ESTIMATION OF THE MEAN AND STANDARD DEVIATION BY ORDER STATISTICS. PART II

By A. E. Sarhan

University of North Carolina

1. Introduction. In a previous paper [3], the best linear estimates of the mean and standard deviation for the rectangular, triangular, double exponential, and the exponential distributions were worked out. The best linear estimates were obtained by ranking the observations in ascending order and finding the best linear combination of them [2]. The variation of the coefficients in the estimates and the efficiencies of some other linear estimates were discussed.

This paper—which is a continuation of the previous one [3]—deals with three distributions: a U-shaped, a parabolic, and a skewed one. The same items were worked out for these distributions as for those in the previous paper. Also, a general idea of the natural sequence of the coefficients in the best linear estimate of the mean as the shape of the distribution undergoes change will be considered.

The mathematical formulae for this work will not be given as they are similar to those given in [3].

2. U-shaped population. The frequency distribution of a U-shaped population is

$$f(y) = \frac{3(y-\theta_1)^2}{2\theta_2^3}, \qquad \theta_1 - \theta_2 \leqq y \leqq \theta_1 + \theta_2 \tag{2.1}$$

where θ_1 is the mean and θ_2 is half the range. Standardizing the variable we get

$$f(x) = \tfrac{3}{2}x^2, \qquad -1 \leqq x \leqq +1. \tag{2.2}$$

The coefficients α_{1i} in the best linear estimates of the mean are given in Table I such that

$$\theta_1^* = \sum_{i=1}^{n} \alpha_{1i} y_{(i)}, \tag{2.3}$$

where $y_{(i)}$ is the ith ordered sample element.

Since

$$V(y) = \tfrac{3}{5}\theta_2^2, \tag{2.4}$$

we can estimate the standard deviation σ by $\sqrt{\frac{3}{5}}\,\theta_2^*$ and the coefficients can be adjusted to give the best linear estimate of the standard deviation σ^*. These adjusted coefficients for which

$$\sigma^* = \sum_{i=1}^{n} \alpha_{2i} y_{(i)} \tag{2.5}$$

are also shown in Table I.

Received August 23, 1954.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

TABLE I

Coefficients in the best linear estimate of the mean and standard deviation, based on the order statistic $y_{(i)}$ in different populations of size n, for the mean $\theta_1^* = \sum_{i=1}^{n} \alpha_{1i} y_{(i)}$ and for the standard deviation $\sigma^* = \sum_{i=1}^{n} \alpha_{2i} y_{(i)}$

Sample size n and population	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{21}	α_{22}	α_{23}	α_{24}	α_{25}
2										
U-shaped.....	.5000000	.5000000				−.9036961	+.9036961			
Parabolic.....	.5000000	.5000000				−.8695889	.8695889			
Skewed.....	.5000000	.5000000				−.8750000	.8750000			
3										
U-shaped.....	.5400000	−.0800000	.5400000			−.6024640	0	.6024640		
Parabolic.....	.4404762	.1190476	.4404762			−.5797259	0	.5797259		
Skewed.....	.4173734	.1539642	.4286624			−.5542721	−.0620262	.6162983		
4										
U-shaped.....	.5530728	−.0530728	−.0530728	.5530728		−.5218559	.0282930	−.0282930	.5218559	
Parabolic.....	.4078157	.0921843	.0921843	.4078157		−.4704178	−.0310128	.0310128	.4704178	
Skewed.....	.3703449	.1237432	.1145196	.3913123		−.4324793	−.0796294	−.0018219	.5139305	
5										
U-shaped.....	.5584792	−.0448589	−.0272406	−.0448589	.5584792	−.4890315	+.0318491	0	−.0318491	.4890315
Parabolic.....	.3862869	.0795401	.0683459	.0795401	.3862869	−.4105156	−.0392644	0	.0392644	.4105156
Skewed.....	.3387590	.1080812	.0903216	.0953199	.3675182	−.3655858	−.0791405	−.0305948	.0192066	.4561145

TABLE II

Variances of the best linear estimates of the mean and standard deviation in different populations ($\sigma = 1$)

Population and sample size	Variance of the estimate of	
	mean	standard deviation
2		
U-shaped	.5000000	.6333333
Rectangular	.5000000	.5000000
Parabolic	.5000000	.5123457
Triangular	.5000000	.5306132
Normal	.5000000	.57079
Double Exponential	.5000000	.7777778
3		
U-shaped	.2501299	.2161616
Rectangular	.3000000	.2000000
Parabolic	.3208101	.2220975
Triangular	.3293975	.2414966
Normal	.3333333	.27548
Double Exponential	.2947532	.4320999
4		
U-shaped	.1279837	.0955036
Rectangular	.2000000	.1111111
Parabolic	.2315500	.1335981
Triangular	.2443499	.1514217
Normal	.2500000	.18005
Double Exponential	.2077706	.2986242
5		
U-shaped	.0675462	.0470213
Rectangular	.1428371	.0714286
Parabolic	.1790064	.0925499
Triangular	.1934059	.1079590
Normal	.2000000	.13332
Double Exponential	.1584266	.2288250

The variances of the estimates of the mean and standard deviation are given in Table II. Furthermore, the relative efficiencies of the sample mean, median, and the midrange as estimates of the population mean are shown in Table III. Similarly the relative efficiencies of the range, the normal estimate, and Gini's estimate are also given in the same table. The efficiencies are calculated relative to the best linear estimate.

Table I shows that the two extreme values in the estimate of the mean have large weights while the middle elements have negative weights.

Comparing the efficiencies (Table III) of the estimates of the mean, we see that the midrange is more efficient than either the sample mean or the median. Again, the range as an estimate of standard deviation has a higher efficiency than either the normal or Gini's estimate. So, the midrange and the range (which are based on the two extreme values) can be used to estimate the popula-

TABLE III

Percentage efficiencies of certain estimates of the mean and standard deviation relative to BLE, *in different populations, from ordered samples of size n*

Population and sample size	Estimates of the mean			Estimates of standard deviation		
	$\bar{y}$	$\tilde{y}$	w	R	N	G
2						
U-shaped	100	100	100	100	100	100
Parabolic	100	100	100	100	100	100
Skewed	100	100	100	100	100	100
3						
U-shaped	75.04	30.57	98.77	100	100	100
Parabolic	96.24	60.25	98.83	100	100	100
Skewed	97.41	61.29	98.59	99.57	99.57	99.57
4						
U-shaped	51.19	23.95	97.16	99.61	86.34	90.99
Parabolic	92.62	65.27	97.35	99.81	52.80	97.73
Skewed	95.91	68.59	97.61	98.84	98.56	97.64
5						
U-shaped	33.77	9.36	95.39	99.01	77.50	69.27
Parabolic	89.50	49.56	95.91	99.55	96.03	91.49
Skewed	91.36	51.58	93.05	97.96	97.40	95.51

Here $\bar{y}$ denotes the sample mean, $\tilde{y}$ denotes median, w denotes the midrange, R denotes the range, N denotes the normal estimates, and G denotes the Gini's mean difference.

tion mean and standard deviation in this distribution for the sample sizes without great loss of accuracy.

3. Parabolic population. The frequency distribution of a parabolic population is

$$f(y) = \frac{6(y - \theta_1 + \frac{1}{2}\theta_2)(\theta_1 + \frac{1}{2}\theta_2 - y)}{\theta_2^3}, \qquad \theta_1 - \tfrac{1}{2}\theta_2 \leqq y \leqq \theta_1 + \tfrac{1}{2}\theta_2, \tag{3.1}$$

where θ_1 is the true mean and θ_2 is the range. Standardizing the variable we get

$$f(x) = 6x(1 - x), \qquad 0 \leqq x \leqq 1. \tag{3.2}$$

The coefficients α_{1i} in the best linear estimate of the mean (θ_1^*) are given in Table I.

Since

$$V(y) = \tfrac{1}{20}\theta_2^2, \tag{3.3}$$

we can estimate the standard deviation σ^* by $(1/\sqrt{20})\theta_2^*$ and the coefficients can be adjusted to give the best linear estimate of the standard deviation σ^*. These adjusted coefficients for which

$$\sigma^* = \sum_{i=1}^{n} \alpha_{2i} y_{(i)} \tag{3.4}$$

are given in Table I. The variances of the estimates of the mean and standard deviation are given in Table II.

Table III gives the percentage efficiencies of the different estimates relative to the best linear estimate. In the best linear estimate of the mean we find that the extreme values have higher weights while the middle elements have smaller positive weights (decreasing towards the middle).

For the given sample sizes, the midrange as an estimate of the population mean is shown to be more efficient than the sample mean (Table III), while the median has low efficiency. Furthermore, the range as an estimate of the standard deviation is more efficient than either the normal or the Gini's estimate as shown in Table III.

4. A skewed population. The frequency distribution of a skewed population is

$$\frac{12}{\theta_2}\left(\frac{y-\theta_1}{\theta_2}+\tfrac{2}{3}\right)^2\left(\tfrac{1}{3}-\frac{y-\theta_1}{\theta_2}\right), \qquad \theta_1-\frac{2\theta_2}{3}\leqq y\leqq\theta_1+\frac{2\theta_2}{3}, \tag{4.1}$$

where θ_1 is the true mode and θ_2 is the true range. Let

$$x=\frac{y-\theta_1}{\theta_2}+\tfrac{2}{3}, \tag{4.2}$$

to get

$$f(x)=12x^2(1-x). \tag{4.3}$$

Since the population mean is $\theta_1-\theta_2/15$, and the population standard deviation is $\theta_2/5$, then the coefficients can be adjusted to give the estimates for the mean μ and the standard deviation σ. These can be obtained from

$$\mu^*=\theta_1^*-\theta_2^*/15, \tag{4.4}$$

$$\sigma^*=\theta_2^*/5. \tag{4.5}$$

The adjusted coefficients in the BLE of the mean μ^* and standard deviation σ^* are given in Table I. The efficiencies of the estimates are given in Table III.

In this case, again, we find that the two extreme sample elements have the greatest numerical weights in the BLE while the other values have smaller weights. It is of interest to see that the least sample value (the extreme value on the side of the long tail) has a smaller coefficient than the largest sample value (the other extreme on the side of the shorter tail). This is to be expected since extreme values from the longer tail occur more often and tend to upset the estimate. It throws some light on the effect of the shape of the distribution or the length of its tails on the coefficients of the BLE. This is not the only relation, however, and the nature of the general relation is not yet well known.

The midrange has a higher efficiency for the given sample sizes than that of the sample mean while the median has a lower efficiency. Again, the range has a higher efficiency than either the normal or the Gini's estimate.

5. Coefficients in the BLE of the mean for symmetric distributions. We have seen in [3], and in the previous sections that the coefficients in the best linear estimate of the mean vary as the parent distribution undergoes change. It is of interest to notice the sequence of this variation. The sample elements may have equal weights or smaller weights at the middle than at the tail, or zero weights at the middle and equal weights at the extremes or large weights on the tails and negative weights in the middle. There is a sequence in which the middle elements are to be equally weighted, zero weighted, and negatively weighted.[1] It seems that the full sequence is missing its natural extension and the complete sequence should read:

(a) negative weights in the middle and large positive weights at the tails,

(b) zero weights in the middle and equal weights at the tails,

(c) less weights in the middle than at the tails,

(d) equal weights throughout,

(e) more weight in the center and less weights in tails, but all positive weights,

(f) middle observations receive all the weight, others nothing,

(g) middle observations receive more than unity and tails take on negative weights.

This is the sequence which might be anticipated. The results show that (a) is U-shaped; (b) is rectangular; (c) is triangular or parabolic; (d) is normal; (e) is double exponential; (f) is the case where the median gets all the weight, which is like a double exponential but not exactly. For (g) the author does not know any example at this time, i.e., a distribution where it would be best to estimate the mean by giving the middle element a weight greater than one and to give the elements on the tails some negative weights. This represents, however, a natural continuity in the sequence.

6. The variances of the best linear estimates. Table II gives the variances of the best linear estimates of the mean and standard deviation in different symmetric distributions with $\sigma = 1$. The variances of the estimates of the normal population are obtained from Tables 5 and 6 in [1] calculated to five decimal places.

The table shows that the variance of the best linear estimate of the mean of a U-shaped population (for $n > 2$) is the least among the given distributions. This raises the theoretical problem of finding the distribution whose mean can be estimated with the least variance.

The same table shows also that the variance of the estimate of the mean increases gradually from the case of the U-shaped distribution to the rectangular, to the parabolic, to the triangular, and then to the normal. The variance of the estimate of the mean then decreases in the case of double exponential.

As to the variance of the estimate of standard deviation, the same table shows that the variance of the estimate increases from the rectangular to the

[1] The author wishes to thank Professor Frederick Mosteller for directing his attention to this particular sequence.

parabolic, to the triangular, to the normal, and then to the double exponential. For the U-shaped distribution, the variance of the estimate is greater than that of the rectangular for $n = 2$ and 3. For $n = 4$ and 5, the variance becomes smaller than that of the rectangular. However, working out the estimates and their variances for $n = 6$ and 7, it has been found that the variance of the estimate of standard deviation for the U-shaped becomes progressively smaller than that of the rectangular. So it seems to the author that as n increases, the variance of the estimate of the standard deviation of the U-shaped distribution tends to be the least among the given distributions.

The author wishes to acknowledge the kind help of Dr. B. G. Greenberg, under whose direction this work was done.

REFERENCES

[1] A. K. GUPTA, "Estimation of the mean and standard deviation of a normal population from a censored sample," *Biometrika*, Vol. 39 (1952), pp. 260–273.

[2] E. H. LLOYD, "Least squares estimation of location and scale parameters using order statistics," *Biometrika*, Vol. 39 (1952), pp. 88–95.

[3] A. E. SARHAN, "Estimation of the mean and standard deviation by order statistics," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 317–328.