ON THE RELATIVE STABILITY OF THE MEDIAN AND ARITHMETIC MEAN, WITH PARTICULAR REFERENCE TO CERTAIN FREQUENCY DISTRIBUTIONS WHICH CAN BE DISSECTED INTO NORMAL DISTRIBUTIONS[1]

By
HARRY S. POLLARD

I.

THE CHOICE OF AN AVERAGE

In any statistical investigation in which an average is to be used as a summarizing figure for a frequency distribution the question arises, which average best describes the distribution. That this is still a debatable question among writers on economic statistics is shown by a perusal of the many papers dealing with the measurement of seasonal variations which have appeared in recent years.[2]

Each of the proposed methods of isolating seasonal variations involves an averaging, either of monthly items or of relatives of monthly items, but whether this averaging is best accomplished by use of the arithmetic mean, the median, or the mean of a middle group of items seems to be a moot point. Persons[3] employs the median of link relatives of monthly items, since by this device the influence of large non-seasonal variations may be greatly moderated. Hart[4] in justifying the use of the arithmetic mean has shown that the method of monthly means gives the actual

[1] A resume of a dissertation, bearing the same title, written under the direction of Professor Mark H. Ingraham and submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University of Wisconsin, 1933.

[2] For a bibliography of literature on this subject see Mills, F. C., Statistical Methods, p. 343.

[3] Persons, W. M., *Correlation of Time Series,* Jour. Amer. Statis. Assn., June, 1923, p. 717.

[4] Hart, W. L., *The Method of Monthly Means for Determination of a Seasonal Variation,* Jour. Amer. Statis. Ass'n., Sept., 1922, pp. 341-349.

monthly values of the seasonal variation in case the seasonal variation is strictly periodic throughout the period of years under consideration and the long term variations are also periodic with integral numbers of years as their periods. The proof of this theorem is based on a property of Fourier series discussed by Bocher[5].

The point of view of this paper is that another factor of importance should influence the choice of an average, that this choice should be guided not alone by consideration of exceptional cases which may arise, nor by theory which assumes a periodicity seldom found in sequences of economic data, but also by a consideration of the stability of the averages. For if a given frequency distribution is regarded as a random sample drawn from a theoretical distribution which contains a very large number of items, the accuracy with which a particular average of the sample will typify the entire theoretical distribution is influenced by the frequency curve for that average. It is the purpose of this paper to compare the stability of the arithmetic means and medians of frequency distributions which may be dissected into two and three normal distributions, and to develop a general method of comparing the relative stability of the mean and median which shall be applicable to any frequency distribution.

The dissection of a frequency curve into two normal components has been discussed by Karl Pearson[6], who has developed methods for determining the values of the parameters of both symmetrical and asymmetrical frequency functions. He has applied these methods to distributions of cranial weights. Crum[7] has used Pearson's method of dissecting a symmetrical distribution in his discussion of the relative stability of the median and mean of link

---

[5] *Annals of Mathematics*, Second Series, vol. 7, p. 135, Formula (63).

[6] Pearson, K., *Contributions to the Mathematical Theory of Evolution, Philosophical Transactions*, Series A, vol. 185, 1894, pp. 71-110.

[7] Crum, W. L., *The Use of the Median in Determining Seasonal Variation*, Jour. Amer. Statis. Ass'n., March, 1923, pp. 607-614.

relatives of monthly figures for the rate of interest on sixty to ninety-day commercial paper for the years 1890-1917, and his results are discussed in section VI of this paper. Our interest in an asymmetrical distribution composed of two normal distributions arises from the fact that such a distribution affords a good fit both to distributions which possess two distinct modes, and to skewed distributions with one mode. The study of a distribution which may be dissected into three normal components is suggested by the occurrence in economic data of tri-modal distributions. This paper will be concerned with only a particular class of three-component distributions, those which are symmetrical.

The hypothesis from which this investigation started was that a good criterion for measuring the stability of an average is its standard deviation. However, a difficulty which soon presented itself was the accurate determination of the standard deviation of the median. The classical formula for expressing the standard deviation, σ_M, of the medians of samples of s items each, drawn from a frequency distribution whose equation is $y = f(x)$ and which satisfies the condition:

$$\int_{-\infty}^{o} f(x)\,dx = \frac{1}{2} = \int_{o}^{\infty} f(x)\,dx \quad \text{is} \quad \sigma_M = \frac{1}{2\sqrt{s}\cdot f(o)}.$$

The approximation to the value of the standard deviation of the median given by this formula is discussed in section IV, where it is shown that, although this approximation is close to the true value of the standard deviation of the median when s is large, it may be a very poor approximation when s is small, particularly for certain types of frequency curves.

Since it became obvious that the relative stability of the medians and arithmetic means of small samples cannot be determined by the methods which are valid when the samples are large, this paper resolved itself into two distinct investigations: a treatment of certain frequency functions using the classical formula for the standard deviation of the median, valid for large values of s;

and the development of a second method of comparing the stability of the arithmetic mean and median which may be applied also when S is small. The first of these topics is considered in sections II and III, the second is taken up in sections IV and V, and is mathematically the more interesting part of the work. In section VI the various methods of comparing the stability of the arithmetic mean and median are applied to a particular sequence of economic data.

II.

THE RELATIVE MAGNITUDE AND STABILITY OF THE ARITHMETIC MEAN OF A FREQUENCY DISTRIBUTION WHICH IS COMPOSED OF TWO NORMAL DISTRIBUTIONS.

1. The Mean and Median and Their Standard Deviations.

In this section a study will be made of the frequency function whose equation is

$$(1) \qquad y = \frac{1}{\sqrt{2\pi}}\left(\frac{\alpha_1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{\alpha_2}{\sigma_2} e^{-\frac{(x-\beta)^2}{2\sigma_2^2}}\right),$$

with the purpose of determining the influence of the five parameters of this equation upon the location of the mean and median of the distribution, and upon the standard deviations of these averages.

The only conditions imposed upon the parameters α_1, α_2, σ_1, σ_2, β are that they shall assume only positive values (since they represent, respectively, the areas of the two component curves, their standard deviations, and the distance between their arithmetic means), and that the first two parameters shall satisfy the equation

$$(2) \qquad \alpha_1 + \alpha_2 = 1,$$

so that the total probability, as represented by $\int_{-\infty}^{\infty} y\,dx$, shall be unity.

The arithmetic mean, $\bar{x}$, of the distribution may be expressed as a function of the parameters by the equation

$$(3) \qquad \bar{x} = \alpha_2 \beta.$$

The median, M, of the distribution satisfies the equation

$$(4) \qquad \int_{M}^{\infty} \frac{1}{\sqrt{2\pi}} \left(\frac{c_1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(x-b)^2}{2\sigma_2^2}} \right) dx = \frac{1}{2} ,$$

and can in general be located only by interpolation in a table of areas under the normal curve. This interpolation can be more easily performed if equation (4) is transformed into

$$(5) \qquad c_1 \int_{0}^{M/\sigma_1} e^{-\frac{t^2}{2}} dt = c_2 \int_{0}^{\frac{b-M}{\sigma_2}} e^{-\frac{t^2}{2}} dt .$$

In distribution (1), σ_1 and σ_2 denote the standard deviations of the component distributions, measured from the means of the respective components. Hence, the standard deviation, σ, of the entire distribution satisfies the equation

$$\sigma = \sqrt{c_1(\sigma_1^2 + \bar{x}^2) + c_2(\sigma_2^2 + [b-\bar{x}]^2)}$$

Therefore the value of the standard deviation of the arithmetic means of samples containing S items each drawn from distribution (1) is

$$(6) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{S}} = \sqrt{\frac{c_1\sigma_1^2 + c_2\sigma_2^2 + c_1c_2b^2}{S}}$$

If we assume that S, the number of items in the sample, is sufficiently large to justify its use, an approximation to the standard deviation of the median may be obtained from the equation

$$(7) \qquad \sigma_M = \frac{1}{2y_M\sqrt{S}} \quad \text{where} \quad y_M = \frac{1}{\sqrt{2\pi}} \left(\frac{c_1}{\sigma_1} e^{-\frac{M^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(b-M)^2}{2\sigma_2^2}} \right)$$

2. The Relative Magnitude of the Median and Mean.

From equations (3) and (5) it is seen that, if four of the five parameters, c_1, c_2, σ_1, σ_2, b, are fixed and the fifth is allowed to vary, both $\bar{x}$ and M will be monotone increasing functions of b and of c_2, and monotone decreasing functions of c_1, and that $\bar{x}$ is independent of the standard deviations of both components, while M is a monotone increasing function of σ_1 and a monotone decreasing function of σ_2.

When $c_1 = c_2$ and $\sigma_1 = \sigma_2$

distribution (1) becomes symmetrical, and

$$\bar{x} = M = \frac{\theta}{2}.$$

To obtain conditions under which $\bar{x}$ shall exceed M, let equations (3) and (5) be differentiated with respect to θ. The inequality

$$\frac{d\bar{x}}{d\theta} > \frac{dM}{d\theta}$$

may be reduced to the form

$$\frac{1}{\sigma_1} e^{-\frac{M^2}{2\sigma_1^2}} > \frac{1}{\sigma_2} e^{-\frac{(\theta - M)^2}{2\sigma_2^2}}$$

It follows from equation (5) that when $c_1 > c_2$,

$$\frac{\theta - M}{\sigma_2} > \frac{M}{\sigma_1}, \text{ whence } e^{-\frac{M^2}{2\sigma_1^2}} > e^{-\frac{(\theta - M)^2}{2\sigma_2^2}}$$

Hence the inequalities

(8) $$c_1 > c_2, \quad \sigma_2 \geq \sigma_1$$

are a sufficient condition that $\frac{d\bar{x}}{d\theta}$ shall exceed $\frac{dM}{d\theta}$, and since $\bar{x} = M = o$ when $\theta = o$, inequalities (8) are sufficient to insure that for positive values of θ, $\bar{x}$ will exceed M.

In the case of many frequency distributions whose form suggests dissection into two normal components it is found that the standard deviations of the smaller component exceeds that of the larger component. Hence, condition (8) is fulfilled, and $\bar{x}$ differs more from the mean of the larger component than does M.

3. *Relative Stability of Median and Mean for the Special Case, $\theta = o$.*

From equations (6) and (7) it is seen that while $\sigma_{\bar{x}}$ and σ_M are both monotone increasing functions of θ, they do not possess a monotone character with respect to the other parameters of equation (1). The development of general conditions which the parameters must satisfy in order that $\sigma_{\bar{x}}$ may exceed σ_M is impeded by

the fact that M is defined in (5) by an equation containing integrals, and its numerical value, for given values of the parameters, can be obtained only by interpolation in a table of areas under the normal curve. We shall therefore determine the relative stability of the median and arithmetic mean, as measured by the standard deviations of these averages, for certain special cases of distribution (1).

If, in equation (1), b is assigned the value zero, the distribution becomes symmetrical and $\bar{x} = M = 0$. Hence the condition for equal stability of median and arithmetic mean, $\sigma_{\bar{x}} = \sigma_M$, may in this special case be written

$$\sqrt{c_1\sigma_1^2 + c_2\sigma_2^2} = \sqrt{\frac{\pi}{2}}\,\frac{\sigma_1\sigma_2}{c_1\sigma_2 + c_2\sigma_1}.$$

Letting the ratio, $\frac{\sigma_2}{\sigma_1}$, be denoted by ρ, we obtain

$$(9)\quad f(\rho) = c_1^2c_2\rho^4 + 2c_1c_2^2\rho^3 + (c_1^3 + c_2^3 - \tfrac{\pi}{2})\rho^2 + 2c_1^2c_2\rho + c_1c_2^2 = 0.$$

This fourth degree equation in ρ possesses two positive real roots, independent of c_1, and c_2, for $f(0)$ and $f(\infty)$ are both positive, while

$$f(1) = (c_1 + c_2)^3 - \tfrac{\pi}{2} = 1 - \tfrac{\pi}{2} < 0.$$

Hence there exist two values, $\rho_1 < 1$ and $\rho_2 > 1$, such that when ρ assumes either of these values the standard deviations of the arithmetic mean and median are equal. For values of ρ in the interval $(\rho_1 < \rho < \rho_2)$, the standard deviation of the arithmetic mean is less than that of the median. For values of ρ outside this interval, the standard deviation of the arithmetic mean is greater than that of the median. Hence it is seen that, for $b = 0$, the relative stability of the median and arithmetic mean of distribution (1) is determined by the ratio of the standard deviations of the two component curves.

Yule[8] has discussed the relative stability of the median and arithmetic mean of distribution (1) when, in addition to the condition $b = 0$, the distribution is subjected to the further restriction

$$c_1 = c_2 = 0.5,$$

[8] Yule, G. U., *An Introduction to the Theory of Statistics*, 8th ed., p. 339.

and has obtained the numerical values of ρ for which the two averages will possess equal standard deviations:

$$\rho_1 = 0.4472 , \quad \rho_2 = 2.2360 .$$

4. Relative Stability of Median and Mean for the Special Case, $c_1 = c_2$.

Now let the restriction $b = 0$ be removed. Let b assume any positive value, and let the condition , $c_1 = c_2 = 0.5$, be imposed. The upper limits of the integrals in equation (5) will then be equal, whence

$$M = \frac{b\sigma_1}{\sigma_1 + \sigma_2} \qquad \bar{x} = \frac{b}{2}$$

$$\sigma_M = \frac{\sqrt{2\pi}\ \sigma_1\sigma_2\, e^{\frac{b^2}{2(\sigma_1+\sigma_2)^2}}}{\sqrt{5}\ (\sigma_1+\sigma_2)} \qquad \sigma_{\bar{x}} = \frac{\sqrt{2\sigma_1^2+2\sigma_2^2+b^2}}{2\sqrt{5}} .$$

The relative magnitude of the median and mean is seen to depend upon the standard deviations of the component distributions, and $\bar{x}$ is greater than, equal to, or less than M according as $\rho = \sigma_2/\sigma_1$ is greater than, equal to, or less than unity.

To obtain conditions for equal stability in the two averages, let $\sigma_{\bar{x}}$ be set equal to σ_M . By introducing the notation,

$$\rho = \sigma_2/\sigma_1 , \qquad k = b/(\sigma_1+\sigma_2) ,$$

this equation may be reduced to the form

$$(10) \quad (k^2+2)\rho^4 + (4k^2+4)\rho^3 + (6k^2+4-8\pi e^{k^2})\rho^2 + (4k^2+4)\rho + (k^2+2) = 0.$$

Taking $\lambda = (\rho + 1/\rho)$ as a new variable, this equation may be written as the quadratic

$$(k^2+2)\lambda^2 + (4k^2+4)\lambda + 4k^2 - 8\pi e^{k^2} = 0,$$

whose roots, are both real for all values of k^2. Furthermore, since $\pi e^{k^2} > 2(k^2+1)$ for all values of k^2, one of these roots is positive and greater than 2, and therefore has a value which $\lambda = (\rho + 1/\rho)$ may assume.

Hence, for all values of k^2 (and therefore for all values of b) there exist two reciprocal values of ρ , (ρ_1 and ρ_2), such

that when ρ assumes either of these values the standard deviations of the arithmetic mean and median are equal. For values of ρ in the interval ($\rho_1 < \rho < \rho_2$), the standard deviation of the arithmetic mean is less than that of the median, and for values of ρ not in this interval, the standard deviation of the arithmetic mean is greater than that of the median.

Yule's results show that when $b = 0$, $\rho_1 = 0.4472$ and $\rho_2 = 2.2360$, and therefore that the mean and median are equally stable when the standard deviation of one component is approximately 2.25 times that of the other. It remains to investigate the behavior of the interval ($\rho_1 < \rho < \rho_2$)as b varies.

Since it is the ratio of the standard deviations of the component curves, and not their actual values, which determines this interval, suppose the unit of measurement to be so chosen that $(\sigma_1 + \sigma_2) = 1$, whence $h = b$, and

$$\lambda = \frac{-2(b^2+1) + 2\sqrt{2\pi e^{b^2}(b^2+2) + 1}}{b^2+2}.$$

Then since $\pi e^{b^2}(b^2+1)^2 > 2$ for all values of b^2, λ may be shown to be a monotone increasing function of b^2, and therefore a monotone increasing function of b (positive). Since, furthermore, λ is an increasing function of ρ (when $\rho > 1$), it appears that the interval ($\rho_1 < \rho < \rho_2$) in which the standard deviation of the arithmetic mean is less than that of the median (i. e., the interval in which the mean is the more stable average), becomes larger as the size of b is increased.

Summarizing, for the special case of distribution (1) in which $\alpha_1 = \alpha_2$, (i. e., in which the areas of the two component normal curves are equal), the relative stability of the median and mean depends upon the value of ρ, the ratio of the standard deviations of the two component curves, and upon the value of b, the distance between their means. When ρ equals one, the mean is the more stable average, independent of b. Furthermore, for all positive values of b there exists an interval of values of ρ, including $\rho = 1$, within which the mean is more stable, at the end points of which the averages are equally stable, and without which the median is more stable. When $b = 0$, this interval is ($.4472 < \rho < 2.2360$), and as b increases the interval becomes larger.

It was stated at the beginning of this section that, on account of the approximation to the value of σ_M which is used, the conclusions will apply to distributions containing a large number of items. It should be noted that, in the special case which has just been considered, to assign a large value to b may cause the median to fall at a point of relatively small frequency, in which case, as will be shown in section IV, the approximation to the standard deviation of the median, $\sigma_M = 1/(2y_M\sqrt{s})$, will exceed its true value, and the superior stability of the arithmetic mean, as obtained from equation (10), may be exaggerated. In such cases the probable errors of the averages should be computed by the method of section V.

5. *Relative Stability of Median and Mean for the General Distribution* (1).

Finally, let the restriction $c_1 = c_2$ be removed from distribution (1). The condition that the standard deviations of the median and mean of this general distribution shall be equal may be written

$$\sqrt{c_1\sigma_1^2 + c_2\sigma_2^2 + c_1c_2b^2} = \sqrt{\frac{\pi}{2}}\,\frac{\sigma_1\sigma_2}{c_1\sigma_2 e^{-\frac{M^2}{2\sigma_1^2}} + c_2\sigma_1 e^{-\frac{(b-M)^2}{2\sigma_2^2}}}.$$

Let the notation

$$\rho = \frac{\sigma_2}{\sigma_1}, \qquad b = k\sqrt{\sigma_1\sigma_2}, \qquad q = e^{-\frac{M^2}{2\sigma_1^2}}, \qquad m = e^{-\frac{(b-M)^2}{2\sigma_2^2}}$$

be introduced. The parameters ρ, k, q, m may thus assume only positive values, and q and m are not greater than unity. Then

$$f(\rho) = c_1^2c_2q^2\rho^4 + (2c_1c_2^2qm + c_1^3c_2q^2k^2)\rho^3 + (c_2^3m^2 + 2c_1^2c_2^2qmk^2 + c_1^3q^2 - \tfrac{\pi}{2})\rho^2 + (c_1c_2^3m^2k^2 + 2c_1^2c_2qm)\rho + c_1c_2^2m^2 = 0.$$

This equation may possess two real positive roots. As ρ approaches zero or positive infinity, it is seen that $f(\rho)$ becomes positive, independent of c_1 and c_2, and therefore that the standard deviation of the mean exceeds that of the median. If the equation possesses two distinct positive roots, there will be an interval of positive values of ρ for which the standard deviation of the median will exceed that of the mean. However, this interval does not necessarily contain the value, $\rho = 1$, as in the special case where $c_1 = c_2$.

III.

The Relative Stability of the Median and Arithmetic Mean of a Symmetrical Frequency Distribution Which Is Composed of Three Normal Distributions.

1. The Relative Magnitude of the Standard Deviations of the Median and Mean.

It will be the purpose of this section to investigate the relative stability of the median and arithmetic mean of a symmetrical frequency distribution which may be dissected into three normal distributions, two of which possess equal areas and equal standard deviations and whose means are translated equal distances to left and right, respectively, of the mean of the third distribution.

The equation of the frequency function which describes such a distribution is of the form

$$(1) \qquad y = \frac{c_1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2\sqrt{2\pi}}\left(e^{-\frac{(x-\beta)^2}{2\sigma_2^2}} + e^{-\frac{(x+\beta)^2}{2\sigma_2^2}}\right),$$

where the areas of the components are connected by the relation

$$c_1 + 2c_2 = 1.$$

Since the distribution is symmetrical with respect to the y-axis, both the median and mean fall at the origin, and, if the approximation, $\sigma_M = 1/(2y_m\sqrt{s})$, is used, the standard deviations of the median and mean are readily expressed in terms of the parameters of equation (1):

$$\sigma_M = \frac{\sqrt{2\pi}\cdot\sigma_1\sigma_2}{2\sqrt{s}\cdot\left(c_1\sigma_2 + 2c_2\sigma_1 e^{-\frac{\beta^2}{2\sigma_2^2}}\right)},$$

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{s}}\cdot\sqrt{c_1\sigma_1^2 + 2c_2(\sigma_2^2+\beta^2)}.$$

To obtain conditions under which the two averages will be equally stable, let the notation,

$$\rho = \frac{\sigma_2}{\sigma_1}, \qquad n\sigma_2 = \beta,$$

be employed and let the standard deviations of median and mean be set equal to each other, whence we obtain the equation

$$\sqrt{1+2c_2[\rho^2(1+\pi^2)-1]} = \frac{\rho\sqrt{2\pi}}{2[\rho+2c_2(e^{-\frac{\pi^2}{2}}-\rho)]},$$

which, if we let

$$(e^{-\frac{\pi^2}{2}}-\rho) = g \quad \text{and} \quad [\rho^2(1+\pi^2)-1] = h,$$

may be written

$$(2)\quad f(c_2) = 16g^2hc_2^3 + 8g(2\rho h+g)c_2^2 + 4\rho(\rho h+2g)c_2 + \rho^2(2-\pi) = 0.$$

Since $c_1+2c_2=1$, only the positive real roots of equation (2) which are less than 0.5 are of interest. Independent of the positive value assigned to ρ and π, this equation may have not more than two real roots in this interval, for both $f(0)$ and $f(0.5)$ are less than zero when $\rho \neq 0$.

If equation (2) possesses two real, distinct, positive roots less than $c_2 = 1/2$, then there will be a subinterval of the interval $(0<c_2<0.5)$ within which the standard deviation of the arithmetic mean will exceed the standard deviation of the median, at the end points of which the averages will be equally stable, and without which the standard deviation of the median will exceed that of the arithmetic mean. If the equation has no real roots in this interval, the standard deviation of the median will exceed that of the arithmetic mean throughout the interval.

The tangents to the curve whose equation is (2) are horizontal when $c_2 = -\rho/2g$ and when $c_2 = (-\rho h-2g)/6gh$. Since $f(-\rho/2g)$ is negative for all values of ρ other than zero, $-\rho/2g$ is a value of c_2 for which the standard deviation of the median is greater than the standard deviation of the arithmetic mean. If there exists an interval of values of c_2 for which the standard deviation of the arithmetic mean is greater than the standard deviation of the median, it will contain the value $c_2 = (-\rho h-2g)/6gh$. Therefore the condition that such an interval exist is

$$0 < \frac{-\rho h-2g}{6gh} < \frac{1}{2}, \qquad f\left(\frac{-\rho h-2g}{6gh}\right) > 0.$$

If the value, $c_2=(-\rho h-2g)/6gh$ lies in the interval $(0<c_2<0.5)$ and if $f(-\rho h-2g)/6gh = 0$, then equation (2) will possess a double root, and the standard deviation of the median will equal the standard deviation of the arithmetic mean for a single value of c_2, $c_2=(-\rho h-2g)/6gh$, and will exceed it for all other values of c_2 . The condition under which $f(-\rho h-2g)/6gh$ will vanish is that $g/\rho h$ assume one of the values: 4.67284, —1.53327, —0.13957, for letting $c_2=(-\rho h-2g)/6gh$, equation (2) becomes

$$8(g-\rho h)^3-27\pi\rho^2h^2g=0, \tag{3}$$

which may be written as a cubic equation in $g/\rho h$ whose roots have the above values.

2. *The Dissection of a Frequency Distribution into Three Normal Components.*

In order to apply the above conclusions in determining the relative stability of the averages of a particular sequence of economic data, it is necessary that the data be dissected into three normal distributions. A general method of determining the values of the five parameters of equation (1) from given frequency data will therefore be developed.

Karl Pearson[9] has described a method for dissecting an asymmetrical frequency curve into two normal curves. He obtains expressions for the first five moments of the curve, which he solves, after lengthy algebraic manipulation, for the parameters. A similar procedure, the solution of moment equations, may be applied to a dissection into three normal curves. However, since the distribution has been assumed to be symmetrical, expressions for the odd moments vanish identically. Hence it is necessary to use moments as high as the eighth in order to obtain five equations from which the values of the parameters may be determined. While Pearson's method of setting up the moment equations may be used, his method of solution will not carry over to this case.

[9] Loc. cit.

Given a frequency distribution of the variable x whose origin has been chosen at the arithmetic mean of the distribution, let M_k denote the k^{th} moment of the distribution, and let M_k be set equal to the corresponding moment of the theoretical distribution whose equation is (1). We have, then, as the equations from which the five parameters of distribution (1) may be determined:

(4)
$$c_1 + 2c_2 = 1$$
$$c_1\sigma_1^2 + 2c_2(\sigma_2^2 + b^2) = M_2$$
$$3c_1\sigma_1^4 + 2c_2(3\sigma_2^4 + 6b^2\sigma_2^2 + b^4) = M_4$$
$$15c_1\sigma_1^6 + 2c_2(15\sigma_2^6 + 45b^2\sigma_2^4 + 15b^4\sigma_2^2 + b^6) = M_6$$
$$105c_1\sigma_1^8 + 2c_2(105\sigma_2^8 + 420b^2\sigma_2^6 + 210b^4\sigma_2^4 + 28b^6\sigma_2^2 + b^8) = M_8.$$

Instead of carrying through the solution of these five equations, it has been found convenient to assign to σ_1 a value equal to the standard deviation of a central group of items, and to retain only the first four moment equations to be solved for the other four parameters. Later the five equations will be used to correct this estimated value of σ_1 .

If we let M_k' denote the k^{th} moment of the given distribution with σ_1 as unit, denote b/σ_2 by u, σ_2/σ_1 by ρ, and eliminate c_2 from these equations, we obtain

(5)
$$c_1 + (1-c_1)\rho^2(1+u^2) = M_2'$$
$$3c_1 + (1-c_1)\rho^4(3+6u^2+u^4) = M_4'$$
$$15c_1 + (1-c_1)\rho^6(15+45u^2+15u^4+u^6) = M_6'.$$

Now eliminating c_1, this system of equations reduces to

$$3[M_2' - \rho^2(1+u^2)] + (1-M_2')\rho^4(3+6u^2+u^4) = M_4'[1-\rho^2(1+u^2)]$$
$$15[M_2' - \rho^2(1+u^2)] + (1-M_2')\rho^6(15+45u^2+15u^4+u^6) = M_6'[1-\rho^2(1+u^2)],$$

and when the notation $\alpha = (1-M_2')$, $\beta = (M_4'-3)$, $\gamma = (3M_2'-M_4') = -3\alpha-\beta$,

$$\delta = (M_6'-15), \qquad \epsilon = (15M_2'-M_6') = -15\alpha-\delta,$$

is introduced and the equations are written in descending powers of ρ^2 they become

$$(6)\qquad \begin{aligned} &\rho^4\alpha(3+6u^2+u^4)+\rho^2\beta(1+u^2)+\gamma = 0,\\ &\rho^6\alpha(15+45u^2+15u^4+u^6)+\rho^2\delta(1+u^2)+\epsilon = 0. \end{aligned}$$

Let Sylvester's method of elimination be applied to these two equations, making use of the property that the resultant, R, of the equations $a_0x^3+a_1x^2+a_2x+a_3=0$ and $b_0x^2+b_1x+b_2=0$ is

$$R = \begin{vmatrix} (a_0b_2) & (a_1b_2)-a_3b_0 & -a_3b_1 \\ (a_0b_1) & (a_0b_2) & -a_3b_0 \\ b_0 & b_1 & b_2 \end{vmatrix},$$

where (a_ib_j) denotes $a_ib_j-a_jb_i$.[10]

Let $(1+u^2)=f_1(u^2)$, $(3+6u^2+u^4)=f_2(u^2)$, $(15+45u^2+15u^4+u^6)=f_3(u^2)$. Then

$$R = \begin{vmatrix} \alpha\gamma f_3-\gamma\delta f_1f_2 & -3\delta f_1^2-\gamma\epsilon f_2 & -\beta\epsilon f_1 \\ \gamma\beta f_1f_3 & \gamma\gamma f_3-\alpha\delta f_1f_2 & -\alpha\epsilon f_2 \\ \alpha f_2 & 3f_1 & \gamma \end{vmatrix} = 0,$$

and expanding and simplifying this determinant we obtain

$$\begin{aligned} &(3\alpha\beta\gamma\epsilon-2\alpha\delta\gamma^2)f_1f_2f_3+(\alpha\gamma\delta^2-\gamma 3\delta\epsilon)f_1^2f_2^2\\ &+(\beta^2\delta\gamma-\beta^3\epsilon)f_1^3f_3+\gamma\gamma^3f_3^2+\gamma^2\epsilon^2f_2^3 = 0. \end{aligned}$$

Since $\alpha, \beta, \gamma, \delta, \epsilon$ are constants, this equation may be written

$$Af_1f_2f_3+Bf_1^2f_2^2+Cf_1^3f_3+Df_3^2+Ef_2^3 = 0.$$

If, finally, f_1, f_2, and f_3 are replaced by their values as functions of u^2, this equation reduces to

$$\begin{aligned} (7)\quad &u^{12}(A+B+C+D+E)+u^{10}(22A+14B+18C+30D+18E)\\ &+u^8(159A+67B+93C+315D+117E)\\ &+u^6(468A+132B+196C+1380D+324E)\\ &+u^4(555A+123B+195C+2475D+351E)\\ &+u^2(270A+54B+90C+1350D+162E)\\ &+(45A+9B+15C+225D+27E) = 0, \end{aligned}$$

[10] Dickson, L. E., *First Course in Theory of Equations*, p. 150.

a sixth degree equation in u^2 upon which the complete solution of the problem now turns, for having obtained a value of u^2 from equation (7), the values of ρ^2 and c_1 may be determined from equations (6) and (5), respectively. Since $u = b/\sigma_2$, $\rho = \sigma_2/\sigma_1$, and $c_1 + 2c_2 = 1$, the parameters of equation (1) may be obtained.

It will be recalled that equation (7) has been obtained from the first four of equations (4), and that we have employed this equation to obtain values of c_1, c_2, σ_2, b corresponding to an assigned value of σ_1. This estimated value of σ_1 may be corrected, and corresponding corrections to the values of the other four parameters may be obtained, by use of the five equations (4).

Let equations (4) be written

$$f_i(c_1, c_2, \sigma_1, \sigma_2, b) = M_{2i-2} \qquad (i = 1, 2, 3, 4, 5).$$

Let c_1', c_2', σ_2', b' denote the values which the four parameters take on when σ_1 is assigned the value σ_1'. Let Δc_1, Δc_2, $\Delta \sigma_1$, $\Delta \sigma_2$, Δb denote the respective corrections which should be applied. Then using Taylor's theorem and neglecting terms which contain derivatives of higher order than the first, we obtain five linear equations in the five corrections:

$$f_i(c_1, c_2, \sigma_1, \sigma_2, b) = f_i(c_1' + \Delta c_1, c_2' + \Delta c_2, \sigma_1' + \Delta \sigma_1, \sigma_2' + \Delta \sigma_2, b' + \Delta b)$$

$$= f_i(c_1', c_2', \sigma_1', \sigma_2', b') + \Delta c_1 \frac{\partial f_i}{\partial c_1} + \Delta c_2 \frac{\partial f_i}{\partial c_2} + \Delta \sigma_1 \frac{\partial f_i}{\partial \sigma_1} + \Delta \sigma_2 \frac{\partial f_i}{\partial \sigma_2} + \Delta b \frac{\partial f_i}{\partial b} = M_{2i-2}.$$

The corrected values of the parameters, $c_1' + \Delta c_1$, $c_2' + \Delta c_2$, $\sigma_1' + \Delta \sigma_1$, $\sigma_2' + \Delta \sigma_2$, $b' + \Delta b$ may be regarded as second approximations to their true values, and further approximations may be obtained in the same fashion.

IV.

The Standard Deviation of the Medians of Small Samples.

1. *The Classical Approximation to the Standard Deviation of the Median.*

In the preceding sections an approximation to the standard deviation of the median has been used, and the conclusions have

been assumed to be valid only when s, the number of items in the sample, is large. We wish, in the present section, to examine this approximation, and to compare the results which it produces with those obtained when other methods of determining the standard deviation of the median are employed.

The formula ordinarily used to compute the standard deviation of the medians of samples of s items each, drawn from a frequency distribution whose equation is $y = f(x)$ and which satisfies the condition

$$(1) \qquad \int_{-\infty}^{0} f(x)\,dx = 0.5 = \int_{0}^{\infty} f(x)\,dx$$

is

$$(2) \qquad \sigma_M = \frac{1}{2 \cdot f(0) \cdot \sqrt{s}} .$$ [11]

That this formula gives only an approximation to the true value of the standard deviation of the median and that the approximation may be rather poor for distributions of certain types is clear from the following derivation of the formula.

Let samples containing s items each be drawn from the distribution $y = f(x)$ which satisfies condition (1). Let the proportion of items above $x = 0$ in each sample be denoted by $(0.5 + d)$. These observed values will tend to cluster around 0.5 as a mean, with a standard deviation of $\frac{1}{2\sqrt{s}}$. Let the deviation of the median of a sample from the median of the theoretical distribution, $x = 0$, be denoted by e. Then if the number of items in the sample is sufficiently large to justify us in assuming that d is so small that we may regard the element of the frequency curve whose base is the interval $(0, e)$, and whose area is d, as approximately a rectangle, we may write

$$e = \frac{d}{f(0)}, \quad \text{whence} \quad \sigma_M = \frac{\sigma_d}{f(0)} = \frac{1}{2 \cdot f(0) \cdot \sqrt{s}} .$$

[11] Rietz, H. L., *Mathematical Statistics*, Carus Monograph III, p. 134. Yule, G. U., *Introduction to the Theory of Statistics*, 8th ed., p. 337.

This replacement of an element of a frequency curve by a rectangle can be justified only when e, the deviation of the median of the sample from the median of the theoretical distribution, is small. Hence there is reason to doubt whether formula (2) will give a close approximation to the value of the standard deviation of the median of samples which do not contain a large number of items. The formula would seem particularly untrustworthy when applied to a theoretical distribution in which the median falls at a point of relatively small frequency.

An expression for the standard deviation of the median which is not liable to the inaccuracies of approximation (2) may be derived as follows. Given the frequency function $y = f(x)$ which satisfies condition (1), if a sample of $(2n+1)$ items is drawn from this distribution, the probability that an item will fall in the interval $(x, x+dx)$ approaches the limit $y\,dx$ as dx approaches zero, the probability that an item will fall below x is $\int_{-\infty}^{x} f(x)\,dx$, and the probability that an item will fall above x is $\int_{x}^{\infty} f(x)\,dx$. Hence the limit, as dx approaches zero, of the probability that the median of the sample will fall in the interval $(x, x+dx)$ is

$$(2n+1)_{2n}C_n \left[\int_{-\infty}^{x} f(x)\,dx\right]^n \cdot \left[\int_{x}^{\infty} f(x)\,dx\right]^n f(x)\,dx,$$

and the square of the standard deviation of the median may be obtained from the equation

$$(3)\qquad \sigma_M^2 = (2n+1)_{2n}C_n \int_{-\infty}^{\infty} x^2 f(x)\left[\int_{-\infty}^{x} f(x)\,dx\right]^n \cdot \left[\int_{x}^{\infty} f(x)\,dx\right]^n dx.$$

The integrations involved in this equation may be difficult to perform unless $f(x)$ is a simple function. Hence, we consider the rectangular distribution whose equations are

$$f(x) = 1, \quad \left(-\tfrac{1}{2} \le x \le \tfrac{1}{2}\right); \qquad f(x) = 0, \quad \left(x < -\tfrac{1}{2}\right), \left(x > \tfrac{1}{2}\right),$$

and obtain

$$\sigma_M^2 = (2n+1)_{2n}C_n \int_{-1/2}^{1/2} x^2 (0.25 - x^2)^n\,dx = \frac{1}{4(2n+3)}.$$

If we denote by σ'_M the approximation to the value of the standard deviation of the median obtained using formula (2), we have for this distribution $\sigma'_M = \frac{1}{2 \cdot f(0) \cdot \sqrt{2n+1}} = \frac{1}{2\sqrt{2n+1}}$, whence we have the relation $\sigma_M = \sigma'_M \sqrt{\frac{2n+1}{2n+3}}$.

It is observed that the approximation, σ'_M, exceeds the true value, σ_M, for all values of n, but that the error factor approaches unity as n increases, and is close to unity even for fairly small values of n.

2. *A General Method of Obtaining Upper and Lower Limits of the Standard Deviation of the Median.*

For distributions composed of two normal components the integrations involved in equation (3) can be performed only approximately, and this equation will serve only to determine upper and lower limits of the true value of the standard deviation of the median. A more straightforward method of obtaining these upper and lower limits, and one which is applicable to any frequency distribution, will be followed.

Let x_i denote the deviation of the i^{th} percentile of a distribution from the median, and let p_i denote the probability that the median of a sample of s items will fall between the i^{th} and $(i+1)^{th}$ percentiles of the distribution from which the samples are drawn. Then a lower limit of the standard deviation of the medians of samples containing s items drawn from this distribution is given by the expression

$$\left[\sum_{i=1}^{49} x_i^2 p_{i-1} + \sum_{i=51}^{99} x_i^2 p_i\right]^{1/2},$$

and an upper limit, by the expression

$$\left[\sum_{i=0}^{49} x_i^2 p_i + \sum_{i=51}^{100} x_i^2 p_{i-1}\right]^{1/2},$$

where, in the case of a distribution in which the zeroth or hundredth percentile is at an infinite distance from the median,

x_0 denotes the largest value of x for which it is true that $\int_{-\infty}^{x_0} f(x)\,dx < \epsilon$, and x_{100} denotes the smallest value of x for which it is true that $\int_{x_{100}}^{\infty} f(x)\,dx < \epsilon$, where ϵ is an arbitrarily small positive constant.

The values of x_i depend on the distribution, and are independent of the number of items in the sample. The values of p_i depend on the number of items in the sample and are independent of the form of the distribution. Approximations to the values of p_i, $(i = 0, 1, 2, \cdots, 99)$, may be obtained by use of the DeMoivre-Laplace theorem[12]. In our notation the theorem may be stated:

The probability that m or more of the items of a sample containing $(2m-1)$ items will fall to the right of the i^{th} percentile of the distribution from which the sample is drawn is

$$P_i = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx \quad \text{where} \quad z = \frac{m-(2m-1)(1-.01i)-0.5}{\sqrt{(2m-1)(.01\,i)(1-.01i)}}.$$

Then $p_i = P_i - P_{i+1}$.

Tables[13] of values of p_i for samples containing 7 and 51 items have been computed, and have been used in calculating upper and lower limits of the standard deviation of the medians of samples containing 7 and 51 items drawn from the distributions whose equations are

$$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$f_2(x) = \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x+2)^2}{2}} + e^{-\frac{(x-2)^2}{2}} \right),$$

$$f_3(x) = \frac{1}{\sqrt{2\pi}} \left(e^{-\frac{(4/3\,x)^2}{2}} + e^{-\frac{(4x)^2}{2}} \right).$$

[12] Rietz, H. L., *Mathematical Statistics*, Carus Monograph III, p. 35.

[13] These tables are included in the author's dissertation, which is filed in the library of the University of Wisconsin.

Approximations to the value of the standard deviation of the median have also been obtained using formula (2). The results are tabulated below.

STANDARD DEVIATION OF THE MEDIAN

	7 items.			51 items.		
	Upper Limit	*Lower Limit*	*Formula (2)*	*Upper Limit*	*Lower Limit*	*Formula (2)*
$f_1(x)$	0.4715	0.4462	0.4737	0.1838	0.1631	0.1755
$f_2(x)$	1.4701	1.4114	3.5003	0.8551	0.7733	1.2968
$f_3(x)$	0.2683	0.2521	0.2368	0.0938	0.0831	0.0877

We conclude that, when applied to samples containing a fairly small number of items, the results obtained using the customary formula for the standard deviation of the median may be very untrustworthy, particularly for a distribution in which the median falls at a point of relatively small frequency,

We therefore shall propose another method for the comparison of the stability of the arithmetic mean and median, one which does not involve the computation of the standard deviation of these averages.

V.

THE RELATIVE STABILITY OF THE MEDIAN AND ARITHMETIC MEAN, DETERMINED FROM THE FREQUENCY DISTRIBUTIONS OF THESE AVERAGES.

1. *The Frequency Distributions of the Median and Mean.*

Since the true value of the standard deviation of medians of samples containing items, drawn from a frequency distribution which is composed of two normal distributions, is not easily determinable, and since the customary approximation is not sufficiently accurate to justify its use in the study of small samples drawn from a distribution of this type, we shall develop a method of comparing the relative stability of the median and arithmetic mean, based not on the standard deviations of these two averages but on their frequency distributions.

Another consideration, aside from expediency, motivates the development of this method, for even if the standard deviations of the arithmetic mean and median could be accurately computed, they would not determine the relative stability of the two averages unless it is assumed that the frequencies of the mean and median are distributed in the same fashion. If, however, the equations of the frequency curves of the mean and median of samples of s items drawn from a given distribution are determined, then by comparing the deviations from the median of corresponding percentiles of these two averages, a judgment as to the relative stability of the two averages may be formed.

We shall assume the frequency curve of the arithmetic means of samples of $(2n+1)$ items to be normal, independent of the form of the theoretical distribution from which the samples are drawn, and to possess a standard deviation of $\sigma/\sqrt{2n+1}$, where σ is the standard deviation of the theoretical distribution.[14] We proceed to determine the equation of the frequency curve of the medians of these samples.

Let the equation of the original distribution be $y = f(x)$, and let the condition

$$\int_{-\infty}^{0} f(x)\,dx = \tfrac{1}{2} = \int_{0}^{\infty} f(x)\,dx$$

be satisfied. Then the probability that the median of a sample of $(2n+1)$ items will fall in the interval $(x, x+dx)$ is the product of the probabilities that an item will fall in this interval and that of the remaining $2n$ items, n will fall above this interval and n below this interval. We let y_M denote the frequency function according to which the medians of the samples are distributed, and obtain

$$\begin{aligned} y_M &= (2n+1)\,{}_{2n}C_n \cdot f(x)\left[\int_{-\infty}^{x} f(x)\,dx\right]^n \cdot \left[\int_{x}^{\infty} f(x)\,dx\right]^n dx \\ &= (2n+1)\,{}_{2n}C_n \cdot f(x)\left[\tfrac{1}{2} + \int_{0}^{\infty} f(x)\,dx\right]^n \cdot \left[\tfrac{1}{2} - \int_{0}^{x} f(x)\,dx\right]^n dx \\ &= (2n+1)\,{}_{2n}C_n \cdot f(x)\left\{.25 - \left[\int_{0}^{x} f(x)\,dx\right]^2\right\}^n dx, \end{aligned} \tag{1}$$

[14] Rietz, H. L., *Mathematical Statistics*, Carus Monograph III, p. 127.

[15] A similar expression for the probability density of the median is

2. *The Stability of an Average Determined from Its Probable Error.*

Expressions for the frequency functions of the median and mean of samples containing $(2n+1)$ items having been determined, we may form a judgement as to the relative stability of these two averages for a given distribution by comparing the deviations, from the median of the given distribution, of corresponding percentiles of the two averages. If some definite criterion of relative stability is desired, it seems natural to select the probable errors of the averages, where the term (probable error) is understood to have its original meaning, and not to denote a fixed multiple of the standard deviation of the average. We shall therefore proceed to determine the deviation, from the median, of a given percentile of the frequency distribution of medians of samples containing $(2n+1)$ items drawn from the distribution whose equation is $y=f(x)$.

found in a paper by E. L. Dodd (Functions of Measurements under General Laws of Error, *Skandinavisk Aktuarietidskrift*, 1922, p. 150), and is there used in comparing the relative stability of the median and arithmetic mean of certain theoretical frequency distributions. However, the method used in Dodd's paper is to compare the probability densities of the two averages at the median of the original distribution, rather than to compare the deviations from the median of specific percentiles of the frequency curves of the two averages, as we shall do. Dodd uses Stirling's formula to obtain an approximation to the probability density at the median,

$$y_M(0) = \sqrt{\frac{2(2n+1)}{\pi}}\, f(0),$$

and represents the probability density of the arithmetic mean at the same point by the expression

$$y_{\bar{x}}(0) = \frac{\sqrt{2n+1}}{\sigma\sqrt{2\pi}},$$

where σ is the standard deviation of the original distribution.

It is readily seen that this method of comparison, when applied to small samples, would lead to exactly the same inaccuracies that would result if the relative stability of the two averages were determined by comparing their standard deviations, the customary approximation formula being used to obtain the value of the standard deviation of the median, since

$$\frac{y_M(0)}{y_{\bar{x}}(0)} = \sqrt{\frac{2(2n+1)}{\pi}}\, f(0) \div \frac{\sqrt{2n+1}}{\sigma\sqrt{2\pi}} = \frac{2\sqrt{2n+1}\, f(0)}{\frac{\sqrt{2n+1}}{\sigma}} = \frac{\sigma_{\bar{x}}}{\sigma_{M'}}.$$

Let S denote that fraction of the area under the frequency curve of the medians which is bounded by ordinates drawn to the curve at the points $x=0$ and $x=\ell$. Our problem is to determine the value of ℓ which corresponds to an assigned value of S, and from (1) the relationship between ℓ and S is seen to be expressible in the form

$$(2)\qquad S = (2n+1)_{2n}C_n \int_0^{\ell} f(x)\left\{\frac{1}{4}-\left[\int_0^{x} f(x)\,dx\right]^2\right\}^n dx,$$

where S may be assigned any value in the interval $(0 \le S \le 0.5)$.

If the transformation

$$t = 2\int_0^{x} f(x)\,dx$$

be applied to equation (2) it becomes

$$(3)\qquad S = \frac{(2n+1)_{2n}C_n}{2^{2n+1}} \int_0^{\alpha} (1-t^2)^n\,dt, \quad \text{where } \alpha \text{ corresponds to } \ell.$$

Then

$$(4)\qquad \frac{S\,2^{2n+1}}{(2n+1)_{2n}C_n} = \frac{S\cdot 2^{2n+1}\cdot (n!)^2}{(2n+1)!} = \int_0^{\alpha} (1-t^2)^n\,dt$$

$$= \alpha - \frac{n}{3}\alpha^3 + \frac{n(n-1)}{2!\cdot 5}\alpha^5 - \frac{n(n-1)(n-2)}{3!\cdot 7}\alpha^7 + \cdots \pm \frac{\alpha^{2n+1}}{2n+1},$$

and using Stirling's approximation, we obtain

$$(5)\qquad \frac{2S\sqrt{\pi n}}{(2n+1)} = \int_0^{\alpha} (1-t^2)^n\,dt$$

$$= \alpha - \frac{n}{3}\alpha^3 + \frac{n(n-1)}{2!\cdot 5}\alpha^5 - \frac{n(n-1)(n-2)}{3!\cdot 7}\alpha^7 + \cdots \pm \frac{\alpha^{2n+1}}{2n+1}.$$

It is observed from equation (3) that, for a fixed value of n, α is a monotone increasing function of S, and that $\alpha = 0$ when $S = 0$, and $\alpha = 1$ when $S = 0.5$. It is also observed that, for a fixed value of S, α is a monotone decreasing function of

n, and that $\lim_{n\to\infty} \alpha = 0$.

Unless S is assigned a value near 0.5, we may obtain an approximation to the value of α from equation (5) by neglecting terms containing powers of α higher than the third. We wish to determine the degree of approximation which is introduced by dropping terms after the second from the second member of equation (5). To this end, we shall first ascertain an interval of values of S for which it is true that α is not greater than the simple, decreasing function of n, $1/\sqrt{n}$; that is, we shall determine the interval of values of S which satisfy the inequality,

$$\frac{2S\sqrt{\pi n}}{2n+1} \leq \frac{1}{\sqrt{n}} - \frac{n}{3}\frac{1}{n\sqrt{n}} + \frac{n(n-1)}{2!\,5}\frac{1}{n^2\sqrt{n}} - \frac{n(n-1)(n-2)}{3!\,7}\frac{1}{n^3\sqrt{n}} + \cdots \pm \frac{1}{(2n+1)\,n^n\sqrt{n}},$$

or

$$S \leq \frac{2n+1}{2n\sqrt{\pi}}\left[1 - \frac{1}{3} + \frac{n-1}{2!\,5n} - \frac{(n-1)(n-2)}{3!\,7n^2} + \cdots \pm \frac{1}{(2n+1)\,n^n}\right].$$

The second member of this inequality is greater than

$$\frac{1}{\sqrt{n}}\left[1 - \frac{1}{3} + \frac{n-1}{2!\,5n} - \frac{(n-1)(n-2)}{3!\,7n^2} + \cdots \pm \frac{1}{(2n+1)\,n^n}\right],$$

and since the terms of the finite alternating series within parentheses obviously decrease in numerical value, their sum will exceed $2/3$ for all positive values of n. Therefore the inequality

$$\alpha \leq \frac{1}{\sqrt{n}}$$

will certainly be satisfied for all values of s in the interval

$$|S| \leq \frac{2}{3\sqrt{\pi}} = 0.3761,$$

and therefore α is not greater than $1/\sqrt{n}$ when S is assigned a value corresponding to a percentile of the frequency distribution of the medians between the 13th and 87th percentiles. Certainly in determining the first and third quartiles of the frequency distribution of the medians, α will be less than $1/\sqrt{n}$.

Since in equation (5) the value of S is given by a finite alternating series whose terms do not increase in numerical value,

the error involved in neglecting terms after the second will be less than the first term neglected:

$$\frac{2n+1}{2\sqrt{\pi n}}\,\frac{n(n-1)}{2!\,5}\,\alpha^5.$$

But, when $|S| \leq 0.3761$, it is true that

$$\frac{2n+1}{2\sqrt{\pi n}}\,\frac{n(n-1)}{2!\,5}\,\alpha^5 \leq \frac{2n+1}{2\sqrt{\pi n}}\,\frac{n(n-1)}{2!\,5}\,\frac{1}{n^2\sqrt{n}} = \frac{2n^2-n-1}{20\,n^2\sqrt{\pi}}$$

$$= \frac{1}{10\sqrt{\pi}}\left(1-\frac{1}{2n}-\frac{1}{2n^2}\right) < \frac{1}{10\sqrt{\pi}} = 0.0564.$$

We conclude, therefore, that the value of α obtained from equation (5) by neglecting powers of α higher than the third corresponds to a percentile of the frequency distribution of the median which differs from the assigned value of S by not more than 0.05.

We see, then, that an approximation to the value, ℓ, of a given percentile of the frequency distribution of the medians of samples containing $(2n+1)$ items drawn from the theoretical distribution whose equation is $y = f(x)$ may be obtained by solving for α the third degree equation

$$\frac{2S\sqrt{\pi n}}{2n+1} = \alpha - \frac{n}{3\ell}\alpha^3, \tag{6}$$

where

$$\alpha = 2\int_0^{\ell} f(x)\,dx.$$

The tables of values of p_i mentioned in the preceding section afford a check on the accuracy of the results of equation (6) when $(2n+1)$ is assigned the values 7 and 51. From these tables it is observed that the third quartile of the frequency distribution of the medians of samples containing 7 items falls at the 62nd percentile of the theoretical distribution from which the samples are drawn, and that for samples containing 51 items the third quartile of the medians falls near the 55th percentile of the original distribution. Hence, when $S = 0.25$, the value of $\int_0^x f(x)\,dx$, accurate to two places of decimals, is 0.12 when

$(2n+1)$ equals 7, and 0.05 when $(2n+1)$ equals 51.

In equation (6) let

$$K = \frac{2S\sqrt{\pi n}}{2n+1},$$

whence we obtain

$$\frac{n}{3}\alpha^3 - \alpha + K = 0.$$

Letting $\alpha = K + \lambda$, this equation becomes

$$\frac{n}{3}\left(K^3 + 3K^2\lambda + 3K\lambda^2 + \lambda^3\right) - \lambda = 0,$$

or as an approximation

$$\lambda = \frac{\frac{n}{3}K^3}{1 - nK^2}.$$

Assigning to n the values 7 and 51 we obtain the following results:

$2n+1$	K	λ	α	$\int_0^x f(x)dx$	Computed Value of $\int_0^x f(x)dx$
7	.2193	.0123	.2316	.1158	.12
51	.0869	.0072	.0941	.0471	.05

Thus we have developed a method of determining the probable error (or any percentile) of the median, which possesses the double advantage of being applicable to distributions which do not contain a very large number of items, and of being applied easily to any distribution, for after (6) has been used to determine the value of α, the corresponding value of ℓ may be obtained either from a table of integrals of a theoretical frequency function, or from an actual distribution by cumulating frequencies beyond the median.

The calculation of the probable error (or any percentile) of the arithmetic mean offers no difficulty if we assume the means of samples to be normally distributed. The relative stability of the two averages may then be determined by comparing the probable errors (or corresponding percentiles) of the two averages. This method of comparison will be applied to a particular distribution in section VI of this paper.

A table[16] of values of α and κ for certain assigned values of n and S is given below.

Values of α and κ when $\kappa = \frac{2S\sqrt{\pi n}}{2n+1} = \int_0^{\alpha}(1-t^2)^n\,dt$

n	S	κ	α
10	.05	.02669	.027
	.10	.05338	.054
	.15	.08007	.082
	.20	.10676	.111
	.25	.13345	.143
	.30	.16014	.177
	.35	.18683	.217
	.40	.21352	.265
	.45	.24021	.334
25	.05	.017377	.017
	.10	.034754	.035
	.15	.052131	.053
	.20	.069508	.073
	.30	.104262	.116
	.35	.121639	.143
	.40	.139016	.176
	.45	.156393	.223
50	.05	.012409	.012
	.10	.024818	.025
	.15	.037227	.038
	.20	.049636	.052
	.25	.062045	.067
	.30	.074454	.083
	.35	.086863	.102
	.40	.099272	.126
	.45	.111681	.161
100	.05	.008818	.0088
	.10	.017636	.018
	.15	.026455	.027
	.20	.035273	.037
	.25	.044091	.047
	.30	.052909	.059
	.35	.061727	.073
	.40	.070546	.090
	.45	.079364	.115

[16] Computed by Miss Beatrice Berberich, university computer, University of Wisconsin.

VI.

An Application to a Particular Sequence of Economic Data of Various Methods of Comparing the Stability of the Arithmetic Mean and Median.

1. *Dissection into a Symmetrical Distribution Composed of Two Normal Distributions.*

In a paper by W. L. Crum[17] a particular sequence of economic data has been examined with the purpose of determining the relative stability of its median and arithmetic mean. The series studied comprises the monthly link relatives of the rate of interest on 60-90 day commercial paper from January, 1890, to January, 1917. A frequency distribution of deviations from their medians of the link relatives for each month is reproduced below, together with the values of the first six moments of the distribution.

Frequencies of Deviations from the Medians

Dev.	*Freq.*	*Dev.*	*Freq.*	*Dev.*	*Freq.*	*Dev.*	*Freq.*	*Dev.*	*Freq.*
—37	1	—18	2	—7	6	4	19	15	1
—32	1	—17	2	—6	23	5	13	16	1
—30	1	—16	2	—5	10	6	13	17	1
—29	1	—15	1	—4	13	7	8	18	2
—28	1	—14	3	—3	19	8	6	23	1
—24	1	—13	6	—2	9	9	5	24	1
—23	1	—12	3	—1	11	10	2	28	1
—22	1	—11	6	0	28	11	4	34	1
—21	2	—10	3	1	22	12	3	35	2
—20	1	— 9	5	2	22	13	1	41	2
—19	2	— 8	11	3	13	14	2	42	1
								45	1

$N = 324$

Moments	*About 0 Deviation*	*About* $\bar{x}$	*With Sheppard Adjustments*
1	—0.46	0.00	0.00
2	107.06	106.85	106.77
3	656.45	793.68	793.68
4	83520	84860	84465
5	3015000	3209000	3208000
6	110890000	119480000	119370000

[17] Loc. cit.

Professor Crum's method of attack is to dissect the series, according to Pearson's method, into two normal components whose means are coincident. He therefore fits to the data a curve whose equation is

(1) $$y = \frac{324}{\sqrt{2\pi}}\left(\frac{c_1}{\sigma_1}e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2}e^{-\frac{x^2}{2\sigma_2^2}}\right),$$

and obtains for the parameters the values:

$$c_1 = .26, \qquad \sigma_1 = 2.46, \qquad c_2 = .74, \qquad \sigma_2 = 11.9$$

This theoretical distribution is of the type discussed in paragraph 3, section II. Its median and mean will be equally stable if $\rho = \sigma_2/\sigma_1$ satisfies the equation

$$c_1^2 c_2 \rho^4 + 2c_1 c_2^2 \rho^3 + (c_1^3 + c_2^3 - \tfrac{\pi}{2})\rho^2 + 2c_1^2 c_2 \rho + c_1 c_2^2 = 0$$

Letting $c_1 = 0.25$ and $c_2 = 0.75$, this equation reduces to

$$\rho^4 + 6\rho^3 - 24\rho^2 + 2\rho + 3 = 0,$$

which has a root between 2.5 and 2.6. Since for the distribution under consideration $\rho = 4.8$, the standard deviation of the arithmetic mean is larger than the standard deviation of the median, and the median is the more stable average.

2. *Dissection into an Asymmetrical Distribution Composed of Two Normal Distributions.*

In the method of dissection employed by Professor Crum, the slight positive skewness which the distribution possesses is ignored. We shall dissect the data into two normal components whose means are not equal, and investigate the relative stability of the median and mean of the resulting asymmetrical distribution:

(2) $$y = \frac{324}{\sqrt{2\pi}}\left[\frac{c_1}{\sigma_1}e^{-\frac{(x-\theta_1)^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2}e^{-\frac{(x-\theta_2)^2}{2\sigma_2^2}}\right].$$

Pearson's method of dissecting an asymmetrical distribution depends on the solution of his "fundamental nonic,"

$$24p_2^9 - 28\lambda_4 p_2^7 + 36\mu_3^2 p_2^6 - (24\mu_3\lambda_5 - 10\lambda_4^2)p_2^5 - (148\mu_3^2\lambda_4 + 2\lambda_5^2)p_2^4$$
$$+ (288\mu_3^4 - 12\lambda_4\lambda_5\mu_3 - \lambda_4^3)p_2^3 + (24\mu_3^3\lambda_5 - 7\mu_3^2\lambda_4^2)p_2^2 + 32\mu_3^4\lambda_4 p_2 - 24\mu_3^6 = 0.$$

in which μ_i denotes the i^{th} moment of the given distribution,

and $\lambda_4 = 9\mu_2^2 - 3\mu_4, \qquad \lambda_5 = 30\mu_2\mu_3 - 3\mu_5.$

A value of p_2 having been obtained from this equation, the parameters of equation (2) are determined by solving, successively, the equations

$$p_3 = \frac{2\mu_3^3 - 2\mu_3\lambda_4 p_2 - \lambda_5 p_2^2 - 8\mu_3 p_2^3}{4\mu_3^2 - \lambda_4 p_2 + 2p_2^3},$$

$$p_1 = p_3/p_2,$$

$b^2 - p_1 b + p_2 = 0$, (the two roots of this equation are denoted b_1, b_2)

$$c_1 = -b_2/(b_1 - b_2),$$

$$c_2 = b_1/(b_1 - b_2),$$

$$\sigma_1^2 = \mu_2 - \frac{1}{3}\frac{\mu_3}{b_2} - \frac{1}{3}p_1 b_1 + p_2,$$

$$\sigma_2^2 = \mu_2 - \frac{1}{3}\frac{\mu_3}{b_1} - \frac{1}{3}p_1 b_2 + p_2.$$

The calculation of the Sturm's functions of the fundamental nonic shows it to have three real roots, two between 0 and —100, and a third between 200 and 300. The values of these roots are found to be

$$p_2 = -5.5517, \quad -11.6140, \quad 210.$$

However, the use of the second and third of these roots leads to imaginary values of certain of the parameters of equation (2), and they are therefore rejected. Using the root, $p_2 = -5.5517$, the parameters of equation (2) are found to have the following values:

$$c_1 = 0.9637, \qquad b_1 = -0.46, \qquad \sigma_1 = 9.02$$

$$c_2 = 0.0363, \qquad b_2 = 12.20, \qquad \sigma_2 = 25.07$$

3. Dissection into a Symmetrical Distribution Composed of Three Normal Distributions.

Finally, let the given data be fitted by a frequency curve whose equation is

$$(3) \qquad y = \frac{324}{\sqrt{2\pi}}\left[\frac{c_1}{\sigma_1}e^{-\frac{x^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2}\left(e^{-\frac{(x-b)^2}{2\sigma_2^2}} + e^{-\frac{(x+b)^2}{2\sigma_2^2}}\right)\right],$$

where the origin is selected at the arithmetic mean of the original series. The dissection depends on the solution of equation (7), section III, which for the distribution under consideration has the form

$$u^{12} - 58.355\,u^{10} - 230.538\,u^8 - 243.922\,u^6 - 60.184\,u^4 - 3.164\,u^2 - 0.527 = 0.$$

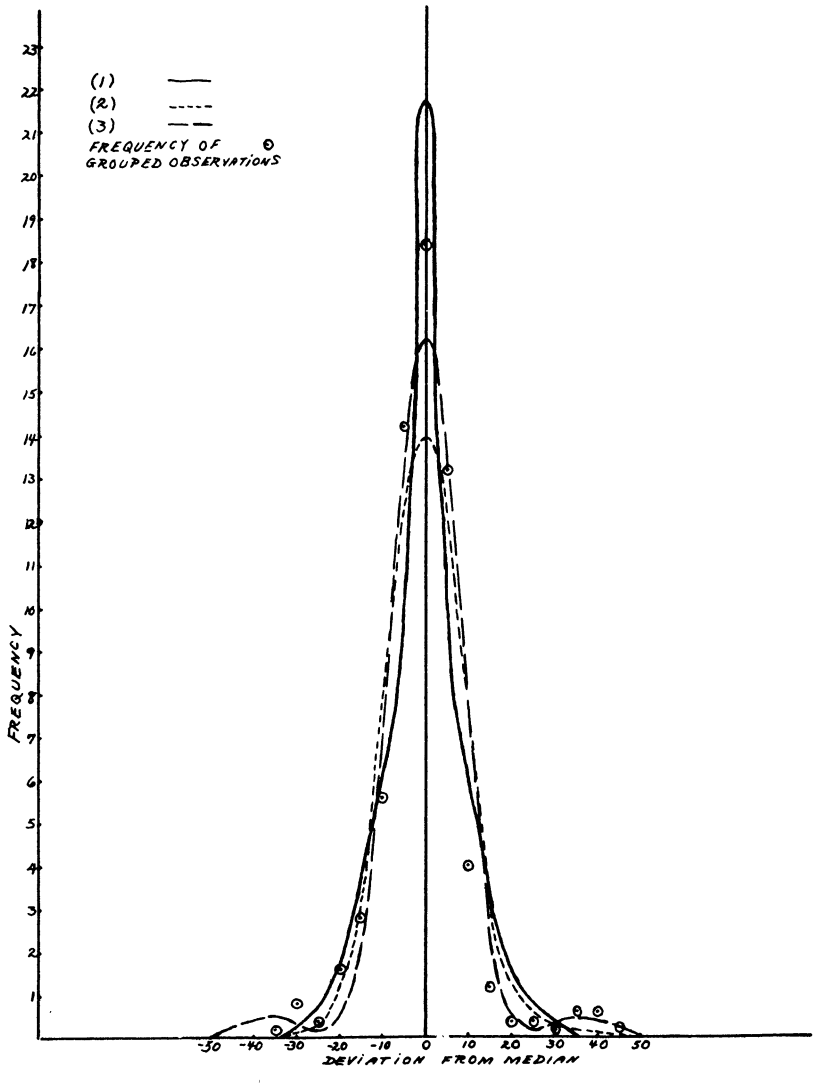
The only positive root of this equation is $u^2 = 62.13$.
Solving, successively, equations (6) and (5) of section III, the values of the parameters of equation (3) are found to be

$$c_1 = 0.96, \quad \sigma_1 = 7.64, \quad c_2 = 0.02, \quad \sigma_2 = 4.69, \quad b = 36.95.$$

The accompanying figure shows the original distribution (grouped into class intervals of five units) and the curves obtained by each of the three methods of dissection, plotted on the same set of axes. It appears from the figure that a distribution of type (3) fits the data more closely than either of the other curves. This fact may be checked by comparing the sums of the squares of the differences between the actual and theoretical frequency of each class. The values of these sums of squares of deviations from theoretical distributions (1), (2), (3) are found to be, respectively, 68.250, 51.604, 19.435.

4.—Relative Stability of Median and Mean for Each of the Methods of Dissection.

We turn now to the problem of comparing the stability of the median and mean of the three theoretical frequency functions


(1)
(2)
(3)
FREQUENCY OF GROUPED OBSERVATIONS
FREQUENCY
DEVIATION FROM MEDIAN
-50 -40 -30 -20 -10 0 10 20 30 40 50

obtained by dissection. Since $N = 324$ is fairly large, and since the median of each of the theoretical distributions is located at a point of relatively large frequency, we shall use the approximation to the standard deviation of the median,

$$\sigma_M = \frac{1}{2 \cdot y_M \cdot \sqrt{s}},$$

where y_M is the ordinate at the median.

For Crum's dissection into two normal distributions, the arithmetic mean and median are both at the origin. Hence

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{324}} \sqrt{c_1 \sigma_1^2 + c_2 \sigma_2^2} = 0.57,$$

$$\sigma_M = \frac{1}{\frac{2\sqrt{324}}{\sqrt{2\pi}} \left(\frac{c_1}{\sigma_1} + \frac{c_2}{\sigma_2} \right)} = 0.42,$$

which verifies his conclusion that the median is more stable than the arithmetic mean.

For the asymmetrical dissection into two normal curves,

$$\bar{x} = c_2 b = 0.46$$

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{324}} \sqrt{c_1 \sigma_1^2 + c_2 \sigma_2^2 + c_1 c_2 b^2} = 0.57.$$

M is a root of the equation

$$c_1 \int_0^{\frac{M}{\sigma_1}} e^{-\frac{t^2}{2}} dt = c_2 \int_0^{\frac{b-M}{\sigma_2}} e^{-\frac{t^2}{2}} dt,$$

and its value, obtained by interpolation in a table of values of the integral $\int_0^t e^{-\frac{t^2}{2}} dt$, is $M = 0.16$. Hence

$$\sigma_M = \frac{1}{\frac{2\sqrt{324}}{\sqrt{2\pi}} \left(\frac{c_1}{\sigma_1} e^{-\frac{M^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2} e^{-\frac{(b-M)^2}{2\sigma_2^2}} \right)} = 0.64.$$

For this method of dissection the arithmetic mean is more stable than the median. This was to be expected, since the second component contains so small a fraction of the total area that the compound curve differs little from a single normal curve. The

figure shows that this curve would not naturally be chosen to represent the given data.

For the symmetrical three normal curve dissection,

$$M = \bar{x} = 0,$$

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{324}} \sqrt{c_1\sigma_1^2 + 2c_2\sigma_2^2 + 2c_2\beta^2} = 0.59,$$

$$\sigma_M = \frac{1}{\sqrt{324}} \frac{\sqrt{2\pi}\cdot\sigma_1\sigma_2}{2(c_1\sigma_2 + 2c_2\sigma_1 e^{-\beta^2/2\sigma_2^2})} = 0.55.$$

This method of dissection bears out Crum's conclusion that the median of the original series is a more stable average than the arithmetic mean, although the difference between the standard deviations of the two averages obtained by this method is considerably smaller than that obtained by the first method of dissection.

5. *The Probable Errors of the Median and Mean, Determined from the Frequency Distributions of these Averages.*

The above discussion has been concerned with certain theoretical frequency curves, rather than with the actual data which these curves are intended to fit. We shall now compare the relative stability of the mean and median by the method of section V, which does not involve a fitting to the data of a theoretical frequency curve.

The method of determining the quartiles of the frequency distribution of the median, developed in section V, assumed the number of items in the sample to be odd. We therefore solve the equations

$$K = \frac{\frac{1}{2}\sqrt{\pi n}}{2n+1}, \qquad \lambda = \frac{\frac{n}{3}K^3}{1-nK^2},$$

using the values 323 and 325 for $(2n+1)$ and obtain roots

$$K = 0.0348, \qquad \lambda = 0.0027$$

in each case, whence $\alpha = \kappa + \lambda = 0.0375$,

and $$\int_0^x f(x)\,dx = 0.0188.$$

For the given distribution, the median falls at zero deviation. The fourteen items in the upper half of the zero class comprise 4.32% of the entire frequency distribution. Hence the third quartile of the distribution of the medians has a value

$$\ell = \frac{1}{2} \cdot \frac{0.0188}{0.0432} = 0.2176.$$

Similar reasoning shows the value of the first quartile of the medians to be—0.2176, whence the semi-interquartile range is 0.2176.

Since $\sigma^2 = 106.85$, the probable error of the arithmetic mean has a value

$$.6745 \sqrt{\frac{106.85}{324}} = 0.3845.$$

Thus the median is again shown to be more stable than the arithmetic mean.

HARRY S. POLLARD,
Miami University,
Oxford, Ohio.

Harry S. Pollard