

UNCLASSIFIED

AD 406 791

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA



UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

1.1 Introduction

Statistical variance has been widely used in propagation studies associated with radio direction finding for many years. With the advent at this laboratory of high speed digital computing machinery, it is advisable to explore methods of on and off line calculation of variance, with particular emphasis placed on on-line computations. It is not the purpose of this note to judge whether or not variance is the correct parameter to use, merely to find ways to calculating it.

It is planned to extend this discussion in future notes to cover means of estimating the confidence which can be placed in a given calculation, basing this estimate on the number of samples, the number of quantizing levels, the power spectrum of the sampled signal, and other relevant factors.

1.2 Variance Calculations

Since propagation problems are involved in most of the work done by this laboratory, it will be assumed that no a priori knowledge of the population distribution with which we are working will be available. This assumption leads immediately to the conclusion that no exact calculation of the population variance can be made. The only meaningful calculation which can be made is that of the sample variance. Possible methods of inferring the population variance from these computations will be treated in a future note.

In all this discussion, variance is used because the resulting equations appear in simpler form than if standard deviation were found. In practice, the standard deviation will be used widely as being more useful intuitively when judging results of an experiment. Since the standard deviation is defined as the square root of the variance, there is no essential difference between the two terms for the purposes of this discussion.

In the calculation of variance, the formulae

$$\sigma_s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.2.1)$$

$$\sigma_k^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.2.2)$$

are used widely. The first gives the sample variance in the strict statistical sense, but is a biased estimate of the variance, while the second is more empirical but results in an unbiased estimate of the true population variance. The expected values of the sample variances are given below, as found from the two expressions above.

$$E(\sigma_s^2) = \sigma^2 \left(1 - \frac{1}{n}\right) \quad (1.2.3)$$

$$E(\sigma_k^2) = \sigma^2 \quad (1.2.4)$$

Thus the estimate of the true population variance obtained from a sample variance is biased by the term $(1 - \frac{1}{n})$ in the first case. The error term is therefore $1/n$, n being the number of samples in the set. This indicates that if an estimate of the true population variance biased by no more than 1% is desired, at least 100 samples must be taken. Ordinarily this will not be too serious a problem, but in certain propagation studies it is desirable to operate with small samples to obtain intermediate results in the shortest possible time. Since use of small sample sizes may result in large amounts of bias being inserted in the variance estimates, we will use calculation methods based on the unbiased estimator in this note.

Consideration of Equation (1.2.2) shows the difficulty involved in trying to use this form directly in the situation in which the value n is not known in advance. The term $\bar{x}$ (sample mean) inside the parentheses requires a new

calculation of the sample mean and recalculation of all the differences as each new number is brought in. Long data runs would thus require a computer with a very large, fast store for computations on any but very short runs to be feasible. An expansion of Equation (1.2.2) will give the form

$$\sigma_s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right] \quad (1.2.5)$$

thus separating the x_i and $\bar{x}$ terms and allowing them to be calculated separately.

1.3 Computer Considerations in Variance Calculations

When programming any of these computation systems for a digital computer, the possibilities of number overflow and accuracy deterioration must be kept in mind. Note in Equation (1.2.5) that there is real danger of number overflow resulting from the accumulation of the squares of the input values. To limit the machine word length requirements for this type of calculation, let us rearrange Equation (1.2.5) into the following form:

$$\sigma_s^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \left(\frac{n}{n-1} \right) \bar{x}^2 \quad (1.3.1)$$

This may now be used in repetitive fashion to calculate a sample variance for a sample of unspecified size, the only requirement being that at least two numbers be read in before the expression is evaluated (this is to avoid division by zero).

Further benefit can be gained by substituting the expression for evaluating the sample mean in terms of the input values. The sample mean is given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.3.2)$$

and its square by:

$$\bar{x}^2 = \frac{1}{n^2} \left(\sum_{i=1}^n x_i \right)^2 \quad (1.3.3)$$

Substituting Equation (1.3.3) into Equation (1.3.1) we obtain

$$\sigma_s^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{1}{n(n-1)} \left(\sum_{i=1}^n x_i \right)^2 \quad (1.3.4)$$

This expression has the disadvantage that the sum of the inputs increases continuously. The recommended form of the equation for solution on a digital computer is:

$$\sigma_s^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 \quad (1.3.5)$$

This allows all numbers to be kept within a specified scaling with little difficulty.

To avoid the necessity of keeping all input numbers in store and forming new sums of squares at each reading time, let us find a method of calculating the current value of a sum of squares from the previous value plus the new element of the measurement set.

$$\text{Let} \quad S_n = \frac{1}{n-1} \sum_{i=1}^n x_i^2 \quad (1.3.6)$$

$$\text{and} \quad N = n + 1 \quad (1.3.7)$$

$$\text{then} \quad S_N = \frac{N-2}{N-1} S_n + \frac{x_N^2}{N-1} \quad (1.3.8)$$

In like manner:

$$\text{let} \quad T_n = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.3.9)$$

$$\text{then} \quad T_N = \frac{N-1}{N} T_n + \frac{x_N}{N} \quad (1.3.10)$$

All the forms that have been given for calculating the sample variance without knowing the sample size in advance require taking the difference of two numbers which may be quite large to find the variance, which may be small. It is necessary to examine the input data to determine its range, then scale this judiciously if a fixed point machine is being used (such as the Packard Bell PB250) to obtain the best possible results. As a trial estimate, let us base our scaling on being able to read a bearing to a tenth of a degree.

Allowing for a reading system which will read through north to eliminate this number discontinuity (bearings read in the range of 000.0° to 399.9° as with the Coleman digitizers now in use), all input numbers will be in the range of zero to 400.0 and their squares will be in the range of zero to 160,000. A variance of 1° would thus be represented by a difference of one part in 160,000. This should present little difficulty even on the PB250 used single precision, for its 21 bits used to hold the magnitude of a number can resolve to better than one part in two million. In a floating point machine such as the G-20, no difficulties will be encountered.

1.4 Typical Program Timing

Running time estimates have been made on programs to evaluate the variance of a time series as discussed above. Two programs were timed; one on the PB250 and one on the G-20. The PB250 is a serial machine with a word time of 12 microseconds for a 22 bit word and the G-20 is a parallel floating point machine with a storage cycle time of 6 microseconds and a multiply time of about 50 microseconds. In both programs it was assumed that the data input has already been done by some other part of the complete system operation, so that the calculation times given are strictly for the variance computation and do not include any sort of input/output processing. Average operation times obtained were 6 milliseconds on the PB250 and 700 microseconds on the G-20. In each case, the exact form of the mathematical expression for the variance was transformed as necessary to reduce scaling and significance problems, and each program calculated a new value of variance based on the previous value plus the new element of input data.