ESTIMATION OF THE MEAN AND STANDARD DEVIATION BY ORDER STATISTICS[1]

By* A. E. Sarhan

*Egyptian Medical Research Laboratories*

**1. Summary.** The aim of this paper is: (i) to find the best linear estimates of the means and standard deviations of the rectangular, triangular, exponential and double exponential populations; (ii) to compare the efficiencies of these estimates with some other estimates for small samples; (iii) to discuss the variation of coefficients in the best linear estimates as the population varies.

**2. Introduction.** In recent literature, linear combinations of the sample ordered values are used to provide estimates from random samples drawn from populations with specified forms. Such statistics are termed systematic by Mosteller [8]. They are now in common use, because they provide very simple solutions of many important parametric problems of statistical estimation. Sometimes they are inefficient in the sense that they do not use all the information contained in the sample as it would be used by the best possible methods, which are however computationally more complicated. In this work that estimate is obtained (called for short "best linear") which is the best linear combination of the ranked observations.

**3. Rectangular population.** The frequency distribution of a rectangular population is

$$f(y) = 1/\theta_2, \qquad \theta_1 - \tfrac{1}{2}\theta_2 \leqq y \leqq \theta_1 + \tfrac{1}{2}\theta_2$$

where $\theta_1$ is the mean and $\theta_2$ is the range. Let $y_1, y_2, \cdots, y_n$ be a sample of size $n$ and let the observations be ordered to give $y_{(1)}, y_{(2)}, \cdots, y_{(n)}$, for $n \geqq 2$. Now consider the linear estimates

$$\theta_1^* = \sum_{i=1}^{n} \alpha_{1i} y_{(i)} \qquad \theta_2^* = \sum_{i=1}^{n} \alpha_{2i} y_{(i)}$$

The method of least squares will provide the best linear estimates of $\theta_1$ and $\theta_2$. The estimates are (Lloyd [7])

(3.1) $$\theta_1^* = \tfrac{1}{2}(y_{(1)} + y_{(n)})$$

(3.2) $$\theta_2^* = (y_{(n)} - y_{(1)})\,(n + 1)/(n - 1)$$

and the variances

Received 5/25/53, revised 9/30/53.

[1] This is a summary of thesis submitted to Liverpool University on July 1, 1952. The estimation of the parameters of the rectangular population was included which was obtained independently of Lloyd's work [7]. The details of this case are not given here.

$$V(\theta_1^*) = \theta_2^2/2(n + 1)(n + 2), \tag{3.3}$$

$$V(\theta_2^*) = 2\theta_2^2/(n + 2)(n - 1). \tag{3.4}$$

The standard deviations ($\sigma$) can be estimated by $\theta_2^*/2\sqrt{3}$. For the special case $f(y) = 1/\theta_2$, $0 \leqq y \leqq \theta_2$, the estimates are (Craig [2]):

$$\theta_1^* = (n + 1)y_{(n)}/2n, \tag{3.5}$$

$$\theta_2^* = (n + 1)y_{(n)}/n \tag{3.6}$$

$$V(\theta_1^*) = \theta_2^2/4n(n + 2), \tag{3.7}$$

$$V(\theta_2^*) = \theta_2^2/n(n + 2). \tag{3.8}$$

The maximum likelihood estimates are in agreement with the best linear estimate in both the general (Fisher [5]) and the special cases.

**4. Triangular population.** The frequency distribution of the triangular population is

$$f(y) = (4/\theta_2^2)(\tfrac{1}{2}\theta_2 - | y - \theta_1 |), \qquad | y - \theta_1 | \leqq \tfrac{1}{2}\theta_2$$

where $\theta_1$ is the mean and $\theta_2$ is the range. Standardizing the variable ($\theta_1 = \frac{1}{2}\,\theta_2 = 1$) to get

$$f(x) = \begin{cases} 4x, & 0 \leqq x \leqq \tfrac{1}{2}, \\ 4(1 - x). & \tfrac{1}{2} \leqq x \leqq 1, \end{cases}$$

we have (Wilks [11])

$$E(x_r) = \int_0^{\frac{1}{2}} x_r f_1(x_r)\, dx_r + \int_{\frac{1}{2}}^1 x_r f_2(x_r)\, dx_r \tag{4.1}$$

where

$$f_1(x_r) = K(2\,x_r^2)^{r-1}(1 - 2x_r^2)^{n-r}4x_r$$

$$f_2(x_r) = K[2(1 - x_r)^2]^{n-r}[1 - 2(1 - x_r)^2]^{r-1}4(1 - x_r)$$

with $K = n!/(r - 1)!(n - r)!$ and

$$E(x_r^2) = \int_0^{\frac{1}{2}} x_r^2 f_1(x_r)\, dx + \int_{\frac{1}{2}}^1 x_r^2 f_2(x_r)\, dx \tag{4.2}$$

with the same notation for $f_1(x_r)$ and $f_2(x_r)$. Also, when $x_n < x_s$,

$$E(x_r x_s) = \int_0^1 \int_0^{x_s} x_r x_s f(x_r, x_s)\, dx_r\, dx_s$$

$$= \int_0^{\frac{1}{2}} \int_0^{x_s} x_r x_s f_1(x_r, x_s)\, dx_r\, dx_s + \int_{\frac{1}{2}}^1 \int_0^{\frac{1}{2}} x_r x_s f_2(x_r, x_s)\, dx_r\, dx_s \tag{4.3}$$

$$+ \int_{\frac{1}{2}}^1 \int_{x_s}^1 x_r x_s f_3(x_r, x_s)\, dx_r\, dx_s$$

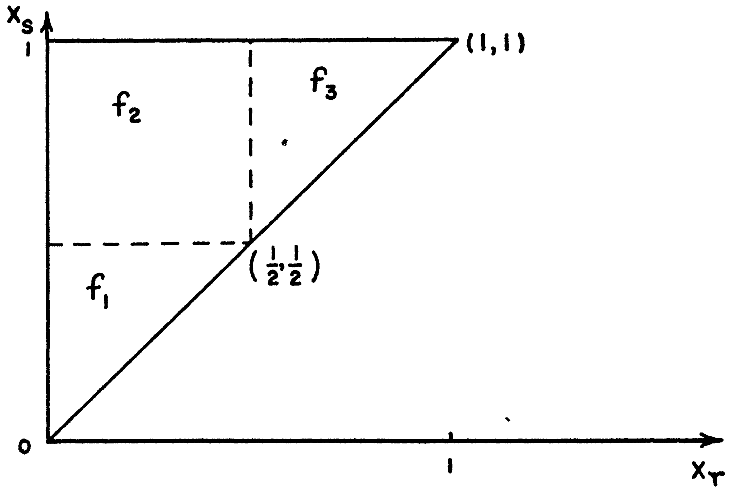



FIG. 1

where

$$f_1(x_r, x_s) = C(2x_r^2)^{r-1}(2x_s^2 - 2x_r^2)^{s-r-1}(1 - 2x_s^2)^{n-s}(4)^2 x_r x_s,$$

$$f_2(x_r, x_s) = C(2x_r^2)^{r-1}[1 - 2x_r^2 - 2(1 - x_s)^2]^{s-r-1}[2(1 - x_s)^2]^{n-s}(4)^2 x_r(1 - x_s)$$

$$f_3(x_r, x_s) = C[1 - 2(1 - x_r)^2]^{r-1}[2(1 - x_r)^2 - 2(1 - x_s)^2]^{s-r-1}[2(1 - x_s)^2]^{n-s} (4)^2 (1 - x_r)(1 - x_s)$$

$$C = n!/(r - 1)!(s - r - 1)!(n - s)!.$$

The expected values, variances and covariances of the order statistic $(x_r)$ are shown in Table I up to sample size 5. For the best linear estimate of the mean,

$$\theta_1^* = \sum_{i=1}^{n} \alpha_{1i} y_{(i)}, \tag{4.4}$$

the coefficients $\alpha_{1i}$ are given in Table III. Since

$$V(y) = \theta_2^2/24 \tag{4.5}$$

we can estimate the standard deviation $(\sigma)$ by $(\theta_2/\sqrt{24})$, and the coefficients may be adjusted to give the best linear estimate of the standard deviation $(\sigma^*)$. These adjusted coefficients for which

$$\sigma^* = \sum_{i=1}^{n} \alpha_{2i} y_{(i)} \tag{4.6}$$

are also shown in Table III.

Cramér [4] found for large samples that the ordinary sample mean is a better estimate of the mean of the distribution that is the midrange; Table IV shows this to be true also for small samples.

TABLE I

*Exact expected values, variances, and covariances of the order statistic ($x_r$) in samples of size n from standardized triangular populations. Each variance and covariance value must be divided by the appropriate divisor given in the last column.*

| $n$ | $r$ | $E(X_r)$ | cov $(X_r, X_1)$ | cov $(X_r, X_2)$ | cov $(X_r, X_3)$ | cov $(X_r, X_4)$ | cov $(X_r, X_5)$ | Divisor |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 23/60 | 101 | 491 | | | | 3,600 |
| 3 | 1 | 13/40 | 726 | 400 | 229 | | | 33,600 |
| | 2 | 20/40 | 400 | 690 | 400 | | | |
| 4 | 1 | 1451/5040 | 452,399 | 261,651 | 170,781 | 105,529 | | 25,401,600 |
| | 2 | 2199/5040 | 261,651 | 418,599 | 275,409 | 170,781 | | |
| 5 | 1 | 527/2016 | 3,405,665 | 2,034,710 | 1,379,952 | 983,242 | 630,871 | 223,534,080 |
| | 2 | 794/2016 | 2,034,710 | 3,146,180 | 2,143,008 | 1,530,940 | 983,242 | |
| | 3 | 1008/2016 | 1,379,952 | 2,143,008 | 2,978,640 | 2,143,008 | 1,379,952 | |

TABLE II

*Exact expected values, variances, and covariances of the order statistic ($x_r$) in samples of size n from standardized double exponential populations. Each variance and covariance value must be divided by the appropriate divisor given in the last column.*

| $n$ | $r$ | $E(X_r)$ | cov$(X_r, X_1)$ | cov $(X_r, X_2)$ | cov $(X_r, X_3)$ | cov $(X_r, X_4)$ | cov $(X_r, X_5)$ | Divisor |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | −3/4 | 23 | 9 | | | | 16 |
| 3 | 1 | −9/8 | 815 | 272 | 185 | | | 576 |
| | 2 | 0 | 272 | 368 | 272 | | | |
| 4 | 1 | −133/96 | 13,287 | 4,315 | 2,533 | 2,137 | | 9,216 |
| | 2 | −11/32 | 4,315 | 4,799 | 2,945 | 2,533 | | |
| 5 | 1 | −305/192 | 1,354,983 | 444,738 | 243,328 | 183,838 | 170,233 | 921,600 |
| | 2 | −55/96 | 444,738 | 463,068 | 258,208 | 197,668 | 183,838 | |
| | 3 | 0 | 243,328 | 258,208 | 323,648 | 258,208 | 243,328 | |

**5. Exponential population.** The frequency distribution of an exponential population is

$$f(y) = e^{-(y-\mu)/\sigma}\sigma/\sigma, \qquad \mu \leqq y \leqq \infty.$$

Let $x = (y - \mu)/\sigma$ to get $f(x) = e^{-x}$ for $0 \leqq x \leqq \infty$. Then

$$(5.1) \quad \begin{cases} E(y_r) = \mu + \sigma E(x_r), \\ E(x_{(r)}) = K \int_0^\infty x_r(1 - e^{-x_r})^{r-1} \cdot e^{-x_r(n-r+1)}\, dx_r = \sum_{i=1}^{r} \frac{1}{(n - i + 1)} \end{cases}$$

## TABLE III

*Coefficients in the best linear estimates, based on the order statistic $y_{(i)}$ in different populations of size n, for the mean, $\sum_{i=1}^{n} \alpha_{1i}\, y_{(i)}$, where $\alpha_{1i} = \alpha_{1(n-i+1)}$ and the standard deviation $\sum_{i=1}^{n} \alpha_{2i}\, y_{(i)}$, where $\alpha_{2i} = -\alpha_{2(n-i+1)}$*

| Sample Size, n, and Population | | $\alpha_{11}$ | $\alpha_{21}$ | $\alpha_{12}$ | $\alpha_{22}$ | $\alpha_{13}$ | $\alpha_{23}$ |
|---|---|---|---|---|---|---|---|
| 2 | Rectangular | .5000000 | −.8660253 | | | | |
| | Triangular | .5000000 | −.8748178 | | | | |
| | Normal | .5000000 | −.8862269 | | | | |
| | Dbl. expon. | .5000000 | −.9428070 | | | | |
| 3 | Rectangular | .5000000 | −.5773503 | 0. | 0. | | |
| | Triangular | .3945578 | −.5832118 | .2108844 | 0. | | |
| | Normal | .3333333 | −.5908179 | .3333333 | 0. | | |
| | Dbl. expon. | .1481481 | −.6222161 | .7037038 | 0. | | |
| 4 | Rectangular | .5000000 | −.4811250 | 0. | 0. | | |
| | Triangular | .3378906 | −.4722486 | .1621094 | −.0541433 | | |
| | Normal | .2500000 | −.4539404 | .2500000 | −.1101807 | | |
| | Dbl. expon. | .0472971 | −.4307352 | .4527030 | −.3003697 | | |
| 5 | Rectangular | .5000000 | −.4330128 | 0. | 0. | 0. | 0. |
| | Triangular | .3060758 | −.3994195 | .1188518 | −.0637213 | .1501447 | 0. |
| | Normal | .2000000 | −.3723816 | .2000000 | −.1352139 | .2000000 | 0. |
| | Dbl. expon. | .0166355 | −.3263380 | .2213003 | −.3169696 | .5241287 | 0. |

## TABLE IV

*Percentage efficiencies of various estimators of the mean in different populations. Sample mean = $\bar{y}$, midrange = $w$, median = $\tilde{y}$*

| Sample size .... | 2 | | | 3 | | | 4 | | | 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimator...... | $\bar{y}$ | $w$ | $\tilde{y}$ | $\bar{y}$ | $w$ | $\tilde{y}$ | $\bar{y}$ | $w$ | $\tilde{y}$ | $\bar{y}$ | $w$ | $\tilde{y}$ |
| Rectangular | 100.00 | 100.00 | 100.00 | 90.00 | 100.00 | 50.00 | 80.00 | 100.00 | 62.50 | 71.45 | 100.00 | 33.33 |
| Triangular.. | 100.00 | 100.00 | 100.00 | 98.82 | 96.58 | 66.83 | 97.72 | 92.71 | 74.53 | 96.70 | 89.25 | 60.47 |
| Normal..... | 100.00 | 100.00 | 100.00 | 100.00 | 91.99 | 73.55 | 100.00 | 83.99 | 83.89 | 100.00 | 76.70 | 69.69 |
| Dbl. expon. | 100.00 | 100.00 | 100.00 | 88.43 | 67.90 | 92.27 | 82.80 | 49.65 | 98.90 | 79.21 | 38.29 | 90.23 |

where $K = n!/(r-1)!(n-r)!$. Also

$$(5.2)\quad \begin{cases} E(x_r^2) = K \int_0^\infty x_r^2 (1 - e^{-x_r})^{r-1} e^{-x_r(n-r+1)}\, dx_r, \\ V(x_r) = E(x_r^2) - [E(x_r)]^2 = \sum_{i=1}^{r} \frac{1}{(n-i+1)^2}, \\ E(x_r x_s) = C \int_0^\infty \int_0^{x_s} x_r x_s (1 - e^{-x_r})^{r-1} (e^{-x_r} - e^{-x_s})^{s-r-1} \\ \qquad\qquad \cdot (e^{-x_s})^{n-s+1} e^{-x_r}\, dx_r\, dx_s \end{cases}$$

where $C = n!/(r-1)!(s-r-1)!(n-s)!$. Finally

(5.3) $$\text{cov}(x_r x_s) = \sum_{i=1}^{r} \frac{1}{(n-i+1)^2}.$$

Therefore,

$$V^{-1} = \begin{bmatrix} n^2+(n-1)^2 & -(n-1)^2 & 0 & 0 & \cdots & 0 \\ & (n-1)^2+(n-2)^2 & -(n-2)^2 & 0 & \cdots & 0 \\ & & (n-2)^2+(n-3)^2 & -(n-3)^2 & \cdots & 0 \\ & & & & \cdots & \cdots \\ & & & & & 1 \end{bmatrix}.$$

Since

$$A' = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \frac{1}{n} & \sum_{i=1}^{2} \frac{1}{(n-i+1)} & \sum_{i=1}^{3} \frac{1}{(n-i+1)} & \cdots & \sum_{i=1}^{n} \frac{1}{(n-i+1)} \end{bmatrix},$$

it follows that

(5.4) $$\begin{cases} (A'V^{-1}A)^{-1} = \dfrac{1}{n(n-1)} \begin{bmatrix} 1 & -1 \\ -1 & n \end{bmatrix} \\ (A'V^{-1}A)^{-1}A'V^{-1} = \dfrac{1}{n(n-1)} \begin{bmatrix} (n^2-1) & -1 & \cdots & -1 \\ n(n-1) & n & \cdots & n \end{bmatrix}. \end{cases}$$

Therefore,

(5.5) $$\mu^* = [ny_{(1)} - \bar{y}]/(n-1),$$

(5.6) $$\sigma^* = [\bar{y} - y_{(1)}]n/(n-1).$$

These estimates are in agreement with the maximum likelihood estimates. From (5.4) we have,

(5.7) $$V(\mu^*) = 1/n(n-1)$$

(5.8) $$V(\sigma^*) = 1/(n-1).$$

Since the mean is equal to $\mu + \sigma$, the estimate of the mean is

(5.9) $$[1 \quad 1] \frac{1}{(n-1)} \begin{bmatrix} ny_{(1)} - \bar{y} \\ n\bar{y} - ny_{(1)} \end{bmatrix} = \bar{y}.$$

Therefore, the best linear estimate of the mean of the exponential population is the sample mean.

**6. Double exponential population.** The frequency distribution of the double exponential population with mean $\mu$ and variance $2\sigma^2$ is

$$f(y) = e^{-|y-\mu|/\sigma}/2\sigma.$$

Let $x = (y - \mu)/\sigma$ giving $f(x) = \frac{1}{2}e^{-|x|}$ for $-\infty \leqq x \leqq \infty$, and $E(y_{(r)}) = \mu + \sigma E(x_{(r)})$. Then,

$$E(x_{(r)}) = \int_{-\infty}^{0} x_r f_1(x_r)\, dx + \int_{0}^{\infty} x_r f_2(x_r)\, dx \tag{6.1}$$

where

$$f_1(x_r) = K(\tfrac{1}{2}e^{x_r})^r[1 - \tfrac{1}{2}e^{x_r}]^{n-r}$$
$$f_2(x_r) = K(1 - \tfrac{1}{2}e^{-x_r})^{r-1}(\tfrac{1}{2}e^{-x_r})^{n-r+1}$$

with $K = n!/(r - 1)!(n - r)!$. Also

$$E(x_{(r)}^2) = \int_{-\infty}^{0} x_r^2 f_1(x_r)\, dx + \int_{0}^{\infty} x_r^2 f_2(x_r)\, dx \tag{6.2}$$

with the same notations for $f_1(x_r)$ and $f_2(x_r)$. Finally,

$$\begin{aligned} E(x_{(r)}\, x_{(s)}) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{x_s} (x_r x_s) f(x_r, x_s)\, dx_r^{\alpha}\, x_s \\ &= \int_{-\infty}^{0} \int_{-\infty}^{x_s} f_1(x_r, x_s) x_r x_s\, dx_r\, dx_s \\ &\quad + \int_{0}^{\infty} \int_{-\infty}^{0} x_r x_s f_2(x_r, x_s)\, dx_r\, dx_s \\ &\quad + \int_{0}^{\infty} \int_{0}^{x_s} x_r x_s f_3(x_r, x_s)\, dx_r\, dx_s \end{aligned} \tag{6.3}$$

where

$$f_1(x_r, x_s) = C(\tfrac{1}{2})^s e^{r x_r}(e^{x_s} - e^{x_r})^{s-r-1}(1 - \tfrac{1}{2}e^{x_s})^{n-s} \cdot e^{x_s}$$
$$f_2(x_r, x_s) = C(\tfrac{1}{2}e^{x_r})^{r-1}(1 - \tfrac{1}{2}e^{x_r} - \tfrac{1}{2}e^{-x_s})^{s-r-1}(\tfrac{1}{2}e^{-x_s})^{n-s} \cdot (\tfrac{1}{2})^2 \cdot e^{-x_r} \cdot e^{-x_s}$$
$$f_3(x_r, x_s) = C(\tfrac{1}{2})^{n-r+1}(1 - \tfrac{1}{2}e^{-x_r})^{r-1}(e^{-x_r} - e^{-x_s})^{s-r-1} \cdot e^{-x_s(n-s+1)} \cdot e^{-x_r}$$

with $C = n!/(r - 1)!(s - r - 1)!(n - s)!$.

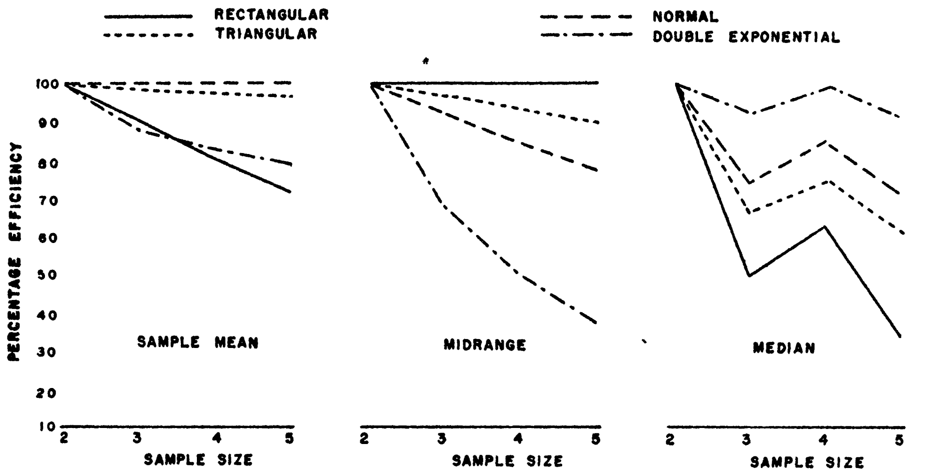
The exact expected values, variances, and coveriences of the order statistic $(x_{(r)})$ are given in Table II for $n = 2, 3, 4, 5$. The missing entries in the table may be obtained by $E(x_{(r)}) = -E(x_{(n-r+1)})$ for expected values and cov $[x_{(r)}x_{(s)}]$ $=$ cov $[x_{(n-r+1)}x_{(n-s+1)}]$, for covariances.

The coefficients $(\alpha_{1i})$ in the best linear estimate $\mu^* = \sum_{i=1}^{n} \alpha_{1i}\, y_{(i)}$ of the mean are given in Table III.

Since $\sigma_D^2 = 2\sigma^2$, then the coefficients in the best linear estimate of $\sigma$ can be adjusted to give the coefficients in the best linear estimate of the standard devia-

GRAPH (1)

PERCENTAGE EFFICIENCES OF THE SAMPLE MEAN, MIDRANGE, AND MEDIAN IN DIFFERENT POPULATIONS




tion ($\sigma_D$). The adjusted coefficients ($\alpha_{2i}$) for which $\sigma_D^* = \sum_{i=1}^{n} \alpha_{2i} y_{(i)}$ are given in Table III.

Comparison of efficiencies of different estimates of the mean shows that the median is more efficient than either the sample mean or the midrange and less efficient than the best linear estimate. The maximum likelihood estimates of the mean and the standard deviation of this population are the median and the mean deviation about the median (Fisher [6]), respectively. Neither is efficient for small samples.

**7. Comparison of different estimates.** Table IV is constructed to give the percentage efficiencies of midrange, median and sample mean as estimates of the population mean (relative to the best linear estimate of the mean). The comparison of efficiencies of the estimates in the different populations may easily be seen in Graph 1.

The sample mean is the best linear estimate of the mean of a normal population, so in general we expect that its efficiency decreases in both platykurtic and leptokurtic populations. The efficiency of the midrange decreases in normal populations and again in leptokurtic populations. The median behaves in a reverse way: its efficiency is high in leptokurtic, decreases in normal and again in platykurtic populations.

Table V gives the expected values of different estimates of the standard deviation in different populations. By the normal estimate in this table is meant that linear estimate of the standard deviation of the given population which is best for a normal population (Godwin [4]). Further, Gini's estimate is that obtained by using Gini's mean difference (Nair [9]) which may be expressed as

$$\Delta_1 = 2[2U - (n + 1)V]/n(n - 1)$$

TABLE V

*Expected values of the normal and Gini's estimates of the standard deviation in different populations*

| Population | n = 2 | | n = 3 | | n = 4 | | n = 5 | |
|---|---|---|---|---|---|---|---|---|
| | Normal | Gini's | Normal | Gini's | Normal | Gini's | Normal | Gini's |
| Rectangular | 1.02333 | 1. | 1.02333 | 1. | 1.01983 | 1. | 1.01611 | 1. |
| Triangular.. | 1.01304 | 1.13221 | 1.01304 | 1.13221 | 1.01213 | 1.13221 | 1.01115 | 1.13221 |
| Normal..... | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| Dbl. expon. | .93999 | 1.06066 | .93999 | 1.06066 | .94296 | 1.06066 | .94612 | 1.06066 |

TABLE VI

*Percentage efficiencies of the range, normal, and Gini's estimates of the standard deviation in different populations, from ordered samples of size n*

| Population | n = 2 | | | n = 3 | | | n = 4 | | | n = 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Range | Normal | Gini's | Range | Normal | Gini's | Range | Normal | Gini's | Range | Normal | Gini's |
| Rectangular... | | 100 | 100 | | 100 | 100 | | 97.38 | 95.23 | | 93.45 | 88.04 |
| Triangular.... | 100 | 100 | 100 | 100 | 100 | 100 | 99.77 | 99.77 | 97.10 | 99.02 | 98.66 | 95.29 |
| Normal....... | 100 | | 100 | 100 | | 100 | 98.78 | | 99.72 | 96.95 | | 99.48 |
| Dbl. expon..... | 100 | 100 | 100 | 100 | 100 | 100 | 94.75 | 97.99 | 99.99 | 89.83 | 96.62 | 99.75 |

where $U = \sum_{j=1}^{n} jy_{(j)}$ and $V = \sum_{j=1}^{n} y_{(j)}$. Table V shows that the normal estimates are biased. Since the normal estimate is the best linear estimate of the standard deviation of a normal population, we may expect in general that with a platykurtic population the estimate is too high and with a leptokurtic population it is too small.

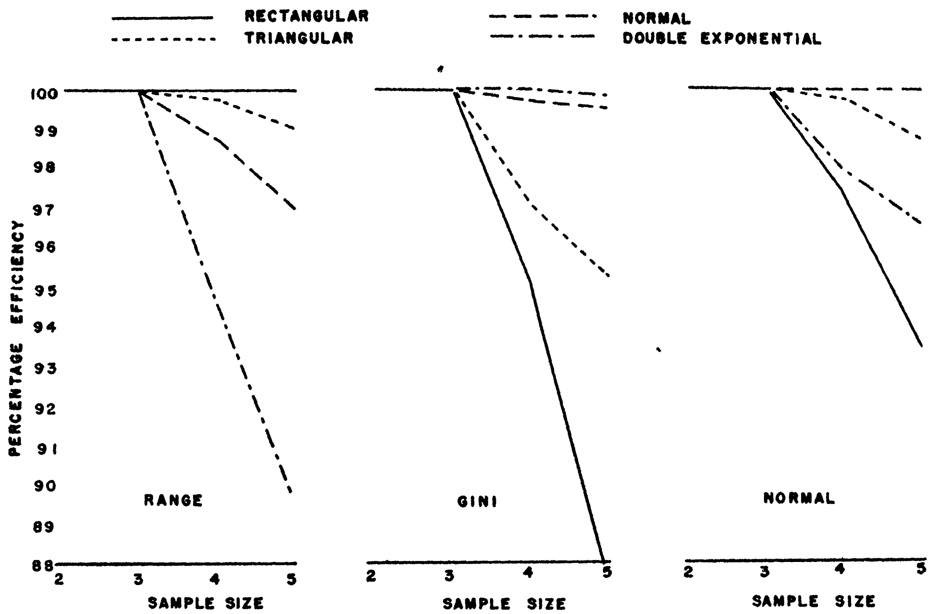
The efficiencies of different estimates of the standard deviation relative to the best linear estimate are given in Table VI, and a graphical representation in Graph 2. The efficiency of the range, as an estimate of the standard deviation, is greater in the rectangular population, decreases in the normal and again in the double exponential. The efficiency should generally be higher in platykurtic populations, decrease in the normal and again in leptokurtic populations. Furthermore, the efficiency of the normal estimate should generally decrease in both platykurtic and leptokurtic populations.

The efficiency of Gini's estimate is higher in the double exponential than in the normal, and decreases in the triangle and again in the rectangular population. In the case of the double exponential, Gini's estimate is more efficient than either the range or the normal estimate, and it is nearly as efficient as the best linear estimate, so we can use it as an estimate of standard deviation in samples from this population. Its coefficients are very simple and so it can be calculated quickly and easily.

In the normal population, Gini's estimate is shown to be nearly as efficient as

GRAPH (2)

PERCENTAGE EFFICIENCES OF THE RANGE, GINI'S ESTIMATE, AND NORMAL ESTIMATE IN DIFFERENT POPULATIONS




the best linear estimate and more efficient than the range. So the estimate $G = \frac{1}{2}\sqrt{\pi}\,\Delta_1$ may be used as an estimate of standard deviation for a normal population; it will be considerably more reliable as $n$ increases than that based on the range. If the functional form of the population distribution is unknown, we may use the best normal estimate or better, Gini's estimate, as an estimate of the standard deviation, because of their high efficiencies. A further advantage of the latter estimate is that its expectation is independent of $n$, so that it can be used as an unbiased estimate.

**8. Variation of coefficients in the best linear estimates.** Table III gives the coefficients in the best linear estimates of different populations. Comparing the coefficients in the best linear estimate of the mean, we can see that equal weights are given to the sample elements in the case of the normal population, while smaller weights are given to the middle elements in samples from a triangular population than those given to the extreme elements, and zero coefficients are attached to all elements other than the two extremes in samples from a rectangular population.

Again in the case of the double exponential population, the weight is largest for the middle sample element, decreases gradually, and becomes least for the extreme elements. In general, we expect that the more platykurtic a population is, the greater should be the weights given to the extreme elements compared with those given to the middle elements.

TABLE VII

*Variances of the different estimates of standard deviation in different population* ($\sigma = 1$)

| Population | $n = 2$ | | | | $n = 3$ | | | | $n = 4$ | | | | $n = 5$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Range | Normal | Gini's | Best | Range | Normal | Gini's | Best | Range | Normal | Gini's | Best | Range | Normal | Gini's |
| Rectangular.. | .5000 | | .5236 | .5000 | .2000 | | .2094 | .2000 | .1111 | | .1188 | .1167 | .0714 | | .0790 | .0800 |
| Triangular..... | .5306 | .5306 | .5445 | .6933 | .2415 | .2415 | .2478 | .3155 | .1514 | .1518 | .1555 | .2000 | .1079 | .1099 | .1117 | .1453 |
| Normal. | .5708 | .5708 | | .5708 | .2755 | .2755 | | .2755 | .1801 | .1823 | | .1806 | .1333 | .1375 | | .1340 |
| Dbl. expon.... | .7778 | .7778 | .6872 | .8750 | .4321 | .4321 | .3818 | .4863 | .2986 | .3152 | .2711 | .3314 | .2288 | .2547 | .2120 | .2599 |
| Expon... | 1.0000 | 1.0000 | .7854 | 1.0000 | .5000 | .5555 | .4363 | .5555 | .3333 | .4049 | .3085 | .3889 | .2500 | .3280 | .2404 | .3000 |

(Best denotes the best linear estimate; Normal, the normal estimate; Gini's, the Gini's mean difference.)

Comparing the coefficients in the best linear estimate of the standard deviation of different populations shows that no weight is given to the middle element, as is to be expected because the populations are symmetric. Similarly, we see that the more platykurtic a population is, the smaller the weights given to the middle elements and the larger the weights given to the extreme elements.

**Conclusions.** I would like to point out a few problems raised but not solved in this paper.

(i) The reverse problem, that is to find the population or the set of populations, if any, for which a given estimate is best linear estimate, is not yet attacked. It would be of interest, for example, to know the population whose best linear estimate for standard deviation is Gini's mean difference.

(ii) Table VII gives the variances of the best linear estimates of the standard deviations for the given populations. It shows that the variances of the estimates for the rectangular population are the least among the given populations. This raises the problem of finding the population whose standard deviation can be estimated with the least variance.

(iii) When general expressions of the expected values, variances and covariances are known, the best linear estimates from samples of size $n$ can generally be obtained. However, in many cases such expressions are not possible, or are very difficult to obtain. In these cases we must find them separately for each value of $n$, which will be tedious for large values. It may be possible to find a new method or approximation, by means of which we can find the linear estimates without these tedious calculations.

(iv) It has been shown that the coefficients in the best linear estimates vary with varying shapes of the distributions, but it seems that this is not the only relationship. One could relate the coefficients directly to known properties of the distribution functions, such as moments or semi-invariants. However, this problem seems to be difficult.

Finally, I would like to express my acknowledgement to Mr. R. L. Plackett for suggesting the problem and for his help during his supervision. I am also greatly indebted to the referee for his comments and to Dr. W. Hoeffding, Dr.

E. L. Lehmann and Dr. B. G. Greenberg for their kind help in revising the manuscript for publication.

## REFERENCES

[1] A. C. AITKEN, "On least squares and linear combination of observations," *Proc. Roy. Soc. Edinburgh,* Vol. 55 (1935), pp. 42–48.
[2] A. T. CRAIG, "A note on the best linear estimate," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 88–90.
[3] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946, pp.178–179, 372–373.
[4] H. J. GODWIN, "On the estimation of the dispersion by linear systematic statistics," *Biometrika*, Vol. 36 (1949), pp. 92–100.
[5] R. A. FISHER, "On the mathematical foundation of theoretical statistics," *Trans. Roy. Soc. London Series A*, Vol. 222 (1921), pp. 309–368.
[6] R. A. FISHER, "A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error," *Contributions to Mathematical Statistics*, 1922, p. 2.769.
[7] E. H. LLOYD, "Least squares estimation of location and scale parameters using order statistics," *Biometrika*, Vol. 39 (1952), pp. 88–95.
[8] F. MOSTELLER, "On some useful inefficient statistics," *Ann. Math. Stat.* Vol. 17 (1946), pp. 377–407.
[9] U. S. NAIR, "The standard error of Gini's mean difference," *Biometrika*, Vol. 23 (1936), pp. 428–434.
[10] R. L. PLACKETT, Lecture notes on the method of least squares.
[11] S. S. WILKS, *Mathematical Statistics*, Princeton University Press, 1943, pp. 89–92.