PERFORMANCE OF THREE AVERAGING METHODS, FOR VARIOUS DISTRIBUTIONS

by

Albert H. Nuttall
Naval Underwater Systems Center
New London, CT 06320 USA

ABSTRACT

The performance of three averaging methods, namely the sample median, the sample arithmetic mean, and the sample geometric mean, are analyzed in terms of their bias, variance, and mean square error. The bias and variance are numerically evaluated for various parent distributions and plotted versus the number, N, of data points employed in the sample statistics. Also, the limiting behaviors, as N increases without limit, are derived. It is found that the best averaging method is very dependent upon the distribution of the data, with the sample median being favored for data with occasional large out-liers.

INTRODUCTION

Estimation of average properties, such as the average power in a particular angular sector and/or frequency bin is often accomplished by taking N independent measurements of such data and calculating a simple arithmetic average. However, when the desired process is subject to random fade-outs or occasional large out-liers, this sample arithmetic mean (SAM) is severely perturbed, and alternative averaging methods should be considered. Two possible candidates are the sample median (SMD) and the sample geometric mean (SGM); these nonlinear processors of the available data have the potential of suppressing the deleterious effects mentioned above. Here we investigate* the performance of all three of these averaging methods in terms of the number, N, of independent data points employed in the pertinent average, and the parent distribution of the data. A wide variety of distributions are considered, some with parameters which allow for significantly different character and shapes of the governing probability functions.

---

*The basic analysis, derivations, and programs are given in Ref. 1.

DEFINITIONS

We have available N statistically-independent identically-distributed samples (random variables) $x_1, x_2, \ldots, x_N$ from some parent population with cumulative distribution function $P(u) = \text{Prob}\{x<u\}$ and probability density function, PDF, $p(u) = P'(u)$. The SAM of the available measurements is

$$a(N) = \frac{1}{N}\left(x_1 + x_2 + \ldots + x_N\right); \tag{1}$$

the SMD is

$$q(N) = \text{middle value of } \{x_1, x_2, \ldots, x_N\}, \text{ for N odd;} \tag{2}$$

and the SGM is (for non-negative random variables)

$$g(N) = \left(x_1\, x_2 \ldots x_N\right)^{1/N} = \exp\left(\frac{\ln(x_1) + \ldots + \ln(x_N)}{N}\right)$$

$$= B^{\left(\frac{1}{A}\,\frac{A\log_B(x_1) + \ldots + A\log_B(x_N)}{N}\right)} \qquad \text{for any base } B > 0 \text{ and scaling } A. \tag{3}$$

The last form in (3) for base B = 10 goes under the name of dB averaging.

As N tends to infinity, the sample quantities above tend to definite (non-random) limits. In particular, as $N \to \infty$,

$$a(N) \to \text{arithmetic mean} = \int du\; u\; p(u);$$

$$q(N) \to \text{median} = u_{1/2}, \text{ where } P\left(u_{1/2}\right) = \frac{1}{2};$$

$$g(N) \to \text{geometric mean} = \exp\left(\int du\, \ln(u)\, p(u)\right); \tag{4}$$

where we drop the prefix 'sample' for these deterministic quantities. The last result in (4) follows from the exponential form of the SGM in (3). If the 1/2 in the median definition is replaced by r, we have for q(N) the sample quantile of order r (Ref. 2, page 181).

The three limiting quantities in (4) will generally not be equal. For example, for an exponential parent PDF

$$p(u) = \frac{1}{m}\exp\left(-\frac{u}{m}\right) \text{ for } u > 0, \tag{5}$$

we have

$$\begin{aligned} &\text{arithmetic mean} = m; \\ &\text{median} = m \ln(2) = m\ .693; \\ &\text{geometric mean} = me^{-\gamma} = m\ .562. \end{aligned} \tag{6}$$

Thus we define the bias of each of the sample statistics (1)-(3) as the difference between their mean value and their asymptotic value:

$$\begin{aligned} \text{bias}\{\text{SAM}\} &= \overline{a(N)} - \text{arithmetic mean}; \\ \text{bias}\{\text{SMD}\} &= \overline{q(N)} - \text{median}; \\ \text{bias}\{\text{SGM}\} &= \overline{g(N)} - \text{geometric mean}. \end{aligned} \tag{7}$$

By virtue of this definition, all three biases will tend to zero as $N \to \infty$; that is, all three estimators, (1)-(3), are asymptotically unbiased, each with respect to its desired value as given by (4), respectively.

It is then convenient to define a normalized bias, NB, for each sample statistic as

$$NB(N) = N \frac{\text{bias}}{\sigma}, \tag{8}$$

where σ is the standard deviation of parent PDF p(u). The scale factor of N leads to a non-zero value of the normalized bias for large N, while the scale factor of σ is convenient in that it eliminates the dependence of the normalized bias on the absolute scale of the input data. For large N, (8) yields

$$\text{bias} \sim \sigma \frac{NB(\infty)}{N} \text{ as } N \to \infty; \tag{9}$$

thus $NB(\infty)$ is an important measure of quality of the particular sample statistic under consideration.

The variances of sample statistics (1)-(3) are defined as

$$\begin{aligned} \text{var}\{\text{SAM}\} &= \overline{a^2(N)} - \overline{a(N)}^2; \\ \text{var}\{\text{SMD}\} &= \overline{q^2(N)} - \overline{q(N)}^2; \\ \text{var}\{\text{SGM}\} &= \overline{g^2(N)} - \overline{g(N)}^2. \end{aligned} \tag{10}$$

Again, since these quantities tend to zero for large N, it is more convenient to define a normalized variance, NV, as

$$NV(N) = N \frac{\text{variance}}{\sigma^2} \quad . \tag{11}$$

Then we can state that

$$\text{variance} \sim \sigma^2 \frac{NV(\infty)}{N} \text{ as } N \rightarrow \infty ; \tag{12}$$

thus $NV(\infty)$ is also an important measure of the quality of a particular sample statistic.

We present results here for NB(N) and NV(N), along with their asymptotic values at $N = \infty$, for a variety of parent distributions P(u). Additional results for the sample quantile with r = .75 and .9, and for the PDF, cumulative distribution function, characteristic function, cumulants, and moments of the various sample statistics are available in Ref. 1.

RESULTS

The first case we consider is the Gaussian PDF with arithmetic mean m and variance σ^2. Since this random variable can go negative, the SGM is undefined. The SAM and SMD are unbiased for all N; thus NB(N) = 0 for all N, for this example. Results for the normalized variance are presented in Fig. 1 for the number of samples, N, between 1 and 51, for both the SAM and the SMD. The normalized variance for the SMD is computed only at odd values of N, indicated by an X, and straight lines drawn between these points for ease of association of values. It is seen that the variance of the SMD is always greater than that for the SAM, the limiting value, $NV(\infty)$, being $\pi/2$ for the SMD; see also Ref. 2, page 369. Observe that the parameters m and σ of this PDF have dropped out of this normalized plot.

It is worth pointing out here, and for similar results to follow, that although the curve for the SMD increases with N, that does not mean that the variance increases with N; rather, the normalizing factor of N in definition (11) causes this behavior. The actual (unnormalized) variance decreases monotonically with N, eventually being of order 1/N.

For a Rayleigh random variable, the results for the normalized bias are given in Fig. 2. The SAM is unbiased for all N, whereas the SGM and SMD are, of course, only asymptotically unbiased. The limiting values, $NB(\infty)$, for both of these latter sample statistics are given by analytically complicated expressions and are not repeated here, for the sake of brevity; they are indicated numerically by horizontal lines at the right edge of the figure. The corresponding results for the normalized variance are given in Fig. 3. They indicate that whereas the SGM has about the same stability as the SAM, the variance for the SMD is about 65% greater.

For an exponential PDF (as given by (5)), the normalized bias and variance results are presented in Figs. 4 and 5 respectively. The biases of the SGM and SMD are comparable, but we observe that the variance for the SGM is twice as small as that for the SAM and the SMD.

For a log-normal PDF,

$$p(u) = \frac{1}{\sigma_y \, u \sqrt{2\pi}} \exp\left[- \frac{\left(\ln(u) - m_y\right)^2}{2\sigma_y^2} \right] \quad \text{for } u > 0 \, , \tag{13}$$

the NB(N) and NV(N) results are independent of location parameter m_y, but they do depend on spread factor σ_y. This may be anticipated by plotting the PDF (13) for various values of σ_y and observing that the shape changes as σ_y does. Since NB(N) and NV(N) depend upon the shape of the PDF (rather than upon absolute location and scale), results will depend on the particular value of σ_y selected. An example of NB(N) and NV(N) for $\sigma_y = 1$ is presented in Figs. 6 and 7. Now we observe that the variance of the SMD is 3 times better, and that of the SGM 4.6 times better, than for the SAM, at least for larger values of N. However, as $\sigma_y \to 0$, the log-normal PDF in (13) approaches a Gaussian PDF about the point $u = \exp(m_y)$, and the behaviors would revert back to Fig. 1 then.

The next example is the Rice PDF; physically, this corresponds to the squared-envelope of the sum* of a sine wave and a centered narrowband Gaussian noise process. That is, the PDF is

$$p(u) = \frac{1}{2\sigma_i^2} \exp\left(- \frac{u + A^2}{2\sigma_i^2} \right) I_0\left(\frac{A\sqrt{u}}{\sigma_i^2} \right) \text{for } u > 0 \quad , \tag{14}$$

where A is the sine wave amplitude and σ_i is the noise standard deviation. Once again, the shape of the PDF depends on a parameter, namely A/σ_i. Results for $A/\sigma_i = 1$ are given in Figs. 8 and 9. The SMD has 14% greater variance than the SAM, but the SGM has about 60% of the SAM variance. As $A/\sigma_i \to 0$, the exponential PDF results are obtained, whereas as $A/\sigma_i \to \infty$, the Gaussian case is realized. Thus (14) represents a transition case between these extremes.

The last example we consider here is an exponential PDF with out-liers. That is, each sample or measurement $\{x_k\}$ in (1)-(3) is given by

$$x_k = x_a + x_b, \tag{15}$$

where x_a has an exponential PDF,

$$p_a(u) = \frac{1}{m_a} \exp\left(- \frac{u}{m_a} \right) \quad \text{for } u > 0, \tag{16}$$

---

*If each of the observed random variables $x_1, x_2, \ldots, x_N$ in (1)-(3) is obtained by first summing up the envelope-squared outputs of M narrowband filters, as for example in diversity reception, the PDF in (14) is replaced by the Q_M distribution. Results for this case are available in Ref. 1, but are not given here, for sake of brevity.

and disturbance x_b is a random variable which is zero most of the time, but occasionally takes on a large value (out-lier) L. That is, its PDF is

$$p_b(u) = (1 - Q)\,\delta(u) + Q\,\delta(u - L), \tag{17}$$

where Q is the probability of an out-lier. Then the parent PDF of observation (sample) x_k is the convolution of (16) and (17):

$$p(u) = \frac{1 - Q}{m_a} \exp\left(-\frac{u}{m_a}\right) U(u) + \frac{Q}{m_a} \exp\left(-\frac{u - L}{m_a}\right) U(u - L), \tag{18}$$

where unit step

$$U(t) = \left\{\begin{array}{l} 1 \text{ for } t > 0 \\ 0 \text{ for } t < 0 \end{array}\right\}. \tag{19}$$

The important parameter now is L/m_a, which obviously affects the shape of PDF (18).

Now however, before we get into the detailed bias and variance results, another consideration is of paramount importance. Our sample statistics, (1)-(3), will no longer extract (estimate) the arithmetic mean, median, and geometric mean, respectively, of the (disturbance-free) exponential PDF (16), but perforce, the corresponding statistics of the measurement PDF (18). If, however, we are really interested in the parameters of (16), then we must inquire into the quantitative disturbance caused by the out-liers described in (17). Here we merely cite the results for one numerical case; additional results are given in Ref. 1.

For probability Q = .05, and out-lier value $L/m_a = 6$, we find that the ratio of arithmetic means, for (18) with respect to (16), is 1.3. The corresponding ratio of medians is only 1.08, whereas the ratio of geometric means is 1.13. Thus the SMD and SGM are more resistant to the presence of infrequent out-liers, insofar as their effects on the particular parameters of median and geometric mean.

The results for the normalized bias and variance are given in Figs. 10 and 11. The variances of the SMD and SGM are again smaller than that for the SAM. The bias of the SMD and SGM are comparable.

If we define a mean-square error as the average value of the squared difference between a sample statistic s and a desired parameter d_a of the disturbance-free PDF, we can develop it as follows:

$$MSE = \overline{(s - d_a)^2} = \overline{(s - \bar{s} + \bar{s} - d_a)^2}$$

$$= \overline{(s - \bar{s})^2} + (\bar{s} - d_a)^2 \,. \tag{20}$$

But the first term is the variance of the sample statistic, and the second term can be expressed as

$$\bar{s} - d_a = (\bar{s} - d) + (d - d_a) = \text{bias} + \text{deflection in desired parameter}, \qquad (21)$$

where d is the modified value of the desired parameter d_a, due to the disturbance. Thus

$$\text{MSE} = \text{variance} + (\text{bias} + \text{deflection})^2 . \qquad (22)$$

Now the bias and variance are $O(N^{-1})$ for large N, whereas the deflection of the desired parameter does not decay with N at all; in fact, it is independent of N. Thus the considerations above, whether for the ratio of arithmetic means or medians or geometric means, are very important, since they dominate the magnitude of the mean-square error for very many samples available.

DISCUSSION

The ability of the SMD and SGM to suppress deleterious effects due to occasional large interferences is very pronounced for some probability density functions. Not only is the deflection of the desired parameter (arithmetic mean or median or geometric mean) decreased, but the bias and variance of the estimate can be markedly reduced in some cases. The exact amounts depend on the magnitude and frequency of the interference.

Another possible approach to alleviate the effects of additive large out-liers is to subject the available samples $x_1, x_2, \ldots, x_N$ to a nonlinear transformation such as saturation, in order to suppress the large contributions, prior to evaluating the SAM or SGM or SMD. Knowledge of relative levels (such as L/m_a for the above example) would be required for optimal adjustment of the saturation level, but performance could be markedly improved. The nonlinear transformation would reduce the deflection, while the averaging of N samples would reduce the bias and variance. This possibility has not yet been pursued.

REFERENCES

1. A. H. Nuttall, "Statistics of Sample Median, Quantile, Geometric Mean, and Arithmetic Mean, for Various Distributions," Technical Report 6689, Naval Underwater Systems Center, New London, CT USA.

2. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J., 1961.

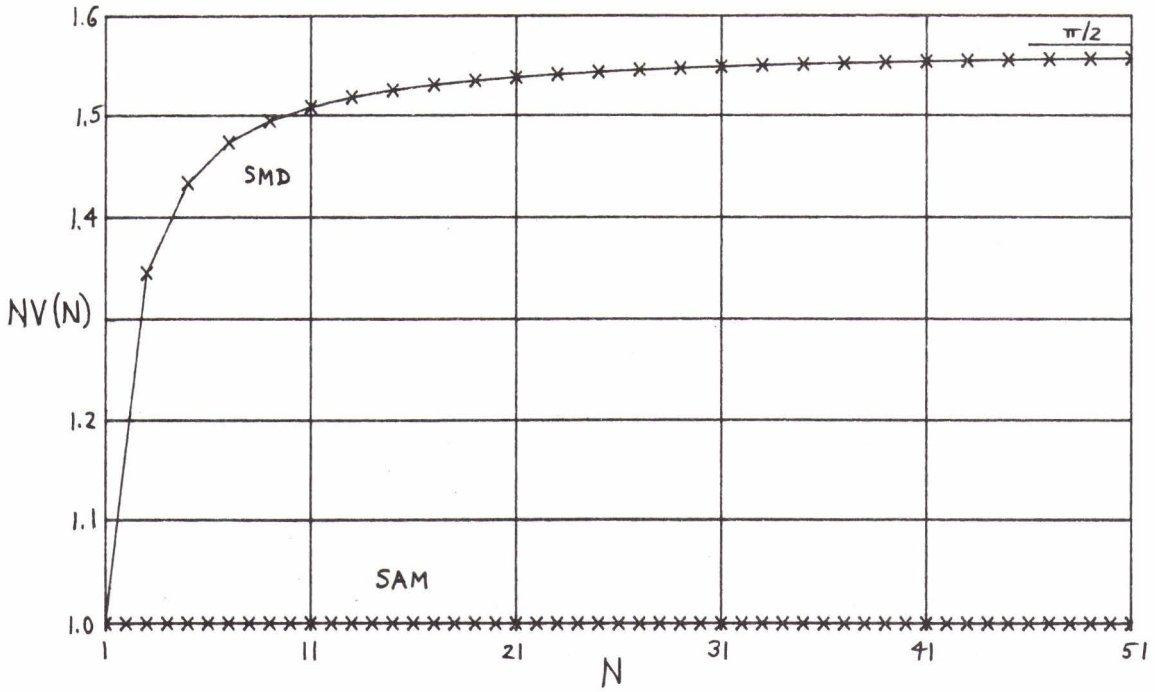

FIG. 1 NORMALIZED VARIANCE; GAUSSIAN PDF

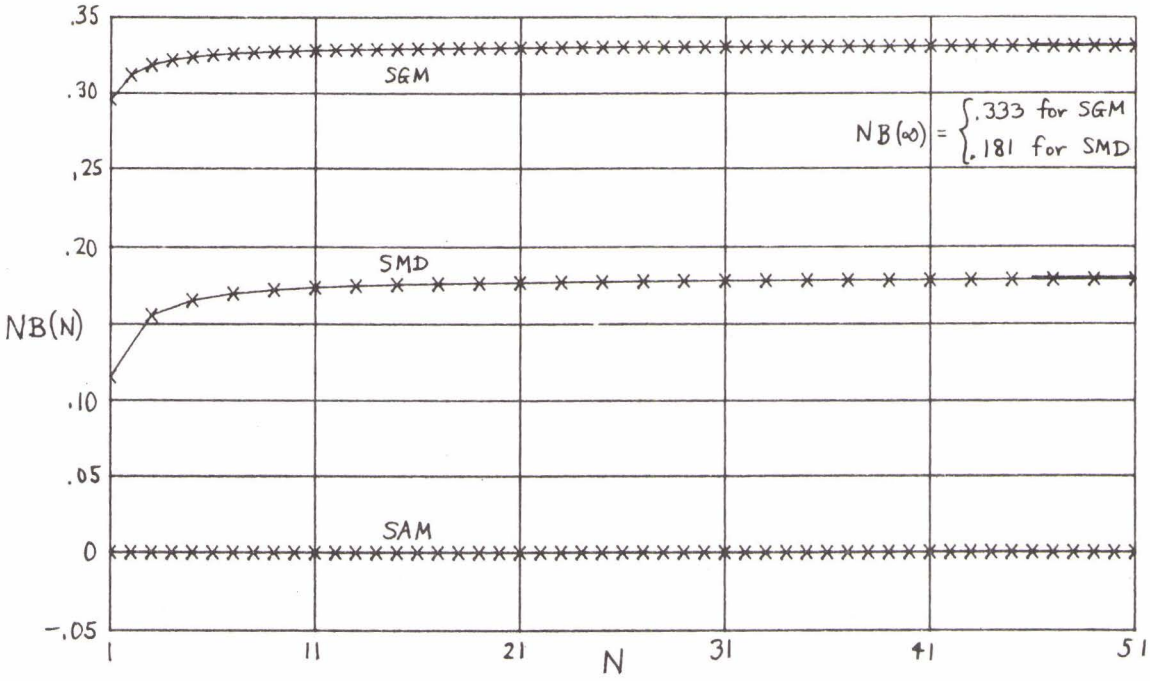

FIG. 2 NORMALIZED BIAS; RAYLEIGH PDF

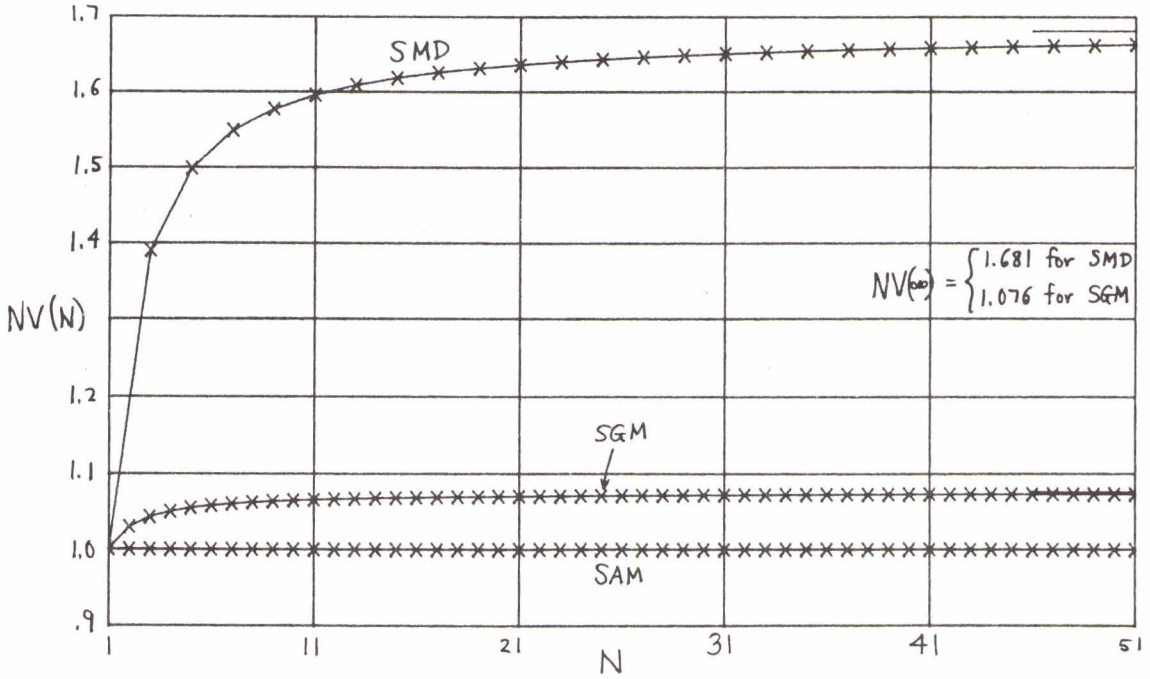

FIG. 3 NORMALIZED VARIANCE; RAYLEIGH PDF

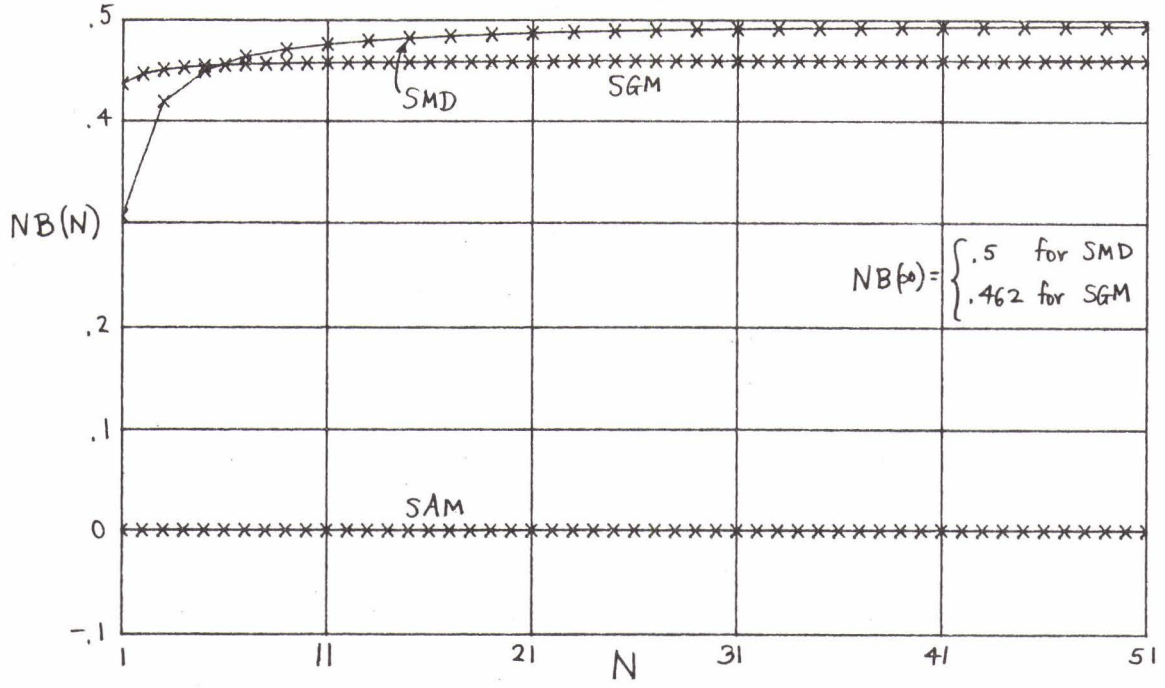

FIG. 4 NORMALIZED BIAS; EXPONENTIAL PDF

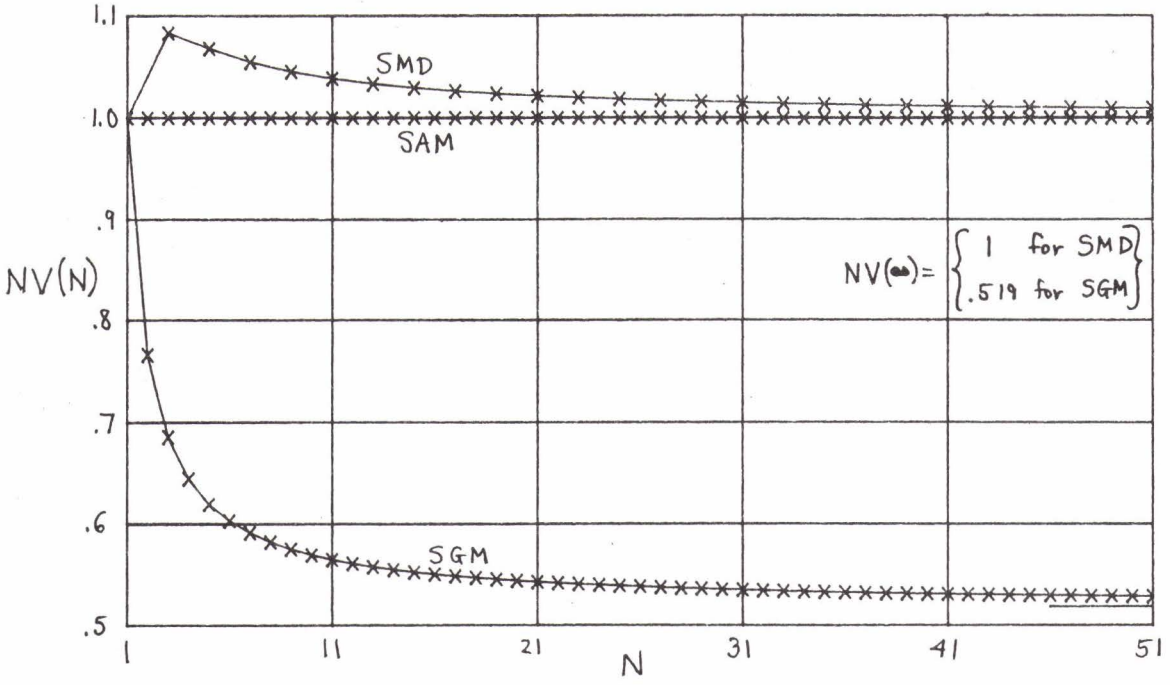

FIG. 5 NORMALIZED VARIANCE; EXPONENTIAL PDF

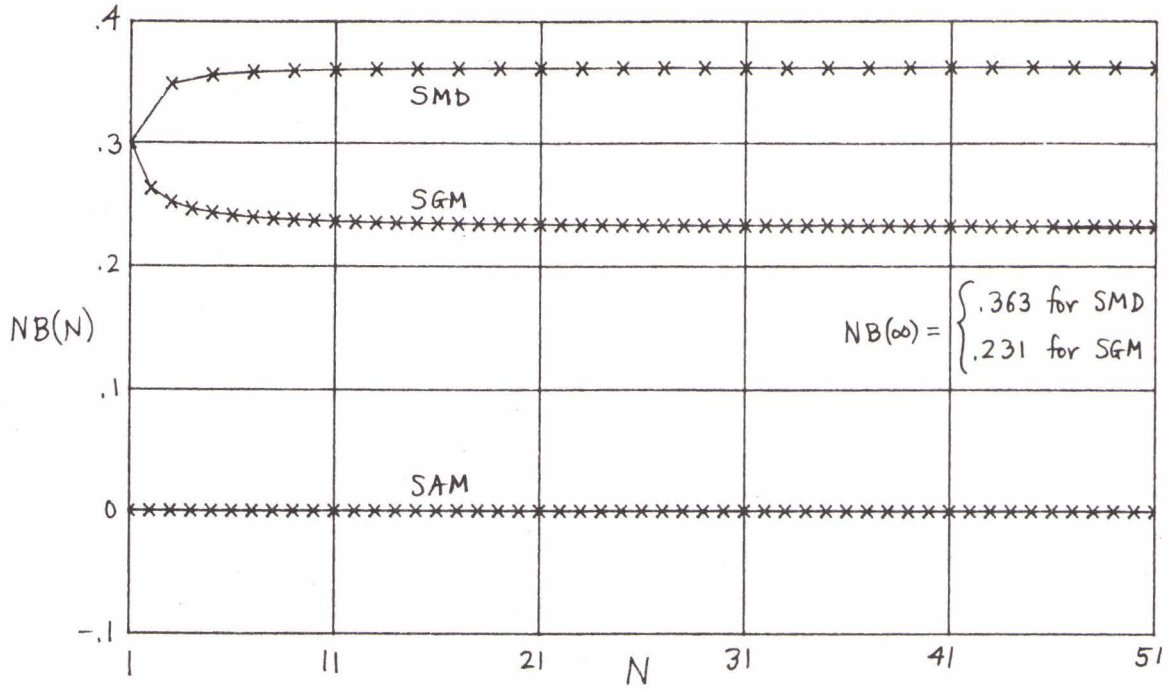

FIG. 6 NORMALIZED BIAS; LOG-NORMAL PDF (σ_y = 1)

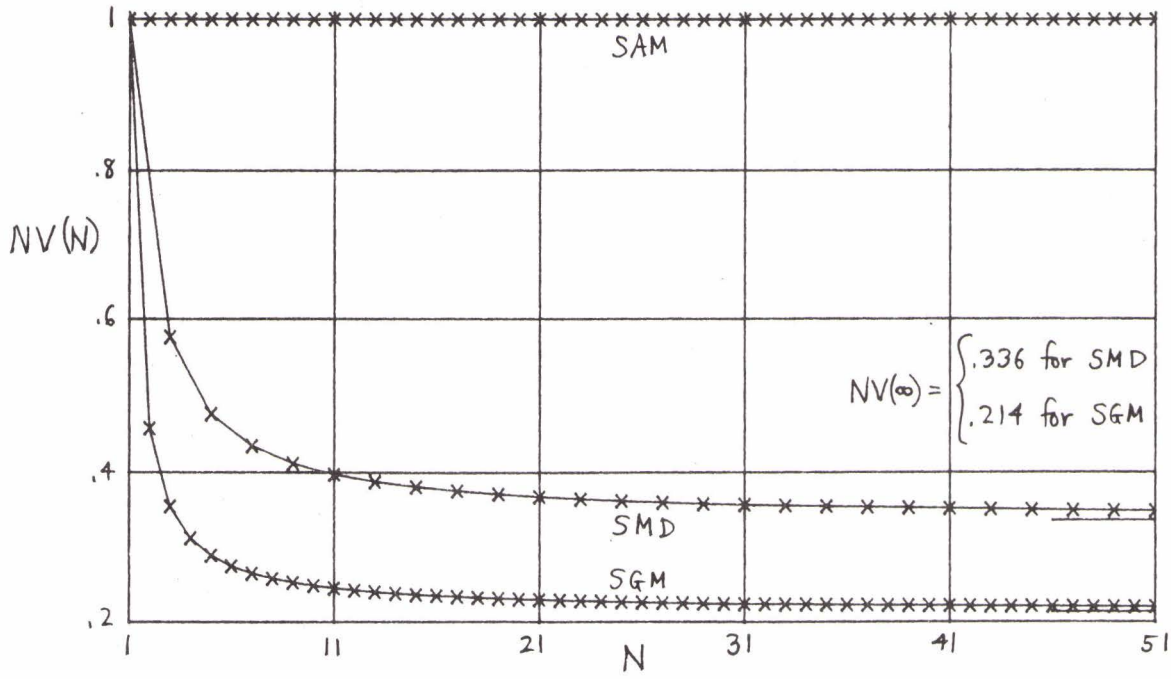

FIG. 7 NORMALIZED VARIANCE; LOG-NORMAL PDF (σ_y = 1)

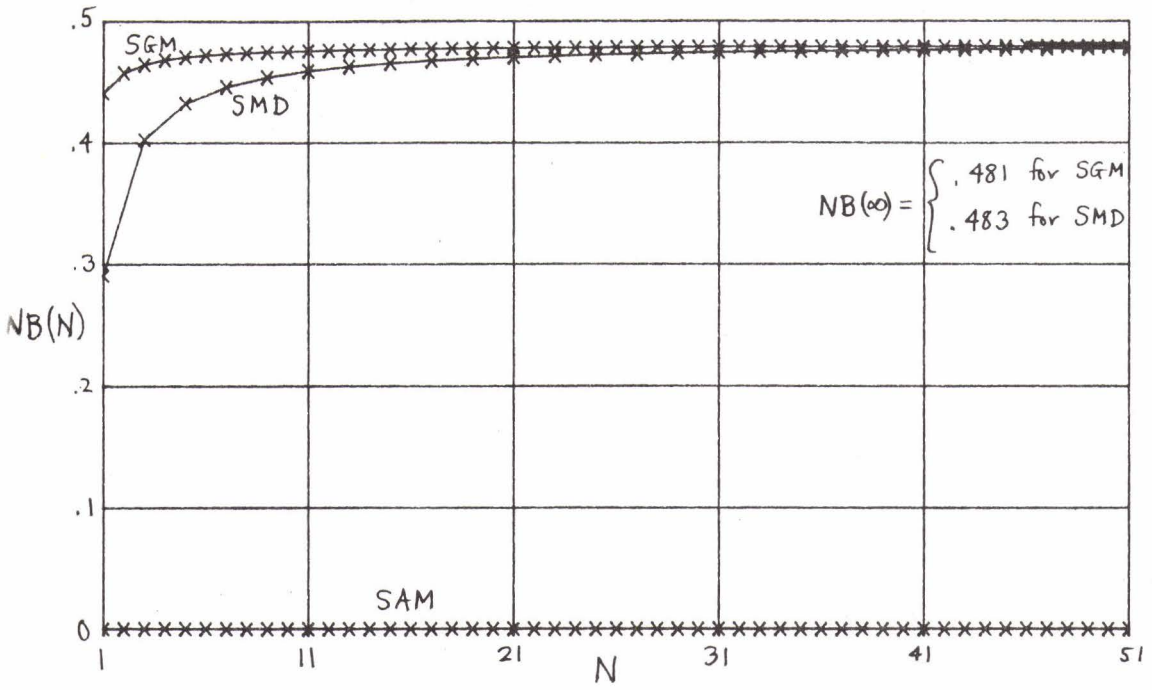

FIG. 8 NORMALIZED BIAS; RICE PDF ($A/\sigma_1 = 1$)

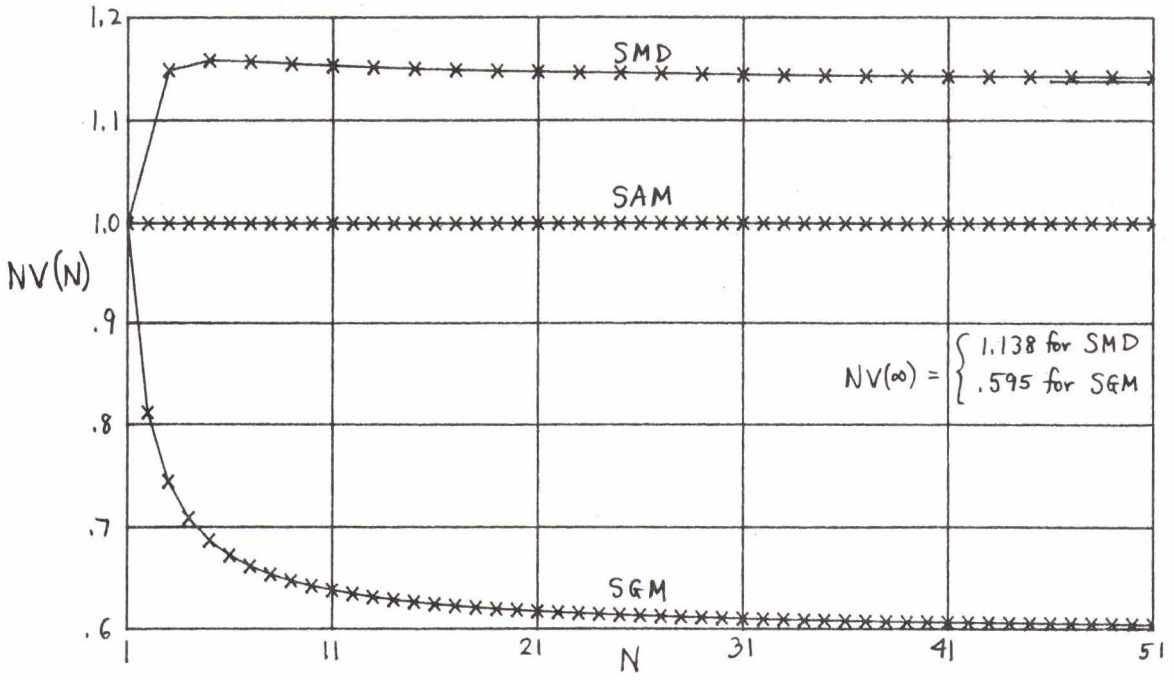

FIG. 9 NORMALIZED VARIANCE; RICE PDF ($A/\sigma_1 = 1$)

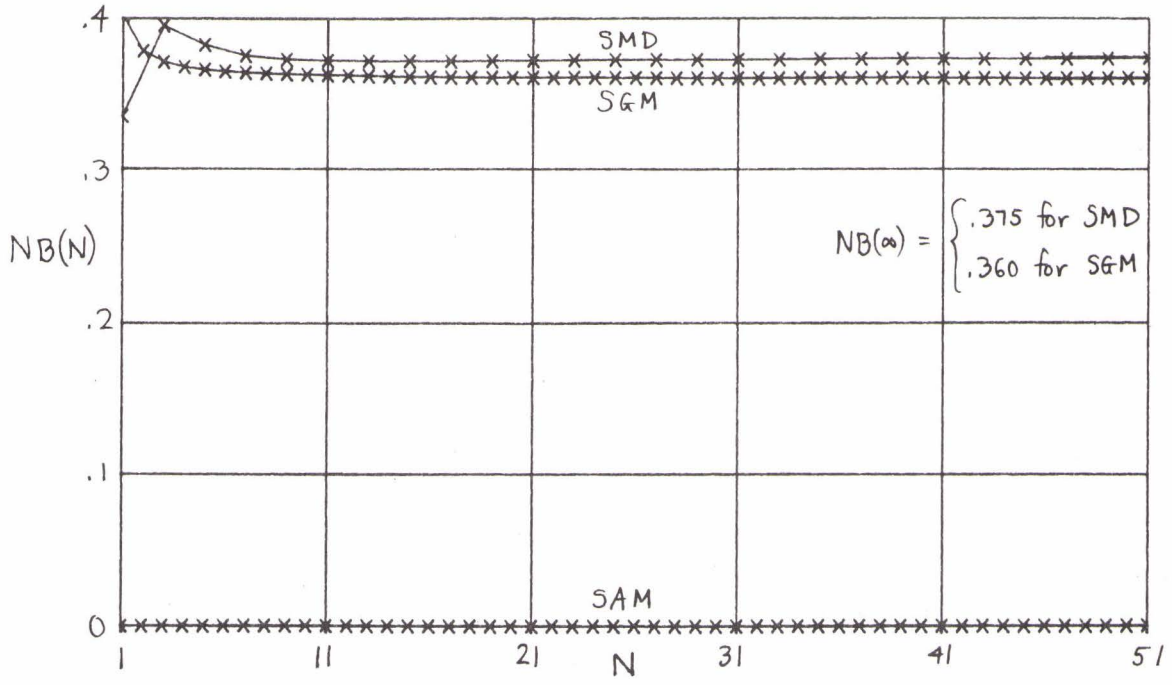

FIG. 10 NORMALIZED BIAS; EXPONENTIAL PDF WITH OUT-LIERS ($\frac{L}{m_a} = 6$, $Q = .05$)

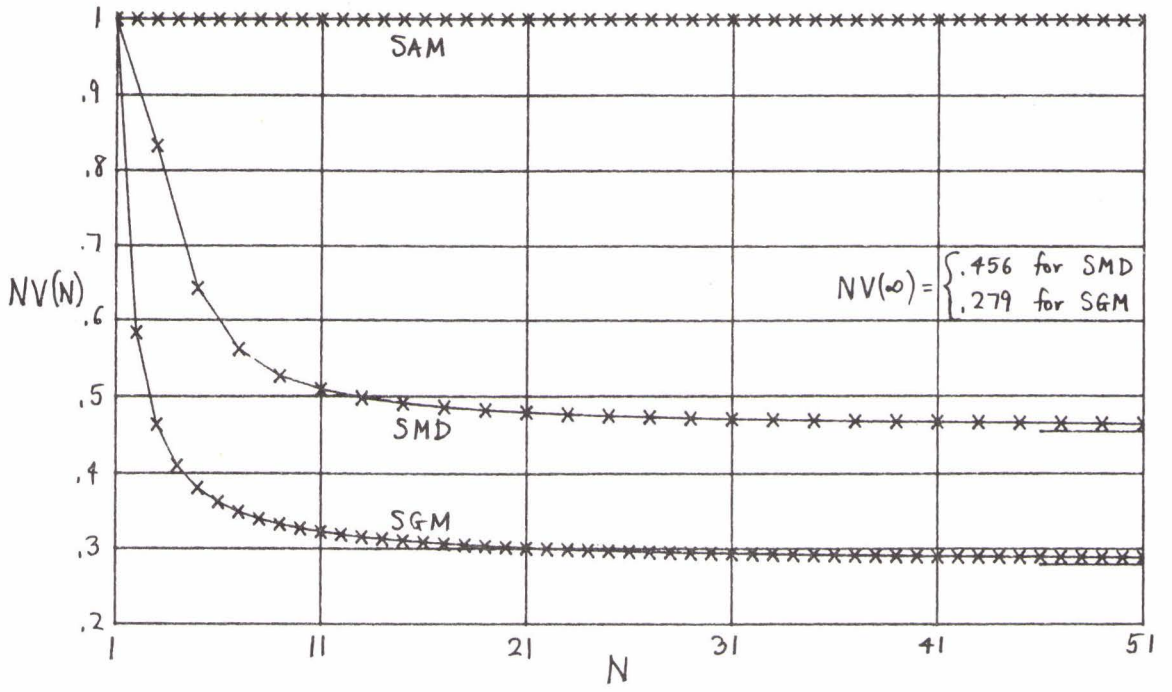

FIG. 11 NORMALIZED VARIANCE; EXPONENTIAL PDF WITH OUT-LIERS ($\frac{L}{m_a} = 6$, $Q = .05$)