EDITORIAL

FUNDAMENTALS OF THE THEORY OF SAMPLING

I. Sampling from a Limited Supply

We shall consider first a population of s individuals, in which each individual possesses a common attribute that can be measured quantitatively. The sum of the associated variates may be expressed as follows:

$$x_1 + x_2 + x_3 + \cdots\cdots x_s = \sum^{s} x = sM_x$$

From this so-called *parent population* it is possible to select $\binom{s}{r}$ different *samples*, each consisting of r individuals, ($r \leq s$). These samples may be ordered after any fashion, and the algebraic sum of the variates for the respective samples may be designated

$$\begin{array}{llllll}
z_1 = & x_1 & +x_2 & +x_3 & \cdots\cdots +x_r & = \sum^{r:1} x \\
z_2 = & x_2 & +x_3 & +x_4 & \cdots\cdots +x_{r+1} & = \sum^{r:2} x \\
\cdot & \cdot & \cdot & \cdot & \cdot\cdot\cdot\cdot\cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot\cdot\cdot\cdot\cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot\cdot\cdot\cdot\cdot & \cdot \\
z_{\binom{s}{r}} = & x_{s-r+1} & +x_{s-r+2} & \cdot & \cdots\cdots +x_s & = \sum^{r:\binom{s}{r}} x
\end{array}$$

Thus, while $\sum^{s} x$ represents the sum of all the s variates in the parent population, $\sum^{r:i} x$ designates the sum of the r variates occurring in the i th sample.

We face now the problem of describing adequately, from a statistical point of view, the distribution of these $\binom{s}{r}$ values of z, that is to say, we must express the moments $\mu_{n:z}$ in terms of the moments of the parent population, $\mu_{n:x}$.

By definition $$M_z = \frac{\sum z}{\binom{s}{r}}$$

Since each value of z will contribute r terms to the value of $\sum z$, this latter expression will consist of $r\cdot\binom{s}{r}$ terms involving each of the s variates of the parent population alike. Therefore, each variate, x_i ($i = 1, 2, 3, \quad . \quad . \quad . \quad s$), will occur in the expression for $\sum z$ exactly $\frac{r}{s}.\binom{s}{r}$ times. Consequently

$$(1) \qquad M_z = \frac{\sum z}{\binom{s}{r}} = \frac{1}{\binom{s}{r}} \cdot \frac{r}{s} \cdot \binom{s}{r} \left\{ x_1 + x_2 + \cdots x_s \right\} = \frac{r}{s} \sum x = r M_x$$

We shall now investigate the values of

$$\mu_{n \cdot z} = \frac{\sum \bar{z}^{\,n}}{\binom{s}{r}}$$

where we choose to represent a deviation from the mean as

$$\bar{z}_i = z_i - M_z$$

Observing that

$$\bar{z}_1 = z_1 - M_z = x_1 + x_2 + \cdot \cdot \; x_r - r M_x = \bar{x}_1 + \bar{x}_2 + \cdot \cdot \cdot \cdot \bar{x}_r$$

we note that

$$\bar{z}_1^{\,2} = \sum^{r:1} \bar{x}^{\,2} + 2 \sum^{r:1} \bar{x}_i \, \bar{x}_j$$
$$\bar{z}_2^{\,2} = \sum^{r:2} \bar{x}^{\,2} + 2 \sum^{r:2} \bar{x}_i \, \bar{x}_j$$
$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$
$$\bar{z}_{\binom{s}{r}}^{\,2} = \sum^{r:\binom{s}{r}} \bar{x}^{\,2} + 2 \sum^{r:\binom{s}{r}} \bar{x}_i \, \bar{x}_j$$

Therefore

$$\mu_{2 \cdot z} = \frac{\sum \bar{z}^{\,2}}{\binom{s}{r}} = \frac{1}{\binom{s}{r}} \left\{ \frac{r \cdot \binom{s}{r}}{s} \sum^{s} \bar{x}^{\,2} + 2 \frac{\binom{r}{2}\binom{s}{r}}{\binom{s}{2}} \sum^{s} \bar{x}_i \, \bar{x}_j \right\},$$

or, writing

$$\rho_i = \frac{r^{(i)}}{s^{(i)}} = \frac{r\,(r-1)(r-2) \cdot \cdot \cdot \cdot \cdot (r - \overline{i-1})}{s\,(s-1)(s-2) \qquad\qquad (s - \overline{i-1})}$$

$$(2a) \qquad \mu_{2 : z} = 2! \left\{ \rho_1 \frac{\sum^{s} \bar{x}^{\,2}}{2!} + \rho_2 \frac{\sum^{s} \bar{x}_i \, \bar{x}_j}{(1!)^2} \right\}$$

By utilizing further the multinomial theorem, it follows easily that

$$(3a)\qquad \mu_{3:x} = 3!\left\{\rho_1 \frac{\sum^s \bar{x}^3}{3!} + \rho_2 \frac{\sum^s \bar{x}_i^2 \bar{x}_j}{2!\,1!} + \rho_3 \frac{\sum^s \bar{x}_i \bar{x}_j \bar{x}_k}{(1!)^3}\right\}$$

$$(4a)\qquad \mu_{4:x} = 4!\left\{\rho_1 \frac{\sum^s \bar{x}^4}{4!} + \rho_2 \frac{\sum^s \bar{x}_i^3 \bar{x}_j}{3!\,1!} + \rho_2 \frac{\sum^s \bar{x}_i^2 \bar{x}_j^2}{(2!)^2}\right.$$

$$\left. + \rho_3 \frac{\sum^s \bar{x}_i^2 \bar{x}_j \bar{x}_k}{2!\,(1!)^2} + \rho_4 \frac{\sum^s \bar{x}_i \bar{x}_j \bar{x}_k \bar{x}_l}{(1!)^4}\right\}$$

etc.

The rule for writing down the terms is as follows: The number of terms in the expression for $\mu_{n:x}$ equals the number of partitions that can be formed from the integer n. The subscript of ρ equals the number of elements in the corresponding partition, and exponents of $\bar{x}$ and the factorials in the denominators are in fact the elements of the partitions.

Our next problem is to express the summations in terms of moments of the parent population, $\mu_{n:x}$.

First order summation

$$\sum^s \bar{x} = s\mu_{1:x} = 0$$

Second order summations

$$\sum \bar{x}^2 = s\mu_{2\cdot x}$$

$$2\sum \bar{x}_i \bar{x}_j = -s\mu_{2\cdot x}$$

since $$\Big(\sum^s \bar{x}\Big)^2 = 0 = \sum^s \bar{x}^2 + 2\sum^s \bar{x}_i \bar{x}_j$$

Third order summations

$$\sum^s \bar{x}^3 = s\mu_{3\cdot x}$$

$$\sum^s \bar{x}_i^2 \bar{x}_j = -s\mu_{3\cdot x}$$

$$3\sum^s \bar{x}_i \bar{x}_j \bar{x}_k = s\mu_{3\cdot x} .$$

since $\sum^{s}\bar{x}^2 \sum^{s}\bar{x} = 0 = \sum^{s}\bar{x}^3 + \sum^{s}\bar{x}_i^2\bar{x}_j$

and $\left(\sum^{s}\bar{x}\right)^3 = 0 = \sum^{s}\bar{x}^3 + 3\sum^{s}\bar{x}_i^2\bar{x}_j + 6\sum^{s}\bar{x}_i\,\bar{x}_j\,\bar{x}_k$

Fourth order summations

$$\begin{aligned}
\sum^{s}\bar{x}^4 &= s\mu_{4:x}\\
\sum^{s}\bar{x}_i^3\,\bar{x}_j &= -s\mu_{4:x}\\
2\sum^{s}\bar{x}_i^2\,\bar{x}_j^2 &= -s\mu_{4:x} + s^2\mu_{2:x}^2\\
2\sum^{s}\bar{x}_i^2\bar{x}_j\,\bar{x}_k &= 2s\mu_{4:x} - s^2\mu_{2:x}^2\\
8\sum^{s}\bar{x}_i\,\bar{x}_j\,\bar{x}_k\,\bar{x}_l &= -2\,s\mu_{4:x} + s^2\mu_{2:x}^2
\end{aligned}$$

Utilizing these summations, (2a), (3a) and (4a) may be written

(2) $\mu_{2:\bar{x}} = s\mu_{2:x}\left\{\rho_1 - \rho_2\right\}$

(3) $\mu_{3:\bar{x}} = s\,\mu_{3:x}\left\{\rho_1 - 3\rho_2 + 2\rho_3\right\}$

(4) $\mu_{4:\bar{x}} = s\mu_{4:x}\left\{\rho_1 - 7\rho_2 + 12\rho_3 - 6\rho_4\right\} + 3s^2\mu_{2:x}^2\left\{\rho_2 - 2\rho_3 + \rho_4\right\}.$

Continuing after this fashion, one can show after a lavish use of symmetric functions that

(5) $$\begin{aligned}\mu_{5:\bar{x}} = {}& s\mu_{5:x}\left\{\rho_1 - 15\rho_2 + 50\rho_3 - 60\rho_4 + 24\rho_5\right\}\\ &+ 10s^2\mu_{3:x}\mu_{2:x}\left\{\rho_2 - 4\rho_3 + 5\rho_4 - 2\rho_5\right\},\end{aligned}$$

(6) $$\begin{aligned}\mu_{6:\bar{x}} = {}& s\mu_{6:x}\left\{\rho_1 - 31\rho_2 + 180\rho_3 - 390\rho_4 + 360\rho_5 - 120\rho_6\right\}\\ &+ 15\,s^2\mu_{4:x}\mu_{2:x}\left\{\rho_2 - 8\rho_3 + 19\rho_4 - 18\rho_5 + 6\rho_6\right\}\\ &+ 10\,s^2\mu_{3:x}^2\left\{\rho_2 - 6\rho_3 + 13\rho_4 - 12\rho_5 + 4\rho_6\right\}\\ &+ 15\,s^3\mu_{2:x}^3\left\{\rho_3 - 3\rho_4 + 3\rho_5 - \rho_6\right\},\end{aligned}$$

$$
\begin{aligned}
(7)\quad \mu_{7:z} = s\mu_{7:x}\{&\rho_1 - 63\rho_2 + 602\rho_3 - 2100\rho_4 + 3360\rho_5 \\
&\qquad - 2520\rho_6 + 720\rho_7\} \\
+ 21s^2\mu_{5:x}\mu_{2:x}\{&\rho_2 - 16\rho_3 + 65\rho_4 - 110\rho_5 \\
&\qquad + 84\rho_6 - 24\rho_7\} \\
+ 35s^2\mu_{4:x}\mu_{3:x}\{&\rho_2 - 10\rho_3 + 35\rho_4 - 56\rho_5 + 42\rho_6 - 12\rho_7\} \\
+ 105s^3\mu_{3:x}\mu^2_{2:x}\{&\rho_3 - 5\rho_4 + 9\rho_5 - 7\rho_6 + 2\rho_7\},
\end{aligned}
$$

$$
\begin{aligned}
(8)\quad \mu_{8:z} = s\mu_{8:x}\{&\rho_1 - 127\rho_2 + 1932\rho_3 - 10206\rho_4 \\
&+ 25200\rho_5 - 31920\rho_6 + 20160\rho_7 - 5040\rho_8\} \\
+ 28s^2\mu_{6:x}\mu_{2:x}\{&\rho_2 - 32\rho_3 + 211\rho_4 - 570\rho_5 \\
&\qquad + 750\rho_6 - 480\rho_7 + 120\rho_8\} \\
+ 56s^2\mu_{5:x}\mu_{3:x}\{&\rho_2 - 18\rho_3 + 97\rho_4 - 240\rho_5 + 304\rho_6 \\
&\qquad - 192\rho_7 + 48\rho_8\} \\
+ 35s^2\mu^2_{4:x}\{&\rho_2 - 14\rho_3 + 73\rho_4 - 180\rho_5 + 228\rho_6 - 144\rho_7 + 36\rho_8\} \\
+ 210s^3\mu_{4:x}\mu^2_{2:x}\{&\rho_3 - 9\rho_4 + 27\rho_5 - 37\rho_6 + 24\rho_7 - 6\rho_8\} \\
+ 280s^3\mu^2_{3:x}\mu_{2:x}\{&\rho_3 - 7\rho_4 + 19\rho_5 - 25\rho_6 + 16\rho_7 - 4\rho_8\} \\
+ 105s^4\mu^4_{2:x}\{&\rho_4 - 4\rho_5 + 6\rho_6 - 4\rho_7 + \rho_8\}.
\end{aligned}
$$

It is convenient, at this point, to define the "*n th sampling polynomial*" as follows:

$$(9)\qquad P_n(\rho) = D_x^n \log(\rho e^x + 1 - \rho)\Big]_{x=0}$$

If we place $y = \log(\rho e^x + 1 - \rho)$, then

$$y^{(1)} = \frac{\rho e^x}{\rho e^x + 1 - \rho} \quad \text{or} \quad (\rho e^x + 1 - \rho)\, y^{(1)} = \rho e^x$$

Taking the n th derivative of both sides by utilizing Leibnitz' Theorem, we obtain

$$(\rho e^x + 1 - \rho)\, y^{(n+1)} + \binom{n}{1} \rho e^x y^{(n)} + \binom{n}{2} \rho e^x y^{(n-1)} + \cdots = \rho e^x.$$

Placing $x = 0$ in this equation yields, by definition,

$$P_{n+1}(\rho) + \binom{n}{1} \rho P_n(\rho) + \binom{n}{2} \rho P_{n-1}(\rho) + \cdots\cdot = \rho$$

That is, for $n = 0, 1, 2, \cdots\cdot\cdot$

$$P_1(\rho) = \rho$$

$$P_2(\rho) + \rho P_1(\rho) = \rho$$

$$P_3(\rho) + 2\rho P_2(\rho) + \rho P_1(\rho) = \rho$$

$$P_4(\rho) + 3\rho P_3(\rho) + 3\rho P_2(\rho) + \rho P_1(\rho) = \rho$$

etc.

Thus:

$$(10) \quad \begin{cases} P_1(\rho) = \rho \\ P_2(\rho) = \rho - \rho^2 \\ P_3(\rho) = \rho - 3\rho^2 + 2\rho^3 \\ P_4(\rho) = \rho - 7\rho^2 + 12\rho^3 - 6\rho^4 \\ P_5(\rho) = \rho - 15\rho^2 + 50\rho^3 - 60\rho^4 + 24 p^5 \\ P_6(\rho) = \rho - 3\rho^2 + 180\rho^3 - 390\rho^4 + 360\rho^5 - 120\rho^6 \\ P_7(\rho) = \rho - 63\rho^2 + 602\rho^3 - 2100\rho^4 + 3360\rho^5 - 2520\rho^6 + 720\rho^7 \\ P_8(\rho) = \rho - 127\rho^2 + 1932\rho^3 - 10206\rho^4 + 25200\rho^5 - 31920\rho^6 \\ \qquad\qquad \left[+ 20160\rho^7 - 5040\rho^8\right] \end{cases}$$

etc.

The law of formation of the coefficients is obvious: for if $c_{i:n}$ designates the coefficient of ρ^i in the expression for $P_n(\rho)$,

$$c_{i:n} = i c_{i:n-1} - (i-1) c_{i-1:n-1}$$

Comparing the polynomials of equations (9) with formulae (2) to (8) inclusive, suggests writing the expressions for $\mu_{n:x}$ in the following symbolic form:

$$
\begin{aligned}
\mu_{2:x} &= 2!\left\{P_2 \frac{s\mu_{2:x}}{2!}\right\} \\
\mu_{3:x} &= 3!\left\{P_3 \frac{s\mu_{3:x}}{3!}\right\} \\
\mu_{4:x} &= 4!\left\{P_4 \frac{s\mu_{4:x}}{4!} + \frac{P_2^2}{2!}\,\frac{s\mu^2_{2:x}}{(2!)^2}\right\} \\
\mu_{5:x} &= 5!\left\{P_5 \frac{s\mu_{5:x}}{5!} + P_3 P_2 \frac{s^2\mu_{3:x}\mu_{2:x}}{3!\,2!}\right\} \\
\mu_{6:x} &= 6!\left\{P_6 \frac{s\mu_{6:x}}{6!} + P_4 P_2 \frac{s^2\mu_{4:x}\mu_{2:x}}{4!\,2!} + \frac{P_3^2}{2!}\,\frac{s^2\mu^2_{3:x}}{(3!)^2}\right. \\
&\qquad \left. + \frac{P_2^3}{3!}\,\frac{s^3\mu^3_{2:x}}{(2!)^3}\right\} \\
&\text{etc.}
\end{aligned}
\tag{11}
$$

By P_n we understand an expression derived from the sampling polynomial, $P_n(\rho)$, by writing ρ^i as ρ_i . Thus,

$$P_4(\rho) = \rho - 7\rho^2 + 12\rho^3 - 6\rho^4 \text{ , whereas}$$

$$P_4 = \rho_1 - 7\rho_2 + 12\rho_3 - 6\rho_4$$

Again, since

$$P_3(\rho)\cdot P_1(\rho)\cdot P_1(\rho) = (\rho - 3\rho^2 + 2\rho^3)\rho\cdot\rho = \rho^3 - 3\rho^4 + 2\rho^5,$$

$$P_3 P_1^2 = \rho_3 - 3\rho_4 + 2\rho_5$$

The number of terms in the expression for $\mu_{n:x}$ will equal the number of partitions that can be formed from the integer n . The subscripts of the P and μ factors for any selected term correspond to the elements of the corresponding partition, and the exponent of s equals the number of elements in the partition. The factorials beneath the μ factors agree with the order of these moments, and the factorials appearing occasionally under the P factors depend upon the number of times that any P is repeated as a factor in that term. All terms arising from a partition in which unity is an element have been neglected, since such terms will contain $\mu_{1:x}$ as a factor and consequently be equal to zero.

Illustration I. For the parent population we shall select the following (it will be noted that graphically the ordinates terminate on the hypotenuse of an isosceles right triangle):

TABLE I

Parent Population

x	f_x
1	24
2	23
3	22
4	21
5	20
..	..
22	3
23	2
24	1
Total	300

The mean, standard deviation and moments about the mean for this distribution are as follows:

$$M_x = 8.666$$
$$\mu_{2:x} = 33.222 \qquad \sigma_x = 5.76387$$
$$\mu_{3:x} = 108.526 \qquad \alpha_{3:x} = .566749$$
$$\mu_{4:x} = 2642.27 \qquad \alpha_{4:x} = 2.39398$$
$$\mu_{5:x} = 20525.2 \qquad \alpha_{5:x} = 5.69279$$
$$\mu_{6:x} = 322570 \qquad \alpha_{6:x} = 27.3878$$

It may well be remarked at this point that the standard variate corresponding to an observed variate, x_i is

$$(12) \qquad t_i = \frac{x_i - M_x}{\sigma_x} = \frac{\bar{x}_i}{\sigma_x} ,$$

and is consequently an *abstract number.* The n th moment of the standard variates is also without unit, i. e.

$$\alpha_{n:x} = \frac{\Sigma t^n}{N} = \frac{1}{N\sigma_x^n} \Sigma \bar{x}_i^n = \frac{\mu_{n:x}}{\sigma_x^n} \tag{13}$$

In dealing with distributions one should always bear in mind that the mean and standard deviation determine merely the *position* of the centroid vertical and the *scale* of the distribution, but that the standard moments are influenced by the *shape* of the distribution alone. Consequently a study of the mathematical representation of frequency distributions is essentially an investigation concerning the standard moments of observed and theoretical distributions.

From the above parent population it would be possible to select $\binom{300}{25}$ samples, each consisting of 25 individuals. To describe the distribution of these sampes, we proceed as follows:

$$\rho_1 = \qquad = .08333$$

$$\rho_2 = \rho_1 \cdot \frac{24}{299} = .0066 \quad 8896 \quad 32$$

$$\rho_3 = \rho_2 \cdot \frac{23}{298} = .0005 \quad 1626 \quad 226$$

$$\rho_4 = \rho_3 \cdot \frac{22}{297} = .0000 \quad 3824 \quad 1649$$

$$\rho_5 = \rho_4 \cdot \frac{21}{296} = .0000 \quad 0271 \quad 3090 \quad 0$$

$$\rho_6 = \rho_5 \cdot \frac{20}{295} = .0000 \quad 0018 \quad 3938 \quad 31$$

$P_2 =$	.0766	4437	0		$P_6 =$	—.0450	5692	2
$P_3 =$	.0642	9896	8		$P_4 =$	.0032	3772	45
$P_4 =$	.0424	7628	8		$P_3^2 =$	.0040	5670	98
$P_2^2 =$	.0056	9468	03		$P_2^3 =$	.0004	0949	264
$P_5 =$	.0065	8261	36					
$P_3P_2 =$	.0048	0969	62					

$M_x = 216.66$	
$\mu_{2.x} = 763.88$	$\sigma_x = 27.6385$
$\mu_{3:x} = 2093.43$	$\alpha_{3:x} = 0991550$
$\mu_{4:x} = 1730700$	$\alpha_{4:x} = 2.96594$
$\mu_{5:x} = 15647600$	$\alpha_{5:x} = .970225$
$\mu_{6:x} = 6503500000$	$\alpha_{6:x} = 14.5900$

As a check on this theory, three hundred Hollerith cards were punched with numbers corresponding to the three hundred variates of the parent population. The cards were thoroughly shuffled and then placed in a tabulating machine. After twenty-five cards had run through this electric tabulator, their total was recorded. By repeating this procedure one thousand samples were readily obtained and the results are presented below.

TABLE II

Distribution of the Totals of Samples of Twenty-five Variates Selected at Random from the Parent Population of Table I

Class	*Frequency*
120–	6
140–	28
160–	78
180–	179
200–	273
220–	229
240–	124
260–	56
280–	20
300–	7
Total	1000

In this observed distribution it is found that

$M = 215.84$	$\sigma = 30.8505$
$\alpha_3 = .1556\ 56$	$\alpha_5 = 1.39471$
$\alpha_4 = 3.18939$	$\alpha_6 = 15.8603$

The significance of the differences that exist between these functions and the values of M_z, σ_z and $\alpha_{n,z}$ given above will be considered in a subsequent paper.

The unmodified moments, ν, for the preceding observed distribution were corrected for grouping by means of the following formula:

$$\begin{aligned}\mu_n = \nu_n - \binom{n}{2}\frac{1-\frac{1}{k^2}}{12}\nu_{n-2} + \binom{n}{4}\frac{(1-\frac{1}{k^2})(7-\frac{3}{k^2})}{240}\nu_{n-4} \\ - \binom{n}{6}\frac{(1-\frac{1}{k^2})(31-\frac{18}{k^2}+\frac{3}{k^4})}{1344}\nu_{n-6} + \cdots\end{aligned} \tag{14}$$

where k represents the number of different equidistant variates that can appear in each class. In our case, $k = 20$. Sheppard's corrections will appear as a special case of this formula by permitting k to approac infinity. Thus

$$\mu_n = \nu_n - \binom{n}{2}\frac{1}{12}\nu_{n-2} + \binom{n}{4}\frac{7}{240}\nu_{n-4} - \binom{n}{6}\frac{31}{1344}\nu_{n-6} + \cdots ^{*} \tag{15}$$

At first thought one is apt to be surprised in observing that the distribution of samples appearing in Table II is so nearly "normal," whereas the samples were taken from a right-triangular parent population. As an even more extreme case, I may mention that a group of students chose arbitrarily the following most unusual distribution for a parent population:

TABLE III

x	f_x
15	9
3	2
29	43
405	189
1710	37
Total	280

*Compare with formulae (2b), page 94, Handbook of Mathematical Statistics.

and found that the distribution of the totals of 1000 samples of twenty-five variates each was as follows:

TABLE IV

Class	Freq.
5000–	2
7000–	54
9000–	203
11000–	310
13000–	254
15000–	130
17000–	36
19000–	9
21000–	2
Total	1000

As a matter of fact, if r is fifty or greater and s is at least ten times as large as r, the parent population has relatively little control over the shape of the distribution of samples. But before investigating the limit towards which distributions of samples approach in shape, it is well to present a second illustration of the theory so far developed.

Illustration II. Pearson's Hypergeometric Series.

If from a bag containing qs black and ps white balls, r balls are withdrawn without replacements, the chances that the r balls withdrawn will contain 0, 1, 2, . . ., x, . . . r white balls are given by the successive terms of the hypergeometric series

$$(16)\quad \frac{1}{\binom{s}{r}}\left\{\binom{qn}{r}+\binom{qn}{r-1}\ \binom{pn}{1}+\cdots\binom{qn}{r-x}\binom{pn}{x}+\cdots\cdots\binom{pn}{r}\right\}$$

A distribution of this type is equivalent to the simplest case that can arise in accordance with the theory of sampling, that is, by assuming that each variate of the parent population is equal to either zero or one, and that ϖ denotes the proportion of the s variates that have unit value. The moments of the parent population are found as follows:

TABLE V

Parent Population for Hypergeometric (and Binomial) Series

x	f_x	xf_x	$x-M_x=\bar{x}$	$(x-M_x)^n f_x = \bar{x}^n f_x$
0	$(1-p)s$	0	$-p$	$(-1)^n p^n (1-p)s$
1	ps	ps	$1-p$	$p(1-p)^n \cdot s$
Total	s	ps		$p(1-p)s\{(1-p)^{n-1}+(-1)^n p^{n-1}\}$

Therefore

$$(17)\quad \mu_{n:x}=p(1-p)\{(1-p)^{n-1}+(-1)^n p^{n-1}\}=pq\{q^{n-1}+(-1)^n p^{n-1}\},$$

where $(p+q=1)$

In numerical problems this formula should be used ordinarily as it stands, although for algebraic purposes we may use frequently the forms

$$\begin{aligned}
\mu_{1:x} &= 0\\
\mu_{2:x} &= pq = p(1-p)\\
\mu_{3:x} &= pq\ (q^2-p^2) = p(1-p)(1-2p)\\
\mu_{4:x} &= pq\ (q^3+p^3) = p(1-p)(1-3p+3p^2)
\end{aligned}$$

etc.

Using formulae 2, . . ., we may write the moments for the hypergeometric series as follows:

$$\begin{aligned}
\mu_{2:z} &= s\mu_{2:x}\ \{\rho_1-\rho_2\}\\
\mu_{3:z} &= s\mu_{3:x}\ \{\rho_1-3\rho_2+2\rho_3\}
\end{aligned}$$

etc.

or if one prefers

$$\mu_{2:x} = spq\left\{\frac{r}{s} - \frac{r^{(2)}}{s^{(2)}}\right\}$$

$$\mu_{3:x} = spq\,(q^2-p^2)\left\{\frac{r}{s} - 3\,\frac{r^{(2)}}{s^{(2)}} + 2\,\frac{r^{(3)}}{s^{(3)}}\right\}$$

$$\mu_{4:x} = spq\,(q^3+p^3)\left\{\frac{r}{s} - 7\,\frac{r^{(2)}}{s^{(2)}} + 12\,\frac{r^{(3)}}{s^{(3)}} - 6\,\frac{r^{(4)}}{s^{(4)}}\right\}$$
$$+3s^2p^2q^2\left\{\frac{r^{(2)}}{s^{(2)}} - 2\,\frac{r^{(3)}}{s^{(3)}} + \frac{r^{(4)}}{s^{(4)}}\right\}$$

etc.

These will be found equivalent to those given by Pearson*, namely

$$\mu_2 = \frac{\alpha\beta\,(s+\alpha)(s+\beta)}{s^2(s-1)}$$
$$\mu_3 = \frac{\alpha\beta(s+\alpha)(s+\beta)(s+2\alpha)(s+2\beta)}{s^3(s-1)(s-2)}$$
$$\mu_4 = \frac{m_2\,(s^2+m_1 s+m_2)}{s(s-1)(s-2)(s-3)}\Big\{s^4+s^3\,(3m_2+6m_1+1)$$
$$+3s^2(m_1m_2+2m_1^2+2m_2)$$
$$+3s\,m_2\,(m_2+6m_1)+18\,m_2^2\Big\}$$

where

$$\alpha = -r \qquad \beta = -ps$$

$$m_1 = \alpha+\beta \qquad m_2 = \alpha\beta$$

II. Sampling from an Unlimited Supply

Referring to the formula of the first part of this paper, we observe that as s approaches infinity, r remaining finite,

*Lond., Edinburgh and Dublin Phil. Mag., Jan.-June, 1899, page 236.

$$(18)\left\{\begin{aligned}
&M_z = rM_x\\
&\mu_{2:z} = r\mu_{2:x}\\
&\mu_{3:z} = r\mu_{3:x}\\
&\mu_{4:z} = r\mu_{4:x} + 3r^{(2)}\mu^2_{2:x}\\
&\mu_{5:z} = r\mu_{5:x} + 10r^{(2)}\mu_{3:x}\mu_{2:x}\\
&\mu_{6:z} = r\mu_{6:x} + 15r^{(2)}\mu_{4:x}\mu_{2:x} + 10r^{(2)}\mu^2_{3:x} + 15r^{(3)}\mu^3_{2:x}\\
&\mu_{7:z} = r\mu_{7:x} + 21r^{(2)}\mu_{6:x}\mu_{2:x} + 35r^{(2)}\mu_{4:x}\mu_{3:x}\\
&\qquad + 105r^{(3)}\mu_{3:x}\mu^2_{2:x}\\
&\mu_{8:z} = r\mu_{8:x} + 28r^{(2)}\mu_{6:x}\mu_{2:x} + 56r^{(2)}\mu_{5:x}\mu_{3:x} + 35r^{(2)}\mu^2_{4:x}\\
&\qquad + 210r^{(3)}\mu_{4:x}\mu^2_{2:x} + 280r^{(3)}\mu^2_{3:x}\mu_{2:x} + 105r^{(4)}\mu^4_{2:x}
\end{aligned}\right.$$

From these the following equations may be obtained:

$$(19)\left\{\begin{aligned}
&\mu_{2:z} = r\mu_{2:x}\\
&\mu_{3:z} = r\mu_{3:x}\\
&\mu_{4:z} - 3\mu^2_{2:z} = r\{\mu_{4:x} - 3\mu^2_{2:x}\}\\
&\mu_{5:z} - 10\mu_{3:z}\mu_{2:z} = r\{\mu_{5:x} - 10\mu_{3:x}\mu_{2:x}\}\\
&\mu_{6:z} - 15\mu_{4:z}\mu_{2:z} - 10\mu^2_{3:z} + 30\mu^3_{2:z} = r\{\mu_{6:x} - 15\mu_{4:x}\mu_{;}\\
&\qquad - 10\mu^2_{3:x} + 30\mu^3_{2:x}\}\\
&\mu_{7:z} - 21\mu_{5:z}\mu_{2:z} - 35\mu_{4:z}\mu_{3:z} + 210\mu_{3:z}\mu^2_{2:z} = r\{\mu_{7:x}\\
&\qquad - 21\mu_{5x}\mu_{2:x} - 35\mu_{4:x}\mu_{3:x} + 210\mu_{3:x}\mu^2_{2:x}\}\\
&\mu_{8:z} - 28\mu_{6:z}\mu_{2:z} - 56\mu_{5:z}\mu_{3:z} - 35\mu^2_{4:z} + 420\mu_{4:z}\mu^2_{2:z}\\
&\qquad + 560\mu^2_{3:z}\mu_{2:z} - 630\mu^4_{2:z} = r\{\mu_{8:x} - 28\mu_{6:x}\mu_{2:x}\\
&\qquad - 56\mu_{6:x}\mu_{3:x} - 35\mu^2_{4:x} + 420\mu_{4:x}\mu^2_{2:x}\\
&\qquad + 560\mu^2_{3:x}\mu_{2:x} - 630\mu^4_{2:x}\}
\end{aligned}\right.$$

In terms of the standard moments of the distributions these equations become

$$(20)\begin{cases}\alpha_{3:z}=\dfrac{1}{r^{1/2}}\alpha_{3:x}\\[2ex] \alpha_{4:z}-3=\dfrac{1}{r}\left\{\alpha_{4:x}-3\right\}\\[2ex] \alpha_{5:z}-10\alpha_{3:z}=\dfrac{1}{r^{3/2}}\left\{\alpha_{5:x}-10\alpha_{3:x}\right\}\\[2ex] \alpha_{6:z}-15\alpha_{4:z}-10\alpha^{2}_{3:z}+30=\dfrac{1}{r^{2}}\left\{\alpha_{6:x}-15\alpha_{4:x}-10\alpha_{3:x}^{2}+30\right\}\\[2ex] \alpha_{7:z}-21\alpha_{5:z}-35\alpha_{4:z}\alpha_{3:z}+210\alpha_{3:z}=\dfrac{1}{r^{5/2}}\left\{\alpha_{7:x}-21\alpha_{5:x}\right.\\ \qquad\left.-35\alpha_{4:x}\alpha_{3:x}+210\alpha_{3:x}\right\}\\[2ex] \alpha_{8:z}-28\alpha_{6:z}-56\alpha_{5:z}\alpha_{3:z}-35\alpha_{4:z}^{2}+420\alpha_{4:z}+56\alpha_{3:z}^{2}-630\\ =\dfrac{1}{r^{3}}\left\{\alpha_{8:x}-28\alpha_{6:x}-56\alpha_{5:x}\alpha_{3:x}-35\alpha_{4:x}+420\alpha_{4:x}+56\alpha_{3:x}^{2}-630\right\}\end{cases}$$

If, without reference to subscripts, we write

$$(21)\begin{cases}\lambda_2=\mu_2\\ \lambda_3=\mu_3\\ \lambda_4=\mu_4-3\mu_2^2\\ \lambda_5=\mu_5-10\mu_3\mu_2\\ \lambda_6=\mu_6-15\mu_4\mu_2-10\mu_3^2+30\mu_2^3\\ \lambda_7=\mu_7-21\mu_5\mu_2-35\mu_4\mu_3+210\mu_3\mu_2^2\\ \lambda_8=\mu_8-28\mu_6\mu_2-56\mu_5\mu_3-35\mu_4^2+420\mu_4\mu_2^2\\ \qquad\qquad+560\mu_3^2\mu_2-630\mu_2^4\end{cases}$$

the distribution of samples from an unlimited supply is defined, so far as moments through the eighth order are concerned, by the relations

$$(22)\qquad\begin{cases}M_z=rM_x\\[1ex] \lambda_{n:z}=r\lambda_{n:x}\end{cases}$$

Working along a different line of approach, Thiele was the first to realize the importance of these λ functions. He made an extensive study of their unusual properties and was thus both directly and indirectly responsible for many important contributions to the theory of mathematical statistics. These values of λ_i are the so-called "semi-Invariants of Thiele."

Again, we may write

$$3)\begin{cases} \gamma_3 = \alpha_3. \\ \gamma_4 = \alpha_4 - 3 \\ \gamma_5 = \alpha_5 - 10\,\alpha_3 \\ \gamma_6 = \alpha_6 - 15\,\alpha_4 - 10\,\alpha_3^2 \quad + 30 \\ \gamma_7 = \alpha_7 - 21\,\alpha_5 - 35\,\alpha_4\alpha_3 + 210\,\alpha_3 \\ \gamma_8 = \alpha_8 - 28\,\alpha_6 - 56\,\alpha_5\alpha_3 - 35\,\alpha_4^2 + 420\,\alpha_4 + 560\,\alpha_3^2 - 630 \end{cases}$$

and observe that the shape of the distribution of samples is determined by the relation

$$(24) \quad \gamma_{n:\bar{x}} = \frac{1}{p^{n/2-1}} \cdot \gamma_{n:x}$$

which follows from equations (20).

The values γ_i are referred to as the "standardized semi-invariants of Thiele."

If now p be permitted to approach infinity as a limit, we observe that in this limiting situation the *shape* of the distribution of samples is entirely independent of the shape of the parent population, since

$$\lim_{r \to \infty} \gamma_{n:\bar{x}} = 0$$

that is

$$\alpha_{3:\bar{x}} = 0$$

$$\alpha_{4:\bar{x}} - 3 = 0$$

$$\alpha_{5:\bar{x}} - 10\,\alpha_{3:\bar{x}} = 0$$

$$\alpha_{6:\bar{x}} - 15\alpha_{4:\bar{x}} - 10\,\alpha_{3:\bar{x}}^2 + 30 = 0$$

etc.

Thus the limiting distribution, which is called "the Normal Curve," must have the following properties:

$$(25)\qquad \begin{cases} \alpha_{3:x} = 0 \\ \alpha_{4:x} = 1\cdot 3 \\ \alpha_{5:x} = 0 \\ \alpha_{6:x} = 1\cdot 3\cdot 5 \\ \alpha_{7:x} = 0 \\ \alpha_{8:x} = 1\cdot 3\cdot 5\cdot 7 \end{cases}$$

The Theorem of Bernoulli

If p denotes the probability that an event will happen in a single trial and $q = 1 - p$ the probability that it will not happen in that trial, then the probability that the event will happen exactly x times during r trials is, by Bernoulli's Theorem

$$(26)\qquad B_{r:x} = \binom{r}{x} q^{r-x} p^{x}$$

From our point of view we need only regard the problem as one of sampling in which we withdraw samples of r variates from an infinite parent population, in which ,as per Table V, p designates the proportion of the variates which are zero in magnitude—the remaining variates being of unit magnitude. Then since

$$\mu_{n:x} = pq\left\{ q^{n-1} + (-1)^n p^{n-1} \right\}$$

we see from formulae (18) that

$$(27)\qquad \begin{cases} M_x = rp \\ \mu_{2:x} = rpq \\ \mu_{3:x} = rpq\{q^2 - p^2\} \\ \mu_{4:x} = rpq\{q^3 + p^3\} + 3r^{(2)}p^2q^2 \\ \mu_{5:x} = rpq\{q^4 - p^4\} + 10\, r^{(2)}p^2q^2\{q^2 - p^2\} \end{cases}$$

etc.

Poisson's Exponential Binomial Limit

If the probability that each of 1000 individuals die in one year were .5, then the expected number of deaths in such a group for one year would be 500. On the other hand, if the probability that each of 10,000 die in the year were .05 then the expected number of deaths would also be 500. Again $r=100000$ and $p=.005$ or $r=1000000$ and $p=.0005$ would give the same value. If we continue after this fashion to let r approach infinity and p zero, but in such a manner that the product $rp=M$ remains constant, then it can be shown quite readilv that (26) becomes

$$\lim_{\substack{r\to\infty\\ p\to 0\\ rp=M}} B_{r:x}=\frac{e^{-M}M^{x}}{x!} \tag{28}$$

This is known as Poisson's Exponential Binomial Limit. For a Poisson distribution it follows from (27) that

$$(29)\quad\begin{cases}\mu_{2:z}=M_z\\ \mu_{3:z}=M_z\\ \mu_{4:z}=M_z+3M_z^2\\ \mu_{5:z}=M_z+10M_z^2\\ \mu_{6:z}=M_z+25M_z^2+15M_z^3\\ \mu_{7:z}=M_z+56M_z^2+105M_z^3\\ \mu_{8:z}=M_z+119M_z^2+409M_z^3+105M_z^4\end{cases}$$

Substituting these values back in the definitions of the semi-invariants (formulae 21), we observe that for a Poisson distribution

$$\lambda_{n:z}=M_z \qquad (z=2,3,\cdots\cdots,8) \tag{30}$$

Discussion of Results

So far as I know, no general method has been worked out which will permit one to express complex summations, such as those on pages

103, 104, in terms of moments. Moreover, I am unable at present to justify the use of the "sampling polynomials" for the moments of the samples of an order higher than the eighth. Laborious computations have established the fact that the apparent law of the sampling polynomials holds for the first eight moments, and hence we have a simple method at our disposal of writing down expressions for these moments of samples withdrawn from finite parent populations. A study of these sampling polynomials should reveal an entirely different approach to the problem. This is but one of many interesting problems of mathematical statistics that require further investigation.

Although we utilized the results of sampling from a limited supply to obtain corresponding formulae for sampling from an unlimited supply, nevertheless it can be shown that for $s=\infty$ a simple method exists for expressing the moments in terms of the moments, $\mu_{n:x}$, as in formulae (18). Moreover, this law holds for any positive integer, n .

Thus

$$\begin{aligned}\mu_{20:x} = {} & \frac{20!}{20!} r^{(1)} \mu_{20:x} + \frac{20!}{18!\,2!} r^{(2)} \mu_{18:x} \mu_{2:x} + \cdots\cdots \\ & + \cdots \frac{20!}{9!7!4!} r^{(3)} \mu_{9:x} \mu_{7:x} \mu_{4:x} + \cdots\cdots \\ & + \cdots \frac{20!}{(4!)^2 (3!)^4} r^{(6)} \frac{\mu_{4:x}^2 \mu_{3:x}^4}{2!\,4!} + \cdots\end{aligned}$$

Since formulae, such as (3a) and (4a) are based on multinomial considerations, the rule for writing down the values of $\mu'_{n:x}$ is valid for any value of n , when $s=\infty$

Proceeding after this fashion, one can show that corresponding to formulae (25) one can write for the limiting distribution, referred to as the Normal Curve,

$$(31) \qquad \begin{cases} \alpha_{2n+1:x} = 0 \\ \alpha_{2n:x} = \dfrac{(2n)!}{2^n (n!)} \end{cases}$$

And since the function

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}^{2}} \qquad (32)$$

satisfies the above conditions, we say that (32) is the equation of the Normal Curve. In the Theory of Least Squares this equation is usually* developed on the so-called Hagen's hypothesis, that is "An error is the algebraic sum of an indefinitely great number of small elementary errors which are all equal, and each of which is equally likely to be positive or negative."

From the results that we have obtained it appears that it is not necessary to impose the restrictions that the elementary errors are all equal and that positive and negative values are equally likely. It is necessary only that

(1) the number of elementary errors be infinite, although of an order less than that of the number of errors in the parent population.

(2) the errors be independent. This restriction is really involved in our assumption that in evaluating summations, each of the s variates of the parent population occurs exactly as many times as every other variate.

Otherwise, the limiting shape of the distribution of samples is independent of the shape of the parent distribution. The fact that tables II and IV, arising from parent distributions that are so extremely abnormal, exhibit distributions of samples that are fairly normal, seems to bear out our point in spite of the fact that we employed in each instance a small value of r, i. e. twenty-five.

*See Merriman's Method of Least Squares. John Wiley and Sons. New York City.