SOME RECENT DEVELOPMENTS IN NONPARAMETRIC STATISTICS *

by

Wassily Hoeffding

University of North Carolina

Institute of Statistics Mimeo Series No. 523

SUMMARY

This is a survey of problems and recent developments in some selected areas of nonparametric statistical theory. The paper is divided into three parts, Nonparametric versus parametric tests, Robust estimates, and Robust analysis of variance.

SOMMAIRE

Ce travail est une revue des problèmes et des développements récents dans quelques branches choisies de la théorie statistique non-paramétrique. L'article est divisé en trois parties: Tests non-paramétriques versus tests paramétriques; estimateurs robustes; analyse robuste de la variance.

*This research was supported in part by the Mathematics Division of the Air Force Office of Scientific Research.

1. NONPARAMETRIC VERSUS PARAMETRIC TESTS

The main motivation for the development of nonparametric statistics was the need for statistical methods that have desirable properties when little is assumed about the population or populations being sampled. For a number of problems tests were designed whose probability of falsely rejecting the hypothesis was equal or at most equal to a specified constant under little or no assumptions beyond that of random sampling and which were consistent (that is, had error probabilities approaching zero with increasing sample size) in a wide class of alternatives. Classical examples are Smirnov's two-sample test, which is consistent against all alternatives of the two-sample problem, and Wilcoxon's two-sample test, whose domain of consistency is more restricted. These two tests depend only on the rank order of the observations and therefore seem to discard much information contained in the sample. It seemed reasonable to expect that a test which is valid under few assumptions, and especially a rank test, could not be nearly as powerful in a parametric class of distributions as an optimal parametric test for that class. It came therefore as a surprise when it was found that often there are nonparametric tests, including rank tests, which compare favorably with corresponding classical parametric tests.

When dealing with parametric tests, the following should be kept in mind. The assumption that the population distribution is normal, or belongs to almost any other parametric class, is, at best, only approximately satisfied in applications. On the other hand, a parametric test is asymptotically nonparametric in large samples. Thus Fisher's two-sample t-test preserves approximately its nominal significance level whenever the variance of the population distribution is finite (and positive), provided that both samples are large enough, and is consistent against a wide class of non-normal alternatives. This is essentially due to the nonparametric character of the central limit theorem. A strictly non-parametric two-sample test has the obvious advantage over the t-test that it better controls the probability of falsely rejecting the hypothesis when the population is not normal: Its significance level is more robust.

Comparisons between tests are most conveniently done in terms of Pitman's notion of asymptotic relative efficiency. Roughly speaking, if A and A* are two tests of the same significance level α, defined for each sample size, and N and N* are the least sample sizes required by the two tests to achieve power $\beta > \alpha$ against a specified alternative, then the limit $e(A,A^*)$ of the ratio N^*/N as the alternative approaches the hypothesis is the asymptotic relative efficiency of test A with respect to test A* (see [23], [15]). In many cases the asymptotic relative efficiency does not depend on α and β.

Consider the two-sample problem with two independent random samples, $X_1, \ldots, X_m$ and $Y_1, \ldots Y_n$, $P\{X_i < x\} = F(x)$, $P\{Y_i < x\} = G(x)$. The Wilcoxon two-sample test [28] for testing the hypothesis $F = G$ is based on the statistic

$$R_1 + \ldots + R_m, \tag{1}$$

where R_i is the rank of X_i in the combined sample. The asymptotic relative efficiency $e(w,t)$ of this test with respect to the t-test has been investigated for shift alternatives, $G(x) = F(x-\Delta)$. Pitman [24] found that

$$e(w,t) = 12\,\sigma^2 \left(\int f^2 dx\right)^2, \tag{2}$$

where σ^2 is the variance and f the probability density of the distribution F. If F is normal, $e(w,t) = \frac{3}{\pi} = 0.955$, which is close to the maximum possible value 1. Hodges and Lehmann [11] proved the remarkable result that $e(w,t) \geq 0.864$ for all continuous F, and $e(w,t)$ can be arbitrarily large. In this sense the Wilcoxon test is never much worse than the t-test (so far as shift alternatives are concerned) and can be much better: Not only its significance level but also its power is more robust than that of the t-test.

An even more striking result was conjectured by Hodges and Lehmann and proved by Chernoff and I. R. Savage [5] for the normal-scores two-sample test. This test is based on the sum

$$c(R_1) + \ldots + c(R_m), \tag{3}$$

where the R_i are as above and $c(j)$ is the expected value of the j-th smallest among $m + n$ independent, standard normal random variables. The test, first proposed by Fisher and Yates [6], has been known to have asymptotic relative efficiency one with respect to the t-test against shift alternatives with F normal (see [8], section 7.5). Chernoff and Savage showed that the asymptotic relative efficiency is strictly greater than one for shift alternatives with any non-normal F having a density and finite second moment. Analogous results hold for related several-samples tests (Puri [25]) and tests of independence (Bhuchongkul [2]); see also Hajek [9].

For testing whether N independent observations $X_1, \ldots, X_N$ are symmetrically distributed about 0, Wilcoxon [28] proposed the test statistic

$$S_1 R_1' + \ldots + S_N R_N', \tag{4}$$

where S_i is the sign of X_i and R_i' is the rank of $|X_i|$ among $|X_1|, \ldots, |X_N|$. If it is assumed that the X_i have the common distribution function $F(x - \theta)$ and the probability density $f(x) = F'(x)$ is symmetric about

zero, then the asymptotic relative efficiency of this one-sample Wilcoxon test relative to the one-sample t-test is also given by (2) (Pitman [24]). An analogous normal-scores test has been considered by Fraser [7].

The normal-scores two-sample test and the related tests just mentioned are examples of rank tests which are most powerful (or asymptotically most powerful) among all rank tests against specified parametric alternatives close to the hypothesis (see [13], [7]). As noted, in a number of cases they are not only asymptotically as efficient as best parametric against the same alternatives, but even asymptotically more efficient than the latter against other alternatives. Tests that are most powerful, or optimal in other ways, against specified parametric classes of alternatives among all distribution-free tests (not only among rank tests) have been obtained for certain hypotheses by Lehmann and Stein [22]. For the two-sample and some other problems these are tests based on permutations of the observations earlier studied by Fisher, Pitman and others. Their practical usefulness is limited by computational difficulties. It has been shown in [14] that in many cases a test of this kind is asymptotically equivalent, for alternatives close to the hypothesis, to the corresponding best parametric test. Thus the distribution-free test for the two-sample problem which is most powerful unbiased against normal shift alternatives behaves asymptotically like the two-sample t-test. Therefore it suffers from the same lack of robustness as the latter.

It seems to be a common feature of the tests discussed in the preceding paragraph that they are consistent only against restricted classes of alternatives. It would be of some interest to find out whether (to give a specific example) there exist tests which are consistent against all alternatives of the two-sample problem (as the Smirov test is) and at the same time are asymptotically as powerful as a best parametric test against, say, normal shift alternatives.

It would seem too much to expect of a test to be asymptotically as powerful as the best parametric test simultaneously against a wide, nonparametric class of alternatives. That tests of this type sometimes exist has been shown by Stein [26] and Hajek [9]. The idea is to replace the density function, on which a best parametric test depends, by a sample estimate of the density. However, it seems that the sample size has to be excessively large in order that the attractive asymptotic properties of such tests be approximately realized.

2. ROBUST ESTIMATES

The mean $\bar{X}$ of a random sample is an unbiased estimate of the population mean $\int x dF$ and has minimum variance among all unbiased estimates if F is normal. If it is assumed that the population distribution F belongs to a sufficiently extensive class $\mathfrak{F}$ of distributions with finite second moments, then $\bar{X}$ is the only symmetric function of the observations which is an unbiased estimate of $\int x dF$ in the entire class $\mathfrak{F}$; and this can be shown to imply that $\bar{X}$ is the unique minimum variance unbiased

estimate. (For more general results see Fraser [8]). However, this kind of nonparametric optimality of the sample mean is deceptive. It depends in an essential way on the requirement that the estimate be strictly unbiased, which is not very important for most purposes. It may also be reasonable to restrict the class of distributions in such a way that the stated result does not hold. It has been noticed long ago that $\bar{X}$ is a poor estimate when the population distribution has heavy "tails". Modified estimates have been proposed which are obtained by discarding outlying observations, such as the "trimmed" and "winsorized" means (see, for example, Tukey [27] and Anscombe [1]) and a related estimate suggested by Huber [17]. Hodges and Lehmann [12] have introduced estimates which are defined in terms of suitable rank test statistics. Although these estimates are not functions of the ranks (they include generalizations of the sample median), their close relation to functions of ranks suggests that they should be insensitive to extreme outliers. In the case of a single random sample Hodges and Lehmann impose the restriction that the population distribution be symmetric, in which case its mean (when it exists) is equal to its median. These authors also introduced analogous estimates of the shift parameter Δ based on two samples from distributions $F(x)$ and $F(x - \Delta)$, where F need not be symmetric. These estimates are as robust compared with the sample mean estimates as the underlying tests are compared with the t-tests. They play an important role in Lehmann's new approach to analysis of variance described in section 3.

I shall mention only the Hodges-Lehmann estimates based on the one-sample and two-sample Wilcoxon statistics (4) and (1). First consider estimating the median θ from a random sample $Z_1, \ldots, Z_N$ with $P\{Z_i \leq x\} = F(x - \theta)$, where F is continuous and symmetric about 0. The statistic (4) has been shown by Tukey to be equivalent to (in the sense of being a linear function of) the statistic

(5) W_1 = number of pairs (i,j) with $1 \leq i \leq j \leq N$ such that $Z_i + Z_j > 0$

The corresponding Hodges-Lehmann estimate of θ is

(6) $$\hat{\theta} = \text{median}\left\{\frac{Z_i + Z_j}{2},\ 1 \leq i \leq j \leq N\right\}.$$

The event $\hat{\theta} < a$ is equivalent or nearly equivalent to the event $W_1(a) \leq \frac{1}{2}\frac{N(N+1)}{2}$, where $W_1(a)$ is defined as W_1 with Z_i replaced by $Z_i - a$ for all i. Thus the distributions of $\hat{\theta}$ and W_1 are closely related. In [12] it is shown that the asymptotic relative efficiency (in the sense of reciprocal ratio of the asymptotic variances) of $\hat{\theta}$ with respect to the sample mean $\bar{X}$ is given by (2). Thus for symmetric distributions $\hat{\theta}$ enjoys the same robustness property compared with $\bar{X}$ as is the case for the corresponding tests.

For estimating the shift parameter Δ from two independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ with

(7) $\quad P\{X_i \leq x\} = F(x), \qquad P\{Y_j \leq x\} = F(x - \Delta),$

one of the estimates proposed by Lehmann and Hodges is

(8) $\quad \hat{\Delta} = \text{median}\{Y_j - X_i, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n\}$

In the same way as $\hat{\theta}$ is related to W_1, $\hat{\Delta}$ is related to the Wilcoxon statistic (1) in its equivalent form (due to Mann and Whitney),

(9) $\quad W_2 =$ number of pairs (i,j) with $1 \leq i \leq m$, $1 \leq j \leq n$, such that $X_i < Y_j$

Its asymptotic relative efficiency with respect to the difference $\bar{Y} - \bar{X}$ of the sample means is again given by (2).

Results obtained by Bickel [3] indicate that, in terms of robustness, the Hodges-Lehmann estimate $\hat{\theta}$ is superior to the trimmed and Winsorized means and Huber's estimate.

The estimates $\hat{\theta}$ and $\hat{\Delta}$ defined by (6) and (8) have a serious drawback: They take much time to compute. Methods to expedite the computation have been discussed in [12] and by Hφyland [16]. Nevertheless, according to Bickel and Hodges [4], the computation of $\hat{\theta}$ seems to require a number of steps which is "prohibitive" when the sample is large. These authors investigate the alternative estimate

(10) $\quad D = \text{median}\left\{\frac{Z^{(i)} + Z^{(N+1-i)}}{2}, \; 1 \leq i \leq \frac{N+1}{2}\right\},$

where $Z^{(1)} \leq \ldots \leq Z^{(N)}$ is the ordered sample. The estimate D was first proposed by Hodges [10]. It is easier to compute than $\hat{\theta}$, and the results obtained in [4] suggest that the robustness properties of the two estimates are quite similar, although the evidence is incomplete. Unfortunately the form of the asymptotic distribution of D depends on the population distribution F and is difficult to obtain explicitly; the authors have been able to do it only for a rectangular and a Laplacian population.

Rank statistics like W_1 and W_2 can be used to obtain distribution-free confidence intervals for θ and Δ in the situations just considered (see Lehmann [20]). For example, let $D^{(1)} \leq \ldots \leq D^{(mn)}$ denote the ordered mn differences $Y_j - X_i$. In the case of model (7) with F continuous the probability of

(11) $\quad D^{(r)} \leq \Delta \leq D^{(s)}$

is equal to the probability of

$$r \leq mn - W_2 \leq s - 1,$$

evaluated for $\Delta = 0$. This probability does not depend on F. In [20] two alternative definitions of asymptotic relative efficiency of confidence

intervals are considered and it is shown that with either definition the asymptotic relative efficiency of the confidence interval (11) with respect to the classical confidence interval based on the two-sample t-statistic is again given by (2).

3. ROBUST ANALYSIS OF VARIANCE

Until recently nonparametric procedures have been available only for rather special statistical problems. In particular, no extensive, unified body of nonparametric methods comparable to the classical analysis of variance had been developed that would be adaptable to a great variety of testing, estimation and multiple decision problems arising in statistical practice. Of course, classical analysis of variance is asymptotically nonparametric, but it lacks robustness. Early attempts to use nonparametric techniques in analysis of variance (which are mentioned in [19]) have not been very satisfactory, due to low efficiency or lack of versatility. It was E. L. Lehmann who, about 1963, initiated the development of an asymptotically nonparametric and relatively robust analogue of traditional analysis of variance. In this section some of the main aspects of this work are described.

First consider the model according to which the observable random variables X_i are of the form

$$(12) \qquad X_{j\alpha} = \xi_i + U_{i\alpha}, \quad \alpha = 1,\ldots, n_i \; ; \; i = 1,\ldots,c,$$

where the $U_{i\alpha}$ are mutually independent with common continuous distribution function F, and $\xi_1,\ldots,\xi_c$ are unknown constants. One of the objects is to estimate "contrasts"

$$\sum_{i=1}^{c} a_i \xi_i \qquad \left(\sum_{i=1}^{c} a_i = 0\right).$$

In [18] Lehmann first considers estimating the difference $\xi_i - \xi_j$ by

$$(13) \qquad Y_{ij} = \text{median } \{ X_{i\alpha} - X_{j\beta}, \; 1 \le \alpha \le n_i \; ; \; 1 \le \beta \le n_j \},$$

which corresponds to $\hat{\Delta}$ in (8). Since a contrast can be represented as a linear combination of differences $\xi_i - \xi_j$, it could be estimated by the corresponding combination of the Y_{ij}. However, these estimates are not unique. For example, $Y_{13} + Y_{24}$ and $Y_{14} + Y_{23}$ are two different estimates of the contrast $\xi_1 + \xi_2 - \xi_3 - \xi_4$. To avoid this ambiguity, Lehmann proposes to replace the raw estimates Y_{ij} by the adjusted estimates

$$(14) \qquad Z_{ij} = Y_{i.} - Y_{j.}, \quad \text{where} \quad Y_{i.} = \frac{1}{c}\sum_{j=1}^{c} Y_{ij}, \; (Y_{ii} = 0).$$

The estimate of a contrast $\Sigma a_i \xi_i$ in terms of the Z's is unique and can be written as $\Sigma a_i Z_{ic}$ or $\Sigma a_i Y_{i.}$.

The classical estimate of $\xi_i - \xi_j$ is

$$T_{ij} = X_{i.} - X_{j.}, \quad \text{where} \quad X_{i.} = \frac{1}{n_i} \sum_{\alpha=1}^{n_i} X_{i\alpha} \tag{15}$$

If $N = n_1 + \ldots + n_c$ tends to infinity such that $n_i/N \to \rho_i > 0$ for all i, then the random variables

$$N^{\frac{1}{2}} (Z_{ij} - \xi_i + \xi_j), \qquad 1 \leq i < j \leq c,$$

and

$$N^{\frac{1}{2}} (T_{ij} - \xi_i + \xi_j), \qquad 1 \leq i < j \leq c,$$

have both jointly normal limit distributions with zero means and respective covariance matrices

$$\Sigma_Z = \tau^2 A, \qquad \Sigma_T = \sigma^2 A,$$

where

$$\tau^2 = 1/\{12 (\int f^2 dx)^2\},$$

σ^2 is the variance of F, and the matrix A depends only on the ρ_i.

This implies that the asymptotic relative efficiency of the Z-estimate, $\Sigma a_i Z_{ic}$, of an arbitrary contrast $\Sigma a_i \xi_i$ with respect to the classical estimate $\Sigma a_i X_{i.}$ is σ^2/τ^2, which is identical with (2).

Now consider a linear hypothesis H which specifies that the vector $(\xi_1, \ldots, \xi_c)$ lies in a (c-r)-dimensional linear subspace Π of c-space. If F is normal with known σ, the classical test statistic is

$$\Sigma n_i (X_{i.} - \tilde{\xi}_i)^2/\sigma^2, \tag{16}$$

where $(\tilde{\xi}_1, \ldots, \tilde{\xi}_c)$ is the projection of $(X_{1.}, \ldots, X_{c.})$ on Π. Now $\tilde{\xi}_i = L_i (X_{i.}, \ldots, X_{c.})$ is a linear function of $X_{1.}, \ldots, X_{c.}$. If we let

$$\xi_i' = L_i (Y_{1.}, \ldots, Y_{c.}),$$

it follows from the preceding (see [19]) that

$$\Sigma n_i (Y_i - \xi_i')^2/\tau^2 \tag{17}$$

has the same limit distribution as (16) when H is true, which is that of χ^2 with r degrees of freedom. Since τ^2 depends on the unknown distribution F, the random variable (17) cannot be used as a test statistic.

However, if t_N^2 is any consistent estimate of τ^2, then the statistic

$$(18) \qquad \Sigma n_i \, (Y_{i.} - \xi_i')^2 / t_N^2$$

has the same limit distribution as (17) and provides an asymptotically nonparametric test. Estimates of τ^2 are proposed in [19] and [20]. An alternative test statistic is also discussed in [19]. These statistics can also be used to construct asymptotically distribution-free confidence intervals for individual contrasts and simultaneous confidence intervals for all contrasts.

In [21] Lehmann considers the model with one observation per cell,

$$(19) \quad X_{i\alpha} = \nu + \xi_i + \mu_\alpha + U_{i\alpha}, \quad i = 1,\ldots,c; \quad \alpha = 1,\ldots, N$$

with $\Sigma \xi_i = \Sigma \mu_\alpha = 0$, where the ξ_i are the parameters of interest. The $U_{i\alpha}$ are assumed to be mutually independent and identically distributed. In this case $\xi_i - \xi_j$ can be estimated by

$$(20) \quad Y_{ij}^* = \text{median} \left\{ \frac{X_{i\alpha} - X_{j\alpha} + X_{i\beta} - X_{j\beta}}{2}, \quad 1 \leq \alpha \leq \beta \leq N \right\},$$

which corresponds to $\hat{\theta}$ in (6), or by the adjusted estimate

$$(21) \quad Z_{ij}^* = Y_{i.}^* - Y_{j.}^*, \qquad Y_{i.}^* = \frac{1}{c} \sum_{j=1}^{c} Y_{ij}^*.$$

Whereas the Y - and Z - estimates for model (12) are asymptotically equivalent, it is shown in [21] that this is not true of the Y^*- and Z^*-estimates and that the latter are asymptotically (slightly) more efficient. The Z^*- estimates and the corresponding tests have asymptotic properties analogous to those for model (12).

As noted in section 2, estimates like Y_{ij} and Y_{ij}^* require much time to compute. Lehmann's approach does not depend on this particular choice of estimates, and analogous methods based on statistics such as D in (10) may prove to be more suitable for applications.

REFERENCES

[1] Anscombe, F. J. (1960), Rejection of outliers. Technometrics 2, 123-147.

[2] Bhuchongkul, S. (1964). A class of nonparametric tests for independence in bivariate populations. Ann. Math. Statist. 35, 138-149.

[3] Bickel, Peter J. (1965). On some robust estimates of location. Ann. Math. Statist. 36, 847-858.

[4] Bickel, P. J., and Hodges, J. L., Jr. (1967). The asymptotic theory of Galton's test and a related simple estimate of location. Ann. Math. Statist. 38, 73-89.

[5] Chernoff, Herman, and Savage, I. Richard (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. Ann. Math. Statist. 29, 972-994.
[6] Fisher, Ronald A., and Yates, Frank (1949). Statistical Tables for Biological, Agricultural and Medical Research (3rd Ed.). Hafner, New York.
[7] Fraser, D. A. S. (1957). Most powerful rank-type tests. Ann. Math. Statist. 28, 1040-1043.
[8] Fraser, D. A. S. (1957). Nonparametric Methods in Statistics. Wiley, New York.
[9] Hajek, Jaroslav (1962). Asymptotically most powerful rank-order tests. Ann. Math. Statist. 33, 1124-1147.
[10] Hodges, J. L., Jr. (1955). Galton's rank-order test. Biometrika 42, 261-262.
[11] Hodges, J. L., Jr., and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. Ann. Math. Statist. 27, 324-335.
[12] Hodges, J. L., Jr., and Lehmann, E. L. (1963). Estimates of location based on rank tests. Ann. Math. Statist. 34, 598-611.
[13] Hoeffding, Wassily (1951). "Optimum" nonparametric tests. Proc. Second Berkeley Symp. on Math. Statistics and Probability, Univ. of California Press, Berkeley and Los Angeles, pp. 83-92.
[14] Hoeffding, Wassily (1952). The large-sample power of tests based on permutations of observations. Ann. Math. Statist. 23, 169-192.
[15] Hoeffding, Wassily, and Rosenblatt, Joan Raup (1955). The efficiency of tests. Ann. Math. Statist. 26, 52-63.
[16] Høyland, Arnljot (1964). Numerical evaluation of Hodges-Lehmann estimates. Det Kong. Norske Videnskabers Selskabs Forh. 37, 42-47.
[17] Huber, Peter J. (1964). Robust estimation of a location parameter. Ann. Math. Stat. 35, 73-101.
[18] Lehmann, E. L. (1963). Robust estimation in analysis of variance. Ann. Math. Statist. 34, 957-966.
[19] Lehmann, E. L. (1963). Asymptotically nonparametric inference: An alternative approach to linear models. Ann. Math. Statist. 34, 1494-1506.
[20] Lehmann, E. L. (1963). Nonparametric confidence intervals for a shift parameter. Ann. Math. Statist. 34, 1507-1512.
[21] Lehmann, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. Ann. Math. Statist. 35, 726-734.
[22] Lehmann, E. L., and Stein, C. (1949). On the theory of some non-parametric hypotheses. Ann. Math. Statist. 20, 28-45.
[23] Noether, Gottfried E. (1955). On a theorem of Pitman. Ann. Math. Statist. 26, 64-67.
[24] Pitman, E. J. G. (1948). Notes on Non-Parametric Statistical Inference. (Unpublished).

[25] Puri, Madan Lal (1964). Asymptotic efficiency of a class of c-sample tests. Ann. Math. Statist. 35, 102-121.

[26] Stein, Charles (1956). Efficient nonparametric testing and estimation. Proc. Third Berkeley Symp. on Math. Statistics and Probability, Univ. of California Press, Berkeley and Los Ang les.

[27] Tukey, John W. (1960). A survey of sampling from contaminated distributions. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford Univ. Press, Stanford, California.

[28] Wilcoxon, Frank (1945). Individual comparisons by ranking methods. Biometrics Bull. 1, 80-83.