THE TWO-SAMPLE SCALE PROBLEM WHEN LOCATIONS ARE UNKNOWN[1]

By M. Raghavachari

Ministry of Commerce, Government of India and University of California, Berkeley

1. Introduction. Let $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ be independent observations from two distributions with cdf $F(x - \nu)/\sigma$ and $F(x - \eta)/\tau$ respectively, where ν and η are location parameters (for example medians), σ and τ are scale parameters and F is an absolutely continuous distribution function. The problem is to test the hypothesis $H: \sigma = \tau$ against one- or two-sided alternatives. In addition to the classical F-test, there are several tests available for this problem (see Klotz [6], [7], Sukhatme [11], [12], Ansari and Bradley [1], Barton and David [2], Siegel and Tukey [10] and Mood [8]). Some of these tests assume the knowledge of the location parameters ν and η and some the equality of ν and η. For the general problem where the location parameters are completely unknown, the possibility of applying the usual tests to the deviations of the observations from certain consistent estimates of the unknown location parameters has been recognized by several workers in the field. See for example [1]. Sukhatme [12] has constructed a test for this case which is asymptotically distribution-free under certain conditions on the underlying distributions. Crouse [4] has recently shown that the test proposed by Mood [8] when modified in the above manner is asymptotically distribution-free under certain conditions.

In this paper we consider the normal scores test proposed by Klotz [7] for the case $\nu = \eta$ (unknown) and modify it to apply for the case when the location parameters are completely unknown. The limiting distribution of the modified test statistic is shown to remain unchanged by this modification if F is symmetric and satisfies certain regularity conditions. It follows that the modified test is asymptotically distribution-free for a fairly general class of distributions. In Section 4 it is shown that the relative asymptotic efficiency of the proposed test and the Studentized F-test also remains unchanged. In the same section this efficiency has been studied for some standard distributions.

2. Assumptions and notations. Let $X_1, X_2, \cdots, X_m$ be independent random variables with common continuous cumulative distribution function $F(x - \nu)$. Let $Y_1, Y_2, \cdots, Y_n$ be independent random variables with common continuous cumulative distribution function $G(x - \eta)$. We shall assume throughout that the distributions F and G have densities f and g respectively and that ν and η are the medians of F and G respectively. Let $N = m + n$; $\lambda_N = m/N$ and assume

Received 13 November 1964; revised 24 February 1965.

[1] This forms the author's dissertation submitted in partial satisfaction of the requirements of the Ph.D. degree in statistics at the University of California, Berkeley. Prepared with the partial support of the Government of India and the U. S. Army Research Office (Durham), Grant DA-31-124-ARO-D-548.

that for all N the inequalities $0 < \lambda_0 \leqq \lambda_N \leqq 1 - \lambda_0 < 1$ hold for some fixed $\lambda_0 \leqq \frac{1}{2}$. Let

$$F_m(x) = (\text{number of } X_i \text{ such that } X_i - \nu \leqq x)/m,$$
$$G_n(x) = (\text{number of } Y_j \text{ such that } Y_j - \eta \leqq x)/n, \tag{2.1}$$
$$H_N(x) = \lambda_N F_m(x) + (1 - \lambda_N)G_n(x).$$

Thus F_m, G_n and H_N are the sample cumulative distributions for the X, Y and the pooled samples respectively. Let $\hat{\nu}(X_1, \cdots, X_m)$ and $\hat{\eta}(Y_1, Y_2, \cdots, Y_n)$ be consistent estimates of ν and η respectively, such that $N^{\frac{1}{2}}(\hat{\nu} - \nu)$ and $N^{\frac{1}{2}}(\hat{\eta} - \eta)$ are bounded in probability. Define

$$F_m^*(x) = (\text{number of } X_i \text{ such that } X_i - \hat{\nu} \leqq x)/m,$$
$$G_n^*(x) = (\text{number of } Y_j \text{ such that } Y_j - \hat{\eta} \leqq x)/n, \tag{2.2}$$
$$H_N^*(x) = \lambda_N F_m^*(x) + (1 - \lambda_N)G_n^*(x).$$

Let

$$H(x) = \lambda_N F(x) + (1 - \lambda_N)G(x), \tag{2.2'}$$
$$h(x) = \lambda_N f(x) + (1 - \lambda_N)g(x).$$

Consider the combined sample $X_i - \nu; Y_j - \eta$, $(i = 1, \cdots, m; j = 1, 2, \cdots, n)$ and define $Z_{Ni} = 1$ if the ith smallest in the above combined sample is an X and $= 0$ otherwise. Define similarly Z_{Ni}^* for the combined sample $X_i - \hat{\nu}$; $Y_j - \hat{\eta}$ $(i = 1, 2, \cdots, m; j = 1, 2, \cdots, n)$. Define also

$$mT_N = \textstyle\sum_{i=1}^N E_{Ni}Z_{Ni}, \qquad mT_N^* = \sum_{i=1}^N E_{Ni}Z_{Ni}^* \tag{2.3}$$

where

$$E_{Ni} = J_N(i/N) = \{\Phi^{-1}[i/(N + 1)]\}^2. \tag{2.4}$$

Here Φ^{-1} is the inverse of the cdf of the standard normal distribution. We shall extend the domain of definition of J_N to the interval $(0, 1]$ by letting J_N be constant on $(i/N, (i + 1)/N]$. Let $J(H) = \lim_{N\to\infty} J_N(H) = [\Phi^{-1}\{H\}]^2$. As in Chernoff and Savage [3], T_N and T_N^* have the following integral representations

$$T_N = \int_{-\infty}^{\infty} J_N(H_N(x))\, dF_m(x), \qquad T_N^* = \int_{-\infty}^{\infty} J_N(H_N^*(x))\, dF_m^*(x).$$

Throughout our proofs K will denote a generic constant which will not depend on m, n, N. Statements involving o_p or O_p will always be uniform in the interval $0 < \lambda_0 \leqq \lambda_N \leqq 1 - \lambda_0 < 1$. Let I_N^* be the (random) interval in which $0 < H_N^*(x) < 1$, and throughout the paper let $J(x) = [\Phi^{-1}(x)]^2$.

3. Asymptotic normality. Using the theorem of Chernoff and Savage [3], Klotz [7] showed the asymptotic normality of the statistic T_N given by (2.3). In this section we prove that the limiting distribution of T_N^* defined by (2.3) is the same as that of T_N under fairly general conditions on the underlying distributions F and G, namely that

(3.1) f and g are symmetric about their respective location parameters, and

$$f(x)/\varphi[\Phi^{-1}\{F(x)\}], \qquad g(x)/\varphi[\Phi^{-1}\{G(x)\}] \tag{3.2}$$

are bounded. Here $\varphi(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)$. The assumption of symmetry of f and g was also made by Sukhatme [12] to prove a similar result on U-statistics. Condition (3.2) will, in particular, be satisfied if (i) f and g are bounded and (ii) $f(x)/\varphi[\Phi^{-1}\{F(x)\}]$ and $g(x)/\varphi[\Phi^{-1}\{G(x)\}]$ are bounded as $x \to \pm\infty$. Condition (ii) is the same as the one given in (b), Lemma 3 of Hodges and Lehmann [5]. It can easily be checked that condition (3.2) is satisfied by the usual distributions such as normal, double exponential, logistic and Cauchy.

Our statistical results will be based on the following principal theorem.

THEOREM 3.1. *If assumptions* (3.1) *and* (3.2) *hold, then*

$$\lim_{N\to\infty} P\{(T_N^* - \mu_N)/\sigma_N \leqq t\} = \int_{-\infty}^{t} (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)\, dx$$

where

$$\mu_N = \int_{-\infty}^{\infty} J\{H(x)\}\, dF(x) \tag{3.3}$$

and

$$N\sigma_N^2 = 2(1 - \lambda_N)\{\textstyle\iint_{-\infty<x<y<\infty} G(x)(1 - G(y))J' \cdot\{H(x)\}J'\{H(y)\}\, dF(x)\, dF(y) + [(1 - \lambda_N)/\lambda_N] \iint_{-\infty<x<y<\infty} F(x) \cdot(1 - F(y))J'\{H(x)\}J'\{H(y)\}\, dG(x)\, dG(y)\}. \tag{3.4}$$

PROOF. Without loss of generality, we can take $\nu = \eta = 0$. We first state two lemmas and indicate their proofs.

LEMMA 3.2. *If* (3.2) *holds, then* $h(x)/\varphi[\Phi^{-1}\{H(x)\}]$ *is also bounded, where* $h(x)$ *and* $H(x)$ *are given by* (2.2′).

INDICATION OF PROOF. From the concavity of $\varphi[\Phi^{-1}\{H(x)\}]$, $-\infty < x < \infty$, it follows that

$$h(x)/\varphi[\Phi^{-1}\{H(x)\}] \leqq \max(f(x)/\varphi[\Phi^{-1}\{F(x)\}], g(x)/\varphi[\Phi^{-1}\{G(x)\}]).$$

LEMMA 3.3. *Let* $X_1, X_2, \cdots, X_m$ *be independent and identically distributed* rv's *with continuous* cdf $F(x - \nu)$, *where* ν *is the location parameter. Let* $\hat{\nu}(X_1, \cdots, X_m)$ *be a consistent estimate of* ν *such that* $m^{\frac{1}{2}}(\hat{\nu} - \nu)$ *is bounded in probability as* $m \to \infty$. *Let* $M(x)$ *be a real-valued function defined on* $(-\infty, \infty)$ *such that the derivative* $M'(x)$ *exists and satisfies* $|M'(x - t)| \leqq KT(x)$ *uniformly in* t *for* $|t| \leqq c$ (*c and K are constants*) *and* $\mathcal{E}[\{T(X)\}^2] < \infty$ *where X has the* cdf $F(x - \nu)$. *Also assume* $\mathcal{E}[M'(X - \nu)] = 0$. *Then*

$$L_m = m^{-\frac{1}{2}} \textstyle\sum_{i=1}^{m} \{M(X_i - \nu) - M(X_i - \hat{\nu})\} \to 0 \quad \textit{in probability.}$$

INDICATION OF PROOF. Apply Tchebychev inequality.

PROOF OF THE THEOREM. From the results of Chernoff and Savage [3] and Klotz [7], it follows that

$$\lim_{N\to\infty} P\{(T_N - \mu_N)/\sigma_N \leqq t\} = \int_{-\infty}^{t} (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)\, dx,$$

where μ_N and σ_N are given by (3.3) and (3.4) respectively. To prove the theorem, we shall show that the limiting distributions of T_N and T_N^* are identical. We can write

$$T_N^* = \int_{-\infty}^{\infty} J_N(H_N^*)\, dF_m^*(x) = \int_{0<H_N^*\leqq 1} [J_N(H_N^*) - J(H_N^*)]\, dF_m^*(x) + \int_{0<H_N^*\leqq 1} J(H_N^*)\, dF_m^*(x).$$

Proceeding as in Chernoff and Savage [3], we can write

$$T_N^* = A^* + B_{1N}^* + B_{2N}^* + \sum_{i=1}^{6} C_{iN}^*$$

where

$$A^* = \int_{0<H<1} J(H)\, dF(x),$$
$$B_{1N}^* = \int_{0<H<1} J(H)\, d[F_m^*(x) - F(x)],$$
$$B_{2N}^* = \int_{0<H<1} (H_N^* - H)J'(H)\, dF(x),$$
$$C_{1N}^* = \lambda_N \int_{0<H<1} (F_m^* - F)J'(H)\, d(F_m^*(x) - F(x)),$$
$$C_{2N}^* = (1 - \lambda_N) \int_{0<H<1} (G_n^* - G)J'(H)\, d(F_m^*(x) - F(x)),$$
$$C_{3N}^* = \int_{I_N^*} [(H_N^* - H)^2/2]J''(\gamma H_N^* + (1 - \gamma)H)\, dF_m^*(x), \qquad 0 < \gamma < 1,$$
$$C_{4N}^* = \int_{H_N^*=1} (-J(H) - (H_N^* - H)J'(H))\, dF_m^*(x),$$
$$C_{5N}^* = \int_{I_N^*} (J_N(H_N^*) - J(H_N^*))\, dF_m^*(x),$$
$$C_{6N}^* = \int_{H_N^*=1} J_N(H_N^*)\, dF_m^*(x).$$

A^* is the same as A of Chernoff-Savage [3] and is finite. It is shown in the Appendix that C_{iN}^*, $i = 1, 2, \cdots, 6$ are of order $o_p(N^{-\frac{1}{2}})$ for $i = 1, 2, \cdots, 6$. It remains to show that $N^{\frac{1}{2}}(B_{1N}^* + B_{2N}^*)$ has the same limiting normal distribution as that of $N^{\frac{1}{2}}T_N$. Proceeding as in Chernoff-Savage [3], we obtain

$$N^{\frac{1}{2}}(B_{1N}^* + B_{2N}^*) - N^{\frac{1}{2}}(B_{1N} + B_{2N}) = N^{\frac{1}{2}}(1 - \lambda_N)\{m^{-1} \sum_{i=1}^{m} [B(X_i - \hat{\theta}) - B(X_i)] - n^{-1} \sum_{j=1}^{n} [D(Y_j - \hat{\eta}) - D(Y_j)]\}$$

where

$$B(x) = \int_0^x J'\{H(y)\}\, dG(y),$$
$$D(x) = \int_0^x J'\{H(y)\}\, dF(y),$$
$$B_{1N} + B_{2N} = (1 - \lambda_N)\{m^{-1} \sum_{i=1}^{m} [B(X_i) - \mathcal{E}B(X)] - n^{-1} \sum_{j=1}^{n} [D(Y_j) - \mathcal{E}D(Y)]\}.$$

Let

$$L_N = N^{-\frac{1}{2}} \sum_{i=1}^{m} [B(X_i - \hat{\theta}) - B(X_i)].$$

Using Lemma 3.2 and the mean value theorem it can be shown easily that $|B'(X - t)| \leqq K\{|\Phi^{-1}\{H(x)\}| + K\}$ for $|t| \leqq C$. Also

$$\int_{-\infty}^{\infty} [\Phi^{-1}\{H(x)\}]^2 \, dF(x) \leqq K \int_{-\infty}^{\infty} [\Phi^{-1}\{H(x)\}]^2 \, dH(x) < \infty.$$

Further

$$\mathcal{E}B'(X) = 2 \int_{-\infty}^{\infty} \{\Phi^{-1}\{H(x)\}/\varphi[\Phi^{-1}\{H(x)\}]\}f(x)g(x) \, dx = 0$$

since $f(x)$ and $g(x)$ are symmetric about zero so that $\Phi^{-1}\{H(x)\}/\varphi[\Phi^{-1}\{H(x)\}]$ is an odd function in x. Also $\mathcal{E}B'(X) < \infty$. Thus $B(x)$ satisfies all the conditions of Lemma 3.3. Hence $L_N \to 0$ in probability. In a similar manner it follows that $N^{\frac{1}{2}} \sum_{j=1}^{n} [D(Y_j - \hat{\eta}) - D(Y_j)] \to 0$ in probability, from which it follows that $N^{\frac{1}{2}}(B_{1N}^* + B_{2N}^*)$ has the same asymptotic distribution of $N^{\frac{1}{2}}(B_{1N} + B_{2N})$. From the results of Chernoff-Savage [3] and Klotz [7] $N^{\frac{1}{2}}(B_{1N} + B_{2N})$ has the same limiting distribution of $N^{\frac{1}{2}}T_N$. This completes the proof of the theorem.

4. A Normal Scores test for the general two-sample scale problem. For the two-sample scale problem defined in Section 1, Klotz [7] proposed a nonparametric test when it is assumed that ν and η are unknown but equal. The test statistic is T_N defined by (2.3). It was also shown by Klotz [7] that the asymptotic efficiency of his test relative to the F-test is unity when the underlying distributions are normal. For the general case, when ν and η are completely unknown, we propose the test based on T_N^* defined by (2.3). From Theorem 3.1, it follows that the modified test is asymptotically distribution-free for the class of distributions satisfying (3.1) and (3.2). It also follows from the same theorem that for the class of distributions covered by the conditions of the theorem, the limiting distributions of T_N and T_N^* are identical for fixed alternatives σ and τ. However, to obtain the efficacy of the T_N^*-test, we need to derive the asymptotic distribution of T_N^* for a sequence of alternatives. The identity of the limiting distributions of T_N and T_N^* in this case follows from the following corollary.

COROLLARY 4.1. *Let $G(x) = F\{(x - \eta)\theta\}$, $0 < \theta < \infty$, in Theorem 3.1. Then the asymptotic normality of T_N^* holds uniformly in λ_N and for θ in some neighborhood of unity.*

PROOF. From the results of Chernoff and Savage [3] and Klotz [7], the limiting normality of the T_N-statistic is uniform in λ_N and for θ in some neighborhood of unity. To prove the corollary, it is, therefore, enough to show that (i) $N^{\frac{1}{2}}(B_{1N}^* + B_{2N}^*) - N^{\frac{1}{2}}(B_{1N} + B_{2N}) \to 0$ in probability, uniformly in λ_N and for θ in some neighborhood of unity and that (ii) $N^{\frac{1}{2}}C_{iN}^* \to 0$ in probability, uniformly in λ_N and for θ in some neighborhood of unity ($i = 1, 2, \cdots, 6$). As in the proof of Theorem 3.1, assume $\nu = \eta = 0$. Note that, when $G(x) = F(x\theta)$, $g(x)/\varphi[\Phi^{-1}\{G(x)\}] = \theta f(x\theta)/\varphi[\Phi^{-1}\{F(x\theta)\}]$. Since $f(x)/\varphi[\Phi^{-1}\{F(x)\}]$ is bounded and θ lies in some neighborhood of unity, we have $g(x)/\varphi[\Phi^{-1}\{G(x)\}]$ is bounded uniformly in θ. Hence, by Lemma 3.2, $h(x)/\varphi[\Phi^{-1}\{H(x)\}]$ is bounded uniformly in θ, for θ lying in some neighborhood of unity. A perusal of the proof of Theorem 3.1 now proves (i) and (ii). (Uniformity in λ_N throughout is clear from the assumption $0 < \lambda_0 \leqq \lambda_N \leqq 1 - \lambda_0 < 1$.) This proves the corollary. It now follows that for the distributions satisfying the condition: $f(x)/\varphi[\Phi^{-1}\{F(x)\}]$ is bounded, and f symmetric about zero, the asymptotic efficiency (Pitman sense) of the T_N^*-test is the same as that of the T_N-test. The expression for the efficiency $e_{T_N^*,F}$ is therefore the same as $e_{T_N,F}$ given by Klotz [7].

APPENDIX

5. Higher order terms. In this appendix, the proof for the negligibility of the C-terms of Theorem 3.1 is briefly indicated. The full details are given in [9]. The method adopted to achieve the result is to replace $\hat{p}$ and $\hat{\eta}$ occurring in these terms by nonrandom $t_1/N^{\frac{1}{2}}$ and $t_2/N^{\frac{1}{2}}$ and to show that the resulting terms are $o_p(N^{-\frac{1}{2}})$ uniformly in t_1, t_2 for $|t_1| \leqq C_1$ and $|t_2| \leqq C_2$. Let $S_{N\epsilon} = \{x: H(1-H) > \eta_\epsilon \lambda_0/N\}$. Since f and g are assumed to be symmetric about zero, the interval $S_{N\epsilon}$ is symmetric about zero. Let $S_{N\epsilon} = (-a_N, a_N)$. Define $S'_{N\epsilon} = (-a_N - t/N^{\frac{1}{2}}, a_N - t/N^{\frac{1}{2}})$ and $S^+_{N\epsilon} = (-a_N - \max(t_1, t_2)/N^{\frac{1}{2}}, a_N - \max(t_1, t_2)/N^{\frac{1}{2}})$. C^*_{1N} can be shown to be $o_p(N^{-\frac{1}{2}})$ by using the general technique mentioned above and by employing the method of Chernoff-Savage. To deal with C^*_{2N}, define

$$C^{**}_{2N} = \int_{-\infty}^{\infty} [G^{**}_n(x) - G(x)] J'\{H(x)\}\, d[F^{**}_m(x) - F(x)]$$

where $G^{**}_n(x) = G_n(x + t_2/N^{\frac{1}{2}})$ and $F^{**}_m(x) = F_m(x + t_1/N^{\frac{1}{2}})$. Splitting the range of integration into $S^+_{N\epsilon}$ and $S^{+c}_{N\epsilon}$, and in the integral over $S^+_{N\epsilon}$, writing $G^{**}_n(x) - G(x) = \{G^{**}_n(x) - G^*(x)\} + \{G^*(x) - G(x)\}$ and $F^{**}_m(x) - F(x) = \{F^{**}_m(x) - F^*(x)\} + \{F^*(x) - F(x)\}$ where $F^*(x) = F(x + t_1/N^{\frac{1}{2}})$ and $G^*(x) = G(x + t_2/N^{\frac{1}{2}})$, and collecting the terms it can be shown as in Chernoff-Savage [3] that $C^*_{2N} = o_p(N^{-\frac{1}{2}})$. For the consideration of the C^*_{3N} term, define

$$C^{**}_{3N} = \int_{0<H^{**}_N<1} [H^{**}_N(x) - H(x)]^2 J''\{\gamma H^{**}_N(x) + (1-\gamma)H(x)\}\, dF^{**}_m(x),$$

$$C^{***}_{3N} = \int_{0<H^{**}_N(x)<1} [H^{**}_N(x) - H(x)]^2 J''\{H(x)\}\, dF^{**}_m(x),$$

where $H^{**}_N(x) = \lambda_N F^{**}_m(x) + (1-\lambda_N) G^{**}_n(x)$.

It can be shown that

(i) $N^{\frac{1}{2}}(C^{**}_{3N} - C^{***}_{3N}) = o_p(1)$ uniformly in t_1, t_2 for $|t_1| \leqq c_1$ and $|t_2| \leqq c_2$,

(ii) $N^{\frac{1}{2}}(C^{***}_{3N}) = o_p(1)$ uniformly in t_1, t_2 for $|t_1| \leqq c_1$ and $|t_2| \leqq c_2$,

where c_1, c_2 are constants. This completes the discussion of the C^*_{3N} term. The discussion of the terms C^*_{4N}, C^*_{5N} and C^*_{6N} is analogous to the argument used by Chernoff and Savage [3] in their treatment of the terms C_{4N}, C_{5N}, C_{6N} respectively. This completes the discussion of all the C-terms.

6. Acknowledgment. The author is grateful to Professor E. L. Lehmann for suggesting the above investigation and for his helpful guidance throughout the preparation of the paper.

REFERENCES

[1] Ansari, A. R. and Bradley, R. A. (1960). Rank sum tests for dispersions. *Ann. Math. Statist.* **31** 1174–1189.

[2] Barton, D. E. and David, F. N. (1958). A test for birth order effects. *Ann. Eugenics* **22** 250–257.

[3] Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29** 972–994.

[4] Crouse, C. F. (1964). Note on Mood's Test. *Ann. Math. Statist.* **35** 1825–1826.

[5] Hodges, J. L., Jr. and Lehmann, E. L. (1961). Comparisons of the Normal scores and Wilcoxon tests. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 307–317.

[6] Klotz, J. (1960). Unpublished doctoral dissertation. Univ. California, Berkeley.

[7] Klotz, J. (1962). Nonparametric tests for scale. *Ann. Math. Statist.* **32** 498–512.

[8] Mood, A. (1954). On the asymptotic efficiency of certain nonparametric tests. *Ann. Math. Statist.* **25** 514–522.

[9] Raghavachari, M. (1964). Unpublished doctoral dissertation. Univ. California, Berkeley.

[10] Siegel, S. and Tukey, J. W. (1960). A non-parametric sum of ranks procedure for relative spread in unpaired samples. *J. Amer. Statist. Assoc.* **55** 429–455.

[11] Sukhatme, B. V. (1957). On certain two-sample nonparametric tests for variance. *Ann. Math. Statist.* **28** 188–194.

[12] Sukhatme, B. V. (1958). Testing the hypothesis that two populations differ only in location. *Ann. Math. Statist.* **29** 60–67.