

# SIMPLIFIED ESTIMATION FROM CENSORED NORMAL SAMPLES

BY W. J. DIXON

*University of California, Los Angeles*

**0. Summary.** Estimators of mean and standard deviation for censored normal samples which are based on linear systematic statistics and which use simple coefficients are almost as efficient as estimators using the best possible coefficients. Estimators are given for samples of size  $N \leq 20$  for censoring at one extreme and for several types of censoring at both extremes.

**1. Introduction.** A censored sample is a sample lacking one or more observations at either or both extremes with the number and positions of the missing observations known. Censoring may take place naturally i.e., an observation has a magnitude known only to be more extreme than the other observations in the sample. Censoring may also be imposed by the experimenter who from past experience knows that extreme observations are so unreliable that their magnitudes should not be used as observed. The experimenter may impose censoring to reduce the duration of an experiment and obtain estimates before the extreme cases are determined. Estimation of the mean and standard deviation of a normal distribution from a sample which is censored has been considered by Sarhan and Greenberg [1], who obtained coefficients for best linear systematic statistics. They also record efficiencies of these estimators compared to the case of no censoring. Winsor [4] and perhaps others have suggested using for the magnitude of an extreme, poorly known, or unknown observation the magnitude of the next largest (or smallest) observation. We shall show that when symmetry is maintained (or proper adjustment is made) this practice results in estimators of the mean whose efficiencies are scarcely distinguishable from those of best linear estimators. For non-symmetrical censoring, it is demonstrated that optimum simple estimators of the mean result from these "Winsorized" estimators. Also presented are estimators of the standard deviations using one or two ranges (not necessarily symmetrical) which have efficiency .94 or greater when compared with the best linear systematic statistics.

The variances of the proposed estimators were computed from an original 21 decimal tabulation of the means variances and covariances of the order statistics made available by Dan Teichroew. These tables are described in reference [5]. The efficiencies are the ratios of variances of corresponding estimators given by Sarhan and Greenberg [1].

**2. Symmetrical censoring. Estimation of mean.** If natural or imposed censoring of the sample results in the same number of observations censored from each extreme of the sample the practice of using for each missing observation the magnitude of its nearest neighbor whose magnitude is known has a minimum

Received August 24, 1959; revised January 4, 1960.

TABLE I

Relative efficiency for estimate  $m_w$  compared with best linear systematic statistic, when censoring involves  $i - 1$  observations at one extreme and  $i$  observation at the other extreme.

| $N \backslash i$ | 1     | 2     | 3     | 4     | 5     | 6     |
|------------------|-------|-------|-------|-------|-------|-------|
| 3                | 1.000 |       |       |       |       |       |
| 4                | .962  |       |       |       |       |       |
| 5                | .964  | 1.000 |       |       |       |       |
| 6                | .969  | .964  |       |       |       |       |
| 7                | .973  | .963  | 1.000 |       |       |       |
| 8                | .977  | .967  | .967  |       |       |       |
| 9                | .980  | .971  | .965  | 1.000 |       |       |
| 10               | .982  | .974  | .968  | .971  |       |       |
| 11               | .984  | .977  | .971  | .968  | 1.000 |       |
| 12               | .986  | .979  | .974  | .970  | .974  |       |
| 13               | .987  | .981  | .976  | .972  | .970  | 1.000 |
| 14               | .988  | .983  | .979  | .975  | .971  | .976  |
| 15               | .989  | .985  | .981  | .977  | .973  | .973  |
| 16               | .990  | .986  | .982  | .979  | .975  | .973  |
| 17               | .991  | .987  | .984  | .980  | .977  | .975  |
| 18               | .991  | .988  | .985  | .982  | .979  | .976  |
| 19               | .992  | .989  | .986  | .983  | .981  | .978  |
| 20               | .992  | .989  | .987  | .984  | .982  | .980  |

relative efficiency of .99912 (this occurs for  $N = 20, i = 4$ ) when compared with the best linear systematic statistic, BLSS, as given by Sarhañ and Greenberg [1] for  $N \leq 20$ . For  $i$  observations censored at each extreme this estimator is

$$m_w = [(i + 1)x_{i+1} + x_{i+2} + \dots + x_{N-i-1} + (i + 1)x_{N-i}] / N$$

Efficiency is defined here as the ratio of  $\text{Var}(\text{BLSS})/\text{Var}(m_w)$ . Table III of reference [1a] and Table II of reference [1b] may be used for  $\text{Var}(m_w)$  to three or four figures of accuracy since the efficiency is virtually 1.000 for all cases of symmetrical censoring for  $N \leq 20$ .

**3. Almost symmetrical censoring. Estimation of mean.** If one more observation is censored from one extreme than from the other extreme one may consider the simple procedure of dropping another observation to symmetrize censorship and proceed as in Section 2. Efficiencies of the resulting estimators compared with BLSS are given in Table I. For each  $i$ , the efficiencies first decrease and then increase with increasing  $N$  and the minimum increases with  $i$  from .962 for  $i = 1$  for  $N \leq 20$  and  $i \leq 6$ . It therefore seems reasonable to assume that the efficiency is never less than .962. In the example of reference [1] and [2] for the sample of ten

—, —, 108, 111, 119, 121, 125, —, —, —

TABLE II

Relative efficiencies of  $m_W$  and  $m_a$  compared with best linear systematic statistic for samples with  $i$  observations censored at one extreme. The coefficient  $a$  is used to obtain the estimate  $m_a$ .

| $i \backslash N$ | 1     |       |      | 2     |       |       | 3     |       |        |
|------------------|-------|-------|------|-------|-------|-------|-------|-------|--------|
|                  | $m_W$ | $m_a$ | $a$  | $m_W$ | $m_a$ | $a$   | $m_W$ | $m_a$ | $a$    |
| 3                | 1.000 | 1.000 | 0    |       |       |       |       |       |        |
| 4                | .962  | .998  | .289 |       | 1.000 | -.866 |       |       |        |
| 5                | .964  | .998  | .426 | .990  | .992  | -.426 |       | 1.000 | -1.703 |
| 6                | .969  | .998  | .506 | .963  | .992  | -.188 |       | .986  | -1.143 |
| 7                | .973  | .999  | .560 | .959  | .993  | -.038 | .984  | .985  | -.821  |
| 8                | .977  | .999  | .599 | .960  | .994  | .066  | .965  | .986  | -.609  |
| 9                | .980  | .999  | .628 | .963  | .995  | .143  | .958  | .988  | -.458  |
| 10               | .982  | .999  | .651 | .966  | .996  | .202  | .957  | .989  | -.345  |
| 11               | .984  | .999  | .669 | .968  | .997  | .249  | .958  | .991  | -.255  |
| 12               | .986  | .999  | .685 | .971  | .997  | .288  | .960  | .992  | -.183  |
| 13               | .987  | .999  | .698 | .973  | .997  | .321  | .962  | .993  | -.123  |
| 14               | .988  | 1.000 | .709 | .975  | .998  | .349  | .964  | .994  | -.073  |
| 15               | .989  | 1.000 | .719 | .977  | .998  | .373  | .966  | .994  | -.029  |
| 16               | .990  | 1.000 | .727 | .979  | .998  | .394  | .968  | .995  | .008   |
| 17               | .991  | 1.000 | .735 | .980  | .998  | .413  | .970  | .995  | .041   |
| 18               | .991  | 1.000 | .742 | .981  | .998  | .429  | .972  | .996  | .070   |
| 19               | .992  | 1.000 | .748 | .982  | .998  | .444  | .973  | .996  | .097   |
| 20               | .992  | 1.000 | .754 | .984  | .999  | .458  | .975  | .996  | .120   |

  

| $i \backslash N$ | 4     |       |        | 5     |       |        | 6     |       |        |
|------------------|-------|-------|--------|-------|-------|--------|-------|-------|--------|
|                  | $m_W$ | $m_a$ | $a$    | $m_W$ | $m_a$ | $a$    | $m_W$ | $m_a$ | $a$    |
| 6                |       | 1.000 | -2.532 |       |       |        |       |       |        |
| 7                |       | .981  | -1.864 |       | 1.000 | -3.361 |       |       |        |
| 8                |       | .977  | -1.466 |       | .977  | -2.591 |       | 1.000 | -4.190 |
| 9                | .981  | .978  | -1.198 |       | .969  | -2.122 |       | .973  | -3.324 |
| 10               | .966  | .980  | -1.002 |       | .969  | -1.800 |       | .963  | -2.787 |
| 11               | .959  | .982  | -.853  | .978  | .971  | -1.562 |       | .960  | -2.412 |
| 12               | .957  | .984  | -.734  | .966  | .973  | -1.378 |       | .962  | -2.133 |
| 13               | .957  | .986  | -.638  | .960  | .976  | -1.231 | .977  | .964  | -1.916 |
| 14               | .957  | .987  | -.557  | .957  | .978  | -1.109 | .967  | .967  | -1.740 |
| 15               | .959  | .988  | -.488  | .956  | .980  | -1.008 | .961  | .970  | -1.596 |
| 16               | .960  | .989  | -.429  | .956  | .982  | -.921  | .958  | .972  | -1.473 |
| 17               | .962  | .991  | -.378  | .957  | .984  | -.846  | .956  | .975  | -1.369 |
| 18               | .964  | .991  | -.332  | .958  | .985  | -.781  | .956  | .977  | -1.278 |
| 19               | .965  | .992  | -.292  | .959  | .987  | -.723  | .956  | .979  | -1.198 |
| 20               | .967  | .993  | -.256  | .960  | .987  | -.671  | .957  | .980  | -1.127 |

the BLSS estimate of mean is 118.9. The estimate

$$m_w = [4(111) + 119 + 121 + 4(125)]/10 = 118.4.$$

**4. Censoring entirely at one extreme. Estimation of mean.** If  $i$  observations are censored at one extreme, one may consider dropping  $i$  observations at the other extreme to produce symmetry and proceed as in Section 2. For  $i \leq 6$  the efficiency of this estimator is never less than .956. Since the efficiencies for each  $i \leq 6$  are increasing at  $N = 20$  it seems reasonable to assume this minimum holds for all  $N$  with  $i \leq 6$ . If fewer observations are dropped, some adjustment must be made to maintain an unbiased estimator. A simple estimator which usually has greater efficiency is

$$m_a = [ax_1 + x_2 + \dots + x_{N-i-1} + (i + 1)x_{N-i}]/(N + a - 1)$$

Here  $a$  is chosen as a coefficient of  $x_1$ , i.e. chosen to satisfy  $E(m_a) = \mu$  and the other extreme is "Winsorized" as in the estimator  $m_w$ . If  $i$  is not large  $m_a$  shows very little loss in efficiency from the BLSS, and of course it is possible to estimate the mean for smaller sample sizes than is possible if one arbitrarily makes the

TABLE III

Relative efficiencies of estimates based on ranges of samples compared with best linear systematic statistic for estimating standard deviation from samples censored of  $i$  observations at each extreme. Estimate is maximum range except where noted.

| $N \backslash i$ | 1      | 2      | 3     | 4     | 5     | 6     |
|------------------|--------|--------|-------|-------|-------|-------|
| 4                | 1.000  |        |       |       |       |       |
| 5                | 1.000  |        |       |       |       |       |
| 6                | .997   | 1.000  |       |       |       |       |
| 7                | .991   | 1.000  |       |       |       |       |
| 8                | .984   | .999   | 1.000 |       |       |       |
| 9                | .975   | .997   | 1.000 |       |       |       |
| 10               | .966   | .993   | 1.000 | 1.000 |       |       |
| 11               | .966*  | .989   | .998  | 1.000 |       |       |
| 12               | .969*  | .984   | .997  | 1.000 | 1.000 |       |
| 13               | .969*  | .979   | .994  | .999  | 1.000 |       |
| 14               | .968*  | .973   | .992  | .998  | 1.000 | 1.000 |
| 15               | .966*  | .967   | .989  | .997  | .999  | 1.000 |
| 16               | .967** | .967*  | .985  | .995  | .999  | 1.000 |
| 17               | .968** | .967*  | .981  | .993  | .998  | 1.000 |
| 18               | .968** | .967*  | .977  | .991  | .997  | .999  |
| 19               | .968** | .966*  | .973  | .988  | .996  | .999  |
| 20               | .966** | .965** | .969  | .986  | .994  | .998  |

\* Efficiency for estimate based on  $(x_{N-i} - x_{i+1}) + (x_{N-i-1} - x_{i+2})$ .

\*\* Efficiency for estimate based on  $(x_{N-i} - x_{i+1}) + (x_{N-i-2} - x_{i+3})$ .

TABLE IV

Relative efficiencies of estimates based on ranges of samples compared with best linear systematic statistic for estimates of standard deviation from samples censored for  $i - 1$  observations at one extreme and  $i$  observations at the other extreme. Estimate is based on maximum range except where noted. Efficiencies and estimates for  $i = 1$  as given in Table V.

| $N \backslash i$ | 2      | 3     | 4     | 5     | 6     |
|------------------|--------|-------|-------|-------|-------|
| 5                | 1.000  |       |       |       |       |
| 6                | .998   |       |       |       |       |
| 7                | .995   | 1.000 |       |       |       |
| 8                | .990   | .999  |       |       |       |
| 9                | .984   | .998  | 1.000 |       |       |
| 10               | .977   | .996  | 1.000 |       |       |
| 11               | .973*  | .993  | .999  | 1.000 |       |
| 12               | .973*  | .990  | .998  | 1.000 |       |
| 13               | .972*  | .986  | .997  | 1.000 | 1.000 |
| 14               | .969*  | .982  | .995  | .999  | 1.000 |
| 15               | .965*  | .977  | .992  | .998  | 1.000 |
| 16               | .966** | .973* | .990  | .997  | .999  |
| 17               | .967** | .972* | .987  | .995  | .999  |
| 18               | .967** | .970* | .983  | .994  | .998  |
| 19               | .967** | .968* | .980  | .992  | .997  |
| 20               | .965** | .965* | .976  | .990  | .996  |

\* Estimate is based on  $(x_{N-i} - x_i) + (x_{N-i} - x_{i+1})$ .

\*\* Estimate is based on  $(x_{N-i} - x_i) + (x_{N-i-1} - x_{i+2})$ .

sample symmetric and uses  $m_w$  as suggested above. Table II lists the efficiencies for these two types of estimators and lists the values of  $a$  for the estimator  $m_a$ .

**5. Estimation of standard deviation. Symmetrical censoring.** Any estimator of the standard deviation based on a sample whose extremes are censored has low efficiency since the observations of greatest importance are not available. For example if one extreme observation in a sample of 10 is missing the BLSS has efficiency .837 compared with the sample standard deviation based on all ten observations; for one extreme observation censored from a sample of five the efficiency is .677. Furthermore, the situation rapidly deteriorates for more observations censored. It seems of interest to investigate whether an estimate of standard deviation based on ranges will more than slightly depress these efficiencies.

For  $i$  observations censored from each extreme an estimate of the standard deviation based on an optimum choice of one or two ranges has minimum relative efficiency .965 compared with the BLSS for  $i \leq 6$  and  $N \leq 20$ . Table III indicates these estimators and efficiencies. For similar estimators for the case of no censoring see [3]. This table and also Tables IV and V indicate the range or



one or two ranges has minimum efficiency .965 compared to BLSS for  $1 < i \leq 6$  and  $N \leq 20$ . These estimators and efficiencies are given in Table IV. Table IV indicates the use of two ranges for certain cases. The increase in efficiency for two ranges can be seen by comparison with the efficiency of the maximum range alone which for  $i = 1$  is .937 for  $N = 15$  and .896 for  $N = 20$ ; and for  $i = 2$  is .950 for  $N = 20$ .

**7. Censoring entirely at one extreme. Estimation of standard deviation.** For  $i$  observations censored at one extreme an estimator of the standard deviation based on an optimum choice of one or two ranges has minimum efficiency .937 compared to BLSS for  $i \leq 6$  and  $N < 20$ . These estimators and efficiencies are given in Table V. The estimators are indicated by the order of the observations used in the estimator. For example, the designation of 1, 3,  $N - 5$ ,  $N - 5$  for  $N = 15$  indicates the estimator  $K(2x_{10} - x_3 - x_1)$  where  $K^{-1} = E(2x_{10} - x_3 - x_1)$  and the expectation applies to the unit normal table. For this example,  $K^{-1} = 2(.33530) + .94769 + 1.73591 = 3.35420$ . The optimum solution for most cases requires the use of an extreme observation at the censored end with doubled weight rather than two different observations.

## REFERENCES

- [1] AHMED E. SARHAN AND BERNARD G. GREENBERG, "Estimation of location and scale parameters by order statistics from singly and doubly censored samples." *Ann. Math. Stat.*, (I) Vol. 27 (1956) pp. 427-451 and (II) Vol. 29 (1958) pp. 79-105.
- [2] BERNARD G. GREENBERG AND AHMED E. SARHAN, "Applications of order statistics to health data," *Amer. J. of Public Health*, Vol. 48 (1958) pp. 1388-94.
- [3] W. J. DIXON, "Estimates of the mean and standard deviation of a normal population," *Ann. Math. Stat.*, Vol. 28 (1957) pp. 806-809.
- [4] CHARLES P. WINSOR, Personal communication.
- [5] D. TEICHROEW, "Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution," *Ann. Math. Stat.* Vol. 27, (1956) pp. 410-426.