THE SAMPLING DISTRIBUTION OF THE COEFFICIENT OF VARIATION

By Walter A. Hendricks with the assistance of Kate W. Robey

National Agricultural Research Center, Beltsville, Maryland

The coefficient of variation does not appear to be of very great interest to statisticians in general. However, its use in biometry is sufficiently extensive for some knowledge of its sampling distribution to be desirable. The present paper is an attempt to satisfy this need.

For the purposes of the following discussion, the coefficient of variation may be defined as the ratio of the standard deviation of a number of measurements to the arithmetic mean:

$$v = \frac{s}{\bar{x}} \qquad (1)$$

As is well known, the probability that the mean of a sample of n measurements, taken at random from a normal universe, lies between $\bar{x}$ and $\bar{x} + d\bar{x}$ and that the standard deviation of the measurements in the same sample lies between s and $s + ds$ is given by the relation:

$$dF_{\bar{x},s} = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n} e^{-\frac{n}{2\sigma^2}[(\bar{x}-m)^2+s^2]} s^{n-2}\,d\bar{x}\,ds \qquad (2)$$

If equation (2) is expressed in terms of polar coördinates by means of the transformation: $\bar{x} = \rho\cos\theta$; $s = \rho\sin\theta$, it becomes a distribution function of ρ and θ in which $\theta = \text{arc}\tan v$:

$$dF_{\rho,\theta} = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n} e^{-\frac{n}{2\sigma^2}(\rho^2-2m\rho\cos\theta+m^2)} \rho^{n-1}\sin^{n-2}\theta\,d\rho\,d\theta \qquad (3)$$

In equation (3), ρ may vary from 0 to ∞ and θ may vary from 0 to π. To find the distribution function of θ, all that is necessary is to write:

$$dF_\theta = k\left[\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho\right]d\theta \qquad (4)$$

in which,

$$k = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{n}{2\sigma^2}m^2\sin^2\theta}\sin^{n-2}\theta,$$

$$a = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\sigma}, \quad \text{and} \quad b = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\sigma}\, m\cos\theta,$$

and to perform the indicated integration.

To evaluate the integral inside the brackets in equation (4), we may write:

$$\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho = \frac{1}{a^n}\int_{-b}^\infty e^{-u^2}\sum_{i=0}^{n-1}\frac{(n-1)!}{(n-1-i)!\,i!}\,u^{n-1-i}\,b^i\,du \ldots\ldots(5)$$

Consider the integral, $\int_{-b}^\infty e^{-u^2}u^{n-1-i}\,du$. If b is sufficiently large, as is the case when the parameters of equation (2) are of such magnitude that practically the entire volume under the frequency surface lies to the right of the s axis, that is to say, if negative and small positive values of $\bar{x}$ occur so infrequently that their effects may be neglected, the lower limit, $-b$, of this integral may be replaced by $-\infty$ without introducing any appreciable error. The value of the integral, $\int_{-\infty}^\infty e^{-u^2}u^{n-1-i}\,du$, is zero when $n-1-i$ is odd and $\Gamma\left(\frac{n-i}{2}\right)$ when $n-1-i$ is even, zero being counted as an even number.

Subject to the above condition that b be sufficiently large, we may, therefore, write equation (5) in the form:

$$\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho = \frac{1}{a^n}\sum_{i=0}^{n-1}{}'\frac{(n-1)!}{(n-1-i)!\,i!}\,\Gamma\left(\frac{n-i}{2}\right)b^i \ldots\ldots(6)$$

in which the symbol, $\sum'$, indicates that the only terms entering into the summation are those in which $n-1-i$ is an even number.

Substituting this expression for the integral inside the brackets in equation (4), replacing k, a, and b by the quantities which they represent, and writing V in place of the ratio, $\frac{\sigma}{m}$, we obtain the following distribution function of θ:

$$dF_\theta = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}\, e^{-\frac{n}{2V^2}\sin^2\theta}\sin^{n-2}\theta\sum_{i=0}^{n-1}{}'\frac{(n-1)!\,\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!\,i!}\,\frac{n^{\frac{1}{2}i}}{2^{\frac{1}{2}i}V^i}\cos^i\theta\,d\theta. (7)$$

Equation (7) may be written in terms of v, if desired, by making the substitution, $\theta = \text{arc}\tan v$:

$$dF_v = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{n}{2V^2}\frac{v^2}{1+v^2}} \frac{v^{n-2}}{(1+v^2)^{\frac{1}{2}n}}$$

$$\sum_{i=0}^{n-1}{}' \frac{(n-1)!\,\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!\,i!} \frac{n^{\frac{1}{2}i}}{2^{\frac{1}{2}i}\,V^i} \frac{1}{(1+v^2)^{\frac{1}{2}i}} dv \quad (8)$$

It must be emphasized that equation (8) has been derived on the hypothesis that negative and small positive values of $\bar{x}$ occur so infrequently that they may be neglected. However, since this condition is satisfied in the vast majority of practical problems in which the coefficient of variation is likely to be used, the limitation is not of much practical importance.

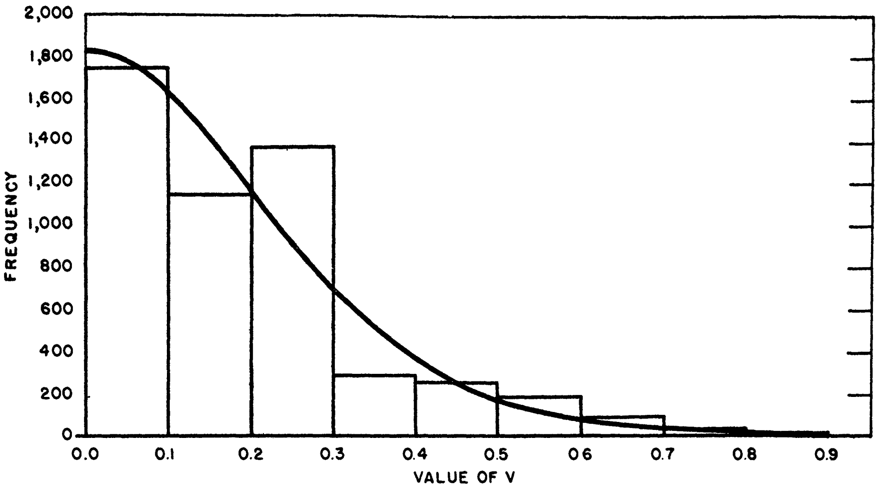

FIG. 1. OBSERVED AND THEORETICAL DISTRIBUTIONS OF VALUES OF v FOR 512 SAMPLES OF NUMBERS OF HEADS APPEARING IN TWO SUCCESSIVE TOSSES OF TEN COINS

As a test of the validity of equation (8), the authors calculated 512 coefficients of variation of the numbers of heads appearing in two successive tosses of ten coins. The coins were tossed 1024 times, thus yielding 512 samples, each consisting of two observations. For these data we have $m = 5$, $\sigma = 1.581$, and $V = 0.3162$.

For the case, $n = 2$, equation (8) reduces to:

$$dF_v = \frac{2}{\pi^{\frac{1}{2}}V} e^{-\frac{1}{V^2}\frac{v^2}{1+v^2}} \frac{dv}{(1+v^2)^{\frac{3}{2}}} \quad (9)$$

Figure 1 shows the distribution of the 512 values of v obtained from the coin tossing experiment, together with the theoretical distribution given by equation (9).

An inspection of Figure 1 indicates that the agreement between the observed and theoretical frequencies is fairly good. An application of the familiar chi test for goodness of fit showed the agreement to be rather poor. According to this test, the degree of discrepancy between theory and observation could have arisen by chance less than once in a hundred trials. However, the discrepancies may be partly due to the fact that data distributed in a discrete fashion were treated by methods appropriate to the analysis of data distributed according to a continuous frequency curve.

As another test of the validity of equation (8), the authors calculated 149 coefficients of variation of "days to maturity," which is the length of time elapsing between the date of hatch of a chicken and the time egg production com-

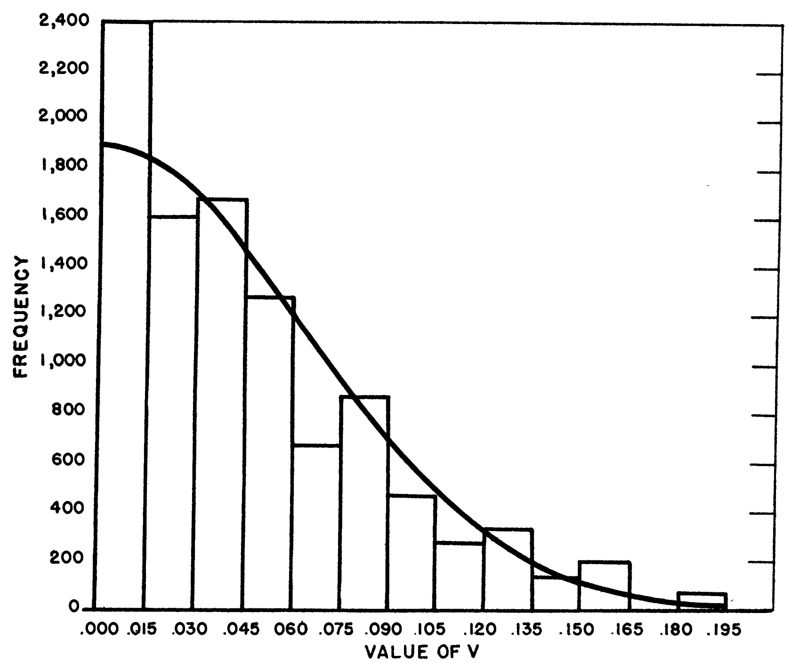

FIG. 2. OBSERVED AND THEORETICAL DISTRIBUTIONS OF VALUES OF v FOR 149 SAMPLES OF "DAYS TO MATURITY" IN RHODE ISLAND RED PULLETS FOR SAMPLES OF TWO OBSERVATIONS

mences, for samples of two observations made upon Rhode Island Red pullets. Figure 2 shows the observed distribution of the 149 coefficients of variation, together with the theoretical distribution given by equation (9).

In applying equation (9) to these data, the parameter, V, had to be evaluated from the data. The best estimates of the values of m, σ, and V which could be obtained from the 298 measurements of "days to maturity" are $m = 210.477$, $\sigma = 18.6991$, $V = 0.0888415$. The theoretical distribution shown in Figure 2 is based on this value of V.

The agreement between theory and observation shown by Figure 2 is very good. In this case, the chi test showed that the degree of discrepancy encountered could have arisen by chance about six times in ten trials.