AN ANALYSIS OF SCIENTIFIC PRODUCTIVITY

By Clarence Zener

TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS

Read before the Academy April 22, 1968

(1) *Introduction and Results.*—Some years ago Shockley[1] studied the productivity distribution of professional people within large laboratories in government, industry, and universities. A typical productivity distribution is shown by the stepped diagram in Figure 1. Here a measure of productivity has been taken to be the number of papers published within a four-year period. According to this measure, more than a third of the personnel showed no productivity. A major fraction of the papers published came from a very small fraction of the personnel.

From the characteristic shape of such productivity distribution curves, Shockley deduced that productivity, as measured by publication rate, can be expressed as the product of several essentially independent factors. The purpose of the present paper is to develop a method of correlating the number of these independent factors with the precise shape of the productivity distribution curves. This correlation is unique provided we assume a given statistical distribution for the individual factors. We here assume that each factor f has a uniform statistical distribution from zero up to a maximum value $f_{\max}$. The consequences of this simple assumption are deduced in section (2).

Figure 2 presents the theoretical statistical distribution of productivity for several values of the number of independent factors. This number, n, will be called the degree of sophistication. The steepness of the statistical distribution is seen to increase rapidly with the increasing degree of sophistication. The area beneath each curve is exactly halved each time we increase by unity the degree of sophistication. Thus

$$\text{Area} = {}^1/_{2^n}.$$

The maximum productivity is hence $2^n \times$ (average productivity).

We have superposed on the observed data in Figure 1*a* the theoretical productivity curve for a degree of sophistication of 3. This curve has no arbitrary constants. The maximum productivity of 13 was determined by the requirement that

$$258 \text{ Papers} = ({}^1/_8) \times (\text{Max. Papers/Man}) \times 161 \text{ Men}.$$

The close agreement in Figures 1*a* and *b* of the smooth theoretical curve for a degree of sophistication of 3 with the stepped, observed curve gives us confidence in the essential validity of our basic assumptions. We are still left, of course, with the problem of identifying the three basic factors corresponding to a degree of sophistication of 3. The author is especially pleased that the degree of sophistication is, indeed, 3. He has for many years told students that success in research demands curiosity, ability to learn, and intellectual vigor.

Most of the productivity distribution curves given by Shockley use a loga-

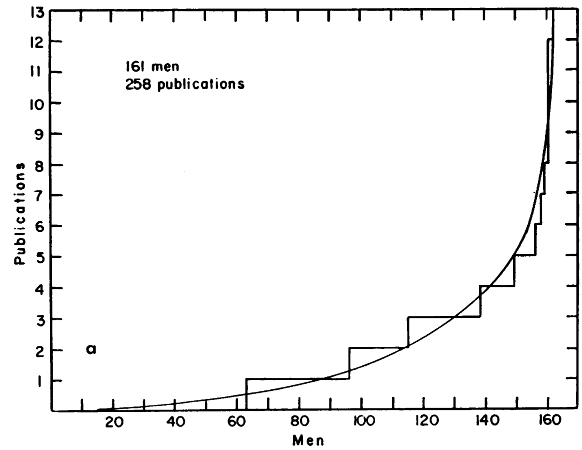

FIG. 1.—(a) *Stepped curve:* Observed distribution of publication rate at Los Alamos, after Shockley.[1]
Smooth curve: Theoretical distribution for productivity with degree of sophistication of 3, drawn for 161 men with a total of 258 publications.

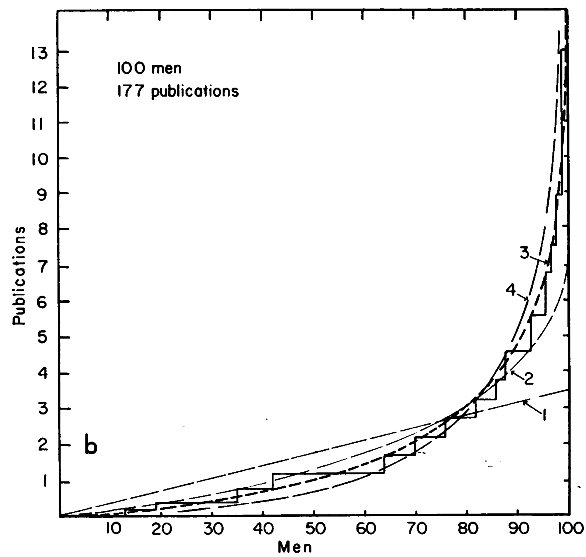

FIG. 1.—(b) *Stepped curve:* "Weighted" observed distribution of publication rate at Brookhaven National Laboratory, after Shockley.[1] (A joint paper of two authors counts $^1/_2$ for each author.)
Smooth curves: Theoretical distribution for productivity assuming a degree of sophistication as labeled for each curve.

rithmic vertical scale. The general shape of both the observed and of the theoretical curves in such a plot is given in Figure 3. Comparison between theory and observations is most readily obtained by comparing the value of S_0 and S_1 as defined by the construction in this figure. The theoretical ratio S_1/S_0 is given below.

n	1	2	3	4	5
S_1/S_0		3.9	6.4	9.1	12.4

All the examples given by Shockley of productivity as measured by publication rate are consistent with a S_1/S_0 ratio of 6.4, and hence with a degree of sophistication of 3.

The general shape of the curve in Figure 3 suggests a log-normal distribution

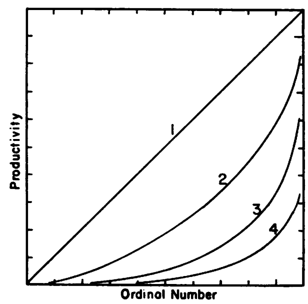

FIG. 2.—Theoretical statistical distribution of productivity. Numbers refer to degree of sophistication.

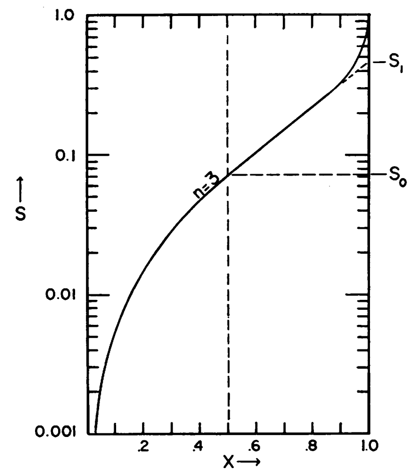

FIG. 3.—Theoretical distribution for productivity, showing how to evaluate rapidly the degree of sophistication from the ratio S_1/S_0.

curve. We find that the theoretical distribution approaches the log-normal distribution only for large values of n. The exact distribution is given by

$$dx/dq = \{1/(n-1)!\}q^{n-1}e^{-q},$$

where x is the normalized ordinal parameter, and

$$q = -\ln\{\text{productivity}/\text{productivity}_{\max}\}.$$

The deviations found by Shockley from a log-normal distribution are in fact reproduced by this theoretical distribution curve.

(2) *Analysis.*—In the first section, we defined success as the product

$$S \equiv f_1, f_2 \ldots f_n,$$

where each factor f_i has a uniform statistical distribution within the range $0 < f_i < 1$. We seek the statistical distribution for S.

Dr. R. R. Hocking has pointed out to the author that for one versed in statistical literature the derivation of the statistical distribution of S is both obvious and trivial. For this purpose, one works with the auxiliary variables $q_1, q_2 \ldots q_n$, Q defined by

$$f_i = e^{-q_i}, \qquad i = 1, 2, \ldots n, \tag{1}$$

$$S = e^{-Q}. \tag{2}$$

Then, if $p(x)$ refers to the statistical distribution function of its argument,

$$p(q_i)(-dq_i) = p(f_i)df_i,$$

and hence, from (1)

$$p(q_i) = e^{-q_i}.$$

We now observe that $p(q_i)$ is precisely a gamma statistical function of degree unity, the general gamma statistical function being defined by

$$\Gamma_n(x) = x^{n-1}e^{-x}/(n-1)! \tag{3}$$

The advantage of this recognition is that the gamma statistical functions are additive in the following sense. If x and x^1 have statistical distributions $\Gamma_n(x)$ and $\Gamma_{n1}(x^1)$ in the positive domain, then the sum $x + x^1$ has the statistical distribution $\Gamma_{n+n1}(x + x^1)$. From the relation

$$Q = q_1 + q_2 + \ldots q_n,$$

we therefore conclude that Q has the gamma statistical distribution of degree n,

$$p(Q) = \Gamma_n(Q). \tag{4}$$

Our final task is to obtain the relation between Q or S and our ordinal parameter x defined as follows. We represent each man in our group by a point in the range $0 < x < 1$. The men are arranged in order of increasing S, and hence decreasing Q, and their points are uniformly spaced. Then S and Q may be regarded as continuous functions of x. From this definition of x it follows that

$$p(Q) = -dx/dQ, \tag{5}$$

and hence

$$x = \int_{Q(x)}^{\infty} p(Q)dQ. \tag{6}$$

The integral on the right is known[2] as the χ^2- integral $p(\chi^2|\surd)$, with

$$\chi^2 = 2Q$$

and

$$\surd = 2n.$$

Tabulated values are given in Tables 7 and 8 of reference 2.

From (2)–(5) we can evaluate the following integral discussed in section (1).

$$\begin{aligned}\int_0^1 S(x)dx &= \int_0^{\infty} Sp(Q)dQ \\ &= \int_0^{\infty} e^{-Q}\Gamma_n(Q)dQ \\ &= {}^1/_2{}^n. \end{aligned} \tag{7}$$

The author wishes to acknowledge partial support by a DOD Themis contract to this university (Office of Naval Research contract N00014-68-A0140).

[1] Shockley, W., *Proc. I.R.E.* (*Inst. Radio Engrs.*), **45**, 279 (1957).

[2] Pearson, E. S., and H. O. Hartley, *Biometrika Tables for Statisticians* (Cambridge University Press, 1966), vol. 1, 3rd ed.