MOMENTS OF RANDOM GROUP SIZE DISTRIBUTIONS[1]

By John W. Tukey

Princeton University

1. Summary. A number of practical problems involve the solution of a mathematical problem of the class described in the classical language of probability theory as follows: "A number of balls are *independently* distributed among a number of boxes, how many boxes contain no balls, 1 ball, 2 balls, 3 balls, and so on." Problems arising in the oxidation of rubber and the genetics of bacteria are discussed as applications.

A method is given of solving problems of this sort when "how many" is adequately answered by the calculation of means, variances, covariances, third moments, etc. The method is applied to a number of the simplest cases, where the number of balls is fixed, binomially distributed or Poisson and where the "sizes" of the boxes are equal or unequal.

2. Introduction. The distribution of the number of empty boxes has been investigated by Romanovsky in 1934 [3], and, apparently independently, by Stevens in 1937 [4]. Romanovsky investigated the case of N equal boxes and m balls for (i) the case where the balls are independent, and (ii) the case where there is a limit to the size of each box. He gives no motivation for the problem, and shows that certain limiting distributions approach normality. Stevens investigated the case of m independent balls for N boxes (i) of equal size, and (ii) of unequal size, and developed a useful approximation for the last case. Stevens was concerned with this problem in order to test box counts for non-randomness by comparing the number of empty boxes with expectation. The reader interested in that problem is referred to his paper.

The results derived in Part II are based on the use of a chance generating function, a technique which applies easily to the case where the balls are independent. Thus Romanovsky's results for the case of boxes of limited size are neither included or extended. For the other cases where the number of empty boxes has been considered, the results below seem to provide simple moments and cross-moments for the numbers of boxes with any number of balls to the extent previously available for the number of empty boxes. Both Romanovsky and Stevens investigated the actual distribution of the number of empty boxes. A similar investigation of the distribution of the number of b-ball boxes has *not* been carried out here.

3. A chemical problem. In studying the oxidation of rubber, Tobolsky and coworkers were led to propose the following problem: "If a mass of rubber originally consisted of N chains of equal length, if each chain can be broken at a

[1] Prepared in connection with research sponsored by the Office of Naval Research.

large number of places by the reaction with one oxygen molecule, if there are m oxygen molecules each equally likely to react at each link, and if mNp molecules have reacted, what is the probable number of original chains which are now in $b + 1$ parts as a result of b oxygen molecules having reacted with b of their links?

Here an original chain plays the role of a box and an oxygen molecule the role of a ball. The sort of numbers which may be taken as characteristic are:

$$N = 10^{18} \quad \text{(number of chains)},$$
$$m = 10^{16} \text{ to } 10^{20} \quad \text{(number of oxygen molecules)},$$
$$mp = 0.01 \text{ to } 100 \quad \text{(average breaks/chain)}.$$

Thus it is almost certainly going to be appropriate to use the results obtained by assuming N and m very large and $p = 1/N$ very small. We shall return to this example after discussing the general results.

4. A bacteriological problem. The experiments of Newcombe [1] on the irradiation and mutation of bacteria have prompted Pittendrigh to propose the following problem: "Suppose a large number of bacteria each contain m enzyme particles, which have been formed by the action of a nuclear gene. Suppose that irradiation destroys the nuclear gene in a certain fraction of the bacteria. Suppose three generations to occur, during which the m original enzyme particles are randomly distributed among the 8 descendants of an original bacterium. If a bacterium without either nuclear gene or enzyme particle is a recognizable mutant, what is the expected distribution of "families" with 0, 1, 2, 3, $\cdots$, 8 mutants?"

Here the enzyme particles are the balls, and the 8 descendants are the N boxes. We are interested in the number of empty boxes—the problem is that discussed by both Romanovsky and Stevens, with the exception of an allowance for cases where the nuclear gene was not lost. We shall return to this problem also after discussing the general results.

5. The case of large numbers. In case the number of "balls" and "boxes" is large, it is natural and has been customary in similar problems to replace discrete variables by continuous, and derive differential equations. The process runs as follows: Let $y_0, y_1, y_2, \cdots, y_b, \cdots$ be the *fractions* of the total number of boxes containing no, one, two, $\cdots$, b, $\cdots$ balls. Let t be the average number of balls per box (artificially made continuous, so that we may, for example, have a total of $13 + 3\pi$ balls). Increase t to $t + dt$, then of the y_0 boxes previously containing no balls, $y_0\,dt$ will receive one. Of the y_1 boxes previously containing one ball each, $y_1\,dt$ will receive a second, and so on. Hence

$$\frac{dy_0}{dt} = -y_0,$$
$$\frac{dy_1}{dt} = y_0 - y_1,$$
$$\cdots$$

$$\frac{dy_b}{dt} = y_{b-1} - y_b,$$

$$\cdots$$

and if we start, when $t = 0$, with $y_0 = 1$, and $y_b = 0$ for $b > 0$, we find

$$y_b = \frac{t^b}{b!} e^{-t}, \qquad b = 0, 1, 2, \cdots. \tag{1}$$

The usefulness of this result has sometimes been in doubt, thus Opatowski [2, p. 164] says in a similar connection: "Consequently $\cdots$ the theory appears less accurate for small values of t."

It is shown in Part II that; where n_b boxes out of the total of N contain exactly b balls: (I) When the number of *balls and boxes* is large and fixed, (1) is a good approximation to the expectation of n_b/N. (II) When the total number of balls has a Poisson distribution, and t is interpreted as the expected number, (1) *reproduces the expectation* exactly. Since it is appropriate in most problems involving chemical reactions or irradiation to take the number of balls as having a

TABLE 1

A fixed or binomial number of balls and equal boxes

HYPOTHESIS

A total of m balls are independently distributed into N boxes or elsewhere, the chance of a particular ball entering a particular box is p. The number of boxes each containing exactly b balls is n_b.

$$\text{Mean of } n_b = E(n_b) = N\binom{m}{b}(1-p)^m\left(\frac{p}{1-p}\right)^b$$

$$\text{Variance of } n_b = E(n_b)(1 - (1 - \Phi(b, b))E(n_b))$$

$$\text{Covariance of } n_b \text{ and } n_c = -(1 - \Phi(b, c))E(n_b)E(n_c)$$

$$\Phi(b, c) = \left(1 - \frac{1}{N}\right)\frac{(m-c)^{(b)}}{m^{(b)}}\left(1 - \left(\frac{p}{1-p}\right)^2\right)^m\left(\frac{1-p}{1-2p}\right)^{b+c}$$

where $m^{(b)} = m(m-1)\cdots(m-b+1)$ involving b factors

Higher moments See Section 14

Mean of $n_0 = N(1-p)^m$

Mean of $n_1 = Nm(1-p)^m\left(\frac{p}{1-p}\right)$

Variance of $n_0 = N(1-p)^m - N^2(1-p)^{2m} + N(N-1)(1-2p)^m$

Variance of $n_1 = N(N-1)m(m-1)(1-2p)^{m-2}p^2 + Nm(1-p)^{m-1}p - N^2m^2(1-p)^{2m-2}p^2$

Covariance of n_0 and $n_1 = N(N-1)m(1-2p)^{m-1}p - n^2m(1-p)^{2m-1}p$

Poisson distribution, the caution suggested by (I) is often shown unnecessary by (II). For this type of problem the differential equation is entirely adequate!

It is further shown in Part II that, in the Poisson case, the second moments are exactly those which correspond to random sampling from an infinite population with the fractions indicated by the mean number of boxes with 0, 1, 2, $\cdots$, b, $\cdots$ balls. This result is not accidental, and it is shown in Part III how we can see directly that the whole distribution in this case is that of random sampling from such a population.

6. The case of small numbers. The results of Part II also allow us to state the means, variances, and covariances, for the cases where the differential equations do not apply. The results are set forth in the following tables: Tables 1 and 2 apply to the cases where m balls are distributed among the given boxes and possibly others. Thus the total number of balls in the given boxes is either fixed, when there are no other boxes, or follows a binomial distribution.

TABLE 2

A fixed or binomial number of balls and unequal boxes

HYPOTHESIS

A total of m balls are independently distributed into N boxes or elsewhere, the chance of a particular ball entering the ith box being p_i. The average of the $p_i = p$. The sum of the squared fractional deviations of p_i from p is Λ. $p_i = p(1 + \lambda_i), \Sigma_i \lambda_i^2 = \Lambda$. Terms in $\Sigma_i \lambda_i^3$, $\Sigma_i \lambda_i^4$, etc, are to be neglected. The number of boxes each containing exactly b balls is n_b.

Mean of $n_b = E(n_b) = N \binom{m}{b} (1 - p)^{m-b} p^b$ times

$$\left\{\left(1 + \frac{\Lambda}{2N(1 - p)^2}\right) ((mp - b)^2 - (m - b)p^2 - b(1 - p)^2)\right\}$$

Variances and covariances as in Table 1, using

$$\Phi(b, c) \approx \left(1 - \frac{1}{N}\right) \frac{(m - c)^{(b)}}{m^{(b)}} \left(1 - \left(\frac{p}{1 - p}\right)^2\right)^m \left(\frac{1 - p}{1 - 2p}\right)^{b+c} \left(1 + \frac{\Lambda\psi}{2N}\right)$$

where $\psi = 2bc\left(2p - \frac{1}{N}\right) +$ terms in p^2 and in $\frac{p}{N}$

The exact value of ψ is given in Section 16.

Mean of $n_0 = N(1 - p)^m \left(1 + \frac{\Lambda p^2 m(m - 1)}{2N(1 - p)^2}\right)$

Mean of $n_1 = Nm(1 - p)^{m-1} p \left(1 + \frac{\Lambda(m - 1)p(1 - mp)}{2N(1 - p)^2}\right)$

TABLE 3

Poisson balls and equal boxes

HYPOTHESIS

A number of balls with the Poisson distribution, and expectation Nt are independently placed in N boxes. The number of boxes each containing exactly b balls is n_b.

Mean of $n_b = E(n_b) = N \frac{t^b}{b!} e^{-t}$

Variance of $n_b = N\left(\frac{t^b}{b!} e^{-t}\right)\left(1 - \frac{t^b}{b!} e^{-t}\right)$

Covariance of n_b and $n_c = -N\left(\frac{t^b}{b!} e^{-t}\right)\left(\frac{t^c}{c!} e^{-t}\right)$

Mean of $n_0 = Ne^{-t}$,
Mean of $n_1 = Nte^{-t}$,
Variance of $n_0 = Ne^{-t}(1 - e^{-t})$,
Variance of $n_1 = Nte^{-t}(1 - te^{-t})$,
Covariance of n_0 and $n_1 = -Nte^{-2t}$.

7. Discussion of the chemical problem. The number of oxygen molecules which have reacted in a given time is, at best, distributed Poisson. Thus the differential equations would give the expected number of cuts, even if the number of balls or boxes were not large.

The fact that the numbers of balls and boxes, are large makes the variances and covariances so small as to be practically unimportant. Thus, for example, with $N = 10^{18}$, $t = 1$ (1 break per chain), we have:

$$\text{mean of } n_0 = \frac{1}{e} \times 10^{18},$$

$$\text{mean of } n_1 = \frac{1}{e} \times 10^{18},$$

$$\text{variance of } n_0 = \frac{1}{e}\left(1 - \frac{1}{e}\right) \times 10^{18},$$

$$\text{variance of } n_1 = \frac{1}{e}\left(1 - \frac{1}{e}\right) \times 10^{18},$$

$$\text{covariance of } n_0 \text{ and } n_1 = -\frac{1}{e^2} \times 10^{18}.$$

Thus the standard deviations are less than 1 part in 100 million of the mean.

TABLE 4

Poisson balls and varied boxes

HYPOTHESIS

A number of balls with the Poisson distribution are independently placed in N unequal boxes. The expected number placed in the ith box is t_i. The average of the t_i is t, $t_i = t(1 + \lambda_i)$ and $\Sigma_i \lambda_i^2 = \Lambda$. Terms in $\Sigma_i \lambda_i^3$, $\Sigma_i \lambda_i^4$, etc. are to be neglected. The number of boxes each containing exactly b balls is n_b.

$$\text{Mean of } n_b = E(n_b) = N \frac{t^b}{b!} e^{-t} \left(1 + \frac{\Lambda}{2N} ((b - t)^2 - b)\right)$$

$$\text{Variance of } n_b = E(n_b) - \frac{1}{N}\left(1 + \frac{\Lambda}{2N} (b - t)^2)(E(n_b))^2\right)$$

$$\text{Covariance of } n_b \text{ and } n_c = -\frac{1}{N}\left(1 + \frac{\Lambda}{2N} ((b - t)(c - t))E(n_b)E(n_c)\right)$$

$$\text{Mean of } n_0 = Ne^{-t}\left(1 + \frac{\Lambda t^2}{2N}\right)$$

$$\text{Mean of } n_1 = Nte^{-t}\left(1 + \frac{\Lambda(t^2 - 2t)}{2N}\right)$$

$$\text{Variance of } n_0 = Ne^{-t}\left(1 + \frac{\Lambda t^2}{2N}\right) - Ne^{-2t}\left(1 + \frac{3\Lambda t^2}{2N}\right)$$

$$\text{Variance of } n_1 = Nte^{-t}\left(1 + \frac{\lambda(t^2 - 2t)}{2N}\right) - Nt^2e^{-2t}\left(1 + \frac{\Lambda(3t^2 - 6t\ 1)}{2N}\right)$$

$$\text{Covariance of } n_0 \text{ and } n_1 = -Nt^2e^{-2t}\left(1 + \frac{\Lambda(3t^2 - 3t)}{2N}\right)$$

8. Discussion of the bacteriological example. Although this example came from an irradiation experiment, we are not entitled to jump to the Poisson case. The balls are not actions of radiation, but rather previously existing enzyme particles. The purpose of the radiation is merely to make a failure to hand down a particle obvious.

For simplicity, let us begin by assuming that the irradiation has been strong enough to knock out all the nuclear genes and none of the enzyme particles. We face the following problem: "If the m enzyme particles are divided by chance among 8 descendants, what should be the distribution of mutants, that is, of boxes with no balls?"

As far as mean and variance, we can answer this question from Table 1, with $N = 8$ and $p = \frac{1}{8}$.

The results are

$$\text{mean number of mutants} = E(n_0) = 8(\tfrac{7}{8})^m,$$

$$\text{variance of same} = 8(\tfrac{7}{8})^m - 64(\tfrac{7}{8})^{2m} + 56(\tfrac{6}{8})^m.$$

For small values of m we get the values tabled below:

TABLE 5

Blanks out of 8

m	mean	variance	$\text{mean}\left(1 - \frac{\text{mean}}{8}\right)$
0	8	0.000	0.000
1	7	0.000	0.875
2	6.125	.109	1.436
3	5.359	.262	1.769
4	4.689	.417	1.941
5	4.103	.556	1.998
6	3.590	.666	1.979
7	3.142	.747	1.908
8	2.749	.799	1.804
9	2.405	.825	1.682
10	2.105	.829	1.551
15	1.079	.663	.934
20	0.554	.426	.515

We notice that the variance is substantially less than the mean.

Now it might be that the number of enzyme particles is not constant from bacterium to bacterium. It would not be unreasonable if it had a Poisson distribution. If this were the case, we would revert to the differential equation solution, which is also given in Table 3. The last column in Table 5 shows the variance which would then arise for the same means. The variance is still somewhat less than the mean. The situation is shown graphically in Figure 1.

If the actual distribution of n_0 is desired, then it can be calculated for the case where m is fixed from the tables in Stevens' paper [4], and when m is distributed Poisson it is merely a binomial distribution.

PART II

DERIVATIONS

9. The chance generating function. We are considering the following class of problems: "balls" are placed *independently* in "boxes" and then the number n_0 of empty compartments, the number n_1 of compartments containing exactly

one ball, $\cdots$, the number n_b of boxes with exactly b balls, and so on, are observed. We are interested in the moments of n_0, n_1, n_2, $\cdots$, n_b, $\cdots$ both simple and mixed.

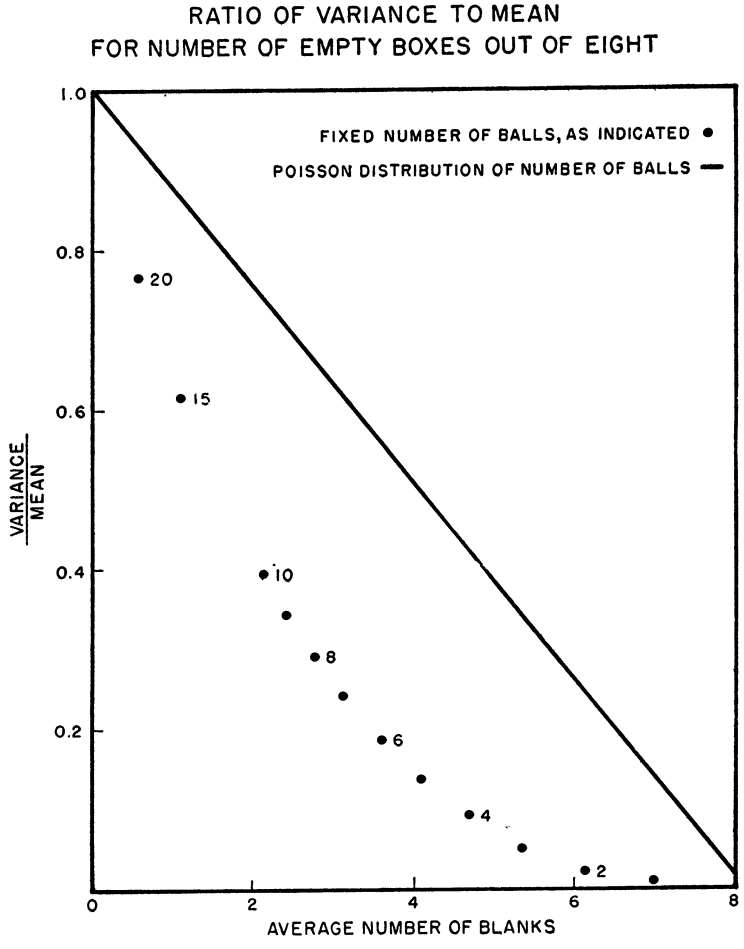

Figure 1

We define chance quantities x_{iq} by

$$x_{iq} = \begin{cases} x, & q\text{th ball in the } i\text{th box}, \\ 1, & \text{otherwise}. \end{cases}$$

Clearly the product of all x_{iq} for fixed i is given by

$$\Pi_q x_{iq} = x^{\text{(number of balls in the } i\text{th box)}}$$

Thus $\Pi_q x_{iq} = x^b$ if and only if there are exactly b balls in the ith box. Hence the coefficient of x^b in $\Sigma_i \Pi_q x_{iq}$, the sum of $\Pi_q x_{iq}$ over all boxes i, is n_b, the number of boxes containing exactly b balls.

We have the relation

$$\Sigma_b n_b x^b = f(x) = \Sigma_i \Pi_q x_{iq},$$

where $f(x)$ is a chance function, and the n_b and the x_{iq} are chance quantities.

Now we take expectations of both sides, and use the fact that the expectation of a sum is the sum of the expectations to obtain

$$\Sigma_b x^b E(n_b) = E(f(x)) = \Sigma_i E(\Pi_q x_{iq}).$$

Now x_{iq} and x_{ir}, for $q \neq r$, are independent since they are determined by different and independent balls. Hence $E(\Pi_q x_{iq}) = \Pi_q(E x_{iq})$ and we have the basic formula

$$E(f(x)) = \Sigma_b x^b E(n_b) = \Sigma_i \Pi_q E(x_{iq}). \tag{1}$$

10. Higher moments. By extending this device, we can obtain generating functions for higher moments. Instead of the x_{iq}, we introduce a whole sequence of chance quantities $x_{iq}, y_{iq}, z_{iq}, \cdots, w_{iq}$, defined by

$$(x_{iq}, y_{iq}, \cdots, w_{iq}) = \begin{cases} (x, y, \cdots, w), & q\text{th ball in } i\text{th box}, \\ (1, 1, \cdots, 1), & \text{otherwise}. \end{cases}$$

We find immediately that

$$\begin{aligned} f(x)f(y)\cdots f(w) &= (\Sigma_i \Pi_q x_{iq})(\Sigma_j \Pi_r y_{jr}) \cdots (\Sigma_n \Pi_p w_{np}) \\ &= \Sigma_i \Sigma_j \cdots \Sigma_n \Pi_q x_{iq} y_{jq} \cdots w_{nq}. \end{aligned}$$

Taking expectations on both sides

$$\begin{aligned} E(f(x)f(y)\cdots f(w)) &= \Sigma_i \Sigma_j \cdots \Sigma_n E(\Pi_q x_{iq} y_{jq} \cdots w_{nq}) \\ &= \Sigma_i \Sigma_j \cdots \Sigma_n \Pi_q E(x_{iq} y_{jq} \cdots w_{nq}), \end{aligned}$$

where we have used the fact that $x_{iq}y_{jq}\cdots w_{nq}$ and $x_{ir}y_{jr}\cdots w_{nr}$ are independent when $q \neq r$ since they are determined by different and independent balls.

On the other hand,

$$\begin{aligned} f(x)f(y)\cdots f(w) &= (\Sigma_b n_b x^b)(\Sigma_c n_c y^c)\cdots(\Sigma_a n_a w^a) \\ &= \Sigma_b \Sigma_c \cdots \Sigma_a (n_b n_c \cdots n_a)(x^b y^c \cdots w^a) \end{aligned}$$

so that

$$E(f(x)f(y)\cdots f(w)) = \Sigma_b \Sigma_c \cdots \Sigma_a (x^b y^c \cdots w^a) E(n_b n_c \cdots n_a).$$

Equating the two expressions for the expectation of $f(x)f(y)\cdots f(w)$, we have, finally, the generating function for $E(n_b n_c \cdots n_n)$ in the form

$$\sum_{b,c,\cdots,a} (x^b y^c \cdots w^a) E(n_b n_c \cdots n_a) = \sum_{i,j,\cdots,n} \Pi_q E(x_{iq} y_{jq} \cdots w_{nq}). \tag{2}$$

Thus a knowledge of $E(x_{iq}y_{jq}\cdots w_{nq})$ will allow us to determine the moments of the n's.

11. A fixed or binomial number of balls and equal boxes. Let there be N boxes, and m balls, each with probability p of entering each box. If $pN = 1$ we have the case where m balls always appear in the boxes taken together—the case of a fixed number of balls. If $pN < 1$, the number of balls appearing in all boxes taken together is a binomial with expectation mpN.

Now x_{iq} equals 1 with probability $1 - p$ and equals x with probability p, hence (1) becomes

$$\Sigma_b x^b E(n_b) = \Sigma_i \Pi_q (1 - p + px) = N(1 - p + px)^m.$$

Using the binomial theorem, the coefficient of x^b is

$$E(n_b) = N\binom{m}{b}(1-p)^{m-b}p^b = N\binom{m}{b}(1-p)^m\left(\frac{p}{1-p}\right)^b. \tag{3}$$

Now if p is small, we may approximate $1 - p$ by e^{-p} and by 1, respectively, in its two occurences, where

$$E(n_b) \approx N\left(\frac{m}{b}\right)e^{-mp}p^b$$

and if m is large compared to b this becomes

$$E(n_b) \approx N\frac{(mp)^b}{b!}e^{-mp}.$$

12. Second moments. We must study $E(x_{iq}y_{jq})$. If $i = j$ then this is $(1 - p + pxy)$ since the qth ball falls into *both* the ith and jth boxes with probability p, otherwise into neither. If $i \neq j$, we immediately find the expectation to be $(1 - 2p + px + py)$.

Hence, since $i = j$ in N cases, and $i \neq j$ in $N(N - 1)$ cases,

$$\Sigma_{ij}\Pi_q E(x_{iq}y_{jq}) = N(1 - p + pxy)^m + N(N - 1)(1 - 2p + px + py)^m,$$

by (2) this equals $\Sigma_{bc}x^b y^c E(n_b n_c)$, and using the multinomial expansion we find

$$E(n_b n_c) = N(N-1)\binom{m}{b\,c}(1-2p)^{m-b-c}p^{bc} + \delta(b,c)N\binom{m}{b}(1-p)^{m-b}p^b,$$

where $\delta(b, c) = 1$ when $b = c$ and is zero otherwise, and where the multinomial coefficient $\binom{m}{b\,c}$ is given by

$$\binom{m}{b\,c} = \frac{m!}{b!c!(m-b-c)!}.$$

We now set

$$E(n_b n_c) = E(n_b)E(n_c)\Phi(b,c) + \delta(b,c)E(n_b), \tag{4}$$

when

$$\Phi(b, c) = \frac{N(N-1)\binom{m}{b\,c}(1-2p)^m\left(\frac{p}{1-2p}\right)^{b+c}}{N\binom{m}{b}(1-p)^m\left(\frac{p}{1-p}\right)^b N\binom{m}{c}(1-p)^m\left(\frac{p}{1-p}\right)^c}$$

$$= \left(1-\frac{1}{N}\right)\frac{\binom{m}{b\,c}}{\binom{m}{b}\binom{m}{c}}\left(\frac{1-2p}{(1-p)(1-p)}\right)^m\left(\frac{1-p}{1-2p}\right)^{b+c} \tag{5}$$

$$= \left(1-\frac{1}{N}\right)\frac{(m-c)^{(b)}}{m^{(b)}}\left(1-\left(\frac{p}{1-p}\right)^2\right)^m\left(\frac{1-p}{1-2p}\right)^{b+c}$$

where $u^{(b)} = u(u-1)\cdots(u-b+1)$ denotes a descending factorial with b factors.

Notice that, if the n_b were independently distributed in Poisson distributions, the second moments would be given by the same formula with $\Phi(b, c) = 1$, while if they were distributed like a multinomial sample from an infinite population the second moments would be given by the same formula with $\Phi(b, c) = 1 - \frac{1}{N}$.

For small p, we have

$$\Phi(b, c) \approx \left(1-\frac{1}{N}\right)\frac{(m-c)^{(b)}}{m^{(b)}},$$

and if m is large compared to b and c, this approaches the multinomial value

$$\Phi(b, c) \approx \left(1-\frac{1}{N}\right).$$

13. Variances and covariances. The variances and covariances are given by

$$\begin{aligned}\text{Variance } (n_b) &= E(n_b n_b) - E(n_b)E(n_b)\\ &= E(n_b)(1-(1-\Phi(b, b))E(n_b)),\end{aligned}$$

and

$$\text{Covariance } (n_b, n_c) = -(1-\Phi(b, c))E(n_b)E(n_c).$$

Thus the covariance of n_b and n_c will vanish when, and only when $\Phi(b, c) = 1$.

Let us suppose $pN = \frac{1}{\beta}$, with p small and m and N large, and see if $\Phi(b, c)$ can be unity. Since a preliminary calculation shows it to be reasonable, let us put $m = \gamma N$. Then

$$\Phi(b, c) \approx (1-\beta p)\frac{(\gamma N - c)^{(b)}}{(\gamma N)^{(b)}}(1-p^2)^{\gamma N}(1+p)^{b+c}.$$

An easy calculation shows that the ratio of descending factorials is nearly

$$e^{-bc/\gamma N} = e^{(-bc\beta/\gamma)p},$$

making further natural approximations,

$$\ln \Phi(b, c) \approx -\beta p - \frac{bc\beta}{\gamma} p - \gamma N p^2 + (b + c)p$$

and this may be written

$$\ln \Phi(b, c) \approx -\frac{\beta p}{4\gamma}\left(\left(2\frac{\gamma}{\beta} - b - c + \beta\right)^2 + 4\beta c - (b - \beta - c)^2\right),$$

and this vanishes for real γ when and only when $|\, b - \beta - c \,| \geq \sqrt{4\beta c}$. This, then, is the condition on b and c which permits the existence of two ratios of m to N so that for either ratio and large N there will be no correlation between n_b and n_c.

14. Higher moments. To deal with the third moments, we need $E(x_{iq}y_{jq}z_{kq})$, which is easily seen to behave as follows:

Relation of ijk	*number of occurrences*	*Expectation of* $x_{iq}\, y_{jq}\, z_{kq}$
$i = j = k$	N	$1 - p + pxyz$
$i = j \neq k$	$N(N - 1)$	$1 - 2p + pxy + pz$
$i = k \neq j$	$N(N - 1)$	$1 - 2p + pxz + py$
$j = k \neq i$	$N(N - 1)$	$1 - 2p + pyz + px$
different	$N(N - 1)(N - 2)$	$1 - 3p + px + py + pz$

Thus we have

$$\Sigma_{bcd}x^b y^c z^d E(n_b n_c n_d) = N(1 - p + pxyz)^m + N(N - 1)(1 - 2p + pxy + pz)^m$$
$$+ N(N - 1)(1 - 2p + pxz + py)^m + N(N - 1)(1 - 2p + pyz + px)^m$$
$$+ N(N - 1)(N - 2)(1 - 3p + px + py + pz)^m$$

from which we can calculate all third moments.

In general if ϵ is a decomposition of the product $xyz \cdots w$ into α monomials $u_1, u_2, \cdots, u_\alpha$, where order is disregarded (for example: $xyz = (xz)y = (zx)y = y(zx) = y(xz)$ is a single decomposition with $\alpha = 2$, $u_1 = xz$, $u_2 = y$), then the generating function becomes

$$\Sigma_\epsilon N^{(\alpha)}(1 + (u_1 + u_2 + \cdots + u_\alpha - \alpha)p)^m.$$

15. Poisson balls and equal boxes. To reach a Poisson distribution we let $m \to \infty$ and $p \to 0$ so that $mNp = tN$, where t is the *average number of balls per box* in the Poisson distribution.

Since

$$p^b\binom{m}{b} \to \frac{t^b}{b!}$$

under these conditions, (3) becomes

$$E(n_b) = N\frac{t^b}{b!}e^{-t} \tag{6}$$

and from (5) it follows that the limit of $\Phi(b, c)$ is $\left(1 - \frac{1}{N}\right)$ so that

$$E(n_b n_c) = N(N-1)\frac{t^{b+c}}{b!\,c!}e^{-2t} + \delta(b, c)N\frac{t^b}{b!}e^{-t}, \tag{7}$$

and hence

$$\text{Variance } (n_b) = N\left(\frac{t^b}{b!}e^{-t}\right)\left(1 - \frac{t^b}{b!}e^{-t}\right), \tag{8}$$

$$\text{Covariance } (n_b, n_c) = -N\left(\frac{t^b}{b!}e^{-t}\right)\left(\frac{t^c}{c!}e^{-t}\right). \tag{9}$$

Notice that these are the moments of the numbers of objects of types $b, c, \cdots$, in a random sample of N from an infinite population where the fraction of b's is $t^b e^{-t}/b!$, just as it should be.

16. Fixed or binomial balls and varied boxes. We now consider the case where the chance of any ball entering the ith box is p_i. We shall again not restrict ourselves to the case $\Sigma_i p_i = 1$.

The expectation of x_{iq} is immediately seen to be $(1 + p_i(x - 1)) = (1 - p_i + p_i x)$, so that the generating function is

$$f(x) = \Sigma_i(1 - p_i + p_i x)^m$$

and the expectation of n_b is

$$E(n_b) = \binom{m}{b}\Sigma_i(1 - p_i)^{m-b}p_i^b = \binom{m}{b}\Sigma_i(1 - p_i)^m\left(\frac{p_i}{1 - p_i}\right)^b. \tag{10}$$

Following Stevens [4] with a slight modification, let us set $p_i = p(1 + \lambda_i)$, where p is the average of the p_i, so that $\Sigma_i\lambda_i = 0$. Then

$$(1 - p_i) = (1 - p(1 + \lambda_i)) = (1 - p)\left(1 - \frac{p\lambda_i}{1 - p}\right),$$

so that

$$\Sigma_i(1 - p_i)^{m-b}p_i^b = (1 - p)^{m-b}p^b\Sigma_i\left(1 - \frac{p\lambda_i}{1 - p}\right)^{m-b}(1 + \lambda_i)^b.$$

Expanding the summand, we find

$$1 + \left\{ -\frac{(n-b)p}{1-p} + b \right\} \lambda_i + \left\{ \frac{(m-b)(m-b-1)p^2}{2(1-p)^2} - \frac{(m-b)bp}{1-p} + \frac{b(b-1)}{2} \right\} \lambda_i^2 + O(\lambda_i^3).$$

Hence, setting $\Sigma_i \lambda_i^2 = \Lambda$ (notice this is not the same as Stevens' Λ!), we have

$$E(n_b) = \binom{m}{b} (1-p)^{m-b} p^b \left\{ N + \tfrac{1}{2}\Lambda \left(\frac{m-b}{m-b-1} \frac{(p(m-1)-b)^2}{(1-p)^2} - \frac{b(m-1)}{m-b-1} \right) \right\} + O(\Sigma_i \lambda_i^3).$$

The expectation for all $p_i = p$ has been modified by multiplication by

$$1 + \frac{\Lambda}{2N} \left\{ \frac{m-b}{m-b-1} \frac{(p(m-1)-b)^2}{(1-p)^2} - \frac{b(m-1)}{m-b-1} \right\} \tag{11}$$

plus terms of higher order. For large N and consequently small p the quantity in braces is nearly

$$b\left(b - \frac{m}{m-b}\right)$$

and more roughly is approximately b^2. Similarly, the expectations of second moments are

$$E(n_b n_c) = \binom{m}{b\,c} \sum_{i \neq j} (1 - p_i - p_j)^{m-b-c} p_i^b p_j^c + \delta(b,c) \binom{m}{b} \textstyle\sum_i (1-p_i)^{m-b} p_i^b,$$

whence

$$\Phi(b,c) = \frac{\binom{m}{b\,c} \sum_{i \neq j} (1 - p_i - p_j)^{m-b-c} p_i^b p_j^c}{\binom{m}{b}\binom{m}{c} \sum_i (1-p_i)^{m-b} p_i^b \sum_j (1-p_j)^{m-c} p_j^c}. \tag{12}$$

Making the same sort of expansion yields

$$\Phi(b,c) \approx \left(1 - \frac{1}{N}\right) \frac{(m-c)^{(b)}}{m^{(b)}} \left(1 - \frac{p^2}{(1-p)^2}\right)^m \left(\frac{1-p}{1-2p}\right)^{b+c} \left(1 + \frac{\Lambda\psi}{2N}\right) \tag{13}$$

where terms in $\Sigma\lambda_i^3$ have been neglected (note that

$$\sum_{i \neq j} \lambda_i \lambda_j = -\sum_i \lambda_i^2 = -\Lambda \cdot),$$

and where

$$\psi = \left\{ \frac{m-b-c}{m-b-c-1} \frac{N-2}{N-1} (1-2p)^{-2} - \frac{m-b}{m-b-1} (1-p)^{-2} \right\} \cdot \{p(m-1)-b\}^2$$

$$+\left\{\frac{m-b-c}{m-b-c-1}\frac{N-2}{N-1}(1-2p)^{-2}-\frac{m-c}{m-c-1}(1-p)^{-2}\right\}$$
$$\cdot\{p(m-1)-c\}^2$$
$$+\frac{1}{2}\frac{m-b-c}{m-b-c-1}\left\{\frac{N}{N-1}-\frac{N-2}{N-1}(1-2p)^{-2}\right\}(b-c)^2$$
$$+\frac{1}{m-b-c-1}\left\{\frac{2bc}{N-1}-\frac{b^2c}{m-b-1}-\frac{c^2b}{m-c-1}\right\}.$$

This can be reduced to

$$\psi = 2bc\left(2p-\frac{1}{N}\right)+O(p^2)+O\left(\frac{b}{N}\right),$$

and for $p = 1/N + O(p^2) + O\left(\frac{b}{N}\right)$.

$$\psi = 2pbc + O(p^2).$$

17. Poisson balls and varied boxes. To reach the Poisson limit, we let $m \to \infty$ and $p_i \to 0$ so that $mp_i = t_i$. The generating function for first moments becomes

$$f(x) = \Sigma_i e^{-t_i+t_ix}$$

and the expectation of n_b is

$$E(n_b) = \Sigma_i \frac{t_i^b}{b!}e^{-t_i}. \tag{15}$$

If we set $t_i = t(1+\Lambda_i)$, this becomes

$$E(n_b) = \frac{t^b}{b!}e^{-t}\Sigma_i(1+\lambda_i)^b e^{-t\lambda_i}$$

The summand expands in the form

$$\left(1+b\lambda_i+\frac{b(b-1)}{2}\lambda_i^2+\frac{b(b-1)(b-2)}{6}\lambda_i^3+\cdots\right)$$
$$\times\left(1-t\lambda_i+\frac{t^2}{2}\lambda_i^2-\frac{t^3}{6}\lambda_i^3+\cdots\right)$$
$$= 1+(b-t)\lambda_i+\left(\frac{b(b-1)}{2}-bt+\frac{t^2}{2}\right)\lambda_i^2+\cdots.$$

If t is chosen as the average of the t_i so that $\Sigma\lambda_i = 0$, the sum becomes

$$N+\left(\frac{(b-t)^2-b}{2}\right)\Sigma\lambda_i^2+\left(\frac{(b-t)^3}{6}-\frac{3b-2}{6}+\frac{bt}{2}\right)\Sigma\lambda_i^3+\cdots.$$

Again setting $\Sigma\lambda_i^2 = \Lambda$ we have

$$E(n_b) \approx \frac{t^b}{b!}e^{-t}\left(N+\left(\frac{(b-t)^2-b}{2}\right)\Lambda\right) \tag{16}$$

which can be written

$$E(n_b) \approx N \frac{t^b}{b!} e^{-t} \left(1 + \frac{\Lambda}{2N} ((b - t)^2 - b)\right).$$

The generating function for the second moments is

$$f(x)f(y) = \Sigma_{ij} e^{-t_i + t_i x - t_j + t_j y}$$

so that the expectation of $n_b n_c$ is

$$E(n_b n_c) = \sum_{i \neq j} \frac{t_i^b t_j^c e^{-t_i - t_j}}{b!c!} + \delta(b, c) \sum_i \frac{t_i^b}{b!} e^{-t_i} \tag{17}$$

which becomes

$$E(n_b n_c) = \frac{t^{bc}}{b!c!} e^{-2t} \sum_{i \neq j} (1 + \lambda_i)^b (1 + \lambda_j)^c e^{-t\lambda_i - t\lambda_j} + \delta(b, c) E(n_b),$$

whence we can derive

$$\Phi(b, c) \approx 1 - \frac{1}{N} - \frac{\Lambda}{2N^2} (b - t)(c - t). \tag{18}$$

Thus

$$\text{Variance } (n_b) \approx E(n_b) - \frac{1}{N} \left\{1 + \frac{\Lambda}{2N} (b - t)^2\right\} (E(n_b))^2, \tag{19}$$

$$\text{Covariance } (n_b n_c) \approx - \frac{1}{N} \left(1 + \frac{\Lambda}{2N} (b - t)(c - t)\right) E(n_b) E(n_c). \tag{20}$$

18. Boxes in a systematic square. Another case which it may be worthwhile to write down arises when the boxes are systematically "rotated" under "spouts" of different probability. That is, the number of balls m is a multiple of the number of boxes N, and the probability of the qth ball entering the ith box depends on the value of $q - i$ taken modulo N. An example for $N = 3$ and $m = 6$ follows:

Probabilities of entry

Box	Ball 1	2	3	4	5	6
1	p_0	p_1	p_2	p_0	p_1	p_2
2	p_2	p_0	p_1	p_2	p_0	p_1
3	p_1	p_2	p_0	p_1	p_2	p_0

If $m = kN$ and the subscript r runs through 0, 1, 2, $\cdots$, $N - 1$, then the expectation of $f(x)$ becomes

$$\Sigma_i \Sigma_q E(x_{iq}) = N\{\Pi_r(1 - p_r + p_r x)\}^m.$$

Thus first moments, and by proceeding similarly higher moments, are available for this case also.

PART III

THE POISSON CASE

19. The Poisson case with equal boxes. The Poisson case is obtained in the limit as $m \to \infty$ and $p \to 0$ with $pm = t$. We wish to show that, in the limit, the number of balls in the different boxes are independent. Let $k_1, k_2, \cdots, k_N$ be the number of balls in the first, second, $\cdots$, Nth box, respectively. Then the distribution of the k's is given by, where we write $k = k_1 + k_2 + \cdots + k_N$,

$$\frac{m^{(k)}}{k_1!k_2! \cdots k_N!} p^k(1 - Np)^{m-k} = \frac{m^{(k)}}{m^k} \frac{(1 - Np)^{m-k}}{e^{-Nmp}} \Pi_i \frac{(mp)^{k_i} e^{-mp}}{k_i!}$$

Now the first two fractions clearly approach unity in the limit, and the independence is proved.

Since the number of balls in each box has an independent Poisson distribution, the distribution of the numbers of boxes each with exactly b balls is that of a random sample of N from an infinite population—namely it is a multivariate distribution with probabilities

$$\frac{(mp)^b e^{-mp}}{b!} .$$

REFERENCES

[1] H. H. NEWCOMBE, "Delayed phenotypic expression of spontaneous mutations in *Escherichia coli*," *Genetics, Vol. 33* (1948), p. 447–476.

[2] I. OPATOWSKI, "Chain processes and their biophysical applications: Part I. General Theory," *Bulletin of Mathematical Biophysics, Vol. 7* (1945), p. 161–180.

[3] V. ROMANOVSKY, "Su due problemi di distribuzione casuale," *Giornale dell'Istituto Italiano degli Attuari, Vol. 5* (1934), p. 196–218.

[4] W. L. STEVENS, "Significance of grouping," *Annals of Eugenics, Vol. 8* (1937–1938), p. 57–59.