NOTE ON THEORETICAL AND OBSERVED DISTRIBUTIONS OF REPETITIVE OCCURRENCES

By P. S. Olmstead

1. A simple problem of repetitive occurrences. Two questions which the engineer often desires to answer whenever he has a new type of apparatus or a new design of an old type of apparatus are: How many times will it perform its intended function without failure? and How many times will it fail to perform its intended function in a given length of time? To do this, he selects a number of what he believes to be identical units of the apparatus and gives each unit a performance test under a uniform test procedure. The number of satisfactory operations prior to the first observed failure to perform this operation is called a "run" and is a measure of the type desired for each unit.

If it is assumed that the probability of failure at any operation is a constant, q, and the probability of satisfactory operation is $1 - q$ or p, then the mathematical probability of runs of 0, 1, 2, 3 $\cdots$ satisfactory operations for any unit are

$$q,\ pq,\ p^2q,\ p^3q,\ \cdots \tag{1}$$

respectively.

Let x denote the number of satisfactory operations in any run. The mean value of x, say m_x, is given by

$$m_x = \frac{p}{q}. \tag{2}$$

The variance of x is

$$\sigma_x^2 = \frac{p}{q^2}. \tag{3}$$

The first step in practice is to determine whether there exists a constant probability, p, by means of the application of the operation of statistical control.[1] Expressions (1), (2), and (3) provide the necessary information for doing this. When a constant probability exists as evidenced by at least 25 consecutive samples of 4 units each the following practical procedure has been found to be satisfactory.

1. An estimate of p (or q), the sole parameter of the distribution, can be obtained from the average length of run in the sample. If p is less than 0.6 and if the sample size is large, a reasonably good estimate of p can be obtained from the proportion of the sample having runs of zero length.
2. The probability of getting runs of length x or more is p^x. Thus, if a minimum (or maximum) value of the probability, p^x, is chosen, a maximum

[1] W. A. Shewhart, "*Statistical Method from the Viewpoint of Quality Control*," The Department of Agriculture Graduate School, Washington, 1939, Chapter I.

(or minimum) expected length of run can be computed for use as a criterion for looking for assignable causes of variation in the length of individual runs by using the estimated value of p.

3. The average and standard deviation to be used in calculating the limits to be applied to successive samples of rational sub-groups in accordance with the Shewhart[2] Criterion I are given by Equations (2) and (3) in which the estimates of p and q are substituted.

2. Application to a signal transmission problem. The theoretical solution given above is a direct answer to the first question at the head of this note.

TABLE I

Observed distributions of runs of x occurrences of event E for various test periods of apparatus life

No. of Occurrences per Period	Freq.	Test Period									
		1	2	3	4	5	6	7	8	11	15
x											
0	n_0	878	1519	961	723	541	407	343	266	160	77
1	n_1	77	226	207	206	171	148	129	97	70	35
2	n_2	2	31	44	55	68	46	52	39	37	27
3	n_3	1	3	8	18	15	19	13	22	19	10
4	n_4		2	1	2	—	6	5	5	7	3
5	n_5			—	1	1	3	1	1	5	2
6	n_6			1			1		—	1	2
7	n_7								1	—	—
8	n_8									2	1
Sample Size	n	958	1781	1222	1005	796	630	543	431	301	157

The second question is also of interest particularly when failure to perform an operation does not impair the apparatus unit for performance of additional operations. In cases of this type, the engineer often lets his test continue for test periods of particular lengths, measured in numbers of operations or sometimes in intervals of time (i.e., time intervals are often considered to be proportional to numbers of operations) and observes the number of failures during the test period for each unit. Thus, he may, after he has assured himself that control exists, arrange his data for each test period to show the frequency of occurrence of 0, 1, 2, 3, $\cdots$ failures per unit.

Data of this type which are typical of those found in other studies made

[2] Loc. cit.

during the past two years are presented in Table I. These were obtained in a signal transmission study in which the data for successive periods were obtained

TABLE II

Comparison of observed and theoretical values of averages and variances for distributions of Table I

Statistic or Parameter		Test Period									
		1	2	3	4	5	6	7	8	11	15
$\bar{q} = \frac{n_0}{n}$	observed	.916	.853	.786	.719	.679	.646	.632	.617	.532	.491
$\bar{x}$	observed	.098	.171	.269	.381	.448	.543	.537	.633	.917	1.026
$m_x = \frac{\bar{p}}{\bar{q}}$	theoretical*	.091	.172	.272	.390	.471	.548	.583	.620	.881	1.039
$\bar{\sigma}_x^2$	observed	.091	.200	.343	.497	.556	.832	.760	1.075	1.783	1.921
$\sigma_x^2 = \frac{\bar{p}}{\bar{q}^2}$	theoretical*	.098	.202	.345	.542	.693	.848	.924	1.005	1.658	2.117

* Based on assumption that $\bar{q}$ is the true value of q.

TABLE III

Theoretical distributions corresponding to distributions of Table I calculated by using $\bar{q} = \frac{n_0}{n}$ as the true value of q

No. of Occurrences per Period	Freq.	Test Period									
x		1	2	3	4	5	6	7	8	11	15
0	n_0*	878.0	1519.0	961.0	723.0	541.0	407.0	343.0	266.0	160.0	77.0
1	n_1	73.3	233.5	205.3	202.8	173.3	144.1	126.4	101.9	74.9	39.2
2	n_2	6.1	32.9	43.8	56.9	55.5	51.0	46.6	39.0	35.1	20.0
3	n_3	.5	4.8	9.4	16.0	17.8	18.0	17.1	14.9	16.5	10.2
4	n_4	.1	.7	2.0	4.5	5.7	6.4	6.3	5.7	7.7	5.2
5	n_5		.1	.4	1.3	1.8	2.3	2.3	2.2	3.6	2.6
6	n_6			.1	.4	.6	.8	.9	.8	1.7	1.4
7	n_7				.1	.2	.3	.3	.3	.8	.7
8	n_8					.1	.1	.1	.1	.4	.3
9 or over	$n_{9-\infty}$								.1	.3	.4
Sample Size	n*	958	1781	1222	1005	796	630	543	431	301	157

* The observed values of n_0 and n form the basis for the calculated distributions.

for separate units. Since each set of these data passed the scrutiny for control, there is justification for assuming that a statistical universe exists and that its functional form may be derived from the observed distribution. It was found that these data were consistent with the assumption that, where the probability of non-occurrence of a failure on a unit in the test period was q, the probability of exactly x failures on a unit was $p^x q$. This set of mathematical probabilities is shown in (1) with q redefined to apply in this case to non-occurrence of a failure.

Observed and "Theoretical" values of the averages and variances for the observed distributions are shown in Table II. The basis for calculating the theoretical values was to take the ratio (designated $\bar{q}$) of n_0 to n for each distribution as the estimate of the true value, q. Distributions as shown in Table III

TABLE IV

Test of fit of theoretical to observed distributions (Table III and Table I, respectively)

	Test Period									
	1	2	3	4	5	6	7	8	11	15
χ^2 *	2.24	0.20	0.32	2.09	9.79	0.65	3.20	6.27	1.07	3.98
Degrees of Freedom	1	2	2	3	3	3	3	3	4	4
P_{χ^2}	.13	.90	.87	.55	.02	.87	.36	.10	.90	.41

* Minimum number in cell for theoretical distribution taken as 5.

were calculated from each $\bar{q}$. These distributions were tested against the ob served distributions by means of the χ^2 test with the results shown in Table IV, which are all within reasonable limits of what might be expected when a constant probability exists.

3. Conclusions. When a constant probability applies to each operation in a repetitive process this note shows how to establish criteria for identifying significantly long or short lengths for individual runs and significantly high or low average lengths for groups of several runs. A problem taken from the field of signal transmission gives assurance of the existence of this type of distribution in practice.

BELL TELEPHONE LABORATORIES,
NEW YORK, N. Y.