THE STOCHASTIC INDEPENDENCE OF SYMMETRIC AND HOMOGENEOUS LINEAR AND QUADRATIC STATISTICS

By Eugene Lukacs

National Bureau of Standards

1. Summary. The following theorem is proved. *If a univariate distribution has moments of first and second order and admits a homogeneous and symmetric quadratic statistic Q which is independently distributed of the mean of a sample of n drawn from this distribution, then it is either the normal distribution* (Q is then proportional to the variance) *or the degenerate distribution* (in this case no restriction is imposed on Q) *or a step function with two symmetrically located steps* (in this case Q is the sum of the squared observations). *The converse of this statement is also true.*

2. Introduction. It is known that the distributions of the mean and of the variance in samples from a continuous population are stochastically independent if and only if the parent distribution is normal. This theorem was proven independently by several authors [1], [2], [3].

The question arises whether there are any distributions having the property that the sampling distributions of the mean and of a symmetric and homogeneous quadratic statistic are stochastically independent. This is the problem to be discussed in the present paper; all distributions with this property will be determined and also the corresponding quadratic statistics. In this discussion we consider a constant as stochastically independent of any random variable.

In an earlier paper [3], dealing with the independence of the mean and the variance, the existence of a frequency function was assumed. To obtain all possible distribution functions it is necessary to refrain from this assumption and to express the formulae in terms of the cumulative distribution functions. Otherwise the derivation of the differential equation of the characteristic function (Section 3) resembles the reasoning of the preceding paper. In Section 4 and Section 5 this differential equation is discussed and the various possible solutions are determined.

3. The differential equation for the characteristic function. We consider a univariate population with a cumulative distribution function $F(x)$. Let $x_1, x_2, \cdots, x_n$ be n independent observations of the variate x. The cumulative distribution function of the sample is then given by $\Phi(x_1, x_2, \cdots, x_n) = F(x_1) F(x_2) \cdots F(x_n)$. Let us set $L = \sum_{j=1}^{n} x_j$ and $S = \sum_{j=1}^{n} x_j^2$. Any symmetric and homogeneous quadratic statistic Q can then be expressed in terms of L and S:

$$Q = aL^2 + bS. \tag{1}$$

The problem under investigation is the determination of all cumulative distribution functions $F(x)$ and statistics Q for which the sampling distributions of Q and of the mean $\bar{x} = \sum_{j=1}^{n} x_j/n$ (or equivalently of $L = n\bar{x}$) are stochastically independent.

The characteristic function of the distribution $F(x)$ is given by

$$\psi(u) = \int_{-\infty}^{+\infty} e^{iux}\, dF(x). \tag{2}$$

The characteristic function of the statistic L is

$$\phi_1(t) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{itL}\, d\Phi(x_1 \cdots x_n) = \prod_{j=1}^{n} \int_{-\infty}^{+\infty} e^{itx_j}\, dF(x_j) = [\psi(t)]^n. \tag{3.1}$$

Similarily the characteristic function of the statistic Q is given by

$$\phi_2(v) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{ivQ}\, d\Phi(x_1, \cdots, x_n), \tag{3.2}$$

and the characteristic function of the joint distribution of L and Q by

$$\phi(t, v) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{itL+ivQ}\, d\Phi(x_1, \cdots, x_n). \tag{3.3}$$

The statistics Q and L are independently distributed if and only if

$$\phi(t, v) = \phi_1(t)\phi_2(v). \tag{4}$$

Differentiating (4) with respect to v and then putting $v = 0$ one has

$$\left.\frac{\partial \phi}{\partial v}\right|_{v=0} = \phi_1(t) \left.\frac{\partial \phi_2}{\partial v}\right|_{v=0} = [\psi(t)]^n \left.\frac{\partial \phi_2}{\partial v}\right|_{v=0},$$

or from (3.1), (3.2) and (3.3)

$$\begin{aligned} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} Qe^{itL}\, d\Phi &= \left\{\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{itL}\, d\Phi\right\}\left\{\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} Q\, d\Phi\right\} \\ &= [\psi(t)]^n \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} Q\, d\Phi. \end{aligned} \tag{5}$$

Any distribution function $F(x)$ and any statistic Q which satisfies (4) also satisfies condition (5). It is worthwhile to remark that (5) implies certain restrictions on the statistic Q.

In the following we determine the distributions $F(x)$ and statistics Q which satisfy (5). This is done by first transforming (5) into a differential equation for the characteristic function $\psi(t)$ and by finding also a statistic Q using (5). Finally we shall investigate which of these distributions and statistics will also satisfy condition (4) and constitute hereby a solution to our problem.

We denote the left-hand side of (5) by $I(t)$. Substituting for Q from (1) we obtain after a simple computation

$$I(t) = -n[\psi(t)]^{n-2}\left[(a+b)\psi(t)\frac{d^2\psi}{dt^2} + a(n-1)\left(\frac{d\psi}{dt}\right)^2\right].$$

This permits one to reduce (5) to a differential equation for the characteristic function

$$(a+b)\left[\frac{1}{\psi(t)}\frac{d^2\psi}{dt^2}\right] + a(n-1)\left[\frac{1}{\psi(t)}\frac{d\psi}{dt}\right]^2 = A, \tag{6}$$

where

$$A = -[(a+b)\alpha_2 + (n-1)a\alpha_1^2], \tag{6.1}$$

with the initial conditions

$$\psi(0) = 1; \qquad \psi'(0) = i\alpha_1. \tag{6.2}$$

Here

$$\alpha_\nu = \int_{-\infty}^{+\infty} x^\nu \, dF(x).$$

From (6) and (6.2) we obtain easily first an equation for the cumulant generating function $g(t) = \ln \psi(t)$ and then the following differential equation for $h(t) = \dfrac{dg}{dt}$:

$$(a+b)\frac{dh}{dt} + (na+b)[h(t)]^2 = A, \tag{7}$$

with

$$h(0) = i\alpha_1. \tag{7.1}$$

4. Discussion of the differential equation (7) if $A \neq 0$. We have to distinguish three possibilities:

$$A \neq 0, \qquad na + b = 0, \qquad a + b \neq 0; \tag{8.1}$$

$$A \neq 0, \qquad na + b \neq 0, \qquad a + b \neq 0; \tag{8.2}$$

$$A \neq 0, \qquad na + b \neq 0, \qquad a + b = 0. \tag{8.3}$$

We first discuss case (8.1); equation (7) reduces to $(a + b)dh/dt = A$; from (6.1) and $na + b = 0$ it follows that

$$\frac{dh}{dt} = -\sigma^2 = -(\alpha_2 - \alpha_1^2).$$

Integrating this equation and considering that $h(t) = dg/dt$ we obtain using the initial condition (7.1), $g(t) = -\frac{1}{2}\sigma^2t^2 + i\alpha_1 t$. This is the cumulant generating function of the normal distribution. Since $b = -na$ we find $Q = -an^2s^2$. Here $s^2 = \sum_{j=1}^n x_j^2/n - (\sum_{j=1}^n x_j)^2/n^2$. In case (8.1) the parent distribution is normal and the statistic Q is—except for a constant factor—the sample variance. Conversely it is known that for the normal distribution variance and mean are stochastically independent.

We start the discussion of case (8.2) with the almost trivial case when $h(t)$ reduces to a constant, and delay dealing with the general case. If $h(t)$ is a constant and if (8.2) holds then equation (7) is satisfied if $h^2(t) = A/(na + b) =$ const. Then $h(t) = h(0) = i\alpha_1$, and since $dg/dt = h(t)$, $g(0) = 0$ we obtain $g(t) = i\alpha_1 t$ and $\psi(t) = e^{i\alpha_1 t}$. The function $\psi(t) = e^{i\alpha_1 t}$ is the characteristic function of the degenerate distribution

$$\varepsilon(x - \alpha_1) = \begin{cases} 1 & \text{if} \quad x \geqq \alpha_1, \\ 0 & \text{if} \quad x < \alpha_1. \end{cases}$$

It is easy to show that for this distribution any symmetric and homogeneous quadratic statistic is independent of the mean.

We proceed with the discussion of case (8.2) by assuming that

$$h^2(t) \neq \frac{A}{na + b}. \tag{9}$$

Equation (7) may then be written as

$$(a + b)\frac{dh}{dt} = A\left[1 - \frac{na + b}{A} h^2(t)\right].$$

We integrate this equation and then compute the function $\psi(t)$ from the relation $h(t) = dg/dt = d/dt \ln \psi(t)$, with $\psi(0) = 1$.
Thus we obtain

$$\psi(t) = [pe^{\frac{1}{2}\beta t} + qe^{-\frac{1}{2}\beta t}]^\lambda, \tag{10}$$

with

$$\begin{aligned} &\lambda = \frac{a + b}{na + b}, \qquad p = \frac{C}{C + 1}, \qquad q = \frac{1}{C + 1}, \\ &\beta = \frac{2}{\lambda}\sqrt{\frac{A}{na + b}} = \frac{2}{\lambda}\sqrt{(\lambda - 1)\alpha_1^2 - \lambda\alpha_2}, \\ &C = \frac{\lambda\beta + 2i\alpha_1}{\lambda\beta - 2i\alpha_1}. \end{aligned} \tag{10.1}$$

In these formulae a and b and therefore also A, β, C as well as λ, p and q may be functions of n.

In the following we have to distinguish two cases. We first assume that

$$\lambda\alpha_2 > (\lambda - 1)\alpha_1^2. \tag{8.2.1}$$

Then β is purely imaginary so that C, p and q are real numbers. Since $p + q = 1$, at least one of the two numbers p and q must be positive. If both p and q

are positive then it can be shown by computing the inverse Fourier transform of (10) that (10) can be a characteristic function only if λ is a positive integer. The corresponding distribution is in this case a binomial distribution. If one of the numbers p and q is positive while the other is negative then we see easily that $\lambda < 0$. In this case (10) is the characteristic function of a negative binomial distribution.

The binomial as well as the negative binomial distribution have the property that the random variable assumes an extreme value with a positive probability. Let L_0 be the extreme value which the statistic L may assume; we have then $P(L = L_0) > 0$. If $L = L_0$ then the value of Q is completely determined. Let us denote it by Q_0. Therefore the conditional probability $P(Q = Q_0 \mid L = L_0) = 1$; on account of the assumed stochastic independence of Q and L this means that $P(Q = Q_0) = 1$ so that Q has necessarily a degenerate distribution. We see then from (1) that x^2 has also a degenerate distribution. We conclude that under the conditions (8.2) and (8.2.1) the case $p > 0$, $q < 0$ (or $p < 0$, $q > 0$) can not occur. The distribution of x has therefore necessarily the form

$$F(x) = p\varepsilon(x - \xi) + q\varepsilon(x + \xi), \tag{10.2}$$

where

$$p > 0, \qquad q > 0, \qquad p + q = 1. \tag{10.3}$$

If we take $a = 0$, then $Q = bS$. From (10.2) it is seen that $S = n\alpha_2 = n\xi^2$ with probability one and hence $Q = bS$ is independent of L. Furthermore, if $n > 1$, $aL^2 + bS$ is not independent of L unless $p = 0$ or $p = 1$, and hence a must be zero and $Q = b\sum_{j=1}^n x_j^2$. On the other hand it is easily seen that for the distribution (10.2) the statistics $Q = \sum_{j=1}^n x_j^2$ and $L = \sum_{j=1}^n x_j$ are stochastically independent. This completes the discussion of the case where (8.2) and (8.2.1) hold.

We still have to consider the possibility that (8.2) is valid while

$$\lambda\alpha_2 \leqq (\lambda - 1)\alpha_1^2. \tag{8.2.2}$$

In this case β is real. Clearly, we must have $\lambda < 0$ since otherwise $\psi(t)$ would not be bounded. We put $\lambda = -\mu(\mu > 0)$ and write

$$\psi(t) = [pe^{\frac{1}{2}\beta t} + qe^{-\frac{1}{2}\beta t}]^{-\mu}. \tag{11}$$

Here p and q must be both positive since otherwise $\psi(t)$ would have a real pole. It is also seen easily that $\psi(t)$ has a maximum for a value t_0 given by $e^{\beta t_0} = q/p$. Furthermore, if $p \neq q$ we see that $\psi(t_0) = [2\sqrt{pq}]^{-\mu} > 1$, while $\psi(t_0) = 1$ if $p = q = \frac{1}{2}$. If $\psi(t)$ is a characteristic function we must therefore have $p = q = \frac{1}{2}$ so that (11) reduces to

$$\psi(t) = \left[\frac{1}{\cosh\frac{\beta t}{2}}\right]^{\mu} = \left[\prod_{j=1}^{\infty} \frac{1}{1 + \frac{\beta^2 t^2}{(2j-1)^2\pi^2}}\right]^{\mu}. \tag{11.1}$$

The representation of $\psi(t)$ as an infinite product shows that (11.1) is a characteristic function. We show finally that for the distributions determined by (11.1) no homogeneous and symmetric quadratic statistic Q exists which is independent of the statistic L. This is proven by demonstrating that $E(Q^2L^2) - E(Q^2)E(L^2)$ can not be zero for any distribution with the characteristic function (11.1). The symbol E denotes here as usual the expected value.

By a somewhat tedious but elementary computation one obtains from (11.1) and (1)

$$E(Q^2L^2) - E(Q^2)\,E(L^2) = 4n\mu\left(\frac{\beta}{2}\right)^6 a^2\left\{(\mu + 1)[(n + 2)\mu + 4]\left(\frac{b}{a}\right)^2 + [(n^2 + 5n)\mu^2 + (6n + 8)\mu + 8]\left(\frac{b}{a}\right) + (n\mu + 1)(3n\mu + 4)\right\}.$$

From $\mu = -\lambda$ and (10.1) we see that $b/a = -(n\mu + 1)/(\mu + 1)$. If we substitute this into the previous equation and simplify we obtain finally

$$E(Q^2L^2) - E(Q^2)E(L^2) = \frac{n\mu^2a^2(n\mu + 1)(n - 1)\beta^6}{8(\mu + 1)}. \tag{11.2}$$

Since $\mu > 0$, this can be zero only if $n = 1$. Then either $Q \equiv 0$ or $a + b \neq 0$. By (10.1) $n = 1$ and $a + b \neq 0$ imply $\lambda = 1$, which contradicts $\mu = -\lambda > 0$. The distribution functions (11.1), derived under the assumptions (8.2) and (8.2.2), do not therefore yield a solution of our problem.

We next consider the case (8.3) by assuming $A \neq 0$, $a + b = 0$, $na + b \neq 0$. Equation (7) then reduces to $[h(t)]^2 = -\alpha_1^2$ which leads to $\psi(t) = e^{\pm i\alpha_1 t}$ and $Q = 2a\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} x_jx_k$. We see therefore that case (8.3) leads to no new distribution as the degenerate distribution appears already in the preceding discussion.

5. Discussion of the differential equation if $A = 0$. In this paragraph we assume

$$A \equiv (a + b)\alpha_2 + (n - 1)a\alpha_1^2 = 0, \tag{12.1}$$

and rewrite equation (7) as

$$(a + b)\frac{dh}{dt} + (na + b)\,h^2(t) = 0. \tag{12.2}$$

We start the discussion of equations (12.1) and (12.2) with four cases which lead either to an already known solution for the characteristic function or to a trivial solution for the statistic Q. We assume first

$$a = 0. \tag{12.1.1}$$

From (12.1.1) and (12.1) we see $b\alpha_2 = 0$. If $b = 0$ we obtain the improper statistic $Q \equiv 0$, which is independent of any other statistic whatever be the parent distribution. If $\alpha_2 = 0$, we obtain the characteristic function $\psi(t) = 1$. The same solutions (either $\psi(t) = 1$ or $Q \equiv 0$) are also derived from (12.1), (12.2) and

$$a + b = 0, \tag{12.1.2}$$

or from (12.1), (12.2) and

$$\alpha_1 = 0. \tag{12.1.3}$$

If we assume

$$na + b = 0, \tag{12.1.4}$$

we obtain in a similar manner either the statistic $Q \equiv 0$ or the characteristic function $\psi(t) = e^{it\alpha_1}$.

We see therefore that if (12.1) holds the cases $a = 0$, $a + b = 0$, $\alpha_1 = 0$, $na + b = 0$ lead again either to a degenerate distribution which was fully treated in the preceding section or to the improper statistic $Q \equiv 0$, which is a trivial solution of our problem.

In the following we therefore discuss equation (12.2) only under the assumption that the four inequalities

$$a \neq 0, \qquad na + b \neq 0, \qquad a + b \neq 0, \qquad \alpha_1 \neq 0 \tag{12.3}$$

hold. In this case (12.2) may be written as $\frac{d}{dt}\frac{1}{h(t)} = -\frac{\sigma^2}{\alpha_1^2}$, where $\sigma^2 = \alpha_2 - \alpha_1^2$ is the variance of the parent distribution. If we integrate this with due regard to the initial conditions we obtain

$$\psi(t) = \left[1 - \frac{\sigma^2 it}{a_1}\right] - \frac{\alpha_1^2}{\sigma^2}. \tag{13}$$

This is the characteristic function of a gamma-distribution with parameters $\alpha = \alpha_1/\sigma^2$, $\lambda = \alpha_1^2/\sigma^2$. Its frequency function is given by

$$f(x, \alpha, \lambda) = \begin{cases} \dfrac{\alpha^2}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x} & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases} \qquad \text{if } \alpha_1 > 0 \tag{13.1}$$

and in case $\alpha_1 < 0$ by

$$g(x, \alpha, \lambda) = \begin{cases} 0 & \text{for } x \geqq 0, \\ \dfrac{(-\alpha)^\lambda}{\Gamma(\lambda)} (-x)^{\lambda-1} e^{-\alpha x} & \text{for } x < 0. \end{cases} \tag{13.2}$$

Equation (12.1), which led to the characteristic function (13) of the gamma-distribution, determines also the statistic Q. It is easily seen from (12.1) that $\sigma^2/\alpha_1^2 = -(na + b)/(a + b) = 1/\lambda$, therefore

$$\frac{b}{a} = -\frac{n\lambda + 1}{\lambda + 1}. \tag{14}$$

We show finally that the gamma distribution determined by (13) as a solution of the differential equation (12.2) does not lead to a solution of our problem since no statistic Q exists which for any sample size is independently distributed of L. Suppose $n > 1$. Then $| Q | = | a(L^2 - (n\lambda + 1)/(\lambda + 1)S)| \leqq | a | (L^2 + (n\lambda + 1)/(\lambda + 1)S)$. But $S \leqq L^2$ with probability one since all x's are of one sign with probability one. Hence $| Q | \leqq | a | [((n + 1)\lambda + 2)/(\lambda + 1)]L^2$ with probability one. The range of the values of Q is unbounded, hence Q can not be independent of L.

REFERENCES

[1] R. C. Geary, "Distribution of Student's ratio for non-normal samples," *Jour. Roy. Stat. Soc. Suppl.*, Vol. 3 (1936), pp. 178–184.

[2] T. Kawata and H. Sakamoto, "On the characterization of the normal population by the independence of the sample mean and the sample variance," *Jour. Math. Soc. Japan*, Vol. 1 (1949), pp. 111–115.

[3] E. Lukacs, "A characterization of the normal distribution," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 91–93.