DISTRIBUTION OF THE RATIO OF SAMPLE RANGE TO SAMPLE STANDARD DEVIATION FOR NORMAL AND COMBINATIONS OF NORMAL DISTRIBUTIONS

By G. A. Baker

College of Agriculture, University of California at Davis

1. Introduction. The distribution of sample ranges in terms of the standard deviation of the sampled population for homogeneous populations has been dealt with in some detail by mathematical methods for the normal parent and by empirical sampling methods for non-normal parents. These results are presented in summary in Tables XXII, XXIII, and XXIV of [1]. Bliss [2] suggests that the range in different sized samples from a normal parent at various levels of significance, in terms of the standard deviation computed with varying degrees of freedom, would be a valuable table. It is not clear whether he means that the standard deviation is to be estimated from the same sample as the range or from a second independent sample, as is done by Newman [3], Pearson and Hartley [4], and Hartley [5].

In natural hybridization of distinct types of plants and subsequent back crossing with parental types distinctly bimodal populations may develop. Heiser [6] has described such a situation for sunflowers. Similar situations may occur in natural and artificial crossing of peaches and apricots as shown by the work of Hesse [7] of this station. In studying such genetical material it often would be helpful to know the expected distributions of the sample ranges in terms of the sample standard deviations estimated from the same sample for certain typical nonhomogenous populations. Applications to such data will be published elsewhere.

Since the mathematical situation for the distributions of the sample range (R) in terms of the sample standard deviation (s) appears somewhat complex, empirical sampling methods were resorted to for obtaining the distributions for a normal parent (N), a symmetrical distinctly bimodal nonhomogeneous parent (A), and a weakly bimodal but strongly skewed parent (B). Populations A and B are pictured in charts A (p. 341) and B (p. 348) of [8].

Population N is approximately represented by

$$(N) \qquad \frac{1296}{5\sqrt{2\pi}} \exp. -\tfrac{1}{2}\frac{(X-15.5)^2}{25},$$

population A by

$$(A) \qquad \frac{648}{5\sqrt{2\pi}}\left(\exp. -\tfrac{1}{2}\frac{(X-15.5)^2}{25} + \exp. -\tfrac{1}{2}\frac{(X-32.5)^2}{25}\right),$$

and population B by

$$(B) \qquad \frac{972}{5\sqrt{2\pi}}\left(\exp. -\tfrac{1}{2}\frac{(X-15.5)^2}{25} + \tfrac{1}{3}\exp. -\tfrac{1}{2}\frac{(X-31.5)^2}{25}\right).$$

The method of drawing samples is the same as that originally described in [9]. N, A, and B each have a total area of 1296. Thus, 1296 integers distributed over a proper range and with the frequencies indicated by the corresponding areas under the curves N, A, and B were entered on charts with 6 big rows and 6 big columns of squares which were subdivided into 6 little rows and 6 little columns. In each case the 1296 integers were distributed in a non-systematic way among the 1296 little squares. By throwing 4 differentiated dice (one die assigned to a big row, one to a big column, one to a little row, and one to a little column) it was possible to draw random individuals from populations that are approximately N, A, and B.

Fisher [10] has defined g_1 which measures the skewness of a distribution and g_2 which measures the flatness. These g's are equivalent to the square root of β_1 and $\beta_2 - 3$, respectively in Karl Pearson's older notation. For population A, $g_1 = 0$ and $g_2 = -1.10$. For population B, $g_1 = 0.62$ and $g_2 = -0.29$.

TABLE 1

Distribution of range in terms of sample standard deviation for samples of specified sizes from a normal parent population (N), $g_1 = 0$, $g_2 = 0$

Sample Size	Number of Samples	Mean	Standard Deviation	g_1	Standard Error of g_1 (Normal)	g_2	Standard Error of g_2 (Normal)
2	...	1.4142	0.0	0.0	...	0.0	...
4	1220	2.2238	0.1564	−0.660	0.0700	0.434	0.1400
16	305	3.5112	0.3879	0.115	0.1396	0.135	0.2783
36	135	4.4014	0.6076	0.607	0.2085	0.332	0.4142
64	76	4.8272	0.6409	0.492	0.2756	−0.751	0.5448
100	48	5.1215	0.6616	−0.077	0.3432	1.038	0.6744

2. Empirical random sampling results. The sample sizes considered are 2, 4, 16, 36, 64, 100. The distribution functions for various sample sizes are characterized by giving means, standard deviations, g_1's, and g_2's. The results are given in Tables 1, 2, and 3. The standard deviations of the samples were computed by dividing the sum of squares by one less than the number in the sample. When the size of the sample is two then the range divided by the standard deviation of the sample is always a constant, square root of 2.

The constants for the distributions for all sample sizes except four were computed without grouping. The constants for the distributions for samples of four were computed from grouped data with a small class interval.

3. Discussion. The mean values of the range divided by the standard deviation of the sample for population A run lower than for populations N and B. The standard deviations of the distributions for all parents increase from zero and continue to increase throughout the range considered for population N.

The standard deviations cut down much more quickly for population A than for population B. The values of g_1 and g_2 show that the distributions are significantly non-normal for certain sample sizes but perhaps not seriously so for other sample sizes.

The distributions of range divided by the sample standard deviation are quite different from the corresponding distributions of range in terms of the standard deviations of the population as can be seen by reference to the tables in [1].

TABLE 2

Distribution of range in terms of sample standard deviation for samples of specified sizes from a bimodal symmetrical population (A), $g_1 = 0$, $g_2 = -1.10$

Sample Size	Number of Samples	Mean	Standard Deviation	g_1	Standard Error of g_1 (Normal)	g_2	Standard Error of g_2 (Normal)
2	...	1.4142	0.0	0.0	...	0.0	...
4	1040	2.2050	0.1551	−0.468	0.0758	−0.356	0.1515
16	259	3.5742	0.5283	1.025	0.1514	1.182	0.3015
36	115	4.0690	0.4604	0.561	0.2255	−0.279	0.4474
64	64	4.3194	0.3377	0.106	0.2993	−1.829	0.5905
100	41	4.4846	0.3194	0.426	0.3695	−0.890	0.7245

TABLE 3

Distribution of range in terms of sample standard deviation for samples of specified sizes from a skewed bimodal population (B), $g_1 = 0.62$, $g_2 = -0.29$

Sample Size	Number of Samples	Mean	Standard Deviation	g_1	Standard Error of g_1 (Normal)	g_2	Standard Error of g_2 (Normal)
2	...	1.4142	...	0.0	...	0.0	...
4	1061	2.2258	0.1459	−0.470	0.0751	−0.142	0.1500
16	265	3.9277	0.5938	0.540	0.1496	0.405	0.2982
36	117	4.4792	0.5476	0.400	0.2236	0.018	0.4437
64	66	4.8485	0.5249	0.534	0.2950	1.028	0.5906
100	42	5.0481	0.3626	−0.092	0.3655	−0.632	0.7166

At the suggestion of the referee it is noted that the empirical results for the means in Table 1 are rather well approximated by $E(R)/E(s)$. It is necessary to remember that $E(s) \neq \sigma$ for small samples. For a discussion of $E(s)$ see Kenney [11] equation 28, page 135.

It is also noted that if

$$X = \log (\log \text{sample size} - \log 2)$$

$$Y = \log \left(\text{mean} \left(\frac{R}{s} \right) - \sqrt{2} \right)$$

then the plots of the (X, Y) values in each case are approximately straight lines for the present range in sample sizes.

The standard deviation and range when determined from the same sample are correlated. For the normal population this correlation decreases and practically disappears for samples of 100 or greater. This is not true for populations A and B. For these populations the correlation between sample range and sample standard deviation decreases much more slowly and seems to be of the order of 0.5 for samples of 100.

REFERENCES

[1] KARL PEARSON, (Editor), *Tables for Statisticians and Biometricians*, Part II, First edition. Cambridge Univ. Press, 1931.

[2] C. I. BLISS. "Review of *Statistical Tables for Biological, Agricultural and Medical Research*, by R. A. Fisher and F. Yates, Second edition," *Science*, Vol. 98 (1943), pp. 346–347.

[3] D. NEWMAN. "The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation," *Biometrika*, Vol. 31 (1939), pp. 20–30.

[4] E. S. PEARSON AND H. O. HARTLEY, "Tables of the probability integral of the studentized range," *Biometrika*, Vol. 33 (1943), pp. 89–99.

[5] H. O. HARTLEY, "Studentization or the elimination of the standard deviation of the parent population from the random sample-distribution of statistics," *Biometrika*, Vol. 33 (1944), pp. 173–180.

[6] CHARLES HEISER, "An analysis of a hybrid swarm of *Helianthus annuus* and *H. petiolaris* near Flagstaff, Arizona," Unpublished data presented at the Genetics Seminar at Davis, Calif. January 7, 1946.

[7] C. O. HESSE, Unpublished data.

[8] G. A. BAKER, "The relation between the means and variances, means squared and variances in samples from the combinations of normal populations," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 333–354.

[9] G. A. BAKER, "Random sampling from nonhomogeneous populations," *Metron.*, Vol. 8 (1930), pp. 68–89.

[10] R. A. FISHER, *Statistical Methods for Research Workers*, 7th edition, Oliver and Boyd, London and Edinburgh, 1938.

[11] J. F. KENNEY, *Mathematics of Statistics*, Part 2, D. Van Nostrand Co., New York, 1939.