* And Pahlavi University, Shiraz.

On The Distribution of a Symmetric Statistic From a Mixed Population

by

Javad Behboodian*

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 732

February, 1971

On the Distribution of a Symmetric Statistic From a Mixed Population

by

Javad Behboodian
University of North Carolina, Chapel Hill
and Pahlavi University, Shiraz

Abstract

The distribution of a symmetric statistic $T = g(x_1,x_2,\ldots,x_n)$, for a random sample from a mixed population with density $f(x) = pf_1(x)+qf_2(x)$, is a binomial mixture of the densities of the statistics $T_k = g(X_{k1},X_{k2},\ldots,X_{kn})$, $k = 0,1,\ldots,n$, where X_{ki}'s are independent with density $f_1(x)$ if $i \le k$ and density $f_2(x)$ if $i > k$. It is shown how to find the distributions of some important symmetric statistics like sample mean, sample variance, and order statistics by using T_k's. The results are applied to normal and exponential mixtures.

1. Introduction

Consider

$$f(x) = pf_1(x) + qf_2(x), \tag{1}$$

where $f_1(x)$ and $f_2(x)$ are two probability density functions with $0 < p < 1$ and $q = 1-p$. The function $f(x)$ is called the density function of a mixture of two densities $f_1(x)$ and $f_2(x)$ with mixing proportions p and q. Finite mix-

tures of distributions often arise in various biological, psychological, and industrial applications, and have received some general attention recently [1,2,4].

The following probabilistic meaning of (1) might be sometimes useful for studying finite mixtures of distributions. Let X be a random variable with distribution $F(x)$, and let Q_j be a population with density $f_j(x)$ and distribution $F_j(x)$, $j = 1,2,$. Suppose D_j is the event that X comes from population Q_j with $P(D_1) = p$ and $P(D_2) = q$. Now

$$P(X \le x) = P(D_1)\, P(X \le x | D_1) + P(D_2)\, P(X \le x | D_2)$$

or

$$F(x) = pF_1(x) + qF_2(x)$$

for any real number x. This last equality says that X comes from a mixed population whose density is given by (1).

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed as X having density (1). Consider a symmetric function $t = g(x_1, x_2, \ldots, x_n)$ of real variables $x_1, x_2, \ldots, x_n$, and let $T = g(X_1, X_2, \ldots, X_n)$ be a statistic. It is clear that the distribution of the statistic T is invariant under any permutation on X_i's; so we call T a symmetric statistic of the random sample. Exampels of such statistics are sample moments and order statistics.

This paper deals with the distribution of T for a finite mixture in general and the distributions of some important statistics like sample mean, sample variance, and order statistics. The results are applied to normal and exponential mixtures as examples.

2. Probability Density Function of T

The distribution of T, particularly when the sample comes from a mixture of distributions, is usually very complicated. However, the following theorem may be helpful in some special cases.

Theorem. Let $X_1,X_2,\ldots,X_n$ be a random sample from a mixture given by (1), and let $T = g(X_1,X_2,\ldots,X_n)$ be a symmetric statistic. Let $T_k = g(X_{k1},X_{k2},\ldots,X_{kn})$, $k = 0,1,\ldots,n$, be a statistic for which X_{ki}'s are independent with density $f_1(x)$ if $i \le k$ and density $f_2(x)$ if $i > k$. Then, we have

$$f_T(t) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} f_{T_k}(t), \tag{2}$$

that is the density of T is a binomial mixture of the densities of T_k's.

Proof: The joint density of the random sample is

$$h(x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} [pf_1(x_i)+qf_2(x_i)]. \tag{3}$$

Let m be a partition of the set $\{1,2,\ldots,n\}$ into two sets A and B, and denote the set of all such partitions by M. For the sake of notational convenience, we assume that the subscripts a, b, and m always run respectively in the sets A, B and M. Expanding the right side of (3), we obtain

$$h(x_1,x_2,\ldots,x_n) = \sum_{m} p^k q^{n-k} h_m(x_1,x_2,\ldots,x_n), \tag{4}$$

where k is the number of elements in the set A and

$$h_m(x_1,x_2,\ldots,x_n) = \prod_{a} f_1(x_a) \prod_{b} f_2(x_b) \tag{5}$$

is the density of an n-dimensional random vector whose components are independent from each other; k of them whose subscripts belong to A have density $f_1(x)$ and the remaining ones have density $f_2(x)$. It follows from (4) that the joint

density of the random sample $X_1, X_2, \ldots, X_n$ is a mixture of 2^n multivariate densities defined by (5).

Now, by (4), the characteristic function of T becomes

$$\phi_T(u) = \sum_m p^k q^{n-k} E_m[exp(iug(X_1, X_2, \ldots, X_n))] \tag{6}$$

where the expectation is taken with respect to the density (5). We observe, by the symmetry of $t = g(x_1, x_2, \ldots, x_n)$ and the structure of the density (5), that the above expectation is invariant under any permutation on $X_1, X_2, \ldots, X_n$ and its dependence on the partition m is only through k, the number of elements in the set A. Therefore, for any of the $\binom{n}{k}$ different partitions which correspond to a fixed integer k, $0 \le k \le n$, the above expectation is the same as the characteristic function of the statistic $T_k = g(X_{k1}, X_{k2}, \ldots, X_{kn})$, where X_{ki}'s are independent with density $f_1(x)$ if $i \le k$ and density $f_2(x)$ if $i > k$. Considering this simple observation, from (6), we have

$$\phi_T(u) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \phi_{T_k}(u). \tag{7}$$

Inverting (7) termwise, we obtain (2). The theorem is proved. □

Intuitively, we can easily derive the density of T otherwise, by using the probabilistic meaning of a finite mixture given in Section 1.

For this purpose, let E_k, $k = 0, 1, \ldots, n$, be the event that exactly k of the X_i's have density $f_1(x)$ and the remaining ones density $f_2(x)$. Using conditional density, we have

$$f_T(t) = \sum_{k=0}^{n} P(E_k) f_T(t|E_k), \tag{8}$$

where

$$P(E_k) = \binom{n}{k} p^k q^{n-k} \tag{9}$$

and by the symmetry of T one can show

$$f_T(t|E_k) = f_{T_k}(t). \tag{10}$$

It should be noted that the generalization of (2) to a finite mixture of more than two densities is a multinomial mixture, which can be obtained by a similar argument.

3. Distributions of Sample Mean and Sample Variance

To find the distributions of sample mean and sample variance, we first find the distributions of

$$\bar{X}_k = \sum_{i=1}^{n} X_{ki}/n \quad \text{and} \quad S_k^2 = \sum_{i=1}^{n} (X_{ki}-\bar{X}_k)/n \tag{11}$$

by breaking $X_{k1},X_{k2},\ldots,X_{kn}$ into independent variables $X_{k1},X_{k2},\ldots,X_{kk}$ with common density $f_1(x)$ and independent variables $X_{k,k+1},X_{k,k+2},\ldots,X_{kn}$ with common density $f_2(x)$. Now, for a sample of size k from $f_1(x)$, we set

$$\bar{X}_{k1} = \sum_{i=1}^{k} X_{ki}/k, \qquad S_{k1}^2 = \sum_{i=1}^{k} (X_{ki}-\bar{X}_{k1})^2/n, \tag{12}$$

assuming that $\bar{X}_{k1} = 0$ if $k = 0$ and $S_{k1}^2 = 0$ if $k = 0$ or $k = 1$. Similarly, for a sample of size $n-k$ from $f_2(x)$, we set

$$\bar{X}_{k2} = \sum_{i=k+1}^{n} X_{ki}/(n-k), \qquad S_{k2}^2 = \sum_{i=k+1}^{n} (X_{ki}-\bar{X}_{k2})^2/(n-k), \tag{13}$$

assuming that $\bar{X}_{k2} = 0$ if $k = n$ and $S_{k2}^2 = 0$ if $k = n-1$ or $k = n$. It follows from (11)-(13), by some simple computation, that

$$n\bar{X}_k = k\bar{X}_{k1} + (n-k)\bar{X}_{k2} \tag{14}$$

and

$$nS_k^2 = kS_k^2 + (n-k)S_{k2}^2 + k(n-k)(\bar{X}_{k1}-\bar{X}_{k2})^2/n. \tag{15}$$

Thus, if we can find the distributions of the sample mean and sample variance for $f_1(x)$ and $f_2(x)$, we may be able to find the distributions of $\bar{X}_k$ and S_k^2 and then, by using (2), the distributions of $\bar{X}$ and S^2.

EXAMPLE. Let (1) be a mixture of two normal densities with means μ_1 and μ_2 and common variance $\sigma^2 > 0$. It follows from (15) that $\bar{X}_k$ is distributed as a normal variable with mean

$$\bar{\mu}_k = (k/n)\mu_1 + (1-k/n)\mu_2 \tag{16}$$

and variance σ^2/n. To botain the distribution of S_k^2, we observe that S_{k1}^2, S_{k2}^2, and $\bar{X}_{k1}-\bar{X}_{k2}$ are independent from each other by normality of $f_1(x)$ and $f_2(x)$. It is also clear that kS_{k1}^2/σ^2 is $\chi^2(k-1)$, $(n-k)S_{k2}^2/\sigma^2$ is $\chi^2(n-k-1)$, and $k(n-k)(\bar{X}_{k1}-\bar{X}_{k2})^2/n\sigma^2$ is non-central chi-square $\chi^2(1,d_k)$ with one degree of freedom and non-centrality parameter

$$d_k = \sqrt{k(n-k)/n}\,|\mu_1-\mu_2|/\sigma. \tag{17}$$

By the reproductive property of chi-square random variables with respect to degrees of freedom and non-centrality parameter, we conclude that nS_k^2/σ^2 is distributed as $\chi^2(n-1,d_k)$.

It should be noted that when the normal densities $f_1(x)$ and $f_2(x)$ have different variances σ_1^2 and σ_2^2, the distribution of $\bar{X}_k$ is again normal with variance $k\sigma_1^2/n^2+(n-x)\sigma_2^2/n^2$. But it can be shown that S_k^2 is a linear combination of two central and one non-central independent chi-square variables which cannot be summarized as one non-central chi-square variable. The distribuiton of S_k^2, when the variances are not equal, must be found by a series expansion [3].

In short, by using (2), for a random sample of size n from a mixture of two normal distributions with common variance σ^2 we have:

(1) The distribution of $\bar{X}$ is a binomial mixture of $n+1$ normal distributions with common variance σ^2/n and means defined by (16).

(2) The distribution of nS^2/σ^2 is a binomial mixture of $n+1$ non-central chi-square variables $\chi^2(n-1,d_k)$ each with $n-1$ degrees of freedom and non-centrality parameters d_k defined by (17).

4. Distributions of Order Statistics

For the sake of simplicity, we just show how to find the distribution of $X_{(1)}$ the first order statistic for a mixture of two absolutely continuous distributions. The distributions of other order statistics can be obtained similarly.

First we find the distribution of $X_{(k1)}$ the first order statistic for $X_{k1}, X_{k2}, \cdots, X_{kn}$.

The event $x < X_{(k1)} \leq x + \Delta x$ may be realized as follows: In k ways, $x < X_{ki} \leq x + \Delta x$ for one X_{ki} with density $f_1(x)$ and $X_{ki} > x + \Delta x$ for the remaining $n-1$ of the X_{ki}'s; or in $n-k$ ways, $x < X_{ki} \leq x + \Delta x$ for one X_{ki} with density $f_2(x)$ and $X_{ki} > x + \Delta x$ for the remaining $n-1$ of the X_{ki}'s. Combining these, we have the probability of the above event. Dividing the probability by Δx and letting $\Delta x \to 0$, we obtain

$$f_{X_{(k1)}}(x) = kf_1(x)[1-F_1(x)]^{k-1}[1-F_2(x)]^{n-k} + (n-k)f_2(x)[1-F_1(x)]^{k}[1-F_2(x)]^{n-k-1}. \quad (18)$$

Now, using (2), we have the density of $X_{(1)}$.

EXAMPLE. Let (1) be a mixture of two exponential densities $f_j(x) = \alpha_j \exp(-\alpha_j x)$ for $x > 0$ and zero elsewhere with $\alpha_j > 0$, $j = 1,2$.

From (18), by simple calculation, we obtain

$$f_{X_{(k1)}}(x) = \beta_k \exp(-\beta_k x) \qquad x > 0$$

$$= 0 \qquad \text{elsewhere}$$

where, for $k = 0,1,\ldots,n$,

$$\beta_k = k\alpha_1 + (n-k)\alpha_2. \tag{19}$$

Now, by using (2), we have the following result:

The distribution of the first order statistic for a random sample of size n from a mixture of two exponential densities with parameters α_1 and α_2 is a binomial mixture of $n+1$ exponential densities with parameters defined by (19).

References

[1] Boes, D. C., "On the estimation of mixing distributions," *Annals of Mathematical Statistics*, vol. 37, (1966), 177-188.

[2] Cox, D. R., "Notes on the analysis of mixed frequency distributions," *British Journal of Mathematical and Statistical Psychology*, vol. 19, (1966), 39-47.

[3] Press, S. J., "Linear combinations of non-central chi-square variates," *Annals of Mathematical Statistics*, vol. 37, (1966), 480-487.

[4] Thomas, E. A. C., "Distributions free tests for mixed probability distributions," *Biometrika*, vol. 56, (1969), 475-484.