NOTES

This section is devoted to brief research and expository articles, notes on methodology and other short items.

A CRITERION FOR TESTING THE HYPOTHESIS THAT TWO SAMPLES ARE FROM THE SAME POPULATION

By W. J. Dixon

1. Introduction. The purpose of this paper is to consider a criterion for testing the hypothesis that two samples have been drawn from populations with the same distribution function, assuming only that the cumulative distribution function common to the two populations is continuous. Let the two samples, O_n and O_m, be of size n and m respectively. We may assume $n \leq m$ without loss of generality. Suppose the elements $u_1, \cdots, u_n$ of O_n are arranged in order from the smallest to the largest, that is, $u_1 < u_2 < \cdots < u_n$. These may be represented as points along a line. The elements of 0_m represented as points on the same line are then divided into $(n + 1)$ groups by the first sample, O_n. Let m_1 be the number of points having a value less than u_1, m_i the number lying between u_i and u_{i+1}, $(i = 1, 2, \cdots, n)$ and m_{n+1} the number greater than u_n, $(m_{n+1} = m - m_1 - m_2 - \cdots - m_n)$. The criterion here proposed is[1]

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

[1] A similar criterion

$$d^2 = \sum_{i=0}^{n} \left(\frac{i}{n} - \frac{\sum_0^i n_j}{n} \right)^2$$

for two samples of the same size was investigated (unpublished) by A. M. Mood. He found the mean and variance to be

$$E(d^2) = \frac{2n+1}{3n}, \qquad \sigma^2_{d^2} = \frac{8(n-1)(2n+1)}{45n^2}.$$

It can be seen that this is the sum of the squares of the differences between the ordinates of the two cumulative sample distributions calculated at the jumps of the first sample distribution.

2. The mean and variance of C^2. The only case of continuous cumulative distribution functions $F(x)$ of any interest in statistics is that in which $dF(x) = f(x)\,dx$, where $f(x)$ is a probability density function. Let us write:

$$p_1 = \int_{-\infty}^{u_1} f(x)\,dx, \qquad p_2 = \int_{u_1}^{u_2} f(x)\,dx, \cdots, p_{n+1} = \int_{u_n}^{\infty} f(x)\,dx,$$

where of course $p_{n+1} = 1 - p_1 - p_2 - \cdots - p_n$.

Now, the joint distribution law of the p_i is

$$P(p_1, \cdots, p_n) = n!\,dp_1 \cdots dp_n \tag{2}$$

and the conditional distribution of the m_i given the p_i is

$$P(m_1, \cdots, m_{n+1} \mid p_1, \cdots, p_n) = \frac{m!}{m_1! \cdots m_{n+1}!} p_1^{m_1} p_2^{m_2} \cdots p_{n+1}^{m_{n+1}}. \tag{3}$$

Therefore the joint probability law of the m_i and p_i is

$$P(m, p) = \frac{n!\,m!}{m_1! \cdots m_{n+1}!} p_1^{m_1} p_2^{m_2} \cdots p_{n+1}^{m_{n+1}}\,dp_1 \cdots \acute{u}p_n. \tag{4}$$

Let $\varphi(\theta) = \varphi(\theta_1, \cdots, \theta_{n+1}) = E\left[\exp \sum_{i=1}^{n+1} \theta_i \left(\frac{1}{n+1} - \frac{m_i}{m}\right)\right]$; then

$$E(C^2) = \sum_{i=1}^{n+1} \frac{\partial^2 \varphi}{\partial \theta_i^2}\bigg]_{\theta=0}, \tag{5}$$

$$E[(C^2)^2] = \sum_{i=1}^{n+1} \frac{\partial^4 \varphi}{\partial \theta_i^4}\bigg]_{\theta=0} + \sum_{i \neq j} \frac{\partial^4 \varphi}{\partial \theta_i^2\,\partial \theta_j^2}\bigg]_{\theta=0} \tag{6}$$

and

$$\varphi(\theta) = \sum\nolimits_m \int \exp\left[\sum_{i=1}^{n+1} \theta_i \left(\frac{1}{n+1} - \frac{m_i}{m}\right)\right] P(m, p), \tag{7}$$

where Σ_m denotes the usual multinomial summation over all integral values of $m_i \geq 0$ for which $\Sigma m_i = m$ and the integration is over the generalized tetrahedron defined by $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_{n+1} \leq 1$. If we perform the summation first, we obtain

$$\varphi(\theta) = n!\,e^{\sum_{i=1}^{n+1} \frac{\theta_i}{n+1}} \int (p_1 e^{-\frac{\theta_1}{m}} + \cdots + p_{n+1} e^{-\frac{\theta_{n+1}}{m}})^m\,dp_1 \cdots dp_n. \tag{8}$$

Differentiating twice with respect to θ_i and setting the θ's equal to zero, we get

$$\frac{\partial^2 \varphi}{\partial \theta_i^2}\bigg]_{\theta=0} = n! \int \left[\left(\frac{1}{n+1}\right)^2 + \left(\frac{1}{m} - \frac{2}{n+1}\right) p_i + \frac{m-1}{m} p_i^2\right] dp_1 \cdots dp_n.$$

If we now integrate and sum from one to $n + 1$, we find

$$E(C^2) = \frac{n(n+m+1)}{m(n+1)(n+2)}. \tag{9}$$

Performing the operations indicated in (6), we obtain $E[(C^2)^2]$ from which we subtract $[E(C^2)]^2$ and have as the variance of C^2,

$$\sigma^2_{C^2} = \frac{4n(m-1)(m+n+1)(m+n+2)}{m^3(n+2)^2(n+3)(n+4)}.$$

3. Significance values of C^2. If we let C^2_α be defined as the smallest value of C^2 for which $P(C^2 \geq C^2_\alpha) \leq \alpha$ then we can compute the value of C^2_α fairly readily for small values of m and n. The values of C^2 for $m, n \leq 10$ are given in Table I for $\alpha = 0.01, 0.05$ and 0.10. Since the distribution of C^2 is not continuous the probabilities $P(C^2 \geq C^2_\alpha)$ will, in general, be less than α.

TABLE I

Values of C^2_α. $\alpha = 0.01, 0.05, 0.10$

m \ n	2	3	4	5	6	7	8	9	10
4	—	—	—						
	—	—	—						
	—	—	.800						
5	—	—	—	—					
	—	—	.800	.833					
	—	.750	.800	.833					
6	—	—	—	—	.857				
	—	.750	.800	.833	.857				
	—	.750	.800	.556	.413				
7	—	—	—	.833	.857	.875			
	—	.750	.800	.588	.612	.467			
	.667	.750	.555	.425	.449	.426			
8	—	—	.800	.833	.857	.656	.670		
	—	.750	.800	.594	.482	.469	.389		
	.667	.531	.425	.413	.357	.375	.358		
9	—	—	.800	.833	.660	.677	.543	.554	
	—	.750	.602	.448	.413	.431	.395	.381	
	.667	.552	.454	.389	.363	.356	.321	.307	
10	—	—	.800	.833	.677	.555	.549	.480	.449
	.667	.750	.480	.493	.437	.415	.349	.340	.349
	.487	.430	.380	.373	.357	.315	.309	.280	.269

It will be seen that if m and n increase indefinitely in the ratio $n/m = \gamma$, then nC^2 converges stochastically to $\gamma + 1$ whereas nC^2 ranges from 0 to $n^2/(n + 1)$ which indicates a tail to the right. This suggests that for larger values of m and n, it is reasonable to try to fit the distribution of nC^2 by the method of moments using a distribution of the form

$$(11) \qquad \frac{(kx^2)^{\frac{1}{2}\nu-1}}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)} e^{-\frac{1}{2}kx^2} d(kx^2)$$

which has

$$E(x^2) = \frac{\nu}{k}, \qquad \sigma_{x^2}^2 = \frac{2\nu}{k^2}.$$

Setting $x^2 = nC^2$, we see that we can consider nkC^2 distributed as χ^2 with ν degrees of freedom. Of course, ν is not necessarily an integer, but χ^2 tables may be used for approximate values of the probability that nkC^2 will exceed certain values,[2] or the values of nkC^2 that will be exceeded a certain per cent of the time.[3] More exact values of these probabilities that nkC^2 will exceed a certain value may be found from a table of the incomplete Gamma function.[4]

To calculate k and ν directly, the following formulas obtained by equating the mean and variance of (11) to the mean and variance of nC^2 may be used:

$$(12) \qquad k = am(n + 2)/n, \qquad \nu = an(n + m + 1)/(n + 1),$$

where

$$a = \frac{m(n + 3)(n + 4)}{2(m - 1)(m + n + 2)(n + 1)}.$$

If the fitted curve (11) is used to obtain significance values of nC^2, there is a tendency toward rejecting slightly over $100\alpha\%$, especially for small values of m and n. The error is probably due to fitting a curve having an infinite range. The discrepancy decreases as m and n increase.

The goodness of fit at the 0.01, 0.05 and 0.10 significance levels was tested for two cases.

Case 1. $n = 9$, $m = 10$; $nk = \frac{2860}{63}$, $\nu = \frac{52}{7}$.

The exact distribution in the region under consideration is the following:

C_0^2	...	.26	.28	.30	.32	.34	.36	.40	.42	.44	.48	...
$P(C^2 \geq C_0^2)$	...	.121	.090	.082	.072	.037	.033	.025	.025	.015	.007	...

The values of C_α^2 from the fitted curve are $C_{.01}^2 = 0.422$, $C_{.05}^2 = 0.323$ and $C_{.10}^2 = 0.277$. The double rule indicates the divisions (from the fitted curve) for $\alpha = 0.01$, 0.05 and 0.10.

[2] Karl Pearson, *Tables for Statisticians and Biometricians*, part 1, Table XII.

[3] R. A. Fisher, *Statistical Methods for Research Workers*, Table III.

[4] *Tables of the Incomplete Gamma Function*, Biometrika Office, London.

Case 2. $n = 12, m = 12; nk = 65.068, \nu = 8.938$.

The important part of the exact distribution for our purposes is:

C_0^2	…	.215	.229	.243	.256	.270	…	.326	.340	.354	.381	…
$P(C^2 \geq C_0^2)$	…	.120	.109	.078	.057	.046	…	.017	.014	.011	.009	…

The values of C_α^2 from the fitted curve are $C_{.01}^2 = 0.3315$, $C_{.05}^2 = 0.2587$ and $C_{.10}^2 = 0.2244$.

4. Examples. *1.* Two samples of ten members each are drawn and it is desired to test, using a rejection region of size α, the hypothesis that these two samples could have originated from the same population about which nothing is assumed except that it is continuous. The first sample was found to divide the second sample into the following groups: 0, 0, 0, 3, 0, 4, 0, 0, 2, 1, 0.

$$C^2 = (\tfrac{1}{11} - \tfrac{4}{10})^2 + (\tfrac{1}{11} - \tfrac{3}{10})^2 + (\tfrac{1}{11} - \tfrac{2}{10})^2 + (\tfrac{1}{11} - \tfrac{1}{10})^2 + 7(\tfrac{1}{11})^2 = .209$$

which we see from Table I is not a significant value even for $\alpha = 0.10$ since $C_{.10}^2 = 0.269$.

2. A sample of 15 divides a second of 25 into the following 16 groups: 0, 1, 0, 0, 5, 4, 1, 3, 9, 0, 0, 1, 0, 1, 0, 0.

$$C^2 = (\tfrac{1}{16} - \tfrac{9}{25})^2 + (\tfrac{1}{16} - \tfrac{5}{25})^2 + (\tfrac{1}{16} - \tfrac{4}{25})^2 + (\tfrac{1}{16} - \tfrac{3}{25})^2 + 4(\tfrac{1}{16} - \tfrac{1}{25})^2 + 8(\tfrac{1}{16})^2$$

$$nC^2 = 2.302 \qquad k = 7.511 \qquad \nu = 10.19$$

$$nkC^2 = 17.295$$

which gives a significant value for $\alpha = 0.10$ but not for $\alpha = 0.05$, since $nkC_{.10}^2 = 16.233$, $nkC_{.05}^2 = 18.568$. Actually $P(nkC^2 > 17.29) = .077$.

5. Remarks. If we set W equal to the number of m_i which are zero and $V = n + 1 - W$ then V is the number of non-zero m_i; further, $2V \cong U$ where U is the total number of runs, the criterion proposed in the paper of Wald and Wolfowitz in the present issue of the *Annals of Mathematical Statistics*. Now,

$$W = \lim_{x_1, \cdots, x_{n+1} \to 0} \sum_{i=1}^{n+1} x_i^{m_i}, \tag{13}$$

so that, setting

$$\Phi = \textstyle\sum_m \displaystyle\int \exp\left[\sum_{i=1}^{n+1} \theta_i \left(\frac{1}{n+1} - \frac{m_i}{m}\right)\right] \sum_{i=1}^{n+1} x_i^{m_i} P(m, p), \tag{14}$$

analogous to (7), we have

$$E(WC^2) = \lim_{x_1, \cdots, x_{n+1} \to 0} \left. \sum_{k=1}^{n+1} \frac{\partial^2 \Phi}{\partial \theta_k^2} \right]_{\theta=0}$$

from which we can find

$$\rho_{VC^2}\sigma_V\sigma_{C^2} = \frac{2n(1 - m)}{m(n + 2)(m + n)}$$

and

$$\rho^2_{UC^2} \cong \rho^2_{VC^2} = \frac{(n + 3)(n + 4)(m + n - 1)}{(n + 1)(m + n + 1)(m + n + 2)}.$$

If $n/m = \gamma$ (a fixed constant) and n is large

$$\rho^2 \cong \frac{n}{n + m}.$$

ρ^2 will be near 1 when n is much larger than m. This corresponds, in computing C^2, to dividing the smaller sample into subgroups by the larger. In this case U and C^2 give essentially the same information. When m and n are more nearly equal the two criteria are quite different. For $n > m$, C^2 has fewer possible values than for $n < m$, and is therefore a more sensitive test when $n < m$.

While it is doubtful that this test is biased for large samples, this question will not be considered in the present note.

Princeton University,
Princeton, N. J.

SIGNIFICANCE TEST FOR SPHERICITY OF A NORMAL n-VARIATE DISTRIBUTION

By John W. Mauchly

1. Introduction. This note is concerned with testing the hypothesis that a sample from a normal n-variate population is in fact from a population for which the variances are all equal and the correlations are all zero. A population having this symmetry will be called "spherical." Under a linear orthogonal transformation of variates, a spherical population remains spherical, and consequently the features of a sample which furnish information relevant to this hypothesis must be invariant under such transformations.

A situation for which this test is indicated arises when the sample consists of N n-dimensional vectors, for which the variates are the n components along coordinate axes known to be mutually perpendicular, but having an orientation which is, a priori at least, quite arbitrary. A specific application for two dimensions, treated elsewhere [1], may be mentioned. Each of N days furnishes a sine and a cosine Fourier coefficient for a given periodicity, and these, when plotted as ordinate and abcissa, yield a somewhat elliptical cloud of N points. The sine and cosine functions are orthogonal, and their variances have