CALCULATION OF EXACT SAMPLING DISTRIBUTION OF RANGES FROM A DISCRETE POPULATION[1]

By Irving W. Burr

Purdue University

1. Introduction. The exact sampling distribution for ranges is known for but few populations, and general information on moments of the range is incomplete. This note gives a method for calculating the exact sampling distribution for discrete universes having a finite range and approximating those for populations with an infinite range.

2. Derivation. Consider a random variable X defined on integers a to b, both finite. Let p_i be the probability that X is i, and $p(R)$ be the probability that the range takes the value R. Then for a sample of n X's from the population (drawn with replacement) we have

$$(1) \qquad p(R) = \sum_{i=a}^{b-R} \sum_{r=1}^{n-1} \sum_{s=1}^{n-r} \frac{n!\, p_i^r p_{i+R}^s}{r!\, s!\, (n-r-s)!} (p_{i+1} + \cdots + p_{i+R-1})^{n-r-s},$$

since the summand contains at least one X at i and at least one X at $i + R$ and those X's not at these values are all between, and the summation is over all possible such samples. To obtain a more useful form we let

$$(2) \qquad M(i, R) = \sum_{j=1}^{i+R} p_j.$$

Then

$$p(R) = \sum_{i=a}^{b-R} \sum_{r=1}^{n-1} \sum_{s=1}^{n-r} \frac{n!\, p_i^r p_{i+R}^s}{r!\, s!\, (n-r-s)!} M^{n-r-s}(i+1, R-2)$$

$$= \sum_{i=a}^{b-R} [\text{terms of } M^n(i, R) \text{ containing at least one } i \text{ and at least one } i + R].$$

To get the desired terms of $M^n(i, R)$, we first subtract from it all of those terms which fail to contain any $i + R$, namely, $M^n(i, R - 1)$. Then we also subtract off those which fail to contain any i, namely $M^n(i + 1, R - 1)$. But these two expressions overlap to the extent of $M^n(i + 1, R - 2)$, that is, terms with neither i nor $i + R$. So this must be added back on. Thus we have

$$(3) \qquad p(R) = \sum_{i=a}^{b-R} [M^n(i, R) - M^n(i, R-1) - M^n(i+1, R-1) + M^n(i+1, R-2)].$$

Received November 27, 1953; revised July 29, 1954.

[1] Presented by title to the Institute, December 27, 1951, at Boston.

To systematize calculation, another form is desirable. Let

$$C_R = \sum_{i=a+1}^{b-R-1} M^n(i, R), \tag{4}$$

$$E_R = M^n(a, R) + M^n(b - R, R). \tag{5}$$

Then we have

$$p(R) = C_R + E_R - 2C_{R-1} - E_{R-1} + C_{R-2}. \tag{6}$$

Formulas (3) and (6) are appropriately modified for $R = 0, 1, b - a - 1$, and $b - a$.

3. Calculation. In computing the $p(R)$, the universe probabilities can best be listed as integer frequencies, as small as possible. Then sums of consecutive frequencies, two at a time, three at a time, etc., are formed, the resulting table being of the same form as a table of differences. Then the C_k and E_k are found by forming sums of nth powers of these table entries. The appropriate modifications of (6) are made by omitting terms naturally absent from this table.

4. An Example. Formula (6) enables us to study the effect on ranges of non-normality in the population. Thus we may compare the following two distributions: One a discrete distribution with probabilities approximately proportional to normal curve areas and the other approximately proportional to those of a well-skewed Pearson Type III.

X	0	1	2	3	4	5	6	7	8	9	10	11
f_1.....	.005	.015	.050	.115	.195	.240	.195	.115	.050	.015	.005	.000
f_2.....	.01	.13	.22	.21	.17	.11	.07	.04	.02	.01	.00	.01

The respective characteristics are

$$\mu = 5.00 \quad \sigma = 1.71 \quad \alpha_3 = 0 \quad \alpha_4 = 3.02$$

$$\mu = 3.45 \quad \sigma = 1.99 \quad \alpha_3 = .99 \quad \alpha_4 = 4.21$$

The respective distributions of range $n = 5$ are the following:

R	0	1	2	3	4	5	6	7	8	9	10	11
$p_1(R)$.....	.001	.031	.146	.239	.251	.179	.096	.040	.013	.003	.0005	
$p_2(R)$.....	.001	.028	.114	.203	.221	.180	.117	.063	.030	.020	.020	.002

The characteristics are respectively

$$\mu_R = 3.93 \quad \sigma_R = 1.53 \quad \alpha_3 = .41 \quad \alpha_4 = 3.01$$

$$\mu_R = 4.44 \quad \sigma_R = 1.94 \quad \alpha_4 = .73 \quad \alpha_4 = 3.47$$

It can be seen that there is much less difference in skewness in the distributions

of R than in the original populations. The R distributions are in fact quite similar if allowance is made for the difference in population standard deviations. Hence we can have quite a bit of confidence in using normal curve constants when making control charts for ranges for moderately skewed populations and small sample sizes.

THE STOCHASTIC CONVERGENCE OF A FUNCTION OF SAMPLE SUCCESSIVE DIFFERENCES[1]

By Lionel Weiss

University of Virginia

1. Summary and introduction. Let $f(x)$ be a bounded density function over the finite interval $[A, B]$ with at most a finite number of discontinuities. Let $X_1, X_2, \cdots, X_n$ be independent chance variables each with the density $f(x)$. Define $Y_1 \leqq Y_2 \leqq \cdots \leqq Y_n$ as the ordered values of $X_1, X_2, \cdots, X_n$, and T_i as $Y_{i+1} - Y_i$. Also define $R_n(t)$ as the proportion of the variates $T_1, \cdots, T_{n-1}$ not greater than $t/(n-1)$. We shall denote $[1 - \int_A^B f(x)e^{-tf(x)}\,dx]$ by $S(t)$, and $\sup_{t \geqq 0} |R_n(t) - S(t)|$ by $V(n)$. Then it is shown that as n increases, $V(n)$ converges stochastically to zero. The relation of this result to other results is discussed.

2. Proof of the stochastic convergence of $V(n)$ to zero.

Lemma 1. *If for each given t, $R_n(t)$ converges stochastically to $S(t)$ as n increases, then $V(n)$ converges stochastically to zero.*

Proof. We must show that for any given positive numbers ϵ and δ, there is a positive integer $N(\epsilon, \delta)$ such that if $n > N(\epsilon, \delta)$, then $P[V(n) < \epsilon] > 1 - \delta$. We can find a finite set of values $t_0 < t_1 < \cdots < t_s$ such that

$$S(t_0) < \tfrac{1}{2}\epsilon, \qquad 1 - S(t_s) < \tfrac{1}{2}\epsilon, \qquad S(t_{i+1}) - S(t_i) < \tfrac{1}{2}\epsilon, \quad i = 0, 1, \cdots, s-1.$$

Also, by the hypothesis of the lemma and other familiar considerations, we can find a positive integer, say $N(\epsilon, \delta)$, such that if $n > N(\epsilon, \delta)$,

$$P[|R_n(t_i) - S(t_i)| < \tfrac{1}{2}\epsilon \text{ for } i = 0, \cdots, s] > 1 - \delta.$$

But then the lemma is proved, for it is easily verified that if $|R_n(t_i) - S(t_i)| < \frac{1}{2}\epsilon$ simultaneously for $i = 0, \cdots, s$, then $|R_n(t) - S(t)| < \epsilon$ simultaneously for all $t \geqq 0$.

Lemma 2. *Let $X_1, \cdots, X_n$ be independent chance variables each with a uniform distribution on* $[0, 1]$. *Let M denote the number of these variables falling in the closed*

Received August 6, 1954.

[1] Research under a grant from the Institute for Research in the Social Sciences, University of Virginia.