EDITORIAL

The Interdependence of Sampling and Frequency Distribution Theory

The object of the theory of sampling is to describe the phenomena exhibited by all the samples that can possibly arise from a parent population of known characteristics. In some cases the desired description can be obtained directly by employing elementary operations of combination theory, in others it is either expedient or necessary to use the indirect attack of the statistical theory of sampling. These two methods are quite different in application, and it is advisable to illustrate the respective peculiarities of the two methods.

Example 1. An auction bridge hand may be regarded as a single sample withdrawn from a parent population of 52 cards. The number of different hands that can be selected equals the number of combinations of 52 things taken 13 at a time, namely, $\binom{52}{13} = 635\ 013\ 559\ 600$. Of these

$$(1) \cdots\cdots\cdots f(z) = \binom{39}{13-z}\binom{13}{z}$$

will contain exactly z cards of any specified suit. Therefore if in this expression we successively place z equal to 0, 1, 2, . . . 13 we shall obtain the frequency of all possible samples ranked according to the number of cards of the specified suit contained in each sample. The results are presented in the following table.

TABLE I

z	$f(z)$	$P_z = f(z)/N$
0	8 122 425 444	.01279
1	50 840 366 668	.08006
2	130 732 371 432	.20587
3	181 823 183 256	.28633
4	151 519 319 380	.23861
5	79 181 063 676	.12469
6	26 393 687 892	.04156
7	5 598 661 068	.00882
8	740 999 259	.00117
9	58 809 465	.00009
10	2 613 754	.00000
11	57 798	.00000
12	507	.00000
13	1	.00000
Total	635 013 559 600	.99999

In this illustration, combination theory has yielded a perfect solution. The frquencies are exact, and the sum of the frequencies between any two limits may likewise be obtained exactly by a simple addition.

Example 2. The bidding strength of hands in auction bridge is often approximated by counting each Jack, Queen, King and Ace as 1, 2, 3 and 4 points, respectively. The total count of a single hand may range, therefore from 0 to 37 inclusive. Required the frequency distribution of all possible hands when they

are classified according to count.

Unlike the preceding problem, we cannot obtain a simple expression for the general term, f_z, of the required distribution. But after rather involved computations the following solution may be obtained:

TABLE II

Count z	Frequency $f(z)$	Count z	Frequency $f(z)$
0	2 310 789 600	19	6 579 838 440
1	5 006 710 800	20	4 086 538 404
2	8 611 542 576	21	2 399 507 844
3	15 636 342 960	22	1 333 800 036
4	24 419 055 136	23	710 603 628
5	32 933 031 040	24	354 993 864
6	41 619 399 184	25	167 819 892
7	50 979 441 968	26	74 095 248
8	56 466 608 128	27	31 157 940
9	59 413 313 872	28	11 790 760
10	59 723 754 816	29	4 236 588
11	56 799 933 520	30	1 396 068
12	50 971 682 080	31	388 196
13	43 906 944 752	32	109 156
14	36 153 374 224	33	22 360
15	28 090 962 724	34	4 484
16	21 024 781 756	35	624
17	14 997 080 848	36	60
18	10 192 504 020	37	4
		Total	635 013 559 600

Example 3. If the mean and the standard deviation of the weights of a group of 200,000 men be 140 lbs. and 20 lbs., respectively, and if in addition it be known that the higher standard moments of this distribution be

$$\alpha_{3:x} = .5 \qquad \alpha_{5:x} = 4.43$$

$$\alpha_{4:x} = 3.17 \qquad \alpha_{6:x} = 17.97,$$

what is the chance that the mean weight of 1000 men chosen at random from the 200,000 will exceed 141 pounds?

It is clear that it would be physically impossible to solve this problem by employing a direct attack by combination theory, even though the weights of each of the 200,000 men were available. Moreover, it is likewise evident that in statistical problems corresponding to the illustrations of examples 1 and 2, the number of individuals in both the parent population and each sample is considerably larger than 52 and 13 respectively, and consequently the calculation of either a single frequency or the sum of any large group of consecutive frequencies by the direct method is quite out of the question.

Let us now consider the three examples above from the point of view of the indirect attack. The parent populations for the first two examples may be interpreted as

Variates	x	0	1
Frequencies	$f(x)$	39	13

and

Variates	x	0	1	2	3	4
Frequencies	$f(x)$	36	4	4	4	4

respectively.

For the first, the mean is at $x = 1/4$, and the moments about the mean of the parent population are obviously

$$\mu_{n:x} = \frac{13}{4^n}\left[3^n + 3(-1)^n\right]$$

For the second, tne mean is at $x = 10/13$, and correspondingly the moments of this parent population are

$$\mu_{n:x} = \frac{1}{13^n}\left[(-10)^n + 3^n + 16^n + 29^n + 42^n\right]$$

If s and r denote the number of individuals in the parent population and each sample respectively, then the moments of the distribution of all samples that can arise from this parent population may be obtained from those of the parent population by means of the relations

$$(2)^1 \begin{cases} M_{\bar{x}} = r \cdot M_x \\ \mu_{2:\bar{x}} = \mu_{2:x}\, s(\rho_1 - \rho_2) \\ \mu_{3:\bar{x}} = \mu_{3:x}\, s(\rho_1 - 3\rho_2 + 2\rho_3) \\ \mu_{4:\bar{x}} = \mu_{4:x}\, s(\rho_1 - 7\rho_2 + 12\rho_3 - 6\rho_4) + 3\mu_{2:x}^2\, s^2(\rho_2 - 2\rho_3 + \rho_4) \\ \mu_{5:\bar{x}} = \mu_{5:x}\, s(\rho_1 - 15\rho_2 + 50\rho_3 - 60\rho_4 + 24\rho_5) \\ \qquad + 10\mu_{3:x}\mu_{2:x}\, s^2(\rho_2 - 4\rho_3 + 5\rho_4 - 2\rho_5) \\ \mu_{6:\bar{x}} = \mu_{6:x}\, s(\rho_1 - 31\rho_2 + 180\rho_3 - 390\rho_4 + 360\rho_5 - 120\rho_6) \\ \qquad + 15\mu_{4:x}\mu_{2:x}\, s^2(\rho_2 - 8\rho_3 + 19\rho_4 - 18\rho_5 + 6\rho_6) \\ \qquad + 10\mu_{3:x}^2\, s^2(\rho_2 - 6\rho_3 + 13\rho_4 - 12\rho_5 + 4\rho_6) \\ \qquad + 15\mu_{2:x}^3\, s^3(\rho_3 - 3\rho_4 + 3\rho_5 - \rho_6) \end{cases}$$

where

$$\rho_i = \frac{r(r-1)(r-2)\cdots \text{to } i \text{ factors}}{s(s-1)(s-2)\cdots \text{to } i \text{ factors}}$$

Since the moments $\mu_{n:x}$ for each of these three examples are now known, and according to the conditions of the problems the values of (r , s) are (13, 52), (13, 52), and (1000, 200000) respectively, it follows that the moments of the desired distributions of samples are as follows:

Function	Example 1	Example 2	Example 3
M_x	13/4	10	M_x =140 lbs.
$\mu_{2:x}$	507/272	290/17	σ_x = .630874 lbs.
$\mu_{3:x}$	6591/13600	288/17	$\omega_{3:x}$ = .0156927.
$\mu_{4:x}$	53591421/5331200	17441114/29155	$\omega_{4:x}$ = 3.0001357
$\mu_{5:x}$	9339447/1066240	2262240/833	$\omega_{5:x}$ = .1569051
$\mu_{6:x}$	71781968037/801812480	2684384074/39151	$\omega_{6:x}$ = 15.026638

It will be observed that the indirect procedure has yielded the moments of the required distributions rather than their frequency functions, and the next step therefore is to obtain with the aid of these moments approximate expressions for the desired frequency functions. In this connection it should be borne in mind that we are not concerned with questions regarding the probable errors of the moments which we are employing, since the moments computed for the distributions of samples are necessarily exact, and their probable errors are therefore zero. For

[1]See Annals, Vol. I, page 104.

this reason arguments tending to limit the number of terms that may be employed in either a Gram-Charlier series, or in the denominator of Pearson's differential equation are not to the point so far as our illustrations are concerned. These remarks hold even for the third example, since if the moments of the parent population are as given, then the moments of the distribution of samples may be determined with any desired degree of accuracy.

Since it is evident that the solution of our problems now depends upon our obtaining approximate expressions for these distributions whose moments are known, we shall at this point develop a general method of representing discrete distributions which is essentially due to the researches of Charlier. Although the results that we shall obtain are practically those that have also been obtained by Gram, Edgeworth and others, the method that we shall employ is that used by Charlier in "Die Strenge Form des Bernoullischen Theorems."

Let $f(x)$ be the frequency function for a discrete distribution ranging from $x = \ell_1$ to $x = \ell_2$. If the ordinates be equidistant at intervals of h, the total frequency of the distribution is

$$(3)\quad N = f(\ell_1) + f(\ell_1 + h) + \cdots + f(x_0 - h) + f(x_0) + f(x_0 + h) + \cdots + f(\ell_2) = \sum_{x=\ell_1}^{\ell_2} f(x).$$

where our interest is focused on a typical ordinate at $x = x_0$. If we now set up the function

$$\sum_{x=\ell_1}^{\ell_2} f(x) \cdot e^{x\omega i} = f(x) + f(x_0 + h) \cdot e^{(x_0+h)\omega i} + \cdots + f(\ell_2) e^{\ell_2 \omega i} + f(x_0 - h) e^{(x_0-h)\omega i} + \cdots + f(\ell_1) e^{\ell_1 \omega i}$$

where $i = \sqrt{-1}$, and multiply each side by $e^{-x_0 \omega i}$ so that

$$e^{-x_0\omega i} \sum_{x=\ell_1}^{\ell_2} f(x) e^{x\omega i} = f(x_0) + f(x_0+h) \cdot e^{h\omega i} + f(x_0+2h) \cdot e^{2h\omega i} + \cdots$$

$$+ f(\ell_2) \cdot e^{(\ell_2 - x_0)\omega i} + f(x_0-h) \cdot e^{-h\omega i} + f(x_0-2h) \cdot e^{-2h\omega i} + \cdots + f(\ell_1) e^{(\ell_1 - x_0)\omega i}$$

we obtain by integrating both members with respect to ω between the limits $\omega = \frac{-\pi}{h}$ and $\omega = \frac{\pi}{h}$

$$\int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} e^{-x_0\omega i} \left\{ \sum_{x=\ell_1}^{\ell_2} f(x) \cdot e^{x\omega i} \right\} d\omega = \int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} f(x_0)\, d\omega,$$

since the integral of all other terms of the right hand member will vanish as follows:

$$\int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} f(x_0 + mh) \cdot e^{mh\omega i} d\omega = f(x_0 + mh) \cdot$$

$$\int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} \left[\cos mh\omega + i \sin mh\omega \right] d\omega = 0$$

(m is an integer.)

It follows therefore that

$$(4) \qquad f(x_0) = \frac{h}{2\pi} \int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} e^{-x_0\omega i} \left\{ \sum_{x=\ell_1}^{\ell_2} f(x) \cdot e^{x\omega i} \right\} d\omega$$

Moreover, since

$$e^{-a\omega i}+e^{-(a+h)\omega i}+\cdots+e^{-b\omega i}=e^{\frac{-(b+h)\omega i}{}}\frac{-a\omega i}{e^{-h\omega i}-1}$$

we see that the sum of all the consecutive frequencies from $x=a$ to $x=b$ may be expressed as the definite integral

$$(5)\quad \sum_{x=a}^{b} f(x)=\frac{h}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \frac{e^{-(b+h)\omega i}-e^{-a\omega i}}{e^{-h\omega i}-1}\left\{\sum_{x=\ell_1}^{\ell_2} f(x)\cdot e^{x\omega i}\right\}d\omega$$

The changing of the order of integration is permitted since the limits are all finite.

Ordinarily frequency distributions are expressed as developments of the integral (4), and the sums of consecutive frequencies obtained by applying the Euler-Maclaurin Sum-Formula to these results. It seems at first sight that it might be well to place a little more emphasis upon the evaluation of (5), since this as it stands affords an exact expression for the sum of any group of consecutive frequencies. For the case of continuous variates we need only permit h to approach zero, replace the sign of summation by the sign of integration, etc., and after justifying the change in the order of integration for the resulting infinite limits obtain

$$(6)\quad \int_a^b f(x)dx=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-b\omega i}-e^{-a\omega i}}{-\omega i}\cdot\left\{\int_{\ell_1}^{\ell_2} f(x)\cdot e^{x\omega i}dx\right\}d\omega$$

We shall now attempt to evaluate the definite integral (4). Let us first observe that the quantity within the parenthesis is a function of ω, since the finite integration with respect to x and the subsequent replacing of x by the limits will cause this distribution variable to disappear.

For reasons which will develop later, let us write

$$\sum_{x=\ell_1}^{\ell_2} f(x)e^{x\omega i} = e^{b_1(\omega i)+b_2\frac{(\omega i)^2}{2!}} \sum_{x=\ell_1}^{\ell_2} f(x)\cdot e^{(x-b_1)\omega i - \frac{b_2(\omega i)^2}{2!}}$$

If in Leibnitz' formula

$$D^n u.v = u\cdot D^n v + \binom{n}{1} Du\cdot D^{n-1}v + \binom{n}{2} D^2 u\cdot D^{n-2}v + \cdots$$

we place $u = e^{\frac{bz^2}{2}}$ and $v = e^{az}$, and note that

$$D^{2n+1} e^{\frac{bz^2}{2}}\Big]_{z=0} = 0$$

$$D^{2n} e^{\frac{bz^2}{2}}\Big]_{z=0} = \frac{(2n)!}{2^n\cdot n!} b^n$$

then

$$(7)\quad D^n e^{az+\frac{bz^2}{2}}\Big]_{z=0} = a^n + \frac{n^{(2)}}{2\cdot 1!}a^{n-2}b + \frac{n^{(4)}}{2^2\cdot 2!}a^{n-4}b^2 + \frac{n^{(6)}}{2^3\cdot 3!}a^{n-6}b^3 + \cdots$$

where $n^{(i)} = n(n-1)(n-2)\cdots$ to i factors.

Thus we may write

$$\sum_{x=l_1}^{l_2} f(x)\cdot e^{(x-b_1)\omega i - b_2\frac{(\omega i)^2}{2}} = N\left[c_0 + c_1(\omega i) + c_2\frac{(\omega i)^2}{2!} + c_3\frac{(\omega i)^3}{3!} + \cdots\right]$$

and employing the notation

$$\sum_{x=l_1}^{l_2}(x-b_1)^n f(x) = N\mu_n'$$

we obtain from (7)

$$(8)\quad \begin{cases} c_0 = 1 \\ c_1 = \mu_1' \\ c_2 = \mu_2' - b_2 \\ c_3 = \mu_3' - 3b_2\mu_1' \\ c_4 = \mu_4' - 6b_2\mu_2' + 3b_2^2 \\ \quad\cdots\cdots\cdots \\ c_n = \mu_n' - \frac{n^{(2)}}{2\cdot 1!}b_2\mu_{n-2}' + \frac{n^{(4)}}{2^2 2!}b_2^2\mu_{n-4}' - \frac{n^{(6)}}{2^3 3!}b_2^3\mu_{n-6}' + \cdots \end{cases}$$

Formula (4) may therefore be written, dropping the subscript on x_0

$$(9)\quad f(x) = N\frac{h}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{-(x-b_1)\omega i - b_2\frac{(\omega i)^2}{2}}$$

$$\left[1 + c_1(\omega i) + c_2\frac{(\omega i)^2}{2} + \cdots\right] d\omega$$

Placing

$$(10)\qquad \Theta(x) = \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{-(x-b_1)\omega i - \frac{b_2\omega^2}{2}}\, d\omega$$

it follows that the nth derivative with respect to x is

$$(11)\qquad \Theta^{(n)}(x) = \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} (-\omega i)^n \cdot e^{-(x-b_1)\omega i - \frac{b_2\omega^2}{2}}\, d\omega\ ;$$

so finally·

$$(12)\qquad f(x) = N\cdot h\left[\Theta(x) - \frac{c_1}{1!}\Theta^{(1)}(x) + \frac{c_2}{2!}\Theta^{(2)}(x) - \frac{c_3}{3!}\Theta^{(3)}(x) + \cdots\right]$$

Let us now investigate the function $\Theta(x)$.

$$\Theta(x) = \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{-b_2\omega^2/2}\cdot\left[\cos(x-b_1)\omega - i\sin(x-b_1)\omega\right] d\omega$$

$$= \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{-b_2\omega^2/2}\cos(x-b_1)\omega\, d\omega$$

[since $e^{-b_2\omega^2/2}\sin(x-b_1)\omega$ is an odd function of ω]

$$= \frac{1}{\pi}\int_0^{\infty} e^{-b_2\omega^2/2}\cos(x-b_1)\omega\, d\omega$$

$$- \frac{1}{\pi}\int_{\frac{\pi}{h}}^{\infty} e^{-b_2\omega^2/2}\cos(x-b_1)\omega\, d\omega$$

$$= \frac{1}{\sqrt{2\pi b_2}} e^{-\frac{(x-b_1)^2}{2b_2}} - \frac{1}{\pi}\int_{\frac{\pi}{h}}^{\infty} e^{-b_2\omega^2/2}\cos(x-b_1)\omega\, d\omega$$

$$= \phi(x) - R_0 .$$

$$\left[\int_0^{\infty} e^{-a^2x^2}\cos mx\, dx = \frac{\sqrt{\pi}}{a} e^{-m^2/4a^2}\right]$$

Likewise we may write

$$\theta^{(n)}(x) = \phi^{(n)}(x) - R_n , \qquad R_n < \frac{1}{\pi}\int_{\frac{\pi}{h}}^{\infty} \omega^n e^{-b_2\omega^2/2}\, d\omega$$

By successive integration by parts it can be shown that

$$\int x^n e^{-\frac{x^2}{2}} dx = -e^{-\frac{x^2}{2}}\left\{x^{n-1} + (n-1)x^{n-3}\right.$$

$$(13)\quad \left. + (n-1)(n-3)x^{n-5} + \cdots + (n-1)(n-3)\cdots(n-2i+3)x^{n-2i+1}\right\} + R_i ,$$

$$R_i = (n-1)(n-3)\cdots(n-2i+1)\int x^{n-2i} e^{-\frac{x^2}{2}} dx$$

so we have that

$$(14)\quad R_n < \frac{1}{b_2}\left(\frac{\pi}{h}\right)^{n-1} e^{-\frac{b_2}{2}\left(\frac{\pi}{h}\right)^2}\left[1+\frac{n-1}{b_2}\left(\frac{h}{\pi}\right)^2+\frac{(n-1)(n-3)}{b_2^2}\left(\frac{h}{\pi}\right)^4+\cdots\right]$$

So far we have said nothing concerning the values of the parameters b_1 and b_2. Referring to formula (8) it is seen that if the origin of x be taken at the mean of the distribution in question, and b_2 equal the second moment about the mean of this distribution, $c_1 = c_2 = 0$, and consequently if the values of R_n may be neglected, the equation of the distribution expressed in standard units becomes

$$(15)\quad f(x) = N\frac{h}{\sigma}\left\{\phi(t) - \frac{A_3}{3!}\phi^{(3)}(t) + \frac{A_4}{4!}\phi^{(4)}(t) - \frac{A_5}{5!}\phi^{(5)}(t) + \cdots\right\}$$

where $t = \frac{x-b_1}{\sqrt{b_2}} = \frac{x-M}{\sigma}$, $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$, and

$$(16)\quad \begin{cases} A_3 = \alpha_3 \\ A_4 = \alpha_4 - 3 \\ A_5 = \alpha_5 - 10\alpha_3 \\ A_6 = \alpha_6 - 15\alpha_4 + 30 \\ \quad\cdots\cdots \\ A_n = \alpha_n - \frac{n^{(2)}}{2\cdot 1!}\alpha_{n-2} + \frac{n^{(4)}}{2^2\cdot 2!}\alpha_{n-4} - \frac{n^{(6)}}{2^3\cdot 3!}\alpha_{n-6} + \cdots \end{cases}$$

By employing the Euler-Maclaurin Sum-Formula we can write

$$f(a)+f(a+h)+f(a+2h)+\cdots+f(b-h)+f(b)$$

(17)

$$=N\left[\int\phi(t)\,dt-A_0'\phi(t)+A_1'\phi^{(1)}(t)-A_2'\phi^{(2)}(t)+A_3'\phi^{(3)}(t)\cdots\right]_{\frac{a-M}{\sigma}}^{\frac{b+h-M}{\sigma}}$$

where

(18)

$$\begin{cases} A_0' = \dfrac{h}{2\sigma} \\ A_1' = \dfrac{h^2}{12\sigma^2} \\ A_2' = \dfrac{\alpha_3}{6} \\ A_3' = \dfrac{\alpha_4-3}{24}+\dfrac{h}{\sigma}\cdot\dfrac{\alpha_3}{12}-\dfrac{h^4}{720\sigma^4} \\ A_4' = \dfrac{\alpha_5-10\alpha_3}{120}+\dfrac{h}{\sigma}\,\dfrac{\alpha_4-3}{48}+\dfrac{h^2}{\sigma^2}\cdot\dfrac{\alpha_3}{72} \\ A_5' = \dfrac{\alpha_6-15\alpha_4+30}{720}+\dfrac{h}{\sigma}\,\dfrac{\alpha_5-10\alpha_3}{240}+\dfrac{h^2}{\sigma^2}\cdot\dfrac{\alpha_4-3}{288}+\dfrac{h_6}{30240\sigma^6} \end{cases}$$

In some cases it may be more convenient to employ a mean and a standard deviation of the generating function that differs somewhat from that of the distribution for which the representation is desired. In this event the coefficients of the first and second derivatives in (15) will not vanish. However, the extra effort

expended in increasing the number of significant terms may be more than offset by the fact that a rather arbitrary choice in the values of b_1 and b_2 may result in simpler values for

$$t = \frac{x - b_1}{\sqrt{b_2}}$$

which in turn may occasionally eliminate difficult interpolations when dealing with tabulations of the generating function and its derivatives.

Formulae (17) and (18) may be regarded as a sort of apology for the fact that the definite integral of formula (5) has never been developed. The need of a satisfactory expression for the sum of any number of consecutive variates is indeed acute.

By permitting h in the foregoing theory to approach zero, one can obtain corresponding formulae for the ordinates and areas of distributions of continuous variates. However, it should be noted that for this case the limits for the integrals in the vicinity of formula (4) are now

$$\lim_{h \to 0} \frac{\pi}{h} = \infty$$

and consequently the changing of the order of integration must be justified.

In conclusion we may state:

I. Answers to problems of statistical sampling are usually expressed as finite or infinitesimal integrals under a function whose moments *only* are known. If known, the function is generally of but little value.

II. It is necessary to approximate the desired integrals by employing frequency functions.

III. Present methods are unsatisfactory from the point of view that remainder or limit of error terms are not available. The χ^2 test, though helpful, does not meet the issue in question.

H. C. Carver.