Revista Colombiana de Estadística

No.35

ON SAMPLE SIZE ESTIMATION OF THE ARITHMETIC MEAN OF A LOGNORMAL DISTRIBUTION WITH AND WITHOUT TYPE I

ADRIANA PÉREZ AND JOHN J. LEFANTE

[a] Profesora Asistente, Unidad de Epidemiología Clínica, Facultad de Medicina, Pontificia Universidad Javeriana y Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Santafé de Bogotá, Colombia
[b] Department of Biostatistics and Epidemiology, School of Public Health and Tropical Medicine, Tulane University Medical Center, New Orleans, LA 70112-2699, USA

ABSTRACT This article presents several formulas to approximate the required sample size to estimate the arithmetic mean of a lognormal distribution with desired accuracy and confidence under and without the presence of type I censoring to the left. We present tables of exact sample sizes which are based on Land's exact confidence interval of the lognormal mean. Monte Carlo estimates of coverage probabilities show the appropriateness of these exact proposed sample sizes at 95% confidence level.

In the case of non censoring, Box-Cox transformations were used to derive formulae for approximating these exact sample sizes and new formulae, adjusting the classic central limit approach, were derived. Each of these formulas as well as other existing formulas (the classical central limit approach and Hewett's formula) were compared to the exact samples size to determine under which conditions they perform optimally and recommendations are given.

KEY WORDS: Confidence interval width, Sample size determination, Box-Cox transformation, Uniformly most accurate unbiased invariant confidence interval, Bias correction, Maximum likelihood estimator.

1. INTRODUCTION

Distributions of concentrations environmental contaminants, occupational exposures, small particles, etc., are often approximately lognormally distributed. In the case of environmental exposure measurements, the choice of a suitable summary measure (arithmetic mean, geometric mean, a tolerance limit, etc.) depends on the investigator's research interest.

Sampling strategies which focus upon the arithmetic mean (Armstrong 1992; Evans and Hawkins 1988; Seixas, Robins and Moulton 1988) often are effective for assessing exposure to toxic materials and is related to the frequency of exposures which exceed particular air concentrations (Rappaport, Selvin and Roach 1988). Emphasis and development in exposure monitoring technology have centered on mechanical aspects such a how to make sampling more convenient and comprehensive and how to make analyses more sensitive and reliable.

The evaluation of the sample size required to achieve statistically credible results is a crucial element in exposure monitoring as well as in a diversity of observational and experimental studies where the interesting is to estimate the most relevant parameter of the lognormal distribution.

For a normally distributed random variable, this minimum sample size is usually determined via the use of simple formulas or from tables. Even the more popular formulas, however, involve large-sample approximations and hence may underestimate required sample sizes. This underestimation phenomenon could be extreme for certain sample size formulas based on confidence interval width (Greenland 1988; Kupper and Hafner 1989).

In the case of a lognormally distributed random variable, there is very little in the statistical literature evaluating the minimum required sample size to estimate the arithmetic mean. Hewett (1995) presented a formula for calculating the approximate sample size needed to estimate the true arithmetic mean within a specified accuracy and with a specified level confidence for non censoring data.

The classical central limit approach has been also used for estimate the minimum required sample size for the non censoring case. However, an evaluation of the accuracy of these formulas has not been made. This article also presents some guidelines for the selection of an adequate formula for estimating the exact sample size for the non censoring case.

A further problem arise, for example, when measuring minute concentrations of environmental pollutants, even state of the art instruments may not be able to detect the actual concentration. When concentration cannot be quantified below a limit of detection (LOD), the value is usually reported as non detectable which leads to left censoring of the sample and new techniques should address to evaluate the minimum required sample size in this type of situations.

In the presence of censoring, Cohen (1950,1959) used the method of maximum likelihood (MLE) to estimate the parameters of normal populations from singly and doubly truncated samples for Type I censoring. Saw (1961) noted that above MLEs were biased and they are not asymptotically unbiased.

Saw (1961) found the leading term in the bias of the estimators of the mean and the standard deviation for a normal random variable, suggesting corrected estimators for singly censored samples. Their bias increases with increasing degree of censoring. Thus, in comparison to the estimators without censoring, an adjustment is required in a censored sample. The bias tends to zero as the sample size tends to infinity, but for small sample sizes the bias is significantly large to warrant consideration.

This paper includes an attempt to address this need, by proposing exact sample sizes to provide statistically credible results for the arithmetic mean of a lognormally distributed random variable when the data contains values below the limit of detection and also when this problem does not exist.

2. NOTATION

X is a lognormal random variable such that the function $f(X) = \ln(X) = Y$ follows a normal distribution with mean m and standard deviation σ. The arithmetic mean, the variance and their minimum variance unbiased estimators (MVUE) (Finney,1941) of this lognormal distribution are respectively for the non-censoring case:

$$\theta = E(X) = \exp\left(\mu + 0.5\sigma^2\right) = \mu_g \exp(0.5\sigma^2) \tag{1}$$

$$\delta = V(X) = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right) \tag{2}$$

$$\hat{\theta}_{MVUE} = \exp\left(\bar{Y}\right) g\left(0.5\ Sy^2\right), \tag{3}$$

and

$$\delta_{MVUE} = \exp(2\bar{y})\left(g\left(2S_y^2\right) - g\left((n-2)\,S_y^2/(n-1)\right)\right), \tag{4}$$

where:

$$g(t) = 1 + \frac{(n-1)\,t}{n} + \sum_{j=2}^{\infty} \frac{(n-1)^{2j-1}}{n^j\,(n+1)(n+3)\ldots(n+2j-1)} \frac{t^j}{j!}$$

The maximum likelihood estimators of the geometric mean ($\mu_g = \exp(\mu)$) and the geometric standard deviation ($\sigma_g = \exp(\sigma)$) of this lognormal distribution are $\hat{\mu}_g = \exp\left(\bar{Y}\right)$ and $\hat{\sigma}_g = \exp\left(S_y\right) = GSD$ respectively; where $\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{(n-1)}$.

As has been noted, the natural logarithm of the geometric mean has the nice property that it is the same value as the mean of the normal distribution. Therefore, required sample size formulas and equivalent tables for estimating the geometric mean are well known. However, there are not straightforward formulae for estimating the sample size for the arithmetic mean.

3. SAMPLES SIZES FORMULAS :NON CENSORING CASE

3.1 Classical Formula

The classical option to generate a formula to obtain the required sample size for a given GSD (estimated from prior information or pilot data) and a desired accuracy level (100 π percentage difference from the true arithmetic mean) is based on confidence interval width and large sample size theory through the Central Limit Theorem. Given a confidence level of α and α two-sided confidence interval, we derive $\pi\theta = Z_{\alpha/2}\left(\sqrt{\delta}/\sqrt{n_{classic}}\right)$ where $n_{classic}$ represents the required sample size, θ and δ were defined above.

Substituting θ by (1) and δ by (2), we derive

$$\pi\left(\exp\left(\mu + 0.5\sigma^2\right)\right) = \left(Z_{\alpha/2}/\sqrt{n_{classic}}\right)\sqrt{\exp\left(2\mu + \sigma^2\right)\left(\exp\sigma^2 - 1\right)}$$

Which can be expressed as

$$n_{classic} = \left(Z_{\alpha/2}/\pi\right)^2\left(\sigma_g^{\ln\sigma_g} - 1\right)$$

An approximate sample size is :

$$n_{classic} = \left(Z_{\alpha/2}/\pi\right)^2 \left(GSD^{\ln GSD} - 1\right), \tag{5}$$

As by expected by the Central Limit Theorem, for most cases this formula underestimate the required sample size. A discussion of this underestimation is provided in Section 5.

3.2 Hewett's Formula

Hewett (1995) published a sample size formula for estimating the true arithmetic mean of a lognormal distribution to within a specified accuracy (± 100π percent difference from the true arithmetic mean) with a specified level of confidence. This formula requires also a priori information from previous data or a pilot study. The approximate sample size can be calculate using the following formula

$$n_{Hewett} \cong \left(t^2_{\alpha/2, n_{pilot}-1}\, \delta_{MVUE}\right) / \left(\pi\theta_{MVUE}\right)^2 \tag{6}$$

where θ_{MVUE} and δ_{MVUE} are given in (3) and (4) respectively, 100π represents the desired accuracy level and t is the value from a t-student distribution for a 1-α confidence level and $(n_{pilot} - 1)$ degrees of freedom. θ_{MVUE} , δ_{MVUE} and n_{pilot} are calculated from prior information or a pilot study.

Using Monte Carlo techniques, Hewett (1995) tested this formula by generating predicted sample sizes for different pilot study sizes, GSD's and several 100π percentage differences. He used pilot study datasets of sizes n_{pilot} = 5, 10, 20 and 50 from lognormal distributions having a true geometric mean of 10 and true geometric standard deviations of 1.5, 2, 3 and 4.

His simulation results indicate that the estimated confidence levels approached the target level of 95% for most combinations of geometric standard deviations and n_{pilot}. The exceptions were for large geometric standard deviations (≥ 3) and small pilot study sample sizes (< 20). Caution is recommended for estimating the appropriate sample size using (6) if n_{pilot} is small and the GSD is large.

3.3 Exact Sample Size

Land (1971, 1972, 1974) developed an exact method for constructing one and two sided confidence intervals for $E(X)$. This method has been described as a special case of estimating confidence intervals for linear functions of the normal mean and variance. The exact method is optimal in the sense that it is defined by uniformly most accurate invariant confidence intervals.

The minimum required sample size can be calculated based on the confidence interval width of Land's exact interval. They are expressed as a function of a specified GSD and within a desired accuracy level (100π) with a specified level of confidence.

Methodology

Land (1973, 1974, 1975, 1988) published tables of standard limits to calculate the exact confidence intervals. These standard limits are based on a computationally tedious method defined in terms of the conditional distribution of a test statistic given the value of another statistic. By using these exact confidence intervals, it is possible to generate exact sample size tables.

In this case, it is easy to compute the percent difference between the upper and/or lower confidence limit and the estimated arithmetic mean. After obtaining these percentages of variation from the arithmetic mean based on GSD, a determination of which "exact" sample size is necessary can be made.

Armstrong (1992) published tables of two sided 95% confidence intervals expressed as a multiple of the geometric mean for different sample sizes and different GSDs. Then, using his result and if we assume a geometric standard deviation of 2.5 and we allow 85.5% variability from the arithmetic mean (upper side percentage difference between upper confidence limit and the estimated arithmetic mean), the "exact" sample size will be 20 for any geometric mean, based in a 95% exact two-sided confidence level.

Therefore, independently of the geometric mean, fixed percentage difference from the true arithmetic mean defines the required sample size. Without loss of generality, a true geometric mean of one was assumed in computations.

Results

Dr. Charles E. Land provided the computer program from which estimate confidence intervals for linear functions of the normal mean and variance are calculated. Exact confidence intervals for a lognormally distributed random variable can be calculated by taking the exponential of the appropriate confidence interval computed by Land's program. The program is written in FORTRAN and has been tested for confidence levels ranging from 0.900 to 0.995 and the degrees of freedom for estimating σ^2 ranging from 2 to 1000.

Table 1 contain the minimum required sample sizes for estimating the true arithmetic mean of a lognormally distributed random variable for the 95% confidence level. These samples sizes were calculated based on the exact confidence interval width. Datasets with GSDs of 1.1, 1.5, 2, 2.5, 3, 3.5 and 4 having a true geometric mean of one for sample sizes of three to 1000 and a confidence level of 95% were generated by Statistic Analysis System (SAS 1985). The degrees of freedom used for estimating σ^2 were (n-1) (Land 1972). These datasets are used in Land's program in order to compute two-sided confidence intervals. Land's program reads these SAS datasets and outputs ASCII datasets.

The percentage difference between the limits of the exact confidence intervals and the true arithmetic mean for the conditions given were computed using SAS on the outputted ASCII datasets. Because the upper sided percentage is always greater than the lower sided, the upper sided percentage is the recommended percentage to used for the estimation of the corresponding sample size. These result appear in the table as a function of the sample size in term of GDS's and 100π percent difference from the true arithmetic mean.

In generating the exact sample size values, the percentages of variation from the true arithmetic mean increase with increasing geometric deviation. This is expected and implies that the large the variability and lower the percentage of variation from the true arithmetic mean, the larger the sample size required, or vice versa, the lower the variability and larger the percentage of variation from the true arithmetic mean, the lower the sample size required.

Table 1. Exact minimum required sample size for 95% two sided confidence level.

	GSD						
100 π	1.1	1.5	2.0	2.5	3.0	3.5	4.0
5	16	251	831				
10	7	77	249	493	794		
15	5	40	125	245	393	564	754
20	4	25	78	151	242	346	461
21	4	24	72	140	223	319	425
22	4	22	67	129	206	295	393
23	4	21	62	120	192	274	365
24	4	20	58	112	179	256	340
25	4	19	55	105	168	239	318
26	4	18	52	99	157	224	298
27	4	17	49	93	148	211	281
28	4	16	46	88	140	199	265
29	4	15	44	84	132	188	250
30	4	15	42	79	126	178	237
31	3	14	40	76	119	169	225
32	3	14	38	72	114	161	214
33	3	13	36	69	108	154	204
34	3	13	35	66	104	147	194
35	3	12	33	63	99	140	186
36	3	12	32	60	95	134	178
37	3	12	31	58	91	129	171
38	3	11	30	56	88	124	164
39	3	11	29	54	84	119	158
40	3	11	28	52	81	115	152
41	3	10	27	50	78	111	146
42	3	10	26	49	76	107	141
43	3	10	25	47	73	103	136
44	3	10	25	45	71	100	132
45	3	9	24	44	69	97	127

NOTE: Result are given for several estimated geometric deviations (GSD) from prior information or pilot data and several percentage differences from the true arithmetic mean (100π).

As an example of how this table works, we used the same example mentioned by Hewett (1995) where a prospective exposure-response study of workers exposed to welding fumes was proposed. For one exposure group from a pilot study, 17 measurements gave an approximately GSD of 1.55, then for a 25% percentage difference from the arithmetic mean at a 95% confidence level a sample size interpolated from table 1, between GSD=1.5 and GSD=2.0 gives a required sample size of 23 observations instead of the 15 measurements suggested by Hewett.

Other of his examples gave a GSD of 2.16 using 18 measurements within 25% percentage difference from the arithmetic mean, at a 95 % confidence level, this requires an interpolated sample size of 71 observations instead of the 51 measurements suggested by Hewett.

Monte Carlo Simulations

Monte Carlo simulations were used to the test above results. Artificial datasets were used to create different scenarios. The computer clock time at execution was used to generate in SAS a seed from the uniform distribution on the interval [0, 1]. The seed's integer value was obtained by multiplying the seed by 1 billion and rounding it to the nearest integer roundoff unit. For convenience, this number will be called a list's seed. Using this list's seed as a seed to generate a lognormal variable with geometric mean given equal to 1 and geometric standard deviation given by $\exp(\sigma)$, with several values, a sample size of size (n) was generated. After taking the natural logarithm of the data, the sample mean and standard deviation of the normalized data were computed. This procedure was repeated 1000 times.

Using Land's program and the sample means and the sample standard deviations, confidence intervals for the arithmetic mean were calculated. After taking the exponential function for these confidence intervals, the number of confidence intervals that contains the true arithmetic mean was counted. This means that the statistic of interest was the observed confidence level of the 1000 datasets that contains the true arithmetic mean.

Coverage probabilities at the target level of 95% for the proportions of the 1000 confidence intervals that contains the true arithmetic mean for several geometric standard deviations and several percentage differences are reported in table 2. For the cases shown, this demonstrates that the sample sizes are adequate at the confidence level specified.

Table 2. Monte Carlo results for 95% two sided confidence level.

GSD	**100π**							
	20		30		40		50	
	n	(1-α)	n	(1-α)	n	(1-α)	n	(1-α)
1.1	4	95.9	4	95.0	3	95.3	3	95.3
1.5	25	95.4	15	95.0	11	94.4	9	96.1
2.0	78	96.0	42	96.1	28	96.7	21	95.1
2.5	151	95.6	79	94.9	52	95.5	38	95.1
3.0	242	94.8	126	93.4	81	94.4	59	95.4
3.5	346	94.8	178	95.7	115	93.8	83	95.8
4.0	461	94.9	237	95.7	152	95.1	109	94.0

NOTE: Results are given for several estimated geometric standard deviations (GSD) form prior information or pilot data and several percentage differences from the true arithmetic mean (100π).

3.4 Proposed Sample Size Formula

Unfortunately, above tables can never be large enough to cover every combination of GSD and percentage difference from the true arithmetic mean. For this reason, we are interested in finding a simple closed linear or nonlinear model that corresponds closely to the exact sample size for estimating the true arithmetic mean of a lognormal distribution with a specified level of confidence. Such formulae $n = f(GSD, \pi) + \epsilon$

will allow researchers to determine the sample size they need in their investigation without relying on sample size tables.

A Box-Cox transformation (Box and Cox 1964) using logarithms and a quadratic term provided:

$$\ln(n) = \beta_o + \beta_1 \ln(GSD) + \beta_2 \ln(GSD)^2 + \beta_3 \ln(\pi)$$

and

$$n = \exp(\beta_0)\, GSD^{\beta_1} GDS^{\beta_2 \ln(GSD)} \pi^{\beta_3}$$

This model performed very well with all the parameters highly significant. Results of these models are presented in table 3 as equations (7) – (9).

Table 3. Parameter estimates for proposed exact formula.

Confidence level	Model	
90%	$\hat{n} = \exp(-0.215269)\, GSD^{3.687867} GSD^{-0.684730 \ln(GSD)} \pi^{-1.185768}$	(7)
95%	$\hat{n} = \exp(-0.172970)\, GSD^{4.297741} GSD^{-0.986961 \ln(GSD)} \pi^{-1.201125}$	(8)
99%	$\hat{n} = \exp(-0.162979)\, GSD^{4.746648} GSD^{-1.213001 \ln(GSD)} \pi^{-1.174062}$	(9)

NOTE: GSD:estimate geometric standard derivation from prior information or pilot data and 100π: percentage difference from the true arithmetic mean.

3.5 Adjusted Classical Formula

Correction factors were sought to improve the classical approximation (5), using linear regression models. Table 4 presents linear regression estimates of the fit of the exact sample sizes values (n_{exact}) on the estimates from equation (5) ($n_{classic}$) for each GSD for 90%, 95% and 99% two-sided confidence level. In short, the model begin used is: $n_{exact} = \beta_0 + \beta_1 n_{classic} + \epsilon$.

All the parameter estimates and models were highly significant and all models correct the under/over estimation of the classic formula. This approach allows a simple adjustment of the classic formula to obtain exact sample sizes values. Furthermore, the method is straightforward and computationally simple to apply.

Table 4. Linear regression coefficients for $\hat{n}_{exact} = \hat{\beta}_0 + \hat{\beta}_1 \hat{n}_{classic}$

GSD	$\hat{\beta}_0$	$\hat{\beta}_1$	
90% two sided confidence level			
1.1	2.9532	0.4714	
1.5	7.5249	0.6926	
2.0	11.3183	0.8509	
2.5	15.5638	0.8794	
3.0	20.1322	0.8499	
3.5	25.9327	0.7731	
4.0	30.3223	0.7033	
95% two sided confidence level			
1.1	3.3331	0.4726	(12)
1.5	7.9237	0.8094	(13)
2.0	14.0744	0.9046	(14)
2.5	20.5406	0.9129	(15)
3.0	27.1563	0.8731	(16)
3.5	33.6865	0.8072	(17)
4.0	40.1084	0.7288	(18)
99% two sided confidence level			
1.1	4.9265	0.4740	
1.5	11.2470	0.8865	
2.0	20.5069	0.9808	
2.5	30.2478	0.9877	
3.0	40.1743	0.9444	
3.5	51.1945	0.8612	
4.0	60.6576	0.7796	

NOTE:GSD estimated geometric standard deviation from prior information or pilot data

4.SAMPLE SIZE ESTIMATION:CENSORING CASE

The approach used for the censoring case is to use the maximum likelihood procedure to estimate the mean and the variance parameters in the transformed scale under censoring and then to use the properties of the MLE's to back transform the MLE's to the original scale (Cohen 1959, 1961). The mayor disadvantage of Cohen's MLE is that when σ is unknown, there are not explicit solutions for the MLE and it is necessary to use Newton-Raphson iteration methods.

To compute the minimum required sample size based on confidence intervals width, Saw's bias correction to Cohen's maximum likelihood estimator was used. The MLE is used because of its nice properties and Saw's bias correction factor was selected because of its low variability in comparison to the other bias correction approaches (Custer 1976, Tiku 1978, Schneider 1986) found in the literature.

Saw's bias correction factors involves complex computations to obtain the leading terms in the bias of μ and σ $(B(\hat{\mu}, p_{n_u})$ *and* $B(\hat{\sigma}, p_{n_u}))$ as a function of fraction of uncensored observations $(p_{n_u} = n_u/(n+1))$. n_u identifies the number of uncensored observations. The relationship between the factors p_{n_u}, $B(\hat{\sigma}, p_{n_u})$ *and* $B(\hat{\mu}, p_{n_u})$ respectively, was investigate to obtain a linear regression model that will model this bias. The final models are shown on equations (10) and (11).

$$\hat{B}(\hat{\mu}, p_{n_u}) = 0.582896 - 0.547792\,(p_{n_u}^{-1.5}) \tag{10}$$

$$B(\sigma, p_{n_u}) = 0.240954 - (1.000859/p_{n_u}) \tag{11}$$

These models performed better than the models proposed by Schneider and Weissfel (1986).

Methodology

In like manner as for the non-censoring case, the fixed percentage of variation (π) from the true arithmetic mean and the assumed, from prior information, geometric standard deviation (GSD) must be specified . In addition, it is also necessary to specify the proportion of expected censoring (percentile in which $Y_0 = \ln(LOD)$ is located in the population). Then, the same methodology that was used for the non-censoring case to estimate the minimum required sample size will be used under the presence of censoring observations.

For a confidence level of 95%, dataset with GSDs of 1.5, 2.0, 2.5, 3.0, 3.5 and 4.0 and true arithmetic mean of 0 were generated by SAS. Under these conditions, datasets with sample sizes ranging from ten to 1000 were generated with combinations of 10% and 20% censoring.

Maximum likelihood estimates of σ corrected for bias using equation (8) were used in Land's procedure to compute two-sided confidence intervals. The number of degrees freedom, used to estimate the maximum likelihood estimator of σ , were over-estimated to be $(n_u - 1)$ using large sample theory through the Central Limit Theorem (Schmee, Gladstein and Nelson 1982, 1985).

For each confidence interval, the percentage difference between the upper and lower confidence limit and the true arithmetic mean was determined. The minimum sample size, in which the confidence interval coincides with the percentage difference needed by the researcher is reported in tables 5 for 95 % confidence level for the GSD coming from pilot data or a priori information, several level of π and several proportions of censoring.

Results

Similarly as in the non censoring case, the percentage of variation from the true arithmetic mean increase with increasing geometric standard deviation at any proportion of censoring. This implies that the larger the variability, the lower the percentage of variation and the larger the percentage of censoring, the larger the sample size required.

Table 5. Exact the minimum required sample size with censoring for estimating the arithmetic mean of a lognormally distributed random variable at 95% two sided confidence interval. Result are given for 10% and 20% levels of censoring, several estimated geometric standard deviations (GSD) and several percentage differences from the true arithmetic mean (100 π).

	10% of Censoring					
	GSD					
100 π	1.5	2.0	2.5	3.0	3.5	4.0
5	284	936				
10	90	286	561	901		
15	48	145	282	450	644	858
20	31	92	177	280	398	530
21	29	86	164	259	368	489
22	29	80	152	240	341	452
23	29	77	142	224	318	421
24	29	70	133	209	297	393
25	29	67	124	196	278	368
26	29	62	118	185	261	345
27	21	59	111	174	246	325
28	20	58	105	165	233	308
29	20	58	100	156	220	291
30	20	51	96	149	209	276
31	20	49	91	141	199	262
32	20	48	87	135	189	250
33	20	48	83	129	181	238
34	20	48	79	123	173	228
35	20	41	77	119	166	218
36	20	40	73	113	160	209
37	20	39	70	109	153	200
38	20	39	68	105	147	193
39	20	39	67	101	142	186
40	20	39	63	98	136	180
41	20	39	61	94	132	173
42	20	39	59	91	128	167
43	20	39	58	89	123	162
44	20	31	58	86	120	156
45	20	30	58	83	115	151

Table 5. Continued

	20% of Censoring					
	GSD					
100 π	1.5	2.0	2.5	3.0	3.5	4.0
5	320					
10	101	321	632			
15	54	163	317	507	725	966
20	35	104	199	315	449	596
21	35	96	184	291	414	550
22	35	90	171	270	384	510
23	30	84	160	251	357	474
24	30	79	150	235	333	443
25	30	74	141	220	314	414
26	25	70	132	208	294	389
27	25	66	125	196	278	367
28	25	64	119	186	261	346
29	25	60	112	176	248	327
30	25	59	107	167	235	311
31	20	55	102	160	224	295
32	20	54	97	152	214	281
33	20	50	94	145	204	269
34	20	49	90	138	196	256
35	20	46	86	133	187	245
36	20	45	82	127	179	235
37	20	44	80	123	172	227
38	20	41	76	119	166	218
39	20	40	74	114	160	209
40	15	40	71	110	153	202
41	15	40	69	106	148	194
42	15	40	69	102	143	188
43	15	40	65	100	138	182
44	15	35	64	96	135	176
45	15	35	61	94	131	171

From Hewett's examples, if we suppose that for some reason we are expecting a 10 % lower undetectable values of exposure and in the first example we assumed that the 17 measurement were detectable, then the minimum required sample size for a 25% percentage difference from the arithmetic mean at a 95% confidence level will approximately be 33 measurements. Let suppose for the second example that a 20 % censoring is expected. Then, under the same conditions, 96 measurements will allow us to estimate the arithmetic mean within a 25% percentage difference of itself at a 95% confidence level.

Monte Carlo Simulations.

Monte Carlo simulations were used to confirm above results. Similar methodology was used over different scenarios with the inclusion of the censoring factor and using bias corrected estimates.

The computer clock time execution was to generate in SAS a seed from the uniform distribution on the interval $[0, 1]$. The seed's integer value was obtained multiplying the seed by 1 million and rounding it to the nearest integer roundoff unit. Again, for convenience, this number will be called a list's seed. Using this list's seed as a seed to generated a lognormal variable with geometric mean of 1 and several GSD's a sample size of size n was generated.

Expected LOD values of 10% and 20% as a specific levels of censoring were set. Any observation below this value was considered missing and the mean and standard deviation of the natural logarithms of the sample were calculated. If no censored observations were found, this sample was excluded and a new sample was generated.

Cohen's estimators were calculated with help of a macro program and this MLE estimators were corrected for bias and were used in Land's procedure. This simulation was repeated 1000 times and confidence intervals for the arithmetic mean were calculated.

After taking the exponential function for these confidence intervals, the number of confidence intervals that contains the true arithmetic mean was counted. These result are reported in table 6 for selected sample sizes, specific GSD, specific $100\pi\%$ of accuracy, and specific percentage of censoring for the 95% confidence level. These results indicate that the estimated confidence levels were higher for the expected target level, especially for high percentage level of censoring. This means a conservative approach in the case of sample size determination. These results are shown in table 6.

Table 6. Monte Carlo simulation results for 95% two-sided confidence interval. Censoring case. Results are given for 10% and 20% levels of censoring, several estimated geometric standard deviations (GSD) and several percentage differences from the true arithmetic mean (100π)

100π	GSD	Percentage Levels of Censoring			
		10		20	
		n	(1-α)	n	(1-α)
10	1.5	90	96.6	101	96.8
	2.0	286	96.4	321	97.5
	2.5	561	96.2	632	97.3
	3.0	901	96.3		
30	1.5	20	97.2	25	97.5
	2.0	51	96.5	59	97.0
	2.5	96	96.1	107	95.6
	3.0	144	94.8	167	97.0
	3.5	209	96.5	235	96.8
	4.0	276	95.7	311	96.7
50	1.5	20	97.7	15	98.2
	2.0	29	96.6	30	96.7
	2.5	48	95.7	54	98.0
	3.0	72	95.9	81	97.1
	3.5	100	95.9	112	95.6
	4.0	131	96.9	147	97.3

5.COMPARISON OF METHODS AND RECOMMENDATIONS

Non-censoring case

Hewett (1995) presents a comparison of sample sizes necessary for estimating different scenarios . The sample sizes were calculated for various combinations of pilot study sample size (n_{pilot}), GSDs, and desired accuracy level (100π). These results are compared with the exact sample sizes and are show in table 7.

Table 7 Hewett's samples sizes and exact sample sizes for 95% two sided confidence level. Non censoring case.

GDS	$n_{pilot}*$	100π					
		20		30		50	
		$n_{Hewett}*$	Exact	$n_{Hewett}*$	Exact	$n_{Hewett}*$	Exact
1.5	5	34		15		6	
	10	23		10		4	
	20	20	25	9	15	3	9
	50	18		8		3	
	>50	17		8		3	
2.0	5	119		53		19	
	10	79		35		13	
	20	68	78	30	42	11	21
	50	62		28		10	
	>50	59		26		9	
3.0	5	452		201		72	
	10	300		133		48	
	20	257	242	114	126	41	59
	50	234		105		38	
	>50	225		100		36	
4.0	5	1124		500		180	
	10	746		332		119	
	20	639	461	284	237	102	109
	50	589		261		94	
	>50	560		249		90	

Note : GSD: estimated geometric standard deviation from prior information, 100π: percentage difference from the true arithmetic mean, n_{pilot}: sample size from pilot data, and n_{Hewett}:approximate sample size computed using Hewett's formula.
Source: Adapted from Paul Hewett (1995), Sample size formulae for estimating the true arithmetic or geometric mean of lognormal distributed exposure distributions. Table III, facing p. 223. Permission granted by the American Industrial Hygiene Association Journal.

Two important results are shown from table 7. First, for small GSD (≤ 2.0) and small pilot sample sizes of $n_{pilot} = 5, 10$, Hewett's method closely approximate the exact sample size. However, for small GSD and large pilot sample sizes, Hewett's method underestimates the exact sample size required. This is especially true as the accuracy decreases (100π increasing). Secondly, accuracy at high GSD's in Hewett's formula requires a large number of observations in the pilot study.

If a two-stage sampling scheme is considered and the investigator, using Hewett's formula, collects an initial sample of size n_{pilot}, calculates the minimum required sample size (n_{Hewett}), but collects only n_{Hewett} -n_{pilot} measurements, the assumption that must first be validated is that the conditions under which the pilot data were collected are similar to the conditions surrounding the collection of the second stage. Then, the total number of collected measurements required to use Hewett's formula is always greater than that required by the exact method ($n_{Hewett} + n_{pilot}$ versus n) and Hewett's method results in higher sampling costs.

Comparison between the exact sample size values and the classic formula (5), using several accuracy levels (100π) and for several GSD's shows that in general, the classic formula underestimates the minimum required sample size for estimating the arithmetic mean of a lognormally distributed random variable for low geometric standard deviations and several reasonable values of accuracy of 100π. The level of underestimation decreases with increasing GSD.

The classic formula starts to overestimate the required sample size for large GSD's (> 3) at large sample sizes, almost independent of the level of accuracy desired. In the case of large accuracy levels, the classic formula always underestimates the required sample size across GSD's.

Comparing at the 95% confidence level the exact sample size, the classical sample size, the proposed model sample sizes and the adjusted classical sample size values, the following rules apply at this confidence level.

a) For a GSD of 1.5 and large desired accuracy levels ($\leq 25\%$) the proposed model from equations (8) is recommended; otherwise for small accuracy levels ($> 25\%$), the classical adjusted model (12) (table 4) is preferable.

b) For medium GSD's (2 and 2.5) and large desired accuracy levels ($\leq 20\%$) the classical adjusted model (13,14) (table 4) is more reliable than the other approaches; for small desired accuracy levels ($> 20\%$) the predicted values from the proposed model in equation (8) is more adequate.

c) The classical formula (5) is recommended for the following combinations of desired accuracy levels ($100\ \pi$) and GSDs: GSD of 3.0 and $100\pi \leq 20\%$, GSD of 3.5 and $100\pi \leq 30\%$, and GSD of 4.0 and $100\pi \leq 40\%$. The proposed model from equation (8) is recommended in estimating the exact sample size required for the following combinations desired accuracy levels and GSDs: GSD of 3.0 and $100\pi > 20\%$, GSD of 3.5 and $100\pi > 30\%$, and GSD of 4.0 and $100\pi > 40\%$. Further research should address the robustness properties of the proposed methodology under non lognormal sampling conditions.

Censoring case

The estimated bias correction factors the maximum likelihood estimates described by equations (10) and (11) performed well and were used in all computations involving censored samples. Independently of which method used, bias correction methods are required and necessarily increase the variance of the maximum likelihood estimates.

A comparison between the minimum sample sizes required for the non censoring case and under the presence censoring at different levels of censoring and for several GSDs and several percentage different from the true arithmetic mean was made using table 1 and table 5. The results shows that a 10% and 20% levels of censoring will increase the sample size by at least 15% and 30% respectively with respect to the non censoring case. This is evidence of the fact that a high degree of censoring will necessitate a large sample size across any percentage difference from the true arithmetic mean . As seen in the table, the required sample size at high accuracy levels is much greater than the sample size required at low desired accuracy levels.

The results of Monte Carlo simulation of 95% confidence intervals shows in table 6 indicate further " fine tuning" of the estimator is possible to more exactly estimate the confidence intervals. As seen in the table, the results are conservative and will lead to higher costs.

Acknowledgments

This work is partly based on Dr Perez's Ph.D. dissertation, prepared at Tulane University and supported by scholarships from the Universidad Nacional de Colombia, Colfuturo, the Clinical Epidemiology Unit at Pontificia Universidad Javeriana and grant 1004-94-4401 from the International Clinical Epidemiology Network. The authors are grateful to Dr. C.E. Land for providing a computer program that estimates the exact confidence intervals. They also wish to thank the editor, associate editor and referees for their helpful comments.

REFERENCES

Armstrong, B.G. (1992), "Confidence Intervals for Arithmetic Means of Lognormally Distributed Exposures" American Industrial Hygiene Association Journal, 53, 481-485.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations. " Journal of the Royal Statistical Society, Series B, 26, 211-252.

Cohen. A.C. (1950), "Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. " Annals of Mathematical Statistics, 21. 557-519.

Cohen, A.C. (1959), "Simplified estimators for the normal distribution when samples are singly censored or truncated. " Technometrics, 1, 217-237.

Cohen, A.C. (1961), "Tables for maximum likelihood estimates singly truncated and singly censored samples. "Technometrics, 3, 535-541.

Custer, S.W. (1976), "Correction for bias in maximum likelihood estimators of σ in a right-censored normal distribution. " Communications in Statistics Theory and Methods. Series B, 5, 15-22.

Evans, J.S. , and Hawkins, N.C. (1988), " The Distribution of Student's t-statistic for Small Samples From Lognormal Exposure Distributions. " American Industrial Hygiene Association Journal , 49, 512-515.

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

Greenland, S. (1988) , "On Sample Size and Power Calculations for Studies Using Confidence Intervals. " American Journal of Epidemiology, 128, 231-237.

Hewett, P. (1995), "Sample Size Formulae for Estimating the True Arithmetic Mean or Geometric mean of Lognormal Exposure Distributions" American Industrial Hygiene Association Journal, 56, 219-25.

Kupper, L.L., and Kafner,K.B. (1989), "How Appropriate are Popular Sample Size Formulas?. "The American Statistician, 43, 101-105.

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

Land, C.E. (1972), "An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means. " Technometrics, 14 145-158.

Land, C.E. (1973), "Standard Confidence Limits for Linear Functions of the Normal Mean and Variance." Journal of the American Statistical Association, 68, 960-963.

Land, C.E. (1974), "Confidence Interval Estimation for Means After Data Transformations to Normality ." Journal of the American Statistical Association, 69, 795-802.

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), eds. H.l. Harter, and D.B. Providence, R. , I.: American Mathematical Society. Washington DC. pp. 385-419

Land, C.E. (1988), "Hypothesis Test and Interval Estimates. "Lognormal Distributions. Theory and Applications. eds. E.L. Crow and K. Shimizu. Marcel Decker, Inc. New york.

Rappaport, S.M., Selvin, S., and Roach, S.A. (1988), "A Strategy for Assessing Exposures With Reference to Multiple Limits." Applied Industrial Hygiene, 3, 310-315.

SAS Institute, Inc. (1985), SAS Language/STAT, Version 5., Cary, NC.

Saw, J.G. (1961), "The bias of the maximum likelihood estimates of location and scale parameters given Type II censored normal data. " Biometrika, 48, 448-451.

Schmee, J., Gladstein, D. and Nelson , W.B. (1982), "Exact confidence limits for (log) normal parameters form maximum likelihood estimates and singly censored samples. " General Electric Co., Corporate Research and Development. TIS Report 82CRD244, 1-30.

Schmee, J., Gladstein, D. and Nelson , W.B. (1985), "Confidence limits for parameters of a normal distribution from singly censored samples, using maximum likelihood. " Technometrics, 27, 119-128.

Schneider, H. (1986), " Truncated and censored samples from normal populations", New York and Basel: Marcel Dekker, Inc.

Schneider, H. and Weissfeld, L. (1986), "Inference based on Type II censored samples," Biometrics, 42, 531-536.

Seixas, N.S., Robins, T.G. and Moulton, L.H. (1988), "The use of Geometric and Arithmetic Mean Exposures in Occupational Epidemiology." American Journal of Industrial Medicine, 14, 465-477.

Tiku, M.L. (1978), "Linear regression model with censored observations." Communications in Statistics Theory and Methods, Series A, 7, 1219-1232.