A METHOD OF TESTING THE HYPOTHESIS THAT TWO SAMPLES ARE FROM THE SAME POPULATION

By Harold C. Mathisen

Princeton University

1. Introduction. There are many cases in testing whether two samples are from the same population in which no assumption about the distribution function of the population can be made except that it is continuous. A. Wald and J. Wolfowitz, [1], have developed a method of testing the hypothesis that two samples come from the same population based on certain kinds of runs of the elements from each sample in the combined ordered sample. W. J. Dixon, [2], has introduced a criterion for testing the same hypothesis based on the number of elements of the second sample falling between each successive pair of ordered values in the first sample.

The problem considered here is that of devising a simple method of testing the hypothesis that two samples come from the same population, based on medians and quartiles, given only that the distribution function of the population is continuous. The simplest method may be described briefly as follows. We observe the number of elements, m_1, in the second sample whose values are lower than the median of the first sample. Since the distribution of m_1 is independent of the population distribution, we are able to compute significance points from the distribution of m_1. These points may then be used for testing the hypothesis at a given significance level. This will be referred to as the case of two intervals.

This method may be easily extended to the case of any number of intervals. In this note we shall consider the extension to four intervals by using the median and the two quartiles of the first sample to establish four intervals into which the elements of the second sample may fall. Then, if the second sample is of size $4m$, it will be shown that, under the hypothesis that the two samples come from the same population, $\frac{1}{4}$ of the second sample, or m elements will be expected to fall in each interval. Let the number in the second sample which actually fall in each interval be m_1, m_2, m_3, and m_4 respectively. The test function here proposed is,

$$C = \frac{(m_1 - m)^2 + (m_2 - m)^2 + (m_3 - m)^2 + (m_4 - m)^2}{9m^2}, \tag{1}$$

where $9m^2$ is a constant, which forces C to lie on the interval 0 to 1. If the m_i, $(i = 1, 2, 3, 4)$, have values quite different from their expected value m, it is apparent that C will be large. Therefore the greater the value of C the more doubtful is the hypothesis that the two samples come from the same population. Significance values of C will be computed for several sample sizes. The question of whether C is the "best" four-interval criterion for testing the hypothesis that two samples come from the same continuous distribution is an open one

which would depend for its answer on an extensive power function analysis. We shall not go into this analysis, however, but shall use C on intuitive grounds. This case will be referred to as the case of four intervals. The extension of the method of the case of four intervals to any number of intervals presents no new difficulties in derivation, however we shall confine our attention to the cases of two and four intervals.

2. The case of two intervals. Suppose $f(x)$ is a continuous distribution function with probability element $f(x)\,dx$. Let us draw a sample of size $2n + 1$ from a population having this probability element. Let the elements in the sample be $x_1, x_2, \cdots, x_{2n+1}$ ordered from least to greatest. The median of this sample will be x_{n+1}. Now consider a second sample of size $2m$, and let m_1 be the number of observations, whose values are less than x_{n+1}. We call $m_2 = 2m - m_1$ the number of elements in the second sample greater than x_{n+1}.

Let $p = \int_{-\infty}^{x_{n+1}} f(x)\,dx$ be the probability of an observation having a value less than x_{n+1}. Then the probability of an element having a value greater than x_{n+1} is $(1 - p)$. Thus we have the relation $f(x_{n+1})\,dx_{n+1} = dp$. The probability law of the median, x_{n+1} given by the multinomial law[1] is

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

The conditional probability law of m_1, given x_{n+1}, is then

$$P_r(m_1 \mid x_{n+1}) = \frac{(2m)!}{m_1!(2m-m_1)!}\,p^{m_1}(1-p)^{2m-m_1}. \tag{3}$$

From this it follows that the joint probability law of x_{n+1} and m_1 is the product of (2) and (3) or

$$P_r(m_1, x_{n+1}) = \frac{(2n+1)!(2m)!}{n!n!m_1!(2m-m_1)!}\,p^{n+m_1}(1-p)^{n+2m-m_1}\,dp. \tag{4}$$

We may integrate (4) with respect to p from 0 to 1 as a Beta Function, leaving the distribution function of m_1 independent of the population probability element $f(x)\,dx$. We get for the distribution of m_1,

---

[1] The multinomial law may be stated briefly as follows:

If a trial results in one and only one of the mutually exclusive events $E_1, E_2, \cdots, E_k$, the probability P that in a total of n trials, n_1 will result in E_1, n_2 in E_2, $\cdots$, n_k in E_k, $\left(\sum_1^k n_i = n\right)$, is given by

$$P = \frac{n!}{n_1!\,n_2!\cdots n_k!}\,p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}$$

where $p_1, p_2, \cdots, p_k$, $\left(\sum_1^k p_i = 1\right)$, are the probabilities of a single trial resulting in $E_1, E_2, \cdots, E_k$ respectively.

$$(5) \qquad P_r(m_1) = \frac{(2n+1)!(2m)!(n+m_1)!(n+2m-m_1)!}{n!n!m_1!(2m-m_1)!(2n+1+2m)!}.$$

From (5) a simple recursion relation between $P_r(m_1)$ and $P_r(m_1+1)$ may be determined from which the probabilities of various values of m may be rapidly computed. For large samples it can be shown that under certain regularity conditions, the ratio, $[m_1 - E(m_1)]/\sigma_{m_1}$ may be approximated by the normal distribution[2] with zero mean and unit variance. The derivation is similar to that of the four-interval case, which is taken up in greater detail. It will be found by the use of (4) that the expected value of m_1 is m, and the variance of m_1 is $m + \frac{m(2m-1)(n+2)}{2n+3} - m^2$. Using this information, values of m_1 for various

TABLE I

The Case of Two Intervals

Lower and upper .01 and .05 percentage points for the distribution of m_1

Sample sizes		Critical values of m_1			
		Lower		Upper	
First $2n+1$	Second $2m$	$m_{1(.01)}$	$m_{1(.05)}$	$m_{1(.05)}$	$m_{1(.01)}$
11	10		1	9	
41	40	10	12	28	30
101	100	34	38	62	66
101	200	72	80	120	128
201	200	77	84	116	123
201	400	160	181	219	240
401	400	167	177	223	233
401	800	353	367	433	447
1001	1000	448	463	537	552

significance levels may be computed. The .01 and .05 percentage points of m_1 for several sample sizes are given in Table I. The values for sample sizes of 10 and 40 are computed directly from the probability law, while the larger samples have limits computed by the normal approximation. Thus for two samples of size 101 and 100, respectively, a value of m_1 less than 38 would be significant at the .05 level. Similarly, at the upper .05 level, the hypothesis would be rejected if a value of m_1 were obtained which was greater than 62. The necessity for the upper limits could easily be eliminated by testing with respect to the smaller of m_1 and m_2. However, for completeness, the upper percentage points

[2] This statement may be proved by showing that as $m, n \to \infty$ such that $m/n = constant$, the limit of the moment generating function for the ratio is identical with the moment generating function of the normal distribution with zero mean and unit variance.

are included to show the range of values of m_1 in which the hypothesis that the two samples come from the same population may be accepted.

3. The case of four intervals. If we let the first sample of size $4n + 3$ be designated by $(x_1, x_2, \cdots x_{4n+3})$, assumed drawn from a population with probability element $f(x)\,dx$ and ordered from least to greatest, then the range of x may be divided into four intervals by x_{n+1}, x_{2n+2}, and x_{3n+3}. The probability element of x_{n+1}, x_{2n+2}, x_{3n+3} is

$$\frac{(4n+3)!}{n!\,n!\,n!\,n!}\left(\int_{-\infty}^{x_{n+1}} f(x)\,dx\right)^n\left(\int_{x_{n+1}}^{x_{2n+2}} f(x)\,dx\right)^n\left(\int_{x_{2n+2}}^{x_{3n+3}} f(x)\,dx\right)^n\left(\int_{x_{3n+3}}^{\infty} f(x)\,dx\right)^n$$
$$\cdot f(x_{n+1})\,dx_{n+1} f(x_{2n+2})\,dx_{2n+2} f(x_{3n+3})\,dx_{3n+3}.$$

TABLE II

The Case of Four Intervals

.95 and .99 percentage points for the distribution of C

Sample sizes				$C_{.95}$	$C_{.99}$
First	Second				
$4n+3$	$4m$	n	m		
15	12	3	3	.446	.582
63	60	15	15	.113	.161
103	100	25	25	.072	.102

Let

$$\int_{-\infty}^{x_{n+1}} f(x)\,dx = p_1,\quad \int_{x_{n+1}}^{x_{2n+2}} f(x)\,dx = p_2,\quad \int_{x_{2n+2}}^{x_{3n+3}} f(x)\,dx = p_3,\quad \int_{x_{3n+3}}^{\infty} f(x)\,dx = p_4.$$

The probability element of p_1, p_2, p_3, and p_4 is

$$(6)\qquad p_r(x_{i(n+1)}) = \frac{(4n+3)!}{n!1!n!1!n!1!n!}\,p_1^n p_2^n p_3^n p_4^n\,dp_1\,dp_2\,dp_3.$$

Now let us consider the second sample, $(x_1', x_2', \cdots x_{4m}')$, of size $4m$. Let the number of observations falling in each of the preassigned intervals be m_i, $(i = 1, 2, 3, 4)$, where $m_4 = 4m - m_1 - m_2 - m_3$. The conditional probability of the m_i, given the values of $x_{i(n+1)}$ is also determined by the multinomial law.

$$(7)\qquad P_r(m_i \mid x_{i(n+1)}) = \frac{(4m)!}{m_1!\,m_2!\,m_3!\,m_4!}\,p_1^{m_1}p_2^{m_2}p_3^{m_3}p_4^{m_4}.$$

The joint distribution of the p_i and the m_i is then

$$(8)\qquad P_r(x_{i(n+1)}, m_i) = \frac{(4n+3)!\,(4m)!}{(n!)^4 m_1!\,m_2!\,m_3!\,m_4!}\,p_1^{n+m_1}p_2^{n+m_2}p_3^{n+m_3}p_4^{n+m_4}\,dp_1\,dp_2\,dp_3.$$

To obtain the distribution of the m_i alone, the p_i will be integrated out by the Dirichlet Integral[3] formula, giving a distribution which is clearly independent of the population distribution function $f(x)$.

$$(9) \quad P_r(m_i) = \frac{(4n+3)!\,(4m)!\,(n+m_1)!\,(n+m_2)!\,(n+m_3)!\,(n+m_4)!}{(n!)^4 m_1!\,m_2!\,m_3!\,m_4!\,(4m+4n+3)!}.$$

To find the expected value of the m_i, the probability law of m_1 will first be derived. The probability function for the value of x_{n+1} is

$$(10) \quad P_r(x_{n+1}) = \frac{(4n+3)!}{1!\,n!\,(3n+2)!}\, p_1^n (1-p_1)^{3n+2}\, dp_1 .$$

Then we have the conditional probability

$$(11) \quad P_r(m_1 \mid x_{n+1}) = \frac{(4m)!}{m_1!\,(4m-m_1)!}\, p_1^{m_1}(1-p_1)^{4m-m_1},$$

and

$$(12) \quad P_r(x_{n+1}, m_1) = \frac{(4n+3)!\,(4m)!}{n!\,(3n+2)!\,m_1!\,(4m-m_1)!}\, p_1^{n+m_1}(1-p_1)^{3n+2+4m-m_1}\, dp_1 .$$

To obtain the expected value of m_1, the joint distribution of m_1 and p_1 is multiplied by m_1, summed on m_1 from 0 to $4m$, and integrated on p_1 from 0 to 1.

$$(13) \quad E(m_1) = \frac{(4n+3)!}{n!\,(3n+2)!} \int_0^1 p_1^n (1-p_1)^{3n+2} \cdot \left[\sum_0^{4m} m_1 \frac{(4m)!}{m_1!\,(4m-m_1)!}\, p_1^{m_1}(1-p_1)^{4m-m_1} \right] dp_1 .$$

This interchange of the order of integration and summation is clearly valid. The quantity in brackets will be recognized as the first moment of the binomial distribution, $(p_1+q)^{4m}$ where $q = 1 - p_1$. Therefore we have

$$(14) \quad E(m_1) = \int_0^1 4m p_1 f(p_1)\, dp_1 = 4m E(p_1).$$

$E(p_1)$ and the higher moments of p_1 are found in the usual way by integrating the distributions as Beta Functions. From this we see that the expected value of m_1 is m. By repeating these operations on m_2, m_3, and m_4, it can be seen that $E(m_i) = m$, which also validates the statement made in the introduction.

---

[3] A discussion of the Dirichlet Integral may be found in Woods—*Advanced Calculus*, p. 167. It may be stated as follows for the problem in which we are interested

$$\iiint x^{l-1} y^{m-1} z^{n-1} (1-x-y-z)^{r-1}\, dx\, dy\, dz = \frac{\Gamma(l)\Gamma(m)\Gamma(n)\Gamma(r)}{\Gamma(l+m+n+r)},$$

where we integrate over the region bounded by $x + y + z = 1$, and the three coordinate planes.

We have previously presented the criterion (1).

The next problem is to find a distribution function to which the distribution of C may be fitted. A reasonable choice appears to be the Pearson Type I curve.

$$f(x) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1-x)^{s-1}. \tag{15}$$

The distribution of C is fitted by equating the first two moments of the two distributions and solving for the constants r and s of the Type I distribution. Using the theorem that the mean value of the sum of variates is equal to the sum of their mean values, we have

$$E(C) = \frac{1}{9m^2}[E(m_1^2) + E(m_2^2) + E(m_3^2) + E(m_4^2) - 4m^2]. \tag{16}$$

Also the second moment may be written as

$$\begin{aligned} E(C^2) = \frac{1}{81m^4}[&E(m_1^4) + E(m_2^4) + E(m_3^4) + E(m_4^4) + 16m^4 + 2E(m_1^2 m_2^2) \\ &+ 2E(m_1^2 m_3^2) + 2E(m_1^2 m_4^2) + 2E(m_2^2 m_3^2) + 2E(m_2^2 m_4^2) \\ &+ 2E(m_3^2 m_4^2) - 8m^2\{E(m_1^2) + E(m_2^2) + E(m_3^2) + E(m_4^2)\}]. \end{aligned} \tag{17}$$

The expected value of m_1^2 is found in the same manner as $E(m_1)$ and here also it can be shown that the $E(m_i^2)$ are all equal. The same procedure holds for $E(m_i^4)$.

$$\begin{aligned} E(m_i^2) &= m + \frac{m(4m-1)(n+2)}{4n+5}, \\ E(m_i^4) &= m + \frac{7m(4m-1)(n+2)}{4n+5} + \frac{6m(4m-1)(4m-2)(n+3)(n+2)}{(4n+6)(4n+5)} \\ &\quad + \frac{m(4m-1)(4m-2)(4m-3)(n+4)(n+3)(n+2)}{(4n+7)(4n+6)(4n+5)}. \end{aligned} \tag{18}$$

By using the moment generating function of the trinomial distribution, the $E(m_i^2 m_j^2)$ may also be found in a similar manner.

$$\begin{aligned} E(m_i^2 m_j^2) = &\frac{m(4m-1)(n+1)}{4n+5} + \frac{2m(4m-1)(4m-2)(n+1)(n+2)}{(4n+6)(4n+5)} \\ &+ \frac{m(4m-1)(4m-2)(4m-3)(n+2)(n+1)(n+2)}{(4n+7)(4n+6)(4n+5)}. \end{aligned} \tag{19}$$

As a result we have

$$E(C) = \frac{4}{9m} + \frac{4(4m-1)(n+2)}{9m(4n+5)}. \tag{20}$$

Let $E(C) = A$ to simplify later relations to be computed. Finally

$$E(C^2) = \frac{4}{81m^3}\Bigg[1 + \frac{7(4m-1)(n+2)}{4n+5} + \frac{6(4m-1)(4m-2)(n+3)(n+2)}{(4n+6)(4n+5)} + \frac{(4m-1)(4m-2)(4m-3)(n+4)(n+3)(n+2)}{(4n+7)(4n+6)(4n+5)} + 4m^3 + \frac{3(4m-1)(n+1)}{4n+5} + \frac{6(4m-1)(4m-2)(n+1)(n+2)}{(4n+6)(4n+5)} + \frac{3(4m-1)(4m-2)(4m-3)(n+2)^2(n+1)}{(4n+7)(4n+6)(4n+5)} - 8m^2 - \frac{8m^2(4m-1)(n+2)}{4n+5}\Bigg]. \tag{21}$$

To simplify later relations we let $E(C^2) = B$.

The first two moments of the Type I distribution are easily found to be

$$\mu_1 = \frac{r}{r+s} = A \qquad \mu_2 = \frac{\mu_1(r+1)}{(r+s+1)} = B. \tag{22}$$

Solving these two simultaneous equations for r and s,

$$r = \frac{B-A}{A - \frac{B}{A}} \qquad s = \frac{r}{A} - r. \tag{23}$$

A number of percentage points for the Type I distribution have been computed by Miss Catherine Thompson, [3]. Using these limits, the hypothesis may be accepted or rejected as to whether or not the two samples come from the same population.

Table II shows the .95 and .99 percentage points of C for three sample sizes.

4. Summary. The problem considered here is that of devising a simple method of testing the hypothesis that two samples are from identical populations having continuous distribution functions. It may be summarized briefly as follows. The first sample is used to establish any desired number of intervals into which the observations of the second sample may fall. A test criterion is proposed which is based on the deviations of the numbers of elements of the second sample which fall in the intervals from the expected values of the respective numbers. Two cases are discussed, that of two intervals and that of four intervals, making use of the median and quartiles in the first sample to determine the intervals. Tables of 1% and 5% points for several sample sizes of both cases are given.

REFERENCES

[1] A. Wald and J. Wolfowitz, *Annals of Math. Stat.*, Vol. 11 (1940), p. 147.
[2] W. J. Dixon, *Annals of Math. Stat.*, Vol. 11 (1940), p. 199.
[3] Catherine Thompson, *Biometrika*, Vol. 32 (1941), p. 151.