NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

TECHNICAL REPORT

Some Elementary Tests for Mixtures of Discrete Distribution

by

J. Tiago de Oliveira

Department of Industrial and Management Engineering
Columbia University, New York, New York

June 1, 1963

This research was supported by the Office of Naval Research under Contract NONR 266(04).

Some elementary tests for mixtures of discrete distribution

by

J. Tiago de Oliveira

The problems dealt with in this paper arose in the following context.

It is known, in some paleontological problems, that the distribution of some countable (meristic) characteristics of neighboring biological species, isolated in well-defined geographical areas, have some fixed (or stable) distribution.

In some cases, however, the analysis of the frequency polygon seemed to suggest the existence of a mixture of two populations and the test in 3) was devised for the purpose of deciding about the hypothesis of the mixture.

After, some of the results were extended.

1. Introduction

Let $P = (p_i)$ and $Q = (q_i)$ be discrete distributions with the same discrete support $\mathfrak{R} = (x_i)$, which is a subset of the real line, and let X be a random variable whose (discrete) distribution is the mixture $(\omega, 1 - \omega)$ $(0 \leq \omega \leq 1)$ of the distributions P and Q, called the components; some of the values p_i or q_i may be zero.

As it is very difficult to work out explicitly the best tests (and estimators) for the existence of the mixture (that is, the hypothesis that $\omega \neq 0, 1$) we will develop some (quick) tests with a lower efficiency (non-evaluated). The tests we will study are based in the first moments of some random variable, dependent on the random variable X as it is usual (Rider, 1961). The technique followed is connected with some previous results of Tiago de Oliveira (1960).

Let, then, g(x) be some function whose domain of definition contains the support R and such that it has mean values and variances relative to P and Q, denoted respectively by μ_P, μ_Q, σ_P and σ_Q . The new random variable $Y = g(X)$ has, then the following mean and variance

$$M(Y) = \omega\, \mu_P + (1 - \omega)\, \mu_Q$$

$$V(Y) = \omega\, \sigma^2{}_P + (1 - \omega)\, \sigma^2{}_Q + 2\, \omega\, (1 - \omega)\, (\mu_P - \mu_Q)^2$$

As it is well known (Cramer, 1946) if μ_k is the k-th moment of Y and

$M_k{}^{(n)} = \frac{1}{n} \sum_1^n y_i{}^k$ is the k-th sample moment,

$\sqrt{n}\, (M_k{}^{(n)} - \mu_k)$ is a random variable with an asymptotically normal distribution with zero mean and variance $\mu_{2n} - \mu n^2$ and the random pair $\left(\sqrt{n}\, (M_k{}^{(n)} - \mu_k),\ \sqrt{n}\, (M_1{}^{(n)} - \mu_1)\right)$ is asymptotically normally distributed with zero means, variances $\mu_{2k} - \mu_k{}^2$ and $\mu_{21} - \mu_1{}^2$ and covariance $\mu_{k+1} - \mu_k \mu_1$.

2. Components fully known: Let's suppose that the components P and Q are fully known and finite. The fact that

$$\sqrt{n} \quad \frac{M_1^{(n)} - (\omega\, \mu_P + (1-\omega)\mu_Q)}{\sqrt{\omega\sigma_P^2 + (1-\omega)\sigma_Q^2 + 2\omega(1-\omega)\mu_P\mu_Q)^2}}$$

is an asymptotically normal random variable with zero mean and unit variance suggests the use of

$$\omega_n^* = \frac{M_1^{(n)} - \mu_Q}{\mu_P - \mu_Q}$$

of an estimator of ω, because the values μ_P, μ_Q, σ_P^2 and σ_Q^2 are known as a consequence of the full knowledge of P and Q.

The asymptotic variance of ω^* is, then,

$$\frac{1}{n}\left[\frac{\omega\ \sigma_P^2 + (1-\omega)\ \sigma_Q^2}{(\mu_P - \mu_Q)^2} + 2\omega\ (1-\omega) \right]$$

and, consequently, to obtain the maximum discrimination we would choose the function $g(x)$ such that the values

$$y_i = g(x_i)$$

would lead to a minimum variance whatever may be the value of ω. This suggests the use of a function of such that

$$\sigma_P^2 = \sigma_Q^2 \text{ and } (\mu_P - \mu_Q)^2 \text{ a maximum.}$$

As we can make a linear transformation of the values y_i without altering the result, which is invariant for these transformations, we can impose then

$$\sigma_P^2 = 1$$

$$\sigma_Q^2 = 1$$

$$\mu_Q = 0$$

as condition equations and, then search the values y_i which maximize μ_P^2.

The technique of the Lagrange multipliers leads to the solution

$$y_i = \frac{\alpha(1-\gamma)p_i - \alpha q_i}{\beta q_i + \gamma p_i}$$

where α, β, γ are determined by the condition equations

$$\Sigma p_i y_i^2 - (\Sigma p_i y_i)^2 = 1$$

$$\Sigma q_i y_i = 0$$

$$\Sigma q_i y_i^2 = 1$$

It is easy to see that $\Sigma p_i y_i = \alpha$ and $\alpha^2 = \beta + \gamma$. As α will be determined from the condition equations except for the sign we can always choose α such that $\mu_P = \Sigma\, p_i y_i > 0$.

In the greater part of the problems the effective determination of the y_i by the solution of the 3 (condition) equations on α, β, γ is a difficult one. In these cases we will only choose the y_i such that

$$\sigma_P^2 = \sigma_Q^2 = 1 \quad \text{and}$$

$$\mu_P \neq \mu_Q = 0$$

which is always possible.

Once determined or chosen the values of y_i we can proceed to the test. As $\omega = 0$ or $\omega = 1$ we know that $\sqrt{n}\ (M_1^{(n)} - \mu_P)$ are asymptotically normal with zero mean and unit variance, the values $M_1^{(n)}$ tend to concentrate around $\mu_Q = 0$ or $\mu_P > 0$ according to $\omega = 0$ or $\omega = 1$. Let's fix then an asymptotic level of significance ϵ and take χ_ϵ such that

$$\int_{-\infty}^{\chi_\epsilon} - \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx = 1 - \epsilon$$

and act as follows:

1) if $M_1^{(n)} \le \frac{\chi_\varepsilon}{\sqrt{n}}$ we will accept the hypothesis $\omega = 0$

2) if $M_1^{(n)} \ge \mu_p - \frac{\chi_\varepsilon}{\sqrt{n}}$ we will accept the hypothesis $\omega = 1$

3) if $\frac{\chi_\varepsilon}{\sqrt{n}} < M_1^{(n)} < \mu_p - \frac{\chi_\varepsilon}{\sqrt{n}}$ we will accept the hypothesis

of the existence of a mixture; steps 1) and 2) led to the rejection of this hypothesis.

Let's denote by $P_1^{(n)}(\omega)$, $P_2^{(n)}(\omega)$ and $P_3^{(n)}(\omega)$ the probabilities of the decisions 1), 2) and 3) if the mixture is $(\omega, 1 - \omega)$. We have $\left(Z_n = \sqrt{n}\ \frac{M_1^{(n)} - \omega\mu_p}{\sqrt{1 + 2\omega(1-\omega)\mu_p^2}}\right)$

$$P_1^{(n)}(\omega) = \text{prob}\left\{ Z_n \le \frac{\chi_\varepsilon}{\sqrt{1 + 2\omega(1-\omega)\mu_p^2}} - \frac{\omega\mu_p\sqrt{n}}{\sqrt{1 + 2\omega(1-\omega)\mu_p^2}} \right\}$$

$$P_2^{(n)}(\omega) = \text{prob}\left\{ Z_n \ge \frac{(1-\omega)\mu_p\sqrt{n}}{\sqrt{1 + 2\omega(1-\omega)\mu_p^2}} - \frac{\chi_\varepsilon}{\sqrt{1 + 2\omega(1-\omega)\mu_p^2}} \right\}$$

and $P_3^{(n)}(\omega) = 1 - P_1^{(n)}(\omega) - P_2^{(n)}(\omega)$ and passing to the limit, we obtain

$$\lim_{n\Rightarrow\infty} P_1^{(n)}(0) = 1 - \epsilon, \ \lim_{n\Rightarrow\infty} P_1^{(n)}(\omega) = 0 \text{ if } \omega \neq 0 \ ;$$

$$\lim_{n\Rightarrow\infty} P_2^{(n)}(1) = 1 - \epsilon, \ \lim_{n\Rightarrow\infty} P_2^{(n)} \ (\omega) = 0 \text{ if } \omega \neq 1 \ ;$$

$$\lim_{n\Rightarrow\infty} P_3^{(n)}(0) = \lim_{n\Rightarrow\infty} P_3^{(n)}(1) = \epsilon, \ \lim_{n\Rightarrow\infty} P_3^{(n)}(\omega) = 1 \text{ if } w \neq 0, 1$$

The test leading to the solution of the trilemma is then a consistent one with the asymptotic level ϵ for the hypotheses $\omega = 0$ and $\omega = 1$.

It is evident that the same technique may be applied even if we don't have random variables but only random attributes, giving to each attribute whose pair of probabilities is (p_i, q_i) the value $y_i = g(x_i)$.

Implicitly we have shown that $\omega_n^* = \dfrac{M_1^{(n)}}{\mu_p}$ is an estimator of ω. Other estimators may also be obtained. The maximum likelihood and minimum χ^2 methods are unworkable. The minimum χ^{1^2}

method leads to the estimator

$$\frac{\sum_i \frac{q_i(q_i-p_i)}{N_i}}{\sum_i \frac{(q_i-p_i)^2}{N_i}}$$

where N_i are the number of observations with the value y_i or the attribute with probabilities (p_i, q_i). As some of the N_i may be zero we can use as a substitute the estimator

$$\omega_n^* = \frac{\sum_i \frac{q_i(q_i-p_i)}{N_i+1}}{\sum_i \frac{(q_i-p_i)^2}{N_i+1}}$$

As $\frac{N_i+1}{n} \xrightarrow{p} \omega p_i + (1-\omega)q_i$ we know (Cramer (1946)) that

$$\omega_n^* \xrightarrow{p} \frac{\sum_i \frac{q_i(q_i-p_i)}{\omega p_i+(1-\omega)q_i}}{\sum_i \frac{(q_i-p_i)^2}{\omega p_i+(1-\omega)q_i}} = \omega ,$$

showing that ω_n^* is an estimator (converging in probability) of ω. As the N_i/n are asymptotically normal the same happens to ω_n^* ; its mean value is asymptotically ω and its variance is a cumbersome one.

3. The translation mixture case: We shall suppose, now, that the discrete components (finite or enumerable) P and Q have the probabilities of the form

$$p_i = \varphi(i - \theta_P) \text{ and } q_i = \varphi(i - \theta_Q)$$

where the outcomes are equally spaced and so that for simplicity we suppose Ω to be the set of the integers; $\varphi(n)$ is a function such that

$$\Sigma\, \varphi(n) = 1 \quad \text{and } \varphi(n) \geq 0$$

and we will suppose, also, that

$$\mu = \Sigma\, n\, \varphi(n) \text{ and}$$

$$\sigma^2 = \Sigma\, n^2\, \varphi(n) - \mu^2$$

exist (and are known). In the sequel we will need also the existence of the 4th moment. Then we have

$$M(x) = \mu + \omega\, \theta_P + (1-\omega)\, \theta_Q$$

$$V(x) = \sigma^2 + 2\omega(1-\omega)\, (\theta_P - \theta_Q)^2 \quad .$$

As the variance increases with the mixture we will use as a test statistic the variance of the sample s_n^2 , the test will be one-sided. As the asymptotic distribution of

$$\sqrt{n}\, \frac{s_n^2 - V(x)}{\sqrt{(\beta_2+2)}\ V(x)}$$

is a normal one with zero mean and unit variance ($\beta_2 = \frac{\mu_4}{\mu_2^2} - 3$ denoting the kurtosis or excess coefficient) we will reject the hypothesis of mixture, on the asymptotic level of significance ϵ, if

$$\sqrt{n}\, \frac{s_n^2 - \sigma^2}{\sqrt{(\beta_2+2)_0}\ \sigma} \leq \chi_\epsilon$$

and accept it if

$$\sqrt{n}\, \frac{s_n^2 - \sigma^2}{\sqrt{(\beta_2+2)_0}\ \sigma} > \chi_\epsilon \text{ , } (\beta_2 + 2)_0 \text{ denoting}$$

denoting the kurtosis coefficient of $\varphi(n)$.

The probability of rejection of the hypothesis of mixture is then, for the mixture $(\omega, 1 - \omega)$,

$$P_n\ (\omega) = \text{prob}\left\{s_n^2 \leq \sigma^2 + \frac{\chi_\epsilon\ \sigma}{\sqrt{n}}\ \sqrt{(\beta_2+2)_0}\ \right\}$$

for $\omega = 0$ or $\omega = 1$ we have

$$\lim_{n \Rightarrow \infty} P_n(0) = \lim_{n \Rightarrow \infty} P_n(1) = 1 - \epsilon$$

and if $\omega \neq 0$ we obtain

$$\lim_{n \Rightarrow \infty} P_n(\omega) = 0$$

the test is then a consistent one.

In concrete cases, such as the biological problems of meristic characteristics, we can use the past experience to obtain a "scheme" of the values of $\varphi(n)$ and then apply this technique.

4. Poisson mixtures: Let now λ_P and λ_Q be the parameters of two Poisson components of a mixture; $\mathcal{R}$ is the set of non-negative integers. As $\mu_P = \sigma_P^2 = \lambda_P$ and $\mu_Q = \sigma_Q^2 = \lambda_Q$ we have

$$M(x) = \omega \lambda_P + (1-\omega) \lambda_Q$$

$$V(x) = \omega \lambda_P + (1-\omega) \lambda_Q + 2\omega(1-\omega)(\lambda_P - \lambda_Q)^2 \quad .$$

We have then, $V(x) = M(x)$ if $\omega = 0$ or $\omega = 1$ and $V(x) > M(x)$ if $\omega \neq 0, 1$. This suggests, as a test statistic, the difference between the sample variance and mean and, also, the use of a one-sided test.

Whatever may be ω we have

$$M(s_n^2 - M_n) = 2\omega(1-\omega)(\lambda_P - \lambda_Q)^2$$

$$V(s_n^2 - M_n) \sim \frac{b(\omega,\lambda_P,\lambda_Q)}{n}$$; for $\omega = 0$ or $\omega = 1$, the function $b(\omega, \lambda_P, \lambda_Q)$ reduces to $\overline{b}(\lambda) = 1 - 2\sqrt{\lambda} + 3\lambda$.

The asymptotic distribution of

$$\sqrt{n}\,\frac{s_n^2 - M_n}{\sqrt{b(\omega,\lambda_P,\lambda_Q)}}$$

is normal with zero mean and unit variance.

As for the hypothesis $\omega = 0$ or $\omega = 1$, b reduces to $\overline{b}(\lambda)$, and an estimator of λ_P (or λ_Q) is the sample mean M_n, $\overline{b}(M_n)$ is an estimator of $\overline{b}(\lambda)$ owing to the continuity of $\overline{b}$ (Cramer, 1946),and the asymptotic distribution of

$$\sqrt{n}\,\frac{s_n^2 - M_n}{\sqrt{\overline{b}\,(M_n)}}$$

is also normal with zero mean and unit variance.

The one-sided test is then the following:

reject the hypothesis of mixture if

$$s_n^2 - M_n \leq \frac{\chi_\epsilon}{\sqrt{n}} \sqrt{b\ (M_n)}$$

and accept it if

$$s_n^2 - M_n > \frac{\chi_\epsilon}{\sqrt{n}} \sqrt{b\ (M_n)} \quad .$$

The probability of rejection of the hypothesis of mixture is then

$$P_n(\omega) = \text{prob} \left\{ s_n^2 - M_n \leq \frac{\chi_\epsilon}{\sqrt{n}} \sqrt{b\ (M_n)} \right\}$$

and by the asymptotic normality referred we see that

$$\lim_{n \Rightarrow \infty} P_n\ (0) = \lim_{n \Rightarrow \infty} P_n(1) = 1 - \epsilon$$

and

$$\lim_{n \Rightarrow \infty} P_n(\omega) = 0 \qquad \omega \neq 0,\ 1.$$

which shows the consistency of the test.

References:

Cramer, H., Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J., 1946.

Rider, P.R., Estimating the parameters of mixed Poisson, binomial and Weibull distributions by the method of moments, Bull. Inst. Intl. Stat., 33rd Session, Paris, 1961, preprint No. 38.

Tiago de Oliveira, J., Invariant estimators of location and dispersion parameters, Rev. Fac. Ciencias Lisboa, 2nd ser., Vol. VIII, 1960, Lisboa.