Erich Lehmann on rank methods

Willem R. van Zwet

Though statistical methods based on ranks, such as the sign test, have a history going back for centuries, it was the introduction of Wilcoxon's tests in Wilcoxon (1945) that started an avalanche of research on this subject. It is worth noting that Wilcoxon needed only a little over three pages introducing and discussing both his two-sample test and his one-sample test for symmetry, to set off a research effort that would fill thousands of journal pages. Almost from the beginning Erich Lehmann played a major role in the development of this subject. Eight of his papers will be discussed here, of which two are joint with J. L. Hodges, Jr. and one with C. Stein. The full content of these papers will not be described here. Instead, the focus will be on the main contributions of each paper to our present knowledge of rank tests. It all adds up to what is more or less a history of the subject. For ease of presentation attention will be focused on the two-sample problem. When needed, appropriate reference will be made to other models that are dealt with in a particular paper.

Let us introduce the necessary notation. The two samples are denoted by $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ and for the combined sample of size $N = m + n$, we write $Z_1, \ldots, Z_N = X_1, \ldots, X_m, Y_1, \ldots, Y_n$. We assume that $Z_1, \ldots, Z_N$ are independent random variables, that $X_1, X_2, \ldots, X_m$ are identically distributed (i.d.) with common distribution function F, and that $Y_1, Y_2, \ldots, Y_n$ are i.d. with distribution function G. We suppose that F and G are continuous so that there are no equal values among $Z_1, \ldots, Z_N$. Let $Z_{(1)} < \cdots < Z_{(N)}$ denote the ordered sequence of $Z_1, \ldots, Z_N$, and R_j the rank of Y_j in this sequence, thus $Y_j = Z_{(R_j)}$. Wilcoxon's two sample test for $H_0 : F = G$ against $H_1 : F > G$ rejects H_0 for large values of the rank sum $W = \sum_{1 \le j \le n} R_j$. Here $F > G$ should be interpreted as $F(x) \ge G(x)$ for all x with strict inequality for at least one value of x, and hence on an interval. More generally, a rank test for this testing problem rejects H_0 for large values of $T = \sum_{1 \le j \le n} a(R_j)$, for an increasing sequence of scores $a(r)$. Obviously, under H_0,

W.R. van Zwet
Mathematical Institute, University of Leiden, P.O.Box 9512, 2300 RA Leiden, The Netherlands
e-mail: vanzwet@math.leidenuniv.nl

J. Rojo (ed.), *Selected Works of E. L. Lehmann*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1412-4_30, © Springer Science+Business Media, LLC 2012

the vector $R = (R_1, \ldots, R_n)$ is distributed as a random sequence of length n drawn without replacement from the set $\{1, 2, \ldots, N\}$ and hence the distribution of W under H_0 is easy to find and independent of the underlying common distribution function F. In other words, rank tests are distribution-free and hence similar, in the sense that the probability of an error of the first kind is constant on the hypothesis H_0. For a given significance level α the critical value for the test based on W is therefore easy to determine.

Perhaps this is the place for a word about terminology. In the classical parametric model unknown distributions have a given parametric form with only a finite dimensional parameter left unspecified. The null hypothesis then sets restrictions on this finite dimensional parameter. In the model discussed above the unknown distributions F and G are left unspecified and under H_0 the common distribution $F = G$ is still unspecified. Neither the general model nor the null hypothesis can be described in terms of finite dimensional parameters, which is why such models and hypotheses are called *nonparametric*. This has led statisticians to call rank tests and other similar tests for such hypotheses *nonparametric tests*, which is clearly a misnomer. There is nothing parametric or nonparametric about a statistical test. The important property of such a test is simply that it deals with such large infinitely dimensional hypotheses by using test statistics that are distribution-free, in the sense that their distribution remains the same throughout the hypothesis. Hence *distribution-free tests* is clearly a more appropriate name for these tests.

Returning to the two-sample rank tests, the main problem is of course to get some idea of the power of these tests against simple alternatives (F, G). The case where F and G are normal distributions differing only in location is of course of special interest. This is the case where Student's test comes into its own and it would be of interest to compare the powers. Initially, the common view of rank tests was that it was a nice idea, but that restricting attention to ranks instead of the variables themselves would probably result in a very serious loss of power. Even Wilcoxon himself didn't sound too optimistic about this, even though he felt that his symmetry test would do better than the sign test.

1 On the theory of some non-parametric hypotheses

The first seven of Lehmann's papers discussed here were written between 1949 and 1963 and in these papers one can trace the progress of the research on this topic almost from its very beginning. These papers will therefore be discussed in chronological order. In 'On the theory of some non-parametric hypotheses' by Lehmann and Stein (1949) a first glimmer is shed on the problem, even though the paper deals primarily with nonparametric hypotheses rather than distribution-free tests. First of all the authors point out that instead of testing H_0, it is often more relevant to test the extended null hypothesis H_0^* that $Z_1, \ldots, Z_N$ are exchangeable, i.e. that their joint distribution is invariant under permutations of the variables. Of course for rank tests this makes no difference as the distribution of the ranks remains the same under the extended hypothesis H_0^*. For general tests, however, optimality properties of a test for H_0^* do not necessarily correspond to the same property of this test among the larger class of tests for the restricted null hypothesis H_0. However, the authors prove that this correspondence does exist if the tests for H_0 are restricted to distribution-free or similar tests. But this means that optimal tests for H_0^* can only be beaten by tests that are not even similar for the restricted hypothesis H_0. One may hope that this means that not much power is lost by using a test for H_0^* — such as a rank test — provided it is in some sense optimal for testing H_0^*. The paper further discusses the same phenomenon for more general pairs of null hypotheses and derives most powerful and most stringent tests for various testing problems.

2 Consistency and unbiasedness of certain nonparametric tests

In the second paper 'Consistency and unbiasedness of certain nonparametric tests' (Lehmann (1951)) the attack on the power of Wilcoxon's test is well under way. Mann and Whitney had proved that Wilcoxon's test is consistent, i.e. its power against a simple alternative with $F > G$ tends to 1 as $m, n \to \infty$. By noting that $F > G$ implies that there exists a function $h(x) > x$ such that $h(X_i)$ has distribution G, and that the test statistic is monotone, it is shown that the test is also unbiased against such alternatives, i.e. its power exceeds the level of significance α. The same proof shows that the power against alternatives $F < G$ is below α and hence the test may also be applied for testing the hypothesis $F \leq G$ against the alternative $F > G$. Unbiasedness and consistency are also proved for tests for the k-sample problem, tests for independence and the symmetry hypothesis.

Because W can be written as

$$W = \sum_{1 \leq i \leq m} \sum_{1 \leq j \leq n} I\{X_i < Y_j\} + \frac{n(n+1)}{2},$$

it is a two-sample U-statistic. Extending Hoeffding's results on one-sample U-statistics (cf. Hoeffding (1948)), asymptotic normality of W under alternatives is proved as m and n tend to infinity at the same rate. The news of this major advance traveled around the world fast. Already before Lehmann's paper appeared, my own teacher David van Dantzig reported to the Royal Netherlands Academy of Sciences that '... Mr. A.M. Mood kindly informed me that Mr. E. Lehmann has proved the important result that the Wilcoxon test criterion is asymptotically normally distributed even when the null hypothesis is not true.' (van Dantzig (1951))

3 The power of rank tests

The problem of power computation is attacked in a novel and imaginative way in the third paper 'The power of rank tests' (Lehmann (1953)). This paper is one of my all-time favorites and I have had the pleasure to speak about it in Berkeley on the occasion of Erich's 80th birthday in 1997. Power computations for alternatives such as shift — i.e. $G(x) = F(x - a)$ for $a > 0$ – look rather forbidding and the result will also depend on F as well as a. Lehmann argued that since one probably has only a vague idea what alternatives would be plausible in any particular case, one might as well choose a mathematically convenient alternative hypothesis. His suggestion is to choose a distribution function K on $(0, 1)$ with density k, take $G(x) = K(F(x))$ and consider the alternative hypothesis $H_1^* : (F, K(F))$ for varying F. Since the ranks are invariant under the monotone transformation $x \mapsto F(x)$ of the Z's, this has the obvious advantage that that the distribution of the ranks under this alternative is the same as for the model (U, K) where the X_i have the uniform distribution function U on $(0, 1)$ and the Y_j have distribution function K. Hence the power is constant on this alternative H_1^*. Application of a result of Hoeffding (1951) yields that under H_1^* and independent of F,

$$P(R_1 = r_1, \dots, R_n = r_n \mid H_1^*) = \frac{m!n!}{(m+n)!}\mathbb{E}\left(\prod_{1 \le j \le n} k(U_{r_j:N})\right), \tag{3.1}$$

where $U_{i:N}$ denotes the i-th order statistic of a sample of size N from the uniform distribution U on $(0, 1)$. If $K(v) < v$ for $0 < v < 1$ we have $G < F$ and H_1^* is a proper subset of the full alternative $H_1 : F > G$. Obvious examples are $K(v) = v^a$ for $a > 1$. Especially for integer a, the right-hand side of (3.1) and consequently the exact power of Wilcoxon's test against H_1^* is easy to compute. One can now traverse the general alternative hypothesis $H_1 : F > G$ almost at will and compute the power of Wilcoxon's test for small samples. In the paper we find graphs of the alternatives that are dealt with and numerical values of the corresponding power. Also moments of W under H_1^* are easy to compute and together with the asymptotic normality proved in the previous paper, this yields large sample approximations for the distribution of W under alternatives, and hence for the power of W.

Finally (3.1) allows us to derive most powerful tests against simple alternatives. Under H_0, the probability of any set of ranks equals $m!n!/(m+n)!$ independent of F, and hence the Neyman-Pearson lemma tells us to reject H_0 for large values of the expected value on the right in (3.1). By taking $K(v) = K_p(v) = (1-p)v + pv^2$ and maximizing the derivative of the power against this alternative at $p = 0$, one finds the locally most powerful rank test which happens to be Wilcoxon's test. Other testing problems are also considered.

4 The efficiency of some nonparametric competitors of the t-test

The fourth paper ‘The efficiency of some nonparametric competitors of the t-test’ (Hodges and Lehmann (1956)) must be the one that caught the most attention, and the number 0.864 is probably engraved in many statisticians’ minds. Consider the problem of testing $G = F$ against $G(x) = F(x - a)$, $a > 0$ for a given F with density f and variance σ^2. Pitman (1949) showed that for this problem the asymptotic relative efficiency (ARE) of Wilcoxon’s test relative to the t-test equals $e_{W,t} = 12\sigma^2 \left(\int f^2(x)dx\right)^2$. It is easy to see that $e_{W,t}$ can be arbitrarily large for certain densities, indicating a superior performance of Wilcoxon’s test. On the other hand the efficiency is less than 1 if $F = \Phi$, the standard normal distribution function, because asymptotically Student’s test can not be beaten on its own turf. However, for the normal case $e_{W,t} = 3/\pi \approx 0.95$, which is still fairly close to 1. Lehmann asked what the worst performance of Wilcoxon’s test relative to Student’s test could be, i.e. what the minimum of $e_{W,t}$ would be as a function of f. He found that this equals $108/125 = 0.864$. This means that for large samples you never throw away more than 14 % of your data if you test for shift by using Wilcoxon’s test instead of Student’s, as was the general custom. For that relatively small price you buy a test that is valid regardless of the underlying distribution. This simple argument put a definite end to the idea that the cost of using ranks would be prohibitive.

5 Rank methods for combination of independent experiments in analysis of variance

In the paper ‘Rank methods for combination of independent experiments in analysis of variance’ by Hodges and Lehmann (1962) the authors turn to generalizations of Wilcoxon’s two-sample test to randomized block designs. In this model the N available subjects are divided into k blocks, with N_i subjects in block i and hence $\sum_{1 \leq i \leq k} N_i = N$. The idea is that within each block subjects are believed to be more homogeneous than in the entire group. In block i, $m_i > 0$ and $n_i > 0$ subjects are randomly assigned to treatments A and B respectively, so $m_i + n_i = N_i$ and the two treatments correspond to two samples of sizes m_i and n_i. The problem is to test the hypothesis that there is no difference between the treatments, against shift alternatives where in each block the response to treatment B is stochastically larger than that to treatment A. One possible test statistic for this problem is a linear combination of Wilcoxon’s two-sample statistics W_i for blocks $i = 1, \ldots, k$, with weights that are in some sense optimal. If the underlying distributions would be known to be normal with a common variance, an optimal statistic would be a linear combination of the differences of the sample means in blocks $1, \ldots, k$. Roughly speaking, for the normal distribution it turns out that if both the number of blocks k and the block sizes N_i tend to infinity, the ARE of this rank test with respect to the normal test is $3/\pi$. This equals the ARE of

Wilcoxon's two-sample test to Student's two-sample test. However, if the block sizes N_i equal 2 for all i and the number of blocks k tends to infinity, then this ARE equals $2/\pi$ which is the ARE of the sign test to Student's test. This is no surprise as $N_i = 2$ implies $m_i = n_i = 1$, and for two samples of size 1, Wilcoxon's test is equivalent to the sign test applied to the difference of the (single) responses to treatments B and A.

Another possibility is to align the blocks by subtracting the average response of all N_i subjects in block i from these N_i responses in that block. Having thus aligned the different blocks, one pools all N subjects and carries out Wilcoxon's two-sample test on the two samples of sizes $\sum_{1 \le i \le k} m_i$ and $\sum_{1 \le i \le k} n_i$ of subjects having received treatment A or B. The test is carried out as a conditional test given the division of the set of ranks $\{1, 2, \dots, N\}$ of the N combined aligned responses over the k blocks. Critical values and powers are computed and asymptotic normality under the hypothesis is proved. If the block sizes N_i equal 2 for all i, then for the normal case the ARE of the aligned block test equals $3/\pi$ which is the ARE of Wilcoxon's test with respect to Student's test.

6 Robust estimation in analysis of variance

The sixth paper 'Robust estimation in analysis of variance' (Lehmann (1963)) is a follow-up on 'Estimates of location based on rank tests' (Hodges and Lehmann (1963)) which appeared in the immediately preceding issue of *Ann. Math. Statist.*. In this earlier paper the authors consider the two-sample shift model with $G(x) = F(x - \xi)$ and propose the median of the mn differences $(Y_j - X_i)$ as a robust estimator of ξ. The ARE of this estimator to the classical estimator which is the difference of the sample means is the same as the ARE of Wilcoxon's test to Student's test.

Now the set-up is generalized to the c-sample problem where for $i = 1, \dots, c$, the i-th sample of size n_i consists of observations $X_{i,\alpha} = \xi_i + U_{i,\alpha}, \alpha = 1, \dots, n_i$. The $U_{i,\alpha}$ are independent with median 0. The Hodges-Lehmann estimator of $(\xi_i - \xi_j)$ is given by $Y_{i,j} = \text{med}(X_{i,\alpha} - X_{j,\beta})$, the median of the $n_i n_j$ differences $(X_{i,\alpha} - X_{j,\beta})$ for $\alpha = 1, \dots, n_i$, $\beta = 1, \dots, n_j$. There is an amusing problem with this definition. It is not compatible in the sense that the estimator $Y_{i,k}$ of $(\xi_i - \xi_k)$ is generally not equal to the sum of the estimators $Y_{i,j}$ of $(\xi_i - \xi_j)$ and $Y_{j,k}$ of $(\xi_j - \xi_k)$. This becomes a nuisance when estimating general contrasts of the form $\sum a_i \xi_i$ with $\sum a_i = 0$ because there are many linear combinations of the $Y_{i,j}$ that can be used to estimate this contrast.

This problem is solved by estimating $(\xi_i - \xi_j)$ by minimizing

$$\sum_{i \neq j} \left(Y_{i,j} - (\xi_i - \xi_j)\right)^2,$$

which yields the difference of the two sample means $Z_{i,j} = Y_{i\cdot} - Y_{j\cdot}$ as an estimate. Now if a contrast is written in two different ways $\sum_{i,j} a_{i,j}(\xi_i - \xi_j) = \sum_{i,j} b_{i,j}(\xi_i - \xi_j)$ identically in $\xi_1, \dots, \xi_c$, then obviously $\sum_{i,j} a_{i,j} Z_{i,j} = \sum_{i,j} a_{i,j}(Y_{i\cdot} - Y_{j\cdot}) =$

$\sum_{i,j} b_{i,j}(Y_{i\cdot} - Y_{j\cdot}) = \sum_{i,j} b_{i,j} Z_{i,j}$, so the estimator is unique. Joint asymptotic normality of appropriately normalized $Y_{i,j}$ is proved and with the same normalization, the $Z_{i,j}$ have the same limit distribution.

7 A class of selection procedures based on ranks

The seventh paper in this sequence 'A class of selection procedures based on ranks' (Lehmann (1963)) is devoted to the use of ranks in yet another area of statistics. Suppose we have samples of size n from each of c populations $\Pi_1, \ldots, \Pi_c$ which differ only in location. So for $i = 1, \ldots, c$ and $j = 1, \ldots, n$, the random variables $X_{i,j}$ are independent and $P(X_{i,j} \leq x) = F(x - \theta_i)$. One wishes to select the *best* population which is the one with the largest value of the parameter θ. The classical procedure is the *means procedure* that selects the population Π_i with largest mean $X_{i\cdot} = n^{-1} \sum_{1 \leq j \leq n} X_{i,j}$. If F is a normal distribution function this is the natural thing to do.

The performance of different selection procedures is compared as follows. Suppose that for $i = 1, 2, \ldots, c - 1$, $\theta_i < \theta_c$ so that population Π_c is uniquely the best. A population Π_i is called *good* if $\theta_i \geq \theta_c - \Delta$ for some $\Delta > 0$. To ensure good behavior of a selection procedure one may require that

$$P(\text{selected population is good}) \geq \gamma, \tag{7.1}$$

for some $\gamma > 0$. To compare the asymptotic performance of two procedures A and B we consider the ratio $e = n_B/n_A$ of smallest sample sizes n_A and n_B needed to satisfy (7.1) for a given Δ and γ with these two procedures. The limit of e as $\Delta \to 0$ is the ARE of procedure A with respect to procedure B.

Let us now consider procedures based on ranks. Let $R_{i,j}$ denote the rank of $X_{i,j}$ among the combined samples $\{X_{i,j} : i = 1, \ldots, c, j = 1, \ldots, n\}$. Define $V_i = \sum_{1 \leq j \leq n} h(R_{i,j})$, for $i = 1, \ldots, c$ and an increasing function h, and select the population Π_i with the largest V_i. For $h(x) \equiv x$ we are dealing with a rank sum test which is similar to Wilcoxon's test. It is shown that in the case of normal F the ARE of this test to the means procedure equals $3/\pi$ which is the ARE of Wilcoxon's test to Student's test. Also the ARE in never smaller than 0.864 for any F and the entire theory for two-sample tests repeats itself is this very different situation.

8 Parametrics versus nonparametrics: two alternative methodologies

After this brief discussion of (a part of) Lehmann's work on rank tests it will be clear that he was one of the driving forces of this development. A masterful picture of what was achieved is given by Lehmann himself in the last paper of this sequence. 'Parametrics versus nonparametrics: two alternative methodologies' (Lehmann (2009)). Here we see Erich Lehmann in a role that we all know so well from his books, that of the superb expositor. In 9 pages he tells us all there is to know about this subject and it makes no sense to produce a further condensed version here. Just read this paper!

References

[1] van Dantzig, D. (1951). On the consistency and the power of Wilcoxon's two sample test. *Proc. Royal Netherlands Acad. Sci., Series A* **54** 1–8.

[2] 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.

[3] Hodges, J. L. Jr. and Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Statist.* **33** 482–497.

[4] Hodges, J. L. Jr. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

[5] Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

[6] Hoeffding, W. (1951). Optimum nonparametric tests. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 83–92.

[7] Lehmann, E. L. and Stein, C. (1949). On the theory of some non-parametric hypotheses. *Ann. Math. Statist.* **20** 28–45.

[8] Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Ann. Math. Statist.* **22** 165–179.

[9] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24** 23–43.

[10] Lehmann, E. L. (1963). A class of selection procedures based on ranks. *Math. Annalen* **150** 268–275.

[11] Lehmann, E. L. (1963). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34** 957–966.

[12] Lehmann, E. L. (2009). Parametrics versus nonparametrics: two alternative methodologies (with discussion). *J. Nonparametr. Stat.* **21** 397–405.

[13] Pitman, E. J. G. (1949). *Lecture notes on nonparametric statistical inference.* Columbia University.

[14] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1** 80–83.