THE REGRESSION SYSTEMS OF TWO SUMS HAVING RANDOM ELEMENTS IN COMMON

By J. F. Kenney

1. **Introduction.** The purpose of this note is to illustrate the power and elegance of the technique of characteristic functions[1] in solving a problem which has been discussed in the literature by Fischer[2] and others.

Let $x_1, x_2, \cdots, x_n$ be n variables independent of each other in the statistical sense, all subject to the same distribution function f, so that the function representing their joint distribution is

$$f(x_1)f(x_2) \cdots f(x_n). \tag{1}$$

Under these conditions a set of values $x_1, x_2, \cdots, x_n$ will be said to constitute a *sample of* n from a population with distribution function $f(x)$ and the function (1) will be said to represent the distribution of samples. It will be understood that $f(x)$ is defined and is non-negative for all real values of x and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

If the actual occurrence of the variable is limited to a finite range, $f(x)$ is defined as identically zero outside that range.

The mathematical expectation of an arbitrary function $\psi(x)$, denoted by application of the operator E, is

$$E[\psi(x)] = \int_{-\infty}^{\infty} \psi(x)f(x)\,dx. \tag{2}$$

This integral will be convergent whenever $\psi(x)$ is absolutely integrable and bounded. In particular, if $\psi(x) = x$ we have the mean

$$a = \int_{-\infty}^{\infty} xf(x)\,dx$$

and it will be assumed that a exists.

Suppose a sample of n is taken from the population represented by $f(x)$ and the sum

$$y = x_1 + x_2 + \cdots + x_k + x_{k+1} + \cdots + x_n \tag{3}$$

---

[1] The writer takes pleasure in acknowledging his indebtedness to Professor A. T. Craig for suggesting this method.

[2] "On correlation surfaces of sums with a certain number of random elements in common," these Annals, vol. 4, no. 2, pp. 103–126.

is formed. From this sample $k < n$ values are chosen at random, and a sample of $m - k$ ($m \leq n$) additional values, x_j', is taken from $f(x)$. The sum

$$z = x_1 + x_2 + \cdots + x_k + x_{k+1}' + \cdots + x_m' \tag{4}$$

is then formed. The problem is to determine the regression systems of z on y and y on z in the population resulting from repeated samples.

Before proceeding with the solution a brief discussion of characteristic functions will be given.

2. **Characteristic functions.** When $\psi(x) = e^{itx}$, where t is a real variable and $i = \sqrt{-1}$, (2) is called the characteristic function of x. Thus if we let $\varphi(t) = E(e^{itx})$ we have

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx.$$

From the conditions imposed on $f(x)$ it follows that the integral defining $\varphi(t)$ is convergent and $|\varphi(t)| \leq 1$. If the kth derivative of $\varphi(t)$ with respect to t exists we have

$$\left.\frac{d^k \varphi(t)}{dt^k}\right|_{t=0} = i^k \nu_k$$

where

$$\nu_k = \int_{-\infty}^{\infty} x^k f(x)\, dx.$$

Thus the characteristic function of x has the property that its kth derivative at the origin (divided by i^k) gives the kth moment of the distribution of x about the origin of x.

The notion of characteristic function extends readily to a distribution of several variables. In particular, let $F(y, z)$ be the joint distribution function of variables y and z subject to the condition

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(y, z)\, dy\, dz = 1.$$

Then the characteristic function of $F(y, z)$ is

$$\varphi(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{it_1 y + it_2 z} F(y, z)\, dy\, dz \tag{5}$$

where y and z are defined in (3) and (4).

3. **Solution of the problem.** The distribution function associated with the population of samples is of the form given by (1). Consequently, the characteristic function of $F(y, z)$ can be written in the form

$$\int \cdots \int \prod_{j=1}^{k} e^{i(t_1+t_2)x_j} f(x_j)\, dx_j \prod_{j=k+1}^{n} e^{it_1 x_j} f(x_j)\, dx_j \prod_{j=k+1}^{m} e^{it_2 x_j'} f(x_j')\, dx_j'$$

the limits of integration being taken over all admissible values of the variables. The above expression reduces to

$$\varphi(t_1, t_2) = [\varphi(t_1 + t_2)]^k [\varphi(t_1)]^{n-k} [\varphi(t_2)]^{m-k}. \tag{6}$$

By the Fourier transform we have from (5),

$$F(y, z) = (1/2\pi)^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-it_1 y - it_2 z} \varphi(t_1, t_2)\, dt_1\, dt_2.$$

Since a distribution is completely determined by its characteristic function, $F(y, z)$ can be exhibited if $f(x)$ is known. However, the solution of the problem does not depend upon exhibiting $F(y, z)$.

Let $g(y)$ and $h(z)$ be the marginal distributions of y and z, respectively. Then the mean value of z for a fixed y is

$$\bar{z}_y = \int \frac{zF(y, z)}{g(y)}\, dz, \tag{7}$$

and the mean value of y for a fixed z is

$$\bar{y}_z = \int \frac{yF(y, z)}{h(z)}\, dy \tag{8}$$

where here and subsequently the integration is taken over all admissible values of the variables.

Let us now take the partial derivative of $\varphi(t_1, t_2)$, as given in (5), with respect to t_2 and evaluate the result at $t_2 = 0$. We obtain

$$\frac{\partial}{\partial t_2} \varphi(t_1, t_2)\Big|_{t_2=0} = \int \int ize^{it_1 y} F(y, z)\, dy\, dz. \tag{9}$$

If we denote the left member of (9) by $G(t_1)$ and utilize (7) in the right member, (9) becomes

$$G(t_1) = \int g(y)\bar{z}_y ie^{it_1 y}\, dy.$$

Application of the Fourier transform yields

$$ig(y)\bar{z}_y = \frac{1}{2\pi} \int e^{-it_1 y} G(t_1)\, dt_1. \tag{10}$$

Now from (6),

$$G(t_1) = k\overline{\varphi(t_1)}^{n-1} \varphi'(t_1) + \overline{\varphi(t_1)}^n ia(m - k).$$

Therefore (10) may be written as follows,

$$ig(y)\bar{z}_y = \frac{k}{2\pi} \int e^{-it_1 y} \overline{\varphi(t_1)}^{n-1} \varphi'(t_1)\, dt_1 + \frac{ia(m - k)}{2\pi} \int e^{-it_1 y} \overline{\varphi(t_1)}^n\, dt_1. \tag{11}$$

To evaluate these integrals, consider

$$\overline{\varphi(t_1)}^n = \int e^{it_1y}g(y)\,dy. \tag{12}$$

Differentiating (12) with respect to t_1 we have

$$n\overline{\varphi(t_1)}^{n-1}\varphi'(t_1) = \int iye^{it_1y}g(y)\,dy. \tag{13}$$

Again using the Fourier transform, we obtain

$$iyg(y) = \frac{nk}{2k\pi}\int e^{-it_1y}\overline{\varphi(t_1)}^{n-1}\varphi'(t_1)\,dt_1$$

from (13) and

$$g(y) = \frac{1}{2\pi}\int e^{-it_1y}\overline{\varphi(t_1)}^n\,dt_1$$

from (12). Therefore (11) reduces to

$$iyg(y)\bar{z}_y = \frac{k}{n}iyg(y) + ia(m-k)g(y)$$

and we have at once the simple result

$$\bar{z}_y = ky/n + a(m-k). \tag{14}$$

In an analogous manner, it may be shown that

$$\bar{y}_z = kz/m + a(n-k). \tag{15}$$

Writing (14) and (15) in the forms

$$\begin{cases} \bar{z}_y - am = c_1(y - an) \\ \bar{y}_z - an = c_2(z - am) \end{cases} \tag{16}$$

where $c_1 = k/n$ and $c_2 = k/m$ are the regression coefficients it follows from the linearity of the regressions that the correlation coefficient is

$$\rho = \sqrt{c_1c_2} = k/\sqrt{mn}.$$

If $m = n$, we have a well known result which is sometimes stated as follows: If y and z are affected by n equally likely causes of which k are common to both, then the correlation coefficient between y and z is equal to k/n.

NORTHWESTERN UNIVERSITY.