SOME NON-PARAMETRIC TESTS FOR LOCATION AND SCALE PARAMETERS IN A MIXED MODEL OF DISCRETE AND CONTINUOUS VARIABLES*

BY SHASHIKALA SUKHATME

New Delhi

1. INTRODUCTION

LET $Z_1, \ldots, Z_N$ where $Z_i = (X_i, Y_i)$ be N independent observations from a bivariate distribution. We assume that the random variable X takes only two values 1 and 0 with $P(X = 1) = p$ and $P(X = 0) = 1 - p = q$. Let the conditional distribution of Y given $X = j$ $(j = 0, 1)$ be $P(Y \leqslant y \mid X = j) = F_j(y)$. The problem considered is that of testing the hypothesis $H: F_1 = F_0$ against the alternative $A: F_1 \neq F_0$.

We divide the observations $Z_1, \ldots, Z_N$ into two groups according as the observed value of X is 1 or 0. Let $U_1, \ldots, U_n$ $(n > 0)$ and $V_1, \ldots, V_{N-n}$ denote those values of Y for which the corresponding X is observed to be 1 and 0 respectively. Since for given n, $U_1, \ldots, U_n$ and $V_1, \ldots, V_{N-n}$ are independent, the problem of testing the hypothesis H is equivalent to testing the hypothesis that the two independent samples come from the same population. However, the problem differs from the usual two-sample problem in that the number of observations in each of the two samples is a random variable.

In what follows, we assume that F_1 and F_0 are absolutely continuous having density functions f_1 and f_0 respectively. We further assume that F_1 and F_0 have the same functional form except that they differ either in the location or the scale parameter.

Several two-sample non-parametric tests have been proposed for testing differences in location, especially those by Wilcoxon,[12] Mood,[8] Wald and Walfowitz[11] and Lehmann.[5] More recently, some non-parametric tests have been proposed for testing differences in dispersion by Mood,[9] Sukhatme[10] and Kamat.[4] The purpose of this

* This research was performed while the author was a graduate student at Michigan State University, East Lansing, Michigan, and was sponsored in part by the Office of Ordnance Research.

paper is to study some of these tests with reference to the problem considered above.

In Sections 2 and 3, we consider the median test and Wilcoxon test for testing differences in location. Sections 4 and 5 are devoted to Mood's rank test and the run test for testing differences in dispersion. For convenience of exposition, the cases when p is known or unknown are treated separately. In the former case both exact and asymptotic properties are investigated. In the latter case, the various test statistics are modified by replacing p by its usual estimator $\hat{p}$ and we investigate whether the tests based on the modified test statistics are asymptotically distribution-free.

2. Two-Sample Median Test

We assume that the sample size is odd, say $N = 2k + 1$. Let W denote the sample median of Y observations and let m be the number of U's which are less than $\tilde{W}$. The hypothesis $H: F_1 = F_0$ is rejected if m is either too large or too small. First, consider the case when the distribution of X is known, *i.e.*, when p is known.

2.1 *Joint and Marginal Distributions of m and* $\tilde{W}$

Henceforth $f(.)$ denotes the probability density function of the random variable written in the parentheses. We first prove the following lemma which gives the joint distribution of m and $\tilde{W}$.

Lemma 2.1.1.—The joint distribution of m and $\tilde{W}$ is

$$f(m, \tilde{w}) = \frac{N!}{m!\,(k-m)!\,k!}\,[pF_1(\tilde{w})]^m\,[qF_0(\tilde{w})]^{k-m} \times [1 - pF_1(\tilde{w}) - qF_0(\tilde{w})]^k\,[p f_1(\tilde{w}) + q f_0(\tilde{w})], \tag{2.1.1}$$

for $m = 0, 1, \ldots, k, -\infty < \tilde{w} < +\infty$.

Proof.—Observing that n is binomial random variable $b(N, p)$, we have

$$f(n, m, \tilde{w}) = f(m, \tilde{w} \mid n) f(n). \tag{2.1.2}$$

From Mood ([8])

$$f(m, \tilde{w} \mid n) = \left\{\frac{n!}{m!\,(n-m-1)!}\,[F_1(\tilde{w})]^m [1 - F_1(\tilde{w})]^{n-m-1}\right\} \times \left\{\frac{(N-n)!\,[F_0(\tilde{w})]^{k-m}}{(k-m)!\,(k+m+1-n)!}\,[1 - F_0(\tilde{w})]^{k+m+1-n}\, f(\tilde{w})\right\}$$

$$+\left\{\frac{n!}{m!\,(n-m)!}\,[F_1(\tilde{w})]^m[1-F_1(\tilde{w})]^{n-m}\right\}$$

$$\times\left\{\frac{(N-n)!}{(k-m)!\,(k+m-n)!}[F_0(\tilde{w})]^{k-m}[1-F_0(\tilde{w})]^{k+m-n}f_0(\tilde{w})\right\} \tag{2.1.3}$$

Using (2.1.3) in (2.1.2) and summing (2.1.2) over all admissible values of n, we obtain $f(m, \tilde{w})$ as given by (2.1.1).

Under the hypothesis $H: F_1 = F_0$, (2.1.1) reduces to

$$f(m, \tilde{w}) = \frac{N!}{m!\,(k-m)!\,k!}\,p^m q^{k-m}[F(\tilde{w})]^k[1-F(\tilde{w})]^k f(\tilde{w}) \tag{2.1.4}$$

for m, $0, 1, \ldots, k$; $-\infty < \tilde{w} < +\infty$.

Integrating (2.1.4) over the domain $0 \leqslant F(\tilde{w}) \leqslant 1$ it is seen that m is a binomial random variable $b(k, p)$. Also it is seen that the marginal distribution of $\tilde{W}$ is

$$f(\tilde{w}) = \frac{N!}{k!\,k!}\,[F(\tilde{w})]^k\,[1-F(\tilde{w})]^k f(\tilde{w}); \quad -\infty < \tilde{w} < +\infty.$$

2.2 *Asymptotic Distribution of m*

Let ξ denote the median of Y, *i.e.*, ξ is the root (assumed unique) of the equation

$$pF_1(\xi) + qF_0(\xi) = \tfrac{1}{2}. \tag{2.2.1}$$

Theorem 2.2.1.—Let

$$\nu = \frac{[m - NpF_1(\xi)]}{[NpF_1(\xi)]^{\frac{1}{2}}}, \quad \eta = N^{\frac{1}{2}}(\tilde{W} - \xi),$$

where ξ satisfies (2.2.1). Assume that in some neighbourhood of ξ, the density function $f_i(x)$ $(i = 0, 1)$ has a continuous derivative. Then the asymptotic joint distribution of (ν, η) is bivariate normal distribution with zero mean vector and covariance matrix $\Sigma = (\sigma_{ij})$ given by

$$\sigma_{11} = \sigma^2(\nu) = qF_0(\xi) + \frac{q\,[pf_1^2(\xi)\,F_0(\xi) + qf_0^2(\xi)\,F_1(\xi)]}{2F_1(\xi)\,[pf_1(\xi) + qf_0(\xi)]^2},$$

$$\sigma_{22} = \sigma^2(\eta) = \frac{1}{4\,[pf_1(\xi) + qf_0(\xi)]^2},$$

$$\sigma_{12}=\sigma_{21}=\text{cov}\,(\nu,\eta)=\frac{pq\,[f_1(\xi)\,F_0(\xi)-f_0(\xi)\,F_1(\xi)]}{(2pF_1(\xi))^{\frac{1}{2}}\,[pf_1(\xi)+qf_0(\xi)]^2}.$$

Proof.—Throughout this proof we write $F_i \equiv F_i(\xi)$ and $f_i \equiv f_i(\xi)$. The joint probability density function of m and $\tilde{W}$ is given by (2.2.1). Expand $F_i(\xi+\eta/N^{\frac{1}{2}})$ in Taylor's series about ξ.

$$F_i\left(\xi+\frac{\eta}{\sqrt{N}}\right)=F_i+\frac{\eta}{\sqrt{N}}f_i+o\left(\frac{\eta^2}{N}\right),\quad i=0,1;$$

$$1-pF_1\left(\xi+\frac{\eta}{\sqrt{N}}\right)-qF_0\left(\xi+\frac{\eta}{\sqrt{N}}\right)$$

$$=\frac{1}{2}-\frac{\eta}{\sqrt{N}}(pf_1+qf_0)-o\left(\frac{\eta^2}{N}\right).$$

Employing the above expansions and substitution of expressions for ν and η in (2.2.1) yields

$$f(m,\tilde{w})=\left\{\frac{N(2k)!}{k!\,k!\,2^{2k}}\right\}\left\{\frac{k!}{m!\,(k-m)!}(2pF_1)^m\,(2qF_0)^{k-m}\right\}$$

$$\times\left\{\left[1+\frac{\eta}{\sqrt{N}}\frac{f_1}{F_1}+o\left(\frac{\eta^2}{N}\right)\right]^m\left[1+\frac{\eta}{\sqrt{N}}\frac{f_0}{F_0}+o\left(\frac{\eta^2}{N}\right)\right]^{k-m}\right.$$

$$\left.\times\left[1-\frac{2\eta}{\sqrt{N}}(pf_1+qf_0)-o\left(\frac{\eta^2}{N}\right)\right]^k\right\}$$

$$\times\left\{pf_1\left(\xi+\frac{\eta}{\sqrt{N}}\right)+qf_0\left(\xi+\frac{\eta}{\sqrt{N}}\right)\right\}.$$

Let $S=\{(\nu,\eta):\ a\leqslant\nu\leqslant b,\ c\leqslant\eta\leqslant d\}$ where a, b, c and d are finit. Now using Stirling's formula for $n!$, taking logarithms and using serie expansion for $\log(1+x)$ it is seen that uniformly in S we have

$$f(m,\tilde{w})\sim\frac{N(pf_1+qf_0)}{\pi k\,(8pqF_1F_0)^{\frac{1}{2}}}\exp.-\frac{1}{2}\left\{\eta^2\left[\frac{pf_1^2}{F_1}+\frac{qf_0^2}{F_0}\right.\right.$$

$$\left.\left.+2\,(pf_1+qf_0)^2\right]+\frac{\nu^2}{4qF_0}-2\nu\eta\,(pF_1)^{\frac{1}{2}}\left(\frac{f_1}{F_1}-\frac{f_0}{F_0}\right)\right\}.$$

Now after making the transformation $(m,\tilde{w})\to(\nu,\eta)$, it is easily seen that

$$\lim_{N\to\infty}P\{a\leqslant\nu\leqslant b,\ c\leqslant\eta\leqslant d\}=\int_a^b\int_c^d f(\nu,\eta)\,d\nu d\eta,$$

where $f(\nu,\eta)$ is the density function of the bivariate normal distribution stated in the theorem.‖

2.3. *Consistency of the Test*

Consider a two-sided test of the hypothesis $H: F_1 = F_0$ against the alternative $A: F_1 \neq F_0$ for which the critical region is given by

$$\left| \frac{(m - kp)}{(kpq)^{\frac{1}{2}}} \right| > t_{N,\alpha}.$$

The sequence $t_{N,\alpha}$ is chosen so that

$$\lim_{N \to \infty} t_{N,\alpha} = t_\alpha$$

here t_α satisfies $1 - \Phi(t_\alpha) = \alpha/2$ and $\Phi(t)$ is the standardized normal distribution function. Then the power of the test is given by

$$P\left\{ \left| \frac{m - kp}{(kpq)^{\frac{1}{2}}} \right| > t_{N,\alpha}, F_1 \neq F_0 \right\}$$

$$= 1 - P\left\{ - \frac{t_{N,\alpha}}{2\,(F_1(\xi)\,F_0(\xi))^{\frac{1}{2}}} + \frac{kp\,(1 - 2F_1(\xi))}{2\,(kpq\,F_1(\xi)\,F_0(\xi))^{\frac{1}{2}}} \right.$$

$$< \frac{m - kp\,F_1(\xi)}{2\,(kpq\,F_1(\xi)\,F_0(\xi))^{\frac{1}{2}}}$$

$$\left. < \frac{t_{N,\alpha}}{2\,(F_1(\xi)\,F_0(\xi))^{\frac{1}{2}}} + \frac{kp\,(1 - 2F_1)(\xi_1)}{2\,(kpq\,F_1(\xi)\,F_0(\xi))^{\frac{1}{2}}} \right\}.$$

If $F_1(\xi) \neq 1/2$, the power tends to 1 as N tends to infinity. Hence the test is consistent.

For alternatives $F_1 > F_0$, $(F_1 < F_0)$ we prove in a similar manner that the test is consistent if $F_1(\xi) > 1/2$, $(F_1(\xi) < 1/2)$.

2.4. *Asymptotic Efficiency of the Test*

Let $F_1(y) = F_0(y - \theta)$, then $H: F_1 = F_0$ is equivalent to $H: \theta = 0$. For alternatives $\theta > 0$, we evaluate the relative asymptotic efficiency of the median test with respect to the corresponding parametric test when F_1 and F_0 are normal distributions with means μ_1 and μ_0 respectively and a common variance σ^2. $F_1(y) = F_0(y)$ if and only if $\mu_1 = \mu_0$ which is equivalent to $\rho = \rho(X, Y) = 0$, where ρ is the correlation coefficient between X and Y, Tate.[12] Let $\triangle = (\mu_1 - \mu_0)/\sigma$. Then we are testing the hypothesis $\triangle = 0$ against $\triangle > 0$. The test is based on the sample correlation coefficient r, between X and Y. Tate[12]

proved that r is asymptotically normally distributed with mean and variance given by

$$\mu_{\Delta}(r) = \Delta\left(\frac{pq}{1+pq\Delta^2}\right)^{\frac{1}{2}}, \quad \sigma_{\Delta}{}^2(r) = \frac{4(1+pq\,\Delta^2) - \Delta^2(6pq-1)}{4N(1+pq\Delta^2)^3}. \tag{2.4.1}$$

The critical region for this test is given by $rN'^{\frac{1}{2}} > t'_{N,'\alpha}$, where $\{t'_{N',\alpha}\}$ is such that $\lim_{N\to\infty} t'_{N',\alpha} = t_\alpha$ and $\Phi(t_\alpha) = 1-\alpha$. Since r is asymptotically normally distributed the asymptotic power of the test is obtained as

$$\lim_{N'\to\infty} \beta'_{N'}(\Delta) = \lim_{N'\to\infty} P\left\{r > \frac{t'_{N',\alpha}}{\sqrt{N'}}\right\}$$

$$= 1 - \Phi\left(\lim_{N'\to\infty} \frac{t'_{N',\alpha} - \mu_{\Delta}(r)\sqrt{N'}}{\sigma_{\Delta}(r)\sqrt{N'}}\right).$$

Now for a sequence of alternatives $\{\Delta_{N'}\}$ where $\Delta_{N'} = \delta'/N'^{\frac{1}{2}}$, $\delta' > 0$.

$$\lim_{N'\to\infty}\left(\frac{1}{[\sqrt{N'}\,\sigma_{\Delta_{N'}}(r)]}\right) = 1$$

and

$$\lim_{N'\to\infty}\left[\frac{\mu_{\Delta_{N'}}(r)}{\sigma_{\Delta_{N'}}(r)}\right] = \frac{\delta'}{\sqrt{pq}}.$$

Therefore

$$\lim_{N'\to\infty} \beta'_{N'}(\Delta_{N'}) = \Phi(-t_\alpha + \delta'\sqrt{pq}). \tag{2.4.2}$$

Now for the median test under consideration the critical region for testing $H: \theta = 0$, against the alternative $\theta > 0$ is given by

$$\frac{(m-kp)}{(kpq)^{\frac{1}{2}}} < t_{N,\alpha}$$

where

$$\lim_{N\to\infty} t_{N,\alpha} = -t_\alpha \text{ with } \Phi(-t_\alpha) = \alpha.$$

The asymptotic power of the test is

$$\lim_{N\to\infty} \beta_N(\theta) = \Phi\left(\lim_{N\to\infty} \frac{-t_{N,\alpha}(kpq)^{\frac{1}{2}} - [NpF_1(\xi) - kp]}{\sigma_\theta(m)}\right).$$

For a sequence of alternatives $\{\theta_N\}$ with $\theta_N = \delta/N^{\frac{1}{2}}$, $\delta > 0$,

$$\lim_{N\to\infty}\left[\frac{(kpq)^{\frac{1}{2}}}{\sigma_{\theta_N}(m)}\right] = 1.$$

Also since ξ satisfies

$$pF_0(\xi - \theta) + qF_0(\xi) = \tfrac{1}{2}, \quad \frac{d\xi}{d\theta} = \frac{p f_0(\xi - \theta)}{pf_0(\xi - \theta) + qf_0(\xi)}.$$

Hence for the sequence $\{\theta_N\}$

$$\lim_{N\to\infty} \frac{NpF_0\left(\xi - \frac{\delta}{N^{\frac{1}{2}}}\right) - kp}{\sigma_{\theta_N}(m)} = \delta f_0(\xi)(2pq)^{\frac{1}{2}}$$

which yields

$$\lim_{N\to\infty} \beta_N(\theta_N) = \Phi\left(-t_a + \delta f_0(\xi)(2pq)^{\frac{1}{2}}\right). \tag{2.4.3}$$

The two sequences $\{\triangle_{N'}\}$ and $\{\theta_N\}$ will be the same if $N'/N = \delta'^2/\delta^2$.

From (2.4.2) and (2.4.3) it is seen that

$$\lim_{N\to\infty} \beta_N(\theta_N) = \lim_{N'\to\infty} \beta'_{N'}(\triangle_{N'})$$

only if $\delta'/\delta = \sqrt{2} f_0(\xi)$. Hence the required efficiency is given by $e(M, r) = 1/\pi$.

2.5 *Case when p is Unknown*

The theory developed so far is not applicable when p is unknown. In this case we consider the test based on the statistic $(m - k\hat{p})/(k\hat{p}\hat{q})^{\frac{1}{2}}$ where $\hat{p}$ is the usual estimator $\hat{p} = n/N$. We now show that the test based on this statistic is asymptotically distribution-free.

Theorem 2.5.1.—Under the hypothesis $H: F_1 = F_0$, the statistic $(m - k\hat{p})/(k\hat{p}\hat{q})^{\frac{1}{2}}$ is asymptotically normally distributed with mean zero and variance $\frac{1}{2}$.

Proof.—Since plim $\hat{p} = p$, by an application of Slutsky's theorem,[1] plim $(\hat{p}\hat{q}/pq) = 1$. Hence the limiting distribution of $(m - k\hat{p})/(k\hat{p}\hat{q})^{\frac{1}{2}}$ is the same as that of $(m - k\hat{p})/(kpq)^{\frac{1}{2}}$.

Write

$$\frac{m - k\hat{p}}{(kpq)^{\frac{1}{2}}} = \frac{m - kp}{(kpq)^{\frac{1}{2}}} - \frac{k(\hat{p} - p)}{(kpq)^{\frac{1}{2}}} \equiv T_1 - T_2.$$

The asymptotic joint distribution of (T_1, T_2) is bivariate normal $n(0, \Sigma)$ with $\Sigma = (\sigma_{ij})$ where $\sigma_{11} = 1$, $\sigma_{12} = \sigma_{21} = \sigma_{22} = 1/2$. Hence the required result follows.

3. Two-Sample Wilcoxon Test

As before, let $Z_i = (X_i, Y_i)$, $i = 1, 2, 3, \ldots, N$, be independent observations from a bivariate population, where X assumes only two values, 1 and 0 with probabilities p and $q = 1 - p$ respectively. The test statistic may then be defined as

$$\bar{U}_N = \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} H(Z_i, Z_j)$$

where

$$H(Z_i, Z_j) = \begin{cases} 1, \text{ if } X_i = 1, X_j = 0, \text{ and } Y_i < Y_j, \\ 0, \text{ otherwise.} \end{cases}$$

If $U_1, \ldots, U_N$ denote those Y observations for which the corresponding values of X are observed to be 1, and $V_1, \ldots, V_{N-n}$ the remaining Y observations, then $N(N-1)\,\bar{U}_N$ is the total number of pairs (U_i, V_j) such that $U_i < V_j$. The hypothesis $H: F_1 = F_0$ is rejected if $\bar{U}_N$ is either too large or too small.

3.1 *Mean and Variance of* $\bar{U}_N$

$$E_p(\bar{U}_N) = E_p H(Z_i, Z_j) = pq \int F_1(y)\, dF_0(y) \tag{3.1.1}$$

To compute the variance of $\bar{U}_N$, write $\bar{U}_N$ as

$$\bar{U}_N = \frac{1}{N(N-1)} \sum_{j=1}^{N-n} \sum_{i=1}^{n} \phi(U_i, V_j), \tag{3.1.2}$$

where

$$\phi(u, v) = \begin{cases} 1, \text{ if } u < v; \\ 0, \text{ otherwise.} \end{cases}$$

Squaring (3.1.2), and taking expected values, we obtain the conditional moment:

$$\begin{aligned} N^2(N-1)^2 E_p(\bar{U}_N{}^2 \mid n) &= n(N-n) \int F_1(y)\, dF_0(y) \\ &+ n(n-1)(N-n) \int F_1{}^2(y)\, dF_0(y) \\ &+ n(N-n)(N-n-1) \int [1 - F_0(y)]^2\, dF_1(y) \\ &+ n(n-1)(N-n)(N-n-1) \left[\int F_1(y)\, dF_0(y)\right]^2. \end{aligned} \tag{3.1.3}$$

Write $N^{(r)} = N(N-1)\ldots(N-r+1)$. Since n has a binomial distribution $b(N, p)$

$$E[n^{(r)}(N-n)^{(s)}] = p^r q^s N^{(r+s)}. \tag{3.1.4}$$

We obtain after using (3.1.3) and (3.1.4)

$$\begin{aligned}\sigma_p^2(\bar{U}_N) = \frac{pq}{N(N-1)}[&\int F_1(y)\,dF_0(y) + (N-2)\,p\int F_1^2(y)dF_0(y)\\ &+ (N-2)\,q\int\{1-F_0(y)\}^2\,dF_1(y)\\ &- 2pq\,(2N-3)\{\int F_1(y)\,dF_0(y)\}^2].\end{aligned} \tag{3.1.5}$$

In particular, under H (3.1.1) and (3.1.5) reduce to

$$E_p(\bar{U}_N \mid H) = \frac{pq}{2}, \tag{3.1.6}$$

$$\sigma_p^2(\bar{U}_N \mid H) = \frac{pq}{N(N-1)}\left[\frac{2N-1}{6} - \frac{(2N-3)}{2}pq\right]. \tag{3.1.7}$$

3.2 *Distribution of* $\bar{U}_N$

Define

$$T_N = N(N-1)\,\bar{U}_N = \{\text{number of pairs } (Z_i, Z_j) \text{ such that } X_i = 1,\ X_j = 0 \text{ and } Y_i < Y_j\}.$$

T_N takes values $0, 1, \ldots, k$, where $k = \max n(N-n) = [N^2/4]$ where $[x]$ denotes the largest integer $\leqslant x$. Let $T_{n,N-n}$ denote the value of T_N when n is fixed. Clearly $T_{n,N-n}$ takes values $0, 1, \ldots, n(N-n)$, and

$$P\{T_N = t\} = \sum_{n=0}^{N}\binom{N}{n}p^n q^{N-n}\,P\{T_{n,N-n} = t\}. \tag{3.2.1}$$

Mann and Whitney[6] have shown that $P\{T_{n,N-n} = t\}$ satisfies the recurrence relation

$$P\{T_{n,N-n} = t\} = \frac{n}{N}P\{T_{n-1,N-n} = t\} + \frac{N-n}{N}P\{T_{n,N-n-1} = t-n\}. \tag{3.2.2}$$

Using this in (3.2.1) we get

$$\begin{aligned}P\{T_N = t\} &= pP\{T_{N-1} = t\}\\ &\quad + q\sum_{n=0}^{t}\binom{N-1}{n}p^n q^{N-n-1}\,P\{T_{n,N-n-1} = t-n\}.\end{aligned} \tag{3.2.3}$$

(3 2.3) is a recurrence relation for $P\{T_N = t\}$, from which we can find the distribution of T_N under the hypothesis H for all N. It is easy to prove by induction from (3.2.3) that,

$$P\{T_N = 0\} = \sum_{r=0}^{N} p^{N-r}q^r, \text{ for all } N.$$

The probability distribution of T_N obtained by using (3.2.3) is given below for $N = 2, 3, 4, 5$.

	t	$P\{T_N = t\}$
$N = 2$	0	$p^2 + pq + q^2$
	1	pq
	2	0
$N = 3$	0	$p^3 + p^2q + pq^2 + q^3$
	1	$p^2q + pq^2$
	2	$p^2q + pq^2$
	3	0
$N = 4$	0	$p^4 + p^3q + p^2q^2 + pq^3 + q^4$
	1	$p^3q + p^2q^2 + pq^3$
	2	$p^3q + 2p^2q^2 + pq^3$
	3	$p^3q + p^2q^2 + pq^3$
	4	p^2q^2
$N = 5$	0	$p^5 + p^4q + p^3q^2 + p^2q^3 + pq^4 + q^5$
	1	$p^4q + p^3q^2 + p^2q^3 + pq^4$
	2	$p^4q + 2p^3q^2 + 2p^2q^3 + pq^4$
	3	$p^4q + 2p^3q^2 + 2p^2q^3 + pq^4$
	4	$p^4q + 2p^3q^2 + 2p^2q^3 + pq^4$
	5	$p^3q^2 + p^2q^3$
	6	$p^3q^2 + p^2q^3$

3.3 *Asymptotic Distribution and Consistency of the Test*

Clearly $\bar{U}_N$ is a U-statistic in the sense defined by Hoeffding.[2] Hence $[\bar{U}_N - E_p(\bar{U}_N)]/\sigma_p(\bar{U}_N)$ is asymptotically $n(0, 1)$, both under the hypothesis H as well as under the alternative.

Consider the two-sided test of the hypothesis $H: F_1 = F_0$ against $A: F_1 \neq F_0$, with critical region $|[\bar{U}_N - E_p(\bar{U}_N)]/\sigma_p(\bar{U}_N)| > t_{N,\alpha}$. The sequence $t_{N,\alpha}$ is chosen so that $\lim_{N\to\infty} t_{N,\alpha} = t_\alpha$, where t_α satisfies $1 - \Phi(t_\alpha) = \alpha/2$. The power of the test is given by

$$P\left\{\left|\frac{\bar{U}_N - E_p(\bar{U}_N)}{\sigma_p(\bar{U}_N)}\right| > t_{N,\alpha} \mid F_1 \neq F_0\right\}$$

$$= 1 - P\left\{-t_{N,\alpha} < \frac{\bar{U}_N - E_p(\bar{U}_N)}{\sigma_p(\bar{U}_N)} < t_{N,\alpha} \mid F_1 \neq F_0\right\}.$$

Proceeding as in Section 2.3, if $\int F_1(y)\, dF_0(y) \neq \frac{1}{2}$, the power tends to 1, as $N \to \infty$, and hence the test is consistent. In a similar manner it can be verified that the test is consistent when $F_1 > F_0$ or $F_1 < F_0$.

3.4 *Asymptotic Efficiency of the Test*

We now find the asymptotic efficiency of the test based on $\bar{U}_N$ with respect to the parametric test based on the sample correlation coefficient between X and Y, described in Section 2.4. We have seen in Section 3.3 that $\bar{U}_N$ is asymptotically normally distributed both under H and the alternative. Proceeding as in Section 2.4 it can be proved that the required relative asymptotic efficiency is given by

$$e(\bar{U}_N, r) = \frac{3[\int f_0^2(y)\, dy]^2}{1 - 3pq} = \frac{3}{4\pi(1 - 3pq)}.$$

The asymptotic efficiency $e(\bar{U}_N, r)$ is a maximum, namely, $3/\pi$ when $pq = \frac{1}{4}$, and is a minimum namely, $3/4\pi$ when $pq = 0$.

3.5 *Case When p is Unknown*

We now estimate p by its usual estimate $\hat{p} = n/N$ and consider the test based on the statistic $[\bar{U}_N - E_{\hat{p}}(\bar{U}_N)]/\sigma_{\hat{p}}(\bar{U}_N)$, where $E_{\hat{p}}(\bar{U}_N)$ and $\sigma_{\hat{p}}(\bar{U}_N)$ are obtained by replacing p and q by $\hat{p}$ and $\hat{q}$ respectively, in (3.1.6) and (3.1.7). It is interesting to note that this test is not asymptotically distribution-free, in that it depends on the distribution of X.

Theorem 3.5.1.—Under the hypothesis $H: F_1 = F_0$, the limiting distribution of the statistic $[\bar{U}_N - E_{\hat{p}}(\bar{U}_N)]/\sigma_{\hat{p}}(\bar{U}_N)$, is normal with mean zero and variance $\{1 - [(\frac{3}{4})(1 - 2p)^2]/(1 - 3pq)\}$.

Proof.—Because plim $\hat{p} = p$, by an application of Slutsky's theorem, plim $(\hat{p}\hat{q}/pq)^{\frac{1}{2}} = 1$, which implies that plim $\sigma_{\hat{p}}(\bar{U}_N) = \sigma_p(\bar{U}_N)$. Hence

by a theorem of (1, *p.* 254), it follows that the asymptotic distribution of $[\bar{U}_N - E_{\hat{p}}(\bar{U}_N)]/\sigma_{\hat{p}}(\bar{U}_N)$ is the same as that of $[\bar{U}_N - E_{\hat{p}}(\bar{U}_N)]/\sigma_p(\bar{U}_N)$. Write

$$\frac{\bar{U}_N - E_{\hat{p}}(\bar{U}_N)}{\sigma_p(\bar{U}_N)} = \frac{\bar{U}_N - E_p(\bar{U}_N)}{\sigma_p(\bar{U}_N)} - \frac{E_{\hat{p}}(\bar{U}_N) - E_p(\bar{U}_N)}{\sigma_p(\bar{U}_N)}.$$

Since

$$\hat{p}\hat{q} - pq = (\hat{p} - p)(1 - 2p) - (\hat{p} - p)^2,$$

we have

$$\frac{\bar{U}_N - E_{\hat{p}}(\bar{U}_N)}{\sigma_p(\bar{U}_N)} = \frac{\bar{U}_N - E_p(\bar{U}_N)}{\sigma_p(\bar{U}_N)} - \frac{(\hat{p} - p)(1 - 2p)}{2\sigma_p(\bar{U}_N)} + \frac{(\hat{p} - p)^2}{2\sigma_p(\bar{U}_N)},$$

where $\sigma_p(\bar{U}_N)$ is given by (3.1.7). As $N^{\frac{1}{2}}(\hat{p} - p)/(pq)^{\frac{1}{2}}$ is bounded in probability and plim $|\hat{p} - p| = 0$ the third term in (3.5.1) tends in probability to zero. By Hoeffding's theorem (2, Theorem 7.2) the asymptotic joint distribution of the first two terms in (3.5.1) is bivariate normal $n(0, \Sigma)$ where $\Sigma = (\sigma_{ij})$ with

$$\sigma_{11} = 1, \sigma_{12} = \sigma_{21} = \sigma_{22} = \frac{[3(1 - 2p)^2]}{[4(1 - 3pq)]}.$$

This proves the result.

4. Rank Test for Dispersion

Now we consider a rank test for dispersion for the problem under consideration which is stated in Section 1. In this section we use the notation employed in the previous sections. For testing the hypothesis $H: F_1 = F_0$ against the alternative that F_1 and F_0 differ only in the scale parameter, we consider the test based on the statistic

$$W = \sum_{i=1}^{n} \left(r_i - \frac{N+1}{2}\right)^2,$$

where r_i denotes the rank of i-th ordered U observation in the combined sample of U's and V's. H is rejected if W is either too large or too small.

4.1 *Mean and Variance of W*

First we find the mean and variance of W under the hypothesis $H: F_1 = F_0$. It has been proved by Mood,[9] that the conditional moments of W for fixed n are,

$$E_p(W|n) = \frac{n(N^2-1)}{12},$$

$$\sigma_p^2(W|n) = \frac{n(N-n)(N+1)(N^2-4)}{180}.$$

Using (3.1.4) we obtain

$$E_p(W) = \frac{Np(N^2-1)}{12}, \tag{4.1.1}$$

$$\sigma_p^2(W) = \frac{pqN(N^2-1)(3N^2-7)}{240}. \tag{4.1.2}$$

Let

$$M_{ij} = \int [F_0(y)]^i [F_1(y)]^j \, dF_1(y).$$

To obtain $E(W)$ under the alternative note that

$$W = \sum_{i=1}^{n} r_i^2 - (N+1)\sum_{i=1}^{n} r_i + \frac{n(N+1)^2}{4} \tag{4.1.3}$$

and use the following results from (10),

$$E\left(\sum_{i=1}^{n} r_i \mid n\right) = n(N-n)\,M_{10} + \frac{n(n+1)}{2},$$

$$E\left(\sum_{i=1}^{n} r_i^2 \mid n\right) = 3n(N-n)\,M_{10} + n(N-n)^{(2)}\,M_{20} + 2n^{(2)}M_{11} + \tfrac{1}{6}n(n+1)(2n+1).$$

After using (3.1.4) to obtain $E_p\left(\sum_{i=1}^{n} r_i\right)$ and $E_p\left(\sum_{i=1}^{n} r_i^2\right)$, $E_p(W)$ is found out to be

$$E_p(W) = \frac{p}{4}N(N-1)^2 + \frac{1}{6}pN(N-1)(N-2)$$
$$[12pq\,M_{11} + 6q^2\,M_{20} - 6q\,M_{10} + 2p^2 - 3p]. \tag{4.1.4}$$

4.2 *Asymptotic Distribution of W*

Define three functions H, K and L as

$$H(Z_i, Z_j) = \begin{cases} 1, & \text{if } X_i = 0,\ X_j = 1 \text{ and } Y_i < Y_j; \\ 0, & \text{otherwise.} \end{cases}$$

$$K(Z_i, Z_j, Z_k) = \begin{cases} 1, & \text{if } X_i = 0,\ X_j = 0,\ X_k = 1 \text{ and } Y_i < Y_k, \\ & \quad Y_j < Y_k; \\ 0, & \text{otherwise,} \end{cases}$$

$$L(Z_i, Z_j, Z_k) = \begin{cases} 1, \text{ if } X_i = 0,\ X_j = 1,\ X_k = 1 \text{ and } Y_i < Y_k, \\ \qquad Y_j < Y_k; \\ 0, \text{ otherwise.} \end{cases}$$

Using (4.1.3) it is seen that W can be expressed in terms of H, K and L as

$$\frac{W}{N^3} = -\frac{N(N-1)}{N^2}\,[U_N^{(1)} - \bar{U}_N^{(2)} - 2\bar{U}_N^{(3)}] + \frac{1}{12N^3}\,[2n(n+1) + 3n(N+1)^2 - 6(N+1)\,n(n+1)] \tag{4.2.1}$$

where $\bar{U}_N^{(1)}$, $\bar{U}_N^{(2)}$, $\bar{U}_N^{(3)}$ are U-statistics defined by

$$\bar{U}_N^{(1)} = \frac{1}{N(N-1)} \sum_{j \neq i=1}^{N} H(Z_j, Z_i),$$

$$\bar{U}_N^{(2)} = \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k=1}^{N} K(Z_j, Z_k, Z_i),$$

$$\bar{U}_N^{(3)} = \frac{1}{N(N-1)(N-2)} \sum_{j \neq k \neq i=1}^{N} L(Z_j, Z_k, Z_i).$$

Theorem 4.2.1.—Let $T = W/N^3$. The asymptotic distribution of $[T - E_p(T)]/\sigma_p(T)$ is $n(0, 1)$ both under the hypothesis as well as the alternative.

Proof.—Observe that the second term of (4.2.1) converges in probability to $p(4p^2 - 6p + 3)/12$. By Hoeffding's theorem (2, Theorem 7.2) it follows that the asymptotic joint distribution of $\bar{U}_N^{(1)}$, $\bar{U}_N^{(2)}$, $\bar{U}_N^{(3)}$ is trivariate normal. The required theorem follows by an application of a theorem of (1, p. 254).‖

4.3. *Case When p is Unknown*

Here we estimate p by $\hat{p} = n/N$ and consider the test based on $[T - E_{\hat{p}}(T)]/\sigma_{\hat{p}}(T)$, where $E_{\hat{p}}(T)$ and $\sigma_{\hat{p}}(T)$ are obtained from (4.1.1) and (4.1.2) by replacing p by $\hat{p}$ and q by $\hat{q}$. It is interesting to note that this test is asymptotically distribution-free.

Theorem 4.3.1.—Under the hypothesis $H: F_1 = F_0$, the limiting distribution of $[T - E_{\hat{p}}(T)]/\sigma_{\hat{p}}(T)$ is $n(0, 4/9)$.

Proof.—Since plim $\hat{p} = p$, the limiting distribution of $[T - E_{\hat{p}}(T)]/\sigma_{\hat{p}}(T)$ is the same as that of $[T - E_{\hat{p}}(T)]/\sigma_p(T)$. Also

$$\frac{T - E_{\hat{p}}(T)}{\sigma_p(T)} = \frac{T - E_p(T)}{\sigma_p(T)} - \frac{E_{\hat{p}}(T) - E_p(T)}{\sigma_p(T)} \equiv a - b. \tag{4.3.1}$$

Note that after using the expressions for $E_{\hat{p}}(T)$, $E_p(T)$ and $\sigma_p(T)$, b in (4.3.1) can be written as

$$b = \frac{\left[\left(\frac{5}{9}\right) N\right]^{\frac{1}{2}} (\hat{p} - p)}{\sigma_p(T)} + c \equiv b_1 + c$$

where c converges in probability to zero. Note that a and b_1 are jointly asymptotically normally distributed with mean vector zero and covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{11} = 1$, $\sigma_{12} = \sigma_{21} = \sigma_{22} = 5/9$. Hence the theorem follows.

5. Two-Sample Run Test

For testing the hypothesis $H: F_1 = F_0$, combine the two samples of U's and V's and arrange them in the order of magnitude. Here we consider the test based on d, the total number of runs of U's and V's. The hypothesis H is rejected if d is too small. Mood[7] has given the exact sampling distribution of d under the hypothesis H when p is known and further proved that under the hypothesis H, the asymptotic distribution of $[d - 2Npq]/[2(Npq\{1 - 3pq\})^{\frac{1}{2}}]$ is $n(0, 1)$. These results are obtained by other authors, *see*, for example, Wishart and Hirshfeld,[14] Iyer.[3]

Here we consider the case when p is unknown. Consider the test based on $[d - 2N\hat{p}\hat{q}]/[2(N\hat{p}\hat{q}\{1 - 3\hat{p}\hat{q}\})^{\frac{1}{2}}]$ where $\hat{p} = n/N$. It is proved in the following theorem, that the test is not asymptotically distribution-free in that the limiting distribution of the statistic depends on p.

Theorem 5.1

Under the hypothesis $H: F_1 = F_0$ the asymptotic distribution of $(d - 2N\hat{p}\hat{q})/[2\{(N\hat{p}\hat{q}(1 - 3\hat{p}\hat{q})\}^{\frac{1}{2}}]$ is normal with mean zero and variance $1 - (1 - 2p)^2/(1 - 3pq)$.

Proof.—As in Theorem 3.5.1 the symptotic distribution of $(d - 2N\hat{p}\hat{q})/[2(N\hat{p}\hat{q}\{1 - 3\hat{p}\hat{q}\})^{\frac{1}{2}}]$ is the same as that of $(d - 2N\hat{p}\hat{q})/[2(Npq\{1 - 3pq\})^{\frac{1}{2}}]$. Since $\hat{p}\hat{q} - pq = (\hat{p} - p)(1 - 2p) - (\hat{p} - p)^2$, we can write

$$\frac{d - 2N\hat{p}\hat{q}}{2[Npq(1 - 3pq)]^{\frac{1}{2}}} = \frac{d - 2Npq}{2[Npq(1 - 3pq)]^{\frac{1}{2}}} - \frac{N^{\frac{1}{2}}(\hat{p} - p)(1 - 2p)}{[pq(1 - 3pq)]^{\frac{1}{2}}} + \frac{N^{\frac{1}{2}}(\hat{p} - p)^2}{[pq(1 - 3pq)]^{\frac{1}{2}}} \tag{5.1}$$

It can be shown that asymptotic joint distribution of the first two terms in the R.H.S. of (5.1) is $n\ (0, \Sigma)$ with

$$\Sigma = (\sigma_{ij}),\ \sigma_{11} = 1,\ \sigma_{12} = \sigma_{21} = \sigma_{22} = (1 - 2p)^2/(1-3pq).$$

Also noting that the 3rd term in the R.H.S. of (5.1) converges in probability to zero, the required theorem follows.

6. Acknowledgement

The author wishes to thank Prof. Ingram Olkin for suggesting the problem and for his keen interest in its solution.

7. References

1. Cramér, H. .. *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

2. Hoeffding, W. .. "A class of statistics with asymptotically normal distribution," *Ann. of Math. Stat.*, 1948, **19**, 293-325.

3. Iyer, P. V. K. .. "Further contributions to the theory of probability distributions of points on a line—I," *Jour. Indian Soc. Agri. Stat.*, 1950, **2**, 141-60.

4. Kamat, A. R. .. "A two-sample distribution free test," *Biometrika*, 1956, **43**, 377-88.

5. Lehmann, E. L. .. "The power of rank tests," *Ann. of Math. Stat.*, 1953, **24**, 23-43.

6. Mann, H. B. and Whitney, D. R. "On a test whether one of the two random variables is stochastically larger than the other," *Ibid.*, 1947, **18**, 50-60.

7. Mood, A. M. .. "Distribution theory of runs," *Ibid.*, 1940, **11**, 367-92.

8. ——— .. *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1950.

9. ——— .. "On the asymptotic efficiency of certain nonparametric two-sample tests," *Ann. of Math. Stat.*, 1954, **25**, 514-22.

10. Sukhatme, Balakrishna V. "On certain two-sample non-parametric tests for variances," *Ibid.*, 1957, **28**, 188-94.

11. ——— .. "A two-sample distribution-free test for comparing variances," *Biometrika*, 1958, **45**, 544-50.

12. Tate, R. F. .. "Correlation between a discrete and continuous variable. Point-biserial correlation," *Ann. of Math. Stat.*, 1954, **25**, 603-07.

13. Wald, A. and Wolfowitz, J. "On a test whether two samples are from the same population, *Ibid.*, 1940, **11**, 147-62.

14. Wilcoxon, F. .. "Individual comparisons by ranking methods," *Biometrics*, 1945, **1**, 80-83.

15. Wishart, J. and Hirshfeld, H. O. "A Theorem concerning the distribution of joins between line segments," *London Math. Soc. Journal*, 1936, **11**, 227.