ON OPTIMUM SELECTIONS FROM MULTIVARIATE POPULATIONS

By DES RAJ

Indian Statistical Institute, Calcutta

1. The problem of selection is frequent in practice. A group of individuals has to be selected on the basis of measurements on p characters $x_1, x_2, \ldots, x_p$, such that in the selected universe another character y, which cannot be directly measured at the time of selection, has some desired properties. For example, an educational institution wants to select some candidates on the basis of scores in a preliminary examination so that the probability that the score in the final examination exceeds a given value is maximum. A wholesale dealer in manufactured goods wants to select some articles from a given lot so that the average of a particular characteristic in the selected lot has a given value and the variance of the characteristic is minimum. A crop physiologist may like to select from a large number of varieties of a certain crop such that the average yield per acre is maximised. The object of this note is to give some theorems regarding optimum rules of selection in certain cases.

2. Let there be a multivariate population

$$f(y, x_1, x_2, \ldots, x_p). \qquad \ldots \quad (1)$$

Let the frequency function of y for fixed $x = (x_1, x_2, \ldots, x_p)$ be given by

$$\phi(y \mid x),$$

so that

$$f(y, x_1, x_2, \ldots, x_p) = \phi(y \mid x) f_1(x). \qquad \ldots \quad (2)$$

We shall denote by $\eta(x)$ the regression of y on x.

Our object is to find a suitable rule of selection based on x such that in the selected portion of the universe, the variate y has some desired properties.

Theorem 1 : *A selection such that in the selected population the probability that y is exceeded by a given value y_0 is α and the fraction retained is maximum is given by*

$$\int_{-\infty}^{y_0} \phi(y \mid x) dy \lesseqgtr \lambda \qquad \ldots \quad (3)$$

according as $\alpha \lesseqgtr P(y < y_0)$ in the original population.

Proof : We have to find a suitable region R in the space of x such that

$$\alpha = \int_{-\infty}^{y_0} \int_R f(y,x) dy dx \Big/ \int_{-\infty}^{\infty} \int_R f(y, x) dy\, dx \qquad \ldots \quad (4)$$

and

$$\int_{-\infty}^{\infty} \int_R f(y, x) dy dx \text{ is maximised.}$$

This means that R is to be selected such that for

$$\int_R \left[\int_{-\infty}^{y_0} \phi(y\,|\,x)dy-\alpha\right] f_1(x)dx = 0,$$

$$\int_R f_1(x)dx \text{ is a maximum.}$$

Using Neyman and Pearson's lemma, the region R is given by

$$g(x) = \int_{-\infty}^{y_0} \phi(y\,|\,x)dy \leqslant a \qquad \dots \quad (5.1)$$

or

$$g(x) \geqslant b \qquad \dots \quad (5.2)$$

where a and b are to be chosen so as to satisfy (4). We note that for the region given by (5.1), the right-hand side of (4) is simply

$$E[g(x)\,|\,g(x) \leqslant a]. \qquad \dots \quad (6)$$

Now (6) is a monotonically increasing function of a and when $a \to \infty$, (6) gives the probability that y is exceeded by y_0 in the original population. Thus for this region $\alpha < P(y < y_0)$ in the original population. Similarly, it can be shown that for the region given by (5.2), $\alpha > P(y < y_0)$ in the original population. It is of interest to note that for the multinormal population the region of selection would be given by $\eta(x) \gtrless c$ so that truncation by means of the linear regression of y on the x's is optimum in this case. This result was obtained by Birnbaum and Chapman (1950).

Theorem 2 : *A selection such that in the selected universe the mean value of y is m_0 and the fraction retained is maximum is given by*

$$\eta(x) \gtrless \lambda \qquad \dots \quad (7)$$

according as $m_0 \gtrless$ mean value of y in the original population.

Proof : R is to be chosen such that

$$m_0 = \int_{-\infty}^{\infty}\int_R yf(y,x)dy\,dx \Big/ \int_{-\infty}^{\infty}\int_R f(y,x)dy\,dx \qquad \dots \quad (8)$$

and

$$\int_{-\infty}^{\infty}\int_R f(y,x)dy\,dx \qquad \text{is a maximum.}$$

This means that

$$\int_R [\eta(x)-m_0]f_1(x)dx = 0$$

and

$$\int_R f_1(x)dx \qquad \text{is maximised.}$$

The required region is obviously given by

$$\eta(x) \lesseqgtr \lambda$$

where λ satisfies (8). As before, we see that for the region $\eta(x) \leqslant \lambda$, the right-hand side of (8) is $E[\eta(x) \mid \eta(x) \leqslant \lambda]$, which is an increasing function of λ and is the mean value of y in the original population when $\lambda \to \infty$, so that for this region $m_0 <$ the mean value of y in the original population. Similar remarks apply to the region $\eta(x) \geqslant \lambda$. We note that for the multinormal population the required truncation is linear—a result obtained earlier by Birnbaum and Chapman (1950).

Theorem 3 : *A selection such that the frequency of selection is α and the probability that y exceeds y_0 in the selected population is maximum, is given by*

$$\int_{y_0}^{\infty} \phi(y \mid x) dy \geqslant \lambda. \qquad \ldots \quad (9)$$

Proof : We have to choose R such that

$$\int_R f_1(x) dx = \alpha \qquad \ldots \quad (10)$$

and

$$\frac{1}{\alpha} \int_R \left[\int_{y_0}^{\infty} \phi(y \mid x) dy \right] f_1(x) dx \text{ is maximised.}$$

Obviously R is given by (9) where λ is chosen to satisfy (10). In case the population is multivariate normal the region is given by $\eta(x) \geqslant \lambda$ so that in this case the truncation is linear. It may be of interest to note that, as shown by Cochran (1951), the truncation $\eta(x) \geqslant \lambda$, where λ is determined from (10), maximises the mean value of y for a given frequency of selection α, for the general class of populations given by (1).

Theorem 4 : *A selection such that the frequency of selection is α, the mean value of y in the selected portion is m_0 and the variance of y in the selected portion is minimum, is given by*

$$\int_{-\infty}^{\infty} (y - m_0)^2 \phi(y \mid x) dy \leqslant \lambda_1 + \lambda_2 \eta(x) \qquad \ldots \quad (11)$$

where λ_1 and λ_2 satisfy

$$\int_R f_1(x) dx = \alpha,$$

$$\int_R [\eta(x) - m_0] f_1(x) dx = 0.$$

The proof follows from the Neyman-Pearson lemma.

In the case of a multinormal population, the selection would be based on the region

$$a < \eta(x) \leqslant b \qquad \dots \quad (12)$$

where a and b are to be chosen such that

$$\left.\begin{aligned} \alpha &= \frac{1}{\sqrt{2\pi}} \int_{a'}^{b'} e^{-t^2/2} dt, \\ \alpha(m_0 - \bar{\eta}) &= \frac{1}{\sqrt{2\pi}}\, \sigma_\eta [e^{-a'^2/2} - e^{-b'^2/2}], \\ \text{where} \qquad a' &= (a - \bar{\eta})/\sigma_\eta, \quad b' = (b - \bar{\eta})/\sigma_\eta. \end{aligned}\right\} \qquad \dots \quad (13)$$

3. It may be noted here that Birnbaum (1950) has considered linear truncations of the multinormal population such that for a fixed mean of y after truncation, the variance of y after truncation is a minimum. In this case, the region R has to be selected such that

$$\int_R [\eta(x) - m_0] f_1(x) dx = 0,$$

and

$$\int_R \left[\int_{-\infty}^{\infty} [y - m_0]^2 \phi(y \mid x) dy \right] f_1(x) dx \Big/ \int_R f_1(x) dx \qquad \dots \quad (14)$$

is a minimum.

Keeping

$$\int_R f_1(x) dx$$

constant, we find that the region of selection belongs to the class

$$a \leqslant \eta(x) \leqslant b.$$

In this class the region minimising (14) is given by

$$\eta(x) - m_0 = 0.$$

We thus notice that some control over the frequency of selection is necessary to get non-trivial truncations. The case when the frequency of selection is assigned beforehand is covered by Theorem 4.

I am grateful to Dr. C. R. Rao under whose guidance this note was written.

References

Birnbaum, Z. W. (1950) : Effect of linear truncation on a multinormal population. *Ann. Math. Stat.*, **21**, 272–279.

Birnbaum, Z. W. and Chapman, D. G. (1950) : On optimum selections from multinormal populations. *Ann. Math. Stat.*, **21**, 443–447.

Cochran, W. G. (1951) : Improvement by means of selection. *Proc. Second Berkeley Symposium on Math. Stat.*, 449–470.

Paper received: April, 1953.