United States
Environmental Protection
Agency

Office of
Research and
Development

Office of Solid Waste
and Emergency Response

EPA/600/s-99/006
November 1999

Technology Support Center Issue

Some Practical Aspects of Sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications

Ashok K. Singh[1], Anita Singh[2], and Max Engelhardt[3]

The Technology Support Projects, Technology Support Center (TSC) for Monitoring and Site Characterization was established in 1987 as a result of an agreement between the Office of Research and Development (ORD), the Office of Solid Waste and Emergency Response (OSWER) and all ten Regional Offices. The objectives of the Technology Support Project and the TSC were to make available and provide ORD's state-of-the-science contaminant characterization technologies and expertise to Regional staff, facilitate the evaluation and application of site characterization technologies at Superfund and RCRA sites, and to improve communications between Regions and ORD Laboratories. The TSC identified a need to provide federal, state, and private environmental scientists working on hazardous waste sites with a technical issue paper that identifies data assessment applications that can be implemented to better define and identify the distribution of hazardous waste site contaminants. The examples given in this Issue paper and the recommendations provided were the result of numerous data assessment approaches performed by the TSC at hazardous waste sites.

This paper was prepared by Ashok K. Singh, Anita Singh, and Max Engelhardt. Support for this project was provided by the EPA National Exposure Research Laboratory's Environmental Sciences Division with the assistance of the Superfund Technical Support Projects Technology Support Center for Monitoring and Site Characterization, OSWER's Technology Innovation Office, the U.S. DOE Idaho National Engineering and Environmental Laboratory, and the Associated Western Universities Faculty Fellowship Program. For further information, contact Ken Brown, Technology Support Center Director, at (702) 798-2270, Ashok K. Singh at (702) 895-0364, Anita Singh at (702) 897-3234, or Max Engelhardt at (208) 526-2100.

Purpose and Scope

Often in Superfund applications of the U.S. EPA, exposure assessment and cleanup decisions are made based upon the mean concentrations of the contaminants of potential concern (COPC) at a polluted site. The objective may be to 1) compare the soil concentrations with site specific or generic soil screening levels (SSLs),

[1] Department of Mathematical Sciences, University of Nevada, Las Vegas, NV 89154
[2] Lockheed Martin Environmental Systems & Technologies, 980 Kelly Johnson Dr., Las Vegas, NV 89119
[3] Lockheed Martin Idaho Technologies, P.O. Box 1625, Idaho Falls, ID 83415-3730

COMMUNICATION • TRAINING • TECHNOLOGY SUPPORT •
Technology
Support
Project

Technology Support Center for
Monitoring and Site Characterization
National Exposure Research Laboratory
Environmental Sciences Division
Las Vegas, Nevada 89193-3478

Technology Innovation Office
Office of Solid Waste and Emergency
Response
U.S. EPA, Washington, D.C.
Walter W. Kovalick, Jr., Ph.D., Director

Printed on Recycled Paper

2) compute the exposure point concentration (EPC) term used as one of several parameters to estimate the contaminant intake for an individual, or 3) verify the attainment of cleanup goals (CUGs) or cleanup standard as set forth in the Record of Decision (ROD) agreed upon by all concerned parties, such as the USEPA and the party responsible for introducing contamination at the site. The CUG of a COPC has been denoted by C_s throughout this article. Suppose that a COPC is believed to be present at a certain site and its concentration varies according to a probability distribution with an unknown mean, μ. The mean, μ, is one of the commonly used measures of the central tendency of a distribution and is often used to represent the EPC term or some cleanup standard at a site. The mean, μ, is typically estimated by the sample mean and some upper confidence limit (UCL) of the mean, which are obtained using the sampled data.

The decisions about the population mean are made using testing of hypotheses about the population mean. In general, there are two hypotheses of interest, the *null hypothesis,* denoted by H_0, and the *alternative hypothesis,* denoted by H_a. In Superfund applications, such as the determination of exposure assessment or the attainment of cleanup levels, it is of interest to test one-sided hypotheses about the population mean; therefore, all sample size and power computation discussions in this paper have been done for one-sided hypotheses. For example, suppose a regulator suspects that the mean concentration of the contaminant exceeds a specified level, say μ_0 (the CUG, C_s), but the party responsible for introducing the contaminant claims that the mean concentration is below μ_0. A mathematical formulation of these hypotheses would be the null hypothesis, H_0: $\mu \geq C_s$, which is the regulator's claim, and the alternative hypothesis, H_a: $\mu < C_s$, which is the potentially responsible party's (PRP's) claim. In these applications, for an alternative value, μ_1 of μ, it is desirable to be able to detect an error margin, $\Delta = C_s - \mu_1$, when the mean of the contaminant is at or below μ_1, with high power and confidence. In other words, the objective is to be able to detect when the population mean approaches μ_1 $(< C_s)$ with pre-specified Type I and Type II error rates. The null hypothesis described here is protective of the environment as it assumes that the area of the site is dirty, and the burden of testing is to show otherwise. This null hypothesis has been used throughout this article.

Some of these issues have been well studied for normally distributed data sets and are documented in USEPA documents (e.g., 1989a, 1989b, 1992). However, data distributions of contaminants originating from environmental applications are often positively skewed, and are invariably modeled by a lognormal distribution. As noted by Singh, Singh, and Engelhardt (1997), the use of decision criteria based on the mean of a lognormal distribution can have undesirable consequences, especially for samples of small sizes. For example, when the UCL of the mean is obtained using a lognormal distribution, one may end up spending more time than is necessary on a Superfund cleanup project in one case and leaving the contamination behind in the other. The later situation can arise when reference or background data based UCLs are obtained using a lognormal distribution.

1.0 Introduction

Hypotheses testing and computation of the UCL of the mean require the availability of an adequate number of data values so that the resulting statistical inference can be considered credible in achieving the pre-specified performance parameters, such as the error rates and power. These data are then used to provide statistical evidence about the truth or falsity of the hypotheses. The problem of obtaining an adequate number of samples for computing the EPC term or the verification of the attainment of cleanup goals satisfying pre-specified performance parameters, such as the Type I (α) and Type II (β) error rates, have been considered in this paper. In statistical terminology, Type I error is the probability of rejecting the null hypothesis when in fact it is true and Type II error represents the probability of not rejecting the null hypothesis when in fact it is false. Type I (level of significance, test size) and Type II error rates are also known as false positive and false negative error rates, respectively. Depending upon how the null hypothesis is designated, an environmental chemist may find this definition of the false positive error rate contradictory to his intuitive definition of a false positive. As adopted in EPA documents (1989b), the statistical convention has been followed in this paper as well.

One of the inherent assumptions required to determine the sample size is that one is dealing only with a single statistical population (e.g., one remediated part of the site). Violation of this assumption can lead to invalid applications of a statistical model (e.g., lognormal) and technique. For example, a normally distributed data set with a few outliers can be incorrectly modeled by the lognormal distribution with the lognormal assumption hiding the outliers. Also, the mixture of two or more datasets with significantly different mean concentrations, such as one coming from a clean part and the other taken from a contaminated part of the site, can also be incorrectly modeled by a lognormal distribution. These are frequent occurrences in environmental applications as discussed by Singh, Singh, and Engelhardt (1997). It appears that the use of the lognormal distribution and the H-statistic based UCL tend to hide contamination rather than find it. Actually, under the assumption of lognormal distribution, one can get away with very little or no cleanup (Bowers, Neil, and Murphy, 1994) at a polluted site. Moreover, there are practical problems which can occur even when the lognormal assumption is correct; especially when the distribution is highly skewed and the number of samples taken (available data) is small. In such a case the H-UCL of the mean can be orders of magnitude greater than the true mean concentration, making the UCL of little practical use for the intended purpose.

The main objective of this article is to discuss some problems regarding the lognormal assumption and how they relate to sample size determination needed to draw reliable inference about hypotheses testing for the population mean with prespecified performance parameters. Methods for computing the number of samples from a normally distributed population are available in the literature (Bain and Engelhardt, 1992). The use of the standard sample-size formula when the population variance is unknown has been discussed by Kupper and Hafner (1989), who proposed a simple adjustment for sample size determination when the population variance is unknown. For a lognormal distribution, many times, the practitioners like to use the standard formula given by equation (5) below, as an approximation, but it has been recommended (Stewart, 1994) that caution should be exercised while using the standard formulas for computing the number of samples.

In general, when a statistical procedure is based on correct assumptions, by taking a sufficiently large number of samples it is possible to make decisions about the parameters (e.g., the mean) with whatever level of confidence is prescribed. However, in real applications, taking a large number of samples may not be practical as it may be time consuming with an unacceptably high cost. When approximate formulas are used for determining the number of samples, or if the lognormal assumption is wrong, it is possible to end up with either too many or too few samples. In the former case, the cleanup expense will be too high, and in the latter case, the actual level of confidence may fall short of what was prescribed by the regulators. It is very well possible that it may not be feasible to achieve the desired performance parameters without taking an enormous number of samples. This is especially true when a lognormal model is assumed. Therefore, it becomes necessary to find a balance between the choice of performance parameters (error rates, power) and the number of samples needed for hypothesis testing. For example, when more (or less) samples are taken, then the gain (or loss) in levels of performance standards for the various approximate formulas need to be investigated. Keeping some of these practical considerations in mind, the regulators may have to settle for reduced values of performance standards. A multi-phase approach may have to be adopted. In this article, several UCL of the mean computation methods have been compared via Monte Carlo simulations.

A convenient way to perform a test of hypotheses about an unknown parameter is to first compute a confidence interval for the parameter, and then reject H_0 if the hypothesized value, in this case the cleanup standard, C_s, lies outside of the interval. For the verification of the attainment of a cleanup goal, the test based upon the one-sided UCL of the mean is typically used. A one-sided UCL is a statistic such that the true population mean is less than the UCL with a prescribed level of confidence, say $(1-\alpha)100\%$. The associated test rejects H_0, suggesting that the site is clean if UCL < C_s. The choice of an appropriate statistical procedure depends on the distributional assumptions and knowledge of the variance, σ^2.

2.0 Normal and Lognormal Distribution

Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with unknown mean, μ, and variance, σ^2. Denote the *sample mean* and *sample variance*, respectively, as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and} \tag{1}$$

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

An important feature of the normal model is that the mean and standard deviation (sd) are location and scale parameters, respectively. In particular, if a normally distributed random variable is transformed by adding a constant, the effect on the density function is a simple linear translation without changing the shape of the density function. For example, given two normal densities with the same sd, if the means differ by 3 units, then the 90th percentiles also differ by 3 units, the 95th percentiles differ by 3 units, and so forth. This makes it possible to derive a sample size formula in terms of the difference, or the error margin (limit), $\Delta = C_s - \mu_1$. Figure 1 exhibits normally distributed densities for three different sites, each with the same sd, $\sigma = 0.5$ ppm, but with different means (designated by dashed vertical lines), $\mu = 2$ ppm at site A, $\mu = 5$ at site B, and $\mu = 10$ at site C.

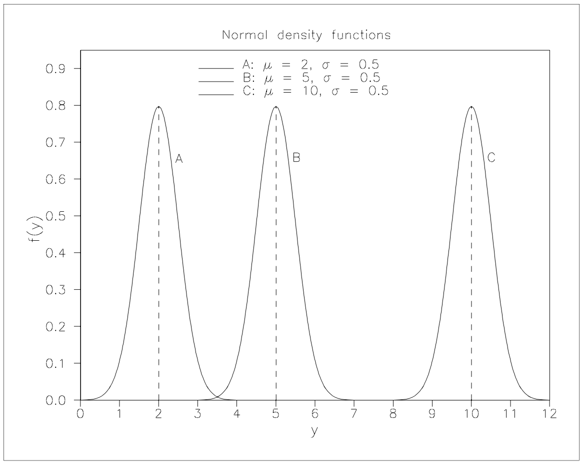


Figure 1. Normal density functions with different means, μ = 2, 5, 10.

However, the mean of a lognormal distribution is not a location parameter. It is possible for lognormal distributions which appear to be located at roughly the same place to have very different means and, conversely, there exist lognormal distributions with the same mean, μ_1, which appear to differ in location. This makes it impossible to derive a sample size formula in terms of the error margin, Δ. Thus, for a lognormal model, the problem reduces to distinguishing between the two populations with mean, μ ($\geq C_s$), and μ_1 ($<C_s$) with pre-specified error rates. Figure 2 has density functions of lognormal populations with different means, μ_1 = 2, 5, and 10, for the log-transformed variable with σ = 1, and varying mean, μ. This figure shows that differences in means for lognormal distributions are not as easily identified by the inspection of graphical displays of the respective density functions.

Another interesting comparison is obtained by studying several lognormal distributions with varying values of the parameters, μ and σ, while holding the lognormal mean, μ_1, constant. This situation is illustrated in Figure 3. For a lognormal model, a small value of σ corresponds to a small skewness as can be seen by the nearly symmetric density function of population A, Figure 3. Furthermore, the other distributions in Figure 3 clearly show greater amounts of skewness, corresponding to larger values of σ. Consequently, it is difficult, based on the usual graphical paradigm, to recognize that all five distributions have the same mean. It can be shown mathematically that larger values of σ yield distributions not only with greater skewness, but also with a thicker right tail. The distribution mean tends toward the thicker tail of a skewed distribution, but tail thickness is generally hard to spot through visual inspection.

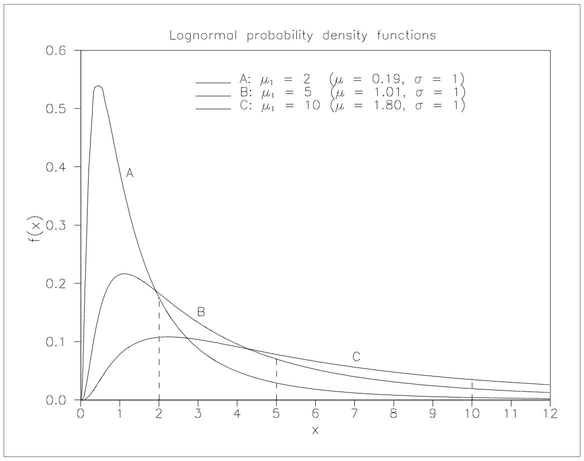


Figure 2. Lognormal density functions with means, μ_1 = 2, 5, 10.

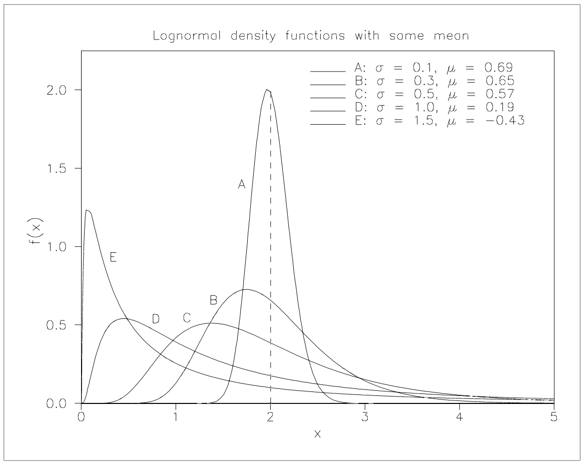


Figure 3. Lognormal density functions, all with μ_1 = 2.

The mean is an intuitive and a commonly used measure of central tendency of a distribution. The sample mean and the associated UCL are often used to verify the attainment of cleanup goals and SSLs, and to estimate the EPC terms in exposure and risk assessment studies. USEPA guidance documents recommend the use of H-statistics to compute a UCL of the mean of a lognormal distribution (EPA, 1989a, 1992, 1996). A detailed discussion of H-statistics is given in Gilbert (1987). Even though, for a lognormal distribution, the test based on the H-statistic is uniformly most powerful (UMP), it has little practical merit for positively skewed data sets of small sizes, such as 20-30 or less. Also, recent work by Gilbert (1993) and Singh, Singh, and Engelhardt (1997) indicates that statistical tests of hypotheses based on H-statistics can yield unusually high false negatives (not rejecting H_0 as defined above, when in fact it is false), which would result in unnecessary cleanup. This is especially true for samples of small sizes from skewed populations with σ exceeding 1 as can be seen in Figures 3A-3C, 4A-4D, 11A-11C, and 12A-12D. These comments suggest that for large values of σ, mean is not a good measure of central tendency for a lognormal distribution. Other parameters which are sometimes used as measures of central tendency are the median (50th percentile) and the mode (maximum of the density) of the distribution. For a unimodal symmetric distribution, such as the normal distribution, the mean, median, and mode are the same. However, the mean, median, and mode can be quite different for a highly skewed distribution. For example, the mean, median, and mode of population E in Figure 3 are 2.0, 0.65, and 0.05, respectively.

Singh, Singh, and Engelhardt (1997) note that ordinarily one would expect the mean and the associated 95% UCL of the mean to be smaller than the 95th percentile of the sampled population. While this is a likely occurrence when the population is normal, the situation can be somewhat different with lognormal populations. For example, it has been observed (see Example 1) that the mean and, consequently, the H-statistic based UCL of the mean can exceed the 90% or 95% percentiles of the lognormal distribution by orders of magnitude, especially for skewed datasets of small sizes. This fact can be easily seen by comparing percentiles with the mean of a lognormal distribution. The *population mean*, μ_1, is greater than x_p, the $100p^{th}$ percentile of a lognormal distribution if and only if $\sigma > 2z_p$, where z_p is the $100\ p^{th}$ percentile of the standard normal distribution. For example, when p = 0.80, $z_p = 0.842$, then μ_1 exceeds $x_{0.80}$, the 80th percentile, if and only if $\sigma > 1.68$, and μ_1 will exceed the 95th percentile if and only if $\sigma > 3.29$. This observation and the simulation results summarized in Section 4 suggest that for σ exceeding 2 and samples of a size as large as 50, the sample mean and the associated H-statistic based UCL for the lognormal mean becomes unrealistically large and cannot be considered a reliable estimate of a cleanup standard or of an EPC term. Due to these reasons, the 1996 EPA Soil Screening Guidance Document abandoned the use of H-UCL to compare the soil concentrations with the SSLs.

Using Monte Carlo simulation experiments, it has been observed that the H-statistic based UCL for the mean is greater than the true mean and the CUG by orders of magnitude even when the sample was drawn from a population with a mean smaller than the cleanup standard, as can be seen in Figures 7A, 7B, 8A-8D, and 15A, 15B, 16A-16D. It is, therefore, desirable to have procedures which work better (achieving the pre-specified performance measures, approximately) than the H-statistic based UCL.

Example 1. Consider a simulated dataset of size *n* = 15 from a lognormal distribution, LN(5, $(1.71)^2$). The generated data range from 16.52 to 1498.61, and are given in Singh, Singh, and Engelhardt (1997). The mean and sd of the lognormal distribution are $\mu_1 = 629.55$, and $\sigma_1 = 2595.18$, and the 80th, 90th, and the 95th percentiles for this lognormal distribution are 626.29, 1329.05, and 2472.41, respectively. Note that the sd, 1.71, exceeds the $2*z_{0.80} = 2* 0.842 = 1.68$; therefore, the mean, μ_1, already exceeds the 80th percentile of the lognormal distribution. The 95% UCL of the mean based on the t-distribution, central limit theorem (CLT), the Chebychev theorem (based on the minimum variance unbiased estimates of mean and sd of the lognormal distribution), and the H-statistic are: 749.31, 731.14, 2059.47, and 4613.32, respectively. Notice that the 95% H-UCL is 4613.32 which exceeds the 95th percentile value of 2472.41 for the lognormal distribution. Thus, even though the H-UCL is theoretically sound and possesses optimal properties, the practical merit of the use of H-UCL is questionable, as it becomes quite large when the sd of the log-transformed variable starts to exceed 1.0. This is especially true for samples of small sizes (viz., <30).

In light of the above remarks, it is crucial that great care should be exercised in choosing an appropriate model and in understanding the potential problems associated with the chosen model when attempting to make decisions about a population mean. Other measure of central tendency, such as the median or some other quantile (e.g., 75%, 85%) and other distributions (e.g., Weibull, Gamma) need to be considered for highly skewed data sets, which will be discussed in a sequel article. In addition to describing the complexity of interpreting statistical evidence about the mean of a lognormal, this article also discusses the difficulties involved in selecting an adequate number of samples when drawing an inference about the mean of a lognormal distribution. In order to shed some light on these issues, the coverage probability, statistical power, and the UCLs for the various procedures have been compared via Monte Carlo simulation experiments. Section 2 has a brief description of normal and lognormal distributions, and Section 3 discusses the methods for computing the UCL of the mean of a lognormal distribution. In Section 4, the power and the UCL of the mean obtained using these procedures have been compared via Monte Carlo simulation experiments, and conclusions have been summarized in Section 5.

2.1 Normally Distributed Datasets

Variance, σ^2, is known

If σ^2 is <u>known</u>, then a $(1-\alpha)100\%$ one-sided UCL for the mean is given by the equation:

$$\text{UCL} = \bar{x} + z_a \sigma/\sqrt{n}, \qquad (3)$$

where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution (SND) and H_0 (defined earlier) is rejected if UCL < C_s with a false positive rate of α. In order to study the false negative rate, β, it is necessary to consider the power function of the test. In general, the *power function*, denoted by $\prod(\mu)$ = P[H_0 is rejected given μ is true] = P[reject $H_0 \mid \mu$], is the probability of rejecting H_0. Using the properties of the SND, the power function is given as follows:

$$\Pi(\mu) = \Phi\left([\sqrt{n}(C_s - \mu)/\sigma] - z_{1-\alpha}\right), \qquad (4)$$

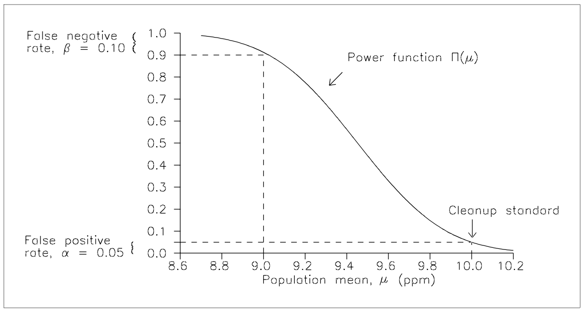


Figure 4. Power function for test of hypothesis H_0: μ > 10 ppm.

where Φ is the standard normal cumulative distribution function. Thus, if it is critical to detect when the difference between the true mean and the hypothesized mean is at least Δ, then it would be desirable to have a low false negative rate, β at μ_1, with $\prod(\mu_1) = 1 - \beta$ and $\mu_1 = C_s - \Delta$. For example, suppose it is required to perform a test of H_0: μ > 10 ppm, and the test is based on a sample from a normal population with known σ = 1, and it is required to have no more than 5% (α = 0.05) false positive decisions and no more than 10% (β = 0.10) false negative decisions if the true mean is $\mu \le 9$ ppm, a scenario given in Figure 4.

For a given choice of the standardized difference, $d = (C_s - \mu)/\sigma$, and *error rate, α*, there is no guarantee that a prescribed false negative rate, β, will be achieved. Note that for normal distributions, the power function in equation (4) is an increasing function of the sample size, *n*, and difference, *d*. Thus the power function can be made arbitrarily close to 1 by choosing sufficiently large *n (this may not be practical)*.

Thus, equating the power in equation (4) to $1 - \beta$ and solving it for *n* results in the following formula,

$$n = \left[\frac{(z_{1-\alpha} + z_{1-\beta})\,\sigma}{C_s - \mu}\right]^2 = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{d}\right)^2 \qquad (5)$$

Note that in equation (5), one can also use the critical values based upon the Chebychev bound, which will only result in a higher sample size as the critical values based on the Chebychev inequality will be higher than those based upon the normal distribution. For example, for $\alpha = 0.05$ and $\beta = 0.10$, $z_{1-\alpha} = 1.645$, and $z_{1-\beta} = 1.282$, whereas the corresponding conservative cut offs based upon the Chebychev inequality are 4.47 and 31.6, respectively.

Example 2. Suppose it is required to have a test in which the false positive rate is $\alpha = 0.05$ and the false negative rate is $\beta = 0.10$ when the true mean, μ, is 0.5 sd below C_s with $d = (C_s - \mu)/\sigma = 0.5$. The sample size required to achieve these performance parameters is = $[(1.645 + 1.282)/0.5]^2 = 34.3$, which when rounded up yield a value of 35. For a less stringent condition, with α and β as 0.05 and 0.10, respectively, and *d* = 1.0, the sample size is given by $n = [(1.645 + 1.282)/1.0]^2 = 8.6$, which can be rounded up to 9.

Variance, σ^2, is unknown

If σ^2 is unknown, then a $(1-\alpha)100\%$ one-sided UCL for the mean is provided by

$$\text{UCL} = \bar{x} + t_{\alpha,n-1} s_x/\sqrt{n}, \qquad (6)$$

where $t_{1-\alpha,\,n-1}$ is the $(1-\alpha)$th quantile of the Student's *t* distribution with $n-1$ degrees of freedom (df). The test rejects H_0 and declares the site clean if UCL < C_s, with a false positive error rate, α. In this case, the power function can still be expressed as the probability that the UCL defined by equation (6) is less than C_s, but its evaluation becomes complicated requiring the use of the noncentral t distribution. Tables of sample size based on the noncentral t distribution are given in Bain and Engelhardt (1992).

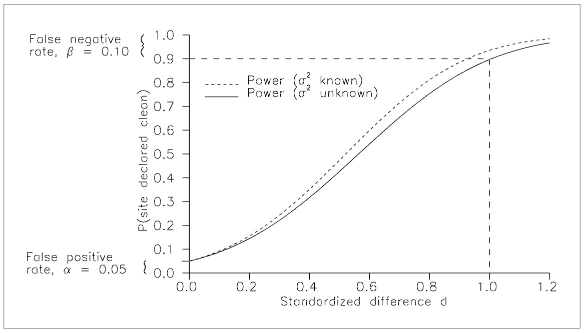


Figure 5. Power functions for tests of mean with n = 11 and α = 0.05.

Example 2 (continued). Consider Example 1, with $d = 1.0$, $\alpha = 0.05$, $\beta = 0.10$, and σ is unknown. Using the non-central t distribution based tables (Bain and Engelhardt), the required sample size comes out to be n =11, which is only two more than the sample size obtained using equation (5). For n =11, the power functions for the two cases, σ known and σ unknown, have been plotted in Figure 5. Kupper and Hafner (1989) note that the actual power attained using sample sizes computed using equation (5) is generally quite close to the desired power, as can also be seen in Figure 5. The sample size required for the test when σ is unknown will always be larger than when it is known. Specifically, when σ is unknown, Kupper and Hafner recommended adjusting the sample size by adding 2 or 3 to the result obtained by using equation (5).

2.2 Positively Skewed Datasets

For a lognormal population, the skewness is a function of σ, as can be seen in Figure 3. A lognormal distribution is typically used for highly positively skewed datasets. If there is evidence that the population distribution is positively skewed, especially when σ is larger than 1, then a lognormal distribution is often assumed. Let x_1, x_2, ... , x_n be a random sample from a lognormal population, LN(μ, σ^2). In other words, the natural logarithm of data are normal with mean μ and variance σ^2, N(μ, σ^2). Let $\overline{y}$ and s_y, denote the sample mean and sample sd, respectively, of the log-transformed data $y_i = \ln(x_i)$; $i = 1, 2, \ldots, n$. For a lognormal population, the mean is $\mu_1 = \exp(\mu + 0.5\sigma^2)$ and the median is given by $M = \exp(\mu)$. Note that μ_1 is the mean of the lognormal distribution and μ is the mean of the transformed distribution, and for positive values of σ, μ_1 is always greater than M.

Variance, σ^2, is known

When σ^2 is known, the problem can be converted to the sample size determination for a normal population by means of a log-transformation. The null hypothesis, H_0: $\mu_1 \geq C_s$, is equivalent to the hypothesis, H_0: $\mu \geq \ln(C_s) - 0.5\sigma^2$, which can be tested by comparing the normal distribution based UCL applied to the transformed data with the hypothesized mean, $\mu_0 = \ln(C_s) - 0.5\sigma^2$.

Variance, σ^2, is unknown

There is no easy solution to compute the sample size, power, and the UCL of the lognormal mean when σ^2 is unknown. Some of the available procedures are discussed in the following.

Test based upon the median, *M*

A test for the median, $M = \exp(\mu)$, is sometimes used to test for the mean of a lognormal distribution. The hypothesis for the median is H_0: $M \geq C_s$, which is equivalent to H_0: $\mu \geq \ln(C_s)$, and the results for the normal population can be applied with $\mu_0 = \ln(C_s)$. If σ^2 is known, then a test (or the UCL) based on the transformed data for a normal distribution may be used; on the other hand, if σ^2 is unknown, then a test based on the Student's - t distribution can be used. It should be stressed that this only provides a test for the median, M, and <u>does not</u> provide a test of the lognormal mean, μ_1. If the amount of skewness is small, then the distribution is approximately symmetrical, as can be seen in Figure 3, M closely approximates the mean, μ_1, and a test of the median may reasonably be used as an approximate test of the mean. To study how well this approximation works, a small numerical study was carried out to compare the nominal and actual significance levels when the test of a median is used as a test of the mean when σ is unknown. The technical details are given in the Appendix. The numerical results are given below in Table 1 for a sample of size n = 10.

From Table 1, it is clear that if the value of σ is small, say less than 0.10, then the nominal and actual significance levels don't appear to differ a great deal. However, for σ in the neighborhood of 0.25, the actual level is about double the nominal level, and if σ is as large as 1.0, they differ by roughly an order of magnitude. This means that even for small values of σ, such as 0.2-0.25, a test based on the median will have a higher false positive rate than the pre-specified α for declaring a site to be clean when, in fact it is polluted. This discrepancy increases dramatically with an increase in σ. Assuming the resulting approximate test is deemed usable (at least for small values of σ), the sample size formula as given by equation (5) above (or using Kupper and Hafner,1989 adjustment) could be applied to determine an appropriate sample size.

As pointed out earlier, a lognormal distribution is often used for large values of σ such as those exceeding 0.75-1.0; therefore, the use of the median-based test for the mean of a lognormal distribution does not seem to be a feasible or desirable option. For values of σ smaller than 0.5, a normal distribution is often used and one needs not consider a lognormal distribution. Also, many exposure and risk assessment applications do not recommend the use of median based estimates of the exposure point concentration term (EPA 1992, 1996). Moreover, for skewed datasets with σ exceeding 0.5, the median is much smaller than the mean of the distribution. For example, when μ = 3.594 and σ = 1, the mean, μ_1, of the lognormal distribution is 60, and median, M, is 36.4. Thus a test based on equation (6) for testing the median does not provide an adequate approximate test of the lognormal mean when σ is larger than 0.5. Therefore, the power and the UCLs for the test based on median have not been plotted in the figures given in Section 4.

3.0 Computation of the UCL of the Mean

Modified t - test for asymmetrical populations

If σ^2 is unknown, then Johnson (1978) and Chen (1995) suggested the use of a modified t-statistic for testing the mean of a positively skewed distribution. Using Johnson's modified t - statistic, a $(1-\alpha)$100% one-sided UCL for the mean is given by

$$UCL = \bar{x} + \hat{\mu}_3/(6s_x^2 n) + t_{\alpha,n-1} s_x/\sqrt{n} \qquad (7)$$

Table 1. Actual significance levels for median test of the mean for a given nominal significance level.

	True Value of σ								
nominal a	0.01	0.05	0.10	0.15	0.20	0.25	0.50	1.0	2.0
0.05	0.051	0.058	0.067	0.077	0.088	0.100	0.181	0.427	0.897
0.10	0.103	0.114	0.129	0.146	0.164	0.183	0.294	0.589	0.957

where $t_{1-\alpha,\, n-1}$ is the $(1-\alpha)$th quantile of the Student's t distribution with $n-1$ df and the moment estimator of the third central moment used in equation (7) is given by

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3$$

The test rejects the hypothesis, H_0: $\mu_1 \geq (C_s)$ if the UCL < C_s. The UCL of the mean given by equation (7) has been used in the simulations discussed in Section 4. It has been observed that, for all practical purposes, the difference between the values of the UCL and the error rates (and, consequently the sample size) based upon the Student's t-statistic given by equation (6) and the modified t-statistic given by equation (7) is negligible. Therefore, in order to avoid the cluttering of graphs, the power and the UCL of the mean based upon the modified t-statistic have not been plotted in the figures given in Section 4.

H-statistics based UCL when the variance, σ^2, is unknown

Land (1971, 1975) derived the uniformly most powerful unbiased (UMPU) test for the mean, μ_1, of a lognormal distribution, which is based on the H-statistic. As mentioned earlier, EPA guidance documents (1980a,1992) recommend the use of the H-statistic based UCL of the mean for positively skewed distributions. The $(1-\alpha)100\%$ UCL for the mean, μ_1, based on the H-statistic is given as follows.

$$UCL = \exp(\bar{y} + 0.5 s_y^2 + s_y H_{1-\alpha} / \sqrt{(n-1)}) \qquad (8)$$

The critical values, $H_{1-\alpha}$, used in equation (8) are not readily available and are not available for n larger than 100. A subset of tables of critical values, $H_{1-\alpha}$, as computed by Land (1975) can be found in Gilbert (1987). Although it is shown by Land (1971, 1975) that the test based on the H-statistic has optimal properties, Singh, Singh, and Engelhardt (1997) point out that this procedure can sometimes lead to UCLs which are too large to be of any practical value. This is particularly true for samples of small sizes and large values of σ. For example, for σ in the interval 0.5-1.0, a sample of size 15 or less is considered small; for σ in the interval, 1-1.5, a sample of size 30 or less can be considered small, and so forth. This sample size requirement increases with σ. This is further discussed in the simulation Section 4.0.

Power and sample size for H-statistic based test

Despite the problems which accompany the lognormal distribution, if the user is **certain** that the population is lognormally distributed, then theoretically, a UCL based on the H-statistic is the optimal (although, it may not be realistic or practical) choice for achieving a pre-specified false positive error rate, α. For a lognormal distribution, no simple established sample size formula, such as given by equation (5), is available which can be used to determine the number of samples needed to test a lognormal mean with a pre-specified error limit, $\Delta = C_s - \mu_2$, where $\mu_2 \leq C_s$ (the symbol μ_1 is reserved as a general term for a lognormal mean in this section). Since a lognormal mean is not a location parameter, the error limit can not be handled simply in terms of the difference, $C_s - \mu_2$. Thus, the problem is to determine an adequate number of samples to collect to be able to detect when the lognormal mean approaches $C_s - \Delta$ with prespecified error rates, α and β. Several approximations are used to determine the sample size. One approximation is simply to use equation (5) on the raw data with the lognormal σ_1. Sometimes, the error limit of ln (Δ) is incorrectly used on log-transformed data. In the log-domain, the error limit is $\delta = \ln(C_s) - \ln(\mu_2)$. Another possibility would be to try, as an approximation, equation (5) with $d = \delta/\sigma$, with the error limit, δ, measured in the log-domain given as follows:

$$\delta = \ln(C_s) - \ln(C_s - \Delta) \qquad (9)$$

Stewart discussed a real application with $\Delta = 100$. In this discussion, a user wanted to incorrectly use ln(100) as the error limit in the log domain. As Stewart noted, this is incorrect because a difference of ln(Δ) in the log domain does not translate to a difference of Δ in the real domain. Two examples were considered with different values of C_s but with the same error factor, Δ. The resulting sample sizes for these two examples differed substantially, and it was concluded that whatever method of conversion to the log domain is used, it must take the value of C_s into account, and not just Δ.

To check the practicality and accuracy of these approximations, the power function of the

H-test is needed, the details for which are given in the Appendix. A search routine with this power function can be used to derive the sample size based on the H-statistic. Some examples illustrating these points follow.

Example 3. This example provides a comparison of power functions for H-tests based on three different sample sizes obtained using the approximation formula discussed above. The example is based on the simple choices: $C_s = 10$, $\alpha = 0.05$, $\beta = 0.10$, $\sigma = 1$, and for $\delta = \ln(C_s) - \ln(C_s - \Delta) = 1$, d = 1, the error limit in the real domain is relatively large, $\Delta = 6.32$. The population sd of the lognormal distribution in real space is $\sigma_1 = 13.11$. Note that the problem is to obtain an adequate number of samples to identify when the mean of the lognormal distribution starts to approach 3.68 given that the hypothesized mean is 10 ppm or more. The power functions are graphed in Figure 6. Using the incorrect error limit as $\ln(\Delta) = \ln(6.32) = 1.84$ in equation (5), one gets the erroneous sample size given by $n = [(1.645 + 1.282)(1)/1.84]^2 = 2.53$, which is rounded up to 3 (the power for this size is not plotted in Figure 6). Using the error limit of $\delta = 1$ in equation (5), the approximate sample size is $n = 9$, and the dashed curve in Figure 6 is the graph of the power function for an H-test based on this sample size. A search routine for the power function of the H-test as given in Appendix A was also used to determine the sample size, $n = 23$. This is the smallest integer such that the power ≥ $1 - \beta = 0.90$ when $\delta = 1$. Another possible approximation would be to apply equation (5) directly mean and sd, μ_1 and σ_1. That is, even though equation (5) is derived for use with a normal population, we plug the error limit and sd in the real domain, $\Delta = 6.32$ and $\sigma_1 = 13.11$ into (5). The result would be $n = [(1.645 + 1.282)(13.11)/6.32]^2 = 36.9$, which is rounded up to 37. While the number of samples for the approximation based on δ was too small, this approximation, as expected yields a value which is too large.

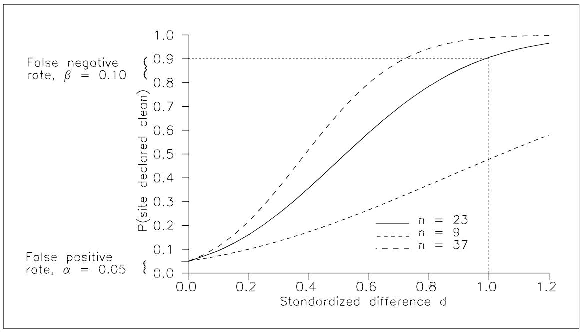


Figure 6. Comparison of power functions of H-test for three sample sizes: $n = 9$, 23 and 37.

From this discussion, one can conclude that using δ in equation (5) does not provide the best approximation because the sample size it yields is less than half of the one actually required and the power resulting from this approximation is 0.46, or roughly half of the nominal value of 0.90. However, it should be pointed out that the sample size came out to be 23 because the alternative mean selected is 3.68 ppm, which is much lower than the hypothesized mean of 10 ppm. In practice, many times, the alternative mean is closer to the hypothesized mean, in which case the number of samples needed to achieve a pre-specified power will increase dramatically. For example, if the alternative mean, μ_2, is 7, then the sample size based on the H-statistic will be 122; whereas, if $\mu_2 = 8$, then the sample size based on the H-statistic will become 286 to achieve a power of 0.9, which may not be practical! Also using δ in equation (5) for $\mu_2 = 7$, the approximate sample size is 67, and for $\mu_2 = 8$, the sample size is 172!

Example 4. A more practical example with overlapping densities and a higher value of σ is discussed. Let $C_s = 100$, $\alpha = 0.05$, $\beta = 0.10$, $\sigma = 1.5$, $\Delta = 100\text{-}80 = 20$ resulting in $\delta = \ln(C_s) - \ln(C_s - \Delta) = 0.2231$, with $d = \delta/\sigma = 0.1488$. The problem is to obtain an appropriate number of samples to distinguish between two overlapping log-normal distributions. As shown below, a large number of samples will be needed to meet these performance standards. Using the incorrect error limit as $\ln(\Delta) = \ln(20) = 2.9957$ in equation (5), the sample size comes out to be 2 (rounded from

2.15). Using $\delta = 0.2231$ as the error limit in equation (5), the sample size comes out be 387, which is fairly large. Next, using the search routine with the H-statistic based power, the sample size comes out to be 877! Taking about 877 samples to achieve a power of 0.9 is probably not practical. Another example from a Superfund site is considered next.

Example 5. This problem was encountered when working on a dataset from a Superfund site with benzo(a)pyrene equivalents (BaPE) being the main COPC. The data came from three areas of the site. Using the historical data, the sample mean and sd are 17.872 ppm and 40.41 ppm. The data do not follow a normal distribution but follow a lognormal distribution where the mean and sd of the log transformed data are 1.449 and 1.708, respectively. The CUG for the site is 60 ppm. The objective is to verify the attainment of the CUG by the three areas of the site. Enough samples are needed to be collected to be able to detect when the mean becomes 50 ppm (so that the areas can be considered clean) with a confidence coefficient of 0.95 and a power of 0.9.

Use of the incorrect error limit, $\ln(\Delta) = \ln(60\text{-}50) = \ln(10) = 2.303$ in equation (5) resulted in n = 4.71; and the consultants for the site suggested that 4.7 ~ 5 samples would be sufficient to verify the attainment of the cleanup standard with the pre-specified performance objectives, which is obviously incorrect. Using $\delta = \ln(C_s) - \ln(C_s - \Delta)$ = ln (60) - ln(50) = 0.182 as the error limit in equation (5) with $\sigma = 1.7$, the sample size comes out to be 752. Next using the search routine with the H-statistic based power, the sample size comes out to be 1808! Apparently taking 752 or 1808 samples to achieve a power of 0.9 is neither practical nor desirable. The influence of the sample size on the power needs to be determined. Is it really worth taking a large number of samples such as 1808 in an effort to achieve the high power of 0.9? *From a practical point of view, one needs to know how to compute the UCL of the mean correctly with maximum practical power for a given sample size and level of significance*.

Obviously, there is no simple solution to the problem. It is observed that serious underestimation of the sample size can occur when $\ln(\Delta)$ is substituted for the error margin in equation (5). However, the error limit, δ, based sample size determination procedure can also lead to a substantial underestimation of the sample size for larger values of δ, especially when C_s is much larger than $C_s - \Delta$, as is the case in Example 3. For large values of sd, the sample size obtained using the normal distribution based power or the H-statistic based power can be unreasonably large to be of practical merit. The question is, can we use approximate tests based upon the t-test, the CLT, or the Chebychev theorem to meet (approximately) the pre-specified performance standards? Does there exist a UCL computation procedure (or model) which yields practical values for the sample size and the UCL? Does there exist a procedure which has higher power (that may not be equal to 0.9, or some other pre-specified level) than the other procedures for a given value of n, the sample size? In order to investigate the power behavior of the various procedures, a Monte Carlo simulation study has been performed which is discussed in Section 4.

Chebychev Inequality based UCL of the mean

The Chebychev theorem as discussed by Singh, Singh, and Engelhardt (1997) to obtain a conservative estimate of the UCL of the mean of a lognormal distribution has also been included in this study. The two-sided Chebychev theorem states that, given a random variable, *X, following any distribution, continuous or discrete,* with a finite mean, μ_1, and a sd, σ_1, we have:

$$\mathrm{P}\left(-k\sigma_1 \le X - \mu_1 \le k\sigma_1\right) \ge 1 - (1/k^2) \tag{10}$$

This result can be applied with the sample mean, $\bar{x}$, to obtain a conservative UCL for the population mean. Specifically, if the right side of equation (10) is equated to 0.95, then $k = 4.47$, and $UCL = \bar{x} + 4.47\sigma_1/n^{1/2}$ is a conservative 95% upper confidence limit for the population mean. Of course, this would require the user to know the value of σ_1. The obvious modification would be to replace σ_1 with an estimate, such as the sample standard deviation, s_x, or the Minimum Variance Unbiased Estimate (MVUE) based upon lognormal theory; but, since this is estimated from data, the result is no longer guaranteed to be conservative.

In general, if $\hat{\mu}_1$ denotes an estimate of the unknown mean, μ_1, and $\hat{\sigma}(\hat{\mu}_1)$ is an estimate of the standard error of $\hat{\mu}_1$, then the quantity $UCL = \hat{\mu}_1 + 4.47\hat{\sigma}(\hat{\mu}_1)$ will give a 95% UCL for μ_1, which should tend to be conservative. The power of the test based on this Chebychev bound has been included in the simulation experiments as well. The Chebychev UCL can be computed using the estimates of mean and sd as given by equations (1) and (2) above (denoted by Cheb_M) or can be obtained using the MVUE given by equations (8) and (10) of Singh, Singh and Engelhardt (1997) based on the lognormal distribution theory (denoted by Cheb_MV) to obtain the estimates $\hat{\mu}_1$ and $\hat{\sigma}(\hat{\mu}_1)$. The power of the test based on the Chebychev UCL has been computed using both sets of estimates, Cheb_M and Cheb_MV.

From the simulation experiments discussed in Section 4.0 below, it is observed that for all sample sizes, the Chebychev - UCL is more conservative (higher) than all other UCLs except for the UCL based on the H-statistic. The H-UCL tends to be much larger than the Chebychev-UCL for large values of sd ($\geq$1.5) and samples of small sizes (n$\leq$30), and as sd increases, this sample size requirement becomes larger than 30. However, it is also observed that the H-UCL based upon samples of small sizes from populations with large values of σ, tend to be unrealistically large (Singh, Singh, Engelhardt, 1997) to be of any practical use. Inference based upon such large H-UCL values would result in a large number of false negatives as can be seen from Figures 3A-3C, 4A-4D, 11A-11C, and 12A-12D.

4.0 Monte Carlo Simulation Experiments

From the discussion presented here, it is obvious that there does not exist a single sample size determination formula which can be preferred over the other formulas while sampling from a lognormal population. Therefore, extensive Monte Carlo simulation experiments were carried out and several procedures to compute the 95% UCL of the mean with a test size of 0.05 have been compared in terms of their power. In order to study the power of the various UCL computation methods, 10,000 samples of sizes 10, 15, 20, 30, 50, and 100 were generated from a variety of lognormal populations for various values of the mean above and below the CUG value. The simulations have been performed for two CUG values:10 and 60. Several combinations of mean and sd have been considered. For CUG = 10, the samples were generated from lognormal populations with means of 15, 12, 10, 8.2, 6.7, and 3.7, and for CUG = 60 ppm, samples were generated from lognormal populations with means of 100, 70, 60, 50, 40, and 30. The sd values were 0.5,1.0,1.5, and 2.0. Higher sample sizes including,150, 200, 300, 400, have been also tried for larger values of σ such as 1.5, and 2.0. The H-UCL and power could not be computed for the larger sample sizes as the critical values for the H-statistic are not available for samples of sizes larger than 100. Therefore, graphs for values of n larger than 100 are not included here.

Procedures considered in the simulation experiments include the Students's t-test, modified t-test for asymmetric distributions, the CLT based normal test, a test based on the median, a test based on the H-statistic, and tests based on the Chebychev theorem. A close look at the simulation results suggests that the differences in the values (power, UCL) obtained using the Student's t-test and the modified t-test are not significant. The UCLs based on the modified t-test are slightly higher than those based on Student's t-test. For example, for a sample of size 10 and the population mean of 100, the 95% UCL of the mean using the Student's t-test are 128.77 (σ = 0.5), 157.79 (σ = 1.0), 190.01 (σ = 1.5), 223.27 (σ = 2.0), and 242.70 (σ = 2.5), whereas the corresponding 95% UCL using the modified t-test are 129.39 (σ = 0.5), 159.99 (σ = 1.0), 194.45 (σ = 1.5), 230.24 (σ = 2.0), and 251.61 (σ = 2.5), respectively. Also, for small samples, the t-test results are preferred over the CLT results, and the t-test results approach the CLT results as the sample size increases. As observed above, the test for the mean based on the median does not work well for values of σ larger than 0.5. Therefore, in order to avoid cluttering the graphs, the computations based upon the CLT, modified t-test, and the median based test are not included in the graphs presented in this paper. The power and the UCLs are plotted only for four methods, namely, the t-test (denoted by t), H-statistic (denoted by H), the Chebychev UCL using the simple sample arithmetic mean and sd (denoted by Cheb_M), and the Chebychev UCL based on the MVU estimates obtained using the lognormal theory (denoted by Cheb_MV).

Graphs for the power and the UCL of the mean for the four procedures have been plotted. For CUG = 10, and σ = 0.5, 1.0, 1.5, and 2.0, the power function is displayed in Figures 1A-1F, 2A-2F, 3A-3F, and 4A-4F, respectively, and for CUG = 60 ppm, and σ = 0.5, 1.0, 1.5, and 2.0, the power function is given in Figures 9A-9F, 10A-10F, 11A-11F, and 12A-12F, respectively. For each sample size, the plotted UCLs are the average values of the respective UCLs over 10,000 iterations for each combination of mean and sd. For CUG = 10, and σ = 0.5, 1.0, 1.5, and 2.0, the UCLs are displayed in Figures 5A-5F, 6A-6F, 7A-7F, and 8A-8F, respectively, and for CUG = 60 ppm, and σ = 0.5, 1.0, 1.5, and 2.0, the UCLs are given in Figures 13A-13F, 14A-14F, 15A-15F, and 16A-16F, respectively. From these graphs, the following observations can be been made.

1. From these graphs, it is clear that the H-statistic based test does possess the pre-specified size of 0.05 (significance level) for all values of n and σ.

2. The size of the t-test is larger than the other three procedures. As the sample size increases, the differences between the H-statistic and the t-test based results decrease.

3. From Figures 1A-1F, 2A-2F, 3A-3F, 9A-9F, 10A-10F, and 11A-11F, it is observed that, for $\sigma \leq 1.5$, the test based on the Chebychev bound has the realized test size (level of significance) smaller than the pre-specified test size of 0.05 for samples of all sizes considered (except for n = 10 and σ = 1.5), which is a desirable property for the Chebychev UCL to possess. This is especially true for the Chebychev results based on the MVU estimates of the mean and sd of the lognormal distribution.

4. For small values of σ ($\leq$0.5), the power and the UCL values based upon the t-test and the H-statistic are quite close, even for samples of a size as small as 15 for both CUG values as can be seen in Figures 1A-1F, 5A-5F, 9A-9F, and 13A-13F. The discrepancy between the UCLs and power based on the t-test and the H-statistic decreases as the sample size increases.

5. From Figures 2A-2C, 6B-6C, 10A-10C, and 14B-14C, it is observed that for $\sigma \approx 1$ and samples of a size smaller than 20, there is not much difference in the power and the 95% UCL of the mean based on the H-test and the Cheb-UCL tests. Actually, the power of H-test is smaller than the Cheb-UCL for samples of a size smaller than 10.

6. For σ exceeding 1.5, the size of the Cheb-MV test becomes larger than the pre-specified level of significance, 0.05 for samples of size 30 or smaller, as can be seen in Figures 4A-4C and 12A-12C. As the sample size increases, the size of the test based on the Cheb-MV comes close to the pre-specified size of 0.05 and then becomes smaller than prespecified size as can be seen in Figures 4A- 4F and 12A-12F. A similar pattern will be observed for larger values of σ and the sample size requirement for the size of the Cheb-MV test to reach 0.05 will also increase.

7. From Figures 3A-3C, 4A-4D, 11A-11C, and 12A-12D, it can be seen that for $\sigma \geq$ 1.5, the test based on the H-statistic yields powers smaller than the Cheb-MV test for samples of sizes smaller than 30, which will result in a large number of false negatives. A similar pattern will be observed for larger values of σ and the sample size requirement for the power of the H-test to reach the power of the Cheb-MV test will also increase.

8. As the sample size increases, the power based on the Chebychev UCL decreases. For small sample sizes, the power comes quite close to the power based on the H-statistics and then

as the sample size increases, the power becomes smaller than the power for the H-test. However, these decrements are not so dramatic as can be seen in Figures 4A-4F and 12A-12F. Also, as the sample size increases, the Cheb-MV UCL becomes larger than the H-UCL. However, those increases are consistent without any dramatic changes. A similar pattern will be observed for larger values of σ and the sample size requirement for the power of the Cheb-MV test to reach the power of the H-test will also increase.

9. For values of σ larger than 1, the H-statistic based UCL becomes unrealistically large, at least for samples of sizes smaller than 30 for both values of the CUG, as can be seen in Figures 7A-7D, 8A-8D, 15A-15D, and 16A-16D. Cleanup decisions based on such unreasonably large H-UCL values cannot be considered reliable. For example, when μ_1 = 60 and σ = 2.0, the 95% quantile of this lognormal distribution is about 218, whereas the 95% UCLs of the mean are about 1300010 and 1488 for samples of sizes 10 and 20, respectively (Figures 16A and 16C). A similar pattern will be observed for larger sample sizes as σ increases. For example, for σ = 2.0, the H-UCL is much higher than the rest of the UCLs even for samples of size 50, as can be seen in Figures 8A-8D and 16A-16D. For μ_1 = 60 and σ = 2.0, the 95% Cheb-MV UCL of mean are 244, 228, and 222 for samples of size 10, 15, and 20, respectively.

5.0 Summary and Recommendations

The discussion presented here leads to the conclusion that there does not exist a single sample size determination formula which can be preferred over the other formulae while sampling from a lognormal population. From the simulation results, it appears that it is not feasible to achieve the desired error rates and critical difference without taking an enormous number of samples. This is especially true when a lognormal model is assumed. Keeping these practical considerations in mind, the regulators may have to settle for reduced values of performance standards.

From equation (9), it is concluded that, even though the H-UCL based test is the test achieving the pre-specified Type I error rate, for samples of small size, the associated H-UCL is unreasonably large, even when samples are obtained from a lognormal distribution. For example, for σ = 2.0, n = 15, and mean = 60, the H-UCL of the mean is 6493 (Fig.16B), which is an unlikely event to happen at a Superfund site. For highly skewed populations, a large number of samples is needed to obtain a UCL of practical value.

It is concluded that for less skewed data with $\sigma \leq 0.5$, a test based on normal distribution (t-test) may be used to determine the sample size and power, and consequently one can use the t-test based UCL of the mean to verify the attainment of cleanup standards. For σ in the interval (0.5-1.0), and for samples of a size smaller than 30, the Chebychev bound gives reasonable and reliable results in terms of power and the UCL of the mean; as the sample size becomes larger than 30, one can use the test based on the t-statistic. For samples of small sizes (less than 50) and σ in the interval (1.0-1.5), the H-UCL of the mean becomes large, and the Cheb-MV UCL can be used to verify the cleanup standard; for samples of large sizes (greater than 50), the central limit theorem can be used to compute the UCL of the mean. For samples of small sizes (less than 100) and σ in the interval (1.5-2.0), the H-UCL of the mean becomes too large, and the Cheb-MV UCL can be used to verify the cleanup standard (with higher Type I error rates); for samples of large sizes (greater than 100) due to the central limit theorem, one can compute the UCL of the mean based on the normal theory. Similar patterns will be observed as the sd increases.

Based on the Monte Carlo simulation results and the authors' experience with Superfund site work, the following recommendations are made.

1. A multi-phase approach may be used when the sample size formulas discussed here result in an unrealistically high number of samples (e.g., exceeding 100-200) needed to achieve the desired performance parameters. One can start with taking a reasonable and economically possible number of samples. The power and size of the tests can then be

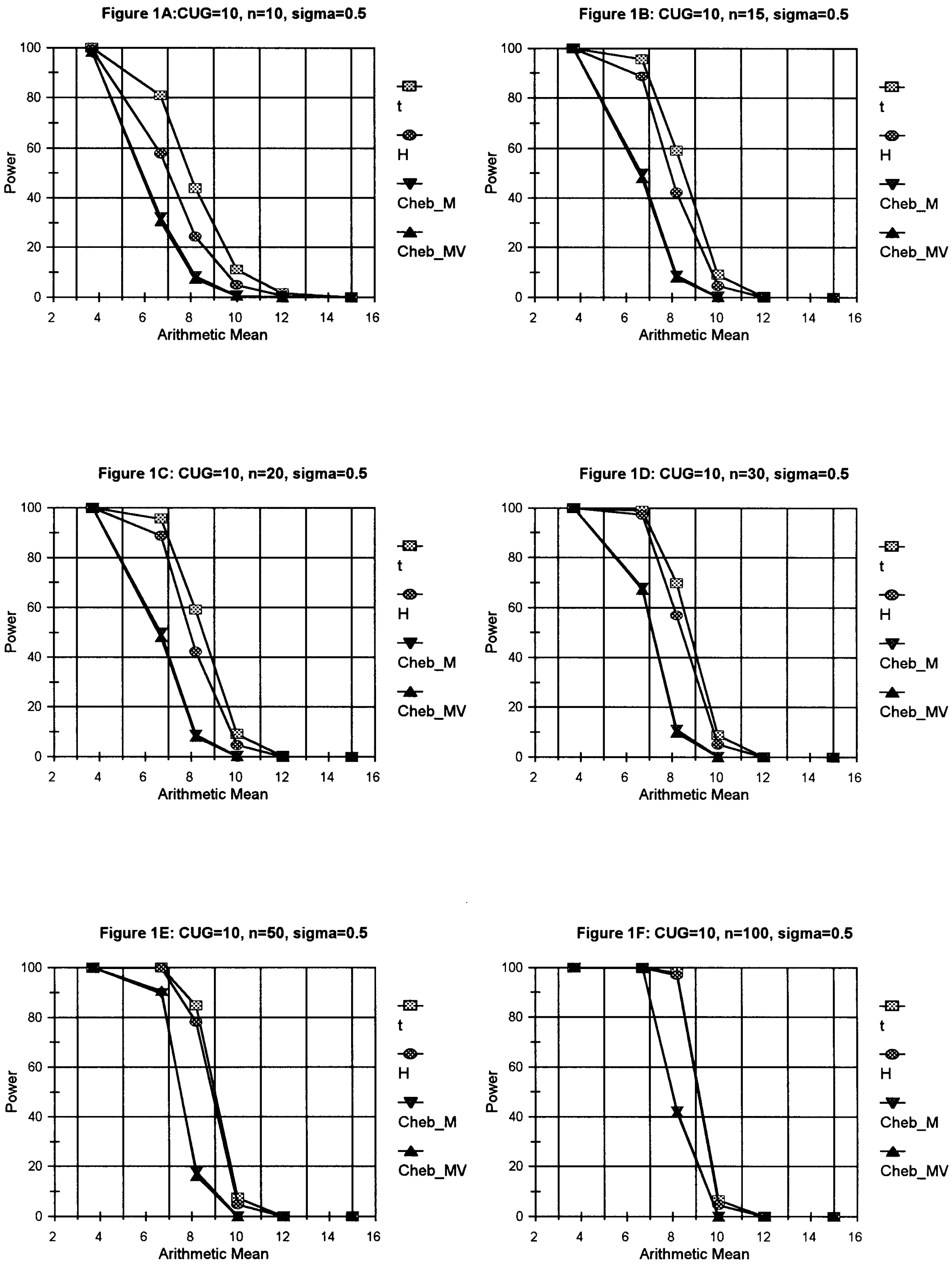
computed. A procedure with maximum power may be chosen to compute the UCL of the mean. If deemed necessary for increased power, more samples can be taken in the next phases and the UCL of the mean can be re-computed using the procedure yielding the maximum power.

2. Low values of error rates, α, and β, and the error margin, Δ, result in large sample sizes which, in reality, may not be achievable. For example, for a lognormal distribution, with $\alpha = 0.05$, $\beta = 0.1$, sd = 1.7, and $\Delta = 10$, the sample size needed is 1808, as seen in Example 5 above, which in not feasible to collect. For a sample of size 100, and sd = 1.5 and 2.0, and $\Delta = 10$, the size of the H-UCL test is about 0.05 but the powers are only about 0.20 and 0.17 (see Figures 11F, 12F), respectively. If possible, one should consider reducing the performance objectives.

3. It is recommended to avoid the use of the lognormal distribution. The appropriate use of the lognormal distribution is not clear to most people and it can very easily be used incorrectly, which may lead to incorrect conclusions.

4. For highly skewed populations, the arithmetic mean becomes larger than higher quantiles of the distribution, such as 90% and 95% (e.g., σ exceeds 3.26), etc. Other measures of central tendency, such as the median, or some other quantile (e.g., 80%, 90%) need to be considered for highly skewed datasets for the verification of the achievement cleanup goals.

It is crucial that great care should be exercised in choosing an appropriate model and in understanding the potential problems associated with the chosen model when attempting to make decisions about a population mean. It is also recommended that some additional Monte Carlo simulations be done to assess the performance of the various methods for a variety of skewed population distributions, such as the Weibull and the Gamma, and of the sort common with contamination data (e.g., mixtures and with outliers).



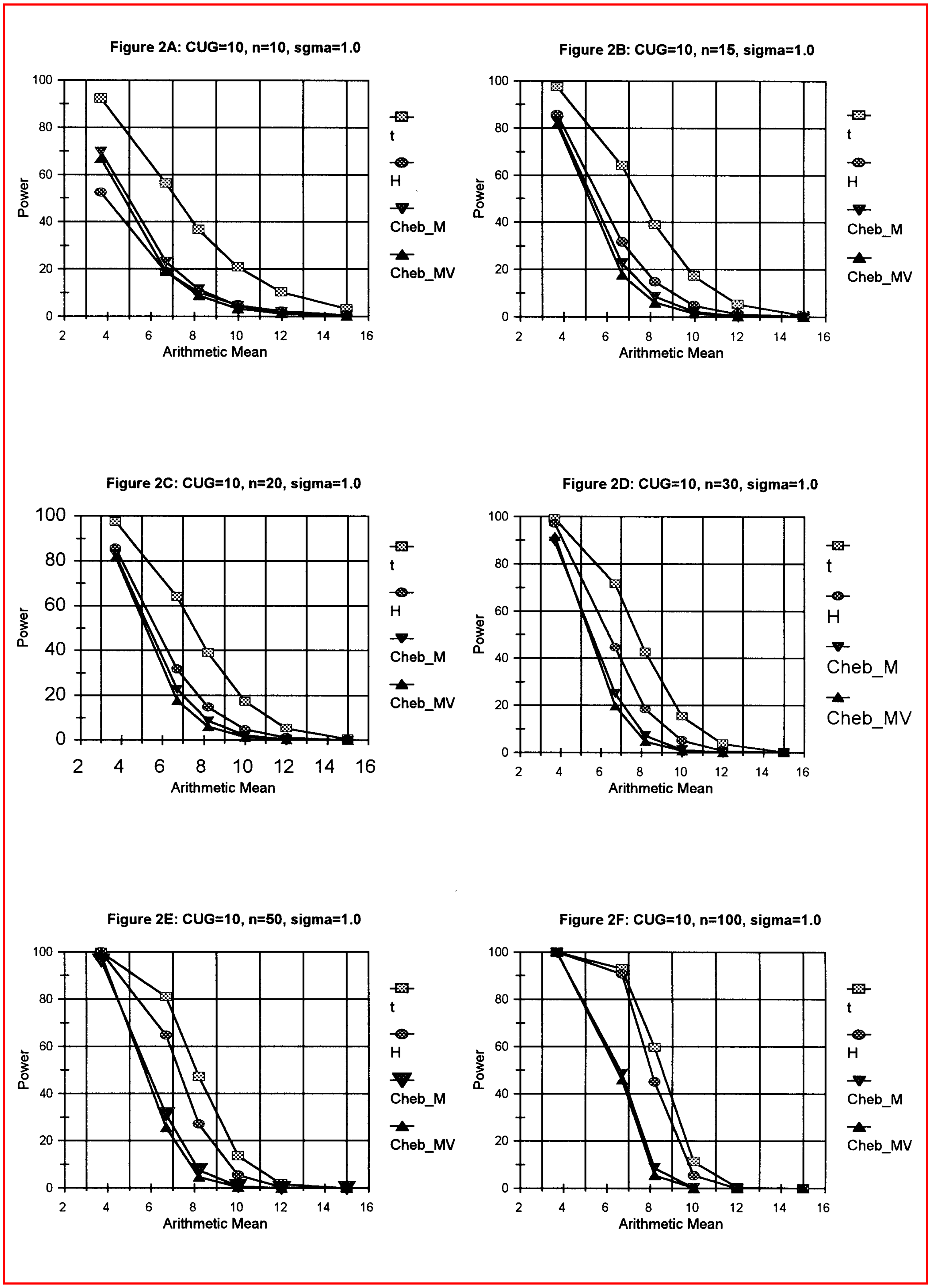


Figure 2A: CUG=10, n=10, sgma=1.0

Figure 2B: CUG=10, n=15, sigma=1.0

Figure 2C: CUG=10, n=20, sigma=1.0

Figure 2D: CUG=10, n=30, sigma=1.0

Figure 2E: CUG=10, n=50, sigma=1.0

Figure 2F: CUG=10, n=100, sigma=1.0

Figures 2A - 2F

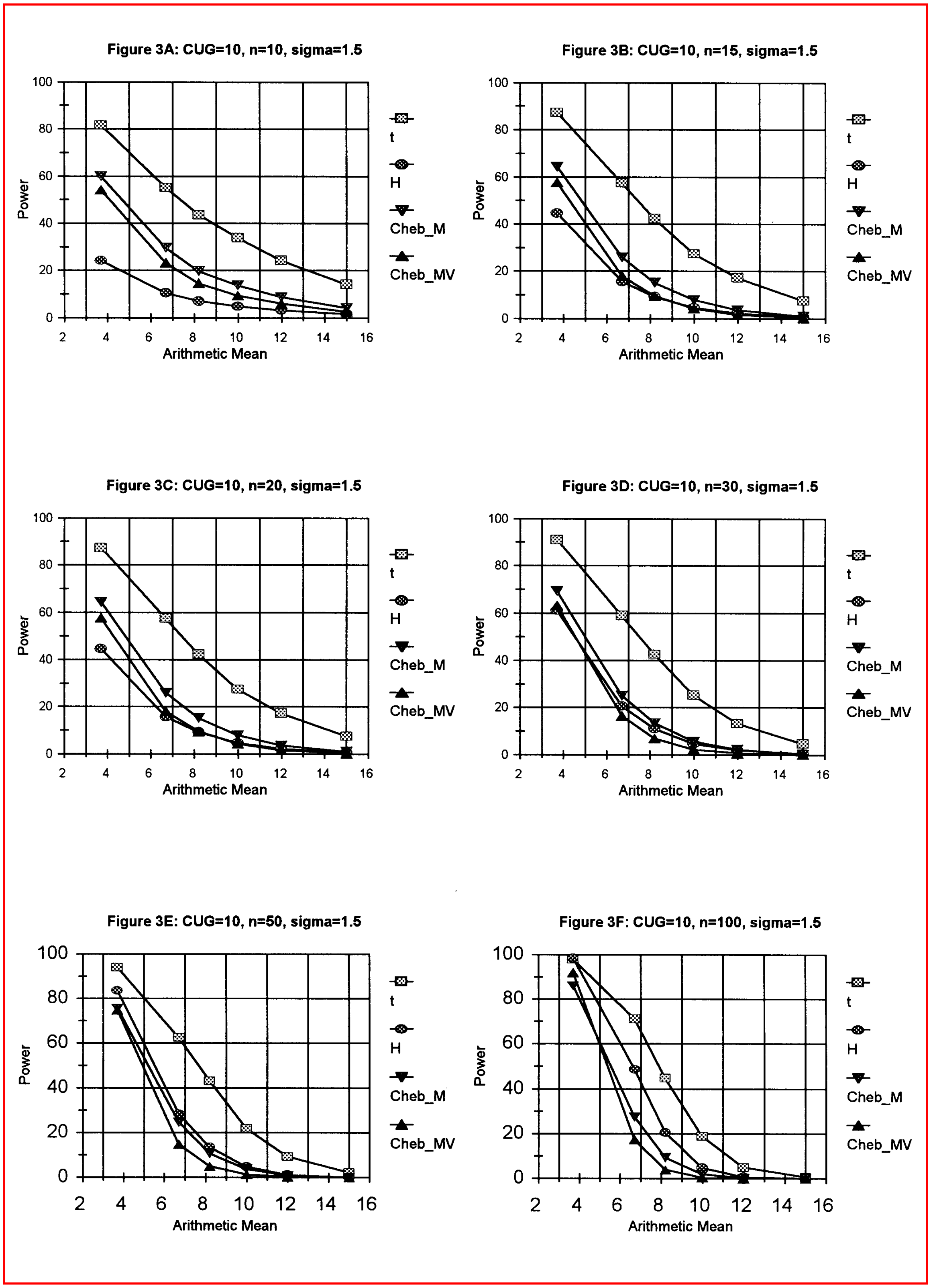


Figure 3A: CUG=10, n=10, sigma=1.5

Figure 3B: CUG=10, n=15, sigma=1.5

Figure 3C: CUG=10, n=20, sigma=1.5

Figure 3D: CUG=10, n=30, sigma=1.5

Figure 3E: CUG=10, n=50, sigma=1.5

Figure 3F: CUG=10, n=100, sigma=1.5

Figures 3A - 3F

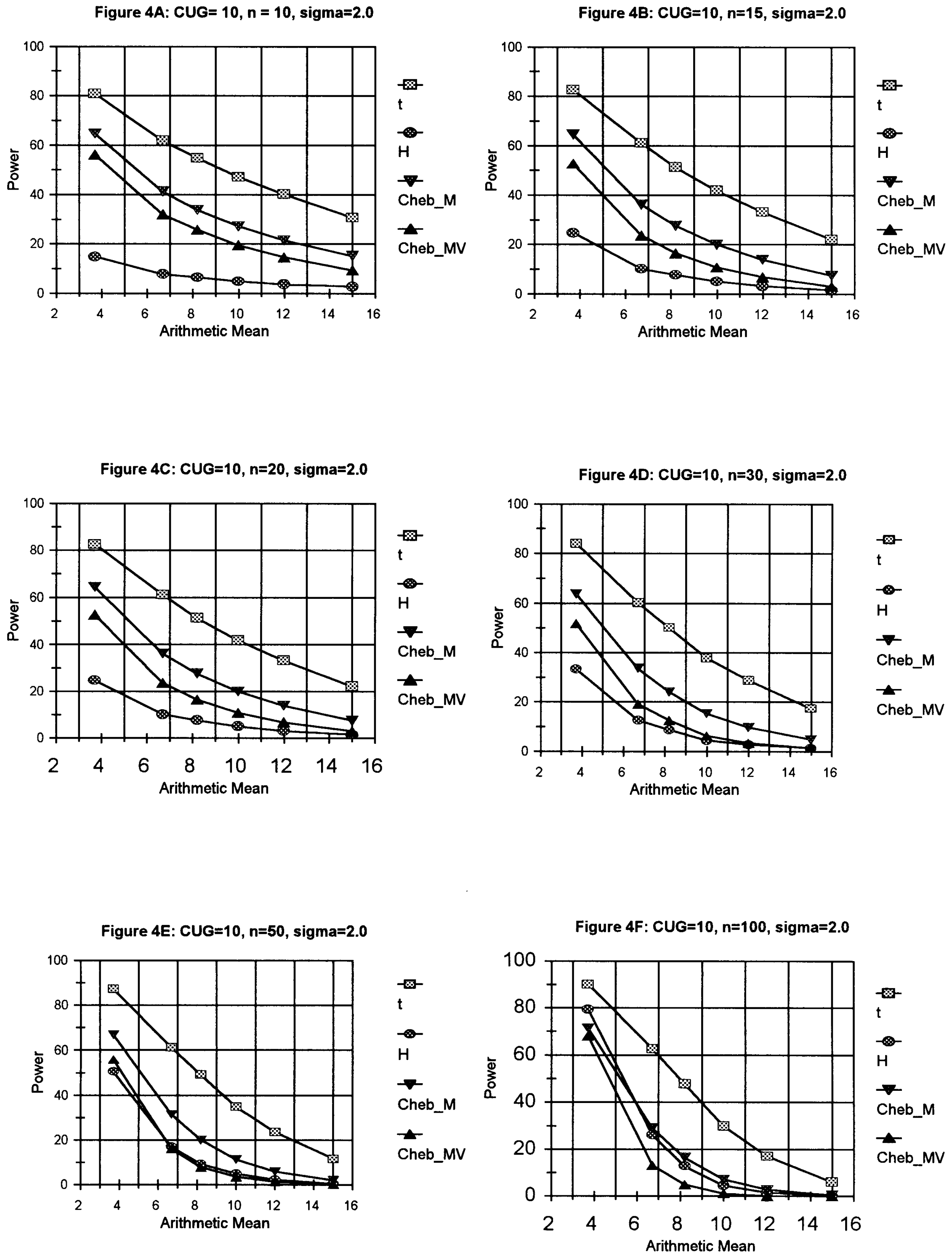


Figure 4A: CUG= 10, n = 10, sigma=2.0

Figure 4B: CUG=10, n=15, sigma=2.0

Figure 4C: CUG=10, n=20, sigma=2.0

Figure 4D: CUG=10, n=30, sigma=2.0

Figure 4E: CUG=10, n=50, sigma=2.0

Figure 4F: CUG=10, n=100, sigma=2.0

Figures 4A - 4F

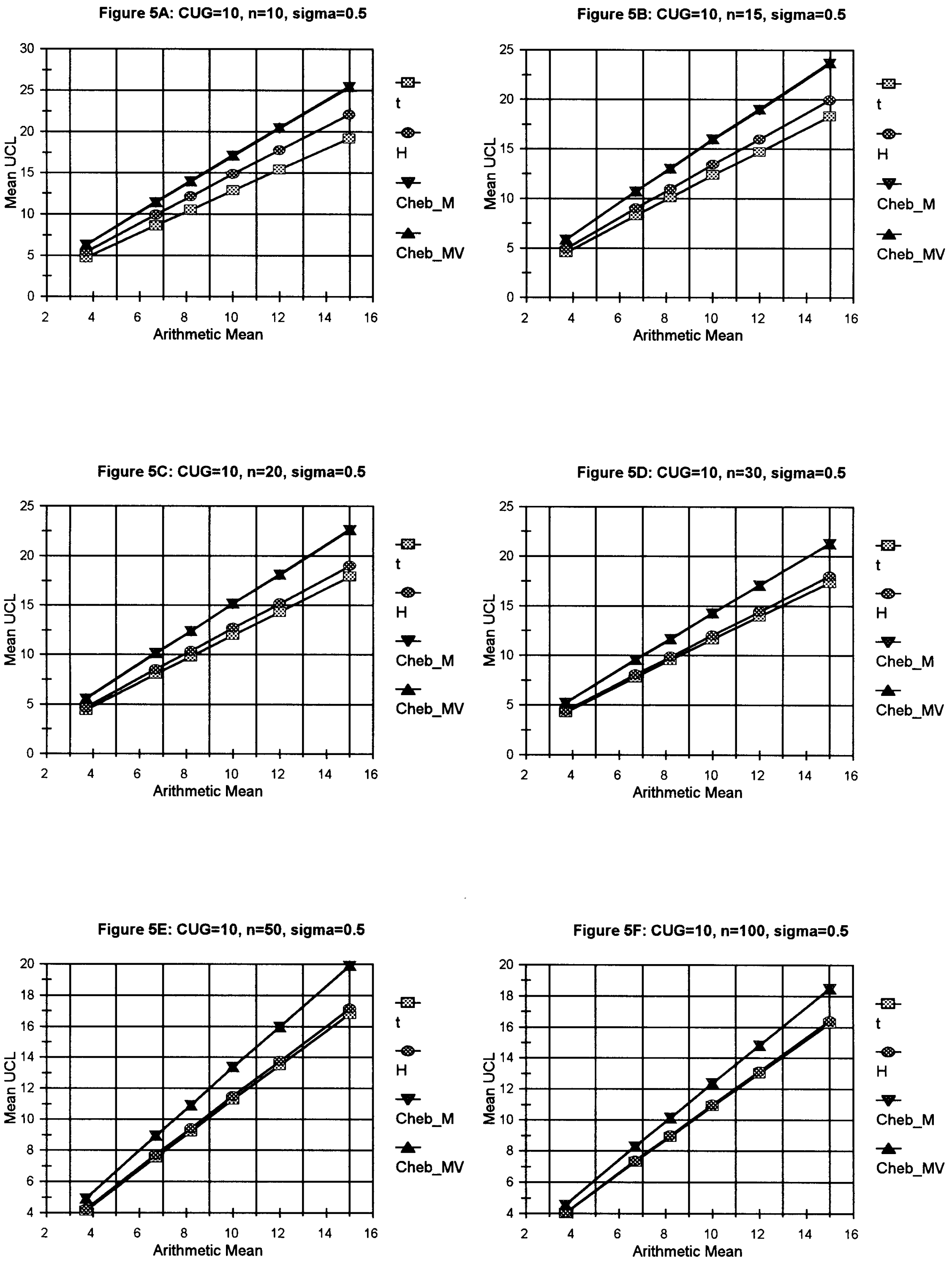


Figure 5A: CUG=10, n=10, sigma=0.5

Figure 5B: CUG=10, n=15, sigma=0.5

Figure 5C: CUG=10, n=20, sigma=0.5

Figure 5D: CUG=10, n=30, sigma=0.5

Figure 5E: CUG=10, n=50, sigma=0.5

Figure 5F: CUG=10, n=100, sigma=0.5

Figures 5A - 5F

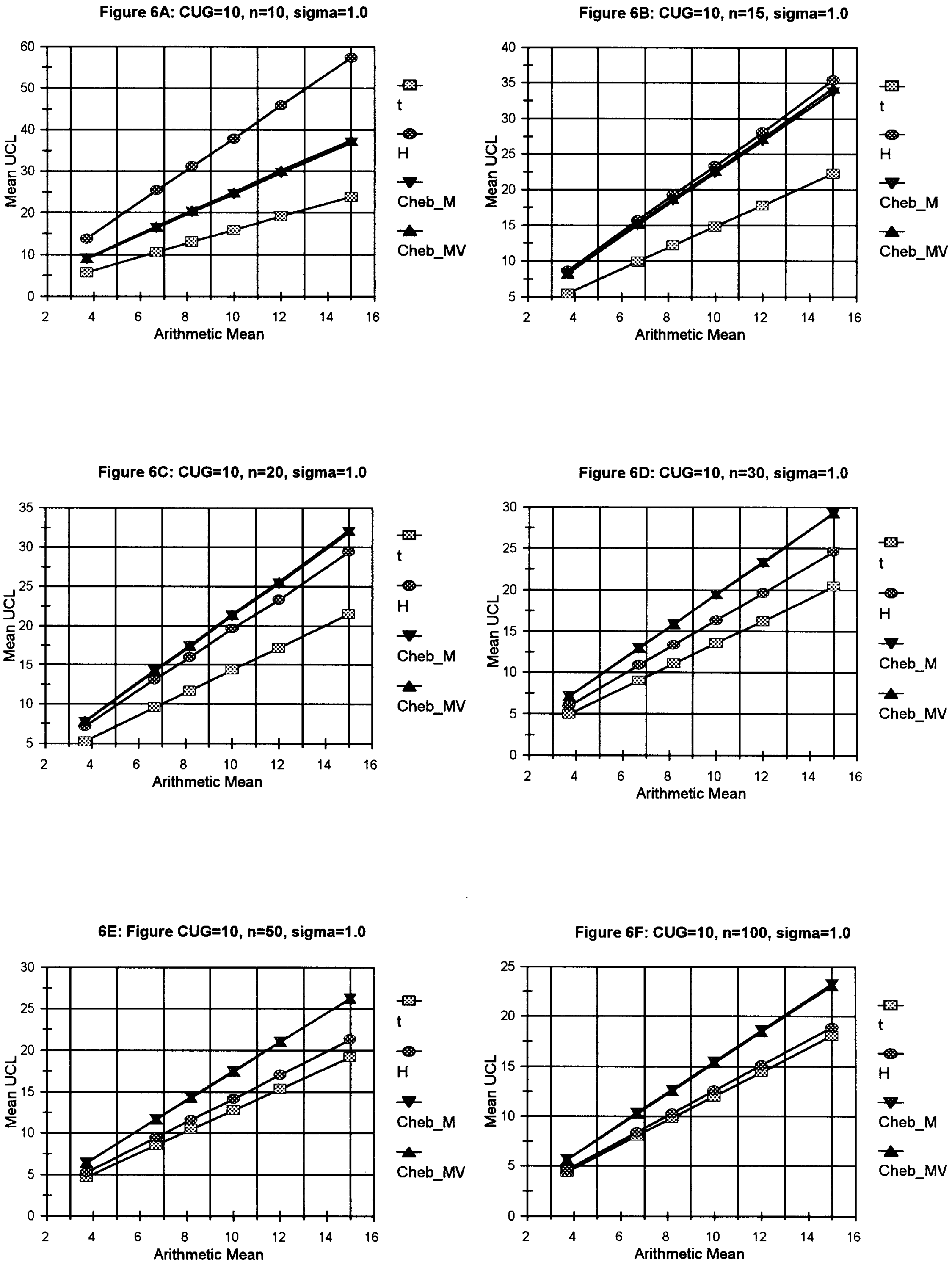


Figure 6A: CUG=10, n=10, sigma=1.0

Figure 6B: CUG=10, n=15, sigma=1.0

Figure 6C: CUG=10, n=20, sigma=1.0

Figure 6D: CUG=10, n=30, sigma=1.0

6E: Figure CUG=10, n=50, sigma=1.0

Figure 6F: CUG=10, n=100, sigma=1.0

Figures 6A - 6F

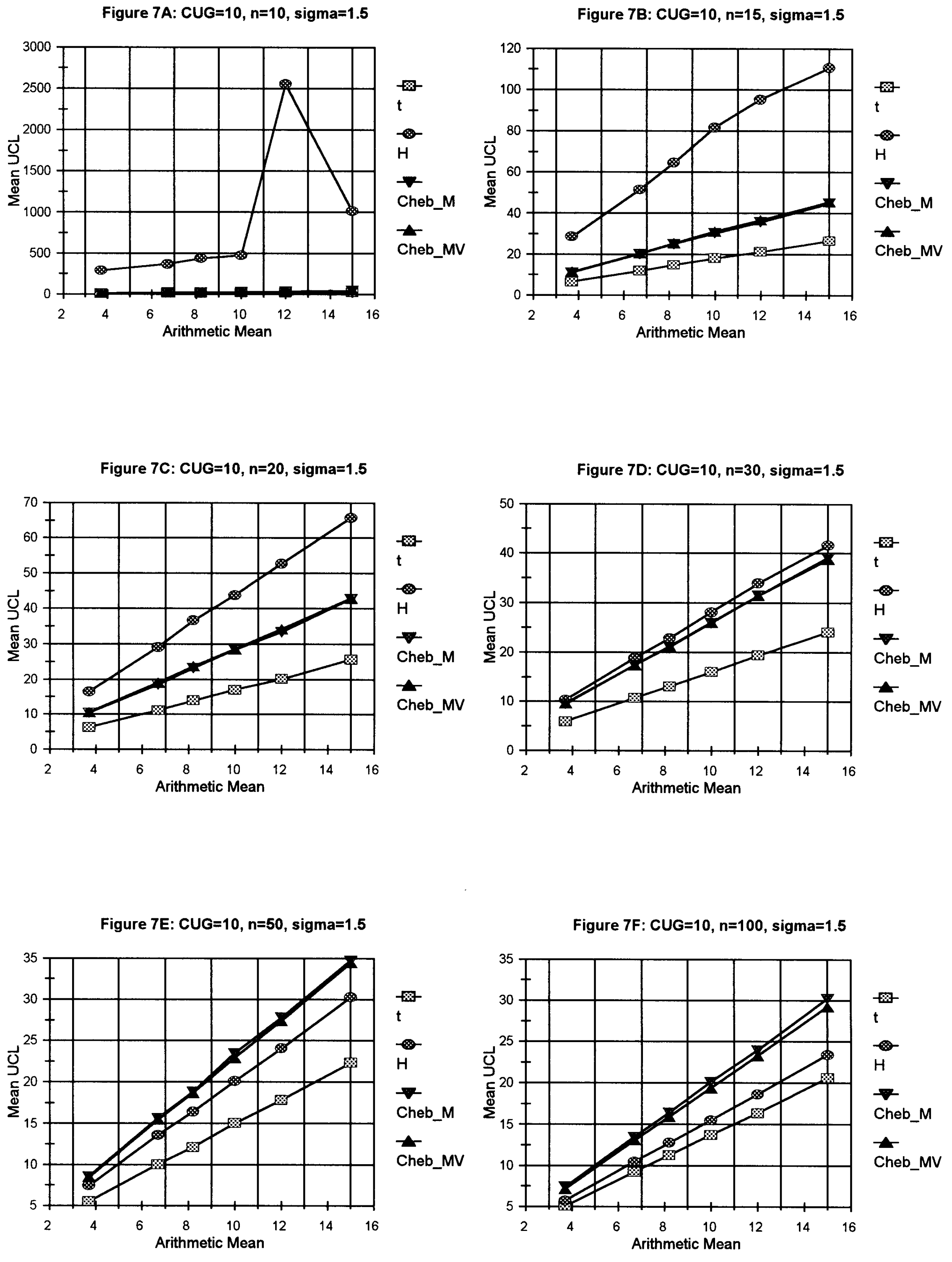


Figure 7A: CUG=10, n=10, sigma=1.5

Figure 7B: CUG=10, n=15, sigma=1.5

Figure 7C: CUG=10, n=20, sigma=1.5

Figure 7D: CUG=10, n=30, sigma=1.5

Figure 7E: CUG=10, n=50, sigma=1.5

Figure 7F: CUG=10, n=100, sigma=1.5

Figures 7A - 7F

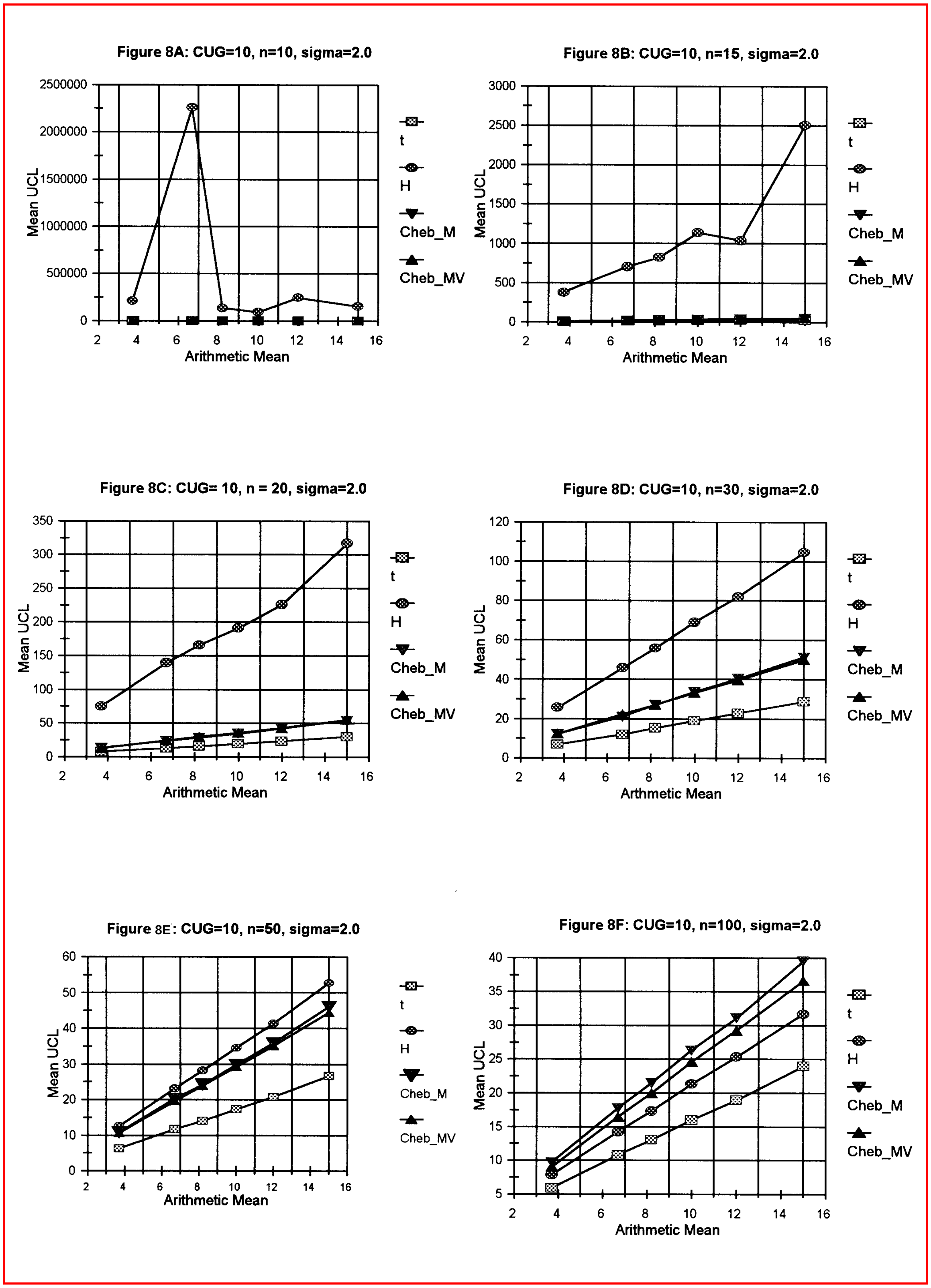


Figure 8A: CUG=10, n=10, sigma=2.0

Figure 8B: CUG=10, n=15, sigma=2.0

Figure 8C: CUG= 10, n = 20, sigma=2.0

Figure 8D: CUG=10, n=30, sigma=2.0

Figure 8E: CUG=10, n=50, sigma=2.0

Figure 8F: CUG=10, n=100, sigma=2.0

Figures 8A - 8F

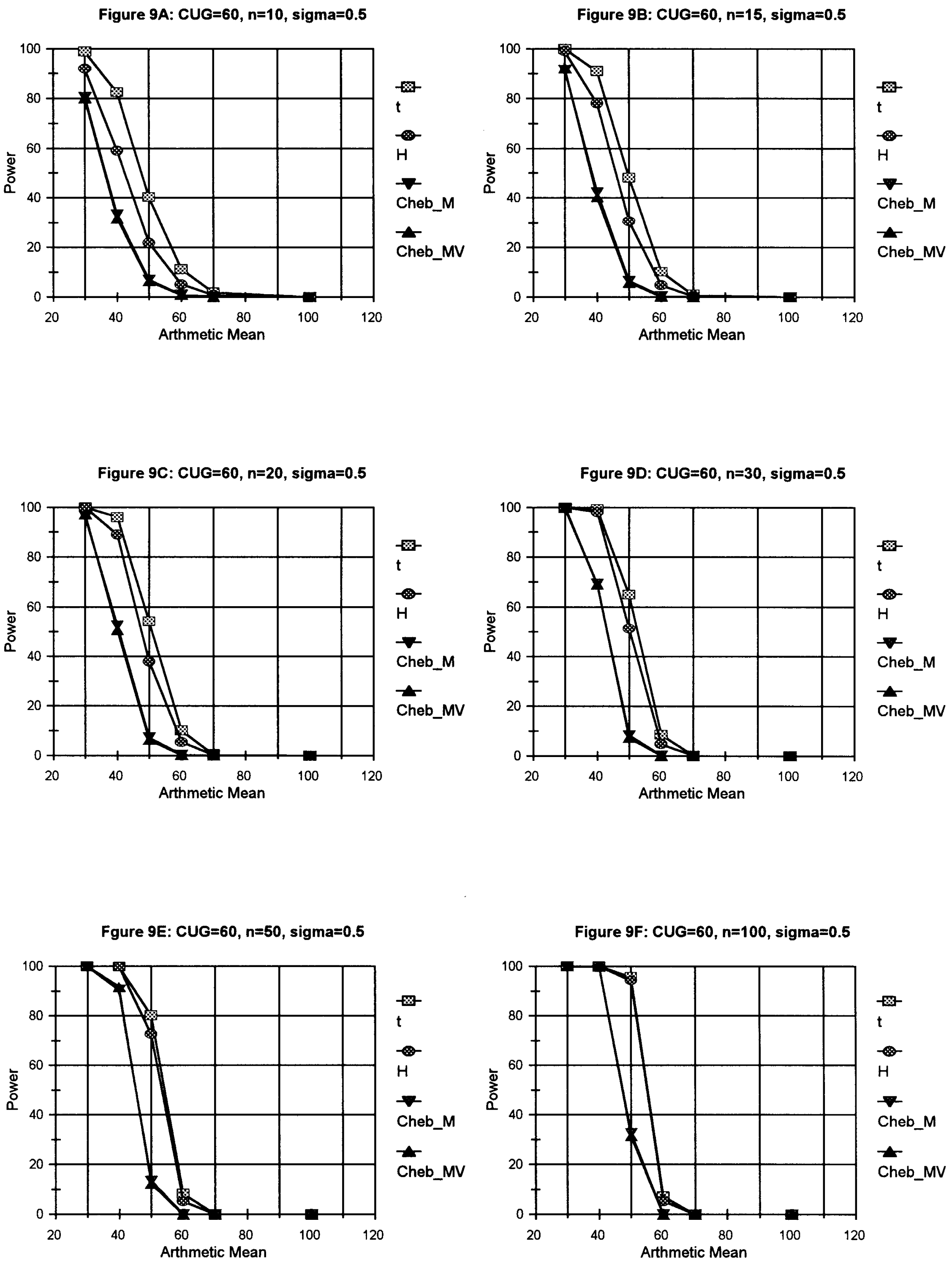


Figure 9A: CUG=60, n=10, sigma=0.5

Figure 9B: CUG=60, n=15, sigma=0.5

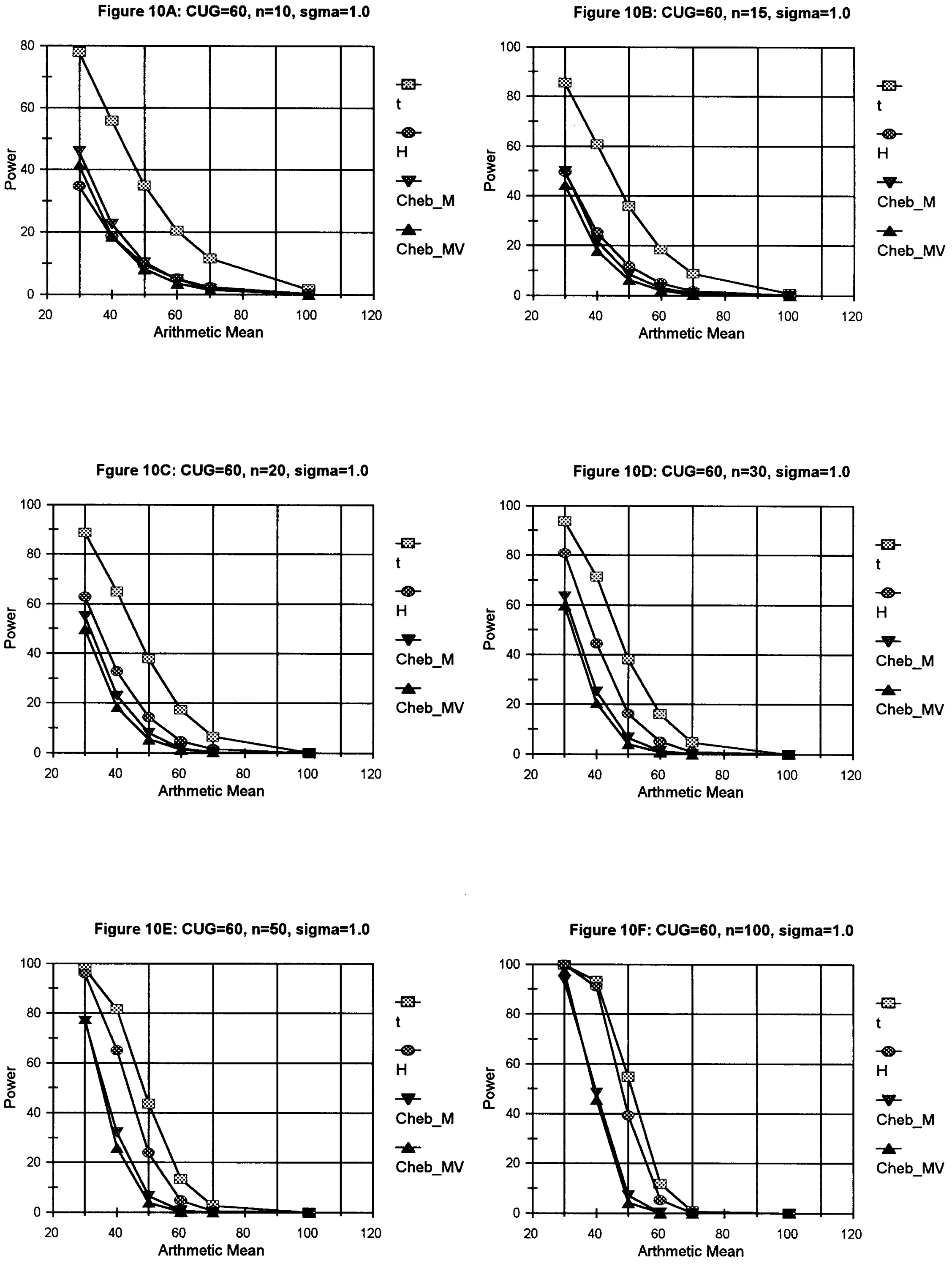
Figure 9C: CUG=60, n=20, sigma=0.5

Fgure 9D: CUG=60, n=30, sigma=0.5

Fgure 9E: CUG=60, n=50, sigma=0.5

Figure 9F: CUG=60, n=100, sigma=0.5

Figures 9A - 9F



Figures 10A - 10F

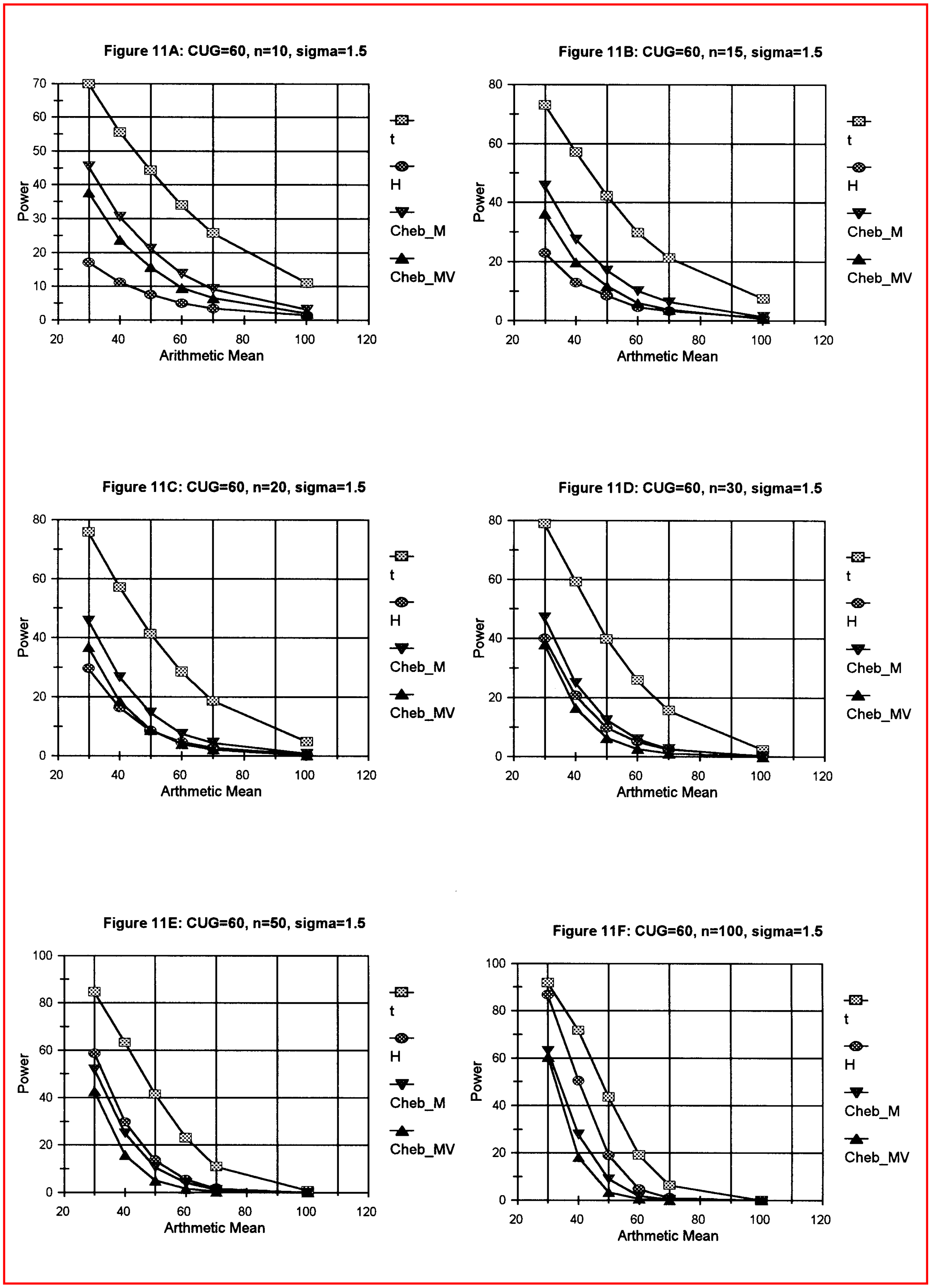


Figure 11A: CUG=60, n=10, sigma=1.5

Figure 11B: CUG=60, n=15, sigma=1.5

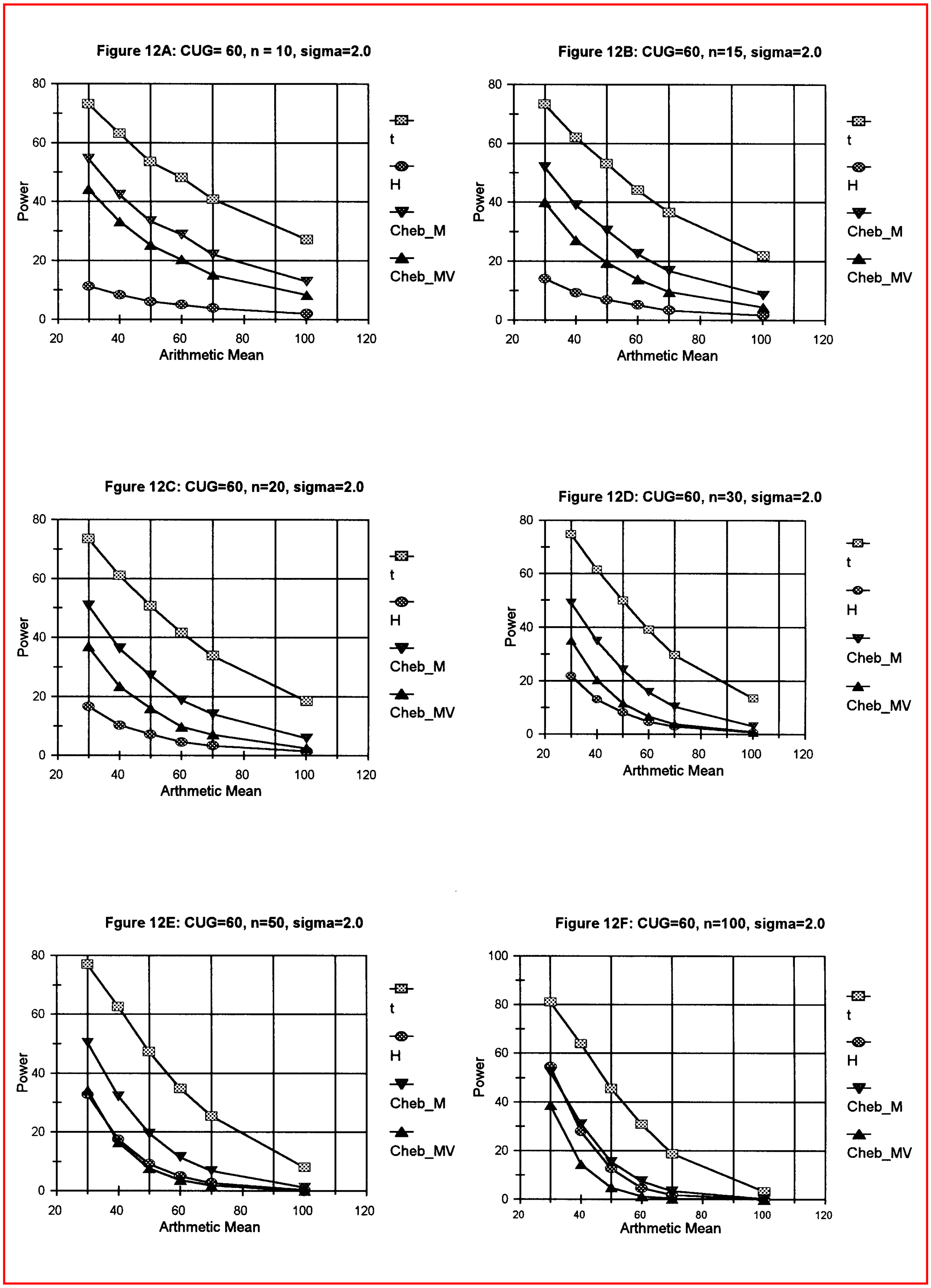
Figure 11C: CUG=60, n=20, sigma=1.5

Figure 11D: CUG=60, n=30, sigma=1.5

Figure 11E: CUG=60, n=50, sigma=1.5

Figure 11F: CUG=60, n=100, sigma=1.5

Figures 11A - 11F



Figures 12A - 12F

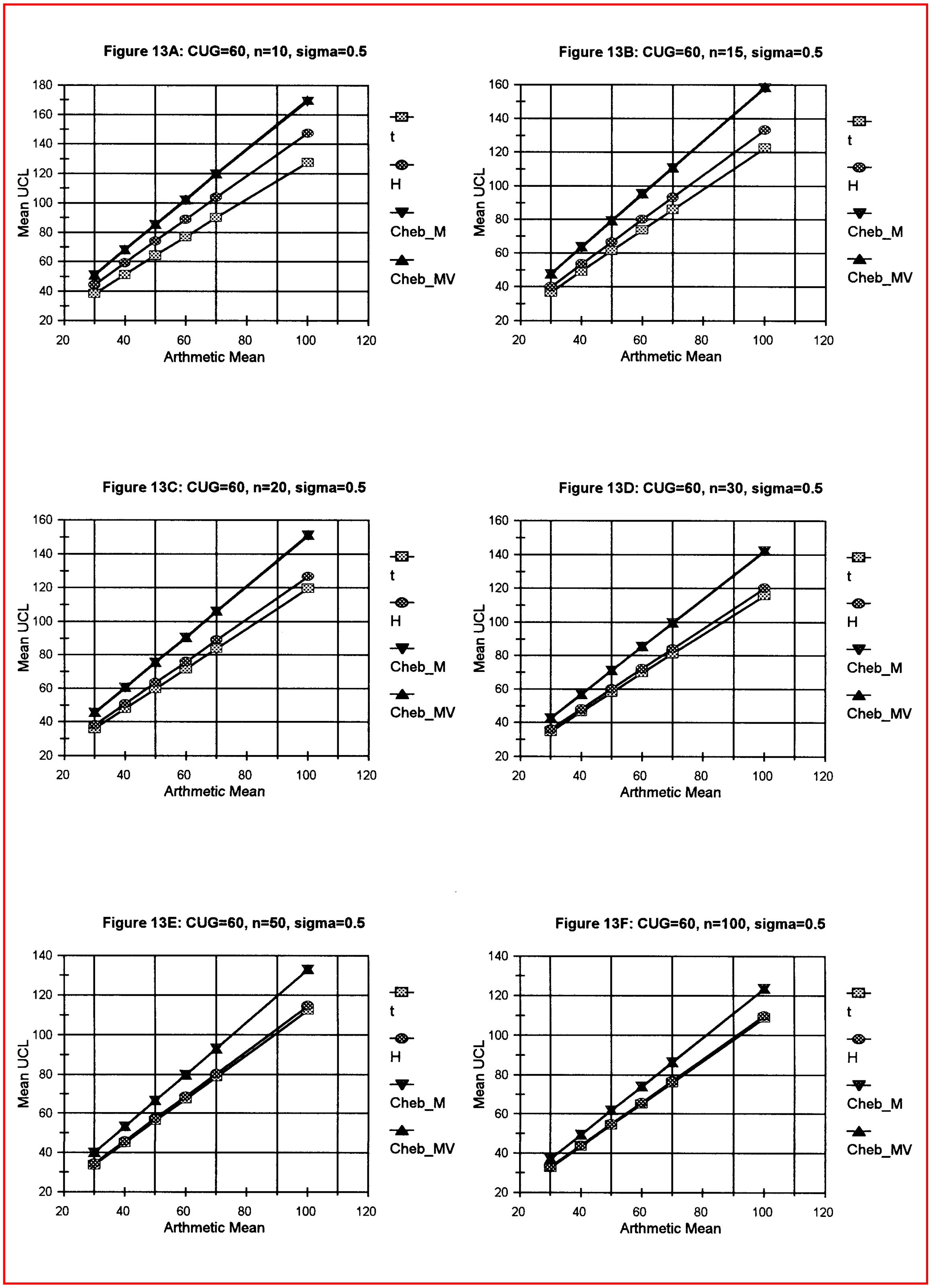


Figure 13A: CUG=60, n=10, sigma=0.5

Figure 13B: CUG=60, n=15, sigma=0.5

Figure 13C: CUG=60, n=20, sigma=0.5

Figure 13D: CUG=60, n=30, sigma=0.5

Figure 13E: CUG=60, n=50, sigma=0.5

Figure 13F: CUG=60, n=100, sigma=0.5

Figures 13A - 13F

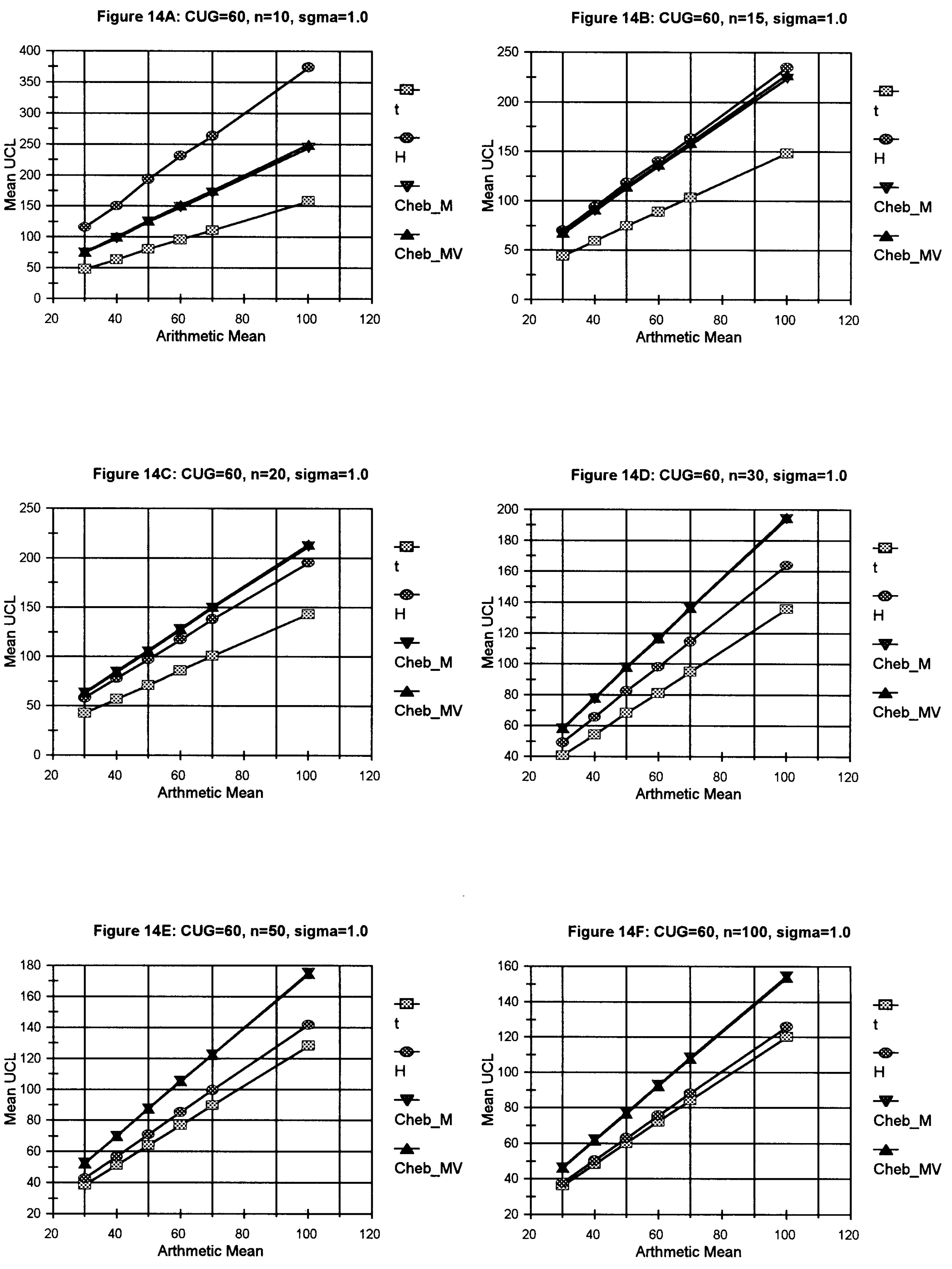


Figure 14A: CUG=60, n=10, sgma=1.0

Figure 14B: CUG=60, n=15, sigma=1.0

Figure 14C: CUG=60, n=20, sigma=1.0

Figure 14D: CUG=60, n=30, sigma=1.0

Figure 14E: CUG=60, n=50, sigma=1.0

Figure 14F: CUG=60, n=100, sigma=1.0

Figures 14A - 14F

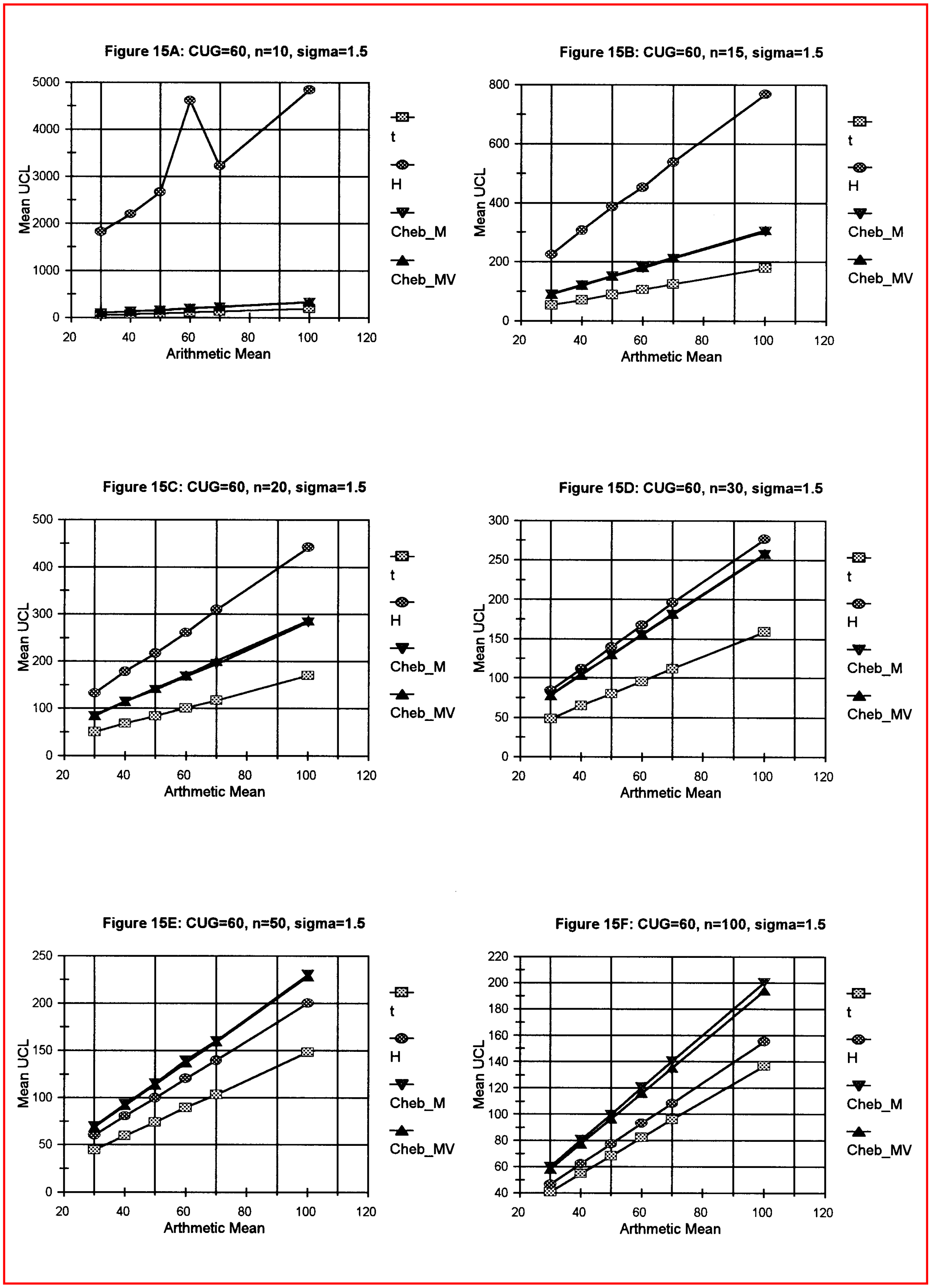


Figure 15A: CUG=60, n=10, sigma=1.5

Figure 15B: CUG=60, n=15, sigma=1.5

Figure 15C: CUG=60, n=20, sigma=1.5

Figure 15D: CUG=60, n=30, sigma=1.5

Figure 15E: CUG=60, n=50, sigma=1.5

Figure 15F: CUG=60, n=100, sigma=1.5

Figures 15A - 15F

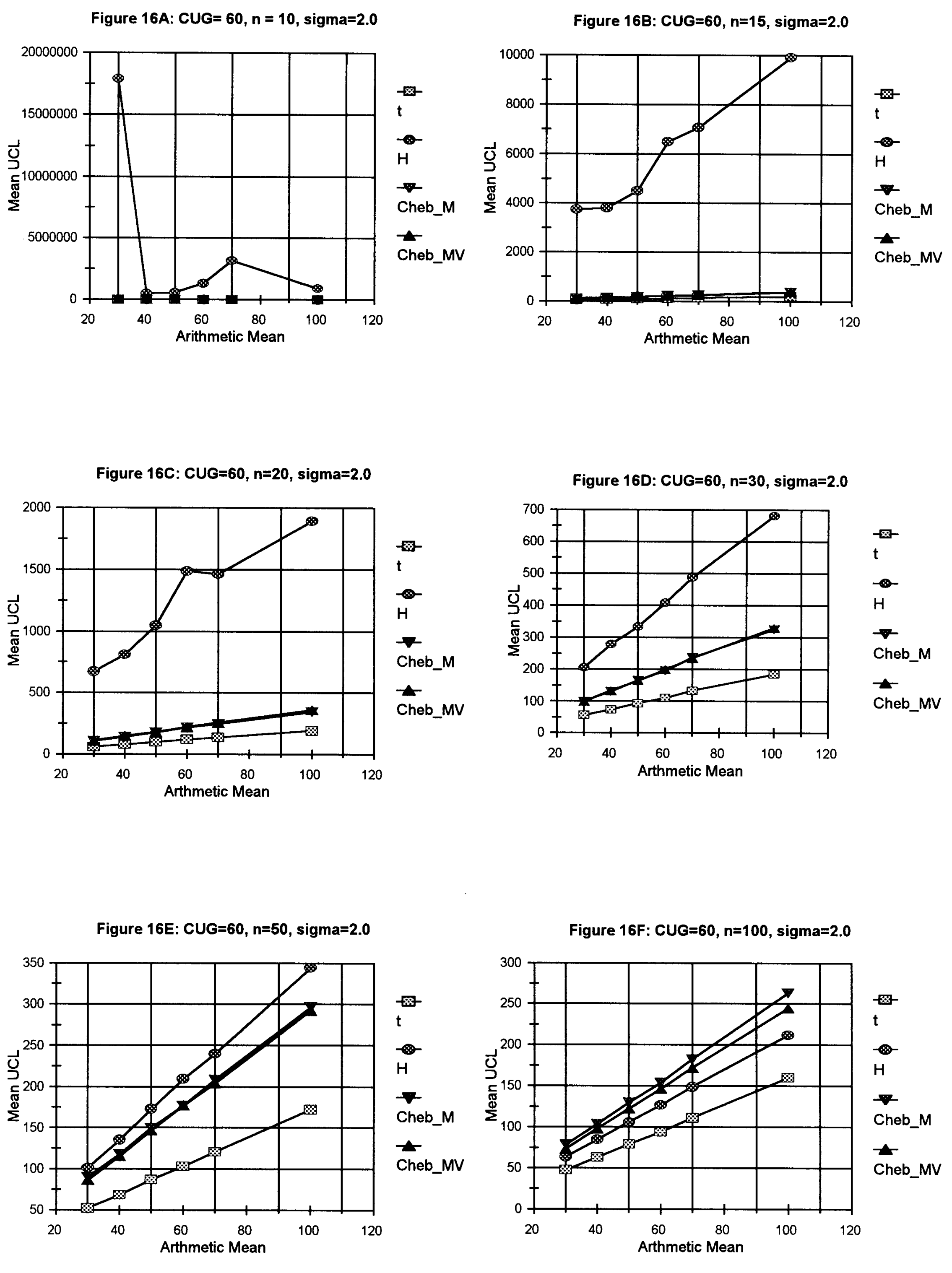


Figure 16A: CUG= 60, n = 10, sigma=2.0

Figure 16B: CUG=60, n=15, sigma=2.0

Figure 16C: CUG=60, n=20, sigma=2.0

Figure 16D: CUG=60, n=30, sigma=2.0

Figure 16E: CUG=60, n=50, sigma=2.0

Figure 16F: CUG=60, n=100, sigma=2.0

Figures 16A - 16F

Notice

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and prepared this Issue Paper. It has been peer reviewed by the EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.

References

Bain, L. J., and Engelhardt, M. (1992), *Introduction to Probability and Mathematical Statistics*, Boston: Duxbury Press.

Blackwood, L. G. (1991), “Assurance Levels of Standard Sample Size Formulas,” *Environmental Science and Technology*, Vol. 25, No. 8, pp. 1366-1367.

Bowers, T., Neil, S., and Murphy, B. (1994), "Applying Hazardous Waste Site Cleanup Levels: A Statistical Approach to Meeting Site Cleanup Goals on Average.” *Environmental Science and Technology*.

Chen, L. (1995), "Testing the Mean of Skewed Distributions," *Journal of the American Statistical Association*, 90, 767-772.

EPA (1989a), “Risk Assessment Guidance for Superfund: Volume 1. Human Health Evaluation Manual (Part A),” Publication EPA 540/1-89/002.

EPA (1989b), “Methods for Evaluating the Attainment of Cleanup Standards. Volume 1. Soils and Solid Media,” Publication EPA 230/2-89/042.

EPA (1992), “Supplemental Guidance to RAGS: Calculating the Concentration Term,” Publication EPA 9285.7-081.

EPA (1991), “A Guide: Methods for Evaluating the Attainment of Cleanup Standards for Soils and Solid Media,” Publication EPA/540/R95/128.

EPA (1996), “A Guide: Soil Screening Guidance: Technical Background Document,” Second Edition, Publication 9355.4-04FS.

Gilbert, R.O. (1987), *Statistical Methods for Environmental Pollution Monitoring*, New York: Van Nostrand Reinhold.

Gilbert, R.O. (1993), "Comparing Statistical Tests for Detecting Soil Contamination Greater than Background," Pacific Northwest Laboratory, Technical Report No. DE 94-005498.

Hogg, R.V., and Craig, A.T. (1978), *Introduction to Mathematical Statistics*, New York: Macmillan Publishing Company.

Kupper, L. L. and Hafner, K. B. (1989), “How Appropriate Are Popular Sample Size Formulas?”, *The American Statistician*, Vol. 43, No. 2, pp. 101-105.

Johnson, N.J. (1978), “Modified t-Tests and Confidence Intervals for Asymmetrical Populations,” *The American Statistician*, Vol. 73, pp. 536-544.

Land, C. E. (1971), "Confidence Intervals for Linear Functions of the Normal Mean and Variance," *Annals of Mathematical Statistics*, 42, 1187-1205.

Land, C. E. (1975), "Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance," in *Selected Tables in Mathematical Statistics*, Vol. III, American Mathematical Society, Providence, R.I., 385-419.

Singh, A. K., Singh, Anita, and Engelhardt, M., “The Lognormal Distribution in Environmental Applications,” EPA/600/R-97/006, December 1997.

Stewart, S. (1994), "Use of Lognormal Transformations in Environmental Statistics," M.S. Thesis, Department of Mathematics, University of Nevada, Las Vegas.

Appendix

Approximate test of the mean of a lognormal distribution

As discussed above, for a lognormal distribution that is not highly skewed, such as $\sigma < 0.5$, it might be reasonable to use a test of the median as an approximate test of the mean. Let the x_i's be a random sample from a lognormal population, with both μ and σ^2 unknown, and with $y_i = \ln(x_i)$. Denote the standardized difference by $d = \delta/\sigma = [\ln(\mathrm{C_s}) - \ln(\mu_1)]/\sigma$. The lognormal mean is $\mu_1 = \exp(\mu + 0.5\sigma^2)$, and the parameter μ can be written in the equivalent form, $\mu = \ln(\mu_1) - 0.5\sigma^2$. The power function is given by

$$\Pi(\mu,\sigma) = \mathrm{P}[\mathrm{UCL} < \ln(\mathrm{C_s}) \mid \mu,\sigma] = \mathrm{P}[\bar{y} + t_{1-\alpha,\, n-1} s_y/\sqrt{n} < \ln(\mathrm{C_s}) \mid \mu,\sigma]$$

$$= \mathrm{P}[(\bar{y} - \mu) + t_{1-\alpha,\, n-1} s_y/\sqrt{n} < \ln(\mathrm{C_s}) - \mathit{ln}(\mu_1) + 0.5\sigma^2 \mid \mu,\sigma]$$

$$= \mathrm{P}[\sqrt{n}(\bar{y} - \mu)/\sigma + t_{1-\alpha,\, n-1}\sqrt{s_y^2/\sigma^2} < \sqrt{n}d + 0.5\sqrt{n}\sigma \mid \mu,\sigma]$$

$$= \mathrm{P}[Z + t_{1-\alpha,\, n-1}\sqrt{W/(n-1)} < \sqrt{n}d + 0.5\sqrt{n}\sigma] \qquad \text{(A1)}$$

where $Z = \sqrt{n}(\bar{x} - \mu)/\sigma$, and $W = (n-1)s_y^2/\sigma^2$ are independent random variables; Z is standard normal and W is chi-square distributed with $n-1$ degrees of freedom. Equation (A1) can be written as follows:

$$\Pi(\mu,\sigma) = \mathrm{P}[Z < \sqrt{n}d - g_1(W)] = \int_0^\infty \Phi(\sqrt{n}d - g_1(w)) f_W(w)\mathrm{d}w \qquad \text{(A2)}$$

where $g_1(w) = t_{1-\alpha,n-1}\sqrt{w/(n-1)} - 0.5\sqrt{n}\sigma$, and

$$f_W(w) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} w^{\nu/2-1}\exp(-w/2)$$

is the probability density function of a chi-square distribution with $\nu = n-1$ degrees of freedom. Note that equation (A2) depends not only on d, but also separately on σ because the function $g_1(w)$ depends on σ. The power can be evaluated by numerical integration and it yields the significance level of the test of a lognormal mean if $d = 0$. This approach was used to compute the significance levels shown in Table 1. Results about the noncentral t distribution, with the appropriate choice of the noncentrality parameter, can also be used to evaluate this power function.

The power function of the H-test

An approach similar to the derivation which yielded equation (A2) can be used to derive the power function of the H-test as a function of σ and the standardized difference, $d = \delta/\sigma = [\ln(\mathrm{C_s}) - \ln(\mu_1)]/\sigma$. Note, this can also be written as $\ln(\mathrm{C_s}) = \ln(\mu_1) + \sigma d = \mu + \sigma^2/2 + \sigma d$. Recall, that the H-factors depend on s_y and n; that is, $H_{1-\alpha} = H_{1-\alpha}(s_y, n)$. It follows that the power function, as a function of both μ and σ, is given by

$$\Pi(\mu,\sigma) = \mathrm{P}[\mathrm{UCL} < \mathrm{C_s}|\mu,\sigma] = \mathrm{P}[\exp(\bar{y} + s_y^2/2 + (s_y H_{1-\alpha}/\sqrt{n-1})) < \mathrm{C_s}|\mu,\sigma]$$

$$= \mathrm{P}[\bar{y} + s_y^2/2 + (s_y H_{1-\alpha}/\sqrt{n-1}) < \ln(\mathrm{C_s}) \,|\, \mu,\sigma]$$

$$= \mathrm{P}[\bar{y} + s_y^2/2 + (s_y H_{1-\alpha}/\sqrt{n-1}) < (\mu + \sigma^2/2 + \sigma d) \,|\, \mu,\sigma]$$

$$= \mathrm{P}[\,(\bar{y} - \mu)/\sigma + (\sigma/2)(s_y^2/\sigma^2 - 1) + (s_y/\sigma)H_{1-\alpha}/\sqrt{n-1} < d \,|\, \mu,\sigma]$$

$$= \mathrm{P}[\,\sqrt{n}(\bar{y} - \mu)/\sigma + \sqrt{n}(\sigma/2)(s_y^2/\sigma^2 - 1) + \sqrt{n}(s_y/\sigma)H_{1-\alpha}/\sqrt{n-1} < \sqrt{n}d \,|\, \mu,\sigma]$$

$$= \mathrm{P}[Z + \sqrt{n}(\sigma/2)[W/(n-1) - 1] + \sqrt{nW}\,H_{1-\alpha}(\sigma\sqrt{W/(n-1)}, n)/(n-1) < \sqrt{n}d]$$

$$= \mathrm{P}[Z < \sqrt{n}d - g_2(W)] = \int_0^\infty \Phi(\sqrt{n}d - g_2(w)) f_W(w)\mathrm{d}w \qquad \text{(A3)}$$

$$g_2(w) = \sqrt{n}(\sigma/2)[w/(n-1) - 1] + \sqrt{nw}\,H_{1-\alpha}(\sigma\sqrt{w/(n-1)}, n)/(n-1),$$

Z and *W* are the random variables defined in equation (A1). The function $g_2(w)$ depends not only on *d*, but also separately on *σ*. Thus, in order to perform numerical evaluation with equation (A3), it is necessary to specify both *d* and *σ*. The integral on the right of equation (A3) can be evaluated by numerical integration in a manner similar to the evaluation of equation (A2).