EFFECT OF NON-NORMALITY ON RESPONSE TO SELECTION IN LARGE POPULATIONS

By

M. Gopinath Rao and J.P. Jain

I.A.S.R.I., New Delhi-12

(Received : July, 1979)

1. Introduction

Very rarely, realised selection responses agree with those expected from the selection practised (Dickerson, [3], [4]; Falconer, [5]; Clayton, Knight, Morris and Robertson, [2]). One of the causes of discrepancy could be the assumption of normality of criterion of selection in computing selection intensity when in fact it follows a non-normal distribution. That quite a few quantitative traits may depart from the normal form has been borne out by empirical studies made by various investigators (Pearl and Miner, [11]; Gowen, [7]; Tocher, [16]; Om Parkash & Mahajan, [10]; and Malhotra, [9]). These studies, in addition, show that Pearsonian system of curves of Type I and Type III provide adequate representation to many types of data in the field of livestock and poultry breeding. In a few cases exponential and lognormal distributions also fitted well (Kapteyn, [8]; Quesenberry *et al.*, [14]). It is the purpose of this paper firstly to derive the expressions of selection intensity for Pearson's Type I distribution including its derivative, namely beta-distribution, Type III including its derived distributions, namely gamma, and exponential, and for lognormal distribution as applicable to large populations, and then to study the percentage discrepancy in response to selection predicted on the assumption of normality relative to these non-normal distributions.

2. Theoretical Procedure

In large populations in which the basis of selection, x is distributed according to density function $f(x)$ with mean μ and variance σ^2, the largest phenotypic selection differential, $(\mu_s - \mu)$ is achieved through truncated selection, *i.e.*, when all individuals with values of

$x>c$, a point of truncation corresponding to a fraction p are chosen, where μ_s is the average of selected individuals obtained as

$$\mu_s=\frac{1}{p}\int_c^\infty x f(x)\,dx,\quad p=\int_c^\infty f(x)\,dx \qquad ...(1)$$

The intensity of selection i, which is the standardized phenotypic selection differential is therefore

$$i=(\mu_s-\mu)/\sigma. \qquad ...(2)$$

When the distribution is normal it is well known that $i=\frac{z}{p}$, where z is the height of the normal ordinate at the cut-off point c.

If i and i^* denote the intensity of selection corresponding to the proportion, p saved from a normal population and that from a particular non-normal population, the predicted response to selection in two situations will be $\Delta G=i\ r_{gx}\ \sigma_g$ and $\Delta G^*=i^*\ r_{gx}\ \sigma_g$, where r_{gx} is the accuracy of selection and σ_g is the genetic standard deviation. The proportionate discrepancy, D_i in the predicted response to selection on the assumption of normality relative to that based on actual distribution, therefore, is

$$D_i=(\Delta G-\Delta G^*)/\Delta G^*=i/i^*-1=R^*-1,\ R^*=i/i^* \qquad ...(3)$$

When $R^*>1$, the progress is over-estimated; when it is less than 1, the progress is underestimated, and when $R^*=0$, the progress remains unaffected.

3. Intensity of Selection in Non-Normal Large Populations

Following the approach outlined in the previous section, the expressions for intensity of selection were derived for Pearson's Type I, beta, Pearson's Type III, gamma, exponential and log-normal distributions and are summarised in Table 1 along with the form of density functions used.

3.1. Pearson's Type I Distribution

When quantitative character x follows this distribution, the proportion of selected individuals, p corresponding to the truncation point c is

$$p=\int_c^{a_2} y_0\left[1+\frac{x}{a_1}\right]^{n_1}\left[1-\frac{x}{a_2}\right]^{n_2}dx$$

TABLE 1

Expressions of selection intensity for different non-normal large populations

Distribution	*Form of density function used*	*First two moments*	*Intensity of selection (i)*
1	2	3	4
Pearson's Type	$f(x)=y_0\left[1+\frac{x}{a_1}\right]^{n_1}\left[1-\frac{x}{a_2}\right]$, $-a_1\leqslant x_1\leqslant a_2$ where $\frac{n_1+1}{a_1}=\frac{n_2+1}{a_2}$; $y_0=\frac{a_1^{n_1}a_2^{n_2}}{(a_1+a_2)^{n_1+n_2}B(n_1+1,n_2+1)}$	$\mu=0$ $\sigma^2=\frac{(a_1+a_2)^2(n_1+1)(n_2+1)}{(n_1+n_2+2)^2(n_1+n_2+3)}$	$i=\frac{(n_1+1)(n_1+n_2+3)}{p\sqrt{n_2+1}}\times[(1-p)-I_d(n_1+2,n_2+1)]$ where $d=\frac{a_1+c}{a_1+a_2}$; $p=1-I_d(n_1+1,n_2+1)$
Beta	$f(x)=\frac{x^{m_1-1}(1-x)^{m_2-1}}{B(m_1,m_2)}$, $0<x<1$, $m_1, m_2>0$	$\mu=\frac{m_1}{(m_1+m_2)}$ $\sigma^2=\frac{(m_1+m_2)}{(m_1+m_2)^2(m_1+m_2+1)}$	$i=\sqrt{\frac{m_1(m_1+m_2+1)}{m_2}}\times\frac{1}{p}[(1-p)-I_c(m_1+1,m_2)]$ where $p=1-I_c(m_1,m_2)$
Pearson's Type III	$f(x)=k_0\left(1+\frac{x}{a}\right)^m e^{-\left(\frac{m+1}{a}\right)x}$, $-a<x<\infty$ where $k_0=\frac{(m+1)^{m+1}}{a\,e^{m+1}\lfloor(m+1)}$	$\mu=0$ $\sigma^2=a^2/(m+1)$	$i=\frac{\sqrt{m+1}}{p}\left[(1-p)-I\left(\frac{b}{\sqrt{m+2}},m+1\right)\right]$ where $b=\frac{(m+1)(a+c)}{a}$, $p=1-I\left(\frac{b}{\sqrt{m+1}},m\right)$

Gamma	$f(x)=\frac{e^{-x}x^{k-1}}{\lceil k}, 0<x<\infty$	$\mu=k$ $\sigma^2=k$	$i=\frac{\sqrt{k}}{p}\left[(1-p)-I\left(\frac{c}{\sqrt{k+1}}, k\right)\right]$ where $p=1-I\left(\frac{c}{\sqrt{k}}, k-1\right)$
Exponential	$f(x)=\theta e^{-\theta x}, \theta>0, x\geqslant 0$	$\mu=1/\theta, \sigma^2=1/\theta^2$	$i=-\log_e p$
Log-normal	$f(x)=\frac{1}{\sqrt{2\pi}x}\exp\left[-\frac{1}{2}(\log x)^2\right], 0<x<\infty$	$\mu=e^{\frac{1}{2}}$, $\sigma^2=e(e-1)$	$i=\frac{1}{p\sqrt{e-1}}\left[(\frac{1}{2}-p)-\phi(\log c-1)\right]$ where $p=\frac{1}{\sqrt{2\pi}}\int_{\log c}^{\infty}\exp(-\frac{1}{2}w^2)\,dw,$ $\phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-x^2/2}\,dx$

p is the proportion saved, and c is the point of truncation corresponding to p.

$$=\int_d^1 \frac{z^{n_1}(1-z)^{n_2}dz}{B(n_1+1, n_2+1)},$$

where $z=\dfrac{x+a_1}{a_1+a_2}$,

and $d=\dfrac{c+a_1}{a_1+a_2}$

$$=1-I_d\,(n_1+1,\ n_2+1) \qquad \ldots(4)$$

where $I_x(p,q)=B_x(p,q)/B(p,q)$ in which $B_x(p,q,)$ is the incomplete beta function tabulated by Karl Pearson [12] for different values of $x=[.00(.01)1]$ and $p,\ q=[0.5(0.5)\ 11.0(1)50]$, with $p \geqslant q$.

The mean value of the selected group (μ_s) which is

$$\frac{y_o}{p}\int_c^{a_2} x\left[1+\frac{x}{a_1}\right]^{n_1}\left[1-\frac{x}{a_2}\right]^{n_2}dx$$

can be shown equal to

$$\mu_s=\frac{(a_1+a_2)(n_1+1)}{p(n_1+n_2+2)}\ [(1-p)-I_d(n_1+2, n_2+1)]$$

Hence, using relation (2), the selection intensity is

$$i=\frac{(n_1+1)(n_1+n_2+3)}{p\sqrt{n_2+1}}\,[(1-p)-I_d\,(n_1+2,\ n_2+1)] \qquad \ldots(5)$$

where $d=(c+a_1)/(a_1+a_2)$ and c is related to p by expression (4).

From the selection intensities computed (Table 2) for different values of p and for different sets of parameters covering their entire range: high/low values of both n_1 and n_2 and high value of one and low value of the other, it is seen that for curves with long tail toward low merit, *i.e.*, when $n_1>n_2$, the selection intensity is more for higher values of p (mild selection) than if the distribution were truly normal. For curves with long tail toward high merit ($n_1<n_2$), the reverse is the case.

3.2. Beta-distribution. If we let $z=(x+a_1)/(a_1+a_2)$, $n_1=m_1-1$ and $n_2=m_2-1$, the Pearson's Type I distribution reduces to beta distribution, and accordingly the expression of intensity of selection shown in Table 1 can be obtained directly from (5).

3.3. Pearson's Type III Dsitribution. For this distribution with the density function as shown in Table 1, the proportion of selected

TABLE 2

Selection intensities for different values of p corresponding to different sets of values of the parameters n_1 and n_2 of Pearson's Type I distribution along with those for normal distribution

Proporation selected p	*Intensity of selection for Pearson Type I distribution*								*Intensity of selection for normal distribution*
	$n_1=1$ $n_2=1$	$n_1=2$ $n_2=1$	$n_1=1$ $n_2=2$	$n_1=48$ $n_2=1$	$n_1=1$ $n_2=48$	$n_1=48$ $n_2=48$	$n_1=3.0$* $n_2=1.5$	$n_1=4.5$* $n_2=1.5$	
0.1	1.663	1.537	1.824	1.185	2.148	1.743	1.552	1.481	1.755
0.2	1.405	1.321	1.482	1.039	1.578	1.396	1.322	1.273	1.399
0.3	1.196	1.143	1.228	0.948	1.219	1.156	1.137	1.104	1.159
0.4	1.010	0.981	1.016	0.843	0.966	0.964	0.938	0.953	0.966
0.5	0.838	0.827	0.827	0.740	0.740	0.800	0.820	0.810	0.798
0.6	0.674	0.677	0.654	0.644	0.562	0.643	0.672	0.670	0.644
0.7	0.513	0.526	0.490	0.522	0.406	0.495	0.524	0.528	0.497
0.8	0.351	0.370	0.330	0.394	0.260	0.349	0.370	0.378	0.349
0.9	0.185	0.203	0.171	0.239	0.132	0.194	0.204	0.214	0.195

*Parameteric values close to those obtained by Malhotra (1973) for rate of lay in poultry.

individuals, p corresponding to the truncation point c is connected by the relation

$$p = k_o \int_c^{\infty} \left(1+\frac{x}{a}\right)^m e^{-\frac{(m+1)x}{a}} dx$$

$$= k_o \frac{e^{m+1}}{a^m} \int_{a+c}^{\infty} z^m e^{\frac{-(m+1)z}{a}} dz, \text{ where } x+a=z$$

$$= 1 - \int_o^b \frac{y^m e^{-y}}{\lfloor (m+1)} \text{ dy, on substituting } (m+1)z=ay$$

where $b=(m+1)(a+c)/a$

$$=1-I(u, m) \qquad \ldots(6)$$

where $u = b/\sqrt{(m+1)}$ and $I(u,m) = \int_0^{\frac{u}{\lfloor m+1}} \frac{e^{-x} x^m}{\lfloor (m+1)} dx$

which has been extensively tabulated by Pearson [13] for different values of u at intervals of 0.1 and $m=-1(0.05)\ 0(0.1)\ 5(0.2)\ 50$.

Likewise, the mean of the selected individuals can be shown as

$$\mu_s = \frac{a}{p}\left[(1-p) - I\left(\frac{}{\sqrt{m+2}}, m+1\right)\right]$$

Hence, $$i = \frac{\sqrt{m+1}}{p}\left[(1-p) - I\left(\frac{b}{\sqrt{m+2}}, m+1\right)\right] \ldots(7)$$

Table 3 gives the intensities of selection for different values of p and m. The selection intensity is more for heavy culling and less for mild culling than if the distribution were truly normal, and the departure is pronounced for small values of m.

3.4. Gamma and exponential distributions. Since these distributions are special cases of Pearson's Type III distribution, the selection intensities for different values of p for gamma distribution can be obtained from Table 3 by taking $m=k-1$ and those for exponential distribution correspond to the value of $m=0$ or can be obtained directly from the relation $i=-\log_e p$ (Burrows [1]).

TABLE 3

Intensities of selection for different values of p and m for Pearson's Type III distribution along with those for normal distribution

Proportion selected p	*Intensity of selection for Pearson's Type III distribution*						*Intensities of selection for normal distribution*
	m=4	*m=5*	*m=10*	*m=15*	*m=20*	*m=48**	
0.1	2.054	2.029	1.964	1.932	1.920	1.866	1.755
0.2	1.542	1.529	1.504	1.490	1.481	1.454	1.399
0.3	1.216	1.215	1.203	1.198	1.194	1.185	1.159
0.4	0.978	0.978	0.977	0.977	0.976	0.976	0.966
0.5	0.776	0.779	0.787	0.791	0.792	0.795	0.798
0.6	0.602	0.607	0.619	0.624	0.627	0.633	0.644
0.7	0.445	0.451	0.465	0.471	0.474	0.483	0.497
0.8	0.298	0.303	0.316	0.322	0.326	0.336	0.349
0.9	0.155	0.158	0.168	0.174	0.176	0.184	0.195

*Parametric value obtained by Pearl and Miner (1919) for lactation yield and fat content in Ayrshire cows.

3.5. **Log-normal distribution.** In this case p and c are related as

$$p=\int_c^\infty \frac{1}{\sqrt{2\pi}\,x} \exp\left[-\tfrac{1}{2}(\log x)^2\right] dx$$

on taking $\log x=w$

$$=\int_{\log c}^\infty \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}w^2\right] dw \qquad ...(8)$$

The average of the selected parents is

$$\mu_s=\frac{1}{p}\int_c^\infty \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(\log x)^2\right] dx$$

$$=\frac{e^{\frac{1}{2}}}{p}\left[\tfrac{1}{2}-\phi(\log c-1)\right]$$

where $$\phi(x)=\int_0^x \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\, dx$$

is the cumulative normal distribution function tabulated by Sheppard [15] and are reproduced in Fisher & Yates [6]. The intensity of selection i, therefore, is

$$i=\frac{1}{p\sqrt{e-1}}\left[(\tfrac{1}{2}-p)-\phi\,(\log c-1)\right] \qquad ...(9)$$

The selection intensities obtained for different values of $p=0.1$ (0.1) 0.9 are 2.21, 1.28, 0.97, 0.71, 0.52, 0.38, 0.26, 0.16 and 0.08, respectively, The selection intensity is more for lower values of p and *vice-versa* than if the distribution were truly normal.

4. Results

Table 4 gives the percentage discrepancy in response to selection relative to beta, gamma, exponential and log-normal distributions. For simplicity, instead of general Type I and Type III distributions their derivatives were considered. This would not affect the general conclusions drawn.

In case the actual distribution is of beta form with $m_1=m_2$, the response under normality is over-estimated to the same extent for extremely heavy and extremely low cullings.

TABLE 4

Percentage discrepancy in response to selection relative to different non-normal large populations

Proportion selected p	*Percentage discrepancy relative to*											
	Beta							*Gamma*			*Exponential*	*Log-normal*
	$m_1=2$ $m_2=2$	$m_1=48$ $m_2=48$	$m_1=3$ $m_2=2$	$m_1=49$ $m_2=2$	$m_1=2$ $m_2=3$	$m_1=2$ $m_2=49$	$m_1=5.5$ $m_2=2.5$	$k=5$	$k=16$	$k=49$		
0.1	5.55	0.66	14.20	48.09	—3.76	—18.29	18.47	—14.54	—9.18	—5.95	—23.80	—20.59
0.2	—0.37	0.30	5.93	34.68	—5.52	—11.29	9.95	— 9.23	—6.05	—3.70	—13.00	9.36
0.3	—3.10	0.27	1.40	22.31	—5.63	— 4.90	4.93	— 4.71	—3.28	—2.16	— 3.74	19.48
0.4	—4.43	0.16	—1.56	14.52	—4.92	0.03	1.31	— 1.29	—1.17	—1.00	5.82	36.04
0.5	—4.85	—0.25	—3.57	7.79	—3.57	7.79	—1.53	2.88	0.93	0.40	15.13	53.43
0.6	—4.43	0.16	—4.92	0.03	—1.56	14.52	—3.91	6.96	3.19	1.74	26.01	69.45
0.7	—3.10	0.27	—5.63	— 4.90	1.40	22.31	—5.86	11.63	5.55	2.84	39.13	91.93
0.8	—0.37	0.30	—5.52	—11.29	5.93	34.68	—7.52	17.63	8.59	4.17	56.92	118.72
0.9	5.55	0.66	—3.76	—18.29	14.20	48.09	—8.71	25.93	12.18	5.81	85.71	143.74

For values of p ranging from 0.2 to 0.8, the progress is underestimated for low values of the parameters and overestimated for high values. For low values of both m_1 and m_2 the underestimation reaches its maximum of 5 per cent at $p=0.5$ and decreases as the value of p deviates from 0.5 on either side. On the other hand, for high values of the two parameters, the overestimation increases as the value of p deviates from half on either side, but never exceeds one per cent. For values of $m_1 > m_2$, the response is overestimated for intense selection and underestimated for mild selection. The overestimation for low values of p is proportionately more than underestimation for high values of p. For the highly peaked curves, for example, when $m_1 = 49$ and $m_2 = 2$ the progress is overestimated by as much as 50 per cent for heavy culling and underestimated by as much as 20 per cent for low culling. When $m_1 < m_2$, the results are exactly the reverse as those for $m_1 > m_2$.

The percentage discrepancy when the basis of selection is distributed as gamma decreases with the increase in the value of k, *i.e.*, with the decrease in the degree of skewness and kurtosis. The response is underestimated for heavy culling and overestimated for mild culling. The under estimation does not exceed 24 per cent even for extreme culling ($p=0.1$) and for extreme departures from symmetry and flatness (when $k=1$). On the other hand, the overestimation of progress in such a situation under mild culling ($p=0.9$) can exceed by as much as 86 per cent of that expected under normal form. For large values of $k \geqslant 49$, the maximum discrepancy is only of the order of 6 per cent for all values of p.

When the criterion of selection follows an exponential distribution, the progress is underestimated by about 25 per cent for intense selection, and overestimated by as much as 86 per cent for mild selection, if normal approximation is assumed. Except when p is in the neighbourhood of 0.3 to 0.4, it is not advisable to use the normal approximation as the discrepancy for other values of p is rather too serious to ignore.

The predicted response using normal approximation for log-normal distribution is always an overestimate for all values of $p \geqslant 0.2$ and the overestimation increases with the decrease in the rigour of selection. For very mild selections the overestimation is as high as 150 per cent.

5. Conclusions

The Pearson's Type I and Type III distributions and their derivatives *viz.*, beta and gamma distributions for parametric values

characterizing common situations can be approximated to normal distribution for moderately heavy and low cullings without any serious error in the predicted response to selection. However, the use of normal approximation for exponential and log-normal distributions is not warranted as the discrepancy in response to selection for almost all values of p, the proportion of individuals saved, is too serious to ignore.

Summary

The expressions for intensity of selection appropriate for large populations have been derived for Pearson"s Type I, beta, Pearson's Type III, gamma, exponential and log-normal distributions and compared with the corresponding expression for normal distribution to investigate the effect of using normal approximation in predicting response to selection when the criterion of selection follows one of these non-normal distributions. Excepting the exponential and log-normal, other distributions can be approximated to normal distribution for moderately heavy and low cullings without any serious discrepency in response to selection,

Acknowledgement

The authors are grateful to Dr. D. Singh, Director, Indian Agricultural Statistics Research Institute, New Delhi, for his keen interest in the investigation.

References

[1] Burrows, P.M. (1972) : Expected selection differentials for directional selection. *Biometrics* 28, 1091-1100.

[2] Clayton, G.A., Knight, C.R., Morris, J.A. and Robertson, A. (1957). : An experimental check on quantitative genetical theory III. Correlated responses. *J. Genet.* 55, 171-180.

[3] Dickerson, G.E. (1951) : Effectiveness of selection for economic characters in Swine. *J. Anim. Sci.* 10, 12-21.

[4] Dickerson, G.E. (1955) : Genetic slippage in response to selection for multiple objectives. *Cold. Spr. Harb. Symp. Quant. Biol.* 20, 213-24.

[5] Falconer, D.S. (1953) : Selection for large and small size in mice. *J. Genet.* 51, 470-501.

[6] Fisher, R.A. and Yates, F. (1938). : *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, London.

[7] Gowen, J.W. (1919) : Study of variation of lactation yield and fat contents of Ayrshire cows. *Jour. Agri. Res.* 16, 79-102.

[8] Kapteyn. J.C. (1903). : *Skew Frequency Curves in Biology and Statistics.* Groningen : Noordhoof.

[9] Malhotra, P.K. (1973). : Estimation of genetic components of variation in economic traits of poultry. Dissertation M.Sc. Degree in Agricultural Stat., (IARS), New Delhi.

[10] Om Parkash and Mahajan, Y.P. (1959). : Study of frequency distributions of the characters relating to milk yield and the consequences of non-normality on standard tests of significance. *J. Indian Soc. Agric. Stat.* 11, 163-79.

[11] Pearl, R. and Miner, J.R. (1919). : Variation of Ayrshire cows in the quantity and fat content of their milk. *Jour Agric. Res.* 17, 285-322.

[12] Pearson, K. (1934). : *Tables of the Incomplete Beta Function.* Camb. Univ. Press.

[13] Pearson, K. (1946). : *Tables of the Incomplete Gamma Function.* Camb. Univ. Press.

[14] Quesnberry, C.P. Whitaker, T.B. and Dickens, J.W. (1976). : On testing normality of using several samples. An analysis of peanut afloatoxin data. *Biometries* 32, 753-759.

[15] Sheppard, W.A. (1903). : New tables of the probability integral. *Biometries* 2, 174-190.

[16] Tocher, J.F. (1928). : An investigation of the milk yield of dairy cows. *Biometries* 20B, 105-244.