situation could easily be avoided by taking C and C' as undefined and defining K as $C+C'$. It is doubtful, however, whether irredundancy of undefined ideas is an especially useful concept.

THE UNIVERSITY OF CALIFORNIA

DISTRIBUTIONS OF GREATEST VARIATES, LEAST VARIATES, AND INTERVALS OF VARIATION IN SAMPLES FROM A RECTANGULAR UNIVERSE*

BY E. G. OLDS

1. *Introduction.* It is proposed to present in this paper the distributions of greatest variates, least variates, and intervals of variation, in samples of size N drawn, without replacement, from the population characterized by the frequency distribution

$$f(x) = \begin{cases} 1 \text{ for } x = 0, 1, 2, \cdots, b, \\ 0 \text{ elsewhere.} \end{cases}$$

This is a finite universe of discrete variates, distributed rectangularly.

The distributions of various statistical parameters, in the case of samples from rectangular distributions, have been investigated by Rietz† and others,‡ but they have been concerned with continuous distributions. The two investigations most closely related to the contents of this paper are those of J. Neyman§ and E. S. Pearson, and of P. R. Rider.¶

* Presented to the Society, December 27, 1934.

† *On a certain law of probability of Laplace*, Proceedings of the International Mathematical Congress, Toronto (1924), vol. 2, pp. 795–799.

‡ Philip Hall, *The distribution of means for samples of size N drawn from a population in which the variate takes values between* 0 *and* 1, *all such values being equally probable*, Biometrika, vol. 19 (1927), pp. 240–244. Allen T. Craig, *On the distributions of certain statistics*, American Journal of Mathematics, vol. 54 (1932), pp. 353–366.

§ *On the use and interpretation of certain test criteria for purposes of statistical inference*, Biometrika, vol. 20A (1928), pp. 175–240.

¶ *On the distribution of the ratio of the mean to standard deviation in small samples from non-normal universes*, Biometrika, vol. 21 (1929), pp. 124–143.

Neyman* and Pearson consider samples of n from a continuous, rectangular distribution of range w. Using W to represent the range in the sample, they find the distribution of W,

$$\phi(W) = n(n-1)\frac{W^{n-2}}{w^{n-1}}\left(1-\frac{W}{w}\right). \tag{1}$$

Rider† assumes an infinite rectangular population of discrete variates and considers the distribution of range in samples of four from a ten-class universe. He obtains the distribution,

$$p(W) = 0.001(-12W^3 + 120W^2 - 2W + 20), \tag{2}$$

and compares it with the distribution of Neyman and Pearson, evaluating (1) for $n=4$ and $w=10$. He notes that the distribution of range and other statistical parameters, for the discrete universe, seem to differ little from those obtained previously for the corresponding continuous universe. It will be interesting to note whether this statement needs to be qualified in the case of the finite universe under consideration.

2. *Greatest Variate.* In a sample of size N, consider, first, the probability, $p(N-1)$, for a greatest variate of size $N-1$. Since the sample must contain the integers from 0 to $N-1$, both inclusive,

$$p(N-1) = \frac{1}{C_{b+1,N}}. \tag{3}$$

If any of the N numbers in this sample be replaced by the number N, the resulting sequence will have the number N as its greatest variate. Corresponding to each sample having $N-1$ as the greatest variate are samples having the number N as greatest variate. This provides a convenient way of enumerating the possible samples which have N as a greatest variate, and we can write

$$p(N) = N\cdot p(N-1). \tag{4}$$

Likewise, the samples with $N+1$ as the greatest variate may be exhibited by replacing each of the numbers in the first-mentioned

* Loc. cit., p. 210.

† Loc. cit., pp. 136–137.

sample, in succession, by the number $N+1$ and choosing the rest of the sample from the numbers less than $N+1$. Continuing this process, it is simple to get the remaining samples, including the cases where the greatest variate is b. We have

$$(5) \qquad p(N+1) = C_{N+1,2}\,p(N-1),$$

and, in general,

$$(6) \qquad p(N+k) = C_{N+k,\,k+1}\,p(N-1), \qquad (k = -1, 0, \cdots, b-N).$$

Replacing $p(N-1)$ in (6) by its value from (3), and replacing $N+k$ by G, we find that the required distribution is

$$(7) \qquad p(G) = \frac{C_{G,\,G-N+1}}{C_{b+1,\,N}},$$

where G may assume all integral values from $N-1$ to b. This may be written more compactly as

$$(8) \qquad p(G) = \frac{N(G)^{(N-1)}}{(b+1)^{(N)}}.$$

The moments of this distribution about the origin may be computed by summing $G^n p(G)$ over the possible values of G, and, by ordinary methods, the moments about the mean may then be obtained. We find

$$(9) \qquad M_G = \frac{N(b+1)}{N+1} - \frac{1}{N+1},$$

and the variance,

$$(10) \qquad \sigma_G^2 = \frac{N(b+1)^2}{(N+2)(N+1)^2} - \frac{(N^2-N)b + (2N^2-N)}{(N+2)(N+1)^2}.$$

It is interesting to compare these results with those obtained from a continuous distribution. Consider the distribution with the probability function

$$\phi(x) = \frac{1}{b+1}, \qquad (0 \leqq x \leqq b+1).$$

Then the probability that the greatest variate in a sample is between G and $G+dG$ is K times the product of the probabilities of $N-1$ members being less than G and one variate lying between G and $G+dG$. Thus*

$$\Phi(G)dG = K\cdot\left[\int_0^G \frac{dx}{b+1}\right]^{N-1}\cdot\frac{dG}{b+1} = \frac{NG^{N-1}}{(b+1)^N}dG, \tag{11}$$

since

$$\int_0^{b+1}\Phi(G)dG = 1.$$

Also

$$\int_0^{b+1} G\Phi(G)dG = \frac{N(b+1)}{N+1}, \tag{12}$$

and

$$\int_0^{b+1}\left[G - \frac{N(b+1)}{N+1}\right]^2\Phi(G)dG = \frac{N(b+1)^2}{(N+2)(N+1)^2}. \tag{13}$$

Compare these three results with (8), (9), (10) above.

3. *Least Variate.* This distribution is found by the same method as used for greatest variate. When the sample consists of the N greatest variates in the population, the smallest is $b-N+1$. It is then apparent that the probability, $r(b-N+1)$, of this least variate is given by

$$r(b-N+1) = \frac{1}{C_{b+1,N}}. \tag{14}$$

Then, reasoning as before,

$$r(b-N) = N\cdot r(b-N+1), \tag{15}$$

and, finally,

$$\begin{aligned} r(b-N-k) &= C_{N+k,\,k+1}\cdot r(b-N+1) \\ &= \frac{N(N+k)^{(N-1)}}{(b+1)^{(N)}}. \end{aligned} \tag{16}$$

* See E. L. Dodd, *Functions of measurement under a general law of error,* Skandinavisk Aktuarietidskrift, 1922.

If L represents the least variate, then, replacing $b-N-k$ by L,

$$r(L) = \frac{N(b-L)^{(N-1)}}{(b+1)^{(N)}}. \tag{17}$$

Proceeding as before, we obtain the mean,

$$M_L = \frac{b-N+1}{N+1}, \tag{18}$$

and the second moment about the mean,

$$\sigma_L^2 = \frac{N(b+1)^2}{(N+2)(N+1)^2} - \frac{(N^2-N)b+(2N^2-N)}{(N+2)(N+1)^2}. \tag{19}$$

It is not surprising that $M_G+M_L=b$ and $\sigma_G^2=\sigma_L^2$.

4. *Interval of Variation.* The least interval of variation, or range, occurs whenever the N numbers in the sample are consecutive. The least number in the sample may be any of the integers $0, 1, 2, \cdots, b-N+1$. Then the probability, $q(N-1)$, of a range $N-1$, is given by

$$q(N-1) = \frac{1}{C_{b+1,N}}(b-N+2). \tag{20}$$

Ranges of size N occur when the largest and smallest variates in the sample differ by N. If the smallest variate is a, the largest is $a+N$, and between the two must lie $N-2$ of the $N-1$ numbers which lie between a and $a+N$. So connected with a and $a+N$ are $N-1$ samples. Furthermore, a may be chosen in $(b-N+1)$ ways, since the greatest admissible value of $a+N$ is b. Therefore

$$q(N) = \frac{(b-N+1)C_{N-1,N-2}}{C_{b+1,N}} = \frac{(b-N+1)C_{N-1,1}}{C_{b+1,N}}. \tag{21}$$

Pursuing the same line of reasoning, we write the general term,

$$q(N+k) = \frac{(b-N-k+1)\,C_{N+k-1,k+1}}{C_{b+1,N}}, \tag{22}$$

which describes the distribution if k varies from -1 to $b-N$. If $N+k$ is replaced by R,

$$q(R) = \frac{(b - R + 1)\, C_{R-1,\, R-N+1}}{C_{b+1,\, N}}, \tag{23}$$

and, after obvious simplification,

$$q(R) = \frac{N(N-1)}{(b+1)^{(N)}} \left[(b+1)(R-1)^{(N-2)} - R^{(N-1)}\right], \tag{24}$$

where R may assume integral values from $N-1$ to b. Proceeding as before, we calculate

$$M_R = \frac{(N-1)(b+1)}{N+1} + \frac{N-1}{N+1}, \tag{25}$$

and

$$\sigma_R{}^2 = \frac{2(N-1)(b+1)^2}{(N+1)^2(N+2)} - \frac{2(N-1)\left[(N-1)(b+1)+N\right]}{(N+1)^2(N+2)}. \tag{26}$$

If in the formulas derived by Neyman and Pearson (see (1) and reference cited), w is replaced by $b+1$, W by R, and n by N, they become

$$\begin{aligned} \phi(R) &= \frac{N(N-1)R^{N-2}}{(b+1)^{N-1}}\left(1 - \frac{R}{b+1}\right) \\ &= \frac{N(N-1)}{(b+1)^N}\left[(b+1)R^{N-2} - R^{N-1}\right], \end{aligned} \tag{27}$$

$$\text{mean} = \frac{(N-1)(b+1)}{N+1}, \tag{28}$$

$$\text{variance} = \frac{2(N-1)(b+1)^2}{(N+1)^2(N+2)}, \tag{29}$$

and the similarity to (24), (25), (26) becomes more apparent.

Also, replacing W by R in (2) and setting N equal to four and b equal to nine in (24) and (27), we have

$$\phi(R) = \frac{1}{1000}(-\ 12R^3 + 120R^2),$$

(for continuous distribution),

$$p(R) = \frac{1}{1000}(-\ 12R^3 + 120R^2 - 2R + 20), \tag{30}$$

(for discrete variates and infinite classes),

$$q(R) = \frac{1}{420}(-\ R^3 + 13R^2 - 32R + 20),$$

(for discrete variates and unit classes).

The third distribution is quite different from the first two. This fact becomes more apparent upon examination of Table I below. If the distribution is of the type assumed in this paper, the true value for the mean range of samples of four is 6.6. Rider's distribution has a mean of 5.93, while that of Neyman-Pearson has a mean of 6. Therefore, either of the latter, if accepted as an estimate for the true mean, gives a result somewhat too small.

5. *Application.* This work was suggested by the problem of sampling automobile license numbers, in states where letters are not combined with the numbers. It was desired to estimate total registration by means of a small sample. If the greatest variate, G_s, observed in a sample, is assumed to be equal to the mean of the greatest variates from all samples, we have

$$G_s = M_G = \frac{N(b+1)}{N+1} - \frac{1}{N+1}. \tag{31}$$

Solving for b, we obtain as one estimate

$$b = \frac{(N+1)(G_s+1)}{N} - 2. \tag{32}$$

TABLE I.* Distribution of Ranges of Samples of Four from a Rectangular Universe.

Range R	$\int\phi(R)dR$	$p(R)$	$q(R)$
0	.0005	.0010	—
1	.0115	.0126	—
2	.0388	.0400	—
3	.0757	.0770	.0333
4	.1150	.1164	.0857
5	.1495	.1510	.1429
6	.1720	.1736	.1905
7	.1753	.1770	.2143
8	.1522	.1540	.2000
9	.0955	.0974	.1333
10	.0140	—	—

$\int\phi(R)dR$ = probability, for a continuous universe, that range will fall in the given class interval.

$p(R)$ = probability of given range for discrete universe of 10 classes, each class having an infinite number of variates.

$q(R)$ = probability of given range for discrete universe of 10 classes, each class having a single variate.

Similarly, the range in a sample, R_s, may be used for a second estimate, giving

$$b = \frac{N+1}{N-1} R_s - 2. \tag{33}$$

CARNEGIE INSTITUTE OF TECHNOLOGY

[*] The first part of this table was given by Rider, (Table XII, loc. cit., p. 136). The last column has been added for the purpose of comparison.