TESTING WHETHER INDEPENDENT TREATMENT GROUPS HAVE EQUAL MEDIANS

RAND R. WILCOX
UNIVERSITY OF SOUTHERN CALIFORNIA

The paper suggests new methods for comparing the medians corresponding to independent treatment groups. The procedures are based on the Harrell-Davis estimator in conjunction with a slight modification and extension of the bootstrap calibration technique suggested by Loh. Alternatives to the Harrell-Davis estimator are briefly discussed. For the special case of two treatment groups, the proposed procedure always had more power than the Fligner-Rust solution, as well as the procedure examined by Wilcox and Charlin. Included is an illustration, using real data, that comparing medians, rather than means, can yield a substantially different conclusion as to whether two distributions differ in terms of some measure of central location.

Key words: L-statistics, Harrell-Davis estimator, bootstrap, kernel density estimates, smoothing.

1. Introduction

Researchers in psychology typically compare treatment groups in terms of their means. However, experience with real data indicates that distributions in psychology can be highly skewed (e.g., Micceri, 1989; Wilcox, in press), and this suggests that some other measure of location might be used to compare groups. A well-known alternative to the mean is the median. One argument in favor of using the median is that the median is closer to the bulk of the data when a distribution is skewed. As illustrated in section 5, a practical consideration is that inferences about medians can yield substantially different conclusions than inferences based on means.

Consider J independent treatment groups with corresponding medians \( \theta_1, \ldots, \theta_J \). The median of the random variable, \( X \), having distribution \( F \), is \( \theta = F^{-1}(0.5) = \inf \{ x \mid F(x) \geq 0.5 \} \). The paper considers the problem of testing

\[
H_0: \theta_1 = \cdots = \theta_J.
\]

Perhaps the best known procedure for testing (1), when \( J = 2 \), is Mood's (1954) median test. However, under general circumstances, Mood's test is neither distribution-free nor asymptotically distribution-free for testing (1), and the actual Type I error probability can exceed the nominal level by an unacceptable amount (Fligner & Rust, 1982). Conservative tests for equal medians are discussed by Hettmansperger (1973), as well as Hettmansperger and Malin (1975), for \( J = 2 \). If distributions are symmetric, methods described by Potthoff (1963), and Fligner and Policello (1981) can be used to test (1), but under asymmetry these procedures are unsatisfactory (Fligner & Rust, 1982).

Fligner and Rust suggest a test of (1), again assuming \( J = 2 \), that is distribution-free when the distributions are identical. However, for unequal sample sizes, their test can be too liberal (Wilcox & Charlin, 1986). In fact, it is possible to have the usual sample medians equal, yet the Fligner-Rust procedure rejects (1). For equal sample sizes,
however, all indications are that the Fligner-Rust procedure provides satisfactory control over Type I errors.

Let $X_1, \ldots, X_n$ be a random sample, and let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the corresponding order statistics. The usual sample median will be denoted by $\hat{\theta}$, which is equal to $X_{((n+1)/2)}$ if $n$ is odd, and $(X_{(n/2)} + X_{((n/2)+1)})/2$ if $n$ is even. The corresponding estimate for the $j$th treatment group is $\hat{\theta}_j$. The standard error of $\hat{\theta}_j$ can be estimated as described by Martiz and Jarrett (1978) with, say, $V_j$. The expression for $V_j$ is rather involved, and it does not play a central role here, so the computational details are omitted. (For some related results, see Hall & Martin, 1988). A simple test of (1), for $J = 2$, is to compute

$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{(V_1 + V_2)^{1/2}},$$

(2)

and assume that $Z$ has a standard normal distribution. This strategy will be called method MJ, and is motivated by results in McKeane and Schrader (1984). It appears to give reasonable control over Type I errors (Wilcox & Charlin, 1986), and generally, but not always, Wilcox and Charlin found that (2) had slightly more power than the Fligner-Rust technique.

In recent years, several alternatives to $\hat{\theta}$, as estimators of the median, have been proposed. Some of these can provide a more accurate estimate of $\theta$, in terms of mean squared error, for a reasonably wide range of situations. This suggests that if these newer estimates of $\theta$ could be incorporated into a test of (1) that controls Type I errors, the power of such a test might be higher than the test based on (2) or the Fligner-Rust technique. The goal in this paper is to describe a test of (1) that has this property, at least for the situations considered in section 4.

Section 2 reviews some alternative estimates of $\theta$. Section 3 describes the new test of (1) for $J = 2$ groups. Section 4 reports some simulations, and using real data, section 5 illustrates the new procedure. Section 6 suggests an extension of the test to $J > 2$; multiple comparison procedures are briefly discussed in section 7.

2. Methods for Estimating Quantiles

This section reviews some recent results on estimating quantiles that include, of course, estimates of the median as a special case. The estimate of $F^{-1}(p)$, $0 < p < 1$, is typically taken to be

$$S_G(p) = (1 - g)X_{(j)} + gX_{(j+1)},$$

where $(n + 1)p = j + g$, and $j$ is the integral part of $(n + 1)p$. As noted above, in terms of mean squared error, it is usually possible to find a more accurate estimate of $\theta$, and more generally, a more accurate estimate of $F^{-1}(p)$. One such estimate of $F^{-1}(p)$ is

$$S_{HD}(p) = \Sigma W_i X_{(i)},$$

where $W_i = I_{\bar{i}n}\{p(n + 1), (1 -p)(n+1)\} - I_{(i-1)/n}\{p(n+1), (1 - p)(n + 1)\}$, and $I_x(a, b)$ is the incomplete beta function (Harrell & Davis, 1982). The corresponding estimate of $\theta$ is denoted by $\theta_{HD}$. Computer programs are widely available for calculating $I_x(a, b)$ such as IMSL subroutine MDBETA, or the algorithm of Majumdar and Bhattacharjee (1973). The motivation for $S_{HD}$ is that it estimates $E(X_{((n + 1)p)})$ which converges to $F^{-1}(p)$. Note that $S_{HD}(p)$ is an L-statistic (i.e., it is a linear combination of the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$). Yoshizawa, Sen, and Davis (1985) show that