ON A CONTAGIOUS DISTRIBUTION

By R. S. G. Rutherford

University of Sydney, Australia

1. Summary. The purpose of this paper is to discuss the probability distribution that arises when the probability of success at any trial depends linearly upon the number of previous successes. Such a scheme has obvious uses in both biological and economic fields.

It will be shown that by assuming a simple linear relationship between the number of previous successes and the probability of success in the next trial, we can derive a distribution that is reasonably easy to handle, provides as good a fit as more usual distributions, and has parameters which are capable of easy physical interpretation. Moreover, for appropriate values of the parameters the negative binomial and the Gram-Charlier systems can be shown to be close approximations.

2. Introduction. Considerable attention has recently been directed to models where previous experience determines the probabilities in the forthcoming trial. This study is particularly indebted to the work of Woodbury [1]. Much of the recent work has developed the probability scheme originally postulated by Polya [2]. Here it is intended to extend that suggested by Woodbury, and it may be well to contrast the two schemes.

In the Polya scheme, we have an urn containing b black and w white balls. After each random drawing, the drawn ball is returned together with c balls of the same colour. Thus the chance of drawing a ball of given colour depends upon both the number of previous successes and of previous failures.

The Woodbury scheme involves the return of the drawn ball only, if the draw be a failure, and in the event of the draw being a success, the reconstitution of the urn, for example, by the replacement of "failure" balls by "success" balls. In this scheme the order of success is important; in the Polya scheme it is not.

Formally the Woodbury scheme involves that if $P(n, x)$ be the probability of exactly x successes in n trials, and p_x be the probability of success after x previous successes, then

$$P(n + 1, x + 1) = p_x P(n, x) + q_{x+1} P(n, x + 1). \tag{1}$$

Woodbury has solved this problem in the general case.

In this article we postulate further that p_x is a simple linear function of x, viz:

$$p_x = p + cx, \qquad 0 \leqq x \leqq n. \tag{2}$$

Since we must have $0 < p < 1$, we have the limiting conditions,

$$c > 0, \quad n < q/c; \qquad c < 0, \quad n < p/|c|. \tag{3}$$

Received 6/6/52, revised 5/3/54.

This involves that c must always be of order n^{-1} or smaller. These conditions do not prove very restrictive.

3. The distribution and its properties. Following Woodbury, the solution of (1) and (2) may be shown to be

$$P(n, x) = \frac{1}{x!}\left[\left(\frac{p}{c}\right)\left(\frac{p}{c}+1\right)\left(\frac{p}{c}+2\right)\cdots\left(\frac{p}{c}+x-1\right)\right]\sum_{r=0}^{x}(-1)^r \cdot\binom{x}{r}(q-cr)^n. \tag{4}$$

The summation term is clearly the coefficient of $\theta^n/n!$ in $e^{q\theta}(1-e^{-c\theta})^x$.

It is desirable to consider the effect of the restrictions of the conditions (2) and (3) upon the value of $P(n, x)$. It will be shown now (a) that $P(n, x)$ is zero for $x > n$, and (b) that $P(n, x)$ is always positive for $0 \leqq x \leqq n$.

(a) Since the term $(1-e^{-c\theta})^x$ if written as $(c\theta - c^2\theta^2/2! + \cdots)^x$ contains only terms of order θ^x and higher powers of θ, the summation term is zero for all $x > n$.

(b) The condition that $P(n, x)$ as given by relation (4) is always positive within the range $x \leqq n$ requires that with $c > 0$ (i.e., with the product term always positive) the summation term should always be positive, while with $c < 0$ (i.e., with the product term alternately positive and negative) the summation should also alternate in sign, being positive when x is even and negative when x odd. Regarding the summation as the leading xth difference of the series q^n, $(q-c)^n, \cdots, (q-nc)^n$, shows immediately that (3) is a necessary and sufficient guarantee for the summation term to have the correct sign.

The generating function of $P(n, x)$ is the coefficient of $\theta^n/n!$ in $e^{q\theta}[1-(1-e^{-c\theta})c]^{-p/c}$ which may be written

$$e^{\theta}[(1-t)e^{c\theta}+t]^{-p/c}. \tag{5}$$

Though this expression has an infinite number of terms, the terms containing θ^n will occur only in the $n+1$ terms containing powers of t from 1 to t^n. Thus the generating function is a finite one. That the sum of all the $P(n, x)$ for $x = 0 \cdots n$ is equal to 1 may be confirmed by putting $t = 1$ in (5) and considering the coefficient of $\theta^n/n!$ in e^{θ}. This gives us, for the factorial moment generating function, the coefficient of $\theta^n/n!$ in

$$e^{\theta}[1-\alpha(e^{c\theta}-1)]^{-p/c}. \tag{6}$$

Denoting the rth factorial moment by f_r, we have then

$$\begin{aligned} f_1 &= (p/c)[(1+c)^n - 1], \\ f_2 &= (p/c)(p/c+1)[(1+2c)^n - 2(1+c)^n + 1], \\ f_3 &= (p/c)(p/c+1)(p/c+2)[(1+3c)^n - 3(1+2c)^n + 3(1+c)^n - 1]. \end{aligned} \tag{7}$$

4. Empiric fitting. For empiric fitting these three moments (7) should be enough to determine the three parameters n, p, and c. Since the present writer

has been unable to derive a method of fitting on maximum likelihood principles, a somewhat cumbersome method of solution is offered. We may write (7) in the form

$$\frac{p}{c} = \frac{f_1}{(1+c)^n - 1} = \alpha_1 f_1,$$

$$\frac{p}{c} + 1 = \frac{f_2}{f_1}\frac{(1+c)^n - 1}{(1+2c)^n - 2(1+c)^n + 1} = \alpha_2 \frac{f_2}{f_1},$$

$$\frac{p}{c} + 2 = \frac{f_3}{f_2}\frac{(1+2c)^n - 2(1+c)^n + 1}{(1+3c)^n - 3(1+2c)^n + 3(1+c)^n - 1} = \alpha_3 \frac{f_3}{f_1}.$$

To obtain an estimate of n we can approximate these further as

$$np = [1 - (n-1)c/2]f_1,$$

$$(n-1)(p+c) = [1 - (n-3)c/2]f_2/f_1,$$

$$(n-2)(p+2c) = [1 - (n-5)c/2]f_3/f_1.$$

From these relations p and c may readily be eliminated, giving a cubic for n. In this cubic, $n = 1$ is always a root, and the relation may be reduced to a quadratic, of which the positive root is the only relevant one. Since n must be integral, the nearest integer may be taken as a trial value. Having obtained n, it is easy to evaluate p and c. The terms of the distribution are very sensitive to small changes in p and c, which should be evaluated carefully.

It is intended to develop tables of the values of the expressions α_1, α_2, and α_3 which will make the fitting less arduous, and more reliable for ranges in which the above approximations are not valid.

5. Comparisons. The results of fitting this distribution to two classical sets of data are given in Tables I and II. In both cases, the fit of the present distribution is at least as good as in the standard fittings. The improvement is not remarkable, but the parameters of the distribution have a clear physical meaning which can never be claimed for the parameters of the negative binomial or the Neyman contagious set [5]. That is the major claim made for this work.

It is intended now to investigate why other distributions appear to be close approximations in certain circumstances. It is important, however, to make plain the purpose of the following sections. There is no intention to discuss the

TABLE I

Accidents to women working on H.E. shells, data of Greenwood and Yule [3]

$n = 6 \quad p = .059886 \quad c = 0.103036$

Number of accidents	0	1	2	3	4	5	Tot.
Observed frequency	447	132	42	21	3	2	647
Negative binomial	442	140	45	14	5	2	648
Neyman contagious distribution	448	128	49	16	5	1	647
Present distribution	447	130	47	17	5	1	647

TABLE II

Yeast cells in 400 squares of a haemcytometer, data of "Student" [4]

$n = 13 \quad p = .046747 \quad c = 0.019088$

Number of yeast cells	0	1	2	3	4	5	Tot.
Observed frequency	213	128	37	18	3	1	400
Negative binomial	214	123	45	13	4	1	400
Present distribution	215	122	45	14	3	1	400
Gram-Charlier Type B	216	119	46	15	3	1	400

minutae of the conditions under which the approximations will be valid. Such conditions may be found by anyone sufficiently interested.

The purpose in this context is merely to explain why certain distributions have provided reasonably good fits to empiric data which may have, in fact, been generated by a system of the type of (1). A second, and perhaps subsidiary, point is that the fitting of the distribution is difficult, particularly as no maximum likelihood method seems available. This may be overcome in certain ranges by fitting these other distributions, where the parameters are easier to determine, if these parameters can be interpreted in terms of those of the present distribution.

To illustrate and confirm the following sections, a number of actual distributions have been evaluated, together with the approximations under discussion. These are given in Section 8.

6. Binomial approximations. In the negative binomial generated by

$$[(1 + P) - Pt]^{-k}, \tag{8}$$

we have, by standard methods

$$f_1 = kP, \quad f_2 = k(k + 1)P^2, \quad f_3 = k(k + 1)(k + 2)P^3. \tag{9}$$

By the method of moments we can then determine the parameters as

$$k = f_1^2/(f_2 - f_1^2), \quad P = (f_2 - f_1^2)/f_1 . \tag{10}$$

Comparing (9) with (7) shows a considerable similarity of form, if we assume that c and hence p/c are positive. If c and n be small enough for us to equate $(1 + 2c)^n$ and $(1 + c)^{2n}$, we will have at once

$$k = p/c, \quad P = (1 + c)^n - 1. \tag{11}$$

The necessary conditions for this to be valid are somewhat complicated but involve that terms in n^2c^2 may be neglected and/or that $p/c < n - 1$. In practice the first of these is rarely likely to be obtained. However, it can still be demonstrated, by some rather cumbersome analysis not shown here, that so long as $p/c < n - 1$ a negative binomial can be fitted, though the parameters no longer bear easy interpretation in terms of those of the original distribution.

A positive binomial generated by, say, $(Q' + P't)^{k'}$ might provide a good fit if c is either negative or positive with $p/c \geqq n - 1$, that is with $f_1^2 > f_2$. The

first possibility is restricted by the fact that it also would appear to require strictly $p/\mid c\mid < n-1$, contrary to (3). The approximation, however, is reasonably good for values of $p/\mid c\mid$ of about the same order as n. The second case also seems to be relevant only when p/c exceeds $n-1$ by only a small amount. Cases where p/c greatly exceeds $n-1$ may be handled more satisfactorily otherwise, as in the following section.

Both cases, however, suffer from the difficulty that the value of k' is not, in general, integral. This will not be an insuperable difficulty if P' be small and k' large, for then we may be in the territory where a Poisson distribution may approximate to the positive binomial and hence to the original distribution. However, again it seems that large values of p/c or $p/\mid c\mid$ (whether exceeding n or not) may be dealt with best by the Gram-Charlier approximations.

7. Gram-Charlier approximations. *Stage I, binomial type.* Let us now consider cases where the ratio p/c is large. Returning to (4), we have already shown that the summation term is the coefficient of $\theta^n/n!$ in $e^{q\theta}(1-e^{-c\theta})^x$. If c be sufficiently small, then we have

$$(13)\qquad (1-e^{-c\theta})^x/c^x = \theta^x(1-c\theta x/2 + c^2\theta^2 x(3x+1)/24 - \cdots).$$

Hence the summation term is

$$(14)\qquad \frac{n!}{n-x!}c^x q^{n-x}\left(1-\frac{cx(n-x)}{2q}+\frac{c^2x(3x+1)(n-x)(n-x+1)}{24q^2}-\cdots\right)$$

Alternatively, with c small, the summation term is

$$(14a)\qquad \begin{aligned}\Delta^x(q-cx)^n &= \frac{d^x}{d\theta^x}\Big|_{\theta=x/2}(q-c\theta)^n = \frac{n!}{n-x!}\left(q-\frac{cx}{2}\right)^{n-x}c^x \\ &= \frac{n!}{n-x!}c^x q^{n-x}\left(1-\frac{cx(n-x)}{2q}+\frac{c^2x^2(n-x)(n-x+1)}{8q^2}\right).\end{aligned}$$

These approximations are, of course, true for all values of p/c.

If we leave the product part of $P(n, x)$ in its original form, obviously we can obtain as an approximation to the whole expression

$$(15)\qquad P(n,x) = \binom{n}{x}p(p+c)(p+2c)\cdots[p+(x-1)c](q-cx/2)^{n-x}.$$

In this form there is more hope of a maximum likelihood fit.

If, however, p/c be sufficiently large, we may write the product term of (4) as

$$(16)\qquad \begin{aligned}&\frac{p(p+c)\cdots[p+(x-1)c]}{x!c^x} \\ &\quad = \frac{p^x}{x!c^x}\left(1+\frac{c}{p}\sum_{r=0}^{x-1}r+\frac{c^2}{p^2}\sum_{r=0}^{x-1}\sum_{s=0}^{x-1}rs\right) \\ &\quad = \frac{p^x}{x!c^x}\left(1+\frac{c}{p}\frac{x(x-1)}{2}+\frac{c^2}{p^2}\cdot\frac{x(x-1)(x-2)(3x-1)}{24}\right),\end{aligned}$$

and hence derive

$$P(n, x) = \binom{n}{x} q^{n-x} p^x \left\{1 + x[x - (np + q)] \frac{c}{2pq} \right.$$

$$(17) \qquad + x\,[p^2(3x + 1)(n - x)(n - x + 1) - 6pqx(x - 1)(n - x)$$

$$\left. + q^2(x - 1)(x - 2)(3x - 1)] \frac{c^2}{24p^2q^2} - \cdots \right\}.$$

The term in c^2 will be at most of order $c^2n^4/8$, and subsequent terms of smaller order. If, therefore, such terms may be neglected, we may write

$$(17a) \qquad P(n, x) = \binom{n}{x} q^{n-x} p^x \left[1 + x[x - (np + 8)] \frac{c}{2pq}\right].$$

It is instructive to compare this distribution with the binomial distribution having constant probability p. With c positive, that is, probability increasing, $P(n, x)$ exceeds the corresponding binomial term for $x > np + q$, and $P(n, x)$ is less than the corresponding binomial term for $0 < x < np + q$; with c negative, the conditions are reversed. (It can also be established that the conditions on c that make the approximations valid also ensure that $P(n, x)$ is always positive.) If, moreover, n and p are of the order to make the binomial symmetrical, the skewness of the distribution is an immediate guide to the sign and magnitude of c.

Stage II-A, Limiting form for large n and large p. As indicated in Section 5, the conditions necessary to ensure that the approximations will be valid for all x have not been elaborated. It is immediately obvious that much less stringent conditions will apply for early terms of the distribution than those required for the whole distribution. For central values of p, the latter terms of the binomial part of the expression for the distribution will in any case be small, and the absolute if not the proportionate error small.

These considerations become important when we examine the limiting form of (17) when n becomes large. By the change of variable $X = (x - np)/\sqrt{npq}$ used to transform the binomial into the normal distribution, we find as the continuous distribution parallel to the normal

$$(18) \qquad dP = \phi(X)[1 - \tfrac{1}{2}nc + \{n(n - 1)pc \,/\, 2\sqrt{npq}\}X + \tfrac{1}{2}ncX^2]\,dX,$$

where $\phi(X) = (2\pi)^{-1/2} \exp\{-\tfrac{1}{2}X^2\}$. If c/p be not quite small enough to make the Stage I approximations valid, there will be discrepancies at the right tail. The fit will be poor at both tails in any case, in the same way as the normal is a poor approximation to the binomial at the tails. But, by and large we may expect to get a good fit with a curve of the form

$$(19) \qquad dP = \phi(X)[(1 - a_2) + a_1X + a_2X^2]\,dX.$$

By transferring the origin to the mean $x = a_1$ and standardising the distribution, we may obtain readily

$$(20) \qquad dP = \phi(X)[1 + \mu_3 H_{(3)}/3! + (\mu_4 - 3)H_{(4)}/4! + \cdots]\,dX,$$

which is the standard Gram-Charlier Type A distribution [6]. As shown earlier, the skewness of the system indicates the type of scheme operating, that is, the sign and relative magnitude of c. Consideration of the order of the terms involved indicates that we need consider only the terms listed.

Stage II-B, Limiting form for large n and small p. The limiting process used in Stage II-A is of course valid only if p is not small. It is interesting to investigate whether with p small (but p/c still large), we obtain the Gram-Charlier Type B distribution. For all p we may write (17) in the form

$$(21)\quad P(n, x) = \binom{n}{x} q^{n-x}p^x - \lambda\left[\binom{n-1}{x-1} q^{n-x}p^{x-1} - \binom{n-2}{x-2}q^{n-x}p^{x-2}\right],$$

where $\lambda = \frac{1}{2}n(n-1)(p/q)c$. If now p be small, and of order n^{-1}, such that

$$np = m_1, \qquad (n-1)p = m_2, \qquad (n-2)p = m_3,$$

when n becomes large we obtain

$$(22)\quad P(n, x) = e^{-m_1}m_1^x/x! - \lambda e^{-m_2}m_2^{x-1}/(x-1)! + \lambda e^{-m_3}m_3^{x-2}/(x-2)!.$$

It is now reasonable to equate the m's and write

$$P(x) = e^{-m}[m^x/x! - \lambda m^{x-1}/(x-1)! + \lambda m^{x-2}/(x-2)!],$$

which again may be written

$$(23)\quad P(x) = (e^{-m}m^x/x!)[1 - \lambda x/m + \lambda x(x-1)/m^2].$$

This is the required Gram-Charlier Type B [6]. It is most easily fitted by means of the relations

$$(24)\quad \mu_1' = m + \lambda, \qquad \mu_2' = m + 3\lambda - \lambda^2,$$

which can be solved readily for m and λ. As an illustration, "Student's" data have been fitted by this distribution also (Table II). We have $m = 0.61567$ and $\lambda = 0.06683$. The fit, though reasonably good, is poorer than those previously considered; this may reasonably be attributed to the relatively low values of $n = 13$ and of $p/c = 2.45$.

8. Calculations. The validity of the approximations suggested above is demonstrated by 16 examples in Table III. We have here a number of distributions calculated with a selection of values for n, p, and c, and the values given by the relevant approximating distributions. All values are quoted to four decimals, though they have been calculated to five or more. The approximating distributions used are identified by roman numerals.

Type I is the negative binomial fitted from the moments of the distribution. In each case the values of the parameters used are given for comparison with those given by (11), which also are given.

Type II is the approximation of the form suggested in (17a). In this case there is no attempt to find n, p, and c from the data to give the closest fitting curve of this type; the values used are those of the original distribution. Table III shows

TABLE III

Comparison of exact probabilities with those of various approximations (denoted by roman numerals as explained in Section 8) for different values of n, p, and c. Final entry in each row is for vales of $x \geqq$ that at head of column, except that entries in $x = 10$ column are for that value only. Asterisk (*) indicates value is $< .00005$.

n	p	c	$p/\lvert c \rvert$	x:	0	1	2	3	4	5	6	7	8	9	10	Parameters
GROUP A																
100	.0005	.005	0.10	Exact:	.9512	.0375	.0081	.0022	.0007	.0003						
				I:	.9513	.0375	.0081	.0022	.0007	.0002						$P = 0.642460$ $k = 0.100655$
100	.0025	.005	0.50	Exact:	.7786	.1538	.0453	.0147	.0050	.0017	.0006	.0002	.0001			
				I:	.7783	.1543	.0451	.0146	.0050	.0017	.0006	.0002	.0001			$P = 0.631132$ $k = 0.512310$
100	.0050	.005	1.00	Exact:	.6058	.2397	.0943	.0369	.0143	.0055	.0020	.0008	.0007			
				I:	.6052	.2407	.0942	.0366	.0142	.0055	.0021	.0008	.0007			$P = 0.625914$ $k = 1.033160$
100	.0500	.005	10.00	Exact:	.0059	.0243	.0544	.0882	.1156	.1295	.1289	.1167	.3426			
				I:	.0056	.0243	.0568	.0932	.1229	.1381	.1375	.1247	.2969			$P = 0.532497$ $k = 12.1441$
				II:	.0059	.0230	.0485	.0838	.1073	.1383	.1561	.1510	.2861			$np + q = 5.95$ $c/2pq = 19$
GROUP B																
20	.0010	.010	0.10	Exact:	.9802	.0179	.0017	.0002	*							
				I:	.9801	.0180	.0017	.0002	*							$P = 0.200494$ $k = 0.109824$
20	.0050	.010	0.50	Exact:	.9046	.0827	.0109	.0015	.0002	.0001						
				I:	.9043	.0843	.0106	.0015	.0002	*						$P = 0.193605$ $k = 0.568657$
20	.0100	.010	1.00	Exact:	.8179	.1503	.0265	.0045	.0007	.0001						
				I:	.8170	.1519	.0259	.0043	.0007	.0002						$P = 0.184377$ $k = 1.119426$
20	.1000	.010	10.00	Exact:	.1216	.2435	.2574	.1894	.1080	.0505	.0296					
				I:	.1136	.2442	.2653	.1902	.1055	.0473	.0339					$P = 0.023194$ $k = 94.88$
				II:	.1216	.2405	.2567	.1933	.1104	.0492	.0283					$np + q = 2.9$ $c/2pq = 18$
				V:	.1353	.2420	.2420	.1804	.1093	.0551	.0359					$m = 2$ $\lambda = 0.211111$
GROUP C																
10	.1000	.010	10.00	Exact:	.3487	.3686	.1950	.0674	.0169	.0034						
				II:	.3487	.3680	.1953	.0677	.0168	.0033						$np + q = 1.9$ $c/2pq = 18$
				IV:	.3513	.3675	.1922	.0686	.0195	.0029						$m = 1.0462$
				V:	.3679	.3485	.1840	.0705	.0214	.0077						$m = 1.00$ $\lambda = 0.05$

10	.2000	.010	20.00	Exact:	.1074	.2538	.2870	.2039	.1002	.0477						$np + q = 2.8$ $c/2pq = 32$
				II:	.1074	.2533	.2859	.2051	.1012	.0471						
10	.7000	.010	70.00	Exact:	*	.0001	.0011	.0066	.0266	.0770	.1625	.2467	.2583	.1687	.0523	$np + q = 7.3$ $c/2pq = 42$
				II:	*	.0001	.0011	.0062	.0252	.0747	.1629	.2535	.2646	.1652	.0464	$X = .69x - 4.83$ $a_1 = .217$ $a_2 = .05$
				III:	*	.0001	.0007	.0050	.0255	.0821	.1765	.2569	.2401	.1436	.0695	
GROUP D																
10	.5000	.010	50.00	Exact:	.0010	.0089	.0382	.1007	.1809	.2317	.2142	.1411	.0634	.0176	.0023	$np + q = 5.5$ $c/2pq = 50$
				II:	.0010	.0089	.0378	.0996	.1805	.2338	.2174	.1418	.0615	.0159	.0018	$X = .6325x - 3.1625$ $a_1 = .1423$ $a_2 = .050$
				III:	.0022	.0099	.0367	.0982	.1765	.2361	.2113	.1382	.0692	.0175	.0042	
10	.5000	.020	25.00	Exact:	.0010	.0082	.0332	.0862	.1579	.2138	.2167	.1625	.0863	.0294	.0049	$np + q = 5.5$ $c/2pq = 25$
				II:	.0010	.0080	.0316	.0820	.1559	.2215	.2297	.1664	.0791	.0221	.0027	$X = .6325x - 3.1625$ $a_1 = .2846$ $a_2 = .100$
				III:	.0022	.0091	.0312	.0803	.1558	.2272	.2268	.1618	.0745	.0250	.0061	
10	.5000	.040	12.50	Exact:	.0010	.0069	.0252	.0626	.1170	.1723	.2030	.1899	.1359	.0679	.0183	$np + q = 5.5$ $c/2pq = 12.5$
				II:	.0010	.0063	.0193	.0469	.1066	.1969	.2543	.2156	.1143	.0344	.0045	$X = .6325x - 3.1625$ $a_1 = .5692$ $a_2 = .200$
				III:	.0022	.0069	.0189	.0465	.1082	.2034	.2502	.2095	.1055	.0487	.0100	
GROUP E																
10	.1000	−.010	10.00	Exact:	.3487	.4074	.1906	.0464	.0065	.0004						$np + q = 1.9$ $c/2pq = -18$
				II:	.3487	.4068	.1910	.0467	.0059	.0009						$m = .9562$
				IV:	.3844	.3675	.1757	.0560	.0131	.0033						
10	.5000	−.010	50.00	Exact:	.0010	.0107	.0505	.1359	.2299	.2559	.1898	.0926	.0284	.0049	.0004	$np + q = 5.5$ $c/2pq = -50$
				II:	.0010	.0106	.0501	.1348	.2297	.2584	.1928	.0926	.0264	.0036	.0001	$X = .6325x - 3.1625$ $a_1 = .1423$ $a_2 = .050$
				III:	.0022	.0121	.0487	.1322	.2227	.2711	.1879	.0922	.0262	.0045	.0002	

the constant $np + q$ and $c/2pq$, used in (17a) with the corresponding values of n and p.

Type III presents the areas cut off in the range $x \pm \frac{1}{2}$ of the continuous distribution of the form of (19). Table III shows the transformation from x to X, and the values of a_1 and a_2 as calculated from the initial values of n, p, and c, with no effort at improvement.

Type IV is the positive binomial, used where the moments make it impossible to fit a negative binomial. Since the exponent is not an integer, it is fitted as a Poisson, of which the parameter m is given.

Type V is a Gram-Charlier Type B of the form of (23), fitted when appropriate. The parameters m and λ, obtained from (24), are given.

The 16 examples fall into five groups, which examine different aspects of the approximations.

Group A has $n = 100$ and $c = .0050$ throughout. The value of p varies from .0005 to .0500 and the ratio p/c from 0.1 to 10.0. With n as high as 100, only the early terms can be evaluated readily. This limits the possible range of p, and of the ratio p/c. The negative binomial provides a very good fit for low values of p/c. It is still good for the earlier terms of the fourth example, and on the whole better than the Type II approximation.

Group B has a smaller n of 20 and a larger c of 0.01, but values of p such that the same four values of p/c are obtained. Again, for low values of p/c the negative binomial gives a very good fit. For larger values of this ratio, however, Types II and V are also relatively good fits.

Group C comprises three examples where p/c is larger than n, which is 10 in all three. The negative binomial can no longer be fitted. The Type II approximations are reasonably good in all three cases. The Type IV and V approximations in the first example suffer from the small values of n, while the poor Type III approximation in the third example reflects the poorness of the normal as an approximation to the binomial with $n = 10$ and p as large as 0.70.

Group D presents three examples designed to examine the validity of the Gram-Charlier Type A approximation, Type III. Since n is small, a limitation produced by the practical difficulties of computing the "exact" series, central values of p have been taken because the binomial-normal approximation is closest at these values. In the first two examples the Type II fit it sufficiently good to make the Type III fit reasonable. In the third example, with a larger $c = 0.04$, the Type II fit is relatively poor and the Type III fit is worse, reflecting the fact that terms in c^2 may no longer be neglected in (17).

Group E contains two examples in which c is negative, -0.01, with n still 10. Types II and IV are used in the first example, with $p = 0.10$, and Types II and III in the second, with $p = 0.50$.

9. Significance of the results. In all fields of scientific investigation the end goal is always explanation rather than mere description. The negative binomial and the Gram-Charlier set have been found to be good descriptive fits for a large

number of empiric distributions. Probability systems to "explain" them have also been available.

The assumption that we are sampling from a population where the probability varies between individual members and is distributed in the form of a gamma variate will produce as the expected distribution the negative binomial. Equally, the Gram-Charlier system may be derived as the resultant of a small number of linearly additive independent causes of about the same order of importance.

In both cases, distributions arise where these explanations are unconvincing. It has been shown above that a much simpler hypothesis will produce distributions that are at least as good a fit, and which in some cases, though perhaps not in all, provides a more convincing "explanation." As with the normal distribution, we can choose which of two alternative "explanations" is most suitable in any particular case.

The fact that the same probability scheme "explains" both types of distribution considerably systematises the field. Moreover, it appears that there are large areas of possible values for the parameters n, p, and c, where the approximations will not be valid. It is hoped that many empiric distributions which previously have appeared to obey no simple law now may become more tractable.

It is possible, with reasonable ease, to establish that both the negative binomial and the Gram-Charlier Type B distributions may be considered as special cases of the Neyman contagious distribution, and hence that our present distribution will often be closely represented by it. It is more difficult to establish a direct connection, but other writers may succeed.

Though Neyman claims for his series that "All the constants introduced have meanings which are easy to interpret," this does not appear to have been general experience. The distribution of the present study may be of equally general application and provide opportunities for much simpler interpretation.

REFERENCES

[1] M. Woodbury, "On a probability distribution," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 311–313.

[2] G. Polya, "Sur quelques points de la théorie des probabilités," *Ann. Inst. H. Poincaré*, Vol. 1 (1931), pp. 117–162.

[3] M. Greenwood and G. U. Yule, "An enquiry into the nature of frequency distributions of multiple happenings, etc.," *J. Roy. Stat. Soc.*, Vol. 83 (1920), pp. 255–304.

[4] "Student," "On the error of counting with a Haemacytometer," *Biometrika*, Vol. 5 (1907), pp. 351–364.

[5] J. Neyman, "On a new class of contagious distributions applicable in entomology and bacteriology," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 35–57.

[6] M. Kendall, *Advanced Theory of Statistics*, Vol. 1., Griffen and Co., 1945, pp. 145–155.