PORTIONS OF THIS DOCUMENT ARE ILLEGIBLE

LA-UR -

TITLE: ESTIMATION OF EXPECTED VALUE FOR LOGNORMAL AND GAMMA DISTRIBUTIONS

AUTHOR(S): Gary C. White

SUBMITTED TO: Proceedings of Plutonium Information Conference, Nevada Applied Ecology Group, San Diego, CA, February 28 - March 2, 1978.

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87544

An Affirmative Action/Equal Opportunity Employer

Form No. 836
St. No. 2629
1/75

UNITED STATES
ENERGY RESEARCH AND
DEVELOPMENT ADMINISTRATION
CONTRACT W-7405 ENG. 36

ESTIMATION OF EXPECTED VALUE

FOR LOGNORMAL AND GAMMA DISTRIBUTIONS

Gary C. White

Environmental Studies Group
Los Alamos Scientific Laboratory
Los Alamos, NM 87545

ABSTRACT

Concentrations of environmental pollutants tend to follow positively skewed frequency distributions. Two such density functions are the gamma and lognormal. Minimum variance unbiased estimators of the expected value for both densities are available. The small sample statistical properties of each of these estimators were compared for its own distribution, as well as the other distribution to check the robustness of the estimator. Results indicated that the arithmetic mean provides an unbiased estimator when the underlying density function of the sample is either lognormal or gamma, and that the achieved coverage of the confidence interval is greater than 75 percent for coefficients of variation less than two. Further Monte Carlo simulations were conducted to study the robustness of the above estimators by simulating a lognormal or gamma distribution with the expected value of a particular observation selected from a uniform distribution before the lognormal or gamma observation is generated. Again, the arithmetic mean provides an unbiased estimate of expected value, and the achieved coverage of the confidence interval is greater than 75 percent for coefficients of variation less than two.

INTRODUCTION

The concentrations of environmental pollutants have been suggested to follow positively skewed frequency distributions by numerous researchers. In particular, Pinder and Smith (1975) investigated the goodness of fit of the lognormal, Weibull, exponential and normal distributions to radiocesium concentrations in soil and biota. They found that the lognormal distribution fit the majority of the data sets. Giesly and Weiner (1977)

found that the lognormal also tended to fit the concentrations of trace metals in fish better than the Weibull, exponential, or normal distributions. Ellett and Brownell (1964) suggest the gamma distribution may be preferred to the lognormal distribution. Eberhardt and Gilbert (1975) made an extensive study of how to distinguish these two distributions, and concluded that this is difficult for less than 200 observations. Extensive Monte Carlo simulations were done to reach this conclusion. Forsythe *et al.* (1973) compared the fit of the gamma and lognormal distributions for the concentration of DDT in earthworms, and concluded that both fit the data equally well. Figure 1 shows the similarity of the lognormal and gamma probability density functions for a variety of coefficients of variation and expected value equal 1.

Given that both these distributions appear to explain contaminant data equally well, I want to explore the implication of selecting one of these two distributions in estimating the expected value of concentration. Some investigators (Eberhardt and Gilbert 1973) have suggested using the median of the observed data to measure central tendecy when a portion of the samples are below detection limits. I believe that the median may be quite useful for answering some questions, but that usually the expected value is the desired measure. This paper presents the results of Monte Carlo simulations studies of estimating the expected value (EX) for these two distributions.

Link and Koch (1975) explored the bias which may result when the lognormal estimator of expected value is used for distributions other than lognormal. They found that a large negative bias (up to 97%) may result when the distribution of the logarithmically transformed variable is heavier tailed than the normal distribution. However, no bias was found when the logarithmically transformed variable has less tail area than the normal distribution. They did not consider lognormal estimation with gamma distributed data.

ESTIMATORS

First the estimation of EX for the lognormal distribution will be considered. The density function is given by Aitchison and Brown (1976)

$$f(x) = \frac{1}{\sigma_y \times \sqrt{2\pi}} \exp\left[- \frac{1}{2\sigma_y^2} (\ln x - \mu_y)^2 \right] dx$$

$$(x > 0;\ \sigma_y > 0,\ -\infty < \mu_y < \infty) ,$$

where μ_y and σ_y^2 are the mean and variance of $y = \ln x$, respectively. Finney (1941) derived a minimum variance unbiased estimator (MVUE) for EX because of the large bias of the maximum likelihood estimator (MLE). Finney's estimator is

$$\hat{E}(x) = \exp(\bar{y})\ g_n(s_y^2/2)$$

where $\bar{y}$ is the arithmetic mean of the log transformed x values, s_y^2 is the variance of the log transformed x values, and the function g is the infinite series

$$g_n(t) = 1 + \frac{(n-1)t}{n} + \frac{(n-1)^3 t^2}{n^2(n+1)2!} + \frac{(n-1)^5 t^3}{n^3(n+1)(n+3)3!} + \ldots$$

A finite sample confidence interval can be estimated for $\hat{E}(x)$ by Cox's direct method (Land, 1972). A confidence interval is calculated for the value $\bar{y} + 1/2\ s_y^2$ and then antiloged to achieve an interval on E(x). This interval is asymmetric, but has the desirable property that the lower confidence bound cannot be negative. Cox's direct method was shown to be easily the best of the approximate methods considered by Land and was recommended by him when dealing with large sample sizes and moderate values of s_y^2. Land's exact method is not used because of the computational difficulties involved.

Estimation of the EX for the gamma distribution is much simpler than for the lognormal distribution. The EX of the gamma distribution is the product of the parameters α and β, where the probability density function is

$$f_x(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \exp(-x/\beta)\ x^{\alpha-1}\ dx \quad (x > 0;\ \alpha > 0,\ \beta > 0) \quad .$$

The maximum likelihood equations to estimate α and β are (Choi and Wette, 1969)

$$\Psi(\hat{\alpha}) + \log \hat{\beta} = \bar{y} \quad , \text{ and } \hat{\alpha}\,\hat{\beta} = \bar{x}$$

where $\Psi(t)$ is the psi (digamma) function. Hence we see by the invariance property of MLE's that $\bar{x}$ is the MLE of EX for the gamma distribution. In addition, $\bar{x}$ is also an MVUE of EX. This result is known because 1) the gamma distribution is a member of the exponential family, 2) the set of minimal sufficient statistics (MSS) is

$$\left\{ \sum_{i=1}^{n} x_i \ , \ \sum_{i=1}^{n} \log x_i \right\} \ ,$$

3) this set of MSS is complete, and 4) any function of the MSS is a MVUE if the function is unbiased. To see that $\bar{x}$ is unbiased,

$$E(\bar{X}) = E(X_1 + X_2 + \ldots + X_n)/n$$

$$= [E(x_1) + E(x_2) + \ldots + E(x_n)]/n$$

$$= (n \mu_x)/n$$

$$= \mu_x$$

Hence $\bar{X}$ is a MVUE of EX. The details of this proof can be obtained in Mood *et al.* (1974). The result that $\bar{x}$ is an MVUE is particularly fortuitous as the arithmetic mean of the sample has often been used to estimate EX for real data. The usual confidence intervals for $\bar{x}$, namely

$$\bar{x} \pm t_{(n-1)} \ s_x/\sqrt{n} \ ,$$

will be used, with the Central Limit Theorem and the asymptotic normality of a MLE to justify the assumption of normality. Because of this assumption, this confidence interval may perform poorly for small sample sizes. The variance estimate thus obtained is not the same as the variance of $\bar{x}$ calculated by the maximum likelihood estimation procedure. However, the calculations are much easier to perform, and this estimator is the one commonly used in the transuranic literature. Therefore, of interest is whether confidence intervals based on this simple variance estimator are valid. Of particular concern is the validity of this approach for small sample sizes, say n = 5.

ROBUSTNESS

Both estimators described above are known to have optimal properties when used with data derived from their respective distributions. In addition the performance of each estimator when applied to other distribution functions is of interest, i.e., how robust the estimator may be.

Neither the gamma or lognormal distribution can be mixed with another gamma or lognormal distribution and the result still be gamma or lognormally distributed. Formally, let $f_1(x;\ \theta_2)$ and $f_2(x;\ \theta_2)$ be either gamma or lognormal probability density functions with parameter vectors θ_1 and θ_2 respectively. Then assume we sample from a population with probability p that the variate is distributed as $f_1(x;\ \theta_1)$. The resulting variate is distributed as

$$p f_1(x;\ \theta_1) + (1-p)\ f_2(x;\ \theta_2) \quad .$$

The concept of mixing two distributions can be extended farther. Suppose that the $EX = \mu_x$ of $f(x)$ is actually drawn from a second density, $g(\mu_x)$. Then the distribution of x is

$$f_{x|\mu_x}(x|\mu_x) = \frac{f_x(x,\mu_x)}{g(\mu_x)}$$

or $$f_x(x,\mu_x) = f_{x|\mu_x}(x|\mu_x)\ g(\mu_x) \quad .$$

$f_x(x,\ \mu_x)$ is a family of distributions indexed by the parameter μ_x (see Mood *et al.*, 1974:122-124).

This result can be applied to transuranic research by conceptualizing the distribution of radioactivity in a fallout pattern. Suppose we stratify the fallout area into n strata, each with mean concentration EX_i, i=1, 2, ..., n. If a random sample of 1-m^2 quadrats are taken from strata i, the expected concentration will be EX_i. However, quadrats closer to ground zero would be expected to have slightly larger concentrations on the average than quadrats farther away from ground zero. However, this process is stochastic so one method of expressing this randomness is to assume the expected value of a quadrat is actually drawn from some distribution, $g(\mu_x)$.

To simulate this process, $g(\mu_x)$ was assumed to be a uniform distribution with density function

$$g(\mu_x) = \frac{1}{b-a}\ d\mu_x$$

Thus the expected value of an observation was first drawn from a uniform distribution, and then a variate generated from a gamma or lognormal distribution with this expected value. The expected value of the resulting

distribution must be evaluated to show that indeed the expected value is $(a + b)/2$:

$$EX = E[E(X|\mu_x)]$$

$$= E(\mu_x)$$

$$= (a + b)/2$$

Comparisons of the lognormal density function and the compound uniform-lognormal density function are made in Fig. 2. Comparisons of the gamma density function and the compound uniform-gamma density function are made in Fig. 3.

MONTE CARLO SIMULATIONS

Random normal deviates were generated by the method suggested by Bell (1968) and then transformed to a lognormal deviate by $x = \exp(y)$. Random gamma deviates with nonintegral shape parameter were generated with the method presented by Fishman (1973). Briefly the method involves summing k (= greatest integer of α) exponential variates, $\mathcal{E}(1)$, adding to this sum a product of a beta variate distributed as $\beta e(\alpha - k, 1 - \alpha + k)$ and an exponential $\mathcal{E}(1)$, and multiplying the total by the parameter β.

Samples of size n = 5, 10, 20, 30, 50, and 100 were drawn from each of the lognormal and gamma distributions. All possible combinations of EX = 1, 5, 10, 50, and 100 and coefficient of variation of c = 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, and 2.0 were used for both distributions. These combinations give a total of 210 cases per distribution. Each case was replicated 1000 times to estimate the bias and achieved coverage (proportion of replicates in which the constructed 95% confidence interval contained the true parameter value) for the two estimators discussed above. In addition, the average length of the confidence interval for $\hat{E}(x)$ was calculated for each estimator.

Parameters were calculated from EX and c for the lognormal distribution as:

$$\sigma_y^2 = \ln (c^2+1)$$

$$\mu_y = 1/2 \ln [(EX)^2/(c^2+1)] \quad .$$

Parameters were calculated from EX and c for the gamma distribution as:

$$\alpha = 1/c^2$$

$$\beta = EX\ c^2$$

In simulations where EX was selected from a uniform distribution, a and b are defined to be ±50% of the desired expected value of the distribution. For example, suppose EX is to be 10.0. Then a = (1-0.5)10 and b = (1+0.5)10, or EX is selected from the interval (5, 15). The parameter values for α, β, μ_y, and σ_y^2 were then calculated with the formulas given above to generate one realization of x.

The infinite series, $g_n(t)$, necessary to calculate $\hat{E}(x)$ assuming lognormality was evaluated to a point where the ratio of an additional term to the summation was less than 1E-7. Values of the t-statistic were obtained from tabled values.

RESULTS AND DISCUSSION

In general, the expected value of the distributions had no effect on the results. Rather, the coefficient of variation tended to explain the observed phenomena. Hence in the following sections, the statements made will apply to the range of expected values simulated. A complete listing of the simulation results is given in White (In Prep.).

Gamma Estimator with Gamma Distributed Data

The arithmetic mean is an unbiased estimator for EX, and so the simulations showed. Of course, individual point estimates may vary widely. Hence the main purpose of simulating this estimator was to check the achieved coverage against the predicted value. A gradual decline in achieved coverage was noted with an increase in the coefficient of variation (Fig. 4). However for all cases simulated, the achieved coverage is greater than 70%. A slight decline in achieved coverage is noted also for decreasing sample sizes. This trend is more apparent for c = 2 than any other case.

Lognormal Estimator for Lognormally Distributed Data

Because this estimator is MVUE for lognormally distributed data, the chief purpose for the simulation was to check the coverage of the confidence interval. The achieved coverage is always close to the predicted 95%

for n = 100. However, the achieved coverage declines with decreasing sample sizes (Fig. 5). The minimum achieved coverage is greater than 80% in all cases. These results are consistent with the findings of Land (1972).

Lognormal Estimator for Gamma Distributed Data

A major finding is that the bias of the lognormal estimator becomes large for gamma distributed data as α becomes small or c becomes large. In particular, a relative bias $(100[Ave(\hat{E}(x)) - EX]/EX)$ of about 25% is present for $\alpha = 1$ (Fig. 6). The case $\alpha = 1$ corresponds to the exponential distribution. The relative bias of the lognormal estimator becomes much worse for $\alpha < 1$. This result would be expected because the shapes of the two distributions differ greatly for $\alpha \leq 1$. However, even for c = 0.75 ($\alpha = 1.78$), the relative bias of the lognormal estimator for gamma distributed data is about 6%. Also the achieved coverage of the confidence interval begins to decrease for c = 1.0 and n = 100 (Fig. 6). Coverage becomes very poor for $c > 1$.

Gamma Estimator for Lognormally Distributed Data

In contrast to the lognormal estimator for gamma distributed data, the gamma estimator for lognormal data does quite well. This estimator is unbiased, and so the simulations showed. Also, the achieved coverage of this estimabor is good for small sample sizes (n = 5) (Fig. 7) whereas the coverage of the lognormal estimator is usually significantly less than the predicted 95% for n = 5. However the average confidence interval width is usually greater for the arithmetic mean. The achieved coverage of the arithmetic mean drops as c increases, but never below 75%, even for n = 5 and c = 2. Also for c = 0.75, the average confidence interval length becomes about the same for the two estimators.

Robustness

The same general conclusions discussed in the preceding four sections also hold when the lognormal and gamma estimators are applied to a compound lognormal or gamma distribution with the expected value selected from a uniform distribution. The arithmetic mean still provides an unbiased estimate of EX in all cases, while the lognormal estimator provides an unbiased estimate when the variate is uniform-lognormal distributed, but not for the case when the variate is uniform-gamma distributed.

Confidence interval coverage for both estimators is always greater than 70% when there is negligible bias. Generally the arithmetic mean had better coverage for the smaller sample sizes, while the lognormal estimator had better coverage for n = 100. Also, the lognormal estimator tended to have better coverage when $c \geq 1.0$. Of course, the average confidence interval width was also greater for the lognormal estimator

when the coverage was larger than the gamma estimator. For the case EX = 1.0 and uniform-lognormal data, Fig. 8 provides the reader with some feel for the relationship between sample size c, and coverage for the two estimators considered.

CONCLUSIONS AND RECOMMENDATIONS

For small values of c ($c < 1$), the differences between the two estimators is relatively small. Generally the confidence interval on the arithmetic mean provides better coverage at the expense of a wider confidence interval. For larger values of c ($1 \leq c \leq 2$), the lognormal estimator becomes very biased for gamma distributed data, and coverage tends to decline with increasing c. If no theoretical reasons are available for selecting one of the distributions simulated, then the arithmetic mean is to be preferred because it is unbiased for either of the distributions and tends to have reasonable coverage.

ACKNOWLEDGMENTS

I thank Drs. R. O. Gilbert, Battelle Pacific Northwest Laboratories, and R. J. Beckman, Los Alamos Scientific Laboratory, for reviewing drafts of this manuscript. Funding for this work was provided under contract No. W-7405-ENG.36 between the Department of Energy and Los Alamos Scientific Laboratory.

REFERENCES

1. Aitchison, J. and J. A. C. Brown. 1976. *The Lognormal Distribution.* Cambridge Univ. Press, Cambridge.

2. Bell, J. R. 1968. "Algorithm 334 Normal Random Deviates [G5]." Communications of the ACM *11*:498.

3. Choi, S. C., and R. Wette. 1969. "Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and their Bias." Technometrics *11*:683-690.

4. Eberhardt, L. L., and R. O. Gilbert. 1973. "Gamma and Lognormal Distributions as Models in Studying Food-Chain Kinetics." BNWL-1747. Battelle Pacific Northwest Laboratories.

5. Eberhardt, L. L., and R. O. Gilbert. (In Press.) "Statistics and Sampling in Transuranic Studies." *In: Transuranic Elements in the Environment.* W. C. Hanson, (Ed.). U. S. Dept. of Energy.

6. Ellett, W. H., and G. L. Brownell. 1964. "The Time Analysis and Frequency Distributions of Cesium-137 Fall-Out in Muscle Samples." *In: Assessment of Radioactivity in Man, Vol. II.* International Atomic Energy Agency, Viena. pp. 155-166.

7. Finny, D. J. 1941. "On the Distribution of a Variate Whose Logarithm is Normally Distributed." J. Royal Statistical Society Supplement 7:144-161.

8. Fishman, G. S. 1973. *Concepts and Methods in Discrete Event Digital Simulation.* Wiley, N. Y.

9. Forsyth , D. J., T. J. Peterle, and G. C. White. 1975. "A Preliminary Model of DDT Kinetics in an Old-Field Ecosystem." *In: Environmental Quality and Safety, Supplement Vol. III.* Georg Thieme, Stuttgart. pp. 754-759.

10. Giesy, J. P., Jr. and J. G. Weiner. 1977. "Frequency Distributions of Trace Metal Concentrations in Five Freshwater Fishes." Trans. Am. Fish Soc. *106*:393-403.

11. Land, C. E. 1974. "An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means." Technometrics *14*:145-158.

12. Link, R. F. and G. S. Koch, Jr. 1975. "Some Consequences of Applying Lognormal Theory to Pseudolognormal Distributions." Mathematical Geology 7:117-128.

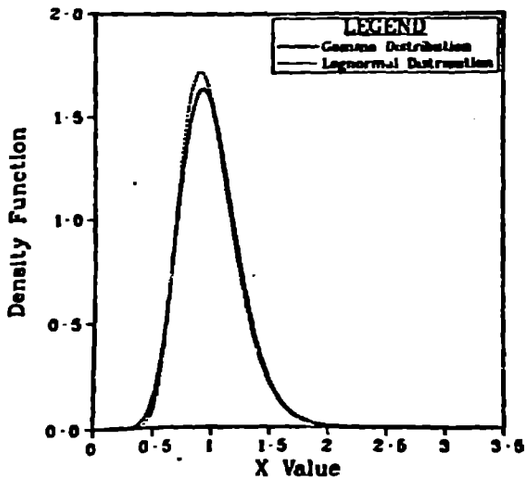
13. Mood, A. M., F. A. Graybill, and D. C. Boes. 1974. *Introduction to the Theory of Statistics*. 3rd ed. McGraw-Hill, N.Y.

14. Pinder, J. E., III, and M. H. Smith. 1975. "Frequency Distributions of Radiocesium Concentrations in Soil and Biota." *In: Mineral Cycling in Southeastern Ecosystems*. F. G. Howell, J. B. Gentry, and M. H. Smith (Eds.). ERDA Symposium Series, CONF-740513. pp. 107-125.

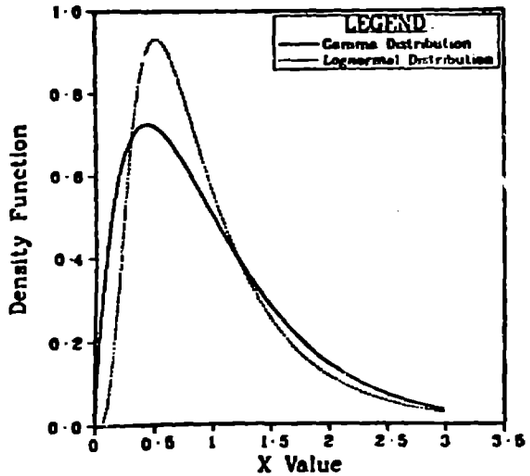
15. White, G. C. (In Prep.) "Estimation of Expected Value and Coefficient of Variation for Lognormal and Gamma Distributions."

Fig. 1. Comparison of gamma and lognormal probability density functions with expected value of unity and four values of the coefficient of variation.

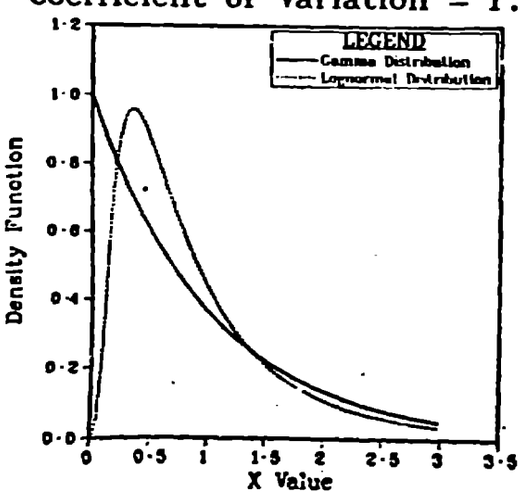
Coefficient of Variation = .25



Coefficient of Variation = .75



Coefficient of Variation = 1.00



Coefficient of Variation = 1.50

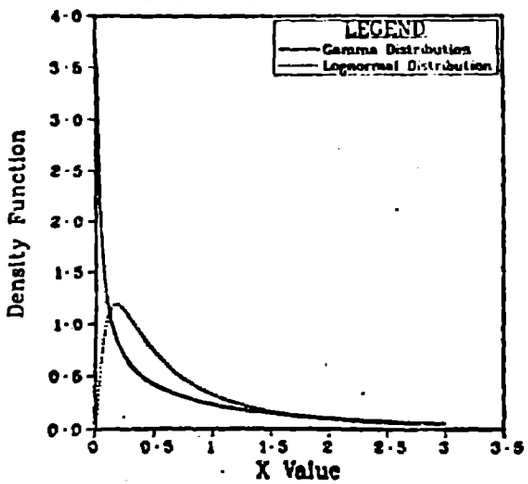
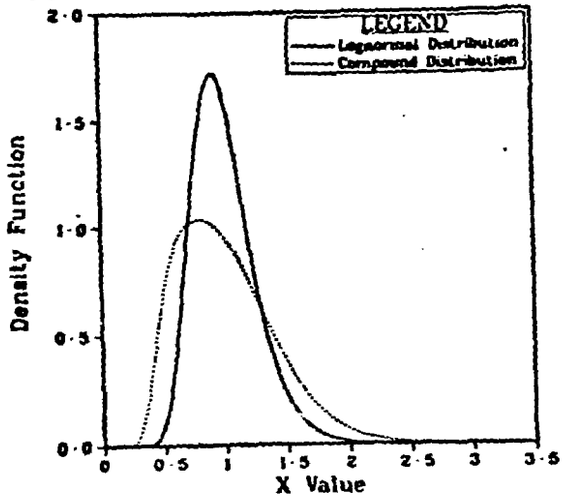
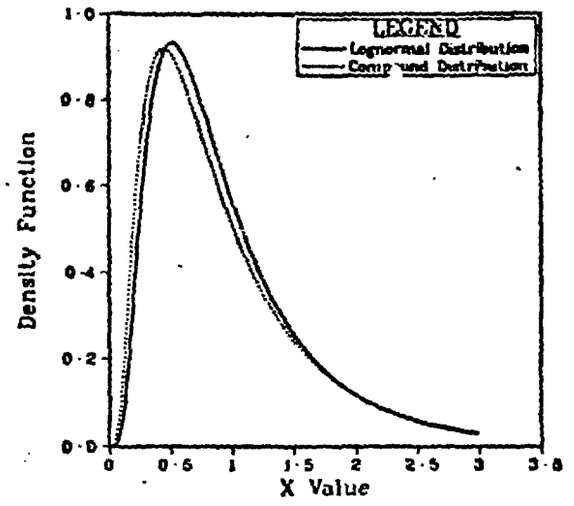


Fig. 2. Comparison of lognormal and compound uniform-lognormal probability density functions with expected value of unity and four values of the coefficient of variation.

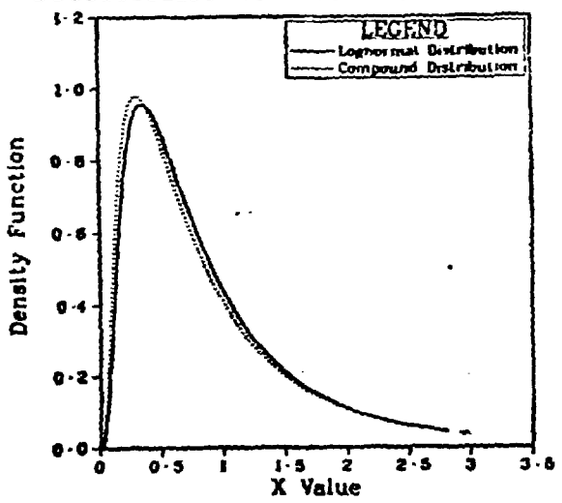
Coefficient of Variation = .25



Coefficient of Variation = .75



Coefficient of Variation = 1.00



Coefficient of Variation = 1.50

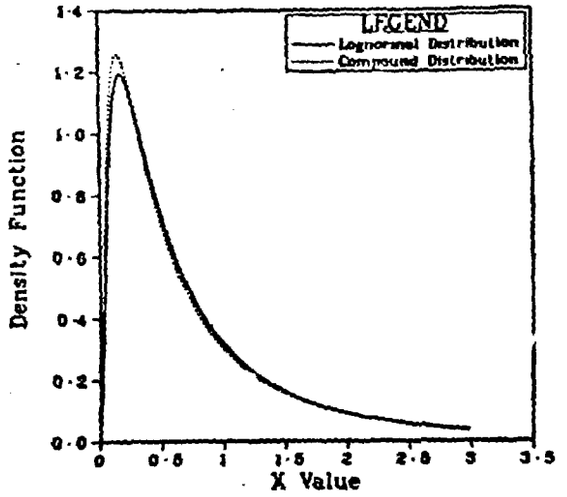
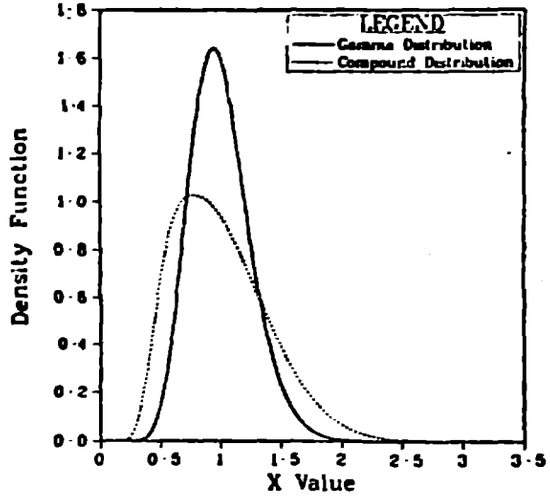
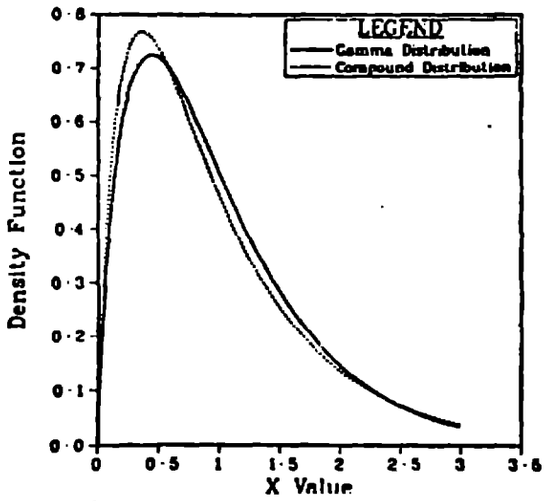


Fig. 3. Comparison of gamma and compound uniform-gamma probability density functions with expected value of unity and four values of the coefficient of variation.

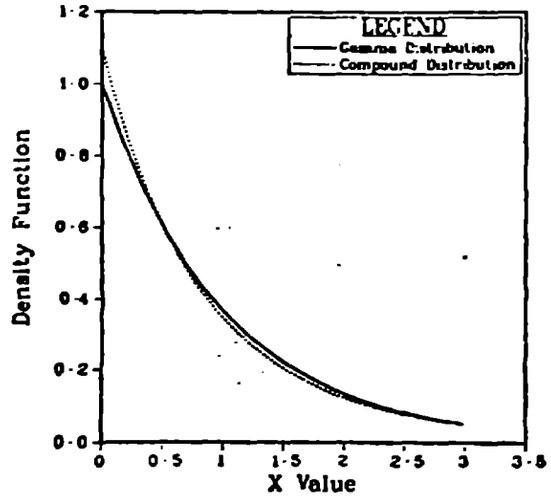
Coefficient of Variation = .25



Coefficient of Variation = .75



Coefficient of Variation = 1.00



Coefficient of Variation = 1.50

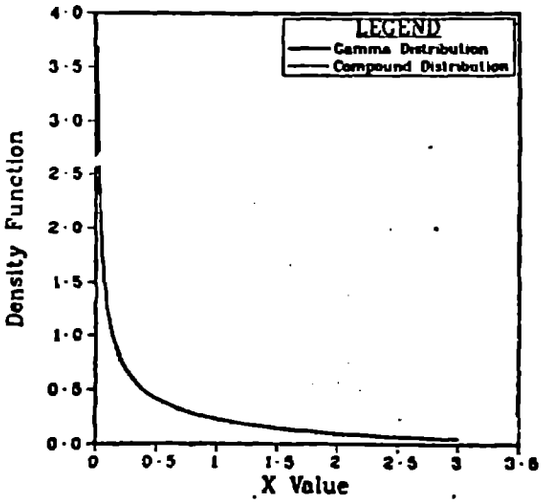


Fig. 4. Monte Carlo simulation results on confidence interval coverage of the gamma estimator (arithmetic mean) with gamma distributed data, EX=1. Values at the intersections of rows and columns are the proportion of 1000 replications where the computed confidence interval included the true expected value.

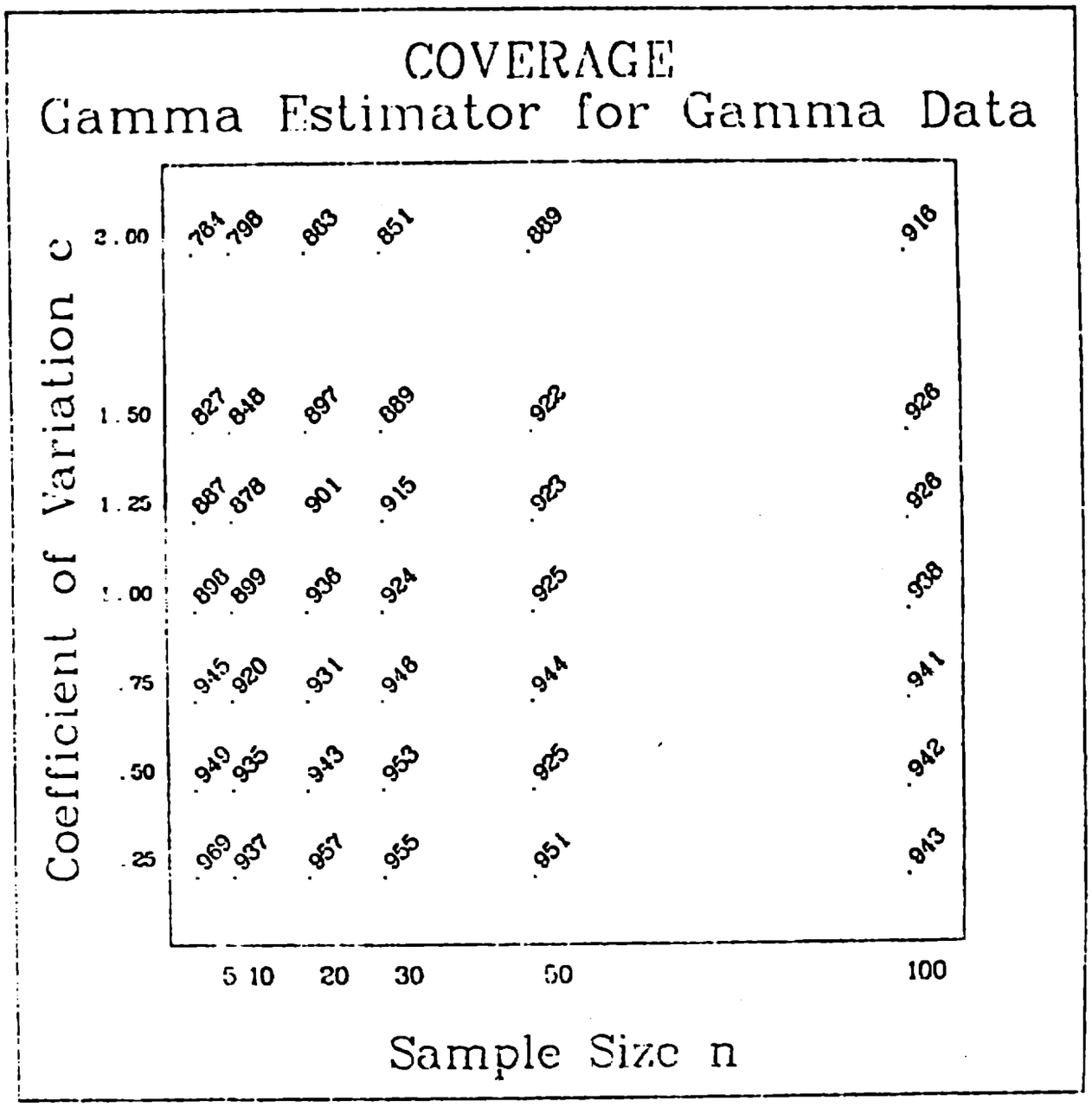

COVERAGE
Gamma Estimator for Gamma Data
Coefficient of Variation c
2.00
1.50
1.25
1.00
.75
.50
.25
784 798 803 851 889 916
827 848 897 889 922 928
887 878 901 915 923 928
898 899 936 924 925 938
915 920 931 918 944 941
949 935 943 953 925 942
969 937 957 955 951 943
5 10 20 30 50 100
Sample Size n

Fig. 5. Monte Carlo simulation results on confidence interval coverage of the lognormal estimator with lognormally distributed data, EX = 1. Values at the intersections of rows and columns are the proportion of 1000 replications where the computed confidence interval included the true expected value.

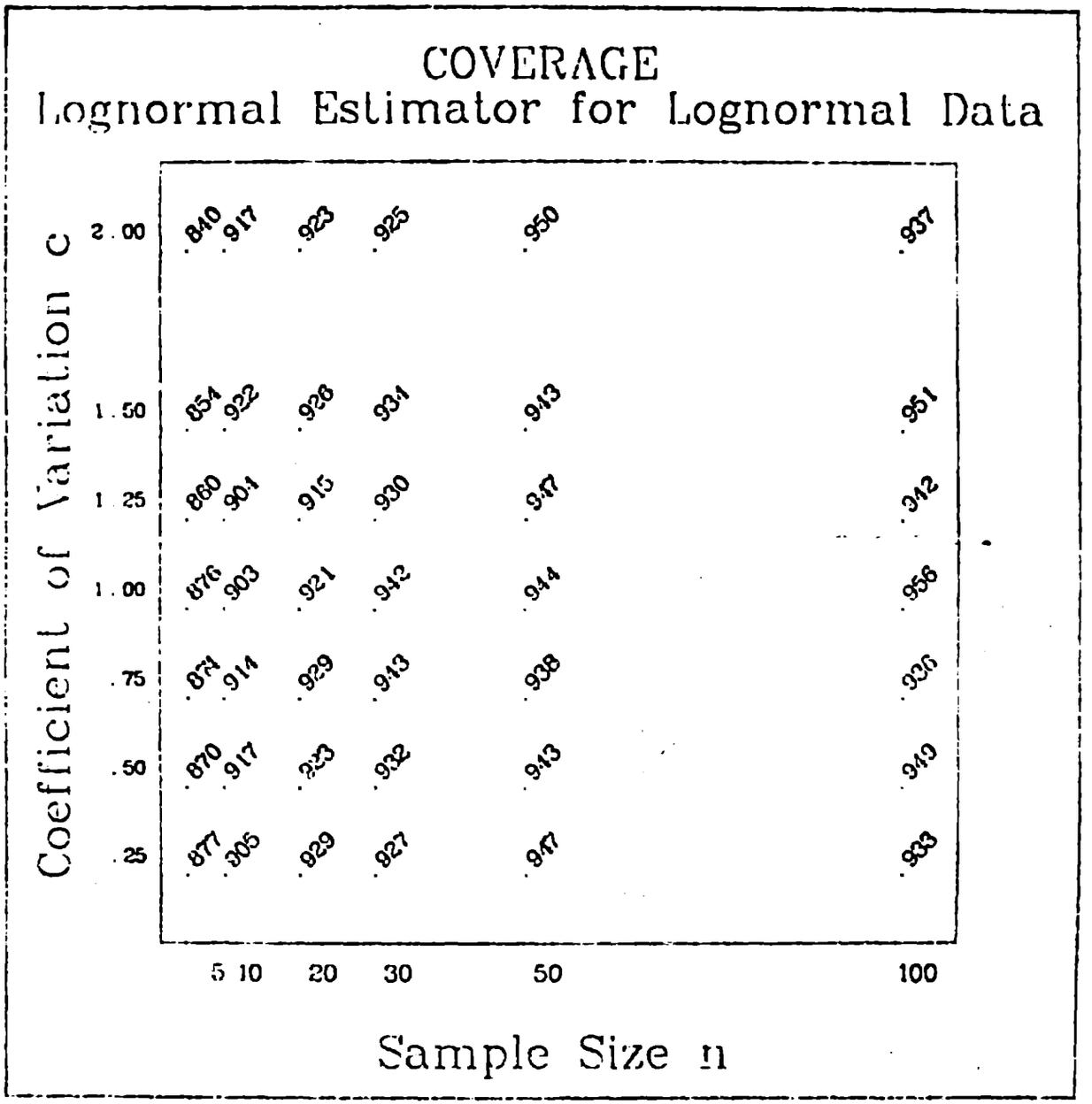
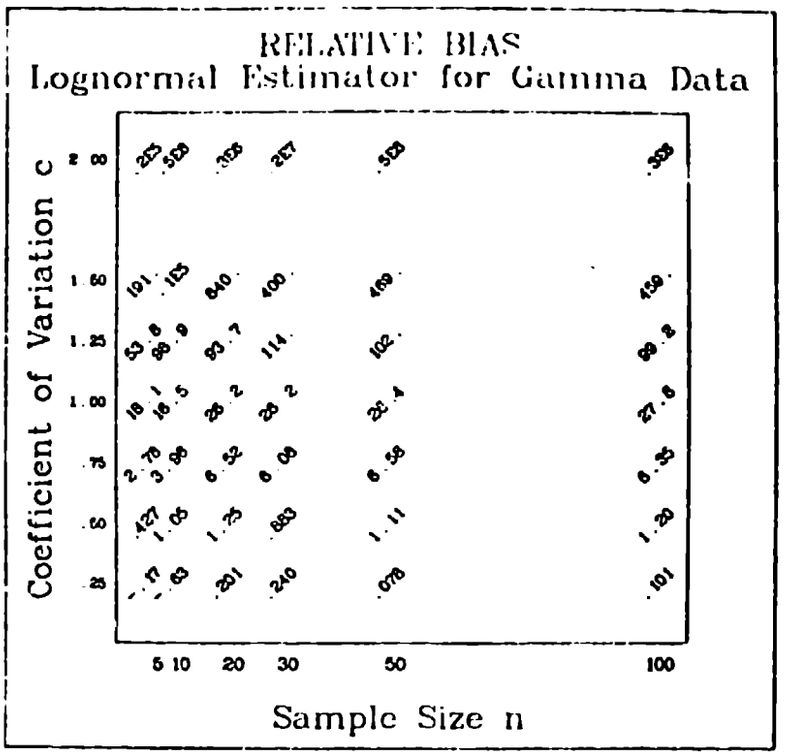

COVERAGE
Lognormal Estimator for Lognormal Data
Coefficient of Variation c
2.00
1.50
1.25
1.00
.75
.50
.25
.840 .917 .923 .925 .950 .937
.854 .922 .926 .931 .943 .951
.860 .904 .915 .930 .947 .942
.876 .903 .921 .942 .944 .956
.874 .914 .929 .943 .938 .936
.870 .917 .923 .932 .943 .949
.877 .905 .929 .927 .947 .933
5 10 20 30 50 100
Sample Size n

Fig. 6. Monte Carlo simulation results on bias and confidence interval coverage of the log-normal estimator with gamma distributed data, EX = 1. Values at the intersections of rows and columns are either 100[Ave($\hat{E}(x)$) - EX]/EX (upper figure) or the proportion of 1000 replications where the computed confidence interval included the true expected value (lower figure).


RELATIVE BIAS
Lognormal Estimator for Gamma Data
Coefficient of Variation c
Sample Size n
5 10 20 30 50 100

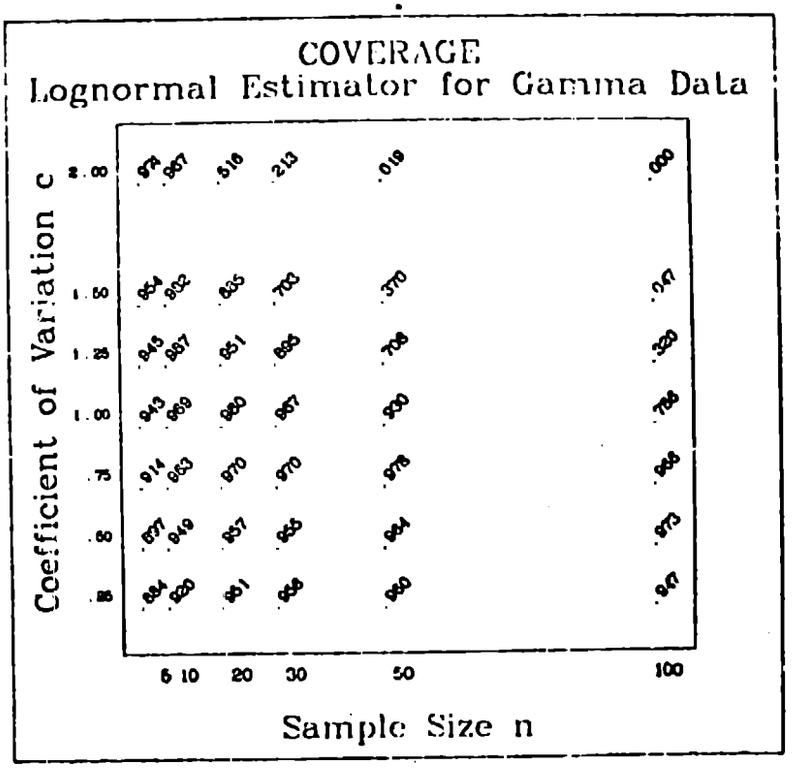

COVERAGE
Lognormal Estimator for Gamma Data
Coefficient of Variation c
Sample Size n
5 10 20 30 50 100

Fig. 7. Monte Carlo simulation results on confidence interval coverage of the gamma estimator (arithmetic mean) with lognormally distributed data, EX = 1. Values at the intersections of rows and columns are the proportion of 1000 replicates where the computed confidence interval included the true expected value.

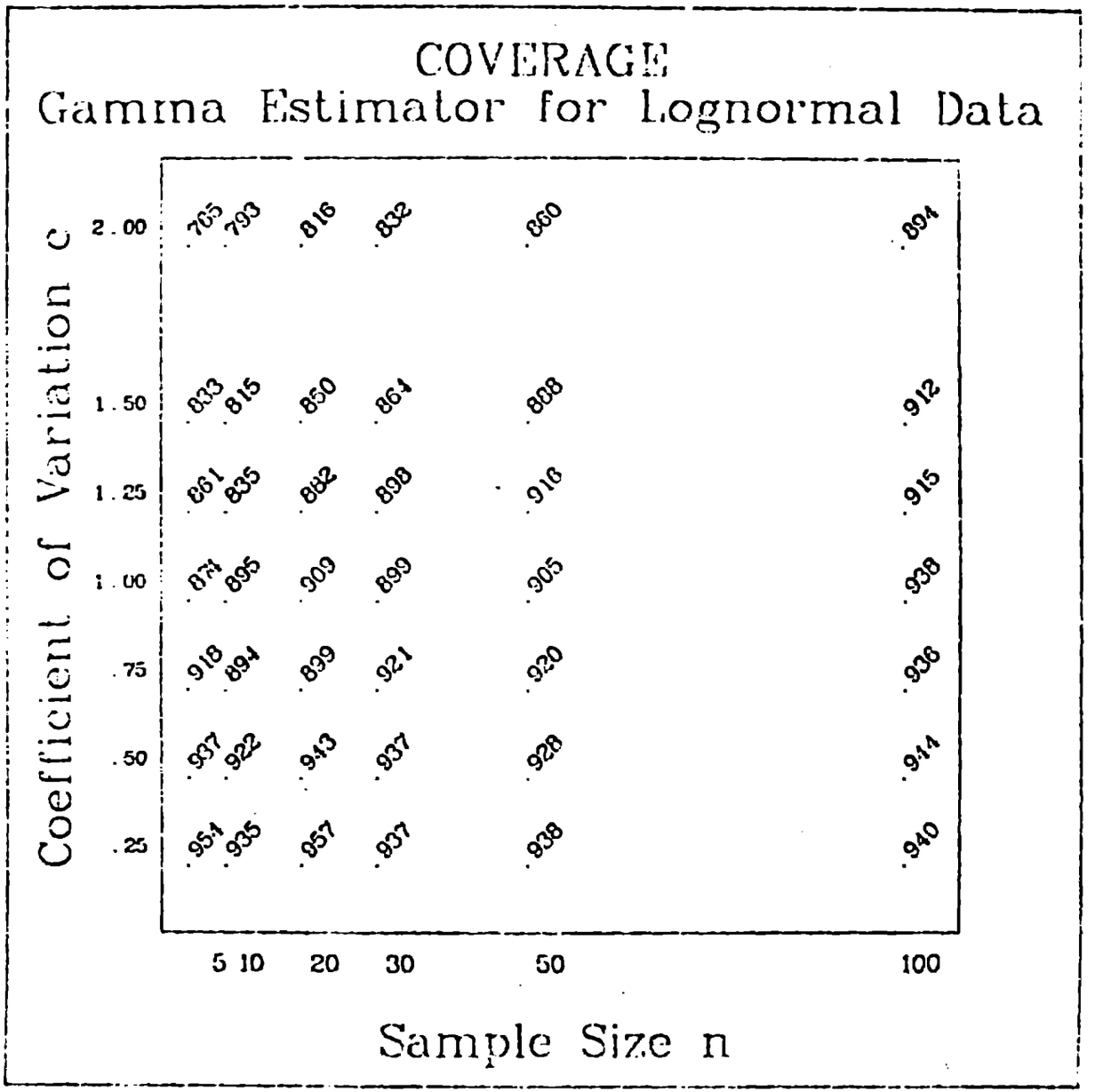
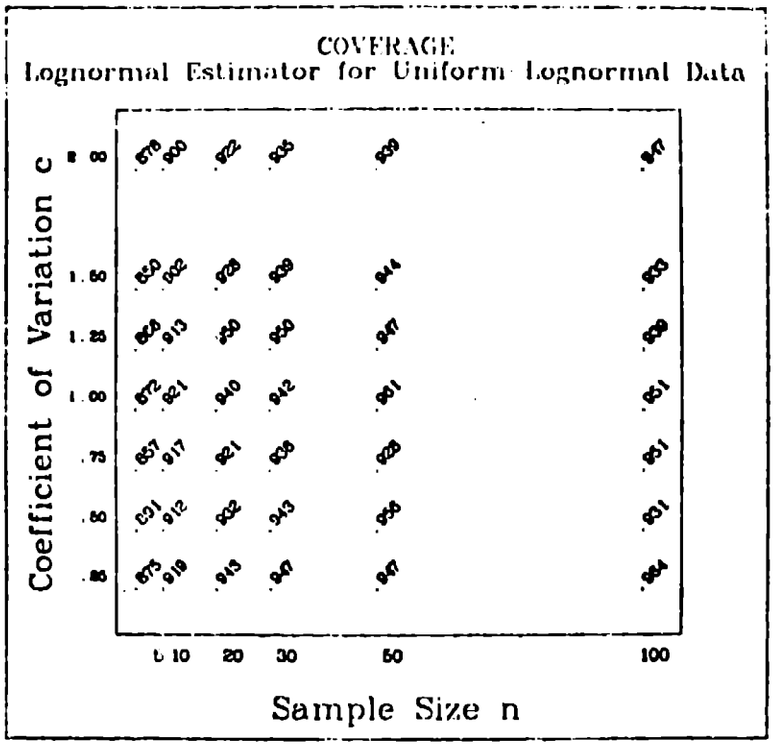
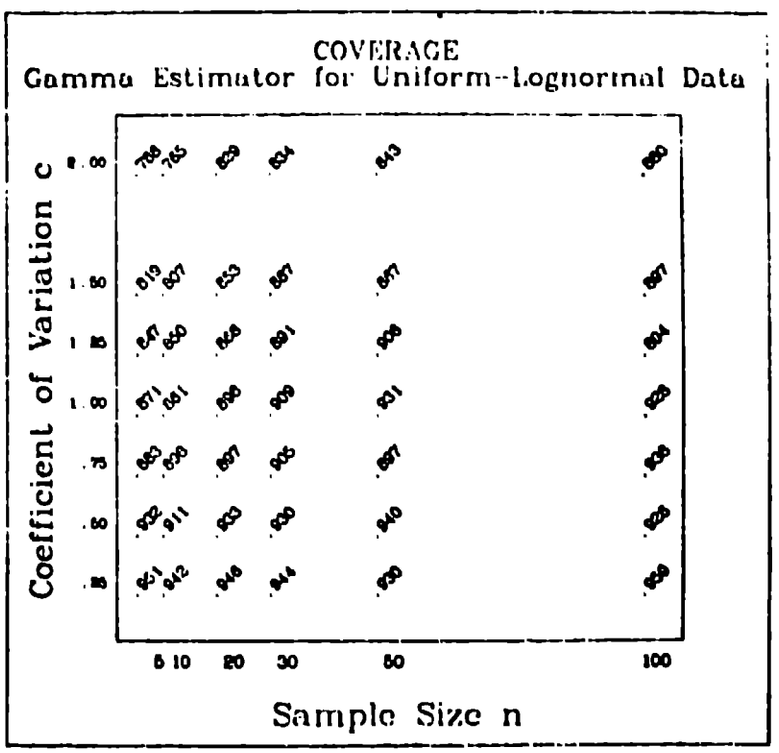

COVERAGE
Gamma Estimator for Lognormal Data
Coefficient of Variation c
2.00
1.50
1.25
1.00
.75
.50
.25
.705 .793 .816 .832 .860 .894
.833 .815 .850 .864 .888 .912
.861 .835 .882 .898 .918 .915
.874 .895 .909 .899 .905 .938
.918 .894 .899 .921 .920 .936
.937 .922 .943 .937 .928 .914
.954 .935 .957 .937 .938 .940
5 10 20 30 50 100
Sample Size n

Fig. 8. Monte Carlo simulation results on confidence interval coverage of the two estimators with the uniform-lognormal compound distributions, $EX = 1$. Values at the intersections of rows and columns are the proportion of 1000 replications where the computed confidence interval included the true expected value.


COVERAGE
Lognormal Estimator for Uniform-Lognormal Data
Coefficient of Variation c
Sample Size n
5 10 20 30 50 100


COVERAGE
Gamma Estimator for Uniform--Lognormal Data
Coefficient of Variation c
Sample Size n
5 10 20 30 50 100