TEST OF HOMOGENEITY FOR NORMAL POPULATIONS

By G. A. Baker

University of California

1. Introduction. In biological experiments it is often of interest to test whether or not all the subjects can be regarded as coming from the same normal population. If they have not come from the same normal population, usually the most plausible alternative is that the subjects have come from a population which is the combination of two or more normal populations combined in some proportions. The combination of normal populations is a "smooth" alternative to the hypothesis of a single normal population. Such non-homogeneous populations are not the only "smooth" alternatives, of course, but are included among the "smooth" alternatives. If there is reason to believe that the only deviation from a normal population is due to non-homogeneity, then the results of Professor Neyman in his paper [1] are available in studying this problem.

It is desirable not to make any hypotheses about the mean and standard deviation of the sampled population, but to base all computations and tests on the data contained in the sample. Such a viewpoint has been stressed in a previous paper [2] where it was shown that if the sampling is from a normal population, the probability of a deviation from the mean of a first sample of n measured in terms of the standard deviation of the sample is proportional to

$$\frac{dv}{\left(1+\frac{v^2}{n+1}\right)^{n/2}}. \tag{1.1}$$

The result (1.1) and Neyman's results give rise to a test of homogeneity which is valid for "large" samples. Empirical results show that fairly conclusive evidence of non-homogeneity may be obtained with samples of 100. Samples of 50 or less may be suggestive but rarely decisive.

2. Development of Test. Suppose that a sample of $n + 1$ is drawn from a normal population. It can be regarded as being made up of a first sample of n and a second sample of one. The value of v corresponding to (1.1) can then be computed and its distribution function is (1.1). This partition, of course, can be made in $n + 1$ ways. That is, $n + 1$ values of v are determined from a random sample of $n + 1$ from the original parent. It is true that these values of v are not independent among themselves. The correlation between the values of v, to a first approximation at least, is of the order of $1/n$ and can be neglected if n is "large."

A suitable transformation as discussed in [3], [1] and elsewhere, transforms (1.1) into a rectangular distribution.

If the same computations are made when the sampled population is not normal, then the resulting values obtained will not be rectangularly distributed. For instance, suppose that the sampled population is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \left(pe^{-\frac{1}{2}(x-m_1)^2/\sigma^2} + qe^{-\frac{1}{2}(x-m_2)^2/\sigma^2}\right) \qquad . \tag{2.1}$$

we find that the distribution of v based on the first sample of 2 is a very complicated expression involving sums of exponentials and definite integrals of exponentials. To obtain a rectangular distribution if the sampled population is normal, the appropriate transformation to make is

$$\begin{aligned} v &= -\sqrt{3}\cot \pi u \\ dv &= \sqrt{3}\,\pi \csc^2 \pi u \, du. \end{aligned} \tag{2.2}$$

The resulting u-distribution for population (2.1) then is to be compared with the rectangular distribution in the interval from zero to one.

For "large" values of $n + 1$ and for symmetrical non-homogeneous populations composed of two normal components, the u-distribution will be symmetrical about $u = \frac{1}{2}$, less than one near the ends, greater than one for values of u moderately far from $\frac{1}{2}$ and less than one for values of u near $\frac{1}{2}$. A Neyman [1] Ψ_k^2 of order 4 will be necessary to detect a difference of this sort. If the non-homogeneous population of two components is skewed, the u-distribution will still show the same two-humped effect but may be skewed instead of symmetrical. A Neyman Ψ_k^2 of order 4 should still be computed, although Ψ_3^2 may be more significant.

The test then consists of:

(a) computing the $n + 1$ quantities

$$x_i' = \frac{x_i - \bar{x}}{\sqrt{n+1}\,s}, \qquad (i = 1, 2, 3, \cdots, n + 1) \tag{2.3}$$

where

$$\begin{aligned} n + 1 &= \text{number in the sample} \\ x_i &= \text{the observed values} \\ x_j &= \text{the observed values except } x_i \end{aligned}$$

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

(b) making the transformation

$$u_i = \int_{-\infty}^{x_i'} \frac{y_0\,dx'}{(1 + x'^2)^{n/2}}, \qquad (i = 1, 2, 3, \cdots, n + 1).$$

(c) computing the first four Ψ_k^2's of Neyman's paper [1]

(d) comparing Ψ_k^2 with $\Psi_\epsilon^2(k)$ as found from the Incomplete Gamma Function Tables.

If n is large, say $n = 100$, then u is given approximately by the normal probability integral.

If n is small, the values of u are obtained from the Table 25 of Vol. 2 of Pearson's Tables.

Neyman's derivation assumes that $n + 1$ is large and that the u's are independent. In this case, if $n + 1$ is large, then the u's are nearly independent, and hence the test is valid. The same procedure can be applied for smaller samples. It can not be expected that small differences from normal in the sampled population can be detected with small samples. Empirical results indicate that samples of 100 are necessary for decisive results even when the differences of the sampled population from a normal homogeneous population are large. Samples of 50 may be suggestive and in very extreme cases might be decisive.

TABLE I

Empirical Sampling Results

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Ψ_k^2's for 51 from population A	.0001	.843	2.009	7.464
Ψ_k^2's for 101 from population A	.086	2.403	4.998	12.868
Ψ_k^2's for 101 from population B	.553	.927	7.472	7.485
Ψ_k^2's for 101 from normal	.017	.082	1.288	1.663
$\Psi^2_{(.05)}(k)$'s (Neyman [1])	3.842	5.992	7.815	9.488
$\Psi^2_{(.01)}(k)$'s (Neyman [1])	6.635	9.210	11.345	13.277

It is to be noted that the test makes no assumption about the parameters of the sampled population and does not group the data. The application of the test gives a unique result that does not depend on the judgment of the computer in any respect, In applying the usual chi-square test the computer must choose groupings. The choice of groupings as indicated in [5] may change the P-values to very different levels of significance.

3. Empirical results. Samples of 51 and 101 from population A, of 101 from population B, and of 101 from a normal population, were drawn by throwing dice. Populations A and B are given in [4]. Population A is symmetrical and distinctly bimodal. Population B is weakly bimodal and strongly skewed.

For samples from population A it is necessary to compute Ψ_4^2. For samples from population B it may be sufficient to compute Ψ_3^2. The non-homogeneity of the type of population A seems to be somewhat more detectable than of the type of population B. The sample from the normal parent shows close conformity with expectation.

In applying the proposed test for homogeneity the u-values for small independent sets of data can be combined to give a much larger number of u-values.

REFERENCES

[1] J. Neyman, "«Smooth Test» for goodness of fit," *Skandinavisk Aktuariedskrift*, (1937), pp. 149-199.

[2] G. A. Baker, "The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 197-201.

[3] G. A. Baker, "Transformations of bimodal distributions," *Annals of Math. Stat.*, Vol. 1 (1930), pp. 334-344.

[4] G. A. Baker, "The relation between the means and variances, means squared and variance in samples from the combinations of normal populations," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 333-354.

[5] G. A. Baker, "The significance of the product-moment coefficient of correlation with special reference to the character of the marginal distributions," *Jour. Am. Stat. Assoc.*, Vol. 25 (1930), pp. 387-396.

A NOTE ON THE POWER OF THE SIGN TEST

By W. Mac Stewart

University of Wisconsin

1. Introduction. Let us consider a set of N non-zero differences, of which x are positive and $N - x$ are negative; and suppose that the hypothesis tested, H_0, implies, in independent sampling, that x will be distributed about an expected value of $N/2$ in accordance with the binomial $(\frac{1}{2} + \frac{1}{2})^N$. As a quick test of H_0, we may choose to test the hypothesis h_0 that x has the above probability distribution. Defining r to be the smaller of x and $N - x$, the test consists in rejecting h_0 and therefore H_0 whenever $r \leq r(\epsilon, N)$, where $r(\epsilon, N)$ is determined by N and the significance level ϵ.

2. Power of a test. In applying such a test it is of interest to know how frequently it will lead to a rejection of H_0 when H_0 is false and the situation H implies that the probability law of x is $(q + p)^N$, with $p \neq \frac{1}{2}$, thereby indicating an expectation of an unequal number of + and − differences. The probability of rejecting H_0 when H_1 implying $p = p_1$ is true, is termed the *power* of the test of H_0 relative to the alternative H_1.[1] Thus, from the point of view of experimental design the power (P) of the test of H_0 may be considered a function of the alternative hypothesis H_1, the significance level ϵ, and N. As such, the following observations may be noted:

1. The power P_2, for an assumed ϵ, N, and H_2 implying $p = p_2$ is greater than or equal to the power P_1 for ϵ, N and H_1 implying $p = p_1$ where $|p_2 - .50| > |p_1 - .50|$.

[1] For an extensive discussion of the power of a test, the reader is referred to J. Neyman and E. S. Pearson, *Statistical Research Memoirs*, Vol. 1 (1936), pp. 3-6.