THE USE OF SAMPLE RANGE IN ESTIMATING THE STANDARD DEVIATION OR THE VARIANCE OF ANY POPULATION

By MOTOSABURO MASUYAMA

Meteorological Research Institute, Tokyo
and
Indian Statistical Institute, Calcutta

1. Introduction

In statistical quality control and sometimes in the analysis of large-scale sample surveys, the sample range is used to estimate the standard deviation or the variance in the population. This requires the use of values of the ratio of the expectation of the sample range to the standard deviation as a function of the sample size which has been tabulated for a normal population and is used extensively in industrial practice. For a general continuous distribution, Plackett (1947) gave an upper bound for this ratio and at the same time gave a discontinuous distribution in which this ratio is nearly zero. Tsumura (1952) computed this ratio for rectangular, triangular, exponential and normal distributions and found that if the sample size is small, these values are nearly equal, and fairly close to the upper bound obtained by Plackett.

A problem that arises is this. Can an estimate of the standard deviation or the variance be obtained by using a constant multiple of the range or its square, with the same multiplier for all distributions? In this paper bounds are derived for the ratio of the expectation of the sample range to the population standard deviation and that of the square of the sample range to the population variance. In the opinion of the author, these bounds are quite close, specially if the sample is small, and the harmonic mean of these bounds is suggested as the appropriate multiplier.

2. Estimation of the standard deviation

Let u denote the standard deviation and R the range in a sample of size n. Neglecting the trivial case $R = 0$, it is easy to show that

$$U = \frac{u}{R_{max}} = \begin{cases} 1/2\sqrt{n/(n-1)} & \text{if } n \text{ is even} \\ 1/2\sqrt{(n+1)/n} & \text{if } n \text{ is odd} \end{cases} \qquad \ldots \quad (2.1)$$

$$L = \frac{u}{R_{min}} = \sqrt{1/2(n-1)}. \qquad \ldots \quad (2.2)$$

If we now define coefficients b and d by

$$E(u) = b\sigma \qquad \ldots \quad (2.3)$$

$$E(R) = d\sigma \qquad \ldots \quad (2.4)$$

σ being the standard deviation in the population, it follows that

$$b/U \leqslant d \leqslant b/L \qquad \cdots \quad (2.5)$$

or

$$L/b \leqslant \frac{1}{d} \leqslant U/b. \qquad \cdots \quad (2.6)$$

Now the harmonic mean of given upper and lower bounds of an unknown quantity gives an approximation which minimises the maximum relative error (Polya, 1950).

In this sense therefore, we suggest the approximation

$$\frac{1}{d} \doteqdot H/b \qquad \cdots \quad (2.7)$$

where H is the harmonic mean of U and L

$$1/H = 1/2\left(\frac{1}{L}+\frac{1}{U}\right). \qquad \cdots \quad (2.8)$$

On the other hand, let us denote the value of d for a normal population by d_2. Tables of values of d_2 are available but a simple expression

$$D = n/\sqrt{(n-1/2)} \qquad \cdots \quad (2.9)$$

serves as quite a good approximation to d_2.

We present below the values of U, L, H, $1/D$ and $1/d_2$ for values of $n = 2$ to $n = 10$

TABLE 1. MULTIPLIER FOR THE RANGE FOR ESTIMATING THE STANDARD DEVIATION

n	U	L	H	$1/D$	$1/d_2$
2	0.7071	0.7071	0.7071	0.612	0.8865
3	0.5774	0.5000	0.5359	0.527	0.5907
4	0.5774	0.4082	0.4782	0.468	0.4857
5	0.5477	0.3536	0.4207	0.429	0.4299
6	0.5477	0.3162	0.4014	0.391	0.3946
7	0.5345	0.2673	0.3749	0.364	0.3698
8	0.5345	0.2673	0.3564	0.343	0.3512
9	0.5270	0.2500	0.3391	0.324	0.3367
10	0.5270	0.2357	0.3257	0.308	0.3249

It will be seen from the above table that :

(i) the difference between the two bounds U and L is not much if n is small,

(ii) $1/D$ and $1/d_2$ lie between the two bounds U and L except when $n = 2$ and $n = 3$,

(iii) for $n > 4$, H is close to $1/d_2$,

(iv) for $n > 4$, $1/D$ is a good substitute for $1/d_2$.

The difficulty now is shifted to the problem that the coefficient b in (2.3) may not be known. However, if the sample size is small, the value of b for a wide class of distributions is expected to be nearly independent of the original distribution, in the sense that bounds b_u and b_l for b not depending on the nature of the original distribution may be available. In this paper, however, we are not concerned with the problem of determining such a class of distributions.

3. Estimation of the variance

Starting with the following bounds for the ratio U^2/R^2

$$U = \frac{u^2}{R^2_{\max}} = \begin{cases} n/4(n-1) & \text{if } n \text{ is even} \\ (n+1)/4n & \text{if } n \text{ is odd} \end{cases} \qquad \ldots \quad (3.1)$$

$$L = \frac{u^2}{R^2_{\min}} = 1/2(n-1) \qquad \ldots \quad (3.2)$$

and defining the coefficient g by

$$E(R^2) = g\sigma^2 \qquad \ldots \quad (3.3)$$

it is easy to show that $1/U \leqslant g \leqslant 1/L$

or

$$L \leqslant \frac{1}{g} \leqslant U. \qquad \ldots \quad (3.4)$$

As before, we suggest the use of the harmonic mean H of U and L as an approximation for the value of $1/g$.

Denoting by g_2 the value of the coefficient g for the normal population, it is found that the simple formula

$$1/G = 0.45/(n-1) + 0.05$$

gives quite a close approximation to $1/g_2$.

In Table 2 below, the values of U, L, H, $1/G$ and $1/g_2$ are presented for $n = 2$ to $n = 10$.

TABLE 2. MULTIPLIER FOR THE SQUARE OF THE RANGE FOR ESTIMATING THE VARIANCE

n	U	L	H	$1/G$	$1/g_2$
2	0.5000	0.5000	0.5000	0.500	0.500
3	0.3333	0.2500	0.2857	0.275	0.274
4	0.3333	0.1667	0.2222	0.200	0.199
5	0.3000	0.1250	0.1765	0.163	0.162
6	0.3000	0.1000	0.1500	0.140	0.140
7	0.2857	0.0833	0.1200	0.125	0.125
8	0.2857	0.07143	0.1143	0.114	0.114
9	0.2778	0.06250	0.1020	0.106	0.106
10	0.2778	0.05556	0.09259	0.100	0.099

It will be seen from Table 2 that (i) the difference between the bounds U and L is not much for small n (ii) $1/G$ and $1/g_2$ both lie within the two bounds and (iii) H or $1/G$ is quite close to $1/g_2$.

However, recently Moriguti (1954) has derived better bounds for the square of the sample range.

4. Estimation in the correlated case

If we put
$$(s/R)^2 = \phi. \qquad \ldots \quad (4.1)$$
then ϕ is a function of a set of n random observations, $x_1, x_2, \ldots, x_n$ of a variate X, where we assume that the correlation coefficients between two different observations are all equal to ρ.

Let the upper and lower bounds of ϕ be U and L respectively, i.e.,
$$U = n/4(n-1), \quad \text{if } n \text{ is even}, \qquad \ldots \quad (4.2)$$
$$= (n+1)/4n, \quad \text{if } n \text{ is odd} \qquad \ldots \quad (4.3)$$
and
$$L = 1/2(n-1). \qquad \ldots \quad (4.4)$$
Then by a simple analysis, we have
$$1/L \geqslant E(R^2)/\sigma^2(1-\rho) \geqslant 1/U, \qquad \ldots \quad (4.5)$$
where we assume naturally the existence of the mean and the variance in the population. However, we need not assume the continuity of the distribution function.

References

Moriguti, S. (1954): Bounds for second moment of the sample range. *Rep. Stat. Appl. Res.*, **3**, 57-64.

Plackett, R. L. (1947): Limits of the ratio of mean range to standard deviation. *Biometrika*, **34**, 120-122.

Polya, G. (1950): On the harmonic mean of two numbers. *Amer. Math. Monthly*, **57**, 26-28.

Tsumura, Y. (1950): Quoted in *Statistical Tables*, Kawade Syoboo, Tokyo.

Paper received: March, 1954.
Revised: December, 1955.