The research in this report is partially supported by the U.S. Army under Contract DAHCO 4-71-C-0042.

SUMMARY

Methods are described for testing whether an observed sample is consistent with it being from a mixture (in unknown proportions) of two specified symmetrical populations. The tests use only very simple sample statistics. Possible tests for use when more complete sample data are available are also mentioned.

SOME SIMPLE TESTS OF MIXTURES WITH SYMMETRICAL COMPONENTS

N. L. Johnson

Department of Statistics
University of North Carolina at Chapel Hill
27514

Institute of Statistics Mimeo Series No. 808

February, 1972

SOME SIMPLE TESTS OF MIXTURES WITH SYMMETRICAL COMPONENTS

N. L. Johnson
University of North Carolina at Chapel Hill

1. Introduction

In this report we consider situations in which it is expected that a given (univariate) population may be a mixture of two known populations in unknown proportions ω: $1 - \omega$. It is desired to construct tests of this hypothesis which are robust with respect to variation in ω. We further restrict ourselves to using only simple functions of the data - the total of all sample values, and proportions between fixed limits.

Two specific test procedures - A and B - will be developed, and their appropriate powers evaluated, in the case when the two component populations are normal with common variance. More powerful tests could be constructed using individual sample values, but the statistics used in tests A and B can sometimes be obtained when more detailed information cannot. For example, total weight of 1000 items is more expeditiously obtained than the 1000 individual weights. So also is the number of items with weight exceeding a specified value. These are the statistics used in test A. In test B, the proportion of items with weight between fixed limits is used.

The research in this report is partially supported by the U.S. Army under Contract DAHCO 4-71-C-0042.

2. Test A

Suppose that X is distributed as a mixture of component distributions in proportions $\omega: 1-\omega$. Suppose further that the component distributions are symmetrical, with expected values ξ_1, ξ_2 respectively, and differ only in respect of location parameter. We denote the common variance by σ^2.

If $X_1, X_2, \ldots, X_n$ are independently distributed as X then $\bar{X}(= n^{-1} \sum_{j=1}^{n} X_j)$ has expected value

$$(1.1) \qquad E[\bar{X}] = \omega\xi_1 + (1-\omega)\xi_2$$

and variance

$$(1.2) \qquad \mathrm{var}(\bar{X}) = n^{-1}[\sigma^2 + \omega(1-\omega)(\xi_1-\xi_2)^2].$$

From (1.1) we see that

$$\hat{\omega}_X = (\bar{X}-\xi_2)/(\xi_1-\xi_2)$$

is an unbiased estimator of ω.

Another unbiased estimator of ω can be based on the number of X_j's which are less than some fixed value θ.

If

$$Y_j = \begin{cases} 1 & \text{if } X_j < \theta \\ 0 & \text{if } X_j \geq \theta \end{cases}$$

then the proportion of X's less than θ is

$$\bar{Y} = n^{-1} \sum_{j=1}^{n} Y_j$$

and

$$(2.1) \qquad E[\bar{Y}] = \omega P_1 + (1-\omega)P_2$$

where

$$P_j = \Pr[X < \theta | \xi_j] \quad (j = 1,2).$$

Also

$$(2.2) \qquad \operatorname{var}(\bar{Y}) = n^{-1}[\omega P_1(1-P_1) + (1-\omega)P_2(1-P_2) + \omega(1-\omega)(P_1-P_2)^2].$$

From (2.1) we see that

$$\hat{\omega}_Y = (\bar{Y}-P_2)/(P_1-P_2)$$

is also an unbiased estimator of ω.

If the two estimators $\hat{\omega}_X$ and $\hat{\omega}_Y$ differ greatly, this may be regarded as evidence that X is not in fact distributed as a mixture of the two specified components.

If n is not too small, it is likely that $\hat{\omega}_X$, $\hat{\omega}_Y$ and $(\hat{\omega}_X-\hat{\omega}_Y)$ will be approximately normally distributed. If X is distributed as a mixture of the two specified distributions, the expected value of $(\hat{\omega}_X-\hat{\omega}_Y)$ is zero, whatever be the value of ω. If $\operatorname{var}(\hat{\omega}_X-\hat{\omega}_Y)$ were also independent of ω, the ratio

$$(\hat{\omega}_X - \hat{\omega}_Y)[\operatorname{var}(\hat{\omega}_X-\hat{\omega}_Y)]^{-\frac{1}{2}}$$

which can be computed without knowing the value of ω, should have approximately a unit normal distribution.

Now

$$\begin{aligned}(3) \quad \operatorname{var}(\hat{\omega}_X - \hat{\omega}_Y) &= \operatorname{var}(\hat{\omega}_X) + \operatorname{var}(\hat{\omega}_Y) - 2\operatorname{cov}(\hat{\omega}_X,\hat{\omega}_Y) \\ &= (\xi_1-\xi_2)^{-2}\operatorname{var}(\bar{X}) + (P_1-P_2)^{-1}\operatorname{var}(\bar{Y}) \\ &\quad - 2(\xi_1-\xi_2)^{-1}(P_1-P_2)^{-1}\operatorname{cov}(\bar{X},\bar{Y}).\end{aligned}$$

Also

$$\operatorname{cov}(\bar{X},\bar{Y}) = n^{-2}\operatorname{cov}\left(\sum_{j=1}^{n} X_j, \sum_{j=1}^{n} Y_j\right)$$

$$= n^{-1} \operatorname{cov}(X,Y)$$

where

$$Y = \begin{cases} 1 & \text{if } X < \theta \\ 0 & \text{if } X \geq \theta. \end{cases}$$

Hence

$$E[XY] = \Pr[X < \theta]\, E[X|X < \theta]$$
$$= \omega P_1 E_1 + (1-\omega) P_2 E_2$$

where

$$E_j = E[X|X < \theta, \xi_j] \quad (j = 1,2).$$

Hence

$$(4) \quad \operatorname{cov}(X,Y) = \omega P_1 E_1 + (1-\omega)\, P_2 E_2 - \{\omega\xi_1 + (1-\omega)\xi_2\}\{\omega P_1 + (1-\omega) P_2\}$$
$$= \omega(P_1(E_1-\xi_1) + (1-\omega)\, P_2(E_2-\xi_2) - \omega(1-\omega)(\xi_1-\xi_2)(P_1-P_2)$$

Inserting (1.2), (2.2) and (4) in (3) we obtain

$$(5) \quad n \operatorname{var}(\hat{\omega}_X - \hat{\omega}_Y) = (\xi_1-\xi_2)^{-2} \sigma^2 + (P_1-P_2)^{-2} [\omega P_1(1-P_1) + (1-\omega) P_2(1-P_2)]$$
$$- 2(\xi_1-\xi_2)^{-1} (P_1-P_2)^{-1} [\omega P_1(E_1-\xi_1) + (1-\omega) P_2(E_2-\xi_2)].$$

If the component distributions are continuous with density functions $\sigma^{-1} f((x-\xi_j)\sigma^{-1})$ $(j = 1,2)$ then

$$(6.1) \quad P_j = \int_{-\infty}^{\theta_j} f(t)\, dt$$

and

$$(612) \quad P_j(E_j-\xi_j) = \sigma \int_{-\infty}^{\theta_j} t f(t)\, dt$$

where

$$\theta_j = (\theta-\xi_j)\sigma^{-1} \quad (j = 1,2).$$

If we take $\theta = \frac{1}{2}(\xi_1+\xi_2)$ and if $\Pr[X = \frac{1}{2}(\xi_1+\xi_2)] = 0$, (which is certainly true if X has a continuous distribution) then because of the symmetry and identity of shape of the component distributions

$$(7.1) \qquad P_1 = 1 - P_2$$

and

$$(7.2) \qquad P_1(E_1-\xi_1) = P_2(E_2-\xi_2).$$

As a consequence

$$(8) \quad n \operatorname{var}(\hat{\omega}_X-\hat{\omega}_Y) = (\xi_1-\xi_2)^{-2}\sigma^2 + (P_1-P_2)^{-2} P_1(1-P_1) - 2(\xi_1-\xi_2)^{-1}(P_1-P_2)^{-1} P_1(E_1-\xi_1)$$

which does not depend on ω. We note that this can also be the case when the component distributions are not symmetrical, provided they are not identical in shape, but are mirror-images of each other.

In the particular case where the component distributions are normal

$$f(t) = (\sqrt{2\pi})^{-1} \exp(-\tfrac{1}{2}t^2) = \phi(t)$$

and, with $\theta = \frac{1}{2}(\xi_1+\xi_2)$

$$P_1 = \Phi(\tfrac{1}{2}\Delta)$$
$$P_1(E_1-\xi_1) = -\sigma\phi(\tfrac{1}{2}\Delta)$$

where

$$\Phi(u) = \int_{-\infty}^{u} \phi(t)\,dt \quad \text{and} \quad \Delta = \tfrac{1}{2}(\xi_2-\xi_1)\sigma^{-1}.$$

So, in this case

$$(9) \quad n \operatorname{var}(\hat{\omega}_X-\hat{\omega}_Y) = (\xi_1-\xi_2)^{-2}\sigma^2 + 2(\xi_1-\xi_2)^{-1}(P_1-P_2)^{-1}\sigma\phi(\tfrac{1}{2}\Delta) + (P_1-P_2)^{-2} P_1(1-P_1)$$
$$= V, \text{ say.}$$

(Note that since $P_1 \lesseqgtr P_2$ according as $\xi_1 \gtreqless \xi_2$, the second term on the right hand side is negative.)

If X is distributed as a mixture, in any proportions, of the two specified normal distributions, then the statistic

$$\sqrt{n}(\hat{\omega}_X - \hat{\omega}_Y)V^{-\frac{1}{2}}$$

has zero mean and unit standard deviation, and is approximately normally distributed for sufficiently large n.

We propose to use, as critical region for test A,

$$(10) \qquad |\sqrt{n}(\hat{\omega}_X - \hat{\omega}_Y)V^{-\frac{1}{2}}| > u_{1-\frac{1}{2}\alpha}$$

where

$$\Phi(u_{1-\frac{1}{2}\alpha}) = 1 - \tfrac{1}{2}\alpha$$

and α is the (approximate) significance level of the test.

3. Power of Test A.

There are so many plausible kinds of departure from hypothesis (of a mixture of two specified distributions) that it is inevitable that power with respect to some alternatives will be low, even when there is substantial discrepancy between the alternative and the hypothesis tested. It is desirable that circumstances where this may happen should be recognized. Such recognition is aided by a knowledge of some features of variation in power in typical cases.

Here, we will study power against the alternative that X has a single homogeneous normal distribution with expected value μ and standard deviation τ. Then, for the test A described at the end of Section 2

$$(10) \qquad E[\hat{\omega}_X] = (\mu-\xi_2)/(\xi_1-\xi_2)$$
$$E[\hat{\omega}_Y] = (P-P_2)/(P_1-P_2)$$

where

$$P = \Phi(\tau^{-1}[\tfrac{1}{2}(\xi_1+\xi_2) - \mu]).$$

Also

$$(11) \qquad n \operatorname{var}(\hat{\omega}_X-\hat{\omega}_Y) = (\xi_1-\xi_2)^{-2}\tau^2 + (P_1-P_2)^{-2}P(1-P) + 2(\xi_1-\xi_2)^{-1}(P_1-P_2)^{-1}\tau\phi(\tau^{-1}[\tfrac{1}{2}(\xi_1+\xi_2)-\mu]).$$

Hence $\sqrt{n}(\hat{\omega}_X-\hat{\omega}_Y)V^{-\frac{1}{2}}$ is distributed approximately normally with expected value

$$(12.1) \qquad \left(\frac{\mu-\xi_2}{\xi_1-\xi_2} - \frac{P-P_2}{P_1-P_2}\right)\sqrt{\frac{n}{V}}$$

and variance

$$(12.2) \qquad V^{-1}[(\xi_1-\xi_2)^{-2}\tau^2 + 2(\xi_1-\xi_2)^{-1}(P_1-P_2)^{-1}\tau\phi(\tau^{-1}[\tfrac{1}{2}(\xi_1+\xi_2)-\mu]) + (P_1-P_2)^{-2}P(1-P)]$$

with P as given in (10).

On this basis approximate powers of test A with approximate 5% significance level (critical region

$$|\sqrt{n}(\hat{\omega}_X-\hat{\omega}_Y)| > 1.96V^{\frac{1}{2}})$$

were calculated, with respect to alternatives defined by μ and τ so chosen that the mean and variance agree with those of a mixture of the two specified distributions. This is achieved by taking

$$(13) \qquad \tau^2 = \sigma^2 + \tilde{\omega}(1-\tilde{\omega})(\xi_1-\xi_2)^2 = \sigma^2 + (\xi_1-\mu)(\mu-\xi_2)$$

with

$$\tilde{\omega}\xi_1 + (1-\tilde{\omega})\xi_2 = \mu.$$

Table 1a shows results of these calculations. It is to be expected that higher

powers will be attained for other values of τ. Table 1b sheds some light on this point.

When $\mu = 0$, the approximate power remains constant as n increases. It is, in fact, equal to

$$2\Phi(-1.96\tau^{-1}/V).$$

Some numerical values (with τ given by (13) and $\sigma = 1$) are

$\frac{1}{2}\lvert\xi_1 - \xi_2\rvert$	0.5	1.0	1.5	2.0
Power	0.864	0.069	0.473	0.739 .

Clearly this is not satisfactory. Test B, to be described in Section 4 will usually have a power increasing with n, even when $\mu = 0$.

Table 1: Approximate Power of Test A

($\sigma = 1$; level of significance 5%; $\xi_2 = -\xi_1$; $\theta = 0$;
the power is the same for $-\mu$ as for μ)

Table 1a ($\tau^2 = 1 + \xi_1^2\{1-(\mu/\xi_1)^2\}$)

ξ_1 =	0.5		1.0		1.5		2.0	
$\lvert\mu/\xi_1\rvert$ n =	100	400	100	400	100	400	100	400
0.2	0.053	0.057	0.086	0.143	0.171	0.347	0.271	0.519
0.4	0.056	0.069	0.127	0.306	0.304	0.717	0.460	0.879
0.6	0.057	0.074	0.151	0.398	0.378	0.844	0.544	0.948
0.8	0.053	0.064	0.115	0.276	0.266	0.668	0.367	0.785

Table 1b

τ =	0.5		1.0		2.0		5.0	
$\lvert\mu/\xi_1\rvert$ n =	100	400	100	400	100	400	100	400
$\xi_1 = 1$								
0.2	0.798	1.000	0.079	0.130	0.250	0.500	0.693	0.753
0.4	1.000	1.000	0.109	0.262	0.505	0.915	0.753	0.895
0.6	1.000	1.000	0.116	0.305	0.858	0.997	0.827	0.975
0.8	1.000	1.000	0.083	0.180	0.918	1.000	0.895	0.997
$\xi_1 = 2$								
0.2	1.000	1.000	0.674	0.989	0.201	0.347	0.685	0.853
0.4	1.000	1.000	0.962	1.000	0.385	0.789	0.853	0.993
0.6	1.000	1.000	0.981	1.000	0.679	0.989	0.958	1.000
0.8	1.000	1.000	0.778	1.000	0.920	1.000	0.993	1.000

4. Test B

Other types of test can be based on the values of $Y_i = |X_i - \frac{1}{2}(\xi_1+\xi_2)|$. For each of the two component destributions, Y_i has the same distribution, which it will also have for a mixture, in any proportion, of these distributions. When the component distributions are normal Y_i has a "folded normal" distribution.

Our test B uses as test statistic the proportion of Y_i values which are less than $\frac{1}{2}|\xi_1-\xi_2|$. This is just the proportion of X_i values falling between the two component expected values ξ_1 and ξ_2. Tables 2a and 2b give approximate powers of this test (at approximate 5% level) with respect to the same alternatives as in Tables 1a and 1b. The critical region uses both tails of the distribution of the test statistic. The number of X_i's between ξ_1 and ξ_2 has a binomial distribution with parameters n, $\Phi(\sigma^{-1}|\xi_1-\xi_2|) - \frac{1}{2}$ if the hypothesis tested is valid.

Comparison of Tables 1 and 2 shows that although test B is mostly more powerful than test A this is not always the case. In Table 2a it is noteworthy that the power of test B does not vary greatly with $|\mu/\xi_1|$ until $|\mu/\xi_1|$ approaches 1 (when, of course, it tends to the significance level).

Table 2: Approximate Power of Test B

(See notes on Table 1)

Table 2a $(\tau^2 = 1 + \xi_1^2\{1-(\mu/\xi_1)^2\})$

ξ_1	0.5		1.0		1.5		2.0	
$\lvert\mu/\xi_1\rvert$ n =	100	400	100	400	100	400	100	400
0	0.051	0.054	0.137	0.410	0.484	0.972	0.739	1.000
0.2	0.051	0.054	0.137	0.409	0.483	0.972	0.738	1.000
0.4	0.051	0.054	0.134	0.400	0.474	0.969	0.728	0.999
0.6	0.051	0.053	0.122	0.354	0.425	0.947	0.670	0.998
0.8	0.051	0.052	0.086	0.219	0.268	0.769	0.521	0.981

Table 2b

τ =	0.5		1.0		2.0		5.0	
$\lvert\mu/\xi_1\rvert$ n =	100	400	100	400	100	400	100	400
$\xi_1 = 1$								
0	1.000	1.000	0.990	1.000	0.135	0.969	1.000	1.000
0.2	1.000	1.000	0.982	1.000	0.143	0.974	1.000	1.000
0.4	1.000	1.000	0.928	1.000	0.168	0.985	1.000	1.000
0.6	1.000	1.000	0.699	0.999	0.215	0.994	1.000	1.000
0.8	0.954	1.000	0.262	0.754	0.291	0.999	1.000	1.000
$\xi_1 = 2$								
0	1.000	1.000	0.966	1.000	1.000	1.000	0.976	1.000
0.2	1.000	1.000	1.000	1.000	0.945	1.000	0.977	1.000
0.4	1.000	1.000	1.000	1.000	0.837	1.000	0.980	1.000
0.6	1.000	1.000	1.000	1.000	0.522	0.983	0.985	1.000
0.8	1.000	1.000	0.886	1.000	0.139	0.410	0.990	1.000

5. Further Tests

The figures in Tables 1 and 2 show that both tests, A and B have "blind spots". It is to be expected that considerably more powerful tests should be available if fuller data are used. In particular, since the distribution of $Y = |X-\frac{1}{2}(\xi_1+\xi_2)|$ is known (a folded normal with known parameters if the components are normal) if the mixture hypothesis is valid, we can use a chi-square or Kolmogoroff-Smirnoff test (preferably the latter) to assess the agreement of data and hypothesis.

Even without supposing individual values of the X_i's (or Y_i's) to be available, the methods of Section 4 may be extended to construct modifications of test B, appropriate to specific circumstances.

Instead of using as a test statistic the proportion of Y_i's less than $\frac{1}{2}|\xi_1-\xi_2|$, we can use other proportions, designed to test for certain kinds of alternative. For example the statistic

$$\begin{aligned}&(\text{Number of } Y_i\text{'s between } \tfrac{1}{3}|\xi_1-\xi_2| \text{ and } \tfrac{2}{3}|\xi_1-\xi_2|)\\ -\ &(\text{Number of } Y_i\text{'s between } 0 \text{ and } \tfrac{1}{3}|\xi_1-\xi_2|)\end{aligned}$$

might be used if good power were required when $\mu = 0$.