DISTRIBUTION OF THE MEANS DIVIDED BY THE STANDARD DEVIATIONS OF SAMPLES FROM NON-HOMOGENEOUS POPULATIONS

By

G. A. BAKER

In a previous paper[1] the distributions of the means and variances, means squared and variances of samples of two drawn from a non-homogeneous population composed of two normal populations have been discussed. It is the purpose of this paper to discuss similarly the distribution of the means of samples of two measured from the mean of the population divided by the standard deviations of the samples for such parent populations and to present experimental results for samples of four.

CASE $n = 2$

Suppose that a population represented by

(1)

$$f(x) = \frac{N_1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x+m_1)^2}{\sigma_1^2}} + \frac{N_2}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-m_2)^2}{\sigma_2^2}}$$

$$-N_1 m_1 + N_2 m_2 = 0$$

is considered. If $n-s$ individuals come from the first component and s from the second in drawing samples of n and if $\bar{m}$ is the mean of the sample measured from the mean of the population and $\bar{\sigma}$ is the standard deviation of the sample,[2] then

(2)

$$\frac{\bar{m}}{\bar{\sigma}} = \frac{-(n-s)\bar{m}_1 + s\bar{m}_2}{\sqrt{n}\left[(n-s)\bar{\sigma}_1^2 + s\bar{\sigma}_2^2 + \frac{(n-s)s}{n}(\bar{m}_1 + \bar{m}_2)^2\right]^{\frac{1}{2}}}$$

[1] *Annals of Mathematical Statistics*, Vol. 2, No. 3, Aug. 1931.

[2] "Random Sampling from Non-Homogeneous Populations"—*Metron*, Vol. 8, No. 3, p. 6.

where $\bar{m}_1$, $\bar{m}_2$, $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$ are estimates of the corresponding parameters of (1).

For the case $n = 2$ when both individuals come from the first component of (1) it is known that the distribution of the means divided by the standard deviations of the samples is proportional to

(3)
$$\frac{du}{1+u^2}$$

the origin of u being taken at $-\frac{m_1}{\sigma_1}$. Similarly, when both individuals of the sample come from the second component, $\frac{\bar{m}}{\bar{\sigma}}$ is distributed as proportional to

(4)
$$\frac{dw}{1+w^2}$$

the origin of w being taken at $\frac{m_2}{\sigma_2}$.

When one individual comes from each component (2) becomes

(5)
$$\frac{\bar{m}}{\bar{\sigma}} = \frac{-\bar{m}_1 + \bar{m}_2}{\sqrt{2}\left[\frac{1}{2}(\bar{m}_1+\bar{m}_2)^2\right]^{\frac{1}{2}}}$$

because no estimate of the standard deviations of the components of (1) can be made from one individual. The distribution of $\bar{m}_1$ is proportional to the first component of (1), and $\bar{m}_2$ is distributed as proportional to the second component. The distributions of $\bar{m}_1$ and $\bar{m}_2$ are independent.

Expression (5) can be rewritten as

(6)
$$\frac{\bar{m}}{\bar{\sigma}} = 1 - \frac{2}{1+\frac{\bar{m}_2}{\bar{m}_1}}.$$

Put

$$\frac{\bar{m}_2}{\bar{m}_1} = v.$$

(7)
$$\frac{\bar{m}_2}{\bar{m}_1} = v.$$

If the distribution of v is found, the distribution of $z = \frac{\bar{m}}{\sigma}$ may be found by making the transformation

(8)
$$z = 1 - \frac{2}{1+v}$$

or

(9)
$$v = -1 - \frac{2}{z-1}$$

and

(10)
$$dv = \frac{2}{(z-1)^2}\, dz.$$

The distribution of v is a special case of the distribution of an index both of whose components follow the normal law. That is, we seek the distribution of

$$v = \frac{y}{x}$$

x and y being distributed as

(11)
$$z_0 e^{-\frac{1}{2}\frac{1}{1-r^2}\left[\frac{(x-\bar{x})^2}{\sigma_1^2} - 2r\frac{(x-\bar{x})}{\sigma_1}\frac{(y-\bar{y})}{\sigma_2} + \frac{(y-\bar{y})^2}{\sigma_2^2}\right]}.$$

This distribution may be obtained as follows.

Lemma I.[2,3] If two variables x and y, $-\infty \leq x \leq \infty$, $-\infty \leq y \leq \infty$

[2] (Loc. cit.)

[3] Baten, W. D. "Combining Constant Probability Functions"—*American Mathematical Monthly*, Oct., 1930.

are so related that the probability of an x in dx and of a y in dy is

$$f(x,y)\,dx\,dy$$

then the probability that $v=x-y$ is in dv is proportional to

$$\left[\int_{-\infty}^{\infty} f(v+y,y)\,dy\right]dv.$$

Consider, first, the portion of (11) in the first quadrant. Put

$$v=\frac{y}{x}$$

and take the logarithm of each side, thus

(12) $$\log v=\log y-\log x.$$

Put

$$I=\log v$$
$$w=\log y$$
$$u=\log x$$

and (12) becomes

(13) $$I=w-u$$

where the range of w and u is $-\infty$ to $+\infty$. The equation of the correlation surface of w and u is proportional to

(14)

$$F(w,u)=e^{u}e^{w}e^{-\frac{1}{2}\frac{1}{1-r^2}\left[\frac{(e^{u}-\bar{x})^2}{\sigma_1^2}-2r\frac{(e^{u}-\bar{x})(e^{w}-\bar{y})}{\sigma_1}\frac{}{\sigma_2}+\frac{(e^{w}-\bar{y})^2}{\sigma_2^2}\right]}$$

Hence, $F(u,I+u)du$ when the transformation $e^{u}=x$ is made, becomes

(15)

$$xe^{I}e^{-\frac{1}{2}\frac{1}{1-r^2}\left[\frac{(x-\bar{x})^2}{\sigma_1^2}-\frac{2r(x-\bar{x})}{\sigma_1}\frac{(xe^{I}-\bar{y})}{\sigma_2}+\frac{(xe^{I}-\bar{y})^2}{\sigma_2^2}\right]}dx$$

where x ranges from 0 to ∞. By the application of Lemma I the proportional probability of a value of v in dv when x and y are both positive is obtained by integrating (15) from 0 to ∞ with respect to x and making the transformation $v=e^{I}$. Thus,

(16)

$$\frac{\sigma_1\sigma_2\sqrt{1-r^2}}{a}\,e^{-\frac{1}{2}\frac{1}{1-r^2}\left[\frac{\bar{x}^2}{\sigma_1^2}-2r\frac{\bar{x}\bar{y}}{\sigma_1\sigma_2}+\frac{\bar{y}^2}{\sigma_2^2}\right]}$$

$$+\frac{b}{a^{3/2}}e^{-\frac{(\bar{x}v-\bar{y})^2}{2a}}\int_0^{\frac{b}{\sigma_1\sigma_2\sqrt{1-r^2}\sqrt{a}}}e^{-\frac{1}{2}z^2}\,dz$$

$$+\frac{\sqrt{\pi}}{\sqrt{2}}\frac{b}{a^{3/2}}\,e^{-\frac{(\bar{x}v-\bar{y})^2}{2a}}$$

where

$$a=\sigma_2^2-2r\sigma_1\sigma_2 v+\sigma_1^2 v^2$$

$$b=(\sigma_1^2\bar{y}-r\sigma_1\sigma_2\bar{x})v+(\sigma_2^2\bar{x}-r\sigma_1\sigma_2\bar{y})$$

is obtained. The distribution of v if both x and y are negative (and hence v positive) is the same as (16) except that the last term is reversed in sign. Thus, for v positive the distribution of v is proportional to two times the first two terms of (16). If v is negative, i.e. x negative and y positive or x positive and y negative, (16) is obtained in one case and (16) with the sign of the last term changed in the other case. That is, the distribution of v is proportional to the first two terms of (16) when v ranges from $-\infty$ to $+\infty$.

In our case $r=0$, $m_1=\bar{x}$, $m_2=\bar{y}$. Hence, the distribution of v becomes proportional to

(17)

$$\frac{\sigma_1\sigma_2\,e^{-\frac{1}{2}\left[\frac{m_1^2}{\sigma_1^2}+\frac{m_2^2}{\sigma_2^2}\right]}}{(\sigma_2^2+\sigma_1^2v^2)}$$

$$+\frac{\sigma_1^2m_2v+\sigma_2^2m_1}{(\sigma_2^2+\sigma_1^2v^2)^{3/2}}\,e^{-\frac{(m_1v-m_2)^2}{2(\sigma_2^2+\sigma_1^2v^2)}}\int_0^{\frac{\sigma_1^2m_2v+\sigma_2^2m_1}{\sigma_1\sigma_2\sqrt{\sigma_2^2+\sigma_1^2v^2}}}e^{-\frac{1}{2}z^2}\,dz.$$

From (8), (9), (10), and (17) z is distributed as proportional to

(18)

$$\frac{\sigma_1\sigma_2}{A}\, e^{-\frac{1}{2}\left[\frac{m_1^2}{\sigma_1^2}+\frac{m_2^2}{\sigma_2^2}\right]} + \frac{B}{A^{3/2}}\, e^{\frac{-[z(-m_1+m_2)-(m_1-m_2)]^2}{2A}} \int_0^{\frac{B}{\sigma_1\sigma_2 A^{1/2}}} e^{-\frac{1}{2}u^2}\, du$$

where

$$A = (\sigma_1^2)+\sigma_2^2)z^2+2(\sigma_1^2-\sigma_2^2)z+(\sigma_1^2+\sigma_2^2)$$

and

$$B = z(\sigma_2^2 m_1 - \sigma_1^2 m_2)-(\sigma_1^2 m_2+\sigma_2^2 m_1).$$

The origin for z is at the mean of population (1).

Thus the distribution of $\frac{m}{\sigma}$ for samples of two drawn from a population represented by (1) has been completely determined as being proportional to

(19)

$$k_1(3)+k_2(4)+k_3(18).$$

Let A_1, A_2 and A_3 be the respective areas under the curves represented by the three terms of (19). Then k_1, k_2, and k_3 are to be so determined that

$$A_1+A_2+A_3 = N$$

where N is the total number of samples considered, and that

$$\frac{1}{A_1} = \frac{k^2}{A_2} = \frac{2k}{A_3}$$

where

$$k = \frac{N_2}{N_1}\ .$$

Expression (19) indicates, in general, that if the means of the components do not coincide and if one component is not large compared with the other, both tails of the distribution of the means of samples measured from the mean of the population divided by the standard deviations of the samples are heavier for populations of the type (1) than for normal populations. In case the means of the components coincide one tail will be heavier $(\sigma_1^2 \neq \sigma_2^2)$. In any case at least one tail will be heavier.

EXPERIMENTAL RESULTS

Samples of four were drawn from a population approximately represented by

(1)

$$f(x) = \frac{648}{5\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-15.5)^2}{25}} + \frac{648}{5\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-32.5)^2}{25}}$$

which is the same as Population I in the first reference. These samples were drawn by throwing dice. The means of these samples were calculated and referred to the mean of (1) as an origin. The standard deviations of the samples were obtained and $\frac{\bar{m}}{\sigma}$ calculated. A grouped frequency distribution of 1038 of these values is presented in Chart I and in Table I. Large values are obtained more frequently than would be expected from a normal population.

TABLE I

Grouped Frequency distribution of 1038 Values of $\overline{m}/\sigma$ for Samples of Four from Population I

Middle of Interval	Frequency
—6.0	1
—5.6	0
—5.2	1
—4.8	2
—4.4	1
—4.0	1
—3.6	4
—3.2	6
—2.8	4
—2.4	10
—2.0	11
—1.6	9
—1.2	42
—0.8	71
—0.4	185
0	294
0.4	183
0.8	100
1.2	40
1.6	23
2.0	14
2.4	9
2.8	9
3.2	8
3.6	4
4.0	0
4.4	1
4.8	1
5.2	1
5.6	0
6.0	1
6.4	2
6.8	0
7.2	0
7.6	0
8.0	0
8.4	1

CHART I

Grouped Frequency Distribution of 1038 Values of $\overline{m}/\overline{\sigma}$ for Samples of Four from Population I