing an international effort to solve an important problem. The senior author of the earlier paper is Igor Zhurbenko of the University of Moscow who was our several months' guest in Berkeley; The junior author is Avinash Singh, at the time our Ph.D. student from India. The subsequent paper by Professor Zhurbenko, represents a version of the same study. It is a great pleasure to note this truly international cooperative research effort.

Acknowledgments

This paper was prepared using the facilities of the Statistical Laboratory with partial support from the Office of Naval Research (ONR N00014 75 C0159), the National Institute of Environmental Health Sciences (2 R01 ES01299-16), and the Department of the Army (Grant DA AG 29 76 G 0167).

REFERENCES

BÜHLER, W. J. and PURI, P. S. (1966): On optimal asymptotic tests of composite hypotheses with several constants. *Z. Wahrscheinlichkeit. Verw. Gebiete*, 5, 71-83.

DAVIES, ROBERT B. and PURI, PREM S. (1967): Some techniques of summary evaluations of several independent experiments. *Proc. Fifth Berkeley Symp. Math. Stat. and Prob.*, University of California Press, Berkeley and Los Angeles, 5, 385-388.

GODSON, W. L., CROZIER, C. L. and HOLLAND, J. D. (1966): Silver iodide cloud seeding by aircraft in western Quebec, Canada, 1959-1963. *J. Appl. Meterorol.*, 5, 500-512.

NEYMAN, J. (1937): Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20, 149-199.

———— (1959): Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics* (The Harald Cramér Volume), Almquist and Wiksells, Uppsala, Sweden, 213-234.

———— (1967): Experimentation with weather control. *J. Royal Stat. Soc.*, 130, 285-326.

———— (1970): Statistical problems in science: The symmetric test of a composite hypothesis. *J.A.S.A.*, 64, 1154-1171.

NEYMAN, J. and SCOTT, E. L. (1965): Asymptotically optimal tests of composite hypotheses for randomized experiments with non-controlled predictor variables. *J.A.S.A.*, 60, 699-721.

———— (1967): Note on techniques of evaluation of single rain stimulation experiments. *Proc. Fifth Berkeley Symp. Math. Stat. and Prob.*, University of California Press, Berkeley and Loss Angeles, 5, 371-384.

SINGH, AVINASH C. and ZHURBENKO, IGOR G. (1975): The power of the optimal asymptotic tests of composite statistical hypotheses. *Proc. Nat. Acad. Sci.*, 72, 577-580.

SMITH, E. J. (1967): Cloud seeding experiments in Australia. *Proc. Fifth Berkeley Symp. Math. Stat. and Prob.*, University of California Press, Berkeley and Loss Angeles, 5, 161-176.

ZHURBENKO, IGOR G. (1976): Estimates for the power of tests of composite hypotheses. *Dokl. Akad. Nark. SSSR*, 226, 1253-1256.