A QUICK NON-PARAMETRIC TWO-SAMPLE TEST FOR COMPARING VARIANCES

By

Vishnu Dayal Jha

Punjabi University, Patiala

(Received : February, 1977)

1. Introduction

Let $X_1, X_2, ..., X_{n_1}$ and $Y_1, Y_2, ..., Y_{n_2}$ be independent ordered samples from two populations with cumulative distribution functions $F(x)$ and $G(y)$ respectively. It is assumed that the respective density functions are differentiable in the neighbourhood of the population quantiles. We also assume that the populations are unimodal and absolutely continuous having the same functional form with identical locations but possibly different variances σ_x^2 and σ_y^2. Without any loss of generality we assume that the population medians are zero. With these assumptions, we consider the problem of testing the hypothesis :

$$H: \sigma_x^2 = \sigma_y^2$$

against

$$A: \sigma_x^2 > \sigma_y^2$$

For testing this hypothesis, in the case when the above two samples come from normal populations, one can use the usual variance ratio F-test but its use is inappropriate in case we ignore the normality assumption since this test lacks robustness with respect to normality. In this contest, Geary [7] and Gayen [6] showed that the F-test is sensitive to changes in kurtosis from the normal distribution theoretical value of zero. Further, in view of the asymptotically distribution-free nature of the F-statistic when suitably normalized, Box & Anderson [1], [3] have shown on the basis of extensive sampling experiments that the F-statistic is insensitive to departures from normality, at least for large samples. Mood [9], Lehmann [8] and Sukhatme [14] constructed non-parametric tests for the above problem. In this paper, based on the intuition that in case the variation in X-population is greater than that in Y-population, then the difference between any two order statistics of X will also be obviously greater than that between the corresponding two order statistics of Y, an asymptotic non-parametric test has been developed by using a finite number of quantiles and their neighbouring observations. In this context, the question of optimality of number of quantiles and the number of neighbouring observations to be used in applying this test has been examined. Further, the asymptotic power of this test has been studied and an example given at the end to illustrate its application. The test is quick in application but appears to be poor theoretically. However, a study on the power efficiency of this test with respect to different non-parametric tests due to Mood, Lehmann and Sukhatme is presently under progress and its results are proposed to be presented in a forthcoming publication of the author.

2. Notations

Consider k non-negative numbers $p_1, p_2, \ldots, p_k$ such that $o<p_1<p_2<\ldots.p_k<1$. Let η_i denote the X-population quantile of order p_i defined by $\eta_i=F^{-1}(p_i)$ for $i=1, 2,\ldots, k$ with X_{α_i} denoting the corresponding sample quantile, where $\alpha_i=[n_1p_i]+1$ and $[n_1p_i]$ is the Guass symbol denoting the largest integer not exceeding n_1p_i. Similarly the Y-population quantile ξ_i is defined by $\xi_i=G^{-1}(p_i)$ where Y_{β_i} denotes the corresponding sample quantile and $\beta_i=[n_2p_i]+1$. Further let

$$U^{(i)}_{n_1,m_{1i}}=\frac{n_1}{2m_{1i}}(X_{\alpha_i+m_{1i}}-X_{\alpha_i-m_{1i}})$$

and

$$V^{(i)}_{n_2,m_{2i}}=\frac{n_2}{2m_{2i}}(Y_{\beta_i+m_{2i}}-Y_{\beta_i-m_{2i}})$$

The notation $X_n\xrightarrow{\alpha}N(\mu, \sigma^2)$ used in the sequel denotes convergence to normal distribution with mean μ and variance σ^2. For notational simplicity we take both the samples to be equal in size, say $n_1=n_2=n$, in which case $\alpha_i=\beta_i$. Also let us take $m_{1i}=m_{2i}=m_i$.

3. The Proposed $T_n^{(k)}$ Test

The test statistic is defined as

$$T_n^{(k)}=\sum_{i=1}^{k}(U^{(i)}_{n,m_i}-V^{(i)}_{n,m_i})$$

with the test function

$$\psi\left(T_n^{(k)}\right)=\begin{cases}1 & \text{if } T_n^{(k)}>\alpha\\ 0 & \text{otherwise}\end{cases}$$

where the critical value t_α for α $(0<\alpha<1)$ is so chosen for large n that $E\left[\psi\left(T_n^{(k)}\right)\right]=\alpha$, the level of significance.

According to the proposed test, $2k$ observations are to be taken from each sample by choosing k to be very small in comparison to m_i. In this context, one has to decide about the optimal spacing of quantiles and optimum choice of m_i's. For determining the optimal spacing of quantiles or equivalently for choosing the spacings suitably in such a manner that the relative efficiency of the estimator attains the maximum value, one can make use of the procedure due to Ogawa [10] (*see* Sarhan & Greenberg [12]). According to this procedure, for an optimum estimator of the standard deviation σ of a normal distribution on the basis of a large sample in case of known population location, the α_i must be so chosen as to minimize the mean square error of σ^*, the estimator of σ. Further regarding optimal choice of m_i's, Ogawa proved that for any symmetric population, the symmetric spacing of quantiles is optimum. Taking this aspect into consideration alongwith the assumptions made about the two populations, we therefore prefer to use the symmetric spacing of quantiles.

Regarding determination of the optimal value of m_i which minimizes the asymptotic mean square error $E\left(U_{n,m_i}^{(i)}-\frac{1}{f(\eta_i)}\right)^2$, we can use the results of various authors as are quoted below without proofs :

Theorem 3.1. [Bloch & Gastwirth (1968)]. Under the assumption that the first three derivatives of $f(x)$ exist in the neighbourhood of η_i $(i=1, 2,\ldots, k)$, the optimal value of m_i is $c_i n^{4/5}$, where,

$$c_i=\left(\frac{9f^8(\eta_i)}{2[3\{f'(\eta_i)\}^2-f(\eta_i)\,f''(\eta_i)]^2}\right)^{\frac{1}{5}}.$$

In order to use this result for a decision on the optimal choice of m_i's, we need to have prior knowledge of the values of $f(\eta_i)$, $f'(\eta_i)$ and $f''(\eta_i)$ but in case the functional form of the distribution is unknown, the choice of m_i's has to be based on the following theorems:

Theorem 3.2. (Bloch and Gastwirth [2]). If $m_i = o(n)$ and $m_i \to \infty$ for all i, then the statistic $U^{(i)}_{n,m_i}$ is a consistent estimator of $1/f(\eta_i)$.

Theorem 3.3. (Siddiqui [13]). Under the conditions of Theorem 3.2, the variate

$$\sqrt{2m_i}\left[U^{(i)}_{n,m_i} - \frac{1}{f(\eta_i)}\right] \Big/ \frac{1}{f(\eta_i)}$$

approximates the standard normal variate and

$$\underset{i \neq j}{Cov}\left(U^{(i)}_{n,m_i}, U^{(j)}_{n,m_j}\right) \to 0 \text{ as } m_i, m_j, n \to \infty$$

Since for fixed m_i and large n, the statistic $2m_i\ [f(\eta_i)]\ (U^{(i)}_{n,m_i})$ is distributed like chi-square with $2m_i$ degrees of freedom, for practical purposes the normal approximation as given in theorem 3.3 holds for $m_i \geqslant 15$. Thus $m_i \geqslant 15$ can be chosen in such a way that $X_{\alpha_i - m_i}$ and $X_{\alpha_i + m_i}$ should converge in probability to the population quantile η_i, Further the number of quantiles k is to be chosen in such a way that

$$X_{\alpha_i + m_i} < X_{\alpha_{i+1} - m_{i+1}} \text{ for } i = 1, 2, \ldots, k.$$

In that case the variates $U^{(1)}_{n,m_1}, U^{(2)}_{n,m_2}, \ldots, U^{(k)}_{n,m_k}$ are asymptotically independent by Theorem 3.3. Suppose that for an ordered sample of size 50, we take $k=2$ with $m_1 = m_2 = 15$ then $\alpha_1 = 17$ and $\alpha_2 = 34$. Now the order statistics to be used are X_2, X_{32}, X_{19} and X_{49}. Here $X_{32} \nless X_{19}$. Also since the order statistic X_{19} will converge to the population quantile η_1 $[F(\eta_1) = \frac{1}{3}]$ as against η_2 $[F(\eta_2) = \frac{2}{3}]$, we cannot take $k=2$ in this case but can take $k=1$.

Now $U^{(1)}_{n,m_1}, \ldots, U^{(k)}_{n,m_k}$ are asymptotically independent and so are $V^{(1)}_{nm_1}, \ldots, V^{(k)}_{n,m_k}$. Also since

$$U^{(i)}_{n,m_i} \xrightarrow{\alpha} N\left(\frac{1}{f(\eta_i)}, \frac{1}{f(\eta_i)\sqrt{2m_i}}\right)$$

and

$$X^{(i)}_{n,m_i} \xrightarrow{\alpha} N\left(\frac{1}{g(\xi_i)}, \frac{1}{g(\xi_i)\sqrt{2m_i}}\right)$$

then

$$U^{(i)}_{n,m_i} - V^{(i)}_{n,m_i} \xrightarrow{\alpha} N\left\{\left(\frac{1}{f(\eta_i)} - \frac{1}{g(\xi_i)}\right), \sqrt{\left[\frac{1}{2m_i}\left(\frac{1}{f^2(\eta_i)} + \frac{1}{g^2(\xi_i)}\right)\right]}\right\}$$

Hence

$$T^{(k)}_n \xrightarrow{\alpha} N\left\{\sum_{i=1}^{k}\left(\frac{1}{f(\eta_i)} - \frac{1}{g(\xi_i)}\right), \sqrt{\left[\sum_{i=1}^{k}\frac{1}{2m_i}\left(\frac{1}{f^2(\eta_i)} + \frac{1}{g^2(\xi_i)}\right)\right]}\right\}$$

or

$$Z^{(k)}_n = \frac{T^{(k)}_n - \sum_{i=1}^{k}\left(\frac{1}{f(\eta_i)} - \frac{1}{g(\xi_i)}\right)}{\sqrt{\left[\sum_{i=1}^{k}\frac{1}{2m_i}\left(\frac{1}{f^2(\eta_i)} + \frac{1}{g^2(\xi_i)}\right)\right]}} \xrightarrow{\alpha} N(0,1)$$

So that under the hypothesis $H : \alpha_x^2 = \alpha_y^2$ or equivalently $H : f(\eta_i) = g(\xi_i)$, $i = 1, 2, \ldots, k$ the variate

$$Z^{(k)}_n = \frac{T^{(k)}_n}{\sqrt{\left[\sum_{i=1}^{k}(m_i f^2(\eta_i))^{-1}\right]}} \xrightarrow{\alpha} N(0, 1)$$

4. Studentization

It is to be noted that the asymptotic variance of the test statistic $T^{(k)}_n$ and as such its asymptotic distribution depends on the functional form of the distribution function $F(x)$ under the hypothesis H. In order to make the test statistic $T^{(k)}_n$ distribution-free, we use the consistent estimator of $f(\eta_i)$. Since $U^{(i)}_{n,m_i}$ is a consistent estimator of $(f(\eta_i))^{-1}$ by Theorem 3.2, using a convergence theorem due to Slutsky [*see* Cramer (1946), pp. 254-55], $\sqrt{\left[\sum_{i=1}^{k}\frac{1}{m_i}(U^{(i)}_{n,m_i})^2\right]}$, is a consistent estimator of

$$\sqrt{\left[\sum_{i=1}^{k}\{m_i f^2(\eta_i)\}^{-1}.\right]}$$

Hence a modified test statistic $Z_n^{*(k)}$ under the hypothesis H given by

$$Z_n^{*(k)} = \frac{T_n^{(k)}}{\sqrt{\left[\sum_{i=1}^{k}\frac{1}{m_i}(U_{n,m_i}^{(i)})^2\right]}}$$

has an asymptotic standard normal distribution and it is thus asymptotically distribution-free. We can thus perform the test of hypothesis $H: \sigma_x^2 = \sigma_y^2$ against the alternative $A: \sigma_x^2 > \sigma_y^2$ with the help of this studentized statistic for which we have the ultimate test function

$$\psi(T_n^{(k)}) = \begin{cases} 1 & \text{if } Z_n^{*(k)} > t_{n,\alpha} \\ 0 & \text{otherwise} \end{cases}$$

and where the sequence $t_{n,\alpha}$ is such that $\lim_{n\to\infty} t_{n,\alpha}=t_\alpha$, while t_α for $0<\alpha<1$ is such that $1-\Phi(t_\alpha)=\alpha$, the level of significance, and where $\Phi(t)$ denotes the standard normal distribution function.

5. Asymptotic Relative Efficiency and Asymptotic Power

It was shown in section 3 that $T_n^{(k)}$ is asymptotically normally distributed both under the hypothesis H and the alternative A. By considering the alternative $A: \theta=\theta_n$, where $\theta_n=1-\frac{d}{\sqrt{n}}, d>0$ such that for $n\to\infty$, $\theta_n\to 1$, it can be easily verified that all the conditions of Pitman-Noether's theorem (*see* Puri & Sen [11], pp. 113-114) are satisfied. We can therefore compute the asymptotic relative efficiency of the $T_n^{(k)}$ test with respect to the variance ratio F-test.

Let $G(x)=F(\theta x)$. Then the hypothesis $H: \sigma_x^2 = \sigma_y^2$ is equivalent to the alternative hypothesis A with $\theta=1$. The mean under the alternative A for large n is

$$E(T_n^{(k)}) = \sum_{i=1}^{k}\left(1-\frac{1}{\theta}\right)\frac{1}{f(\eta_i)}$$

so that

$$\left(\frac{dT_n^{(k)}}{d\theta}\right)_{\theta=1}=\sum_{i=1}^{k}\frac{1}{f(\eta)_i}$$

The variance of $T_n^{(k)}$ under the hypothesis H is given by

$$\text{Var}\,(T_n^{(k)})=\sum_{i=1}^{k}[m_i f^2(\eta_i)]^{-1}$$

Thus as in Sukhatme (1957), the efficiency of the $T_n^{(k)}$ test is given by

$$\frac{\left(\sum_{i=1}^{k}\frac{1}{f(\eta_i)}\right)^2}{\sum_{i=1}^{k}[m_i f^2(\eta_i)]^{-1}}$$

It is well known (*see* Geary [7], Sukhatme [14]) that for two samples of equal size n, the efficiency of the variance ratio F-test is $2n(\beta_2-1)$, where β_2 denotes the kurtosis. Hence the asymptotic relative efficiency of the $T_n^{(k)}$ test with respect to the F-test denoted by $ARE\ (T^{(k)}_k, F)$ is

$$ARE\,(T_n^{(k)}, F)=\frac{\left(\sum_{i=1}^{k} f[(\eta_i)]^{-1}\right)^2}{\sum_{i=1}^{k}[m_i f^2(\eta_i)]^{-1}}\left\{\frac{\beta_2-1}{2n}\right\}$$

For the normal distribution if we take $k=1$ and η_i as the population median, then writing m_i and c_i for $i=1$ as m and c respectively, we have

$$ARE\ (T_n^{(k)}, F)=\frac{m}{n}=\frac{c\,n^{4/5}}{n}=\frac{0\cdot5}{n^{1/5}}$$

For the double exponential distribution with $pdf\, f(x)=\frac{1}{2}\, exp\,(-|x|)$, if we take $k=1$ and η_i as the population median, then

$$ARE\,(T_n^{(k)}\,,\,F)=\frac{5m}{2n}$$

In a similar way the ARE of $T_n^{(k)}$ test can also be obtained with respect to Mood's test, Lehmann's test and Sukhatme's test.

Further it is of interest to study the limiting power function of the test. By a straightforward calculation, it can be seen that the statistic $Z_n^{*(k)}$ under the alternative A approximates the normal distribution with mean

$$\frac{\sum_{i=1}^{k}\left(\frac{\theta-1}{\theta}\right)\frac{1}{f(\eta_i)}}{\sqrt{\left[\sum_{i=1}^{k}\frac{1}{m_i}\left(U_{n,m_i}^{(i)}\right)^2\right]}}=\frac{\mu(\theta,\,k)}{\sigma(k)},\text{ say}$$

and variance

$$\frac{\sum_{i=1}^{k}\frac{1}{2m_i}\left(1+\frac{1}{\theta^2}\right)\left(U_{n,m_i}^{(i)}\right)^2}{\sum_{i=1}^{k}\frac{1}{m_i}\left(U_{n,m_i}^{(i)}\right)^2}=\frac{\sigma^2(\theta,k)}{\sigma^2(k)},\text{ say}$$

Thus the limiting power function $P_\phi(\theta)$ is given by

$$P_\psi(\theta)=\int_{t_\alpha}^{\infty}\frac{1}{\sqrt{2\pi}}\,\frac{\sigma(k)}{\sigma(\theta,k)}\,exp\left(-\frac{1}{2}\frac{\sigma^2(k)}{\sigma^2(\theta,k)}\left(t-\frac{\mu(\theta,k)}{\sigma(k)}\right)^2\right)dt$$

$$=1-\Phi\left[\frac{\sigma(k)}{\sigma(0,k)}\left(t_\alpha-\frac{\mu(\theta,k}{\sigma(k)}\right)\right]$$

Consistency of the Test. Let us consider the statistic $Z_n^{*(k)}/\sqrt{m}$ rather than $Z_n^{*(k)}$. Since

$$E\left(Z_n^{*(k)}/\sqrt{m}\right)\simeq\frac{\mu(\theta,k)}{\sqrt{m}\,\sigma(k)}$$

and Var $\left(Z_n^{*(k)}/\sqrt{m}\right)\simeq\frac{1}{m}\,\frac{\sigma^2(\theta,k)}{\sigma^2(k)}\longrightarrow 0$ as $m\longrightarrow\infty$

Applying the Chebyshev inequality with $\delta>0$ and arbitrarily small, we find that

$$P\left(\left|\frac{Z_n^{*(k)}}{\sqrt{m}}-\frac{\mu(\theta,k)}{\sqrt{m}\ \sigma(k)}\right|>\delta\right)\leqslant\frac{\mathrm{Var}\left(Z_n^{*(k)}\right)}{m}\to 0 \text{ as } m\to\infty$$

So $Z_n^{*(k)}/\sqrt{m}$ converges in probability to $\mu(\theta,k)/\sqrt{m}\ \sigma(k)$ as $m\to\infty$,

Since under the hypothesis $H:\theta=1$, or $\frac{\mu(\theta,k)}{\sqrt{m}\ \sigma(k)}=0$, the test that rejects H for large positive values of $T_n^{(k)}$ is consistent against the alternative for which $\theta>1$.

6. Illustrative Example

Suppose we have two independent samples of size 50 each and we arrange their elements in ascending order of magnitude as shown below :

Sample A

−1.156,	−1.141,	−1.086,	−0.998,	−0.964,	−0.780,	−0.708,	−0.703
−0.665,	−0.629,	−0.596,	−0.565,	−0.556,	−0.537,	−0.469,	−0.451
−0.434,	−0.347,	−0.337,	−0.299,	−0.262,	−0.189,	−0.149,	−0.142
−0.119,	0.026,	0.159,	0.165,	0.205,	0.264,	0.337,	0.439
0.548,	0.564,	0.666,	0.680,	0.690,	0.811,	0.852,	1.146
1.156,	1.275,	1.305,	1.346,	1.455,	1.613,	1.709,	1.899
2.323,	2.480						

Sample B

−2.714,	−1.400,	−1.169,	−1.083,	−1.017	−1·016,	−0.837,	−0.828
−0.710,	−0.644,	−0.597,	−0·539,	−0.432,	−0.416,	−0.378,	−0.318
−0.313,	−0.310,	−0.305,	−0.275,	−0.220,	−0.078,	−0.068,	−0.059
−0.007,	+0.019,	0.060,	0.071,	0.121,	0.194,	0.209,	0.229
0.239,	0.266,	0.311,	0.471,	0.506,	0.606,	0.610,	0.733
0.738,	0.744,	0.824,	0.921,	0.925,	1.045,	1.115,	1.254
1.747,	1.774						

Here $n_1=n_2=n=50$, and we take $m=15$, accordingly $k=1$, since $i=1$ in which case we write $U^{(i)}_{n,m_i}=U_{n,m}$ and $V^{(i)}_{n,m_i}=V_{n,m}$. Then for sample A :

$$U_{n,m}=\frac{n}{2m}(X_{\alpha_i+m}-X_{\alpha_i-m}) =\frac{50}{30}(1.146+0.596)=2.903$$

and $$\sqrt{[U^2_{n,m}/15]}=0.743.$$

Likewise for sample B :

$$V_{n,m}=\frac{50}{30}(0.733+0.597)=2.217$$

so that we can compute the value of the test statistic

$$Z_n^{*(k)}=\frac{U_{n,m}-V_{n,m}}{\sqrt{[U^2_{n,m}/15]}}=0.923$$

which being less than the critical normal value $t_{0.05}=1.645$ suggests that these two samples come from the populations with equal variances. The observed difference between $U_{n,m}$ and $V_{n,m}$ can be explained as being due to sampling fluctuations. As a matter of fact, we constructed this example by taking two independent samples of size 50 each from the Table (Fisher & Yates [5]) of random normal numbers having mean $\mu=0$ and standard deviation $\sigma=1$, and we obtained the result as expected.

SUMMARY

A quick non-parametric test has been developed in this paper by using a finite number of quantiles and their neighbouring observations for comparing the variances of two samples drawn from two populations both assumed to be unimodal and absolutely continuous having the same functional form with identical locations but possibly different variances. The power of the test and its asymptotic relative efficiency has been studied and in addition an example given at the end to illustrate its application.

ACKNOWLEDGEMENT

I wish to thank the refrees for their suggestions and comments.

REFERENCES

[1] Anderson, S.L. and Box, G.E.P. (1955) : Permutation theory in the derivation of robust criterion and the study of departures from assumptions *Jour. Roy. Statist. Soc. B* **17**, 1-34.

[2] Bloch, D.A. and Gastwirth, J.L. (1968) : On a simple estimate of reciprocal of the density function. *Ann. Math. Statist.* **39**, 1083-85.

[3] Box, G.E.P. (1953) : Non-normality and tests on variances. *Biometrika.* **40**, 318-35.

[4] Cramar, H. (1946) : *Mathematical Methods of Statistics.* Princeton University Press.

[5] Fisher, R.A. and Yates, F. (1963) : *Statistical Tables for Biological and Medical Research.* Oliver and Boyd. London.

[6] Gayen, A.K. (1950) : The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika.* **37**, 236-55.

[7] Geary, R.C. (1947) : Testing for normality. *Biometrika.* **34**, 209-41.

[8] Lehmann, E.L. (1951) : Consistency and unbiasedness of certain non-parametric tests. *Ann. Math. Statist,* **22**, 165-79.

[9] Mood, A.M. (1954) : On the asymptotic efficiency of certain non-parametric tests. *Ann. Math. Statist.* **25**. 514-22.

[10] Ogawa, J. (1951) : Contribution to the theory of systematic statistics. *Osaka Math. Jour.* **3** No. 2, 175-213.

[11] Puri, M.L. and Sen, P.K. (1971) : *Non-parametric Methods in Multivariate Analysis.* John Wiley and Sons. Inc., New York.

[12] Sarhan, A.E. and Greenberg, B.G. (1962) : *Contributions to Order Statistics.* John Wiley and Sons. Inc., New York.

[13] Siddiqui, M.M. (1960) : Distribution of quantiles in samples from a bivariate population. *Jour. Res. Nat. Bur. Stand.* **64 B**, 145-50.

[14] Sukhatme, B.V. (1957) : On certain two-sample non-parametric tests for variances. *Ann. Math. Statist.* **28**, 188-94.