SAMPLING IN THE CASE OF CORRELATED OBSERVATIONS

By

CECIL C. CRAIG

National Research Fellow

Dr. E. C. Rhodes, in a paper in the *Journal of the Royal Statistical Society*,[1] has considered the distribution of characteristics of samples of N when the individual observations are not assumed to be independent. As he points out, there are many important cases in which the usual assumption of independence or randomness in the observations is not justifiable. In the present paper will be explained a method based on the semi-invariants of Thiele for the calculation of the characteristics of the sought distributions in this case which is especially to be preferred to the method based on moments when it is supposed that the observations are normally correlated. In the case it is further assumed that only consecutive observations are correlated, in addition to Dr. Rhodes' results, the third semi-invariant (which is the same as the third moment about the mean) of the variance and the mean and the variance of the third and fourth moments about the mean are given.

Let the N observations composing a sample be given by values of $x_1, x_2, \cdots, x_N$ respectively and let $F_N(x_1, x_2, \cdots, x_N)$ be the n-way probability function of $x_1, x_2, \cdots,$ and x_N.

[1]The Precision of Means and Standard Deviations When the Individual Errors Are Correlated, Vol. 90 (1927), pp. 135-143.

Then the semi-invariants, $\lambda_{rst}\cdots$ of $x_1, x_2, \cdots, x_N$ are defined by

$$(1)\qquad e^{(\sum_1^N \lambda_i t_i)+\frac{1}{2}(\sum_1^N \lambda_i t_i)^{(2)}+\frac{1}{3!}(\sum_1^N \lambda_i t_i)^{(3)}+\cdots} = \int_{-\infty,\cdots,-\infty}^{\infty,\cdots,\infty} dF_N(x_1, x_2, \cdots\cdots, x_N)\, e^{(\sum_1^N x_i t_i)} \quad {}^1$$

which is to be regarded as a formal identity in $t_1, t_2, \cdots, t_N$. $(\sum_1^N \lambda_i t_i)^{(k)}$ is first expanded by the multinomial law and then each term $\lambda_1^r, \lambda_2^s, \lambda_3^t \cdots$ in the result is replaced by $\lambda_{rst}\cdots$.

We shall pass over the characteristics of distributions of means, since the method of semi-invariants is equivalent to that of moments in this case, and take up the distribution of moments about the mean in samples of N. Following the method previously used by the author in the case of independent observations,[2] let

$$(2)\qquad \begin{aligned} \delta_i &= x_i - \sum_1^N \frac{x_i}{N} \\ &= \sum_{j=1}^N a_{ij} x_j \quad \text{with} \quad \begin{cases} a_{ij} = -\frac{1}{N} \\ a_{ii} = \frac{N-1}{N} \end{cases} \end{aligned}$$

Then let $V(\delta_1, \delta_2, \cdots\cdots, \delta_{N-1})$ be the probability function of $\delta_1, \delta_2, \cdots, \delta_{N-1}, (\sum_i \delta_i = 0.$ The semi-invariants $\lambda'_{rst}\cdots$ of $\delta_1, \delta_2, \cdots\cdots, \delta_{N-1}$ are defined by

[1]Following Cramér, I distinguish between probability and frequency functions. $F_N(x_1, x_2, \cdots, x_N)$ is the "cumulative" frequency function and thus the integral is an n-way Stieltjes integral.

[2]An Application of Thiele's Semi-invariants to the Sampling Problem; Metron, Vol. 7, No. 4 (1928), pp. 3-75.

$$e^{\left(\sum_1^{N-1}\lambda'_i t_i\right)+\frac{1}{2}\left(\sum_1^{N-1}\lambda'_i t_i\right)^{(2)}+\frac{1}{3!}\left(\sum_1^{N-1}\lambda'_i t_i\right)^{(3)}+\cdots}$$

$$(3)\qquad =\int_{-\infty,\cdots,-\infty}^{\infty,\cdots,\infty} dV(\delta_1,\cdots,\delta_{N-1})\,e^{\left(\sum_1^{N-1}\delta_i t_i\right)}$$

$$=\int_{-\infty,\cdots,-\infty}^{\infty,\cdots,\infty} dF_N(x_1,x_2,\cdots,x_N)\,e^{\left(\sum_{i=1}^{N-1}\sum_{j=1}^{N} a_{ij}x_j t_i\right)}$$

We have at once,

$$\left(\sum_{i=1}^{N-1} t_i \sum_{j=1}^{N}\lambda_j a_{ij}\right)^{(K)} = \left(\sum_{i=1}^{N-1}\lambda'_i t_i\right)^{(K)}$$

and as the author has previously remarked,[1] we can also write

$$(4)\qquad \left(\sum_1^{N} t_i \sum_{j=1}^{N}\lambda_j a_{ij}\right)^{(K)} = \left(\sum_1^{N}\lambda'_i t_i\right)^{(K)}$$

so long as the relation is only used to find the values of $\lambda'_{rst\cdots}$'s in which at least one of the subscripts is zero.

Then $S_K(\nu_n)$, the k'th semi-invariant of the n'th moment about the mean in samples of N, is given by the formula

$$S_K(\nu_n) =$$

$$\frac{1}{N^K}\sum\sum\cdots\frac{(-1)^{(r+s+t+\cdots)-1}\left[(r+s+t+\cdots)-1\right]!\,K!\,\nu'^{\,r}_{a_1 n, a_2 n,\cdots}\,\nu'^{\,s}_{b_1 n, b_2 n,\cdots}\,\nu'^{\,t}_{c_1 n, c_2 n,\cdots}\cdots}{\left[a_1!\,a_2!\cdots\right]^r\left[b_1!\,b_2!\cdots\right]^s\left[c_1!\,c_2!\cdots\right]^t\,r!\,s!\,t!\cdots}$$

[1] loc. cit., pp. 18, 19.

the notation $\nu'_{u\,v\,w}$ referring to moments of $\delta_1, \cdots \delta_{N-1}, \delta_N$, the summation including all terms for which

$$r(a_1+a_2+\cdots)+s(b_1+b_2+\cdots)+t\,(c_1+c_2+\cdots)+\cdots\cdots=k,$$

$$a_1 \geq a_2 \geq \cdots\cdots,$$

$$b_1 \geq b_2 \geq \cdots\cdots$$

$$c_1 \geq c_2 \geq \cdots\cdots$$

$$\cdots\cdots\cdots$$ [1]

$$(a_1+a_2+\cdots) > (b_1+b_2+\cdots) > (c_1+c_2+\cdots) > \cdots\cdots.$$

In particular:

$$S_1(\nu_n) = \frac{1}{N}\Sigma\,\nu'_{n,0}\,, \quad (\Sigma\nu'_{n,0} = \nu'_{n,0\cdots 0} + \nu'_{0,n,0\cdots 0} + \nu'_{0,0,n,0\cdots 0} + \cdots),$$

$$S_2(\nu_n) = \frac{1}{N^2}\left[\Sigma\,\nu'_{2n,0} + 2\Sigma\,\nu'_{n,n,0} - (\Sigma\,\nu'_{n,0})^2\right],$$

(5)

$$S_3(\nu_n) = \frac{1}{N^3}\left[\Sigma\,\nu'_{3n,0} + 3\Sigma\,\nu'_{2n,n,0} + 6\Sigma\,\nu'_{n,n,n,0} - 3(\Sigma\,\nu'_{2n,0})(\Sigma\,\nu'_{n,0})\right.$$
$$\left. - 6(\Sigma\,\nu'_{n,n,0})(\Sigma\,\nu'_{n,0}) + 2(\Sigma\,\nu'_{n,0})^3\right],$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

On writing out the moments $\nu'_{u\,v\,w}$ in terms of the semi-invariants λ'_{rst}.[2] and then using (4) the sought semi-invariants are obtained.

In the case that the N observations are normally correlated and $F_N(x_1, x_2, \cdots, x_N)$ is the N-dimensional normal probability function, the left-hand member of (4) vanishes for $k \geq 3$.

If we suppose that the standard deviations of $x_1, x_2, \cdots, x_N$ are all equal (which we shall always do) and take as the simplest case that $x_1, x_2, \cdots, x_N$ are normally correlated and that

[1]See the author's paper cited, p. 21, formula (25).

[2]For a detailed explanation of this kind of calculation see the author's paper cited, pp. 23-27.

the correlation as measured by the Pearsonian coefficient, $r_{x_i x_j}$, is the same for each pair, x_i, x_j, of the set of N observations, we get

$$\lambda'_{20} = \lambda'_{020} = \lambda'_{0020} = \cdots\cdots = \frac{N-1}{N}(\lambda_{20} - \lambda_{11}) = \frac{N-1}{N}(1-r)\lambda_{20},$$

$$\lambda'_{110} = \lambda'_{1010} = \lambda'_{0110} = \cdots\cdots = -\frac{1}{N}(\lambda_{20} - \lambda_{11}) = -\frac{1}{N}(1-r)\lambda_{20},$$

if the common value of $r_{x_i x_j}$ be denoted simply by r. But if the observations are independent and the parent population is normal we have

$$\lambda'_{20} = \lambda'_{020} = \lambda'_{0020} = \cdots\cdots = \frac{N-1}{N}\lambda_{20},$$

$$\lambda'_{110} = \lambda'_{1010} = \lambda'_{0110} = \cdots\cdots = -\frac{1}{N}\lambda_{20}.$$

Thus it follows that the distributions of the characteristics of samples of N in this particular case of dependent observations are the same as if the observations were independent and taken from a normal population of variance $(1-r)\lambda_{20}$.

In case $F_N(x_1, x_2, \cdots, x_N)$ is normal it is convenient to express the right hand members of (5) directly in terms of the semi-invariants $\lambda'_{rst\ldots}$ for $n = 2, 3, 4$. For that purpose we shall adopt the following notation. Let the linear form $\sum_{j=1}^{N} a_{ij}\lambda_j$ be denoted by A_i. Then (4) becomes

$$\left(\sum_1^N A_i t_i\right)^{(\kappa)} = \left(\sum_1^N \lambda'_i t_i\right)^{(\kappa)}. \tag{6}$$

Thus in a symbolic sense A_i's and λ_i's are equivalent. But with regard to the subscriots of the A terms in the expansion of the left member of (6) we use a different convention than for the subscripts of the λ's. We set

[1] See the author, loc. cit., p. 19.

$$\lambda'_{20} = A_{11}\,,\ \lambda_{020} = A_{22}\,,\ \cdots$$

$$\lambda'_{110} = A_{12}\,,\ \lambda_{1010} = A_{13}\,,\ \cdots$$

We get

$$S_1(\nu_2) = \frac{1}{N}\Sigma A_{ii}\,,$$

$$S_2(\nu_2) = \frac{1}{N^2}\left[3\Sigma A_{ii}^2 + 2\Sigma A_{ii}A_{jj} + 4\Sigma A_{ij}^2 - (\Sigma A_{ii})^2\right], i \neq j,$$

the summations, of course, running over all values of i and j from 1 to N. But since

$$\Sigma A_{ii}^2 + 2\Sigma A_{ii}A_{jj} = (\Sigma A_{ii})^2$$

the second relation reduces to

$$S_2(\nu_2) = \frac{2}{N^2}\left(\Sigma A_{ii}^2 + 2\Sigma A_{ij}^2\right).$$

Similarly

$$S_3(\nu_2) = \frac{8}{N^3}\left(\Sigma A_{ii}^3 + 3\Sigma A_{ii}A_{ij}^2 + 6\Sigma A_{ij}A_{ik}A_{jk}\right),$$

$$S_4(\nu_2) = \frac{48}{N^4}\left(\Sigma A_{ii}^4 + 4\Sigma A_{ii}^2 A_{ij}^2 + 4\Sigma A_{ii}A_{jj}A_{ij}^2 + 2\Sigma A_{ij}^4\right.$$

$$\left. + 8\Sigma A_{ii}A_{ij}A_{ik}A_{jk} + 4\Sigma A_{ij}^2 A_{ik}^2 + 8\Sigma A_{ij}A_{ik}A_{jl}A_{kl}\right), \tag{7}$$

. . . .

$$S_1(\nu_3) = 0,$$

$$S_2(\nu_3) = \frac{3}{N^2}\left(5\Sigma A_{ii}^3 + 6\Sigma A_{ii}A_{jj}A_{ij} + 4\Sigma A_{ij}^3\right),$$

$$S_3(\nu_3) = 0,$$

. . . .

$$S_1(V_4) = \frac{3}{N}\Sigma A_{ii}^2 ,$$

$$S_2(V_4) = \frac{48}{N^2}\left(2\Sigma A_{ii}^4 + 3\Sigma A_{ii} A_{jj} A_{ij}^2 + \Sigma A_{ij}^2\right).$$

To illustrate the use of these formulas and to give some results in a case of practical interest, let us suppose that the set of N observations composing a sample may be assigned an order in which only consecutive observations are correlated and in a constant degree. Thus our observations might be prices or indices taken at the ends of consecutive time intervals. We suppose, then, that

$$\lambda_{110} = \lambda_{0110} = \lambda_{00110} = \cdots = r\lambda_{20} ,$$

$$\lambda_{101} = \lambda_{1001} = \lambda_{0101} = \cdots = 0$$

The first step in the calculation is to obtain the values of the various A 's which enter into the formulas (7). A_{11} is found from A_1^2, A_{12} from $A_1 A_2$ and so on. We get

(8)

$$A_{11} = A_{N,N} = \left(1 - \frac{1}{N} - \frac{2r}{N^2}\right)\lambda_{20} ,$$

$$A_{22} = A_{33} = \cdots = A_{N-1,N-1} = \left(1 - \frac{1+2r}{N} - \frac{2r}{N^2}\right)\lambda_{20} ,$$

$$A_{12} = A_{N-1,N} = \left(r - \frac{1+r}{N} - \frac{2r}{N^2}\right)\lambda_{20} ,$$

$$A_{23} = A_{34} = \cdots = A_{N-2,N-1} = \left(r - \frac{1+2r}{N} - \frac{2r}{N^2}\right)\lambda_{20} ,$$

$$A_{13} = A_{14} = \cdots = A_{1,N-1} ,$$

$$= A_{2,N} = A_{3,N} = \cdots = A_{N-2,N} = \left(-\frac{1+r}{N} - \frac{2r}{N^2}\right)\lambda_{20} ,$$

$$A_{1,N} = \left(-\frac{1}{N} - \frac{2r}{N^2}\right)\lambda_{20} ,$$

$$A_{ij} = \left(-\frac{1+2r}{N} - \frac{2r}{N^2}\right)\lambda_{20} \qquad \begin{cases} 1 < i < N-1 \\ 1 < j < N-1 \\ |i-j| > 1 \end{cases}$$

Then, on substitution in (7), we have finally

$$S_1(V_2) = \left(1-\frac{1}{N}\right)\left(1-\frac{2r}{N}\right)\lambda_{20},$$

$$S_2(V_2) = \frac{2}{N}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{4r}{N}\right)+2r^2\left(1-\frac{3}{N}+\frac{2}{N^2}+\frac{2}{N^3}\right)\right]\lambda_{20}^2 .$$

These two results are given by Dr. Rhodes, loc. cit., though there is a slight misprint in the second one as given there. The remainder of the results given here are believed to be new.

$$S_3(V_2) = \frac{8}{N^2}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{6r}{N}\right)+6r^2\left(1-\frac{3}{N}+\frac{2}{N^2}+\frac{2}{N^3}\right)-\frac{4r^3}{N}\left(2-\frac{3}{N}-\frac{3}{N^2}-\frac{2}{N^3}\right)\right]\lambda_{20}^3,$$

$$S_1(V_3) = 0 ,$$

$$S_2(V_3) = \frac{6}{N}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\left(1-\frac{6r}{N}\right)-\frac{6r^2}{N}\left(1-\frac{5}{N}+\frac{12}{N^3}\right)+\frac{2r^3}{N^2}\left(1-\frac{7}{N}+\frac{14}{N^2}\right.\right.$$

$$\left.\left.+\frac{2}{N^3}-\frac{24}{N^4}-\frac{40}{N^5}\right)\right]\lambda_{20}^4 ,$$

$$S_3(V_3) = 0 ,$$

$$S_1(V_4) = 3\left[\left(1-\frac{1}{N}\right)^2\left(1-\frac{4r}{N}\right)+\frac{4r^2}{N^2}\left(1-\frac{3}{N^2}\right)\right]\lambda_{20}^2 ,$$

$$S_2(V_4) = \frac{24}{N}\left[\left(1-\frac{1}{N}\right)\left(4-\frac{9}{N}+\frac{6}{N^2}\right)\left(1-\frac{8r}{N}\right)+6r^2\left(1-\frac{5}{N}+\frac{25}{N^2}-\frac{23}{N^3}-\frac{44}{N^4}+\frac{08}{N^5}\right)\right.$$

$$-\frac{8r^3}{N}\left(4-\frac{19}{N}+\frac{33}{N^2}+\frac{30}{N^3}-\frac{54}{N^4}-\frac{108}{N^5}\right)+2r^4\left(1-\frac{9}{N}+\frac{44}{N^2}-\frac{64}{N^3}\right.$$

$$\left.\left.-\frac{114}{N^4}+\frac{192}{N^5}+\frac{360}{N^6}+\frac{288}{N^7}\right)\right]\lambda_{20}^4 .$$

It should be observed that the expressions for $S_1(V_n)$ for $N < 3$ and for $S_k(V_n)$, $k \geqslant 2$ for $N < 5$ are in general not valid, since it can be seen by reference to (8) that all the types of A 's used in the formulas (7) do not exist for values of N so small. But for these small values of N, the values of the characteristics for which expressions are given above can be readily computed directly.

C. C. Craig

Stanford University.