EDITORIAL: A. L. O'TOOLE

ON A BEST VALUE OF R IN SAMPLES OF R FROM A FINITE POPULATION OF N.

In recent years the problem of finding the moment coefficients for samples of n drawn from a finite population of N has been of interest to so many writers[1] that it seems worthwhile to make a few further observations[2] concerning these moment coefficients—particularly with respect to their dependence on n. In many instances the value of n to be used is at the discretion of the investigator and he would like to know if there is one value of n which is better than any other. An answer to that question will be given here.

---

[1] H. C. Carver, On the fundamentals of the theory of sampling, Annals of Mathematical Statistics, Vol. I, No. 1, pp. 101-121; Vol. I, No. 3, pp. 260-274.

C. C. Craig, An application of Thiele's semi-invariants to the sampling problem, Metron, Vol. VII, No. 4, 1928, pp. 3-74.

R. A. Fisher, Moments and product moments of sampling distributions, Proc. London Math. Soc., Series 2, xxix, 1929, pp. 309-321; xxx, 1929, pp. 199-238.

L. Isserlis, On a formula for the product moment coefficients of any order of a normal frequency distribution in any number of variables. Biometrika, xii, 1918-19, pp. 134-139.

P. R. Rider, Moments of moments, Proc. of the National Academy of Sciences, Vol. 15, 1929, pp. 430-434.

H. E. Soper, Sampling moments of samples of n units each drawn from an unchanging sampled population, from the point of view of semi-invariants, Journal of the Royal Statistical Soc., Vol. 93, 1930, pp. 104-114.

A. A. Tchouproff, On the mathematical expectation of the moments of frequency distributions, Biometrika, xii, 1918-19, pp. 140-169 and 184-210; xiii, 1920-21, pp. 283-295.

A. L. O'Toole, On symmetric functions and symmetric functions of symmetric functions, Annals of Mathematical Statistics, Vol. II, No. 2, May 1931, pp. 102-149. See Chapter III.

[2] These observations arose as a result of some very far-reaching suggestions on the theory of sampling made by Professor Carver, during recent conversations with the writer.

The differential operator method developed by this writer[3] for finding the moment coefficients not only was a very simple method but had the added advantage of leading directly to some theorems whose generality had not been established previously.

Using the notation of the previous paper let the finite parent population of N be composed of the N variates $x_1, x_2, x_3, \ldots, x_N$. From this population draw all of the ${}_NC_n$ different samples and let $z_i = \sum^{n:i} x$, $i = 1, 2, 3, \ldots, {}_NC_n$, where $\sum^{n:i} x$ designates the sum of the n values of x which appear in the i^{th} sample. With this notation it has been shown in the paper cited that

$$(1) \qquad S_{t:z} = t! \sum \frac{P_i^I \cdot P_j^J \cdot P_k^K \cdots S_{i:x}^I \, S_{j:x}^J \, S_{k:x}^K \cdots}{(i!)^I (j!)^J (k!)^K \cdots (I!)(J!)(K!)\cdots}$$

where $S_{t:z} = \sum_{i=1}^{{}_NC_n} z_i^t$, $\qquad t = 1, 2, 3, \ldots$

and $S_{w:x} = \sum_{i=1}^{N} x_i^w$, $\qquad w = 1, 2, 3, \ldots$

The summation in (1) is to be taken over terms such that $Ii + Jj + Kk + \ldots = t$ where $I, J, K, \ldots, i, j, k, \ldots$ are positive integers, and where P_m is obtained from the m^{th} sampling polynomial $P_m(p)$ by replacing the exponents of the polynomial by corresponding subscripts.

$$(2) \qquad P_m(p) = \sum_{i=0}^{m-1} (-1)^i \left(\Delta^i 1^{m-1}\right) p^{i+1}, \quad \text{or}$$

$$(3) \qquad P_m(p) = \frac{d^m}{dx^m} \log \left(pe^x + 1 - p\right)\Big]_{x=0}.$$

[3] Loc. cit.

In particular $P_1 = p_1$, $P_2 = p_1 - p_2$, $P_3 = p_1 - 3p_2 + 2p_3$

$$P_4 = p_1 - 7p_2 + 12p_3 - 6p_4$$

$$P_5 = p_1 - 15p_2 + 50p_3 - 60p_4 + 24p_5$$

$$P_6 = p_1 - 31p_2 + 180p_3 - 390p_4 + 360p_5 - 120p_6$$

$$P_7 = p_1 - 63p_2 + 602p_3 - 2100p_4 + 336p_5 - 2520p_6 + 720p_7$$

where $p_K = {}_{N-K}C_{r-K}$, $\quad K \leq r$.

It must be kept in mind that the multiplication of these operators is symbolic. For example, to find $P_i^I P_j^J$ first multiply the polynomials $P_i^I(p)$ and $P_j^J(p)$ by ordinary multiplication and then the result when the exponents in this product are replaced by corresponding subscripts is $P_i^I P_j^J$.

Since in this paper it is desired to consider moments rather than power sums, replace $S_{t:z}$ by $({}_NC_r)\mu'_{t:z}$ and $S_{w:x}$ by $N\mu'_{w:x}$. Then (1) becomes, after dividing by ${}_NC_r$,

$$(4) \qquad \mu'_{t:z} = \frac{t!}{{}_NC_r} \sum \frac{P_i^I P_j^J P_k^K \cdots}{(i!)^I (j!)^J (k!)^K \cdots} \frac{N^{I+J+K}}{I! \cdot J! \cdot K!} \mu'^{I}_{i:x} \mu'^{J}_{j:x} \mu'^{K}_{k:x} \cdots .$$

Now $p_k = {}_{N-k}C_{r-K}$, $\quad k \leq r$, hence,

$$\frac{p_k}{{}_NC_r} = \frac{r(r-1)(r-2)\cdots(r-k+1)}{N(N-1)(N-2)\cdots(N-k+1)} .$$

Substituting this value for each $p_k/{}_NC_r$ in (4) the result is

Equations 5.

$$\mu'_{1:z} = r\,\mu'_{1:x}$$

$$\mu'_{2:z} = \frac{r}{N-1}\left[N(r-1)\mu'^{2}_{1:x} + (N-r)\mu'_{2:x}\right]$$

$$\mu'_{3:z} = \frac{r}{(N-1)(N-2)}\left[N^2(r-1)(r-2)\mu'^{3}_{1:x} + 3N(r-1)(N-r)\mu'_{1:x}\mu'_{2:x} + (N-r)(N-2r)\mu'_{3:x}\right]$$

$$\mu'_{4:z} = \frac{r}{(N-1)(N-2)(N-3)}\Big[\; N^3(r-1)(r-2)(r-3)\,\mu'^{4}_{1:x}$$
$$+ 6N^2(r-1)(r-2)(N-r)\,\mu'^{2}_{1:x}\,\mu'_{2:x}$$
$$+ 4N(r-1)(N-r)(N-2r+1)\,\mu'_{1:x}\,\mu'_{3:x}$$
$$+ 3N(r-1)(N-r)(N-r-1)\,\mu'^{2}_{2:x}$$
$$+ (N-r)(N^2+6Nr+6r^2+N)\,\mu'_{4:x}\Big]$$

etc.

Now let $N = ar$. Then

Equations 6.

$$\mu'_{1:z} = r\,\mu'_{1:x}$$

$$\mu'_{2:z} = \frac{r^2}{ar-1}\Big[a(r-1)\,\mu'^{2}_{1:x} + (a-1)\,\mu'_{2:x}\Big]$$

$$\mu'_{3:z} = \frac{r^3}{(ar-1)(ar-2)}\Big[a^2(r-1)(r-2)\,\mu'^{3}_{1:x} + 3a(r-1)(a-1)\,\mu'_{1:x}\,\mu'_{2:x}$$
$$+ (a-1)(a-2)\,\mu'_{3:x}\Big]$$

$$\mu'_{4:z} = \frac{r^3}{(ar-1)(ar-2)(ar-3)}\Big[\; a^3r(r-1)(r-2)(r-3)\,\mu'^{4}_{1:x}$$
$$+ 6a^2r(r-1)(r-2)(a-1)\,\mu'^{2}_{1:x}\,\mu'_{2:x}$$
$$+ 4a(r-1)(a-1)\{r(a-2)+1\}\,\mu'_{1:x}\,\mu'_{3:x}$$
$$+ 3a(r-1)(a-1)\{r(a-1)-1\}\,\mu'^{2}_{2:x}$$
$$+ (a-1)\{r(a^2-6a+6)+a\}\,\mu'_{4:x}\Big]$$

etc.

A partial check at this point is to note that for $a = 1$ only the first term of each of these moment coefficients remains.

Let $a = 2$. Then the above moment coefficients become

$$(7)\quad \begin{cases} \mu'_{1:z} = n\,\mu'_{1:x} \\ \mu'_{2:z} = \dfrac{n^2}{2n-1}\left[2(n-1)\mu'^{2}_{1:x} + \mu'_{2:x}\right] \\ \mu'_{3:z} = \dfrac{n^3}{2n-1}\left[2(n-2)\mu'^{3}_{1:x} + 3\mu'_{1:x}\mu'_{2:x}\right] \\ \mu'_{4:z} = \dfrac{n^3}{(2n-1)(2n-3)}\left[4n(n-2)(n-3)\mu'^{4}_{1:x} + 12n(n-2)\mu'^{2}_{1:x}\mu'_{2:x} + 4\mu'_{1:x}\mu'_{3:x} + 3(n-1)\mu'^{2}_{2:x} - \mu'_{4:x}\right] \\ \mu'_{5:z} = \dfrac{n^4}{(2n-1)(2n-3)}\left[n(n-3)(n-4)\mu'^{5}_{1:x} + 20n(n-3)\mu'^{3}_{1:x}\mu'_{2:x} - 20\mu'^{2}_{1:x}\mu'_{3:x} + 15(n-1)\mu'_{1:x}\mu'^{2}_{2:x} - 5\mu'_{1:x}\mu'_{4:x}\right] \end{cases}$$

etc.

It is observed that when $a = 2$, i.e. when $n = 2k$, the moment coefficient $\mu'_{3:z}$ is independent of the moment coefficient $\mu'_{3:x}$. Also $\mu'_{5:z}$ is independent of $\mu'_{5:x}$. But one must not assume that all the odd moment coefficients of z are independent of the corresponding odd moment coefficients of x. For $\mu'_{7:z}$ is not independent of $\mu'_{7:x}$ as is seen by evaluating P_7 which is the coefficient of $\mu'_{7:x}$ in the expression for $\mu'_{7:z}$.

So far the moments considered have been the moments of x with respect to the origin from which x is measured and the moments of z with respect to the origin from which z is measured. Consider now the moments of x about the mean value of x and the moments of z about the mean value of z. That is let

$$\bar{z}_i = z_i - M_z, \qquad i = 1, 2, 3, \cdots, {}_N C_n;$$

$$\bar{x}_i = x_i - M_x, \qquad i = 1, 2, 3, \cdots, N.$$

Then

$$\bar{z}_i = z_i - M_z = \sum^{n:i} x - n M_x \text{ since } M_z = n M_x \text{ by (5)},$$
$$= \sum^{n:i} (x - M_x) = \sum^{n:i} \bar{x}.$$

e.g. $$\bar{z}_1 = z_1 - M_z = x_1 + x_2 + x_3 + \cdots + x_n - n M_x$$
$$= (x_1 - M_x) + (x_2 - M_x) + (x_3 - M_x) + \cdots + (x_n - M_x)$$
$$= \bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \cdots + \bar{x}_n$$
$$= \sum^{n:1} \bar{x}.$$

Hence it is clear that $\bar{z}$ is the same function of $\bar{x}$ as z is of x. In other words $\mu_{t:z}$ —(the moment of z about the mean of z)— is the same function of $\mu_{1:x}, \mu_{2:x}, \mu_{3:x}$ ....—(the moments of x about the mean of x)—as $\mu'_{t:z}$ was of $\mu'_{1:x}, \mu'_{2:x}, \ldots\ldots$ There is one important simplification however due to the fact that $\mu_{1:x} = 0$ and hence all terms which involve $\mu_{1:x}$ vanish. With this in mind (6) becomes

$$(8)\quad \begin{cases} \mu_{1:z} = 0, \qquad \mu_{2:z} = \dfrac{n^2(a-1)}{an-1}\,\mu_{2:x}, \\ \mu_{3:z} = \dfrac{n^3(a-1)(a-2)}{(an-1)(an-2)}\,\mu_{3:x} \\ \mu_{4:z} = \dfrac{3an^3(n-1)(a-1)[n(a-1)-1]}{(an-1)(an-2)(an-3)}\,\mu^2_{2:x} \\ \qquad + \dfrac{n^4(a^3-7a^2+12a-6)+n^3a(a-1)}{(an-1)(an-2)(an-3)}\,\mu_{4:x} \\ \mu_{5:z} = \dfrac{10a(a-1)(a-2)(n-1)(an-n-1)}{(an-1)(an-2)(an-3)(an-4)}\,\mu_{2:x}\mu_{3:x} \\ \qquad + \dfrac{n^4(a-1)(a-2)(a^2n-12an+12n+5a)}{(an-1)(an-2)(an-3)(an-4)}\,\mu_{5:x} \end{cases}$$

etc.

Here again it is noticed that for $a = 2$, i.e. for $N = 2n$, $\mu_{3:z}$ is independent of $\mu_{3:x}$. In other words the skewness of the distribution of z is independent of the skewness of the parent

population of x. Similarly $\mu_{5;\bar{x}}$ is independent[4] of $\mu_{5;x}$ and also independent of $\mu_{3;x}$. But since P_7 is not zero for $a = 2$, $\mu_{7;\bar{x}}$ is not independent of $\mu_{7;x}$.

Now consider the variance of $\bar{x}$,

$$\mu_{2;\bar{x}} = \frac{r^2(a-1)}{ar-1}\mu_{2;x}$$

$$= \frac{Nr-r^2}{N-1}\mu_{2;x} \quad \text{(since } N = ar\text{)}.$$

Obviously it would be very desirable to have the variance (squared standard deviation) a minimum. Since the variance is a function of r differentiate $\mu_{2;\bar{x}}$ with respect to r.

$$\frac{d}{dr}\mu_{2;\bar{x}} = \frac{N-2r}{N-1}\mu_{2;x}.$$

To make $\mu_{2;\bar{x}}$ a minimum $\frac{N-2r}{N-1}\mu_{2;x} = 0$ and hence $N = 2r$ or, that is, $a = 2$.

When $a = 2$, $\mu_{2;\bar{x}} = \frac{N^2}{4(N-1)}\mu_{2;x}$, $\sigma_{\bar{x}} = \frac{N}{2\sqrt{N-1}}\sigma_x$.

In conclusion it may be said that there would seem to be good reason to suggest that, when possible, the investigator arrange to have twice as many variates in the control group or parent population as in each of the samples to be analyzed. Taking $r = \frac{N}{2}$ will insure that the skewness of the samples will be independent of the skewness of the parent population and also that the fifth moment of the samples will be independent of the fifth moment of the sampled population. In addition, taking $r = \frac{N}{2}$ will cause the variance (squared standard deviation) of the samples to be a minimum. Choosing $r = \frac{N}{2}$ presumes, of course, that N is an even number. But in most instances it should be possible to arrange that N be even. For if an odd number of observations are given either another observation may be added or one of the given observations deleted to make N even.

[4] P_5 vanishes with P_3 because $P_5 = P_3(1 - 12P_2)$. But P_3 is not a factor of P_7.