NON-PARAMETRIC TWO SAMPLE TESTS OF STATISTICAL HYPOTHESES

by

Everett Edgar Hunt

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in the Department

of

MATHEMATICS

We accept this thesis as conforming to the standard required from candidates for the degree of MASTER OF ARTS.

Members of the Department of Mathematics

THE UNIVERSITY OF BRITISH COLUMBIA
April, 1951

Abstract

The testing of statistical hypotheses concerning two populations consists in determining the relationship between the cumulative distribution functions on the basis of random samples from each population. In the non-parametric case the only assumption made regarding the populations is that the two c.d.f's. are continuous. Thus the distribution of any statistic proposed to test the two samples must be independent of the functional form of the c.d.f's. One method of approach is based on the order relations of the sample values. A survey is made of such tests recently proposed and a new test is suggested based on sampling without replacement from a population of the positive integers 1, 2, 3, ... N .

Table of Contents

Introduction

The numbers which characterize the distribution of a population or universe are called population parameters. In most cases which arise in practice it is impossible to determine the values of these parameters. Thus they are predicted or estimated by statistics which are functions of the sample values drawn from the population. In the past fifty years a general theory of estimating these parameters and of testing hypotheses concerning their values has been developed [2].

One important problem which has received much attention is to test whether two random samples are drawn from the same population. Tests of this hypothesis are based on the classical Student's t distribution which gives a criterion for testing whether the difference between two sample means is significant and on the F distribution which tests whether the difference between the variances is significant. Both these tests and most of the others in common use assume that the population distributions are normal. Since this hypothesis is very restrictive much effort has been expended by statisticians in attempting to show that the commonly used distributions are at least asymptotically normal. However, not all distributions have this property and further, if the sample is small, the normality assumption will not hold even approximately.

An important statistical problem, then is to derive methods which can be used to test hypotheses assuming nothing about the population distributions except that the cumulative distribution functions are continuous. Such tests are termed non-parametric or distribution free. These tests use qualitative rather than quantitative aspects of the sample values. For example, instead of setting up a criterion to test the difference between the means and variances of the two samples, a test criterion is established concerning the rank or order relations of the data.

It may be argued that the efficiency of a test is reduced by neglecting quantitative relations since all of the available information has not been utilized. This loss in efficiency should be judged against the possibility of making an incorrect assumption concerning the normality of the population distribution. For this reason non-parametric tests have a place in the theory of testing hypotheses.

A good test should have a high probability of rejecting a false hypothesis. The power of a test is the probability of rejecting the null hypothesis when actually it is false and an alternative hypothesis is true [2]. Thus the power is a function of the parameters of the distribution involved in the true alternative hypothesis. Therefore, in non-parametric theory a difficulty arises. However, an alternative method of evaluating a test has been proposed. A test is called consistent if the probability of rejecting a false null hypothesis against certain alternatives approaches one as the size of the sample increases indefinitely [7]. Thus a test may be consistent with respect to one particular alternative hypothesis but not to others.

Many new non-parametric tests for comparing two samples have been proposed recently. The object of this paper is to present a survey of these tests and put forward another.

Classification of non-parametric tests based on order relations of the sample values

By order relations of the sample values is meant the ordered set of values in a random sample from least to greatest. Non-parametric two sample tests using this property can be considered as being one of three types:

i) those based on a comparision of the two population distributions along the whole real line,

ii) those based on a comparison at a finite number of fixed points such as the quantile points of the distributions,

iii) those based on the method of randomization.

In what follows representative tests of these three types are considered.

The Wald-Wolfowitz Run Test

A test of the first type is the Run test of A. Wald and J. Wolfowitz [7]. Let O_m be a sample of observations $X_1, X_2, \ldots X_m$ from a population with continuous cumulative distribution function, $F(X)$ and let O_n be a sample $Y_1, Y_2, \ldots Y_n$ from a population with continuous distribution function, $G(X)$. It is required to derive a test of the null hypothesis that $F(X) = G(X)$. Let O_{m+n} denote the combined sample, the observations being ordered from the least to the greatest.

$$O_{m+n} : Z_1, Z_2, \ldots Z_{m+n} \quad \text{where} \quad Z_i < Z_{i+1}$$

Wald and Wolfowitz proceed as follows: replace Z_i in O_{m+n} by zero or by one depending on whether Z_i comes from the sample O_m or from sample O_n. Define a run to be a sequence of zeros uninterrupted by ones or a sequence of ones uninterrupted by zeros and consider the number of runs in O_{m+n}. The statistic proposed in this test is U, the number of runs.

Naturally before any statistic can be used as a test criterion, its distribution function must be determined. Under the null hypothesis that $F(X) = G(X)$, the distribution of U will be the probability of obtaining a particular number of runs under the assumption that all of the arrangements of the m values of the O_m sample, and all of the arrangements of the n values of the O_n sample have equal probabilities. This probability is the ratio of the number of the arrangements of the X's and the Y's with $m + n$, m, n and U held fixed to the total number of arrangements with $m + n$, m, n constant.

The denominator of this ratio is $C(m + n, n)$ since this is the number of arrangements of $m + n$ elements, m of which are alike and n of which are alike.

To determine the numerator of the ratio, two cases must be considered according as U is odd or even. First, let $U = 2k$. Then there will be k runs of zeros and also k runs of ones in any arrangement in which the exact number of runs equals U. Now the problem of determining the number of arrangements of K runs with m x's is the same as that of finding the number of ways of putting m zeros into k cells, none of which is empty. Consider the cells to be spaces between $k + 1$ bars. Then since each arrangement must start and end with a bar, there are $k - 1$ remaining bars to permute. Further, since the cells are non-empty there must be at most one bar between any two zeros. Thus there are $m - 1$ spaces between the zeros and $k - 1$ places to put the bars, hence there are $C(m - 1, k - 1)$ possible arrangements. Similarly, the number of ways of obtaining exactly k runs with n y's is equal to $C(n - 1, k - 1)$. Now for every given arrangement of the X's, there are two arrangements possible with the Y's depending on whether the combination O_{m+n} begins with zero or one. Then the probability that there are exactly U runs $(U = 2k)$ equals

$$\frac{2\,C(m-1,\ k-1)\ C(n-1,k-1)}{C(m+n,\ n)}$$

For the case $U = 2k + 1$, there are either $k + 1$ runs of the X's and k runs of the Y's or k runs of the X's and $k + 1$ runs of the Y's. Then the probability of $U = 2k + 1$ runs equals

$$\frac{C(m-1,\ k)\ C(n-1,\ k-1) + C(m-1,\ k-1)\ C(n-1,\ k)}{C(m+n,\ n)}$$

The region of rejection for the null hypothesis consists of the values of U such that $U \leq U_\alpha$ where U_α, the critical value of U_α depends on the level of significance α that is desired by the experimenter. U_α is pre-determined and is such that $\text{Prob}\ (U \leq U_\alpha) = \alpha$.

Thus small values of U are judged significant implying that when there are too few runs, there is poor mixing of the data of the two samples. The worst case would occur when $U = 2$. This would mean that all the observations of the one sample are greater than those of the other.

Tables giving values of U_α for $m, n \leq 20$ at the .005, .01, .025 significance levels have been prepared by F. Swed and C. Eisenhart [6]. Values of U_α for $m, n > 20$ have not been computed. However, since the distribution of U has been proved asymptotically normal with mean

$$\frac{2mn}{m+n} + 1$$

and variance

$$\frac{2mn\ (2mn - m - n)}{(m+n)^2(m+n-1)}$$

the critical values can be computed approximately for large samples [7].

The Run test has been shown to be consistent with respect to alternative hypotheses with minor restrictions [7]. Let m, n increase without limit such that the ratio, $m/n = \lambda$, a constant. The expected value of U is approximately $2m/(1 + \lambda)$ when the null hypothesis, $F(X) = G(X)$ is true. The statistic, U/m converges stochastically to its expected value, $2/(1 + \lambda)$ under the null hypothesis. This means that the probability of the expected value of U/m differing from $2/(1 + \lambda)$ by less than any given amount approaches one as m increases indefinitely. Then it is shown that under true alternative hypotheses, U/m converges to its expected value which is less than $2/(1 + \lambda)$. Thus

$$\text{Prob}\ (U/m < 2/(1 + \lambda)\ ) \to 1$$

if the null hypothesis is false.

The following example illustrates the use of the Run test. Given the two samples (5.8, 2.9, 7.2, 3.1, 2.5, 6.1) and (4.9, 3.3, 5.7, 4.1, 4.6, 5.6), test the hypothesis that these are random samples drawn from the same population about which nothing is assumed except that it is continuous. Combine the date and order the values from the least to the greatest. Then assigning the values 0 and 1 to the observations according as they come from the first or second sample, we obtain 000111111000. The observed value of U is 3. From tables [6] the critical value $U_{.05} = 3$ for $m,n = 6$. Thus $U = 3$ is significant and the null hypothesis is rejected on the basis of this particular example.

The Mathisen Test

The following test proposed by H.C. Mathisen [4] is an example of the second type of order relations tests. Two methods of comparing the samples are considered, one involving the median and the other, the quartiles.

Let O_{2n+1} be a sample composed of $2n + 1$ elements drawn from a continuous population. The sample values $X_1, X_2, \ldots X_{2n+1}$ are ordered so that $X_i < X_{i+1}$. The median of the $2n + 1$ observations is X_{n+1}. Let O_{2m} be a sample consisting of elements $Y_1, Y_2, \ldots Y_{2m}$ drawn from another continuous population.

As before, it is required to test the hypothesis that these two samples came from the same population. Let m_1 equal the number of values of sample O_{2m} which are less than X_{n+1}, the median of sample O_{2n+1}. Let $m_2 = 2m - m_1$ equal the number of observations greater than X_{n+1}.

The statistic proposed by Mathisen for testing the null hypothesis is the value of m_1. In order to determine the critical values of m_1, its distribution is obtained. Let the probability that $X < X_{n+1}$ be

$$p = \int_{-\infty}^{X_{n+1}} f(X)\, dX .$$

Then

$$\text{Prob } (X_{n+1} < X) = 1 - p .$$

Since X_{n+1} is the median of O_{2n+1}, there will be n values less than X_{n+1} and n values greater than it. By the multinomial distribution the probability element for X_{n+1} will be

$$\frac{(2n+1)!}{n!\ 1!\ n!}\ p^n(1-p)^n\,dp\ .$$

Also using the multinomial distribution, the conditional probability of m_1 for a given X_{n+1} will be

$$\frac{(2m)!}{m_1!\ (2m-m_1)!}\ p^{m_1}\ (1-p)^{2m-m_1}\ .$$

Then the probability of obtaining particular values for m_1 and X_{n+1} is

$$\frac{(2n+1)!\ (2m)!}{n!\ n!\ m_1!\ (2m-m_1)!}\ p^{n+m_1}\ (1-p)^{n+2m-m_1}\ dp\ .$$

To obtain the probability of a given value for m_1, integrate the above expression in the interval $0 \leq p \leq 1$. Then the distribution of m_1 is

$$\frac{(2n+1)!\ (2m)!\ (n+m_1)!\ (n+2m-m_1)!}{n!\ n!\ m_1!\ (2m-m_1)!\ (2n+2m+1)!}\ .$$

The test criterion is the value of m_1. Either large or small values of m_1 are judged significant. Critical values of the statistic can be computed from the distribution function for any desired significance level, α. A small table of the .01, .05 critical values for a few pairs of values of m, n has been included in the description of the test [4].

Mathisen has proposed an extension of the method just described. Instead of dividing the one sample into two parts it is suggested to make four divisions. This is done by considering the quartile points of the sample O_{2n+1} . For convenience let the second sample be O_{4m} instead of O_{2m} . Let the number of values of O_{4m} falling in each of the four intervals of the quartiles of O_{2n+1} be m_1, m_2, m_3, m_4 respectively.

Then

$$\sum_{i=1}^{4} m_i = 4m$$

The statistic proposed for this test is

$$T_4 = \frac{\sum_{i=1}^{4} (m_i - m)^2}{9m^2}$$

where $9m^2$ is a normalizing factor to ensure that $0 \leq T_4 \leq 1$.

It should be noted that there is an error in the expression T_4 since the maximum value of the numerator is $12m^2$.

Again, unusually large or small values of m_i will indicate a poor comparison of the two samples. Thus such values are judged significant. The distribution function of the statistic T_4 is determined in much the same manner as was employed in the first method. Critical values of T_4 can be computed for various values of α . The computation of the critical values of the statistic for both the median and the quartile method become rather laborious for large m and n . However, in both cases the distribution functions of the statistics can be approximated by other well known distributions for which tables are available. For the median method, the distribution of m_1 has been found to be asymptotically normal. Let $E\left[m_1\right]$ denote the mean of m_1

and $D^2[m_1]$ the variance of m_1. As $m, n \to \infty$ such that $m/n = \lambda$, a constant, the limiting form of the moment generating function for the ratio

$$\frac{m_1 - E[m_1]}{D[m_1]}$$

is shown to be identical with the moment generating function of the standard normal distribution with zero mean and unit variance. Also the distribution of the statistic T_4 used in the quartile method can be approximated by the distribution defined by a Pearson type I curve. It is conjectured that since T_4 is the sum of squares its distribution could be approximated by the chi-squared distribution.

Another non-parametric test, proposed by W.J. Dixon [3] can be shown to be an extension of the method of using the median or quartile points as in the Mathisen test. Let O_m, O_n be the two samples. Consider the $n + 1$ intervals on the real line created by the n ordered observations of O_n,

$$-\infty < X_1 < X_2 < \dots < X_n < \infty .$$

Let the number of values of O_m in these intervals be m_i where where $i = 1, 2, \dots, n + 1$. The test criterion suggested by Dixon is

$$D^2 = \sum_{i=1}^{n+1} \left(\frac{1}{n+1} - \frac{m_i}{m}\right)^2 .$$

Extending the quartile method of Mathisen so that the n quantile points of O_n are considered, the statistic would be

$$T_{n+1} = \frac{\sum_{i=1}^{n+1} \left(m_i - \frac{m}{n+1}\right)^2}{\frac{n}{n+1} \cdot m^2} .$$

Essentially the two statistics are the same since

$$T_{n+1} = \frac{n+1}{n} \cdot D^2 .$$

The distribution of $n D^2$ has been shown by Dixon to be approximately the chi-squared distribution with ν degrees of freedom where

$$\nu = \frac{mn(n+m+1)(n+3)(n+4)}{2(m-1)(m+n+2)(n+1)^2}$$

Thus $(n^2/(n+1))T_{n+1}$ will have the same distribution.

Under certain conditions the Dixon test criterion, D^2, and the run test statistic, U, have been shown to give the same information. In his paper [3], Dixon shows that the correlation between the two criteria approaches one for large n compared to m. In this case the Dixon test can be considered as a test of type one since the two population distributions will be compared at an infinite number of points along the real line. Such should also be true of the extension of the Mathisen test using T_{n+1}.

A.H. Bowker [1] has shown that the median test suggested by Mathisen is not consistent for all alternative hypotheses regarding the two population distribution functions $F(X)$, $G(X)$. This implies that the probability of the false null hypothesis being rejected, when the size of the samples increases indefinitely, does not approach one. In particular, if the null hypothesis is tested against the alternative hypothesis that $F(X)$ and $G(X)$ are different except in the region of their medians, the test will not consistently reject the null hypothesis. As before let O_{2n+1}

and O_{2m} be the two samples. The proof is based on the fact that the sequences $m_\alpha/2n$ and $m_\varepsilon/2n$ each converge to one-half where m_α, m_ε are the upper and lower critical values of m_1 such that under the null hypothesis,

$$\text{Prob } (m_\alpha < m_1) = \alpha \quad \text{and Prob } (m_1 < m_\varepsilon) = \varepsilon < 1 - \alpha \, .$$

Then, even though the alternative hypothesis is true, the probability of rejecting the null hypothesis approaches $\alpha + \varepsilon$ as n increases indefinitely.

The following example illustrates the use of the Mathisen and Dixon tests. Given the two samples (.651, .602, .584, .601, .639, .572, .604, .625, .573, .586) and (.575, .605, .550, .579, .563, .552, .591, .576, .567, .588), test the hypothesis that these are random samples drawn from the same population. Since $n = m = 10$, the median of either sample must be estimated by averaging the two middle numbers. The median of the first sample is .6015. The observed value of $m_1 = 9$. Using tables [4] we find this value of m_1 is significant at the $\alpha = .05$ level. However, using the median of the second sample we obtain a different result. The estimated median is .5755. The observed value of $m_1 = 2$ which is not significant at the $\alpha = .05$ level.

Using the Dixon test the first sample divides the second sample into the following groups: 4, 0, 3, 0, 2, 0, 0, 1, 0, 0, 0. Then

$$D^2 = \left(\frac{1}{11} - \frac{4}{10}\right)^2 + \left(\frac{1}{11} - \frac{3}{10}\right)^2 + \left(\frac{1}{11} - \frac{2}{10}\right)^2 + \left(\frac{1}{11} - \frac{1}{10}\right)^2 + 7\left(\frac{1}{11}\right)^2 = .209 \, .$$

Using the table [3] we find this result is not significant at the $\alpha = .05$ level.

The Pitman Randomization Test

A test based on the method of randomization has been proposed by E.J.G. Pitman [5]. As before, let O_m, O_n be two samples with elements $X_1, X_2, \ldots X_m$ and $Y_1, Y_2, \ldots Y_n$ respectively. Combine and order the data of the two samples so that O_{m+n} consists of the values $Z_1, Z_2, \ldots Z_{m+n}$ where $Z_i < Z_{i+1}$. Again it is required to test the null hypothesis $F(X) = G(X)$.

Define a separation of O_{m+n} to be a division of the $m + n$ observations into two parts, one containing m values and the other, n values. The total number of possible separations will be $C(m+n, m)$. One such separation will be that determined by the two samples O_m, O_n. Call this particular separation R. The spread of this separation R is defined as $|\bar{X} - \bar{Y}|$ where $\bar{X}$ and $\bar{Y}$ are the mean values of O_m, O_n respectively.

Let M = the number of separations of O_{m+n} with a spread equal to or greater than that of R. Let M_α be a fixed integer such that $M_\alpha < C(m+n, m)$. The value of M_α depends on the amount of probability, α desired in the rejection region under the null hypothesis. If $M \leq M_\alpha$, then the spread of R is judged significant and the null hypothesis is rejected. Thus the test criterion is the number of separations of O_{m+n} with spread greater or equal to that of R. If this number M is comparatively small then $|\bar{X} - \bar{Y}|$ is considered too great for the null hypothesis to be true.

For values of m, n as large as 10 there would be considerable computation to determine all the separations with a spread greater or equal to $|\bar{X} - \bar{Y}|$. For this reason a statistic is suggested by Pitman which is related to the previous one with the added property that its distribution function can be approximated by the beta distribution.

Define

$$W = \frac{\frac{mn}{m+n}(\bar{X}-\bar{Y})^2}{S_1+S_2+\frac{mn}{m+n}(\bar{X}-\bar{Y})^2}$$

where

$$S_1 = \sum_{i=1}^{m} (X_i - \bar{X})^2 \quad \text{and} \quad S_2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

The first three moments of the distribution of W are shown to be approximately equal to those of the beta distribution,

$$\beta(\frac{1}{2} \quad \frac{m+n}{2} - 1) \; .$$

Since large values of W will be judged significant, the region of rejection for this test is $W > W_\alpha$ where W_α is the critical value of W for a particular value of α . W_α is is determined by

$$\alpha = \frac{1}{\beta(\frac{1}{2}, \frac{m+n}{2} -1)} \int_{W_\alpha}^{1} X^{\frac{1}{2}-1} (1 - X)^{\frac{m+n}{2}-2} dx$$

As an illustration of the Pitman test apply the statistic $|\bar{X} - \bar{Y}|$ to test the hypothesis that (0, 11, 12, 20) and (16, 19, 22, 24) are two random samples from the same population. There are C(8, 4) = 70 possible separations. $M_\alpha = M_{.057} = 4$.

Since there are $M = 6$ separations with a spread equal to or greater than $|\bar{X} - \bar{Y}|$ the result is not significant and we conclude there is no evidence against the null hypothesis on the basis of these samples.

A New Test: "The Integer Test"

The following new test which will be called the Integer test is based on the principle of randomization, and thus is related to the Pitman test.

As before, suppose O_m and O_n are two samples drawn from populations with continous distribution functions, $F(X)$ and $G(X)$. The null hypothesis is $F(X) = G(X)$. Let O_{m+n} be the ordered combination of the two samples

$$O_{m+n} : Z_1, Z_2, \ldots Z_{m+n} \text{ where } Z_i < Z_{i+1} .$$

Replace the sample values Z_i of O_{m+n} by their corresponding subscript, i, where $i = 1, 2, \ldots m+n$, so that to each element of the two samples O_m, O_n there is assigned a positive integer which indicates the rank or order of the element in the combined sample O_{m+n}. If $Z_i = Z_{i+1} = Z_{i+2} = \ldots = Z_{i+r}$, replace each of these equal sample values by the number, $i + r/2$.

Now consider as a population the integers 1, 2, 3, ..., $m + n = N$. Suppose samples of n integers are drawn from this population so that none of the integers are selected more than once for each sample. These samples will be random in the sense that each has equal probability. In practice, the observations of a sample are actually drawn without replacement from a population but since the size of the population is often very much greater than the size of the sample it can be assumed that the sample data are independent. However, in the Integer test the sample data must be considered as dependent since n and N are of the same order. That is, the sampling is done without replacement from a finite population. Now consider all possible divisions of the N integers into two sets of n and m values respectively. The number of such combinations is $C(N, n)$. One of these divisions will represent the samples O_m, O_n.

The test criteria will be the two means of the sets of m and n integers for the particular division determined by O_m and O_n, Since the two means are dependent a study of one of them will be sufficient. For convenience, let the larger of the two, $\bar{U}$ be the statistic proposed in this test. If $\bar{v}$ denotes the other mean, note that

$$n\bar{U} + (N - n)\bar{v} = \frac{N(N+1)}{2}$$

where $\frac{N(N+1)}{2}$ is the sum of the integers $1, 2, 3, \ldots N$.

Values of $\bar{U}$ greater than $(N + 1)/2$ are judged significant, and for a given level of significance α the region of rejection consists of those values of $\bar{U}$ such that $\bar{U}_\alpha < \bar{U}$, where $\bar{U}_\alpha$ is the critical value of $\bar{U}$ for a given probability α. As is suggested in the Pitman test all the means of the $C(N, n)$ combinations greater than $\bar{U}_\alpha$ can be computed. Then $\bar{U}_\alpha$ is a particular value of the mean such that a proportion, α of the means is greater than $\bar{U}_\alpha$.

Unfortunately, while the computation is simpler for this test than for the Pitman test, this method of determining the critical values for N greater than ten is not practical. It is advisable therefore to obtain the distribution function of $\bar{U}$ and thus determine the critical values $\bar{U}_\alpha$. For independent variables the means of samples are normally distributed, exactly if the population is normal and approximately if the samples are large. However, since $\bar{U}$ is the mean of a sample of dependent integers, the well known central limit theorem can not be applied in this case. Fortunately, A. Wald and J. Wolfowitz [8] have proved a general theorem for the limiting distribution of linear forms where the population consists of all divisions of $m + n$ observations. Now the distribution of $\bar{U}$ will be the same as the distribution of the linear form,

$$\sum_{i=1}^{n} U_i \ .$$

The Wald-Wolfowitz theorem states that as $N \to \infty$, the

$$\text{Prob.} \left(\sum_{i=1}^{n} U_i - E\left[\sum U_i\right] < t \cdot D\left[\sum U_i\right] \right)$$

is approximately

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-x^2/2\right) dx$$

where t is a real number and

$$E\left[\sum U_i\right] \quad \text{and} \quad D^2\left[\sum U_i\right]$$

are the mean and variance of $\sum U_i$ respectively. Before this theorem may be applied a certain condition must be satisfied. Let μ_r be the rth moment about the mean of the integers $1, 2, 3, \ldots N$; the condition is that

$$\frac{\mu_r}{(\mu_2)^{r/2}}$$

must be of the order of one. Since μ_r is of the order of N^r

and μ_2 is of the order of N^2 for a population of N integers, the theorem holds for this case. Thus the limiting distribution of the statistic $\bar{U}$ is normal.

The expected value of $\bar{U}$, $E[\bar{U}]$ equals

$$\frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2}$$

where $(N + 1)/2$ is the population mean of N integers.
The variance of $\bar{U}$, $D^2[\bar{U}]$ is

$$\frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^{\,2} + \frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho_{ij}\sigma_i\sigma_j$$

where $\sigma_i = \sigma_j = \sigma$ and ρ_{ij} denotes the correlation between two integers drawn in succession. Now ρ_{ij} equals

$$\text{(A)} \qquad \frac{1}{\sigma^2} E\left[(U_i - \frac{N+1}{2})(U_j - \frac{N+1}{2})\right] .$$

By definition, (A) is equal to

$$\text{(B)} \qquad \frac{1}{\sigma^2}\,\frac{1}{C(N,2)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(U_i - \frac{N+1}{2})(U_j - \frac{N+1}{2}) .$$

Since

$$0 = \left[\sum_{i=1}^{N}(U_i - \frac{N+1}{2})\right]^2$$

$$= \sum_{i=1}^{N}(U_i - \frac{N+1}{2})^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(U_i - \frac{N+1}{2})(U_j - \frac{N+1}{2})$$

the expression (B) equals

$$-\frac{1}{\sigma^2}\,\frac{\sigma^2}{N-1} .$$

Then

$$D^2[\bar{U}] = \frac{1}{n^2}\left[n\sigma^2 - 2C(n,2)\frac{\sigma^2}{N-1}\right] = \frac{\sigma^2}{n}\,\frac{N-n}{N-1} .$$

Thus the statistic $\bar{U}$ is asymptotically normally distributed with mean $(N + 1)/2$ and variance

$$\frac{\sigma^2}{n} \frac{N-n}{N-1}$$

where σ^2, the population variance equals $(N^2 - 1)/12$.

In order to use the tables of the standard normal distribution the test criterion will be

$$t = \frac{\bar{U} - \frac{N+1}{2}}{\sqrt{\frac{\sigma^2}{n} \frac{N-n}{N-1}}} .$$

The region of rejection becomes $t_\alpha < t$ where t_α is the critical value of t corresponding to the probability α of rejecting the null hypothesis when it is actually true.

If two samples are symmetric about the same mean the statistic $\bar{U}$ will be equal to $(N + 1)/2$ since the integral representatives of the values of the samples will also be symmetric. Now suppose the alternative hypothesis is that the population distributions $F(X)$ and $G(X)$ have the same means but different variances. It would be possible that the Integer test would not detect the falsehood of the null hypothesis as some pairs of samples would have means which differed by very little. For this reason when the value of the observed t is close to zero it is suggested that the sample variances of the two sets of integers be compared with the population variance of N integers. Since

$$\sum_{i=1}^{n} U_i^2 + \sum_{i=1}^{N-n} v_i^2 = \sum_{i=1}^{N} i^2 ,$$

the two sample variances are dependent and thus only one of them, say the larger, need be considered as the test criterion.

As before, the distribution of this statistic must be determined to obtain its critical values. It will be shown that the distribution of this sample variance S^2 can be approximated by the chi-squared distribution.

To determine the particular chi-square distribution the first two moments of S^2 are obtained [2]. By definition the expected value of S^2 is

$$E\left[\frac{1}{n}\sum_{i=1}^{n} (U_i - \overline{U})^2\right] .$$

Previously it was shown that

$$E\left[\overline{U} - \frac{N+1}{2}\right]^2 = \frac{\sigma^2}{n} - \frac{n-1}{n}\,\frac{\sigma^2}{N-1}$$

where σ^2 is the variance of the integers 1, 2, 3, ... N .
By definition

$$E\left[\sum_{i=1}^{n} (U_i - \frac{N+1}{2})^2\right] = n\sigma^2 .$$

Then using the identity

$$\frac{1}{n}\sum_{i=1}^{n} U_i^2 - \overline{U}^2 = \frac{1}{n}\sum_{i=1}^{n} (U_i - \frac{N+1}{2})^2 - (\overline{U} - \frac{N+1}{2})^2$$

we obtain

$$E\left[S^2\right] = \frac{n\sigma^2}{n} - \frac{\sigma^2}{n} + \frac{n-1}{n}\cdot\frac{\sigma^2}{N-1} = \frac{n-1}{n}\cdot\frac{N}{N-1}\cdot\sigma^2 .$$

It can be shown that the variance of S^2 equals

$$\frac{N(N-n)(n-1)\,\sigma^4}{(N-1)^2(N-2)(N-3)\,n^3}\left[2nN^2 - 6(n+1)(N-1) + (n\,N-N-n-1)(N-1)\lambda_2\right]$$

where λ_2 is the coefficient of excess defined as

$$\frac{\mu_4}{\sigma^4} - 3 .$$

Let $$q^2 = \frac{Nn\ S^2}{N-n} .$$

Then $$E\left[\frac{q^2}{\sigma^2}\right] = \frac{N(n-1)\ N}{(N-n)(N-1)} = \frac{N(n-1)}{N-n}\left[1 + O(\tfrac{1}{N})\right]$$

and $$D^2\left[\frac{q^2}{\sigma^2}\right]$$
$$= \frac{2N\ (n-1)\ N^4}{(N-n)\ (N-1)^2(N-2)(N-3)} + \frac{2N(n-1)\ (n-1)\ \lambda 2\ N^2}{(N-n)\ 2n\ (N-2)(N-3)}$$
$$- \frac{2N(n-1)\ N^2\ 3(n+1)}{N-n\ (N-2)(N-3)\ n}$$

where $(n\ N-N-n-1)$ is approximately equal to $(n-1)(N-1)$.

Then
$$D^2\left[\frac{q^2}{\sigma^2}\right] = \frac{2N(n-1)}{N-n}\left[1 + \frac{n-1}{2n}\lambda_2 + O(\tfrac{1}{N})\right] .$$

Thus for large N and λ_2 equal to zero, the mean of q^2/σ^2 is $N(n-1)/(N-n)$ and the variance is $2N(n-1)/(N-n)$. Hence the distribution of q^2/σ^2 can be approximated by a chi-square distribution with $N(n-1)/(N-n)$ degrees of freedom.

The statistic proposed for a comparison of the variances in the Integer test is

$$\chi^2 = \frac{Nn\ S^2}{(N-n)\sigma^2} .$$

The region of rejection will be the values of χ^2 such that $\chi^2_\alpha < \chi^2$.

As an illustration of the Integer test consider the two samples used in the application of the Wald-Wolfowitz Run Test. On ordering the values of the two samples and assigning the appropriate integers the samples become (1, 2, 3, 10, 11, 12) and

and (4, 5, 6, 7, 8, 9) . Then $\bar{U} = 6.5$, $\bar{v} = 6.5$ and $(N+1)/2 = 6.5$

$$\sigma^2 = (N^2-1)/12 = 11.9$$

$$\sigma_{\bar{U}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1} = 1.08 \ .$$

Then

$$t = \frac{\bar{U} - \frac{N+1}{2}}{\sigma_{\bar{U}}} = \frac{6.5 - 6.5}{1.04} = 0 \ .$$

This value of t is certainly not significant. Now if we are testing against the alternative hypothesis that $G(X)$ is a translation of $F(X)$ the statistic t would be valid. However, if the alternative hypothesis is such that the two populations differ in other respects besides their means, the statistic q^2/σ^2 should also be used. For the above example,

$$S^2 = \frac{7N^2-4}{48} = 20.9 \ .$$

Note that this formula for S^2 hold only if $n = N/2$ and N is divisible by 4 . Then

$$\chi^2 = \frac{q^2}{\sigma^2} = \frac{Nn\ S^2}{(N-n)\sigma^2} = \frac{(12)(20.9)}{11.9} = 21.1 \ .$$

The number of degrees of freedom,

$$\nu = \frac{N(n-1)}{N-n} = \frac{(12)(5)}{6} = 10 \ .$$

From tables for the chi-squared distribution

$$\text{Prob. } (\chi^2 > 21.161) = .02 \ .$$

Thus the observed value of $\chi^2 = 21.1$ is significant at the $\alpha = .05$ level and the null hypothesis is rejected.

We note that since the Integer test consists of two parts the total probability in the rejection region will be $\alpha + (1-\alpha)\varepsilon$ where Prob $(t_\alpha < t) = \alpha$ and Prob $(\chi^2_\varepsilon < \chi^2) = \varepsilon$.

A good test should have a high probability of rejecting the null hypothesis when it is actually false. As stated previously this probability, called the power of a test, cannot be determined for distribution free tests. An alternative criterion for the non-parametric case is that a good test is consistent with respect to all couples of continuous $F(X)$, $G(X)$.

It is conjectured that the Integer test is consistent with respect to the alternative hypothesis that $G(X) = F(X + d)$, a translation of $F(X)$ where d is a constant. To prove this it should be shown that the statistic $\bar{U}/N$ converges stochastically to its expected values when either hypothesis is true. It can be shown that if the null hypothesis is true, $\bar{U}/N$ converges stochastically to $(N + 1)/2N$. Let ε be an arbitrarily small positive number. Using Tchebycheff's inequality,

$$\text{Prob.} \left\{ \left| E\left[\frac{\bar{U}}{N}\right] - \frac{N+1}{2N} \right| < \varepsilon \right\} > 1 - \frac{\sigma^2_{\bar{U}/N}}{\varepsilon^2} .$$

Thus for N sufficiently large, $\sigma^2_{\bar{U}/N}$ approaches zero and hence

$$E\left[\frac{\bar{U}}{N}\right] - \frac{N+1}{2N}$$

converges in probability to zero.

A difficulty arises in connection with any attempt to show that $\bar{U}/N$ converges to its expected value when the alternative hypothesis is true, since the distribution of the statistic is not known. Thus the expressions for the expected value and variance of $\bar{U}/N$ cannot be stated explicitly although it is surmised that the expected value depends on the constant d and is greater than $(N + 1)/2N$ and that the variance approaches zero for large N.

Similar difficulties arise in the consideration of the consistency of the Integer test with respect to other alternative hypotheses regarding $F(X)$ and $G(X)$.

Conclusion

In the example used to illustrate Mathisen's and Dixon's tests, conflicting results were obtained. Mathisen's median test rejected the null hypothesis, whereas, Dixon's test indicated there was no evidence against it.

Applying the Wald-Wolfowitz Run test to the same example, the observed value of U is 8. From the tables in [6] the probability $(U = 8) = .1276$ for $n = m = 10$. There is no evidence against the null hypothesis on the basis of these two samples.

Now apply the Integer test to this example. The division of integers is (5, 6, 10, 11, 14, 15, 16, 18, 19, 20) and (1, 2, 3, 4, 7, 8, 9, 12, 13, 17). $\bar{U} = 13.4$, $(N+1)/2 = 10.5$ and $\sigma_{\bar{U}}^2 = 1.755$. The observed t is 2.20. From tables for the normal distribution the probability $(2.20 < t) = .0139$. Thus the null hypothesis is rejected for $\alpha = .05$.

In two of the non-parametric tests the Mathisen and the Integer tests, a significant result is obtained while in the other two, the Dixon and the Run tests, the observed value of the statistic is not significant. If the Mathisen and Integer tests are at fault, it means the probabilities in the rejection region for these tests are too small and conversely for the case the other two give incorrect results.

It is interesting to note what happens if we assume that the populations from which these samples were drawn are normally distributed. In this case we can apply the test statistics based on the Student's t and F distributions. The observed value of F is 2.44 for $\nu_1 = \nu_2 = 9$ degrees of freedom. This value is not significant for $\alpha = .05$. Thus we may assume that the two normal populations have a common variance and thus can apply to the Student's t test. The observed value of t is 2.86 with ν equal to 18. This value is almost significant for $\alpha = .01$, and we therefore reject the null hypothesis

In defense of the Run and Dixon tests which give opposite results to that of Student's t it must be emphasized that we were considering just one particular example. On the other hand examples can be found in which the Run test has smaller probabilities in the rejection region than the Student's t test.

Suggested applications of these tests are as follows: If the population distributions are normal or such that they may be approximated by normal distributions, then Student's t test should be used. For other cases the choice of a test depends on the alternative hypotheses and the demands of the experimenter. If the experiment is such that a comparison of the measures of central tendency is desired the Mathisen, Pitman and Integer tests can be used. If we wish to compare the first two moments of the distributions the Integer test is applicable. For all other non-parametric cases the Run test should be used.

In evaluating non-parametric tests and comparing them with the classical tests consideration should be made of the fact that the latter are limited in their application due to the restrictive assumption that the population distributions are normal. Thus while it is apparent that non-parametric tests do not use as much of the available information as the classical tests they are good substitutes in the cases where the populations are unknown.

References

1. A. H. Bowker, "Note on consistency of a proposed test for the problem of two samples." Ann. Math. Stat. vol. 15 (1944) pp. 98 - 101.

2. H. Cramer, Mathematical Methods of Statistics. Princeton, 1946.

3. W.J. Dixon, "A criterion for testing the hypothesis that two samples are from the same population," Ann. Math. Stat. vol. 11 (1940), pp. 199 - 204.

4. H.C. Mathisen, "A method of testing the hypothesis that two samples are from the same population," Ann. Math. Stat. vol. 14 (1943), pp. 188 - 194.

5. E.J.G. Pitman, "Significance tests which may be applied to samples from any population," Journal Roy. Statist. Soc. Supplement, vol. 4 (1937), pp. 119 - 130.

6. Frieda S. Swed and C. Eisenhart, "Tables for testing randomness of grouping in a sequence of alternatives," Ann. Math. Stat. vol. 14 (1943), pp. 66 - 87.

7. A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," Ann. Math. Stat. vol. 11 (1940), pp. 147 - 162.

8. A. Wald and J. Wolfowitz, "Statistical tests based on permutations of the observations," Ann. Math. Stat. vol. 5 (1944), pp. 358 - 372.