Use of Order Statistics in Estimating Standard Deviations

By H. F. Huddleston

Although many statistical surveys and experimental studies have been and are being conducted for many segments of the economy, including agriculture, standard errors are seldom computed. Estimates of these errors are frequently needed to evaluate survey results and for the planning of future studies. The computation of standard errors is frequently omitted because of the time-consuming procedures required. The use of rank, or order, methods of analysis has increased rapidly in the last few years. These methods provide the analyst with a quick, effective, and inexpensive tool for making many statistical estimates. In this paper, the use of order statistics to estimate standard deviations for certain agricultural series is described and the results are compared with those obtained by the root mean square method.

AT the Bureau of the Census the author needed to compute some 100,000 standard deviations within a period of a few months. Using order statistics, as described in this article, it was possible to get the job done with only a few clerks and desk calculators.

The need for estimating standard deviations frequently arises in statistical work in the Department of Agriculture and cooperating agencies. The procedures based on order statistics are easy to apply, relatively unbiased, and efficient, and they are appropriate for a large class of distributions. Too often they are discounted by statisticians who prefer more "powerful" statistics. They are little used by workers except in quality control, despite their simplicity, and the economies that frequently result from their use.

Within a period of 15 minutes a clerk can "search" or machine sort a sample of 200 items to obtain the ordered values required to estimate the standard deviation and divide by an appropriate constant. Because of the labor involved in their estimation, standard errors are frequently not computed at all or only "guessed at" in the planning of surveys and evaluating results of probability samples.

Order statistics offer considerable economy in estimating levels of sampling errors in connection with large-scale operations, such as a sample census of agriculture or other multipurpose surveys. Thus it becomes practicable to indicate the degree of precision of surveys at the time results are published, and to provide estimates of variability for many problems of sample design. The author's experience with order statistics suggests that other workers might find similar procedures useful.

Results for several items in the 1950 Census of Agriculture are given, together with results from the root mean square method. The nature of the bias that may be associated with such estimates is examined for certain populations. Whereas the use of only two or four observations out of a sample of n may appear grossly inefficient on intuitive grounds, order statistics characterize the shape of sample distribution in the tails where the contributions to the variability are greatest.

Procedures for Estimating Standard Deviation

Estimates of the standard deviation are constructed by selecting one or more pairs of order statistics which specify a given proportion from the respective tails of the distribution. However, the best known and most widely used estimate is based on the sample range, defined as follows:

$$\hat{\sigma} = (X_n - X_1)/C_{1/n}$$

where $C_{1/n}$ is expected value of the difference $(Y_n - Y_1)$, Y_n and Y_1 being the greatest and least observations drawn from a sample of size n from a normal distribution with unit variance. That is, $C_{1/n}$ is the mean value of the ratio of the range to the standard deviation. Tables of $C_{1/n}$ for various size samples are available for estimating the standard deviation from the simple range R_u (i. e., $X_n - X_1$) which have been published in tables for statisticians and biometricians and in various quality control texts.[1]

We may likewise consider the use of various other pairs of order statistics such as $(X_{n-m+1}$

[1] See for example: GRANT, E. L. STATISTICAL QUALITY CONTROL. Appendix III, table B.

$-X_m$) or R_m. For instance, we can use the statistics

$$\hat{\sigma}_m = (X_{n-m+1} - X_m)/C_{m/n}$$

where $C_{m/n}$ is the expected value of the difference $(Y_{n-m+1} - Y_m)$, and we count in m observations from each end of a sample of size n ordered according to the magnitudes of the items and the Y's are drawn from the normal distribution with unit variance. The problem of which of the various pairs of order statistics or combinations of pairs is most appropriate has been resolved by Mosteller[3] for the normal distribution. If we are interested in the optimum spacing of the order statistics in the minimum variance sense, we find for large sample sizes when $\lambda = m/n$ that the minimum value of the variance of $\hat{\sigma}$, occurs when $\lambda = 0.0694$. However, the value of the variance of $\hat{\sigma}$ changes slowly in this neighborhood. Hence, varying λ by 0.01 or 0.02 will make little difference in the efficiency of the estimate $\hat{\sigma}$. For practical purposes, the optimum values for λ are 0.07 and 0.93. The value of λ from the lower tail is denoted as $m/n = \lambda_1$ and the value for the upper tail by $1 - m/n = \lambda_2$.

However, if we wish, we can construct an estimate based upon four order statistics. For the normal distribution Mosteller has shown that if we hold the first two selected order statistics at their optimum values, i. e., $\lambda_1 = 0.07$ and $\lambda_2 = 0.93$, the two additional observations should be more centrally located. Under these conditions the variance of $\hat{\sigma}$ is minimized for λ_3 in the neighborhood of 0.20 and $\lambda_4 = 1 - \lambda_3$. The unbiased estimate of $\hat{\sigma}$ is:

$$\hat{\sigma}_{r,s} = (X_{n-r+1} + X_{n-s+1} - X_s - X_r)/C_{rs/n}$$

where $C_{rs/n}$ is the expected value of the difference of

$$(Y_{n-r+1} + Y_{n-s+1} - Y_s - Y_r),$$

Y_{n-r+1}, Y_{n-s+1}, Y_s and Y_r being observations drawn from a sample of size n from a normal distribution with unit variance.

Tabled values of $C_{1/n}$ are available only for the normal distribution. As the mean value of the $C_{m/n}$ is not available for the numerous sample sizes and distributions encountered in practice, we need to know the utility of the norming constants of $C_{m/n}$ based on large samples drawn from a normal population.

[3] Mosteller, Frederick. On some useful "inefficient" statistics. Annals of Mathematical Statistics. 17: 377. 1946.

Biases Associated With Estimates of Standard Deviations

The use of the norming constants $C_{m/n}$ for the normal population with respect to large samples from several different types of populations will be examined. In particular, we would like to know the nature of any biases which may be encountered when we use the value of $C_{m/n}$ corresponding to the optimum percentage points derived for the normal population. Where $\lambda_1 = 0.07$ and $\lambda_2 = 0.93$ the norming constant, $C_{m/n}$, is 3.0. With respect to the theory the question is, How good is this mean value for general use? On intuitive grounds it would appear that R_m corresponding to very small values of λ_1 for populations having finite ranges will be less than the corresponding range or quasi-range for a normal population. This suggests that $C_{m/n}$ for populations having finite ranges may be less than for the same λ corresponding to a normal distribution. If such is the case, underestimates of $\hat{\sigma}$ will result by using the larger mean value of $C_{m/n}$ (i. e. divisor) corresponding to the normal distribution.

Comparisons with the normal distribution were made by the author for four distributions in table 1. The distributions are the right triangular $f(x) = \frac{2}{a}\left(1 - \frac{x}{a}\right)$, $X \geq 0$; the isosceles triangular $f(x) = \frac{2}{a}\left(1 - \frac{2x}{a}\right)$, $-\frac{a}{2} \leq X \leq \frac{a}{2}$; the rectangular $f(x) = \frac{1}{a}$, $o \leq X \leqq a$; and the exponential $f(x) = e^{-x}$, $X \geq 0$. e^{-x}.

TABLE 1.—*Values of $C_{m/n}$ for large samples corresponding to $\lambda_1 = 0.01$ to 0.10*

λ_1	Type of distribution				
	Normal	Isosceles triangle	Right triangle	Rectangular	Exponential
0.01	4.7	4.2	3.8	3.4	4.6
0.02	4.1	3.9	3.6	3.3	3.9
0.03	3.8	3.7	3.4	3.3	3.5
0.04	3.5	3.5	3.3	3.2	3.2
0.05	3.3	3.4	3.2	3.1	2.9
0.06	3.1	3.2	3.1	3.1	2.8
0.07	3.0	3.1	3.0	3.0	2.6
0.08	2.8	3.0	2.9	2.9	2.4
0.09	2.7	2.8	2.8	2.8	2.3
0.10	2.6	2.7	2.7	2.8	2.2

In addition, comparisons for the chi-square family for different degrees of freedom show a result similar to that found for e^{-x}. As the vast majority of distributions encountered in agriculture are covered by the types examined, the use of the norming constant $C_{m/n}$ near the 7 and 93 percentage points would appear to yield relatively unbiased estimates of the standard deviation. For highly skewed populations possessing some extremely large units or a "contaminated" tail, the norming constant for the normal distribution may underestimate σ. But in sampling agricultural populations or in census enumerations it is a common practice to develop special procedures for handling extremely large units; consequently an estimate of the standard deviation for the remaining portion of the population may be obtained by using the values of $C_{m/n}$ given for the normal distribution.

Working Rules and a Numerical Example

The preceding investigation of several distributions indicates that for large samples the norming constants, for say $\lambda_1=0.07$, may be used in most situations to obtain relatively unbiased estimates of the standard deviation. For moderate size samples there appears to be no reason apriori to believe that the expected values of $C_{m/n}$ for fixed percentage points would be very sensitive to or depend on the sample size, except for $C_{1/n}$. The author knows of no investigation of expected values for various sample sizes greater than 10, and he has not computed them. But the results obtained appear to agree rather well with results to be expected from the large sample values indicated in table 1.

To illustrate which of these "inefficient statistics" should be used for estimating standard deviations, table 2 is given, along with corresponding constants in table 3, as compiled by the author. It is necessary to arrange the two tails in ascending order of magnitude, using either a machine sort or "search" procedure so that $X_1 \leqq X_2 \leqq X_{n-1} < X_n$. Table 2 gives the appropriate X_{ni} and table 3 the $C_{m/n}$ values when only one pair of order statistics is used, except for samples highly skewed to the right, in which case, two pairs of order statistics are used. For moderate size in sample distributions with highly skewed right tails, it has been found worthwhile to use four order statistics.

TABLE 2.—*Pairs of order statistics used in estimating standard deviation for various sample sizes*

Sample size	Select the following sample values (X_{n-s+1} and X_s)	Additional order statistics to be used for sample distributions when $X_{n-r+1} \geq 2X_{n-s+1}$ ($r<s$)
2–25	Largest (X_n) and smallest (X_1).	None.
26–40	2d largest (X_{n-1}) and 2d smallest (X_2).	None.
41–60	3d largest (X_{n-2}) and 3d smallest (X_3).	None.
61–100	5th largest (X_{n-4}) and 5th smallest (X_5).	Largest (X_n) and smallest (X_1) (i. e., $r=1$).
101–200	10th largest (X_{n-9}) and 10th smallest (X_{10}).	2d largest (X_{n-1}) and 2d smallest (X_2) (i. e., $r=2$).
250–500	25th largest (X_{n-24}) and 25th smallest (X_{25}).	3d largest (X_{n-2}) and 3d smallest (X_3) (i. e., $r=3$).
500 or greater.	Use value of X_{ni} corresponding to $n-s+1 \approx 0.93n$ and $s \approx 0.07n$.	Use values of X_{ni} corresponding to $n-r+1 \approx 0.995n$ $r \approx 0.005n$.

That is, whenever the relationship between the sample values is such that $X_{n-r+1} \geqq 2X_{n-s+1}$ the use of two additional order statistics farther out in the tails than $\lambda_1=0.07$ and $\lambda_2=0.93$ tends to eliminate much of the bias that may exist between the estimated standard deviation given by the order statistics and the root mean square method. A comparison of the values of $C_{m/n}$ for $\lambda=0.01$ for the normal and exponential distributions indicate this is to be expected. But in such cases, the root mean square method may also not provide an accurate measure of the population variability.

Table 3 was constructed for use in the situation in which the size of sample was continuously changing and it was desirable to standardize the "searching" or ranking procedure. For instance, the machine operator or clerk was instructed to obtain the 5 largest and 5 smallest values whenever a sample of 61 to 100 items was encountered.

The following example illustrates the technique. For the variable "land in farms" in a sample of 70 farms of a given class in one county the 5 smallest and 5 largest values were:

TABLE 3.—*Values of norming constants to be used with pairs of order statistics given in table 2*

Sample size	One pair used $C_{s/n}$	Two pairs used $C_{s/n}+C_{r/n}$
5	2.3	
10	3.1	
15	3.5	
20	3.7	
25	3.0	
30	3.3	
40	3.5	
50	3.1	
60	3.2	
70	3.0	7.8
80	3.2	8.0
90	3.3	8.2
100	3.4	8.4
110	2.7	6.9
150	3.0	7.4
200	3.3	7.9
250	3.5	8.3
200	2.6	7.1
300	2.8	7.4
400	2.9	7.9
500	3.3	8.3
Over 500	3.0	8.1

X_1, X_2, X_3, X_4, X_5 10, 26, 35, 37, 40 and
$X_{66}, X_{67}, X_{68}, X_{69}, X_{70}$ 100, 120, 150, 200, 240
Here an estimate of σ is

$$(X_{66}-X_5)/C_{5/70} \text{ or } (100-40)/3=20 \text{ acres},$$

while the root mean square estimate was 33 acres. The divisor, 3, comes from line 10, column 2 of table 3 while the ordered values to be used in the numerator are specified by column 2 of table 2 for sample sizes of 61–100. But note that $X_{70}/X_{66}>2$. When this is the case, an estimate based on four order statistics, with the two additional values coming from near the upper and lower 1 percent point, is usually better. In this example such an estimate would be

$$(X_{70}+X_{66}-X_5-X_1)/C_{5,1/70}$$
$$(240+100-40-10)/(3+4.8)=37.2$$

The divisor, 7.8, is given in column 3 of table 3 and the values used in the numerator are specified by column 3 of table 2. When more than one pair of order statistics are used in the estimate, $C_{5,1/70}$ is the sum of the $C_{m/n}$ for each pair; in this example

$$C_{5,1/70}=C_{5/70}+C_{1/70}.$$

Some Results of Empirical Studies

The procedures in the preceding section were used for several agricultural items. Results for six items for various classes of farms are given for the 1950 Census of Agriculture (within-strata σ's). The six items were: Land in farm, crop land harvested, land rented from others, other pasture, unpaid family workers, and tractor repairs. The coefficients of variation rather than the standard deviations are plotted because of the differences in the magnitude of the variables. They are based on sample sizes varying from about 10 to 200.

For the great majority of distributions encountered in agricultural populations, the skewness is of the type found in the chi-square family of curves. But extremely large units are usually eliminated from such populations and samples, as in 1950 Census Enumeration. It is common practice in sampling studies to enumerate extremely large units completely as constituting a separate stratum.

In general, table 1 suggests that estimates of σ would be too low because $C_{m/n}$ for the normal distribution is larger than for the exponential distribution. For the range of λ values used, table 1 would indicate a downward bias of 12–15 percent.

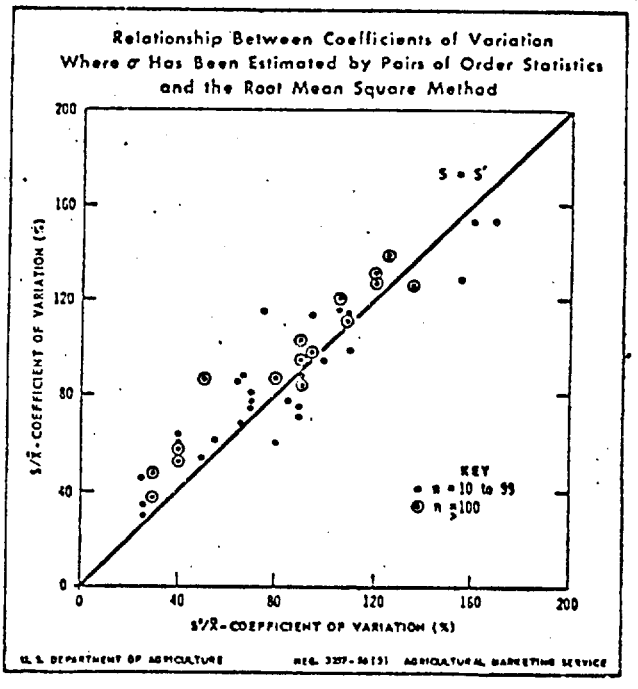

FIGURE 1.

Figure 1 indicates a similar bias. As this is evident for the larger samples ($n \geqq 100$), it appears that a 12–15 percent bias is a good estimate of the bias that can ordinarly be expected for an exponential type of distribution. Other distributions examined in table 1 and data for small samples indicate little or no bias compared with the root mean square estimate.

Book Reviews

Statistics: A New Approach. By W. Allen Wallis and Harry V. Roberts. The Free Press, Glencoe, Ill. 646 pages. 1956. $6.

HERE WE HAVE NOT ONLY a new approach but a most welcome departure from the unappetizing pottage of re-used and worn-out ingredients so often served as a course in "elementary statistical methods." The reader's interest is aroused from the beginning, both by the diversity of statistical problems and applications set before him and by the attractive way in which they are presented. If any pedagogical device can imbue a prospective student with a desire to study this subject, this text should be it. With all this to commend it, it would be disappointing indeed if its technical quality were to fall short of expectations. Any possible misgivings on that score are groundless.

Although the treatment is largely non-mathematical, in the sense that much of the algebraic symbolism cluttering up some texts is happily absent, even a cursory reading makes it clear that the basic concepts of modern statistical thinking are covered in admirable fashion. A surprisingly large amount of material that authors of other elementary texts regard as "too advanced" for beginners is included here without fuss or apology. The treatment confirms what the more astute among beginning students probably have long suspected—the important principles are not hard to grasp when explained in the vernacular by someone who understands them himself.

It is difficult to summarize the content of a work of such scope in a short review. The subject is introduced in a thought-provoking presentation of the nature of statistics and its application to many different subject-matter fields, the planning of statistical investigations, and the interpretation of data. This is a lively discussion with numerous case histories to illustrate ideas.

Tabulation of data and the use of ordinary descriptive statistical measures are covered in the next section. These also are presented in a manner that is a far cry from the usual drab recital with which most of us are all too familiar. Part III takes up sampling theory, statistical inference, probability, sampling distributions, the theory of testing hypotheses, decision functions, and the theory of estimation. The last section, entitled "Special Topics," treats experimental design, sample design, statistical quality control, regression analysis, and time series analysis. An appendix includes tables of squares, square roots, and random digits. Tables of the Normal Probability Integral are pasted to the insides of the two covers.

This reviewer finds only one possible fault with the book if it is used as a text: All of the material is presented so lucidly and challengingly that an instructor will find it hard to restrain himself from trying to crowd it all into a single one-year course. The intent of the authors is clearly that the instructor exercise judgment in the selection of topics for any one course. The wealth of material provided permits a selection that should fit the needs of almost any group with which an instructor is likely to be confronted.

Many mature practicing statisticians trained in the old school could read this work with profit. This reviewer knows of no way in which one could bring himself up to date on modern statistical thinking with less effort or mental strain. A reader at any level of statistical maturity can find something of interest in it.

Walter A. Hendricks