For $i = 1, \cdots, k - 1$ let $m_i^{(N)}$ denote the largest integer $\leqq N y_i^{(N)}$ and let $m_k^{(N)} = N - m_1^{(N)} - \cdots - m_{k-1}^{(N)}$. Define the point $z^{(N)}$ by

$$N z_i^{(N)} = m_i^{(N)} + 2 \quad \text{if} \quad a_i^{(N)} \geqq 0, \qquad N z_i^{(N)} = m_i^{(N)} - 1 \quad \text{if} \quad a_i^{(N)} < 0,$$

for $i \leqq k - 1$ and $z_k^{(N)} = 1 - z_1^{(N)} - \cdots - z_{k-1}^{(N)}$. Then

$$|z_i^{(N)} - y_i^{(N)}| < 2k/N, \qquad i = 1, \cdots, k. \tag{A.11}$$

Since $y^{(N)} \varepsilon \Omega_\epsilon$, $z^{(N)}$ is in $\Omega_{\epsilon/2}$ for $N > N_1 \geqq N_0$. Moreover,

$$a_i^{(N)}(z_i^{(N)} - y_i^{(N)}) \geqq N^{-1} |a_i^{(N)}|, \qquad i = 1, \cdots, k - 1.$$

Hence for $N > N_1$

$$\begin{aligned} f(z^{(N)}) - c_N &\geqq N^{-1} \textstyle\sum_{i=1}^{k-1} |a_i^{(N)}| + O(N^{-2}) \\ &\geqq \tfrac{1}{2} N^{-1} \max_{ij} |f_i'(y^{(N)}) - f_j'(y^{(N)})| + O(N^{-2}). \end{aligned}$$

Condition (A.10) implies that for N large enough we have $f(z^{(N)}) > c_N$, that is, $z^{(N)} \varepsilon A_N$.

Thus the conditions of Lemma A.1 are satisfied. The proof is complete.

REFERENCES

[1] Bahadur, R. R. and Rao, R. Ranga (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027.

[2] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507.

[3] Hoeffding, Wassily (1963). Large deviations in multinomial distributions. (Abstract.) *Ann. Math. Statist.* **34** 1620.

[4] Hoeffding, Wassily (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[5] Matusita, Kameo (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. *Ann. Math. Statist.* **26** 631–640.

[6] Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20-A** 175–240, 264–299.

[7] Sanov, I. N. (1957). On the probability of large deviations of random variables (Russian) *Mat. Sbornik N. S.* **42** (**84**), 11–44. English translation: *Select. Transl. Math. Statist. and Probability* **1** (1961) 213–244.

[8] Wald, Abraham (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

DISCUSSION OF HOEFFDING'S PAPER

Jerzy Neyman[1]: Professor Hoeffding is to be heartily congratulated on his very interesting paper. His results as explicitly formulated are important enough. It *is* important to know that out of the several tests of the same hypothesis, the tests whose certain asymptotic properties are identical and which, therefore, were considered equivalent, one particular test has an asymptotic property, not pre-

[1] Prepared with the partial support of the National Science Foundation Grant GP-2593.

viously considered, that makes it superior to other tests. However, Professor Hoeffding's paper goes further than merely proving the superiority of a particular test. In fact, it is my expectation that Professor Hoeffding's paper is the first section of a new and a very important chapter in the theory of statistics.

As is frequently the case in the history of research, Professor Hoeffding's present success is due to his seizing upon, and to his making an effective use of, a novel tool, the theory of "large deviations". For some time a few statisticians have been aware of the "probabilities of large deviations", particularly of the pioneer work of Harald Cramér [2] and of its extension by Feller [4]. In fact, these results have already been used in a limited way by Herman Chernoff [1] and a little earlier by Charles M. Stein ([1], p. 18). However, the favorite tools in the study of asymptotic properties of tests and of estimtaes remained the various versions of the classical central limit theorem.

The importance of the present paper of Professor Hoeffding is not limited to the utilization of the novel tool. Even more important is his initiative to abandon or, perhaps, to extend the device, which I may call the device of alternatives infinitely close to the hypotheses tested, as a tool in deducing optimal asymptotic tests. This device has a respectable history and I, personally, have an emotional attachment to it. Yet, Professor Hoeffding's paper clearly indicates that the potential of the device of infinitely close alternatives as means of deducing optimal tests is already spent and that it should be replaced by some other more effective device and "probabilities of large deviations" [5] seems an excellent promise.

My own first use of the device of infinitely close alternatives was made in 1936 when I attempted to formulate the problem of an optimal asymptotic test [6]. This was done with reference to a particular simple hypothesis relating to a sequence of independent and identically distributed random variables. "Optimality" was understood as the property of maximizing the power function of the test and here I encountered a difficulty that appeared staggering. If one keeps the level of significance (or, at least its limit) constant and considers the power function with respect to a fixed alternative, with the increase of the number n of observations this power function would tend to unity, except for such tests as no one would consider decent. The device I adopted consisted in considering a sequence $\{h_n\}$ of hypotheses alternative to the one tested which, as $n \to \infty$, approached the hypotheses tested. The corresponding power function of any given test criterion may then tend to a limit and the value of this limit was taken as a criterion of optimality.

To my knowledge, the next use of the same device is due to Churchill Eisenhart [3] who applied it to the deduction of the asymptotic power of Karl Pearson's χ^2 test.

Subsequently, the same device was used by a number of other authors. One of the most fruitful uses is due to E. J. G. Pitman who introduced the concept of the relative asymptotic efficiency of tests. As is well known, Pitman's idea resulted in a series of important studies, particularly of non-parametric tests, whose mere enumeration would take more space than the present article.

While the early idea of optimality of an asymptotic test worked for simple hypotheses, its extension to composite hypotheses required a new effort to produce a computable family of critical regions similar to the sample space, at least asymptotically. In a preliminary fashion this was done in 1954 [7] and the theory of locally asymptotically most powerful tests of composite hypotheses [8] appeared a few years later.

The usefulness of this method is felt when dealing with "live" problems of applied character, where as a rule the observable random variables, perhaps vectors, have non-standard distributions. If a test of some hypothesis is required, and invariably this would be a composite hypothesis, such test had to be based on a guess or, alternatively, one could use the test deduced to have at least the local asymptotic optimality property, even though the latter is limited to a special family. This special family of comparison tests is determined by convenience in applying the classical central limit theorem.

While the device of infinitely close alternatives worked in the above fashion in a number of cases with which I had to deal, a recent experience showed its lack of sharpness and here Professor Hoeffding's paper serves as an indication of where to look for an alternative method.

In a paper, joint with Professor Elizabeth L. Scott, now submitted for publication, we used the device of infinitely close alternatives to treat a fairly complex situation of randomized experiments. One randomization considered was "unrestricted": for each of a sequence $\{U_n\}$ of experimental units (e.g. patients in a clinic) a coin is tossed to decide whether or not this particular unit be subject to a treatment. The other randomization scheme considered consisted in randomizing successive pairs of experimental units. A coin is tossed only for each "odd" unit U_{2k-1}, for $k = 1, 2, \cdots$. If the coin falls heads, U_{2k-1} is subjected to the treatment but U_{2k} is not, etc. For both these randomization schemes the locally asymptotically optimal test of class $C(\alpha)$ was deduced. Also, we found the asymptotic powers of these tests and then asked the question: Suppose that the same number $2n$ of experimental units, with identical distributions of the relevant observable variables, are used in a randomized experiment alternatively with unrestricted randomization and with randomization of pairs; how would the corresponding asymptotic power functions differ? Naturally, we expected that the asymptotic power corresponding to randomization of pairs would exceed that corresponding to the unrestricted randomization. To our surprise and regret we found this not to be the case: the asymptotic power functions corresponding to the two cases proved identical! Further analysis of the two test criteria, say $Z_1(n)$ for unrestricted randomization and $Z_2(n)$ for randomization of pairs, an analysis divorced from infinitely close alternatives and referring to a somewhat simplified situation, showed the following. For a fixed alternative, both criteria are asymptotically normal with the same means, say $\xi\sqrt{n}$. However, the asymptotic variances of the criteria, σ_1^2 and σ_2^2, differ. Namely

$$\sigma_1^2 = \sigma_2^2 + \xi^2,$$

so that the randomization of pairs appears more effective than the unrestricted randomization, as expected. The above finding indicates the mechanism behind the apparent paradox of the asymptotic power functions of the two criteria being identical. These two power functions were obtained through the device of infinitely close alternatives, that is through the passage to the limit as $n \rightarrow \infty$ in which ξ does not remain constant but diminishes $O(n^{-\frac{1}{2}})$. Under these circumstances the difference between σ_1^2 and σ_2^2 is of the order of $1/n$ and in the limit disappears. This is, then, another instance, parallel to that indicated by Professor Hoeffding, indicating that the device of infinitely close alternatives, while having the advantage of being easy to use, has the disadvantage of being not sufficiently sharp to catch the distinctions that may be important.

The next step, foreshadowed by the important paper of Professor Hoeffding, is now to devise a method of using the probabilities of large deviations for a workable deduction of asymptotic tests of composite hypotheses that are, in a well defined sense, optimal for a fixed alternative. Another outstanding problem suggested by Professor Hoeffding's paper is to re-examine the many recent results indicating the asymptotic equivalence of various tests. When the method of infinitely close alternatives indicates that a test T_1 is relatively more efficient than another test T_2, this result appears to be worthy of being taken at its face value. However, when the asymptotic relative efficiency of T_1 compared to T_2 is found to be unity, there is room for doubt as to what this may mean for any fixed alternative, even if the number of observations is large.

Each deviation from the old routine of thought opens new possibilities, frequently in some unanticipated directions. It is possible, therefore, that Professor Hoeffding's approach will open the way for the asymptotic treatment of a problem for which I was not able to deduce a fixed sample optimal test and for which the method of infinitely close alternatives did not prove effective. Briefly, the problem is as follows.

Consider two normal populations Π_1 and Π_2 with underlying distributions $N(\xi_i, \sigma_i^2)$. Let S_i represent a sample of n_i independent observations on Π_i, with $i = 1, 2$. Here $\xi_1, \xi_2, \sigma_1^2, \sigma_2^2$ are unknown numbers, except that the parameter point (ξ_1, σ_1^2) is likely to be rather different from (ξ_2, σ_2^2). (This presumption is one of the obstacles to the use of the "old" asymptotic approach.) Let x stand for $(n_1 + n_2 + 1)$st observation of which it is only known that its distribution is either $N(\xi_1, \sigma_1^2)$ or $N(\xi_2, \sigma_2^2)$, but nothing else.

The problem is a classification problem. It is required to find an optimal rule to decide whether x belongs to Π_1 or Π_2. I tried the following fixed sample size approach. Let X stand for the totality of the observable variables (S_1, S_2, x) and $\mathfrak{X}$ for the corresponding sample space. Further, let $\mathfrak{X}_1$ and $\mathfrak{X}_2 = \mathfrak{X} - \mathfrak{X}_1$ be measurable subsets of $\mathfrak{X}$. The decision rule will be to assert x belongs to Π_i if $X \,\varepsilon\, \mathfrak{X}_i$. The conditions to which I tried to subject $\mathfrak{X}_1$ and $\mathfrak{X}_2$ are as follows:

(i) *Symmetry of errors of misclassification.*

$$(1)\qquad \begin{aligned} P\{X \,\varepsilon\, \mathfrak{X}_2 \mid x \,\varepsilon\, \Pi_1\} &= P\{X \,\varepsilon\, \mathfrak{X}_1 \mid x \,\varepsilon\, \Pi_2\} \\ &= 1 - P\{X \,\varepsilon\, \mathfrak{X}_2 \mid x \,\varepsilon\, \Pi_2\} \end{aligned}$$

or

$$\tfrac{1}{2}[P\{X \varepsilon \mathfrak{X}_2 \mid x \varepsilon \Pi_1\} + P\{X \varepsilon \mathfrak{X}_2 \mid x \varepsilon \Pi_2\}] = \tfrac{1}{2} \tag{2}$$

for all ξ_1, ξ_2, σ_1^2 and σ_2^2. This symmetry condition, in Formula (2) is, in effect, the condition that the subset $\mathfrak{X}_2$ be similar to the sample space $\mathfrak{X}$ with respect to the density corresponding to the left hand side of (2), with unspecified parameters ξ_1, ξ_2, σ_1^2 and σ_2^2, and of "size" one-half.

Let Φ denote the family of all subsets $\mathfrak{X}_2$ satisfying the symmetry condition.

(ii) *Optimality condition.* In order that $\mathfrak{X}_2^0 \varepsilon \Phi$ be called optimal, it must satisfy the condition

$$P\{X \varepsilon \mathfrak{X}_2^0 \mid x \varepsilon \Pi_2\} \geqq P\{X \varepsilon \mathfrak{X}_2 \mid x \varepsilon \Pi_2\} \tag{3}$$

for all $\mathfrak{X}_2 \varepsilon \Phi$.

My specific question is: can one think of a plausible rewording of the above fixed sample size problem so it could be treated asymptotically from the point of view that Professor Hoeffding was so successful in initiating?

Naturally, there is nothing sacred in the assumption of normality of the two populations Π_1 and Π_2. It is the similarity of the subset $\mathfrak{X}_2$ indicated in (2) that is important.

REFERENCES

[1] Chernoff, Herman (1956). Large sample theory: parametric case. *Ann. Math. Statist.* **27** 1–22.

[2] Cramér, Harald (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Sci. Ind.*, No. 736.

[3] Eisenhart, Churchill (1938). The power function of the χ^2 test. *Bull. Amer. Math. Soc.* **44** 32.

[4] Feller, W. (1943). Generalization of a probability theorem of Cramér. *Trans. Amer. Math. Soc.* **54** 361–372.

[5] Linnik, Yu. V. (1961). On the probability of large deviations for the sums of independent variables. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **2** 289–306. Univ. of California Press.

[6] Neyman, J. (1937). Smooth test for goodness of fit. *Skand. Aktuarietidskr.* **20** 149–199.

[7] Neyman, J. (1954). Sur une famille de tests asymptotiques des hypothèses statistiques composées. *Trabajos Estadist.* **5** 161–168.

[8] Neyman, Jerzy (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics—The Harald Cramér Volume.* 213–234. Almqvist and Wiksell, Stockholm.

Herman Chernoff: Professor Hoeffding has contributed a remarkably fresh insight into the applicability of the theory of the probability of large deviations to hypothesis testing. Using a single crude approximation, powerful conclusions are simply derived for the multinomial distribution. This approach promises to extend to other families of distributions. I shall take the liberty of paraphrasing the Hoeffding development, and by dropping details required for rigor, try to expose the simple underlying approach.

The function $I(x, p)$ is a special case of the Kullback Leibler Information number and some slight additional insight may be gained by keeping this in mind. In general, consider the problem of testing a simple hypothesis $H_1:f = f_1(x)$ versus a simple alternative $H_2:f(x) = f_2(x)$ on the basis of N independent observations $X_1, X_2, \cdots, X_n$, with density $f(x)$, with respect to a measure μ. If w_1 and w_2 are *a priori* probabilities for H_1 and H_2 the *a posteriori* probabilities w_1^* and w_2^* satisfy

$$\frac{w_2^*}{w_1^*} = \frac{w_2}{w_1}\frac{\prod f_2(X_i)}{\prod f_1(X_i)} = \frac{w_1}{w_2} e^{-S_N}$$

where $S_N = \sum \log [f_1(X_i)/f_2(X_i)]$. If H_1 is true S_N is approximately $NI(f_1, f_2)$ where

$$I(f_1, f_2) = \int f_1(x) \log [f_1(x)/f_2(x)] \, d\mu(x)$$

is the natural generalization of the formula used in the multinomial case. Thus I measures the exponential rate at which $w_2^* \to 0$ when $f = f_1$. That is to say I is a measure of the ability to discriminate against f_2 when f_1 is the true density and hence I may be regarded as an asymmetric measure of distance between f_1 and f_2.

That I is asymmetric is easy to see but the example of $p^0 = (1, 0)$ and $p^1 = (.9, .1)$ is informative. Here $I(p^0, p^1) < \infty$ while $I(p^1, p^0) = \infty$. The statistical explanation for this case is the following. If p^1 is the true distribution, a finite number of observations will yield an observation in the second cell completely disproving p^0. If p^0 is the true distribution, the fact that no observations occur in the second cell will build up evidence against p^1 in a more gradual fashion. In general points on the boundary of the p simplex are infinitely far from interior points but not vice versa.

Keeping this distance interpretation in mind the results described by Hoeffding flow from the crude probabilistic approximation

$$P\{Z^{(N)} \varepsilon A \mid p\} = \exp\{-NI(A^{(N)}, p) + O(\log N)\}$$

where $I(A^{(N)}, p)$ is the shortest distance from $A^{(N)}$ to p, and the analytic properties of I. Disregarding fine details such as the remainder term and the distinction between $A^{(N)}$ and A, the above approximation states that the probability of falling in A when p^0 is the true probability is not substantially increased by adjoining to A all points which are at least as distant to p^0 as $I(A, p^0)$. Statistically speaking, the size α_N of the test "reject $H_0:p = p^0$ when $Z^{(N)} \varepsilon A$" is not much increased by increasing the critical region to reject when $I(Z^{(N)}, p^0) \geqq I(A, p^0)$. The latter test is at least as powerful since its critical region contains more points. Furthermore the latter test is simply the likelihood-ratio test. The ratio in the power functions of the two tests for an alternative $p \neq p^0$ is determined by d_N the difference in distances to p from the two acceptance regions.

While this reasoning combined with an analytic study of $I(z, p^0)$, $z \varepsilon A$, for the chi-square test represents a heuristic outline of the proof of Hoeffding's result it requires care and modification to be done rigorously. Unfortunately the basic approximation does not seem refined enough to prove the conjecture that the L.R. test is more powerful than the χ^2 test when $\alpha_N \to 0$ slowly.

An obvious generalization that suggests itself but is not explicitly mentioned in the paper is the efficiency of the likelihood ratio test for composite hypotheses versus composite hypotheses. For testing $H_0 : p \varepsilon \Lambda_0$ vs. $H_1 : p \varepsilon \Lambda_1$, the likelihood ratio test consists of rejecting H_0 when

$$I(Z^N, \Lambda_1) - I(Z^N, \Lambda_0) \geqq c.$$

A suggestion that a theory of large deviations could be fruitful for the discrimination problem posed by Neyman in his discussion may derive from the following remark concerning a much simpler problem. It is desired to find a linear function of an observation X to discriminate between the two specified alternatives. Normal multivariate populations $\mathfrak{N}(\xi_0, \Sigma_0)$, $\mathfrak{N}(\xi_1, \Sigma_1)$. The linear function $a'X$ has the alternative normal distribution $\mathfrak{N}(\mu_0, \sigma_0^2)$ and $\mathfrak{N}(\mu_1, \sigma_1^2)$ where $\mu_i = a'\xi_i$, $\sigma_i^2 = a'\Sigma_i a$, $i = 1, 2$.

Since a normally distributed variable with a small variance may be regarded as the average of many observations, the theory of large deviations for sums of random variables may be applied if $|\xi_1 - \xi_0|$ is "large" compared to Σ_0 and Σ_1. From this theory it follows that if one desires to minimize the sum of the two error probabilities, then a should be selected to minimize $|\mu_1 - \mu_0|/\sigma_1 + \sigma_0$.

Finally the problem of further generalization is of interest. What is the natural extension of the basic approximation to problems other than those involving the multinomial distribution? For what families of distributions classified by a parameter θ, can we write an expression of the form

$$P\{\hat{\theta} \varepsilon A\} \approx \exp\{-NI(A, \theta) + O(\log N)\}$$

where $\hat{\theta}$ is an estimate of θ, and I is either the Kullback-Liebler Information or some other appropriate measure of distance?

D. G. Chapman: For many tests knowledge of the asymptotic distribution of the test statistic when the null hypothesis is true has served as a useful practical tool in application but for such tests study of the power is difficult. For if the sequence of tests is consistent for fixed alternatives and fixed test size, the asymptotic power of the sequence is one. As Hoeffding points out, to obtain a meaningful comparison it is customary to consider either a sequence of alternatives tending to the null hypothesis or a sequence of tests of size tending to zero. The former approach leads to Pitman's asymptotic relative efficiency which has operational meaning but unfortunately primarily for alternatives "close" to the null hypothesis and these may not be of practical importance. The alternative approach has however been limited by an inadequtae theory of probabilities of large deviations. In the absence of such a theory, Bahadur's stochastic comparison of tests [2] may be thought of as representing an approximation to this method. While Bahadur's "asymptotic slope" yields a "functional on the family of power functions associated with the sequence of statistics" which may be easily evaluated for alternatives of interest, it is not clear what operational meaning this functional has, though some useful properties were noted by

Bahadur. It is therefore of interest to ask what relationship Bahadur's stochastic comparison has to the results obtained by Hoeffding based on the exact study of the probability of large deviations.

For simplicity we consider only the case of the simple hypothesis of the form $p_1 = p_1^0, \cdots, p_k = p_k^0$. It is straight forward to verify that

$$T_n^{(1)} = [\textstyle\sum_{i=1}^k (n_i - Np_i^0)^2/Np_i]^{\frac{1}{2}} \qquad T_n^{(2)} = [2 \textstyle\sum_{i=1}^k n_i \log (n_i/Np_i^0)]^{\frac{1}{2}}$$

are standard sequences as defined by Bahadur with $a = 1$ in both cases and

$$[b^{(1)}(p)]^2 = \textstyle\sum_{i=1}^k (p_i - p_i^0)^2/p_i^0 \qquad [b^{(2)}(p)]^2 = 2 \textstyle\sum_{i=1}^k p_i \log (p_i/p_i^0).$$

The stochastic comparison of the χ^2 test based on the statistic $[T_n^{(1)}]^2$ and the likelihood test based on $[T_n^{(2)}]^2$ reduces to a comparison of $[b^{(1)}(p)]^2$ and $[b^{(2)}(p)]^2$. It is easily seen that we may have inequality in either direction. For example, for $p_1 = 1, p_i = 0, i > 0$, it is observed that $[b^{(1)}(p)]^2 \gtrless [b^{(2)}(p)]^2$ according as $p_1^0 \lesseqgtr 0.2847$.

Slightly more generally if $p_i^0 = 1/k$, $p_i = (1/k) + \delta_i$, where $\sum_{i=1}^k \delta_i = 0$,

$$[b^{(1)}(p)]^2 = \textstyle\sum_{i=1}^k \delta_i^2 \qquad [b^{(2)}(p)]^2 = 2 \textstyle\sum_{i=1}^t [(1/k) + \delta_i] \log (1 + k\delta_i).$$

If the latter function is expanded in a Taylor series retaining the first three terms, it follows that to this approximation

$$[b^{(2)}(p)]^2 = [b^{(1)}(p)]^2 - \textstyle\sum_{i=1}^k [(k^2\delta_i^3)/3] + \textstyle\sum_{i=1}^k [(k^3\delta_i^4)/12].$$

The sum of the last two terms may be either negative or positive. It appears therefore that there is little, if any, relationship between the general results proven by Hoeffding for multinomial tests and the stochastic comparison approach.

It should be noted that the stochastic comparison of tests was anticipated by Anderson and Goodman [1] and applied to comparison of the likelihood ratio test to a χ^2 test, specifically dealing with Markov chains.

REFERENCES

[1] Anderson, T. W. and Goodman, Leo A. (1957). Statistical inference about Markov chains. *Ann. Math. Statist.* **28** 89–109.

[2] Bahadur, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Statist.* **31** 276–295.