ON A MEASURE OF DIVERGENCE BETWEEN TWO MULTINOMIAL POPULATIONS.

By A. BHATTACHARYYA
Statistical Laboratory, Calcutta.

Introduction

The problem of discrimination in Statistics is not new. It has repeatedly been stressed by Mahalanobis (1930) that mere "tests of significance" can not always answer all our questions and for this reason he introduced the idea of divergence of two populations. His "generalised distance" measures the divergence of two normal correlated multivariate populations with the same sets of variances and covariances. Bose and Roy (1938) found the distribution of D^2, "Mahalanobis's generalised distance" in its studentised form.

2. Definition of divergence

Let two multinomial populations be characterised by the two sets of probabilities $(\pi_1, \ldots \pi_k)$ and $(\pi'_1, \ldots \pi'_k)$. Then, as $\Sigma\pi = 1$ and $\Sigma\pi' = 1$, $(\sqrt{\pi_1}, \ldots \sqrt{\pi_k})$ and $(\sqrt{\pi'_1}, \ldots \sqrt{\pi'_k})$ may be considered to be the direction cosines of two st. lines through the origin in a k-dimensional space. The square of the angle between these two lines may be considered to be an appropriate measure of divergence between the two multinomial populations. Thus, if the measure of divergence be denoted by Δ^2 then

$$\cos\Delta = \sqrt{\pi_1\pi_1'} + \sqrt{\pi_2\pi_1'} + \quad . \quad . \quad . \quad + \sqrt{\pi_k\pi_k'} \qquad .. \quad (2.1)$$

This can be written as

$$4\sin^2\frac{\Delta}{2} = (\sqrt{\pi_1} - \sqrt{\pi_1'})^2 + (\sqrt{\pi_2} - \sqrt{\pi_2'})^2 + \; . \; . \; . \; + (\sqrt{\pi_k} - \sqrt{\pi_k'})^2 \qquad .. \quad (2.2)$$

From this latter expression it is evident that when $\pi_i = \pi_i'$ $(i = 1, 2, \ldots k)$ Δ vanishes. Also when Δ vanishes, we have conversely $\pi_i = \pi'_i$ $(i = 1, 2, \ldots k)$ that is, the two populations are identical.

3. Divergence between a sample and the population, its relation to Pearsonian χ^2 for large n

Let n be the size of a sample and (p_i) be the observed proportions taken from a population with proportions (π_i). Then, from the definition of divergence, the divergence β^2 of the sample from the population is given by

$$\cos\beta = \sum_{i=1}^{k}\sqrt{p_i\pi_i} \quad = \sum_{i=1}^{k}\pi_i\left[1 + \frac{p_i - \pi_i}{\pi_i}\right]^{\frac{1}{2}} \qquad .. \quad (3.1)$$

$$= \sum_{i=1}^{k}\pi_i\left[1 + \frac{p_i - \pi_i}{2\pi_i} - \frac{1}{8}\,\frac{(p_i - \pi_i)^2}{\pi_i^2} + \epsilon_i\right]$$

where $|\epsilon_i|$ is of the order of $\frac{1}{n^{3/2}}$

$$= \sum_{i=1}^{k}\pi_i + \frac{1}{2}\sum_{i=1}^{k}(p_i - \pi_i) - \frac{1}{8}\sum_{i=1}^{k}\frac{(p_i - \pi_i)^2}{\pi_i} + \sum_{i=1}^{k}\epsilon_i\pi_i$$

$$= 1 - \frac{1}{8}\frac{\chi^2}{n} + \epsilon \qquad .. \quad (3.2)$$

where $\chi^2=\Sigma(np_i-n\pi_i)^2/n\pi_i$ the Pearsonian χ^2 and ϵ is $O(1/n^{3/2})$. Thus

$$2\sin^2\frac{\beta}{2}=\frac{1}{2}\,\frac{\chi^2}{n} \qquad .. \quad (3.3)$$

approximately, when n is very large. Now we know that, for large values of k, $(\sqrt{2\chi^2}-\sqrt{2k-3})$ is distributed normally with unit standard deviation. So, χ^2 is of the order of k and χ^2/n is very small.

So, finally we get

$$\beta^2=\frac{1}{4}\,\frac{\chi^2}{n}\ \ ,i.e.,\ \beta=\frac{1}{2}\frac{\chi}{\sqrt{n}} \qquad .. \quad (3,4)$$

4. ASYMPTOTIC FORM OF THE MULTINOMIAL PROBABILITY WHEN n IS LARGE

We know that the multinomial probability

$$\frac{n\,!}{(np_1)\,!\,(np_2)\,!\ldots\ldots(np_k)!}\,\pi_1^{\ np_1}\ \pi_2^{\ np_2}\ldots\ldots\ldots\ldots\pi_k^{\ np_k}$$

$$\longrightarrow \text{Const}.\ e^{-\chi^2/2}\ .\frac{dp_1}{\sqrt{\pi_1}}.\frac{dp_2}{\sqrt{\pi_2}}\ldots\ldots\frac{dp_k}{\sqrt{\pi_k}}$$

$$\longrightarrow \text{Const}.e^{-2n\beta^2}\ d\sqrt{p_1}.\,d\sqrt{p_2}\ldots.d\sqrt{p_k}$$

for $\sqrt{p_1}=\sqrt{\pi_1}+(p_1-\pi_1)\left[\frac{d\sqrt{\pi_1}}{d\pi_1}\right]+\frac{(p_1-\pi_1)^2}{2!}\left[\frac{d^2\sqrt{\pi_1}}{d{\pi_1}^2}\right]+\ldots\ldots\ldots$

(by Taylor's theorem) and so $d\sqrt{p_1}=dp_1/\sqrt{\pi_1}$ approximately, for $\pi_1\neq o$ and $(p_1-\pi_1)$ is $O(1/n^{3/2})$

$$\longrightarrow \text{Const}.\ e^{-2n\beta^2}\ ds \qquad .. \quad (4.1)$$

where ds is an elementary volume given by $d\sqrt{p_1}\,d\sqrt{p_2}\ldots d\sqrt{p_k}$ subject to $\Sigma p=1$.

The above could be seen to be true also from the following considerations. We know that the multinomial distribution may be considered to be composed of a number of Poisson distributions.

Thus

$$\frac{n\,!}{(np_1)\,!\,(np_2)\,!\ldots\ldots(np_k)\,!}\pi_1^{\ np_1}\ \pi_2^{\ np_2}\ldots\ldots\pi_k^{\ np_k}$$

$$=\frac{n\,!}{e^{-n}\,.\,n^n}\,.\,\frac{e^{-n\pi_1}\,(n\pi_1)^{np_1}}{(np_1)!}\,.\,\frac{e^{-n\pi_2}\,(n\pi_2)^{np_2}}{(np_2)!}\ldots\ldots\frac{e^{-n\pi_k}\,(n\pi_k)^{np_k}}{(np_k)\,!}$$

We also know (Bartlet : 1936) that in Poisson's distribution if x is the observed and a the expected value, then $\sqrt{x}$ is distributed about $\sqrt{a}$ normally with standard deviation 1/2. The above expression becomes

$$\text{Const}.\ e^{-2n(\sqrt{p_1}-\sqrt{\pi_1})^2}\,d\sqrt{p_1}\,.\,e^{-2n(\sqrt{p_2}-\sqrt{\pi_2})^2}\,d\sqrt{p_2}\ldots.e^{-2n(\sqrt{p_k}-\sqrt{\pi_k})^2}\,d\sqrt{p_k}$$

$$=\ \ \text{Const}\ e^{-2n\sum\limits_{i=1}^{k}(\sqrt{p_i}-\sqrt{\pi_i})^2}\ d\sqrt{p_1}\,.\,d\sqrt{p_2}\ldots\ldots\ldots\ldots d\sqrt{p_k}$$

$$\sim\ \ \text{Const}.\ e^{-2n.\,4\sin^2\frac{\beta}{2}}\,ds$$

$$\sim\ \ \text{Const}.\ e^{-2n\beta^2}\,ds \qquad .. \quad (4.2)$$

he same as (4.1), on the supposition that β is very small.

5. Geometrical interpretation of the volume element

As the point $(\sqrt{p_1} \ldots \sqrt{p_k})$ lies on the unit hypersphere in the k-dimensional space, $d\sqrt{p_1} \ldots d\sqrt{p_k}$ is an elementary surface volume of a sphere round the sample point, say S. With this geometrical representation, the probability of a sample is completely representable on the unit hypersphere. Assuming the law to hold for all values of β, we may find the distribution of β. For keeping β fixed if we integrate out over the available space on the sphere, we get something like a zone of annular space. The volume of this space is proportional to $\beta^{k-2}d\beta$ (for the volume of the zone is proportional to $\sin^{k-2}\beta$ and so approximately to $\beta^{k-2}d\beta$). So, the distribution of β is

$$\text{Const.}\, e^{-2n\beta^2} \beta^{k-2}\, d\beta \qquad \ldots \quad (5.1)$$

which is also obvious from the relation between β^2 and χ^2.

6. Definition of divergence between two amples

Corresponding o Δ^2 in the populations, we have the statistic defined by

$$\cos D' = \sqrt{p_1 p_1'} + \sqrt{p_2 p_2'} + \ldots\ldots\ldots\ldots + \sqrt{p_k p_k'}$$

where n, n' are the sizes and (p_i) and (p'_i) are the sample proportion. This is not an unbiassed estimate of Δ^2.

7. Relation of D'^2 with Pearsonian χ^2 when $\Delta = o$ and n and n' are large and equal to one another

When the sample sizes are the same, namely n, we can express D'^2 in terms of Pearsonian χ^2. For

$$\cos D' = \sum_{i=1}^{k} \sqrt{p_i p_i'} = \tfrac{1}{2}\sum_{i=1}^{k} \sqrt{(p_i - p_i')^2 - (p_i - p_i')^2}$$

$$= \tfrac{1}{2} \sum_{i=1}^{k} (p_i + p_i') \left\{ 1 - \frac{(p_i - p_i')^2}{(p_i + p_i')^2} \right\}^{\frac{1}{2}}$$

$$= \tfrac{1}{2} \sum_{i=1}^{k} (p_i + p_i') \left\{ 1 - \tfrac{1}{2} \frac{(p_i - p_i')^2}{(p_i + p_i')^2} - \tfrac{1}{8} \cdot \frac{(p_i - p_i')^4}{(p_i + p_i')^4} - \ldots \right\}$$

by the binomial theorem, because $\dfrac{(p_i - p_i')}{(p_i + p_i')} < 1$,

$$= \tfrac{1}{2}\sum_{i=1}^{k} (p_i + p_i') - \tfrac{1}{4} \sum_{i=1}^{k} \frac{(p_i - p_i')^2}{(p_i + p_i')} - \tfrac{1}{16}\Sigma r_i$$

$$= 1 - \tfrac{1}{4}\sum_{i=1}^{k} \frac{(p_i - p_i')^2}{(p_i + p_i')} - r \qquad \ldots \quad (7.1)$$

Now, Δ^2 being zero, $(p_i - p'_i)$ is of the order of $1/\sqrt{n}$ and so r is of the order of $1/n^2$. So, when n is very large,

$$2 \sin^2 \frac{D'}{2} = \tfrac{1}{4}\sum_{i=1}^{k} \frac{(p_i - p_i')^2}{p_i + p_i'} \qquad \ldots \quad (7.2)$$

The sample values are np_i and np'_i $(i=1, 2, \dots k)$. So from the usual method of calculating χ^2 we get

$$\chi^2=\sum_{i=1}^{k}\frac{\left(np_i-\frac{np_i+np_i'}{2}\right)^2}{\frac{np_i+np_i'}{2}}+\sum_{i=1}^{k}\frac{\left(np_i'-\frac{np_i+np_i'}{2}\right)^2}{\frac{np_i+np_i'}{2}}$$

$$=\sum_{i=1}^{k}\frac{(np_i-np_i')^2}{np_i+np_i'} \qquad \dots \quad (7.3)$$

This χ^2, we know, is of the order of $2k$. So, χ^2/n is small and as from (7.2) $2 \sin^2 D'/2=\chi^2/4n$ we get $D'^2=\chi^2/2n$. From this relation it is evident that on the null hypothesis, the distribution of D'^2 follows the distribution of χ^2 for very large samples.

8. Distribution of D'^2 when $\Delta^2\neq 0$

Let there be two samples with sizes n and n' and proportions (p_i) and (p'_i) taken respectively from the two populations characterised by (π_i) and (π'_i). Let β^2 and β'^2 be the divergences (i.e. square of the angular distances) of the two samples from their respective populations. Then the joint distribution of the two samples is

$$\text{Const.}\, e^{-2\beta n^2} ds \,.\, e^{-2n'\beta'^2} ds' \qquad \dots \quad (8.1)$$

where ds and ds' are elementary surface volumes, on the unit hypersphere, about the sample points. Let P and P' be the population points (i,e, where the population lines meet the unit sphere), S and S' the sample points (with the same convention as the population points). Then Δ is represented by P P', D' by SS', and β and β' by S P and S'P'. From the distributions of β and β' we know, when n and n' are very large, that these tend to become very small. That is, the probability of a large β is so small that, for all practical purposes, β may be regarded to lie within a narrow zone about the arc PP'. Similar is the case for β'. Outside this zone the probability of getting samples is negligible. This narrow zone may be conceived to be a portion of a cylinder (that is, the projection of this zone on the enveloping cylinder osculating th . unit sphere along the curve PP' may be regarded to be coincident with the zone itself). We may develope out this cylindrical zone into a flat of $(k-1)$ dimensions where the geodesics β, β', Δ and D' all become right lines and ds and ds' become elementary areas (i,e. volumes) in this flat of $(k-1)$ dimensions.

With the help of the above geometrical device we may proceed to find the distribution of D' which now represents a rectilinear length. In the $(k-1)$ flat space let the coordinates of the points P, P', S and S' be (a_i), (a'_i), (x_i) and (x'_i), respectively $(i=1, 2, \dots k-1)$. The joint probability of the two samples is

$$\text{const.}\, e^{-2n\beta^2-2n'\beta'^2} ds\, ds'$$

$$=\ \text{const.}\, e^{-2n\sum_{i=1}^{k-1}(x_i-a_i)^2\ -2n'\sum_{i=1}^{k-1}(x_i'-a_i')^2} \prod_{i=1}^{k-1} dx_i \prod_{i=1}^{k-1} dx_i'$$

From this it is easy to deduce the distribution of $(x_i - x'_i - a_i + a'_i)$ in the form

$$\text{Const.}\ e^{-\frac{2n\,n'}{n+n'}\left\{\sum_{i=1}^{k-1}(x_i - x_i' - \overline{a_i - a_i'})^2\right\}}\ \prod_{i=1}^{k-1} d\,(x_i - x_i')$$

Or, which is the same as

$$\text{Const.}\ e^{-\frac{2nn'}{n+n'}\left\{\sum_{i=1}^{k-1}(x_i - x'_i)^2 - 2\sum_{i=1}^{k-1}(x_i - x_i')(a_i - a_i') + \sum_{i=1}^{k-1}(a_i - a_i')^2\right\}}\prod_{i=1}^{k-1}(x_i - x_i')$$

From this by adopting the method employed by Mr. R. C. Bose for getting the distribution of the classical D² (Bose 1936) we get from the above

$$\text{Const.}\ e^{--\frac{2nn'}{n+n'}\left\{D'^2 - 2D'\Delta\cos\theta + \Delta^2\right\}}(D'\sin\theta)^{k-3}D'dD'd\theta$$

Where $D'^2 = \sum_{i=1}^{k-1}(x_i - x_i')^2$, $\Delta^2 = \sum_{i=1}^{k-1}(a_i - a'^i)^2$ and $D'\Delta\cos\theta = \sum_{i=1}^{k-1}(x_i - x_i')(a_i - a_i')$

Integrating over θ from 0 to 2π we get the distribution of D as

$$\text{Const.}\ D'^{(k-1)/2}\ e^{-\frac{2nn'}{n+n'}(D'^2 + \Delta^2)}\ I_{\frac{k-3}{2}}\left(\frac{4nn'}{n+n'}\ D'\Delta\right)$$

where I is a Bessel function of purely imaginary argum.nt.

9. Approximation of Δ and D' when Δ is very small

We know

$$\cos\Delta = \Sigma\sqrt{\pi_i\pi'_i} = \tfrac{1}{2}\ \Sigma(\pi_i + \pi'_i)\left[\ 1 - \left(\frac{\pi_i - \pi'_i}{\pi_i + \pi_i'}\right)^2\right]^{\frac{1}{2}}$$

Repla ing cosΔ by $1 - \Delta^2/2$ to a first approximation we have $\Delta^2 = \Sigma(\pi_i - \pi_i')^2/(\pi_i + \pi'_i)$. Similarly $D'^2 = \Sigma(p_i - p'_i)^2/(p_i + p'_i)$. These wer suggested by Prof. Mahalanobis as a measure of divergence between two multinomial populations and samples therefrom.

10. Special case of binomial populations

Let there be two binomial populations efined by (π_1, π_2) and (π_1', π'_2) from which two samples of sizes n and n' with pro.ortions (p_1, p_2) and (p'_1, p'_2) are taken respectively. Then we know that $n(p_1 - \pi_1)\sqrt{n\pi_1(1-\pi_1)}$ is distributed normally about zero with unit standard deviation subject to the condition that n is sufficiently large and both π_1 and k_1 not far differe.t from 1/2. Now, un.er t e same assumptions, we have $(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{\pi_1})$ $= (p_1 - \pi_1)/2\sqrt{\pi_1(\ -\pi_1)}$ approximately (by expanding in Taylor's series). Therefore $2\sqrt{n}\,(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{\pi_1})$ is distributed normally about zero with unit standard deviation. Similarly, under the same set of conditions, $2\sqrt{n'}(\sin^{-1}\sqrt{p'_1} - \sin^{-1}\sqrt{\pi'_1})$ is distributed normally with unit standard deviation. So, from these two it easily follows that

$$2\sqrt{\frac{n\,n'}{n+n'}}\left\{\left(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_1'}\right) - \left(\sin^{-1}\sqrt{\pi_1} - \sin^{-1}\sqrt{\pi_1'}\right)\right\}$$

is distributed with unit standard deviation. Here the population divergence is clearly $\sin^{-1}\sqrt{\pi_1}-\sin^{-1}\sqrt{\pi'_1}$ and the sample estimate is $\sin^{-1}\sqrt{p_1}-\sin^{-1}\sqrt{p'_1}$. Clearly this is a particular case of he general problem which degenerates into it when $k=2$.

References

1. Bartlett, M. S. (1936). "The Square-root Transformation in the Analysis of Variance", *J. R. S. S. (Supplement)*, 3. Pp. 68-78.

2. Bose, R. C. On tne exact distribution and moment co-efficients of the D^2 statistic, *Sankhya* 1(2) Pp. 143-54.

3. Bose, R. C. and Roy, S. N. (1938). "The Distribution of the Studentised D^2-Statistics", *Sankhya*, 4(1). Pp. 19-38.

4. Cochran, W. G. (1940). "Analysis of variance when experimental errors follow the Poisson or Binomial Laws". *Annals of Math. Stat.* 11(3). Pp. 335-347.

5. Mahalanobis, P. C. (1930). "On the tests and measures of group divergences". *Jour. and Proc. of the Asiatic Society of Bengal.* (New series), 26, No. 4, Pp. 541-588.

6. ——— "On the generalised distance in statistics". *Proc. of the Nat. Inst. of Sciences of India*, 2(1), 1936, Pp. 49-55.

Paper received 3 July 1941